The New Cosmic Universe

Collection edited by: Steven Mellema
Content authors: Steven Mellema, OpenStax, OpenStax Astronomy, and OpenStax University Physics
Based on: The Cosmic Universe <**http://legacy.cnx.org/content/col12169/1.81**>.
Online: <**https://legacy.cnx.org/content/col31908/1.2**>

Collection structure revised: 2020/08/13
PDF Generated: 2020/08/13 13:27:05

Table of Contents

PREFACE

Welcome to *The New Cosmic Universe*. This textbook was written in collaboration with the OpenStax project, whose purpose is to increase student access to high-quality learning materials, while maintaining the highest standards of academic rigor at little to no cost. It was created by re-organizing and editing some material from the OpenStax textbooks *University Physics*[1] and *Astronomy*[2], and by the addition of new material.

About OpenStax

OpenStax is a nonprofit based at Rice University, and it's our mission to improve student access to education. Our first openly licensed college textbook was published in 2012 and our library has since scaled to over 20 books used by hundreds of thousands of students across the globe. Our adaptive learning technology, designed to improve learning outcomes through personalized educational paths, is currently being piloted for K–12 and college. The OpenStax mission is made possible through the generous support of philanthropic foundations. Through these partnerships and with the help of additional low-cost resources from our OpenStax partners, OpenStax is breaking down the most common barriers to learning and empowering students and instructors to succeed.

About *The New Cosmic Universe*

The New Cosmic Universe is based upon the first semester of the four-semester, calculus-based, introductory physics-course sequence at Gustavus Adolphus College. The text has been developed to meet the scope and sequence of that course. The entire four-semester sequence at Gustavus, like virtually all university physics courses, provides a foundation for a career in mathematics, science, or engineering. This book provides a unique combination and ordering of material centered on a theme of astrophysics. It is distinctive in that it:

intersperses material from both "classical" and "modern" physics;

treats the use of calculus in ways that are largely conceptual; and

serves as a one-semester introduction to astrophysics suitable for physics, astronomy or pre-engineering majors.

Motivation for The Cosmic Universe Course

Over the years at Gustavus, the successful retention of good students in the physics major has always been a concern. Certainly, a physics major is one of the most challenging paths through college, and it is definitely not for everyone. As professors of the liberal arts, we are always happy if a student finds another major that is truly their passion. We also understand that not every college student has the mathematical aptitude for our subject. And, to be honest, we also accept the fact that some will leave physics because they do not wish to put in the level of effort required to succeed in such a rigorous major. Our concerns over the years have focused on a group of hard-working, interested students who seem to leave the physics major sometime in the first year. Although these particular students have demonstrated both the ability and the work ethic to succeed in our program, and even though they have yet to find another major interest, they seem to leave for one of three reasons:

1. They do not find the material in the first-year courses to be particularly interesting or relevant in the world of the 21st century.
2. They are bored by the early repetition of material that they recently covered in high school (especially classical mechanics).
3. They feel overwhelmed by the immediate use of calculus, especially because they are taking their first semester of college calculus, while some of their peers in the physics class have had more advanced high-school mathematics preparation.

Scope, Coverage and Organization

Context for The Cosmic Universe

The four-semester, introductory course sequence at Gustavus Adolphus College consists of:

1. The Cosmic Universe

---

1. https://legacy.cnx.org/content/col11994/1.1/
2. https://legacy.cnx.org/content/col11992/1.10/

2. The Mechanical Universe
3. The Electromagnetic Universe
4. The Quantum Universe

The intent of the sequence is to cover 100% of the topical material normally taught in any undergraduate physics program (and found in any good undergraduate textbook). However, we have re-ordered the topics by using a sequence of themes (obvious from the course names). In that way, the "story lines" for each course are coherent and facilitate better student understanding of the underlying physics concepts.

By incorporating the history of ideas, the thematic approaches lead directly to an understanding of the scientific method - not as some dry set of steps, but as an actual, evolving human experience. As teachers at a liberal arts college, we feel that every physics major should also understand how history, science, politics, religion, and ethics interact as parts of that experience.

And, we also feel that it is important for physics majors to encounter, early in their college careers, the major unanswered questions in physics, many of which are part of the study of astrophysics and cosmology - e.g. dark matter and dark energy.

The New Cosmic Universe textbook contains material from various subfields of physics, from classical mechanics to optics to relativity to quantum mechanics. The choice of topics and of the astrophysics theme were made to provide a first-semester experience that is:

Interesting, because it deals with a very active field of current study in physics

New to virtually all of the students, because it involves topics not taught in most high-school physics courses

Rigorously mathematical in its approach, at the level of algebra and introductory calculus, more so than a traditional college textbook in introductory astronomy

Not dependent upon previous fluency in the use of **calculus**

Presented using a **coherent story line**

Calculus in *The New Cosmic Universe*

The fact that The Cosmic Universe course is taken mostly by students in their first semester of college strongly influences our use of calculus in this book. Many introductory calculus-based physics texts will, in an early chapter, derive the equations of one-dimensional kinematics using integrals. For our student audience, where up to 50% are simultaneously enrolled in their first semester of college calculus, such an approach can be discouraging and therefore counterproductive.

Knowing that they are studying (first) limits and (next) derivatives and then (perhaps by the end of the semester) integration influences our use of calculus. We attempt to include calculus, conceptually at first, and hope that its physical significance (of the derivative in particular) and practical applications can enhance the students' understanding of both the physics and the math. As we are fond of asking our students, "Why did Newton invent the calculus in the first place?"

The book is organized as follows:

Chapter 16: Introductory Thermodynamics
Chapter 17: Kinetic Theory
Chapter 18: Nuclear Energy and the Solar System
Chapter 19: Comparative Planetology
Chapter 20: Exoplanets
Chapter 21: The Sun
Chapter 22: Stellar Properties
Chapter 23: Celestial Distances
Chapter 24: Stellar Life Cycles
Chapter 25: The Deaths of Stars
Chapter 26: The Milky Way Galaxy
Chapter 27: Galaxies
Chapter 28: The Evolution and Distribution of Galaxies
Chapter 29: Big Bang Cosmology
Chapter 30: Geometric Optics - Light as Rays
Chapter 31: Image Formation
Chapter 32: Interference
Chapter 33: Diffraction
Chapter 34: Astronomical Instruments

Assessments That Reinforce Key Concepts

Many of the assessments that were built into the OpenStax books *University Physics* and *Astronomy* have been retained.

In-chapter **Examples** generally follow a three-part format of Strategy, Solution, and Significance to emphasize how to approach a problem, how to work with the equations, and how to check and generalize the result. Examples are often followed by **Check Your Understanding** questions and answers to help reinforce for students the important ideas of the examples. **Problem-Solving Strategies** in each chapter break down methods of approaching various types of problems into steps students can follow for guidance. The book also includes exercises at the end of each chapter so students can practice what they've learned.

Conceptual questions do not require calculation but test student learning of the key concepts.

Problems categorized by section test student problem-solving skills and the ability to apply ideas to practical situations.

Additional Problems apply knowledge across the chapter, forcing students to identify what concepts and equations are appropriate for solving given problems. Randomly located throughout the problems are **Unreasonable Results** exercises that ask students to evaluate the answer to a problem and explain why it is not reasonable and what assumptions made might not be correct.

Challenge Problems extend text ideas to interesting but difficult situations.

For Further Exploration. This section offers a list of website and videos so students can delve into topics of interest, whether for their own learning, for homework, extra credit, or papers.

Answers for selected exercises are available in an **Answer Key** at the end of the book.

About the Authors

The New Cosmic Universe

Steven H. Mellema, Gustavus Adolphus College
Dr. Steve Mellema has taught introductory and advanced physics for over 30 years at Gustavus Adolphus College, where he is currently Professor of Physics. He began his teaching career as a Peace Corps volunteer in Malaysia, where he taught both chemistry and physics in secondary schools, worked on curriculum development, and conducted technical training for science teachers. Dr. Mellema has a PhD in Physics from Ohio University, and he was a post-doctoral research associate in nuclear physics at both Ohio University and the University of Wisconsin - Madison, before joining the Gustavus faculty. Dr. Mellema has experience with research in Physics Education, including two Fulbright Professorships in Malaysia, the latest of which involved bringing active-learning methods into university physics classrooms there.

Senior Contributing Authors - *Open Stax University Physics*

Samuel J. Ling, Truman State University

Jeff Sanny, Loyola Marymount University
Bill Moebs, PhD

Contributing Authors - *OpenStax University Physics*

David Anderson, Albion College
Daniel Bowman, Ferrum College
Dedra Demaree, Georgetown University
Gerald Friedman, Santa Fe Community College
Lev Gasparov, University of North Florida
Edw. S. Ginsberg, University of Massachusetts
Alice Kolakowska, University of Memphis
Lee LaRue, Paris Junior College
Mark Lattery, University of Wisconsin
Richard Ludlow, Daniel Webster College
Patrick Motl, Indiana University–Kokomo
Tao Pang, University of Nevada–Las Vegas
Kenneth Podolak, Plattsburgh State University
Takashi Sato, Kwantlen Polytechnic University
David Smith, University of the Virgin Islands
Joseph Trout, Richard Stockton College
Kevin Wheelock, Bellevue College

Senior Contributing Authors - *OpenStax Astronomy*

Andrew Fraknoi, Foothill College
David Morrison, National Aeronautics and Space Administration
Sidney C. Wolff, National Optical Astronomy Observatories (Emeritus)

Contributing Authors - *OpenStax Astronomy*

John Beck, Stanford University
Susan D. Benecchi, Planetary Science Institute
John Bochanski, Rider University
Howard Bond, Pennsylvania State University, Emeritus, Space Telescope Science Institute
Jennifer Carson, Occidental College
Bryan Dunne, University of Illinois at Urbana-Champaign
Martin Elvis, Harvard-Smithsonian Center for Astrophysics
Debra Fischer, Yale University
Heidi Hammel, Association of Universities for Research in Astronomy
Tori Hoehler, NASA Ames Research Center
Douglas Ingram, Texas Christian University
Steven Kawaler, Iowa State University
Lloyd Knox, University of California, Davis
Mark Krumholz, Australian National University
James Lowenthal, Smith College
Siobahn Morgan, University of Northern Iowa
Daniel Perley, California Institute of Technology
Claire Raftery, National Solar Observatory
Deborah Scherrer, retired, Stanford University
Phillip Scherrer, Stanford University
Sanjoy Som, Blue Marble Space Institute of Science, NASA Ames Research Center
Wes Tobin, Indiana University East
William H. Waller, retired, Tufts University, Rockport (MA) Public Schools
Todd Young, Wayne State College

1 | INTRODUCING PHYSICS

Chapter Outline

Introduction

You are hopefully reading this book (and taking the course with which it is associated) because you want to get a mathematically rigorous introduction to the science of physics. To physicists, this means you want to learn to understand how the world works (and how the things in it work). You likely want to do so because your eventual goal is a career in physics, engineering, or a related field.

This book is like no other textbook in terms of its approach to introductory physics. The traditional undergraduate physics course sequence is quite historical. It begins with the 17th and 18th-century development of ideas about motion from Galileo and Newton, continues with the 19th-century study of energy and heat (by Joule and Carnot), moves on to the study of electricity and magnetism (by Faraday and Maxwell), and follows through with the 20th-century ideas of Einstein (especially in relativity) and of Planck, Bohr, Schrodinger and Heisenberg (in quantum physics).

Such an historical approach can have the strength of helping the reader to appreciate the process of science – how each new discovery is built upon those that came before. A typical textbook's ordering of the major subdisciplines in physics might be:

- Classical Mechanics
- Waves and Sound
- Heat and Thermodynamics
- Electricity and Magnetism
- Optics
- Relativity
- Quantum Physics

But, because a rigorous, introductory study of physics can only realistically take place over a protracted period of time (typically three or four semesters), this traditional, topical approach can sometimes lose track of the big picture. There are important ideas that cross multiple subdisciplines, have taken many centuries to develop and, in fact, continue to be refined even in the 21st century.

This book is fashioned around one such big-picture "story line" – our understanding of the Universe. Where did the Universe (and we) come from? Where are we going? It is the oldest and, to us, the most interesting story in the history of humankind. It is also multidisciplinary within physics, involving almost all of the traditional subtopics listed above.

Why begin your study of physics with astrophysics? For one thing, in the 21st century, no branch of physics is more active and dynamic in its pursuit of scientific knowledge. Chances are you may not yet have spent much time in your formal education studying this subject, but you can surely read almost weekly about new discoveries being made about the nature and evolution of the Universe. New planets in astonishing numbers have been detected orbiting distant stars. Some are considered "Earth-like" – are they possibly homes to life? Even some bodies in our own solar system are being revealed to be very unlike our previous ideas about them. The recent NASA New Horizons mission revealed information about the dwarf planet Pluto (see **Figure 1.1**) that has required astronomers to completely rethink the theories of its formation and evolution.

Figure 1.1 In 2015 NASA's New Horizons spacecraft sent back this stunning color image of Pluto's surface. Images like this one have completely overturned astronomers' ideas about the composition and history of Pluto.

Even more recently, NASA's Juno mission making close approaches to Jupiter has made exquisitely precise images of magnetic storms there (see **Figure 1.2**). Those discoveries not only make us re-think our understanding of the planet, but may in fact lead to new models for the basic formation of planetary magnetic fields.

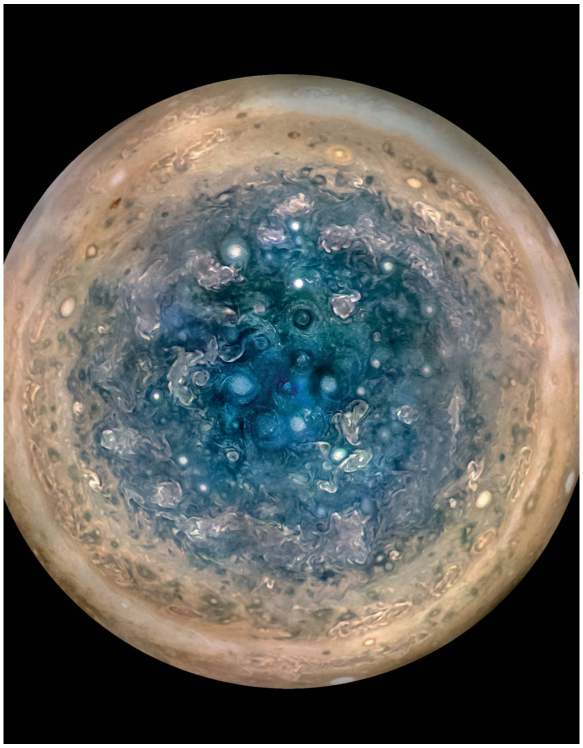


Figure 1.2 In 2017, NASA's Juno spacecraft sent back this image of Jupiter's south pole covered in Earth-sized swirling storms that are densely clustered and rubbing together.

No questions in science are older than those from astronomy. The recorded history of science begins with astronomy. Thousands of years ago, from the West (the ancient Greeks) and the East (the Chinese), we find written records of astronomical data and theories of the Universe. But even before that, it isn't hard to speculate that human beings' earliest questions about their world must have included:

- What causes the regular cycles of the Sun and Moon?
- What are we seeing when we look up in the sky on a clear night?
- Are all those "dots" the same kind of thing? What kind of thing?
- How far away are they?

Indeed, astronomy has been a primary scientific endeavor for well over 2000 years. In the 20^{th} and 21^{st} centuries, the search for answers to these questions has become "astrophysics", involving multiple areas of physics. Our story line will include the study of:

I. Mechanics (the study of motion)
   A. Kinematics (a description of motion)
   B. Dynamics (an explanation of why things move as they do)
      1. Mass, Forces and Newton's Laws
      2. Energy (and its conservation)
         a. Kinetic energy
         b. Potential energy
      3. Momentum (and its conservation)
      4. Angular Momentum (and its conservation)

II. Thermodynamics
   A. Temperature
   B. Heat as transfer of thermal energy
   C. Thermal energy as the internal kinetic energy of molecules
   D. Phase changes with temperature
      1. Liquid-Solid transition: the melting point
      2. Gas-Liquid transition: the boiling point

III. Optics (Everything that we know about what's "out there" comes from studying the light that reaches us here on Earth.)
   A. Geometric (ray) optics (used to construct lenses, mirrors and telescopes)
   B. Physical (wave) optics
   C. Electromagnetic waves and the electromagnetic spectrum – from radio waves to gamma rays
   D. Types of light (continuous sources vs. line sources)
   E. Interference
   F. Diffraction
   G. The Doppler shift

IV. Quantum Physics (After 1900, we learned that the physics that applies at very small scales is not Newtonian.)
   A. The ultimate sources of light are molecules, atoms, electrons, and nuclei.
   B. Are things in our Universe ultimately particles or waves? (Yes!)

As you can see, if we followed this story line through a typical introductory physics textbook, it would take the whole book to complete it – insofar as we can ever say it is "complete". (Perhaps a better way to put it would be "up to its present state of understanding.")

It may sound a bit overdramatic when we say that the ultimate goal of this book (and this course) is to understand the

Universe in its entirety. From an astrophysical point of view, we will work from the inside out – beginning with our own solar system, moving on to consider stars other than our own Sun, and then to study galaxies of stars and what lies beyond, the large-scale structures that form out of clusters of galaxies.

We will also be faced with the fact that telescopes are time machines, i.e. the farther out in space we look at objects, the farther back in time we see them. The term ***cosmology*** means the study of the history of the Universe, from its beginning to now and into the future beyond. Our outward journey, then, will help us to paint a picture of that history.

We will study the evidence for the beginning of our Universe in a “Big Bang”, and trace the evolution of stars, solar systems and galaxies over the roughly 14 billion years of its existence. Through our understanding of the processes that have shaped the past and the present, we will be able to examine and discuss possible fates (or futures) of the Universe. Does the future depend upon unseen “stuff” referred to as dark matter? Does it involve an unexplained repulsive force (like an anti-gravity) called dark energy?

Whether or not the preceding material is enough to convince you that a course in astrophysics is an interesting place to begin your study of physics, we hope you will leave this chapter with one important take-home thought. We paraphrase the late astronomer Carl Sagan who, through his ground-breaking television series, *Cosmos*, did so much to popularize and explain astrophysics to the general public. Just think about this: “Every atom inside your body, at this instant, once lived inside a star.”

Does that statement make sense to you? Because it’s absolutely a true, scientific fact. By the end of “The New Cosmic Universe”, you will know both why that is true and how we know it to be so.

1.1 | Introducing Astrophysics

Learning Objectives

By the end of this section you will be able to:

- List some of the levels of structure in the Universe that are studied in astrophysics.

Physics + Astronomy = Astrophysics

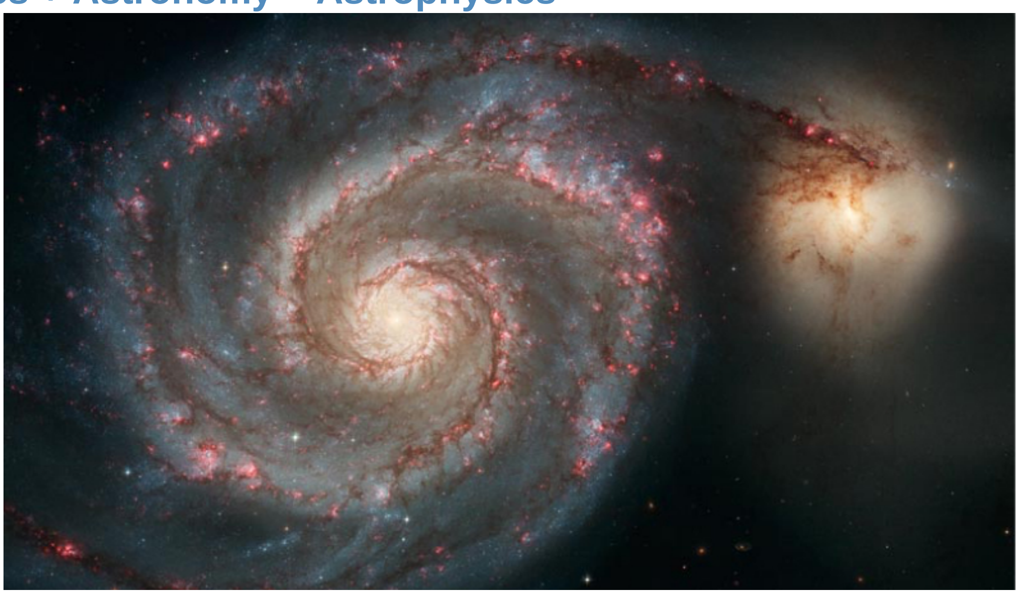


Figure 1.3 This image might be showing any number of things. It might be a whirlpool in a tank of water or perhaps a collage of paint and shiny beads done for art class. Without knowing the size of the object in units we all recognize, such as meters or inches, it is difficult to know what we’re looking at. In fact, this image shows the Whirlpool Galaxy (and its companion galaxy), which is about 60,000 light-years in diameter (about $6 \times 10^{17}\,\text{km}$ across). (credit: S. Beckwith (STScI) Hubble Heritage Team, (STScI/AURA), ESA, NASA)

What's It All About?

What do you make of this picture?

As noted in the figure caption, this image is of the Whirlpool Galaxy. Galaxies are as immense as atoms are small, yet the same laws of physics describe both, along with all the rest of nature—an indication of the underlying unity in the universe. The laws of physics are surprisingly few, implying an underlying simplicity to nature’s apparent complexity. In this text, you learn about the laws of physics. Galaxies and atoms may seem far removed from your daily life, but as you begin to explore this broad-ranging subject, you may soon come to realize that physics plays a much larger role in your life than you first thought, no matter your life goals or career choice.

Figure 1.4 These two interacting islands of stars (galaxies) are so far away that their light takes hundreds of millions of years to reach us on Earth (photographed with the Hubble Space Telescope). (credit: modification of work by NASA, ESA, the Hubble Heritage (STScl/AURA)-ESA/Hubble Collaboration, and K. Noll (STScl))

We invite you to come along on a series of voyages to explore the universe as astronomers and physicists understand it today. Beyond Earth are vast and magnificent realms full of objects that have no counterpart on our home planet. Nevertheless, we hope to show you that the evolution of the universe has been directly responsible for your presence on Earth today.

Along your journey, you will encounter:

- a canyon system so large that, on Earth, it would stretch from Los Angeles to Washington, DC (**Figure 1.5**).

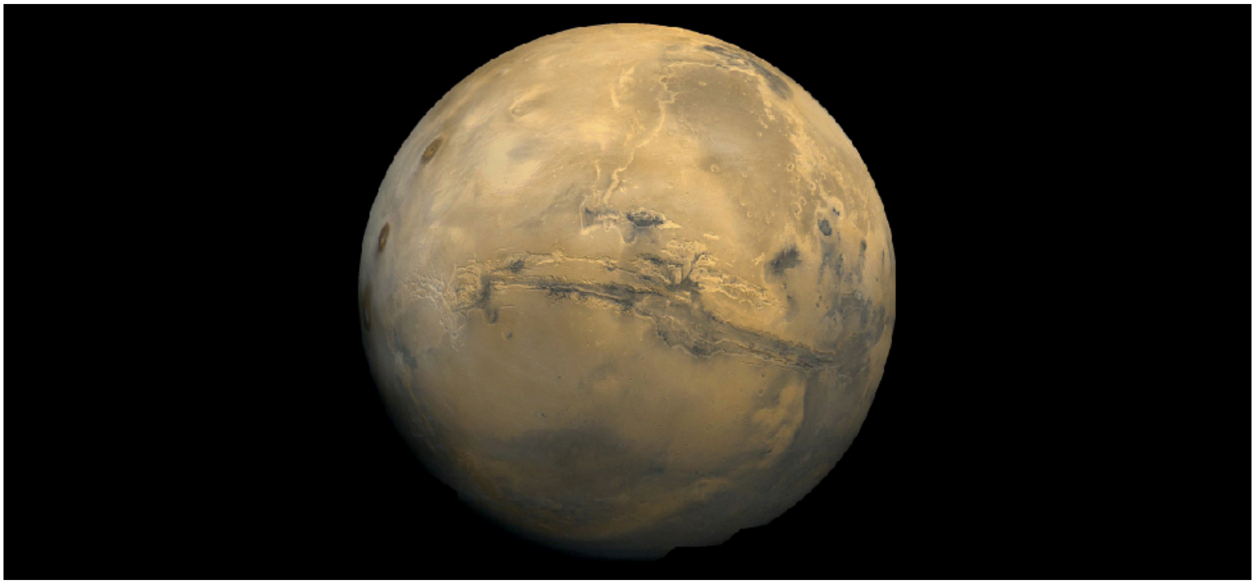


Figure 1.5 This image of Mars is centered on the Valles Marineris (Mariner Valley) complex of canyons, which is as long as the United States is wide. (credit: modification of work by NASA)

- a crater and other evidence on Earth that tell us that the dinosaurs (and many other creatures) died because of a cosmic collision.
- a tiny moon whose gravity is so weak that one good throw from its surface could put a baseball into orbit.
- a collapsed star so dense that to duplicate its interior we would have to squeeze every human being on Earth into a single raindrop.
- exploding stars whose violent end could wipe clean all of the life-forms on a planet orbiting a neighboring star (**Figure 1.6**).

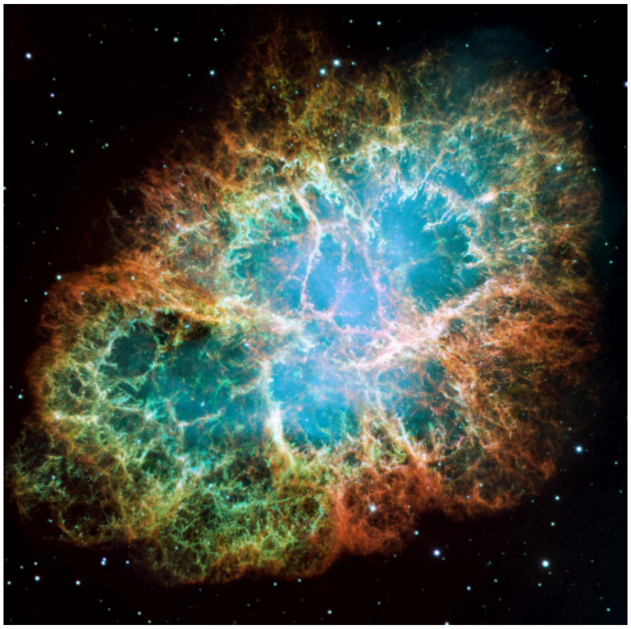


Figure 1.6 We observe the remains of a star that was seen to explode in our skies in 1054 (and was, briefly, bright enough to be visible during the daytime). Today, the remnant is called the Crab Nebula and its central region is seen here. Such exploding stars are crucial to the development of life in the universe. (credit: NASA, ESA, J. Hester (Arizona State University))

- a “cannibal galaxy” that has already consumed a number of its smaller galaxy neighbors and is not yet finished finding new victims.

- a radio echo that is the faint but unmistakable signal of the creation event for our universe.

Such discoveries are what make astronomy such an exciting field for scientists and many others—but you will explore much more than just the objects in our universe and the latest discoveries about them. We will pay equal attention to the *process* by which we have come to understand the realms beyond Earth and the tools we use to increase that understanding.

We gather information about the cosmos from the messages the universe sends our way. Because the stars are the fundamental building blocks of the universe, decoding the message of starlight has been a central challenge and triumph of modern astronomy. By the time you have finished reading this text, you will know a bit about how to read that message and how to understand what it is telling us.

1.2 | The Scope and Scale of Physics

Learning Objectives

By the end of this section, you will be able to:

- Describe the scope of physics.
- Calculate the order of magnitude of a quantity.
- Compare measurable length, mass, and timescales quantitatively.
- Describe the relationships among models, theories, and laws.

Physics is devoted to the understanding of all natural phenomena. In physics, we try to understand physical phenomena at all scales—from the world of subatomic particles to the entire universe. Despite the breadth of the subject, the various subfields of physics share a common core. The same basic training in physics will prepare you to work in any area of physics and the related areas of science and engineering. In this section, we investigate the scope of physics; the scales of length, mass, and time over which the laws of physics have been shown to be applicable; and the process by which science in general, and physics in particular, operates.

The Scope of Physics

Take another look at the **image of the Whirlpool galaxy**. That galaxy contains billions of individual stars as well as huge clouds of gas and dust. Its companion galaxy is also visible to the right. This pair of galaxies lies a staggering billion trillion miles $(1.4 \times 10^{21}\ \text{mi})$ from our own galaxy (which is called the *Milky Way*). The stars and planets that make up the Whirlpool Galaxy might seem to be the furthest thing from most people's everyday lives, but the Whirlpool is a great starting point to think about the forces that hold the universe together. The forces that cause the Whirlpool Galaxy to act as it does are thought to be the same forces we contend with here on Earth, whether we are planning to send a rocket into space or simply planning to raise the walls for a new home. The gravity that causes the stars of the Whirlpool Galaxy to rotate and revolve is thought to be the same as what causes water to flow over hydroelectric dams here on Earth. When you look up at the stars, realize the forces out there are the same as the ones here on Earth. Through a study of physics, you may gain a greater understanding of the interconnectedness of everything we can see and know in this universe.

Think, now, about all the technological devices you use on a regular basis. Computers, smartphones, global positioning systems (GPSs), MP3 players, and satellite radio might come to mind. Then, think about the most exciting modern technologies you have heard about in the news, such as trains that levitate above tracks, "invisibility cloaks" that bend light around them, and microscopic robots that fight cancer cells in our bodies. All these groundbreaking advances, commonplace or unbelievable, rely on the principles of physics. Aside from playing a significant role in technology, professionals such as engineers, pilots, physicians, physical therapists, electricians, and computer programmers apply physics concepts in their daily work. For example, a pilot must understand how wind forces affect a flight path; a physical therapist must understand how the muscles in the body experience forces as they move and bend. As you will learn in this text, the principles of physics are propelling new, exciting technologies, and these principles are applied in a wide range of careers.

The underlying order of nature makes science in general, and physics in particular, interesting and enjoyable to study. For example, what do a bag of chips and a car battery have in common? Both contain energy that can be converted to other forms. The law of conservation of energy (which says that energy can change form but is never lost) ties together such topics as food calories, batteries, heat, light, and watch springs. Understanding this law makes it easier to learn about the various forms energy takes and how they relate to one another. Apparently unrelated topics are connected through broadly applicable physical laws, permitting an understanding beyond just the memorization of lists of facts.

Science consists of theories and laws that are the general truths of nature, as well as the body of knowledge they encompass.

Scientists are continuously trying to expand this body of knowledge and to perfect the expression of the laws that describe it. **Physics**, which comes from the Greek *phúsis*, meaning "nature," is concerned with describing the interactions of energy, matter, space, and time to uncover the fundamental mechanisms that underlie every phenomenon. This concern for describing the basic phenomena in nature essentially defines the *scope of physics*.

Physics aims to understand the world around us at the most basic level. It emphasizes the use of a small number of quantitative laws to do this, which can be useful to other fields pushing the performance boundaries of existing technologies. Consider a smartphone (**Figure 1.7**). Physics describes how electricity interacts with the various circuits inside the device. This knowledge helps engineers select the appropriate materials and circuit layout when building a smartphone. Knowledge of the physics underlying these devices is required to shrink their size or increase their processing speed. Or, think about a GPS. Physics describes the relationship between the speed of an object, the distance over which it travels, and the time it takes to travel that distance. When you use a GPS in a vehicle, it relies on physics equations to determine the travel time from one location to another.

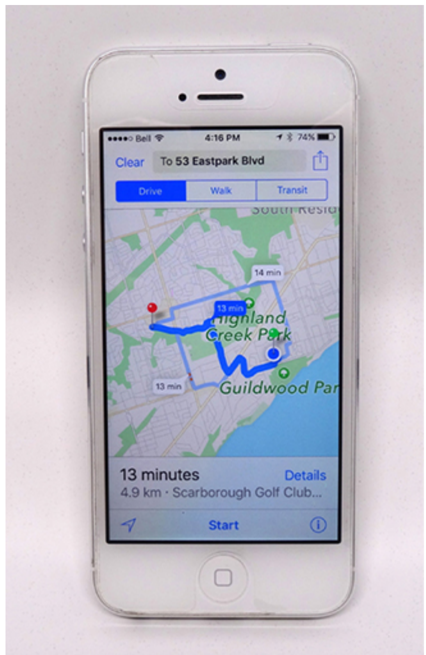


Figure 1.7 The Apple iPhone is a common smartphone with a GPS function. Physics describes the way that electricity flows through the circuits of this device. Engineers use their knowledge of physics to construct an iPhone with features that consumers will enjoy. One specific feature of an iPhone is the GPS function. A GPS uses physics equations to determine the drive time between two locations on a map.

Knowledge of physics is useful in everyday situations as well as in nonscientific professions. It can help you understand how microwave ovens work, why metals should not be put into them, and why they might affect pacemakers. Physics allows you to understand the hazards of radiation and to evaluate these hazards rationally and more easily. Physics also explains the reason why a black car radiator helps remove heat in a car engine, and it explains why a white roof helps keep the inside of a house cool. Similarly, the operation of a car's ignition system as well as the transmission of electrical signals throughout our body's nervous system are much easier to understand when you think about them in terms of basic physics.

Physics is a key element of many important disciplines and contributes directly to others. Chemistry, for example—since it deals with the interactions of atoms and molecules—has close ties to atomic and molecular physics. Most branches of engineering are concerned with designing new technologies, processes, or structures within the constraints set by the laws of physics. In architecture, physics is at the heart of structural stability and is involved in the acoustics, heating, lighting, and cooling of buildings. Parts of geology rely heavily on physics, such as radioactive dating of rocks, earthquake analysis, and heat transfer within Earth. Some disciplines, such as biophysics and geophysics, are hybrids of physics and other disciplines.

Physics has many applications in the biological sciences. On the microscopic level, it helps describe the properties of cells and their environments. On the macroscopic level, it explains the heat, work, and power associated with the human body and its various organ systems. Physics is involved in medical diagnostics, such as radiographs, magnetic resonance imaging, and ultrasonic blood flow measurements. Medical therapy sometimes involves physics directly; for example, cancer radiotherapy uses ionizing radiation. Physics also explains sensory phenomena, such as how musical instruments make sound, how the eye detects color, and how lasers transmit information.

It is not necessary to study all applications of physics formally. What is most useful is knowing the basic laws of physics and developing skills in the analytical methods for applying them. The study of physics also can improve your problem-solving skills. Furthermore, physics retains the most basic aspects of science, so it is used by all the sciences, and the study of physics makes other sciences easier to understand.

The Scale of Physics

From the discussion so far, it should be clear that to accomplish your goals in any of the various fields within the natural sciences and engineering, a thorough grounding in the laws of physics is necessary. The reason for this is simply that the laws of physics govern everything in the observable universe at all measurable scales of length, mass, and time. Now, that is easy enough to say, but to come to grips with what it really means, we need to get a little bit quantitative. So, before surveying the various scales that physics allows us to explore, let's first look at the concept of "order of magnitude," which we use to come to terms with the vast ranges of length, mass, and time that we consider in this text (**Figure 1.8**).

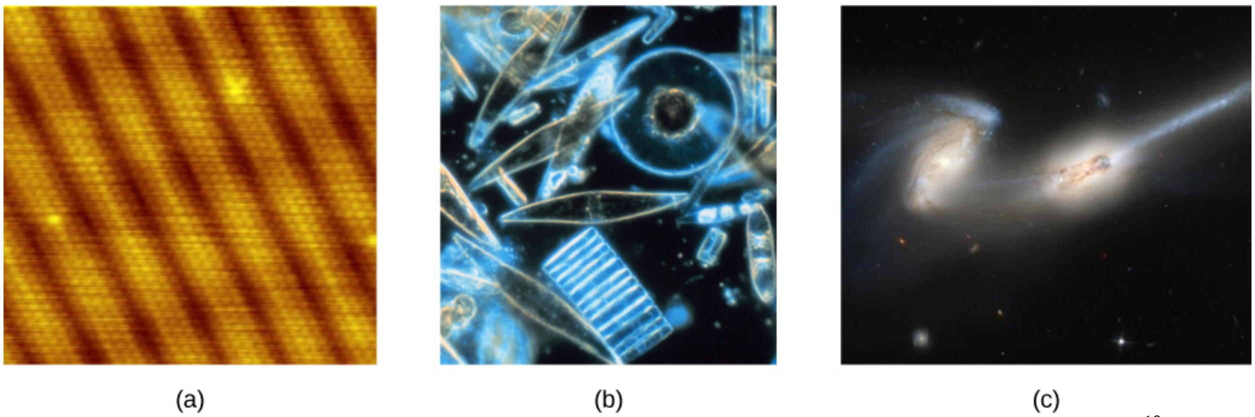


Figure 1.8 (a) Using a scanning tunneling microscope, scientists can see the individual atoms (diameters around 10^{-10} m) that compose this sheet of gold. (b) Tiny phytoplankton swim among crystals of ice in the Antarctic Sea. They range from a few micrometers (1 μm is 10^{-6} m) to as much as 2 mm (1 mm is 10^{-3} m) in length. (c) These two colliding galaxies, known as NGC 4676A (right) and NGC 4676B (left), are nicknamed "The Mice" because of the tail of gas emanating from each one. They are located 300 million light-years from Earth in the constellation Coma Berenices. Eventually, these two galaxies will merge into one. (credit a: modification of work by Erwinrossen; credit b: modification of work by Prof. Gordon T. Taylor, Stony Brook University; NOAA Corps Collections; credit c: modification of work by NASA, H. Ford (JHU), G. Illingworth (UCSC/LO), M. Clampin (STScI), G. Hartig (STScI), the ACS Science Team, and ESA)

Order of magnitude

The **order of magnitude** of a number is the power of 10 that most closely approximates it. Thus, the order of magnitude refers to the scale (or size) of a value. Each power of 10 represents a different order of magnitude. For example, $10^1, 10^2, 10^3,$ and so forth, are all different orders of magnitude, as are $10^0 = 1, 10^{-1}, 10^{-2},$ and $10^{-3}.$ To find the order of magnitude of a number, take the base-10 logarithm of the number and round it to the nearest integer, then the order of magnitude of the number is simply the resulting power of 10. For example, the order of magnitude of 800 is 10^3 because $\log_{10} 800 \approx 2.903,$ which rounds to 3. Similarly, the order of magnitude of 450 is 10^3 because $\log_{10} 450 \approx 2.653,$ which rounds to 3 as well. Thus, we say the numbers 800 and 450 are of the same order of magnitude: 10^3. However, the order of magnitude of 250 is 10^2 because $\log_{10} 250 \approx 2.397,$ which rounds to 2.

An equivalent but quicker way to find the order of magnitude of a number is first to write it in scientific notation and then check to see whether the first factor is greater than or less than $\sqrt{10} = 10^{0.5} \approx 3.$ The idea is that $\sqrt{10} = 10^{0.5}$ is halfway between $1 = 10^0$ and $10 = 10^1$ on a log base-10 scale. Thus, if the first factor is less than $\sqrt{10},$ then we round it down to 1 and the order of magnitude is simply whatever power of 10 is required to write the number in scientific notation. On the other hand, if the first factor is greater than $\sqrt{10},$ then we round it up to 10 and the order of magnitude is one power of 10 higher than the power needed to write the number in scientific notation. For example, the number 800 can be written in scientific notation as $8 \times 10^2.$ Because 8 is bigger than $\sqrt{10} \approx 3,$ we say the order of magnitude of 800 is $10^{2+1} = 10^3.$

The number 450 can be written as 4.5×10^2, so its order of magnitude is also 10^3 because 4.5 is greater than 3. However, 250 written in scientific notation is 2.5×10^2 and 2.5 is less than 3, so its order of magnitude is 10^2.

The order of magnitude of a number is designed to be a ballpark estimate for the scale (or size) of its value. It is simply a way of rounding numbers consistently to the nearest power of 10. This makes doing rough mental math with very big and very small numbers easier. For example, the diameter of a hydrogen atom is on the order of 10^{-10} m, whereas the diameter of the Sun is on the order of 10^9 m, so it would take roughly $10^9/10^{-10} = 10^{19}$ hydrogen atoms to stretch across the diameter of the Sun. This is much easier to do in your head than using the more precise values of $1.06 \times 10^{-10}\,\text{m}$ for a hydrogen atom diameter and $1.39 \times 10^9\,\text{m}$ for the Sun's diameter, to find that it would take 1.31×10^{19} hydrogen atoms to stretch across the Sun's diameter. In addition to being easier, the rough estimate is also nearly as informative as the precise calculation.

Known ranges of length, mass, and time

The vastness of the universe and the breadth over which physics applies are illustrated by the wide range of examples of known lengths, masses, and times (given as orders of magnitude) in **Figure 1.9**. Examining this table will give you a feeling for the range of possible topics in physics and numerical values. A good way to appreciate the vastness of the ranges of values in **Figure 1.9** is to try to answer some simple comparative questions, such as the following:

- How many hydrogen atoms does it take to stretch across the diameter of the Sun?
  (Answer: 10^9 m/10^{-10} m = 10^{19} hydrogen atoms)
- How many protons are there in a bacterium?
  (Answer: 10^{-15} kg/10^{-27} kg = 10^{12} protons)
- How many floating-point operations can a supercomputer do in 1 day?
  (Answer: 10^5 s/10^{-17} s = 10^{22} floating-point operations)

In studying **Figure 1.9**, take some time to come up with similar questions that interest you and then try answering them. Doing this can breathe some life into almost any table of numbers.

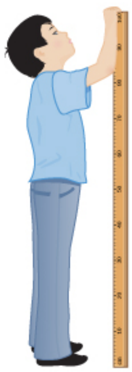
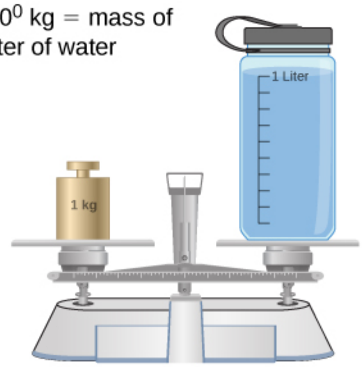
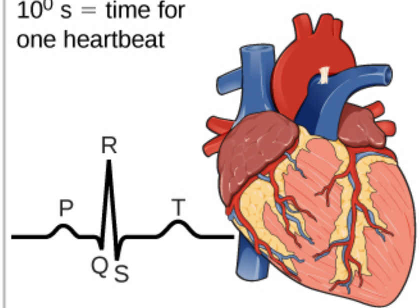
Length in Meters (m)	Masses in Kilograms (kg)	Time in Seconds (s)
10^{-15} m = diameter of proton	10^{-30} kg = mass of electron	10^{-22} s = mean lifetime of very unstable nucleus
10^{-14} m = diameter of large nucleus	10^{-27} kg = mass of proton	10^{-17} s = time for single floating-point operation in a supercomputer
10^{-10} m = diameter of hydrogen atom	10^{-15} kg = mass of bacterium	10^{-15} s = time for one oscillation of visible light
10^{-7} m = diameter of typical virus	10^{-5} kg = mass of mosquito	10^{-13} s = time for one vibration of an atom in a solid
10^{-2} m = pinky fingernail width	10^{-2} kg = mass of hummingbird	10^{-3} s = duration of a nerve impulse
10^{0} m = height of 4 year old child 	10^{0} kg = mass of liter of water 	10^{0} s = time for one heartbeat 
10^{2} m = length of football field	10^{2} kg = mass of person	10^{5} s = one day
10^{7} m = diameter of Earth	10^{19} kg = mass of atmosphere	10^{7} s = one year
10^{13} m = diameter of solar system	10^{22} kg = mass of Moon	10^{9} s = human lifetime
10^{16} m = distance light travels in a year (one light-year)	10^{25} kg = mass of Earth	10^{11} s = recorded human history
10^{21} m = Milky Way diameter	10^{30} kg = mass of Sun	10^{17} s = age of Earth
10^{26} m = distance to edge of observable universe	10^{53} kg = upper limit on mass of known universe	10^{18} s = age of the universe

Figure 1.9 This table shows the orders of magnitude of length, mass, and time.

Visit **this site (https://openstaxcollege.org/l/21scaleuniv)** to explore interactively the vast range of length scales in our universe. Scroll down and up the scale to view hundreds of organisms and objects, and click on the individual objects to learn more about each one.

Building Models

How did we come to know the laws governing natural phenomena? What we refer to as the laws of nature are concise descriptions of the universe around us. They are human statements of the underlying laws or rules that all natural processes follow. Such laws are intrinsic to the universe; humans did not create them and cannot change them. We can only discover and understand them. Their discovery is a very human endeavor, with all the elements of mystery, imagination, struggle, triumph, and disappointment inherent in any creative effort (**Figure 1.10**). The cornerstone of discovering natural laws is observation; scientists must describe the universe as it is, not as we imagine it to be.

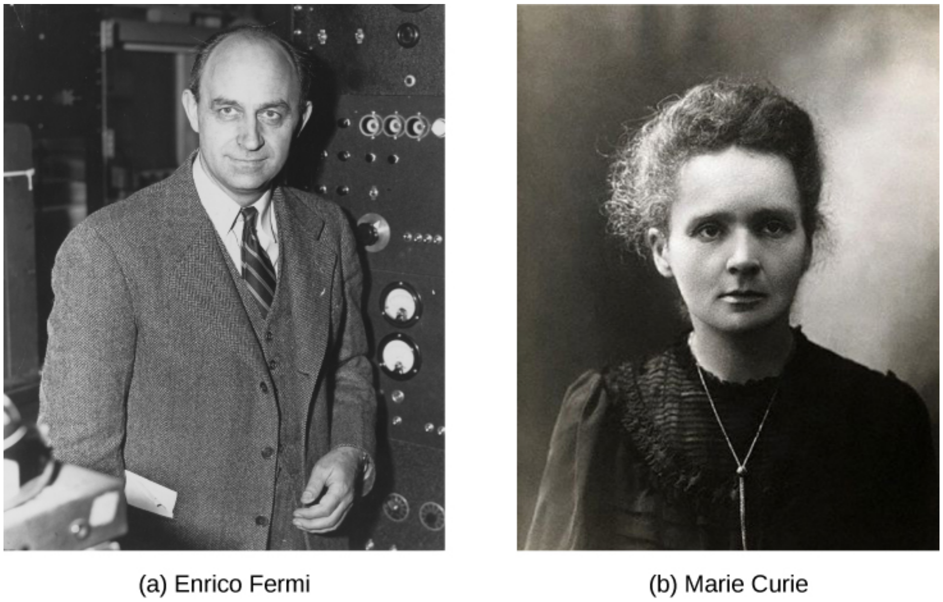


Figure 1.10 (a) Enrico Fermi (1901–1954) was born in Italy. On accepting the Nobel Prize in Stockholm in 1938 for his work on artificial radioactivity produced by neutrons, he took his family to America rather than return home to the government in power at the time. He became an American citizen and was a leading participant in the Manhattan Project. (b) Marie Curie (1867–1934) sacrificed monetary assets to help finance her early research and damaged her physical well-being with radiation exposure. She is the only person to win Nobel prizes in both physics and chemistry. One of her daughters also won a Nobel Prize. (credit a: United States Department of Energy)

A **model** is a representation of something that is often too difficult (or impossible) to display directly. Although a model is justified by experimental tests, it is only accurate in describing certain aspects of a physical system. An example is the Bohr model of single-electron atoms, in which the electron is pictured as orbiting the nucleus, analogous to the way planets orbit the Sun (**Figure 1.11**). We cannot observe electron orbits directly, but the mental image helps explain some of the observations we can make, such as the emission of light from hot gases (atomic spectra). However, other observations show that the picture in the Bohr model is not really what atoms look like. The model is "wrong," but is still useful for some purposes. Physicists use models for a variety of purposes. For example, models can help physicists analyze a scenario and perform a calculation or models can be used to represent a situation in the form of a computer simulation. Ultimately, however, the results of these calculations and simulations need to be double-checked by other means—namely, observation and experimentation.

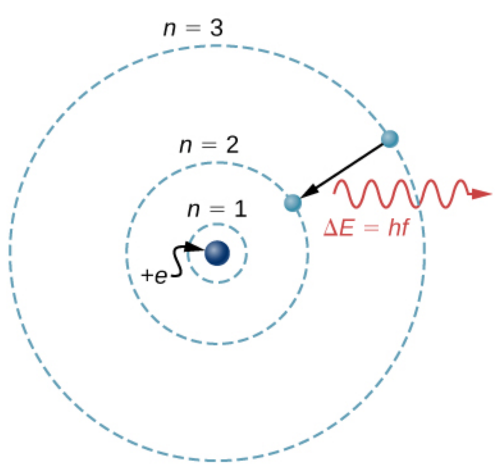


Figure 1.11 What is a model? The Bohr model of a single-electron atom shows the electron orbiting the nucleus in one of several possible circular orbits. Like all models, it captures some, but not all, aspects of the physical system.

The word *theory* means something different to scientists than what is often meant when the word is used in everyday conversation. In particular, to a scientist a theory is not the same as a "guess" or an "idea" or even a "hypothesis." The phrase "it's just a theory" seems meaningless and silly to scientists because science is founded on the notion of theories. To a scientist, a **theory** is a testable explanation for patterns in nature supported by scientific evidence and verified multiple times by various groups of researchers. Some theories include models to help visualize phenomena whereas others do not. Newton's theory of gravity, for example, does not require a model or mental image, because we can observe the objects directly with our own senses. The kinetic theory of gases, on the other hand, is a model in which a gas is viewed as being composed of atoms and molecules. Atoms and molecules are too small to be observed directly with our senses—thus, we picture them mentally to understand what the instruments tell us about the behavior of gases. Although models are meant only to describe certain aspects of a physical system accurately, a theory should describe all aspects of any system that falls within its domain of applicability. In particular, any experimentally testable implication of a theory should be verified. If an experiment ever shows an implication of a theory to be false, then the theory is either thrown out or modified suitably (for example, by limiting its domain of applicability).

A **law** uses concise language to describe a generalized pattern in nature supported by scientific evidence and repeated experiments. Often, a law can be expressed in the form of a single mathematical equation. Laws and theories are similar in that they are both scientific statements that result from a tested hypothesis and are supported by scientific evidence. However, the designation *law* is usually reserved for a concise and very general statement that describes phenomena in nature, such as the law that energy is conserved during any process, or Newton's second law of motion, which relates force (F), mass (m), and acceleration (a) by the simple equation $F = ma$. A theory, in contrast, is a less concise statement of observed behavior. For example, the theory of evolution and the theory of relativity cannot be expressed concisely enough to be considered laws. The biggest difference between a law and a theory is that a theory is much more complex and dynamic. A law describes a single action whereas a theory explains an entire group of related phenomena. Less broadly applicable statements are usually called principles (such as Pascal's principle, which is applicable only in fluids), but the distinction between laws and principles often is not made carefully.

The models, theories, and laws we devise sometimes imply the existence of objects or phenomena that are as yet unobserved. These predictions are remarkable triumphs and tributes to the power of science. It is the underlying order in the universe that enables scientists to make such spectacular predictions. However, if experimentation does not verify our predictions, then the theory or law is wrong, no matter how elegant or convenient it is. Laws can never be known with absolute certainty because it is impossible to perform every imaginable experiment to confirm a law for every possible scenario. Physicists operate under the assumption that all scientific laws and theories are valid until a counterexample is observed. If a good-quality, verifiable experiment contradicts a well-established law or theory, then the law or theory must be modified or overthrown completely.

The study of science in general, and physics in particular, is an adventure much like the exploration of an uncharted ocean. Discoveries are made; models, theories, and laws are formulated; and the beauty of the physical universe is made more sublime for the insights gained.

1.3 | Units and Standards

Learning Objectives

By the end of this section, you will be able to:

- Describe how SI base units are defined.
- Describe how derived units are created from base units.
- Express quantities given in SI units using metric prefixes.

As we saw previously, the range of objects and phenomena studied in physics is immense. From the incredibly short lifetime of a nucleus to the age of Earth, from the tiny sizes of subnuclear particles to the vast distance to the edges of the known universe, from the force exerted by a jumping flea to the force between Earth and the Sun, there are enough factors of 10 to challenge the imagination of even the most experienced scientist. Giving numerical values for physical quantities and equations for physical principles allows us to understand nature much more deeply than qualitative descriptions alone. To comprehend these vast ranges, we must also have accepted units in which to express them. We shall find that even in the potentially mundane discussion of meters, kilograms, and seconds, a profound simplicity of nature appears: all physical quantities can be expressed as combinations of only seven base physical quantities.

We define a **physical quantity** either by specifying how it is measured or by stating how it is calculated from other measurements. For example, we might define distance and time by specifying methods for measuring them, such as using a meter stick and a stopwatch. Then, we could define *average speed* by stating that it is calculated as the total distance traveled divided by time of travel.

Measurements of physical quantities are expressed in terms of **units**, which are standardized values. For example, the length of a race, which is a physical quantity, can be expressed in units of meters (for sprinters) or kilometers (for distance runners). Without standardized units, it would be extremely difficult for scientists to express and compare measured values in a meaningful way (**Figure 1.12**).

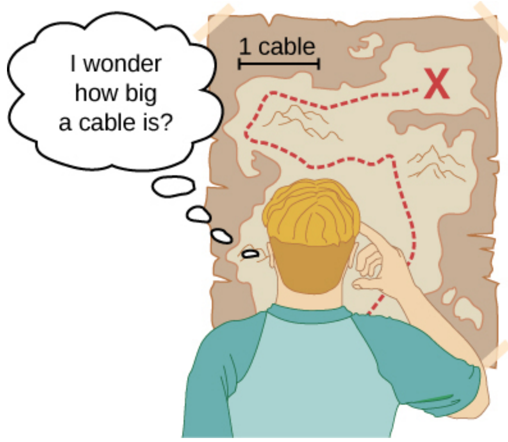


Figure 1.12 Distances given in unknown units are maddeningly useless.

Two major systems of units are used in the world: **SI units** (for the French *Système International d'Unités*), also known as the *metric system*, and **English units** (also known as the *customary* or *imperial system*). English units were historically used in nations once ruled by the British Empire and are still widely used in the United States. English units may also be referred to as the *foot–pound–second* (fps) system, as opposed to the *centimeter–gram–second* (cgs) system. You may also encounter the term *SAE units*, named after the Society of Automotive Engineers. Products such as fasteners and automotive tools (for example, wrenches) that are measured in inches rather than metric units are referred to as *SAE fasteners* or *SAE wrenches*.

Virtually every other country in the world (except the United States) now uses SI units as the standard. The metric system is also the standard system agreed on by scientists and mathematicians.

SI Units: Base and Derived Units

In any system of units, the units for some physical quantities must be defined through a measurement process. These are called the **base quantities** for that system and their units are the system's **base units**. All other physical quantities can then be expressed as algebraic combinations of the base quantities. Each of these physical quantities is then known as a **derived quantity** and each unit is called a **derived unit**. The choice of base quantities is somewhat arbitrary, as long as they are independent of each other and all other quantities can be derived from them. Typically, the goal is to choose physical quantities that can be measured accurately to a high precision as the base quantities. The reason for this is simple. Since the derived units can be expressed as algebraic combinations of the base units, they can only be as accurate and precise as the base units from which they are derived.

Based on such considerations, the International Standards Organization recommends using seven base quantities, which form the International System of Quantities (ISQ). These are the base quantities used to define the SI base units. **Table 1.1** lists these seven ISQ base quantities and the corresponding SI base units.

ISQ Base Quantity	SI Base Unit
Length	meter (m)
Mass	kilogram (kg)
Time	second (s)

Table 1.1 ISQ Base Quantities and Their SI Units

ISQ Base Quantity	SI Base Unit
Electrical current	ampere (A)
Thermodynamic temperature	kelvin (K)
Amount of substance	mole (mol)
Luminous intensity	candela (cd)

Table 1.1 ISQ Base Quantities and Their SI Units

You are probably already familiar with some derived quantities that can be formed from the base quantities in **Table 1.1**. For example, the geometric concept of area is always calculated as the product of two lengths. Thus, area is a derived quantity that can be expressed in terms of SI base units using square meters $(\text{m} \times \text{m} = \text{m}^2)$. Similarly, volume is a derived quantity that can be expressed in cubic meters (m^3). Speed is length per time; so in terms of SI base units, we could measure it in meters per second (m/s). Volume mass density (or just density) is mass per volume, which is expressed in terms of SI base units such as kilograms per cubic meter (kg/m^3). Angles can also be thought of as derived quantities because they can be defined as the ratio of the arc length subtended by two radii of a circle to the radius of the circle. This is how the radian is defined. Depending on your background and interests, you may be able to come up with other derived quantities, such as the mass flow rate (kg/s) or volume flow rate (m^3/s) of a fluid, electric charge $(\text{A} \cdot \text{s})$, mass flux density $[\text{kg}/(\text{m}^2 \cdot \text{s})]$, and so on. We will see many more examples throughout this text. For now, the point is that every physical quantity can be derived from the seven base quantities in **Table 1.1**, and the units of every physical quantity can be derived from the seven SI base units.

For the most part, we use SI units in this text. Non-SI units are used in a few applications in which they are in very common use, such as the measurement of temperature in degrees Celsius $(°\text{C})$, the measurement of fluid volume in liters (L), and the measurement of energies of elementary particles in electron-volts (eV). Whenever non-SI units are discussed, they are tied to SI units through conversions. For example, 1 L is $10^{-3}\ \text{m}^3$.

Check out a comprehensive source of information on **SI units (https://openstaxcollege.org/l/21SIUnits)** at the National Institute of Standards and Technology (NIST) Reference on Constants, Units, and Uncertainty.

Units of Time, Length, and Mass: The Second, Meter, and Kilogram

The initial chapters in this textbook are concerned with mechanics, fluids, and waves. In these subjects all pertinent physical quantities can be expressed in terms of the base units of length, mass, and time. Therefore, we now turn to a discussion of these three base units, leaving discussion of the others until they are needed later.

The second

The SI unit for time, the **second** (abbreviated s), has a long history. For many years it was defined as 1/86,400 of a mean solar day. More recently, a new standard was adopted to gain greater accuracy and to define the second in terms of a nonvarying or constant physical phenomenon (because the solar day is getting longer as a result of the very gradual slowing of Earth's rotation). Cesium atoms can be made to vibrate in a very steady way, and these vibrations can be readily observed and counted. In 1967, the second was redefined as the time required for 9,192,631,770 of these vibrations to occur (**Figure 1.13**). Note that this may seem like more precision than you would ever need, but it isn't—GPSs rely on the precision of atomic clocks to be able to give you turn-by-turn directions on the surface of Earth, far from the satellites broadcasting their location.

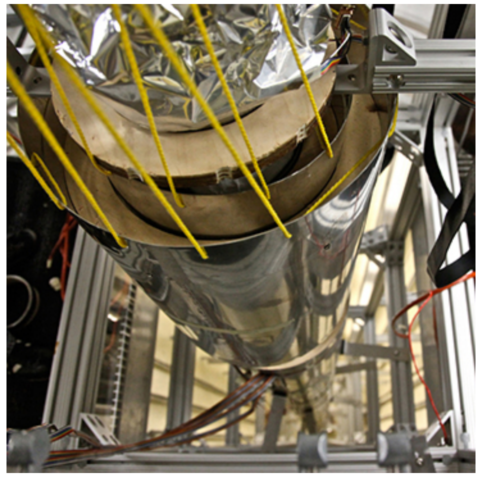


Figure 1.13 An atomic clock such as this one uses the vibrations of cesium atoms to keep time to a precision of better than a microsecond per year. The fundamental unit of time, the second, is based on such clocks. This image looks down from the top of an atomic fountain nearly 30 feet tall. (credit: Steve Jurvetson)

The meter

The SI unit for length is the **meter** (abbreviated m); its definition has also changed over time to become more precise. The meter was first defined in 1791 as 1/10,000,000 of the distance from the equator to the North Pole. This measurement was improved in 1889 by redefining the meter to be the distance between two engraved lines on a platinum–iridium bar now kept near Paris. By 1960, it had become possible to define the meter even more accurately in terms of the wavelength of light, so it was again redefined as 1,650,763.73 wavelengths of orange light emitted by krypton atoms. In 1983, the meter was given its current definition (in part for greater accuracy) as the distance light travels in a vacuum in 1/299,792,458 of a second (**Figure 1.14**). This change came after knowing the speed of light to be exactly 299,792,458 m/s. The length of the meter will change if the speed of light is someday measured with greater accuracy.

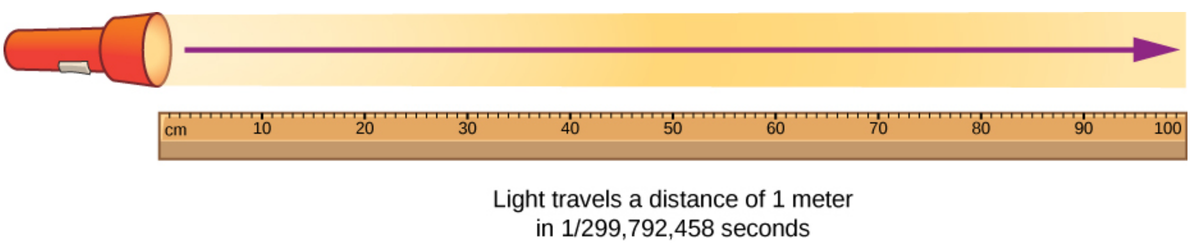


Figure 1.14 The meter is defined to be the distance light travels in 1/299,792,458 of a second in a vacuum. Distance traveled is speed multiplied by time.

The kilogram

The SI unit for mass is the **kilogram** (abbreviated kg); it is defined to be the mass of a platinum–iridium cylinder kept with the old meter standard at the International Bureau of Weights and Measures near Paris. Exact replicas of the standard kilogram are also kept at the U.S. National Institute of Standards and Technology (NIST), located in Gaithersburg, Maryland, outside of Washington, DC, and at other locations around the world. Scientists at NIST are currently investigating two complementary methods of redefining the kilogram (see **Figure 1.15**). The determination of all other masses can be traced ultimately to a comparison with the standard mass.

There is currently an effort to redefine the SI unit of mass in terms of more fundamental processes by 2018. You can explore the history of mass standards and the contenders in the quest to devise a new one at the **website (https://openstaxcollege.org/l/21redefkilo)** of the Physical Measurement Laboratory.

(a) (b)

Figure 1.15 Redefining the SI unit of mass. Complementary methods are being investigated for use in an upcoming redefinition of the SI unit of mass. (a) The U.S. National Institute of Standards and Technology's watt balance is a machine that balances the weight of a test mass against the current and voltage (the "watt") produced by a strong system of magnets. (b) The International Avogadro Project is working to redefine the kilogram based on the dimensions, mass, and other known properties of a silicon sphere. (credit a and credit b: National Institute of Standards and Technology)

Metric Prefixes

SI units are part of the **metric system**, which is convenient for scientific and engineering calculations because the units are categorized by factors of 10. **Table 1.2** lists the metric prefixes and symbols used to denote various factors of 10 in SI units. For example, a centimeter is one-hundredth of a meter (in symbols, 1 cm = 10^{-2} m) and a kilometer is a thousand meters (1 km = 10^3 m). Similarly, a megagram is a million grams (1 Mg = 10^6 g), a nanosecond is a billionth of a second (1 ns = 10^{-9} s), and a terameter is a trillion meters (1 Tm = 10^{12} m).

Prefix	Symbol	Meaning	Prefix	Symbol	Meaning
yotta-	Y	10^{24}	yocto-	y	10^{-24}
zetta-	Z	10^{21}	zepto-	z	10^{-21}
exa-	E	10^{18}	atto-	a	10^{-18}
peta-	P	10^{15}	femto-	f	10^{-15}
tera-	T	10^{12}	pico-	p	10^{-12}
giga-	G	10^9	nano-	n	10^{-9}
mega-	M	10^6	micro-	μ	10^{-6}
kilo-	k	10^3	milli-	m	10^{-3}
hecto-	h	10^2	centi-	c	10^{-2}
deka-	da	10^1	deci-	d	10^{-1}

Table 1.2 Metric Prefixes for Powers of 10 and Their Symbols

The only rule when using metric prefixes is that you cannot "double them up." For example, if you have measurements in petameters (1 Pm = 10^{15} m), it is not proper to talk about megagigameters, although $10^6 \times 10^9 = 10^{15}$. In practice, the only time this becomes a bit confusing is when discussing masses. As we have seen, the base SI unit of mass is the kilogram (kg), but metric prefixes need to be applied to the gram (g), because we are not allowed to "double-up" prefixes. Thus, a thousand kilograms (10^3 kg) is written as a megagram (1 Mg) since

$$10^3\,\text{kg} = 10^3 \times 10^3\,\text{g} = 10^6\,\text{g} = 1\,\text{Mg}.$$

Incidentally, 10^3 kg is also called a *metric ton*, abbreviated t. This is one of the units outside the SI system considered acceptable for use with SI units.

As we see in the next section, metric systems have the advantage that conversions of units involve only powers of 10. There are 100 cm in 1 m, 1000 m in 1 km, and so on. In nonmetric systems, such as the English system of units, the relationships are not as simple—there are 12 in. in 1 ft, 5280 ft in 1 mi, and so on.

Another advantage of metric systems is that the same unit can be used over extremely large ranges of values simply by scaling it with an appropriate metric prefix. The prefix is chosen by the order of magnitude of physical quantities commonly found in the task at hand. For example, distances in meters are suitable in construction, whereas distances in kilometers are appropriate for air travel, and nanometers are convenient in optical design. With the metric system there is no need to invent new units for particular applications. Instead, we rescale the units with which we are already familiar.

Example 1.1

Using Metric Prefixes

Restate the mass $1.93 \times 10^{13}\,\text{kg}$ using a metric prefix such that the resulting numerical value is bigger than one but less than 1000.

Strategy

Since we are not allowed to "double-up" prefixes, we first need to restate the mass in grams by replacing the prefix symbol k with a factor of 10^3 (see **Table 1.2**). Then, we should see which two prefixes in **Table 1.2** are closest to the resulting power of 10 when the number is written in scientific notation. We use whichever of these two prefixes gives us a number between one and 1000.

Solution

Replacing the k in kilogram with a factor of 10^3, we find that

$$1.93 \times 10^{13}\,\text{kg} = 1.93 \times 10^{13} \times 10^3\,\text{g} = 1.93 \times 10^{16}\,\text{g}.$$

From **Table 1.2**, we see that 10^{16} is between "peta-" (10^{15}) and "exa-" (10^{18}). If we use the "peta-" prefix, then we find that $1.93 \times 10^{16}\,\text{g} = 1.93 \times 10^1\,\text{Pg}$, since $16 = 1 + 15$. Alternatively, if we use the "exa-" prefix we find that $1.93 \times 10^{16}\,\text{g} = 1.93 \times 10^{-2}\,\text{Eg}$, since $16 = -2 + 18$. Because the problem asks for the numerical value between one and 1000, we use the "peta-" prefix and the answer is 19.3 Pg.

Significance

It is easy to make silly arithmetic errors when switching from one prefix to another, so it is always a good idea to check that our final answer matches the number we started with. An easy way to do this is to put both numbers in scientific notation and count powers of 10, including the ones hidden in prefixes. If we did not make a mistake, the powers of 10 should match up. In this problem, we started with $1.93 \times 10^{13}\,\text{kg}$, so we have 13 + 3 = 16 powers of 10. Our final answer in scientific notation is 1.93×10^1 Pg, so we have 1 + 15 = 16 powers of 10. So, everything checks out.

If this mass arose from a calculation, we would also want to check to determine whether a mass this large makes any sense in the context of the problem. For this, **Figure 1.9** might be helpful.

1.1 Check Your Understanding Restate $4.79 \times 10^5\,\text{kg}$ using a metric prefix such that the resulting number is bigger than one but less than 1000.

1.4 | Unit Conversion

Learning Objectives

By the end of this section, you will be able to:

- Use conversion factors to express the value of a given quantity in different units.

It is often necessary to convert from one unit to another. For example, if you are reading a European cookbook, some quantities may be expressed in units of liters and you need to convert them to cups. Or perhaps you are reading walking directions from one location to another and you are interested in how many miles you will be walking. In this case, you may need to convert units of feet or meters to miles.

The Power of 1

Even though it will be expressed in new units, we do NOT want to change the actual value of the quantity. In algebra, there is a safe way to keep the value of a quantity unchanged: multiply it by 1. So, if all we ever do is multiply our quantity by 1, we are assured that we keep the same value.

The secret is in a clever use of the many ways there are in which to write the quantity 1. In particular, any fraction whose numerator and denominator are equal does in fact have the value 1. The particular fractions we will choose are called **conversion factors**.

Let's consider a simple example of how to convert units. Suppose we want to convert 80 m to kilometers. The first thing to do is to list the units you have and the units to which you want to convert. In this case, we have units in *meters* and we want to convert to *kilometers*. Next, we need to determine a conversion factor relating meters to kilometers. A **conversion factor** is a ratio that expresses how many of one unit are equal to another unit. For example, there are 12 in. in 1 ft, 1609 m in 1 mi, 100 cm in 1 m, 60 s in 1 min, and so on. Refer to **Appendix B** for a more complete list of conversion factors. In this case, we know that there are 1000 m in 1 km. Now we can set up our unit conversion. We write the units we have and then multiply them by the conversion factor so the units cancel out, as shown:

$$80\,\cancel{\text{m}} \times \frac{1\,\text{km}}{1000\,\cancel{\text{m}}} = 0.080\,\text{km}.$$

Why did the actual quantity (the distance involved) not change? Because all we did, mathematically, was to multiply it by 1. Our conversion factor is a fraction, the value of whose numerator (1 km) is equal to the value of its denominator (1000 m). So, it is just another way to write 1.

Note that the unwanted meter unit cancels, leaving only the desired kilometer unit. You can use this method to convert between any type of unit. Of course, the conversion of 80 m to kilometers is simply the use of a metric prefix, as we saw in the preceding section, so we can get the same answer just as easily by noting that

$$80\,\text{m} = 8.0 \times 10^{1}\,\text{m} = 8.0 \times 10^{-2}\,\text{km} = 0.080\,\text{km},$$

since "kilo-" means 10^3 (see **Table 1.2**) and $1 = -2 + 3.$ However, using conversion factors is handy when converting between units that are not metric or when converting between derived units, as the following examples illustrate.

Example 1.2

Converting Nonmetric Units to Metric

The distance from the university to home is 10 mi and it usually takes 20 min to drive this distance. Calculate the average speed in meters per second (m/s). (*Note:* Average speed is distance traveled divided by time of travel.)

Strategy

First we calculate the average speed using the given units, then we can get the average speed into the desired units by picking the correct conversion factors and multiplying by them. The correct conversion factors are those that cancel the unwanted units and leave the desired units in their place. In this case, we want to convert miles to meters, so we need to know the fact that there are 1609 m in 1 mi. We also want to convert minutes to seconds, so we use the conversion of 60 s in 1 min.

Solution

1. Calculate average speed. Average speed is distance traveled divided by time of travel. (Take this definition as a given for now. Average speed and other motion concepts are covered in later chapters.) In equation form,

$$\text{Average speed} = \frac{\text{Distance}}{\text{Time}}.$$

2. Substitute the given values for distance and time:

$$\text{Average speed} = \frac{10\text{ mi}}{20\text{ min}} = 0.50\,\frac{\text{mi}}{\text{min}}.$$

3. Convert miles per minute to meters per second by multiplying by the conversion factor that cancels miles and leave meters, and also by the conversion factor that cancels minutes and leave seconds:

$$0.50\,\frac{\cancel{\text{mile}}}{\cancel{\text{min}}} \times \frac{1609\text{ m}}{1\,\cancel{\text{mile}}} \times \frac{1\,\cancel{\text{min}}}{60\text{ s}} = \frac{(0.50)(1609)}{60}\text{ m/s} = 13\text{ m/s}.$$

Significance

Check the answer in the following ways:

1. Be sure that each conversion factor is a fraction whose numerator and denominator are equal. This ensures that all you ever do is multiply your quantity by 1 (sometimes repeatedly).
2. Be sure the units in the unit conversion cancel correctly. If the unit conversion factor was written upside down, the units do not cancel correctly in the equation. We see the "miles" in the numerator in 0.50 mi/min cancels the "mile" in the denominator in the first conversion factor. Also, the "min" in the denominator in 0.50 mi/min cancels the "min" in the numerator in the second conversion factor.
3. Check that the units of the final answer are the desired units. The problem asked us to solve for average speed in units of meters per second and, after the cancellations, the only units left are a meter (m) in the numerator and a second (s) in the denominator, so we have indeed obtained these units.

1.2 Check Your Understanding Light travels about 9 Pm in a year. Given that a year is about 3×10^7 s, what is the speed of light in meters per second?

Example 1.3

Converting between Metric Units

The density of iron is 7.86 g/cm^3 under standard conditions. Convert this to kg/m^3.

Strategy

We need to convert grams to kilograms and cubic centimeters to cubic meters. The conversion factors we need are $1\text{ kg} = 10^3\text{ g}$ and $1\text{ cm} = 10^{-2}\text{m}.$ However, we are dealing with cubic centimeters $(\text{cm}^3 = \text{cm} \times \text{cm} \times \text{cm})$, so we have to use the second conversion factor three times (that is, we need to cube it). The idea is still to multiply by the conversion factors in such a way that they cancel the units we want to get rid of and introduce the units we want to keep.

Solution

$$7.86\frac{\cancel{\text{g}}}{\cancel{\text{cm}}^3} \times \frac{\text{kg}}{10^3\,\cancel{\text{g}}} \times \left(\frac{\cancel{\text{cm}}}{10^{-2}\text{m}}\right)^3 = \frac{7.86}{(10^3)(10^{-6})}\text{kg/m}^3 = 7.86 \times 10^3\text{ kg/m}^3$$

Significance

Remember, it's always important to check the answer.

1. Be sure that each conversion factor is a fraction whose numerator and denominator are equal. In this case, the first conversion factor has a numerator of 1 kg and a denominator of 10^3 g. The second conversion factor has a numerator of 1 cm and a denominator of 10^{-2} m.

2. Be sure to cancel the units in the unit conversion correctly. We see that the gram ("g") in the numerator in 7.86 g/cm^3 cancels the "g" in the denominator in the first conversion factor. Also, the three factors of "cm" in the denominator in 7.86 g/cm^3 cancel with the three factors of "cm" in the numerator that we get by cubing the second conversion factor.
3. Check that the units of the final answer are the desired units. The problem asked for us to convert to kilograms per cubic meter. After the cancellations just described, we see the only units we have left are "kg" in the numerator and three factors of "m" in the denominator (that is, one factor of "m" cubed, or "m^3"). Therefore, the units on the final answer are correct.

1.3 Check Your Understanding We know from **Figure 1.9** that the diameter of Earth is on the order of 10^7 m, so the order of magnitude of its surface area is 10^{14} m^2. What is that in square kilometers (that is, km^2)? (Try doing this both by converting 10^7 m to km and then squaring it and then by converting 10^{14} m^2 directly to square kilometers. You should get the same answer both ways.)

Unit conversions may not seem very interesting, but not doing them can be costly. One famous example of this situation was seen with the *Mars Climate Orbiter*. This probe was launched by NASA on December 11, 1998. On September 23, 1999, while attempting to guide the probe into its planned orbit around Mars, NASA lost contact with it. Subsequent investigations showed a piece of software called SM_FORCES (or "small forces") was recording thruster performance data in the English units of pound-seconds (lb-s). However, other pieces of software that used these values for course corrections expected them to be recorded in the SI units of newton-seconds (N-s), as dictated in the software interface protocols. This error caused the probe to follow a very different trajectory from what NASA thought it was following, which most likely caused the probe either to burn up in the Martian atmosphere or to shoot out into space. This failure to pay attention to unit conversions cost hundreds of millions of dollars, not to mention all the time invested by the scientists and engineers who worked on the project.

1.4 Check Your Understanding Given that 1 lb (pound) is 4.45 N, were the numbers being output by SM_FORCES too big or too small?

1.5 | Dimensional Analysis

Learning Objectives

By the end of this section, you will be able to:

- Find the dimensions of a mathematical expression involving physical quantities.
- Determine whether an equation involving physical quantities is dimensionally consistent.

The **dimension** of any physical quantity expresses its dependence on the base quantities as a product of symbols (or powers of symbols) representing the base quantities. **Table 1.3** lists the base quantities and the symbols used for their dimension. For example, a measurement of length is said to have dimension L or L^1, a measurement of mass has dimension M or M^1, and a measurement of time has dimension T or T^1. Like units, dimensions obey the rules of algebra. Thus, area is the product of two lengths and so has dimension L^2, or length squared. Similarly, volume is the product of three lengths and has dimension L^3, or length cubed. Speed has dimension length over time, L/T or LT^{-1}. Volumetric mass density has dimension M/L^3 or ML^{-3}, or mass over length cubed. In general, the dimension of any physical quantity can be written as $\text{L}^a\text{M}^b\text{T}^c\text{I}^d\Theta^e\text{N}^f\text{J}^g$ for some powers $a, b, c, d, e, f,$ and g. We can write the dimensions of a length in this form with $a = 1$ and the remaining six powers all set equal to zero: $\text{L}^1 = \text{L}^1\text{M}^0\text{T}^0\text{I}^0\Theta^0\text{N}^0\text{J}^0$. Any quantity with a dimension that can be written so that all seven powers are zero (that is, its dimension is $\text{L}^0\text{M}^0\text{T}^0\text{I}^0\Theta^0\text{N}^0\text{J}^0$) is called **dimensionless** (or sometimes "of dimension 1," because anything raised to the zero power is one). Physicists often call dimensionless quantities *pure numbers*.

Base Quantity	Symbol for Dimension
Length	L
Mass	M
Time	T
Current	I
Thermodynamic temperature	Θ
Amount of substance	N
Luminous intensity	J

Table 1.3 Base Quantities and Their Dimensions

Physicists often use square brackets around the symbol for a physical quantity to represent the dimensions of that quantity. For example, if r is the radius of a cylinder and h is its height, then we write $[r] = \text{L}$ and $[h] = \text{L}$ to indicate the dimensions of the radius and height are both those of length, or L. Similarly, if we use the symbol A for the surface area of a cylinder and V for its volume, then $[A] = \text{L}^2$ and $[V] = \text{L}^3$. If we use the symbol m for the mass of the cylinder and ρ for the density of the material from which the cylinder is made, then $[m] = \text{M}$ and $[\rho] = \text{ML}^{-3}$.

The importance of the concept of dimension arises from the fact that any mathematical equation relating physical quantities must be **dimensionally consistent**, which means the equation must obey the following rules:

- Every term in an expression must have the same dimensions; it does not make sense to add or subtract quantities of differing dimension (think of the old saying: "You can't add apples and oranges"). In particular, the expressions on each side of the equality in an equation must have the same dimensions.
- The arguments of any of the standard mathematical functions such as trigonometric functions (such as sine and cosine), logarithms, or exponential functions that appear in the equation must be dimensionless. These functions require pure numbers as inputs and give pure numbers as outputs.

If either of these rules is violated, an equation is not dimensionally consistent and cannot possibly be a correct statement of physical law. This simple fact can be used to check for typos or algebra mistakes, to help remember the various laws of physics, and even to suggest the form that new laws of physics might take. This last use of dimensions is beyond the scope of this text, but is something you will undoubtedly learn later in your academic career.

Example 1.4

Using Dimensions to Remember an Equation

Suppose we need the formula for the area of a circle for some computation. Like many people who learned geometry too long ago to recall with any certainty, two expressions may pop into our mind when we think of circles: πr^2 and $2\pi r$. One expression is the circumference of a circle of radius *r* and the other is its area. But which is which?

Strategy

One natural strategy is to look it up, but this could take time to find information from a reputable source. Besides, even if we think the source is reputable, we shouldn't trust everything we read. It is nice to have a way to double-check just by thinking about it. Also, we might be in a situation in which we cannot look things up (such as during a test). Thus, the strategy is to find the dimensions of both expressions by making use of the fact that dimensions follow the rules of algebra. If either expression does not have the same dimensions as area, then it cannot possibly be the correct equation for the area of a circle.

Solution

We know the dimension of area is L^2. Now, the dimension of the expression πr^2 is

$$[\pi r^2] = [\pi] \cdot [r]^2 = 1 \cdot \text{L}^2 = \text{L}^2,$$

since the constant π is a pure number and the radius r is a length. Therefore, πr^2 has the dimension of area. Similarly, the dimension of the expression $2\pi r$ is

$$[2\pi r] = [2] \cdot [\pi] \cdot [r] = 1 \cdot 1 \cdot \text{L} = \text{L},$$

since the constants 2 and π are both dimensionless and the radius r is a length. We see that $2\pi r$ has the dimension of length, which means it cannot possibly be an area.

We rule out $2\pi r$ because it is not dimensionally consistent with being an area. We see that πr^2 is dimensionally consistent with being an area, so if we have to choose between these two expressions, πr^2 is the one to choose.

Significance

This may seem like kind of a silly example, but the ideas are very general. As long as we know the dimensions of the individual physical quantities that appear in an equation, we can check to see whether the equation is dimensionally consistent. On the other hand, knowing that true equations are dimensionally consistent, we can match expressions from our imperfect memories to the quantities for which they might be expressions. Doing this will not help us remember dimensionless factors that appear in the equations (for example, if you had accidentally conflated the two expressions from the example into $2\pi r^2$, then dimensional analysis is no help), but it does help us remember the correct basic form of equations.

1.5 Check Your Understanding Suppose we want the formula for the volume of a sphere. The two expressions commonly mentioned in elementary discussions of spheres are $4\pi r^2$ and $4\pi r^3/3$. One is the volume of a sphere of radius r and the other is its surface area. Which one is the volume?

Example 1.5

Checking Equations for Dimensional Consistency

Consider the physical quantities s, v, a, and t with dimensions $[s] = \text{L}$, $[v] = \text{LT}^{-1}$, $[a] = \text{LT}^{-2}$, and $[t] = \text{T}$. Determine whether each of the following equations is dimensionally consistent: (a) $s = vt + 0.5at^2$; (b) $s = vt^2 + 0.5at$; and (c) $v = \sin(at^2/s)$.

Strategy

By the definition of dimensional consistency, we need to check that each term in a given equation has the same dimensions as the other terms in that equation and that the arguments of any standard mathematical functions are dimensionless.

Solution

a. There are no trigonometric, logarithmic, or exponential functions to worry about in this equation, so we need only look at the dimensions of each term appearing in the equation. There are three terms, one in the left expression and two in the expression on the right, so we look at each in turn:

$$\begin{aligned}[s] &= \text{L} \\ [vt] &= [v] \cdot [t] = \text{LT}^{-1} \cdot \text{T} = \text{LT}^0 = \text{L} \\ [0.5at^2] &= [a] \cdot [t]^2 = \text{LT}^{-2} \cdot \text{T}^2 = \text{LT}^0 = \text{L}.\end{aligned}$$

All three terms have the same dimension, so this equation is dimensionally consistent.

b. Again, there are no trigonometric, exponential, or logarithmic functions, so we only need to look at the dimensions of each of the three terms appearing in the equation:

$$[s] = \mathrm{L}$$
$$[vt^2] = [v] \cdot [t]^2 = \mathrm{LT}^{-1} \cdot \mathrm{T}^2 = \mathrm{LT}$$
$$[at] = [a] \cdot [t] = \mathrm{LT}^{-2} \cdot \mathrm{T} = \mathrm{LT}^{-1}.$$

None of the three terms has the same dimension as any other, so this is about as far from being dimensionally consistent as you can get. The technical term for an equation like this is *nonsense*.

c. This equation has a trigonometric function in it, so first we should check that the argument of the sine function is dimensionless:

$$\left[\frac{at^2}{s}\right] = \frac{[a] \cdot [t]^2}{[s]} = \frac{\mathrm{LT}^{-2} \cdot \mathrm{T}^2}{\mathrm{L}} = \frac{\mathrm{L}}{\mathrm{L}} = 1.$$

The argument is dimensionless. So far, so good. Now we need to check the dimensions of each of the two terms (that is, the left expression and the right expression) in the equation:

$$[v] = \mathrm{LT}^{-1}$$
$$\left[\sin\left(\frac{at^2}{s}\right)\right] = 1.$$

The two terms have different dimensions—meaning, the equation is not dimensionally consistent. This equation is another example of “nonsense.”

Significance

If we are trusting people, these types of dimensional checks might seem unnecessary. But, rest assured, any textbook on a quantitative subject such as physics (including this one) almost certainly contains some equations with typos. Checking equations routinely by dimensional analysis save us the embarrassment of using an incorrect equation. Also, checking the dimensions of an equation we obtain through algebraic manipulation is a great way to make sure we did not make a mistake (or to spot a mistake, if we made one).

1.6 Check Your Understanding Is the equation $v = at$ dimensionally consistent?

One further point that needs to be mentioned is the effect of the operations of calculus on dimensions. We have seen that dimensions obey the rules of algebra, just like units, but what happens when we take the derivative of one physical quantity with respect to another or integrate a physical quantity over another? The derivative of a function is just the slope of the line tangent to its graph and slopes are ratios, so for physical quantities v and t, we have that the dimension of the derivative of v with respect to t is just the ratio of the dimension of v over that of t:

$$\left[\frac{dv}{dt}\right] = \frac{[v]}{[t]}.$$

Similarly, since integrals are just sums of products, the dimension of the integral of v with respect to t is simply the dimension of v times the dimension of t:

$$\left[\int v dt\right] = [v] \cdot [t].$$

By the same reasoning, analogous rules hold for the units of physical quantities derived from other quantities by integration or differentiation.

1.6 | Estimates and Fermi Calculations

Learning Objectives

By the end of this section, you will be able to:

- Estimate the values of physical quantities.

On many occasions, physicists, other scientists, and engineers need to make *estimates* for a particular quantity. Other terms sometimes used are *guesstimates*, *order-of-magnitude approximations*, *back-of-the-envelope calculations*, or *Fermi calculations*. (The physicist Enrico Fermi mentioned earlier was famous for his ability to estimate various kinds of data with surprising precision.) Will that piece of equipment fit in the back of the car or do we need to rent a truck? How long will this download take? About how large a current will there be in this circuit when it is turned on? How many houses could a proposed power plant actually power if it is built? Note that estimating does not mean guessing a number or a formula at random. Rather, **estimation** means using prior experience and sound physical reasoning to arrive at a rough idea of a quantity's value. Because the process of determining a reliable approximation usually involves the identification of correct physical principles and a good guess about the relevant variables, estimating is very useful in developing physical intuition. Estimates also allow us perform "sanity checks" on calculations or policy proposals by helping us rule out certain scenarios or unrealistic numbers. They allow us to challenge others (as well as ourselves) in our efforts to learn truths about the world.

Many estimates are based on formulas in which the input quantities are known only to a limited precision. As you develop physics problem-solving skills (which are applicable to a wide variety of fields), you also will develop skills at estimating. You develop these skills by thinking more quantitatively and by being willing to take risks. As with any skill, experience helps. Familiarity with dimensions (see **Table 1.3**) and units (see **Table 1.1** and **Table 1.2**), and the scales of base quantities (see **Figure 1.9**) also helps.

To make some progress in estimating, you need to have some definite ideas about how variables may be related. The following strategies may help you in practicing the art of estimation:

- *Get big lengths from smaller lengths.* When estimating lengths, remember that anything can be a ruler. Thus, imagine breaking a big thing into smaller things, estimate the length of one of the smaller things, and multiply to get the length of the big thing. For example, to estimate the height of a building, first count how many floors it has. Then, estimate how big a single floor is by imagining how many people would have to stand on each other's shoulders to reach the ceiling. Last, estimate the height of a person. The product of these three estimates is your estimate of the height of the building. It helps to have memorized a few length scales relevant to the sorts of problems you find yourself solving. For example, knowing some of the length scales in **Figure 1.9** might come in handy. Sometimes it also helps to do this in reverse—that is, to estimate the length of a small thing, imagine a bunch of them making up a bigger thing. For example, to estimate the thickness of a sheet of paper, estimate the thickness of a stack of paper and then divide by the number of pages in the stack. These same strategies of breaking big things into smaller things or aggregating smaller things into a bigger thing can sometimes be used to estimate other physical quantities, such as masses and times.
- *Get areas and volumes from lengths.* When dealing with an area or a volume of a complex object, introduce a simple model of the object such as a sphere or a box. Then, estimate the linear dimensions (such as the radius of the sphere or the length, width, and height of the box) first, and use your estimates to obtain the volume or area from standard geometric formulas. If you happen to have an estimate of an object's area or volume, you can also do the reverse; that is, use standard geometric formulas to get an estimate of its linear dimensions.
- *Get masses from volumes and densities.* When estimating masses of objects, it can help first to estimate its volume and then to estimate its mass from a rough estimate of its average density (recall, density has dimension mass over length cubed, so mass is density times volume). For this, it helps to remember that the density of air is around 1 kg/m^3, the density of water is 10^3 kg/m^3, and the densest everyday solids max out at around 10^4 kg/m^3. Asking yourself whether an object floats or sinks in either air or water gets you a ballpark estimate of its density. You can also do this the other way around; if you have an estimate of an object's mass and its density, you can use them to get an estimate of its volume.
- *If all else fails, bound it.* For physical quantities for which you do not have a lot of intuition, sometimes the best you can do is think something like: Well, it must be bigger than this and smaller than that. For example, suppose you need to estimate the mass of a moose. Maybe you have a lot of experience with moose and know their average mass offhand. If so, great. But for most people, the best they can do is to think something like: It must be bigger than a person (of order 10^2 kg) and less than a car (of order 10^3 kg). If you need a single number for a subsequent

calculation, you can take the geometric mean of the upper and lower bound—that is, you multiply them together and then take the square root. For the moose mass example, this would be

$$\left(10^2 \times 10^3\right)^{0.5} = 10^{2.5} = 10^{0.5} \times 10^2 \approx 3 \times 10^2\,\text{kg}.$$

The tighter the bounds, the better. Also, no rules are unbreakable when it comes to estimation. If you think the value of the quantity is likely to be closer to the upper bound than the lower bound, then you may want to bump up your estimate from the geometric mean by an order or two of magnitude.

- *One "sig. fig." is fine.* There is no need to go beyond one significant figure when doing calculations to obtain an estimate. In most cases, the order of magnitude is good enough. The goal is just to get in the ballpark figure, so keep the arithmetic as simple as possible.
- *Ask yourself: Does this make any sense?* Last, check to see whether your answer is reasonable. How does it compare with the values of other quantities with the same dimensions that you already know or can look up easily? If you get some wacky answer (for example, if you estimate the mass of the Atlantic Ocean to be bigger than the mass of Earth, or some time span to be longer than the age of the universe), first check to see whether your units are correct. Then, check for arithmetic errors. Then, rethink the logic you used to arrive at your answer. If everything checks out, you may have just proved that some slick new idea is actually bogus.

Example 1.6

Mass of Earth's Oceans

Estimate the total mass of the oceans on Earth.

Strategy

We know the density of water is about 10^3 kg/m^3, so we start with the advice to "get masses from densities and volumes." Thus, we need to estimate the volume of the planet's oceans. Using the advice to "get areas and volumes from lengths," we can estimate the volume of the oceans as surface area times average depth, or $V = AD$. We know the diameter of Earth from **Figure 1.9** and we know that most of Earth's surface is covered in water, so we can estimate the surface area of the oceans as being roughly equal to the surface area of the planet. By following the advice to "get areas and volumes from lengths" again, we can approximate Earth as a sphere and use the formula for the surface area of a sphere of diameter *d*—that is, $A = \pi d^2$, to estimate the surface area of the oceans. Now we just need to estimate the average depth of the oceans. For this, we use the advice: "If all else fails, bound it." We happen to know the deepest points in the ocean are around 10 km and that it is not uncommon for the ocean to be deeper than 1 km, so we take the average depth to be around $(10^3 \times 10^4)^{0.5} \approx 3 \times 10^3\,\text{m}$. Now we just need to put it all together, heeding the advice that "one 'sig. fig.' is fine."

Solution

We estimate the surface area of Earth (and hence the surface area of Earth's oceans) to be roughly

$$A = \pi d^2 = \pi(10^7\,\text{m})^2 \approx 3 \times 10^{14}\,\text{m}^2.$$

Next, using our average depth estimate of $D = 3 \times 10^3\,\text{m}$, which was obtained by bounding, we estimate the volume of Earth's oceans to be

$$V = AD = (3 \times 10^{14}\,\text{m}^2)(3 \times 10^3\,\text{m}) = 9 \times 10^{17}\,\text{m}^3.$$

Last, we estimate the mass of the world's oceans to be

$$M = \rho V = (10^3\ \text{kg/m}^3)(9 \times 10^{17}\,\text{m}^3) = 9 \times 10^{20}\,\text{kg}.$$

Thus, we estimate that the order of magnitude of the mass of the planet's oceans is 10^{21} kg.

Significance

To verify our answer to the best of our ability, we first need to answer the question: Does this make any sense? From **Figure 1.9**, we see the mass of Earth's atmosphere is on the order of 10^{19} kg and the mass of Earth is on the order of 10^{25} kg. It is reassuring that our estimate of 10^{21} kg for the mass of Earth's oceans falls somewhere between these two. So, yes, it does seem to make sense. It just so happens that we did a search on the Web for "mass of oceans" and the top search results all said 1.4×10^{21} kg, which is the same order of magnitude as our estimate. Now, rather than having to trust blindly whoever first put that number up on a website (most of the other sites probably just copied it from them, after all), we can have a little more confidence in it.

1.7 Check Your Understanding Figure 1.9 says the mass of the atmosphere is 10^{19} kg. Assuming the density of the atmosphere is 1 kg/m^3, estimate the height of Earth's atmosphere. Do you think your answer is an underestimate or an overestimate? Explain why.

How many piano tuners are there in New York City? How many leaves are on that tree? If you are studying photosynthesis or thinking of writing a smartphone app for piano tuners, then the answers to these questions might be of great interest to you. Otherwise, you probably couldn't care less what the answers are. However, these are exactly the sorts of estimation problems that people in various tech industries have been asking potential employees to evaluate their quantitative reasoning skills. If building physical intuition and evaluating quantitative claims do not seem like sufficient reasons for you to practice estimation problems, how about the fact that being good at them just might land you a high-paying job?

For practice estimating relative lengths, areas, and volumes, check out this PhET (https://openstaxcollege.org/l/21lengthgame) simulation, titled "Estimation."

1.7 | Uncertainties and Significant Figures

Learning Objectives

By the end of this section, you will be able to:

- Determine the correct number of significant figures for the result of a computation.
- Describe the relationship between the concepts of accuracy, precision, uncertainty, and discrepancy.
- Calculate the percent uncertainty of a measurement, given its value and its uncertainty.
- Determine the uncertainty of the result of a computation involving quantities with given uncertainties.

Figure 1.16 shows two instruments used to measure the mass of an object. The digital scale has mostly replaced the double-pan balance in physics labs because it gives more accurate and precise measurements. But what exactly do we mean by *accurate* and *precise*? Aren't they the same thing? In this section we examine in detail the process of making and reporting a measurement.

(a) (b)

Figure 1.16 (a) A double-pan mechanical balance is used to compare different masses. Usually an object with unknown mass is placed in one pan and objects of known mass are placed in the other pan. When the bar that connects the two pans is horizontal, then the masses in both pans are equal. The "known masses" are typically metal cylinders of standard mass such as 1 g, 10 g, and 100 g. (b) Many mechanical balances, such as double-pan balances, have been replaced by digital scales, which can typically measure the mass of an object more precisely. A mechanical balance may read only the mass of an object to the nearest tenth of a gram, but many digital scales can measure the mass of an object up to the nearest thousandth of a gram. (credit a: modification of work by Serge Melki; credit b: modification of work by Karel Jakubec)

Accuracy and Precision of a Measurement

Science is based on observation and experiment—that is, on measurements. **Accuracy** is how close a measurement is to the accepted reference value for that measurement. For example, let's say we want to measure the length of standard printer paper. The packaging in which we purchased the paper states that it is 11.0 in. long. We then measure the length of the paper three times and obtain the following measurements: 11.1 in., 11.2 in., and 10.9 in. These measurements are quite accurate because they are very close to the reference value of 11.0 in. In contrast, if we had obtained a measurement of 12 in., our measurement would not be very accurate. Notice that the concept of accuracy requires that an accepted reference value be given.

The **precision** of measurements refers to how close the agreement is between repeated independent measurements (which are repeated under the same conditions). Consider the example of the paper measurements. The precision of the measurements refers to the spread of the measured values. One way to analyze the precision of the measurements is to determine the range, or difference, between the lowest and the highest measured values. In this case, the lowest value was 10.9 in. and the highest value was 11.2 in. Thus, the measured values deviated from each other by, at most, 0.3 in. These measurements were relatively precise because they did not vary too much in value. However, if the measured values had been 10.9 in., 11.1 in., and 11.9 in., then the measurements would not be very precise because there would be significant variation from one measurement to another. Notice that the concept of precision depends only on the actual measurements acquired and does not depend on an accepted reference value.

The measurements in the paper example are both accurate and precise, but in some cases, measurements are accurate but not precise, or they are precise but not accurate. Let's consider an example of a GPS attempting to locate the position of a restaurant in a city. Think of the restaurant location as existing at the center of a bull's-eye target and think of each GPS attempt to locate the restaurant as a black dot. In Figure 1.17(a), we see the GPS measurements are spread out far apart from each other, but they are all relatively close to the actual location of the restaurant at the center of the target. This indicates a low-precision, high-accuracy measuring system. However, in Figure 1.17(b), the GPS measurements are concentrated quite closely to one another, but they are far away from the target location. This indicates a high-precision, low-accuracy measuring system.

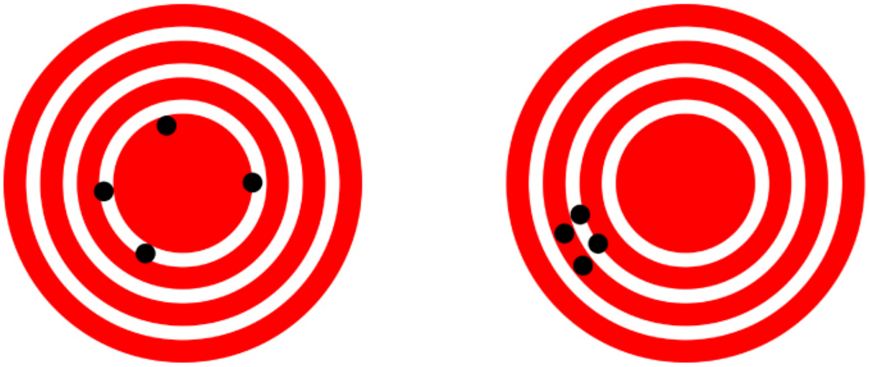


Figure 1.17 A GPS attempts to locate a restaurant at the center of the bull's-eye. The black dots represent each attempt to pinpoint the location of the restaurant. (a) The dots are spread out quite far apart from one another, indicating low precision, but they are each rather close to the actual location of the restaurant, indicating high accuracy. (b) The dots are concentrated rather closely to one another, indicating high precision, but they are rather far away from the actual location of the restaurant, indicating low accuracy. (credit a and credit b: modification of works by Dark Evil)

Accuracy, Precision, Uncertainty, and Discrepancy

The precision of a measuring system is related to the **uncertainty** in the measurements whereas the accuracy is related to the **discrepancy** from the accepted reference value. Uncertainty is a quantitative measure of how much your measured values deviate from one another. There are many different methods of calculating uncertainty, each of which is appropriate to different situations. Some examples include taking the range (that is, the biggest less the smallest) or finding the standard deviation of the measurements. Discrepancy (or "measurement error") is the difference between the measured value and a given standard or expected value. If the measurements are not very precise, then the uncertainty of the values is high. If the measurements are not very accurate, then the discrepancy of the values is high.

Recall our example of measuring paper length; we obtained measurements of 11.1 in., 11.2 in., and 10.9 in., and the accepted value was 11.0 in. We might average the three measurements to say our best guess is 11.1 in.; in this case, our discrepancy is 11.1 – 11.0 = 0.1 in., which provides a quantitative measure of accuracy. We might calculate the uncertainty in our best guess by using the range of our measured values: 0.3 in. Then we would say the length of the paper is 11.1 in. plus or minus 0.3 in. The uncertainty in a measurement, A, is often denoted as δA (read "delta A"), so the measurement result would be recorded as $A \pm \delta A$. Returning to our paper example, the measured length of the paper could be expressed as 11.1 ± 0.3 in. Since the discrepancy of 0.1 in. is less than the uncertainty of 0.3 in., we might say the measured value agrees with the accepted reference value to within experimental uncertainty.

Some factors that contribute to uncertainty in a measurement include the following:

- Limitations of the measuring device
- The skill of the person taking the measurement
- Irregularities in the object being measured
- Any other factors that affect the outcome (highly dependent on the situation)

In our example, such factors contributing to the uncertainty could be the smallest division on the ruler is 1/16 in., the person using the ruler has bad eyesight, the ruler is worn down on one end, or one side of the paper is slightly longer than the other. At any rate, the uncertainty in a measurement must be calculated to quantify its precision. If a reference value is known, it makes sense to calculate the discrepancy as well to quantify its accuracy.

Percent uncertainty

Another method of expressing uncertainty is as a percent of the measured value. If a measurement A is expressed with uncertainty δA, the **percent uncertainty** is defined as

$$\text{Percent uncertainty} = \frac{\delta A}{A} \times 100\%.$$

Example 1.7

Calculating Percent Uncertainty: A Bag of Apples

A grocery store sells 5-lb bags of apples. Let’s say we purchase four bags during the course of a month and weigh the bags each time. We obtain the following measurements:

- Week 1 weight: 4.8 lb
- Week 2 weight: 5.3 lb
- Week 3 weight: 4.9 lb
- Week 4 weight: 5.4 lb

We then determine the average weight of the 5-lb bag of apples is 5.1 ± 0.2 lb. What is the percent uncertainty of the bag’s weight?

Strategy

First, observe that the average value of the bag’s weight, A, is 5.1 lb. The uncertainty in this value, $\delta A,$ is 0.2 lb. We can use the following equation to determine the percent uncertainty of the weight:

$$\text{Percent uncertainty} = \frac{\delta A}{A} \times 100\%. \quad (1.1)$$

Solution

Substitute the values into the equation:

Percent Uncertainty

$$\frac{\delta A}{A} \times 100\% = \frac{0.2\text{ lb}}{5.1\text{ lb}} \times 100\% = 3.9\% \approx 4\%.$$

Significance

We can conclude the average weight of a bag of apples from this store is 5.1 lb ± 4%. Notice the percent uncertainty is dimensionless because the units of weight in $\delta A = 0.2$ lb canceled those inn A = 5.1 lb when we took the ratio.

1.8 Check Your Understanding A high school track coach has just purchased a new stopwatch. The stopwatch manual states the stopwatch has an uncertainty of ±0.05 s. Runners on the track coach’s team regularly clock 100-m sprints of 11.49 s to 15.01 s. At the school’s last track meet, the first-place sprinter came in at 12.04 s and the second-place sprinter came in at 12.07 s. Will the coach’s new stopwatch be helpful in timing the sprint team? Why or why not?

Combining uncertainties in calculations

Addition or Subtraction of Quantities

What happens if two different measurements need to be combined by addition or subtraction to calculate a final quantity? For example, suppose we want to determine the weight of a suitcase using a bathroom scale whose range can only measure objects weighing more than 50 lb. One method is to first take a measurement of the total weight of a person standing on the scale holding the suitcase. Let's call that measurement W_t. Perhaps we obtain a value of 193±2 lb. (The sources of the uncertainty may come from multiple causes: how hard it was for the person to read the scale, the fact that the needle on the scale may have been jiggling somewhat, etc.) Then, take another measurement of the person alone, W_p. Suppose that has a value of 175±1 lb. (Since he was not holding the suitcase at the same time he read the scale, perhaps the uncertainty was less this time.) The weight of the suitcase can obviously be calculated from $W_s = W_t - W_p$. And, 193 lb - 175 lb = 18 lb. But, there must be some uncertainty in the weight of the suitcase, because the individual measurements that we subtracted each had uncertainties.

How do we combine the uncertainties from the individual measurements, $W_t \pm \delta W_t$ and $W_p \pm dW_p$ to arrive at a final result,

$W_s \pm \delta W_s$? We've already seen that the value of W_s is found by simply subtracting W_t - W_p. But what do we do with the uncertainties? The simplest way might be to just add the uncertainties, using the logic that they both contribute to the uncertainty in the final quantity. However more sophisticated analysis reveals that, if the uncertainties in the individual measurements were independent of one another, we are likely overestimating the uncertainty in our final result if we simply add them. A more accurate final answer is obtained by taking the square root of the sum of the squares of the individual uncertainties. That is:

Adding Uncertainties in Quadrature

$$\delta W_s = \sqrt{\delta W_t^2 + \delta W_p^2} \tag{1.2}$$

So, $\delta W_s = \sqrt{2^2 + 1^2} = \sqrt{5} = 2.236$

In our example, then, the uncertainty in the weight of the suitcase, is 2.236 lb. However, we must follow the rules for significant figures (discussed in detail below). Since the individual weight measurements are expressed to the nearest lb, we will round this calculated number to the nearest lb, and express the weight of the suitcase as 18 ± 2 lb.

Multiplication or Division of Quantities

Uncertainty exists in anything calculated from measured quantities. For example, the area of a floor calculated from measurements of its length and width has an uncertainty because the length and width each have uncertainties. How big is the uncertainty in something you calculate by multiplication or division? In this case, the measurements may not even have the same dimensions. In this case, assuming that the uncertainties in the individual measurements are independent of one another, the **percent uncertainty** in a quantity calculated by multiplication or division is the quadrature sum of the percent uncertainties in the items used to make the calculation. Equivalently, the **relative uncertainty** in a quantity calculated by multiplication or division is the quadrature sum of the relative uncertainties in the items used to make the calculation. In our case, if $A = L \times W$, then

Adding Relative Uncertainties in Quadrature

$$\frac{\delta A}{A} = \sqrt{\left(\frac{\delta L}{L}\right)^2 + \left(\frac{\delta W}{W}\right)^2} \tag{1.3}$$

For example, if a floor has a length of $L = 4.00$ m and a width of $W = 3.00$ m, with uncertainties of 1% and 2%, respectively, then the area of the floor is 12.0 m^2 and has an uncertainty of $\sqrt{1^2 + 2^2} = \sqrt{5} = 2.236\,\%$. (Expressed as an area, this is 0.268 m^2 $[\ 12.0\ \text{m}^2 \times 0.02236\ ]$, which we round to 0.3 m^2 since the area of the floor is given to a tenth of a square meter.) So, we express the area of the floor as $A = 12.0 \pm 0.3$ m^2.

Precision of Measuring Tools and Significant Figures

An important factor in the precision of measurements involves the precision of the measuring tool. In general, a precise measuring tool is one that can measure values in very small increments. For example, a standard ruler can measure length to the nearest millimeter whereas a caliper can measure length to the nearest 0.01 mm. The caliper is a more precise measuring tool because it can measure extremely small differences in length. The more precise the measuring tool, the more precise the measurements.

When we express measured values, we can only list as many digits as we measured initially with our measuring tool. For example, if we use a standard ruler to measure the length of a stick, we may measure it to be 36.7 cm. We can’t express this value as 36.71 cm because our measuring tool is not precise enough to measure a hundredth of a centimeter. It should be noted that the last digit in a measured value has been estimated in some way by the person performing the measurement. For example, the person measuring the length of a stick with a ruler notices the stick length seems to be somewhere in between 36.6 cm and 36.7 cm, and he or she must estimate the value of the last digit. Using the method of **significant figures**, the rule is that *the last digit written down in a measurement is the first digit with some uncertainty*. To determine the number of significant digits in a value, start with the first measured value at the left and count the number of digits through the last digit written on the right. For example, the measured value 36.7 cm has three digits, or three significant figures. Significant figures indicate the precision of the measuring tool used to measure a value.

Zeros

Special consideration is given to zeros when counting significant figures. The zeros in 0.053 are not significant because they are placeholders that locate the decimal point. There are two significant figures in 0.053. The zeros in 10.053 are not placeholders; they are significant. This number has five significant figures. The zeros in 1300 may or may not be significant, depending on the style of writing numbers. They could mean the number is known to the last digit or they could be placeholders. So 1300 could have two, three, or four significant figures. To avoid this ambiguity, we should write 1300 in scientific notation as 1.3×10^3, 1.30×10^3, or 1.300×10^3, depending on whether it has two, three, or four significant figures. *Zeros are significant except when they serve only as placeholders*.

Significant figures in calculations

When combining measurements with different degrees of precision, *the number of significant digits in the final answer can be no greater than the number of significant digits in the least-precise measured value*. There are two different rules, one for multiplication and division and the other for addition and subtraction.

1. *For multiplication and division, the result should have the same number of significant figures as the quantity with the least number of significant figures entering into the calculation*. For example, the area of a circle can be calculated from its radius using $A = \pi r^2$. Let's see how many significant figures the area has if the radius has only two—say, $r = 1.2$ m. Using a calculator with an eight-digit output, we would calculate

$$A = \pi r^2 = (3.1415927\ldots) \times (1.2\text{ m})^2 = 4.5238934\text{ m}^2.$$

   But because the radius has only two significant figures, it limits the calculated quantity to two significant figures, or

$$A = 4.5\text{ m}^2,$$

   although π is good to at least eight digits.
2. *For addition and subtraction, the answer can contain no more decimal places than the least-precise measurement.* Suppose we buy 7.56 kg of potatoes in a grocery store as measured with a scale with precision 0.01 kg, then we drop off 6.052 kg of potatoes at your laboratory as measured by a scale with precision 0.001 kg. Then, we go home and add 13.7 kg of potatoes as measured by a bathroom scale with precision 0.1 kg. How many kilograms of potatoes do we now have and how many significant figures are appropriate in the answer? The mass is found by simple addition and subtraction:

$$\begin{array}{r} 7.56\text{ kg} \\ -6.052\text{ kg} \\ \underline{+13.7\text{ kg}} \\ 15.208\text{ kg} \end{array} = 15.2\text{ kg}.$$

   Next, we identify the least-precise measurement: 13.7 kg. This measurement is expressed to the 0.1 decimal place, so our final answer must also be expressed to the 0.1 decimal place. Thus, the answer is rounded to the tenths place, giving us 15.2 kg.

Significant figures in this text

In this text, most numbers are assumed to have three significant figures. Furthermore, consistent numbers of significant figures are used in all worked examples. An answer given to three digits is based on input good to at least three digits, for example. If the input has fewer significant figures, the answer will also have fewer significant figures. Care is also taken that the number of significant figures is reasonable for the situation posed. In some topics, particularly in optics, more accurate numbers are needed and we use more than three significant figures. Finally, if a number is *exact*, such as the two in the formula for the circumference of a circle, $C = 2\pi r$, it does not affect the number of significant figures in a calculation. Likewise, conversion factors such as 100 cm/1 m are considered exact and do not affect the number of significant figures in a calculation.

1.8 | Solving Problems in Physics

Learning Objectives

By the end of this section, you will be able to:

- Describe the process for developing a problem-solving strategy.
- Explain how to find the numerical solution to a problem.
- Summarize the process for assessing the significance of the numerical solution to a problem.

Figure 1.18 Problem-solving skills are essential to your success in physics. (credit: “scui3asteveo”/Flickr)

Problem-solving skills are clearly essential to success in a quantitative course in physics. More important, the ability to apply broad physical principles—usually represented by equations—to specific situations is a very powerful form of knowledge. It is much more powerful than memorizing a list of facts. Analytical skills and problem-solving abilities can be applied to new situations whereas a list of facts cannot be made long enough to contain every possible circumstance. Such analytical skills are useful both for solving problems in this text and for applying physics in everyday life.

As you are probably well aware, a certain amount of creativity and insight is required to solve problems. No rigid procedure works every time. Creativity and insight grow with experience. With practice, the basics of problem solving become almost automatic. One way to get practice is to work out the text’s examples for yourself as you read. Another is to work as many end-of-section problems as possible, starting with the easiest to build confidence and then progressing to the more difficult. After you become involved in physics, you will see it all around you, and you can begin to apply it to situations you encounter outside the classroom, just as is done in many of the applications in this text.

Although there is no simple step-by-step method that works for every problem, the following three-stage process facilitates problem solving and makes it more meaningful. The three stages are strategy, solution, and significance. This process is used in examples throughout the book. Here, we look at each stage of the process in turn.

Strategy

Strategy is the beginning stage of solving a problem. The idea is to figure out exactly what the problem is and then develop a strategy for solving it. Some general advice for this stage is as follows:

- *Examine the situation to determine which physical principles are involved*. It often helps to *draw a simple sketch* at the outset. You often need to decide which direction is positive and note that on your sketch. When you have identified the physical principles, it is much easier to find and apply the equations representing those principles.

Although finding the correct equation is essential, keep in mind that equations represent physical principles, laws of nature, and relationships among physical quantities. Without a conceptual understanding of a problem, a numerical solution is meaningless.

- *Make a list of what is given or can be inferred from the problem as stated (identify the "knowns")*. Many problems are stated very succinctly and require some inspection to determine what is known. Drawing a sketch can be very useful at this point as well. Formally identifying the knowns is of particular importance in applying physics to real-world situations. For example, the word *stopped* means the velocity is zero at that instant. Also, we can often take initial time and position as zero by the appropriate choice of coordinate system.
- *Identify exactly what needs to be determined in the problem (identify the unknowns)*. In complex problems, especially, it is not always obvious what needs to be found or in what sequence. Making a list can help identify the unknowns.
- *Determine which physical principles can help you solve the problem*. Since physical principles tend to be expressed in the form of mathematical equations, a list of knowns and unknowns can help here. It is easiest if you can find equations that contain only one unknown—that is, all the other variables are known—so you can solve for the unknown easily. If the equation contains more than one unknown, then additional equations are needed to solve the problem. In some problems, several unknowns must be determined to get at the one needed most. In such problems it is especially important to keep physical principles in mind to avoid going astray in a sea of equations. You may have to use two (or more) different equations to get the final answer.

Solution

The solution stage is when you do the math. *Substitute the knowns (along with their units) into the appropriate equation and obtain numerical solutions complete with units*. That is, do the algebra, calculus, geometry, or arithmetic necessary to find the unknown from the knowns, being sure to carry the units through the calculations. This step is clearly important because it produces the numerical answer, along with its units. Notice, however, that this stage is only one-third of the overall problem-solving process.

Significance

After having done the math in the solution stage of problem solving, it is tempting to think you are done. But, always remember that physics is not math. Rather, in doing physics, we use mathematics as a tool to help us understand nature. So, after you obtain a numerical answer, you should always assess its significance:

- *Check your units.* If the units of the answer are incorrect, then an error has been made and you should go back over your previous steps to find it. One way to find the mistake is to check all the equations you derived for dimensional consistency. However, be warned that correct units do not guarantee the numerical part of the answer is also correct.
- *Check the answer to see whether it is reasonable. Does it make sense?* This step is extremely important: –the goal of physics is to describe nature accurately. To determine whether the answer is reasonable, check both its magnitude and its sign, in addition to its units. The magnitude should be consistent with a rough estimate of what it should be. It should also compare reasonably with magnitudes of other quantities of the same type. The sign usually tells you about direction and should be consistent with your prior expectations. Your judgment will improve as you solve more physics problems, and it will become possible for you to make finer judgments regarding whether nature is described adequately by the answer to a problem. This step brings the problem back to its conceptual meaning. If you can judge whether the answer is reasonable, you have a deeper understanding of physics than just being able to solve a problem mechanically.
- *Check to see whether the answer tells you something interesting. What does it mean?* This is the flip side of the question: Does it make sense? Ultimately, physics is about understanding nature, and we solve physics problems to learn a little something about how nature operates. Therefore, assuming the answer does make sense, you should always take a moment to see if it tells you something about the world that you find interesting. Even if the answer to this particular problem is not very interesting to you, what about the method you used to solve it? Could the method be adapted to answer a question that you do find interesting? In many ways, it is in answering questions such as these that science progresses.

CHAPTER 1 REVIEW

KEY TERMS

accuracy the degree to which a measured value agrees with an accepted reference value for that measurement

base quantity physical quantity chosen by convention and practical considerations such that all other physical quantities can be expressed as algebraic combinations of them

base unit standard for expressing the measurement of a base quantity within a particular system of units; defined by a particular procedure used to measure the corresponding base quantity

conversion factor a ratio that expresses how many of one unit are equal to another unit

derived quantity physical quantity defined using algebraic combinations of base quantities

derived units units that can be calculated using algebraic combinations of the fundamental units

dimension expression of the dependence of a physical quantity on the base quantities as a product of powers of symbols representing the base quantities; in general, the dimension of a quantity has the form $\text{L}^a\text{M}^b\text{T}^c\text{I}^d\Theta^e\text{N}^f\text{J}^g$ for some powers a, b, c, d, e, f, and g.

dimensionally consistent equation in which every term has the same dimensions and the arguments of any mathematical functions appearing in the equation are dimensionless

dimensionless quantity with a dimension of $\text{L}^0\text{M}^0\text{T}^0\text{I}^0\Theta^0\text{N}^0\text{J}^0 = 1;$ also called quantity of dimension 1 or a pure number

discrepancy the difference between the measured value and a given standard or expected value

English units system of measurement used in the United States; includes units of measure such as feet, gallons, and pounds

estimation using prior experience and sound physical reasoning to arrive at a rough idea of a quantity's value; sometimes called an "order-of-magnitude approximation," a "guesstimate," a "back-of-the-envelope calculation", or a "Fermi calculation"

kilogram SI unit for mass, abbreviated kg

law description, using concise language or a mathematical formula, of a generalized pattern in nature supported by scientific evidence and repeated experiments

meter SI unit for length, abbreviated m

method of adding percents the percent uncertainty in a quantity calculated by multiplication or division is the sum of the percent uncertainties in the items used to make the calculation.

metric system system in which values can be calculated in factors of 10

model representation of something often too difficult (or impossible) to display directly

order of magnitude the size of a quantity as it relates to a power of 10

percent uncertainty the ratio of the uncertainty of a measurement to the measured value, expressed as a percentage

physical quantity characteristic or property of an object that can be measured or calculated from other measurements

physics science concerned with describing the interactions of energy, matter, space, and time; especially interested in what fundamental mechanisms underlie every phenomenon

precision the degree to which repeated measurements agree with each other

second the SI unit for time, abbreviated s

SI units the international system of units that scientists in most countries have agreed to use; includes units such as meters, liters, and grams

significant figures used to express the precision of a measuring tool used to measure a value

theory testable explanation for patterns in nature supported by scientific evidence and verified multiple times by various groups of researchers

uncertainty a quantitative measure of how much measured values deviate from one another

units standards used for expressing and comparing measurements

KEY EQUATIONS

Percent uncertainty	$\text{Percent uncertainty} = \frac{\delta A}{A} \times 100\%$
Adding uncertainties in quadrature	If $z = x \pm y$, then $\delta z = \sqrt{\delta x^2 + \delta y^2}$
Adding relative uncertainties in quadrature	If $z = x \times y$ or $z = x \div y$ then $\frac{\delta z}{z} = \sqrt{\left(\frac{\delta x}{x}\right)^2 + \left(\frac{\delta y}{y}\right)^2}$

SUMMARY

1.2 The Scope and Scale of Physics

- Physics is about trying to find the simple laws that describe all natural phenomena.
- Physics operates on a vast range of scales of length, mass, and time. Scientists use the concept of the order of magnitude of a number to track which phenomena occur on which scales. They also use orders of magnitude to compare the various scales.
- Scientists attempt to describe the world by formulating models, theories, and laws.

1.3 Units and Standards

- Systems of units are built up from a small number of base units, which are defined by accurate and precise measurements of conventionally chosen base quantities. Other units are then derived as algebraic combinations of the base units.
- Two commonly used systems of units are English units and SI units. All scientists and most of the other people in the world use SI, whereas nonscientists in the United States still tend to use English units.
- The SI base units of length, mass, and time are the meter (m), kilogram (kg), and second (s), respectively.
- SI units are a metric system of units, meaning values can be calculated by factors of 10. Metric prefixes may be used with metric units to scale the base units to sizes appropriate for almost any application.

1.4 Unit Conversion

- To convert a quantity from one unit to another, multiply by conversions factors in such a way that you cancel the units you want to get rid of and introduce the units you want to end up with.
- Be careful with areas and volumes. Units obey the rules of algebra so, for example, if a unit is squared we need two factors to cancel it.

1.5 Dimensional Analysis

- The dimension of a physical quantity is just an expression of the base quantities from which it is derived.
- All equations expressing physical laws or principles must be dimensionally consistent. This fact can be used as an aid in remembering physical laws, as a way to check whether claimed relationships between physical quantities are possible, and even to derive new physical laws.

1.6 Estimates and Fermi Calculations

- An estimate is a rough educated guess at the value of a physical quantity based on prior experience and sound physical reasoning. Some strategies that may help when making an estimate are as follows:

- Get big lengths from smaller lengths.
- Get areas and volumes from lengths.
- Get masses from volumes and densities.
- If all else fails, bound it.
- One “sig. fig.” is fine.
- Ask yourself: Does this make any sense?

1.7 Uncertainties and Significant Figures

- Accuracy of a measured value refers to how close a measurement is to an accepted reference value. The discrepancy in a measurement is the amount by which the measurement result differs from this value.
- Precision of measured values refers to how close the agreement is between repeated measurements. The uncertainty of a measurement is a quantification of this.
- The precision of a measuring tool is related to the size of its measurement increments. The smaller the measurement increment, the more precise the tool.
- Significant figures express the precision of a measuring tool.
- When multiplying or dividing measured values, the final answer can contain only as many significant figures as the least-precise value.
- When adding or subtracting measured values, the final answer cannot contain more decimal places than the least-precise value.

1.8 Solving Problems in Physics

The three stages of the process for solving physics problems used in this book are as follows:

- *Strategy*: Determine which physical principles are involved and develop a strategy for using them to solve the problem.
- *Solution*: Do the math necessary to obtain a numerical solution complete with units.
- *Significance*: Check the solution to make sure it makes sense (correct units, reasonable magnitude and sign) and assess its significance.

CONCEPTUAL QUESTIONS

1.2 The Scope and Scale of Physics

1. What is physics?

2. Some have described physics as a “search for simplicity.” Explain why this might be an appropriate description.

3. If two different theories describe experimental observations equally well, can one be said to be more valid than the other (assuming both use accepted rules of logic)?

4. What determines the validity of a theory?

5. Certain criteria must be satisfied if a measurement or observation is to be believed. Will the criteria necessarily be as strict for an expected result as for an unexpected result?

6. Can the validity of a model be limited or must it be universally valid? How does this compare with the required validity of a theory or a law?

1.3 Units and Standards

7. Identify some advantages of metric units.

8. What are the SI base units of length, mass, and time?

9. What is the difference between a base unit and a derived unit? (b) What is the difference between a base quantity and a derived quantity? (c) What is the difference between a base quantity and a base unit?

10. For each of the following scenarios, refer to **Figure 1.9** and **Table 1.2** to determine which metric prefix on the meter is most appropriate for each of the following scenarios. (a) You want to tabulate the mean distance from the Sun for each planet in the solar system. (b) You want to compare the sizes of some common viruses to design a mechanical filter capable of blocking the pathogenic ones. (c) You want to list the diameters of all the elements on the periodic table. (d) You want to list the distances to all the stars that have now received any radio broadcasts sent from Earth 10 years ago.

1.7 Uncertainties and Significant Figures

11. (a) What is the relationship between the precision and the uncertainty of a measurement? (b) What is the relationship between the accuracy and the discrepancy of a measurement?

1.8 Solving Problems in Physics

12. What information do you need to choose which equation or equations to use to solve a problem?

13. What should you do after obtaining a numerical answer when solving a problem?

PROBLEMS

1.2 The Scope and Scale of Physics

14. Find the order of magnitude of the following physical quantities. (a) The mass of Earth's atmosphere: $5.1 \times 10^{18}\,\text{kg}$; (b) The mass of the Moon's atmosphere: 25,000 kg; (c) The mass of Earth's hydrosphere: $1.4 \times 10^{21}\,\text{kg}$; (d) The mass of Earth: $5.97 \times 10^{24}\,\text{kg}$; (e) The mass of the Moon: $7.34 \times 10^{22}\,\text{kg}$; (f) The Earth–Moon distance (semimajor axis): $3.84 \times 10^{8}\,\text{m}$; (g) The mean Earth–Sun distance: $1.5 \times 10^{11}\,\text{m}$; (h) The equatorial radius of Earth: $6.38 \times 10^{6}\,\text{m}$; (i) The mass of an electron: $9.11 \times 10^{-31}\,\text{kg}$; (j) The mass of a proton: $1.67 \times 10^{-27}\,\text{kg}$; (k) The mass of the Sun: $1.99 \times 10^{30}\,\text{kg}$.

15. Use the orders of magnitude you found in the previous problem to answer the following questions to within an order of magnitude. (a) How many electrons would it take to equal the mass of a proton? (b) How many Earths would it take to equal the mass of the Sun? (c) How many Earth–Moon distances would it take to cover the distance from Earth to the Sun? (d) How many Moon atmospheres would it take to equal the mass of Earth's atmosphere? (e) How many moons would it take to equal the mass of Earth? (f) How many protons would it take to equal the mass of the Sun?

For the remaining questions, you need to use **Figure 1.9** to obtain the necessary orders of magnitude of lengths, masses, and times.

16. Roughly how many heartbeats are there in a lifetime?

17. A generation is about one-third of a lifetime. Approximately how many generations have passed since the year 0 AD?

18. Roughly how many times longer than the mean life of an extremely unstable atomic nucleus is the lifetime of a human?

19. Calculate the approximate number of atoms in a bacterium. Assume the average mass of an atom in the bacterium is 10 times the mass of a proton.

20. (a) Calculate the number of cells in a hummingbird assuming the mass of an average cell is 10 times the mass of a bacterium. (b) Making the same assumption, how many cells are there in a human?

21. Assuming one nerve impulse must end before another can begin, what is the maximum firing rate of a nerve in impulses per second?

22. About how many floating-point operations can a supercomputer perform each year?

23. Roughly how many floating-point operations can a supercomputer perform in a human lifetime?

1.3 Units and Standards

24. The following times are given using metric prefixes on the base SI unit of time: the second. Rewrite them in scientific notation without the prefix. For example, 47 Ts would be rewritten as $4.7 \times 10^{13}\,\text{s}$. (a) 980 Ps; (b) 980 fs; (c) 17 ns; (d) $577\,\mu\text{s}$.

25. The following times are given in seconds. Use metric prefixes to rewrite them so the numerical value is greater than one but less than 1000. For example, $7.9 \times 10^{-2}\,\text{s}$ could be written as either 7.9 cs or 79 ms. (a) $9.57 \times 10^{5}\,\text{s}$; (b) 0.045 s; (c) $5.5 \times 10^{-7}\,\text{s}$; (d) $3.16 \times 10^{7}\,\text{s}$.

26. The following lengths are given using metric prefixes on the base SI unit of length: the meter. Rewrite them in scientific notation without the prefix. For example, 4.2 Pm would be rewritten as $4.2 \times 10^{15}\,\text{m}$. (a) 89 Tm; (b) 89 pm; (c) 711 mm; (d) $0.45\,\mu\text{m}$.

27. The following lengths are given in meters. Use metric prefixes to rewrite them so the numerical value is bigger than one but less than 1000. For example, $7.9 \times 10^{-2}\,\text{m}$ could be written either as 7.9 cm or 79 mm. (a) $7.59 \times 10^{7}\,\text{m}$; (b) 0.0074 m; (c) $8.8 \times 10^{-11}\,\text{m}$; (d) $1.63 \times 10^{13}\,\text{m}$.

28. The following masses are written using metric prefixes on the gram. Rewrite them in scientific notation in terms of the SI base unit of mass: the kilogram. For example, 40 Mg would be written as $4 \times 10^{4}\,\text{kg}$. (a) 23 mg; (b) 320 Tg; (c) 42 ng; (d) 7 g; (e) 9 Pg.

29. The following masses are given in kilograms. Use metric prefixes on the gram to rewrite them so the numerical value is bigger than one but less than 1000. For example, $7 \times 10^{-4}\,\text{kg}$ could be written as 70 cg or 700 mg. (a) $3.8 \times 10^{-5}\,\text{kg}$; (b) $2.3 \times 10^{17}\,\text{kg}$; (c) $2.4 \times 10^{-11}\,\text{kg}$; (d) $8 \times 10^{15}\,\text{kg}$; (e) $4.2 \times 10^{-3}\,\text{kg}$.

1.4 Unit Conversion

30. The volume of Earth is on the order of 10^{21} m^3. (a) What is this in cubic kilometers (km^3)? (b) What is it in cubic miles (mi^3)? (c) What is it in cubic centimeters (cm^3)?

31. The speed limit on some interstate highways is roughly 100 km/h. (a) What is this in meters per second? (b) How many miles per hour is this?

32. A car is traveling at a speed of 33 m/s. (a) What is its speed in kilometers per hour? (b) Is it exceeding the 90 km/h speed limit?

33. In SI units, speeds are measured in meters per second (m/s). But, depending on where you live, you're probably more comfortable of thinking of speeds in terms of either kilometers per hour (km/h) or miles per hour (mi/h). In this problem, you will see that 1 m/s is roughly 4 km/h or 2 mi/h, which is handy to use when developing your physical intuition. More precisely, show that (a) $1.0\,\text{m/s} = 3.6\,\text{km/h}$ and (b) $1.0\,\text{m/s} = 2.2\,\text{mi/h}$.

34. American football is played on a 100-yd-long field, excluding the end zones. How long is the field in meters? (Assume that 1 m = 3.281 ft.)

35. Soccer fields vary in size. A large soccer field is 115 m long and 85.0 m wide. What is its area in square feet? (Assume that 1 m = 3.281 ft.)

36. What is the height in meters of a person who is 6 ft 1.0 in. tall?

37. Mount Everest, at 29,028 ft, is the tallest mountain on Earth. What is its height in kilometers? (Assume that 1 m = 3.281 ft.)

38. The speed of sound is measured to be 342 m/s on a certain day. What is this measurement in kilometers per hour?

39. Tectonic plates are large segments of Earth's crust that move slowly. Suppose one such plate has an average speed of 4.0 cm/yr. (a) What distance does it move in 1.0 s at this speed? (b) What is its speed in kilometers per million years?

40. The average distance between Earth and the Sun is $1.5 \times 10^{11}\,\text{m}$. (a) Calculate the average speed of Earth in its orbit (assumed to be circular) in meters per second. (b) What is this speed in miles per hour?

41. The density of nuclear matter is about 10^{18} kg/m^3. Given that 1 mL is equal in volume to cm^3, what is the density of nuclear matter in megagrams per microliter (that is, $\text{Mg}/\mu\text{L}$)?

42. The density of aluminum is 2.7 g/cm^3. What is the density in kilograms per cubic meter?

43. A commonly used unit of mass in the English system is the pound-mass, abbreviated lbm, where 1 lbm = 0.454 kg. What is the density of water in pound-mass per cubic foot?

44. A furlong is 220 yd. A fortnight is 2 weeks. Convert a speed of one furlong per fortnight to millimeters per second.

45. It takes 2π radians (rad) to get around a circle, which is the same as 360°. How many radians are in 1°?

46. Light travels a distance of about $3 \times 10^{8}\,\text{m/s}$. A light-minute is the distance light travels in 1 min. If the Sun is $1.5 \times 10^{11}\,\text{m}$ from Earth, how far away is it in light-minutes?

47. A light-nanosecond is the distance light travels in 1 ns. Convert 1 ft to light-nanoseconds.

48. An electron has a mass of 9.11×10^{-31} kg. A proton has a mass of 1.67×10^{-27} kg. What is the mass of a proton in electron-masses?

49. A fluid ounce is about 30 mL. What is the volume of a 12 fl-oz can of soda pop in cubic meters?

1.5 Dimensional Analysis

50. A student is trying to remember some formulas from geometry. In what follows, assume A is area, V is volume, and all other variables are lengths. Determine which formulas are dimensionally consistent. (a) $V = \pi r^2 h$; (b) $A = 2\pi r^2 + 2\pi rh$; (c) $V = 0.5bh$; (d) $V = \pi d^2$; (e) $V = \pi d^3/6$.

51. Consider the physical quantities s, v, a, and t with dimensions $[s] = \text{L}$, $[v] = \text{LT}^{-1}$, $[a] = \text{LT}^{-2}$, and $[t] = \text{T}$. Determine whether each of the following equations is dimensionally consistent. (a) $v^2 = 2as$; (b) $s = vt^2 + 0.5at^2$; (c) $v = s/t$; (d) $a = v/t$.

52. Consider the physical quantities m, s, v, a, and t with dimensions $[m]$ = M, $[s]$ = L, $[v]$ = LT^{-1}, $[a]$ = LT^{-2}, and $[t]$ = T. Assuming each of the following equations is dimensionally consistent, find the dimension of the quantity on the left-hand side of the equation: (a) $F = ma$; (b) $K = 0.5mv^2$; (c) $p = mv$; (d) $W = mas$; (e) $L = mvr$.

53. Suppose quantity s is a length and quantity t is a time. Suppose the quantities v and a are defined by $v = ds/dt$ and $a = dv/dt$. (a) What is the dimension of v? (b) What is the dimension of the quantity a? What are the dimensions of (c) $\int v dt$, (d) $\int a dt$, and (e) da/dt?

54. Suppose [V] = L^3, $[\rho] = \text{ML}^{-3}$, and [t] = T. (a) What is the dimension of $\int \rho dV$? (b) What is the dimension of dV/dt? (c) What is the dimension of $\rho(dV/dt)$?

55. The arc length formula says the length s of arc subtended by angle Θ in a circle of radius r is given by the equation $s = r\Theta$. What are the dimensions of (a) s, (b) r, and (c) Θ?

56. Sometimes it's good to use dimensional analysis to solve a problem for which there's no simple mathematical formula that leads directly to an answer. The next few problems provide an opportunity to do just that. Suppose that you own a car which gets 30 miles per gallon when driving on the highway at 60 miles per hour, but only gets 20 miles per gallon when driving on city streets at 30 miles per hour.

You drive for one hour on the highway, followed by one hour on the city streets. What would be your average fuel consumption (in miles per gallon) for the entire trip?

57. Suppose you drive the car described in **Exercise 1.56** for 60 miles along the highway followed by 60 miles on city streets. What would be your average fuel consumption (in miles per gallon) for the entire trip?

58. Suppose you drive the car described in **Exercise 1.56**, first using up 5 gallons of fuel driving along the highway followed by using up 5 gallons of fuel on city streets. What would be your average fuel consumption (in miles per gallon) for the entire trip?

1.6 Estimates and Fermi Calculations

59. Assuming the human body is made primarily of water, estimate the volume of a person.

60. Assuming the human body is primarily made of water, estimate the number of molecules in it. (Note that water has a molecular mass of 18 g/mol and there are roughly 10^{24} atoms in a mole.)

61. Estimate the mass of air in a classroom.

62. Estimate the number of molecules that make up Earth, assuming an average molecular mass of 30 g/mol. (Note there are on the order of 10^{24} objects per mole.)

63. Estimate the surface area of a person.

64. Roughly how many solar systems would it take to tile the disk of the Milky Way?

65. (a) Estimate the density of the Moon. (b) Estimate the diameter of the Moon. (c) Given that the Moon subtends at an angle of about half a degree in the sky, estimate its distance from Earth.

66. The average density of the Sun is on the order 10^3 kg/m^3. (a) Estimate the diameter of the Sun. (b) Given that the Sun subtends at an angle of about half a degree in the sky, estimate its distance from Earth.

67. Estimate the mass of a virus.

68. A floating-point operation is a single arithmetic operation such as addition, subtraction, multiplication, or division. (a) Estimate the maximum number of floating-point operations a human being could possibly perform in a lifetime. (b) How long would it take a supercomputer to perform that many floating-point operations?

1.7 Uncertainties and Significant Figures

69. Consider the equation 4000/400 = 10.0. Assuming the number of significant figures in the answer is correct, what can you say about the number of significant figures in 4000 and 400?

70. Suppose your bathroom scale reads your mass as 65 kg with a 3% uncertainty. What is the uncertainty in your mass (in kilograms)?

71. A good-quality measuring tape can be off by 0.50 cm over a distance of 20 m. What is its percent uncertainty?

72. An infant's pulse rate is measured to be 130 ± 5 beats/min. What is the percent uncertainty in this measurement?

73. (a) Suppose that a person has an average heart rate of 72.0 beats/min. How many beats does he or she have in 2.0 years? (b) In 2.00 years? (c) In 2.000 years?

74. A can contains 375 mL of soda. How much is left after 308 mL is removed?

75. State how many significant figures are proper in the results of the following calculations: (a) $(106.7)(98.2)/(46.210)(1.01)$; (b) $(18.7)^2$; (c) $\left(1.60 \times 10^{-19}\right)(3712)$

76. (a) How many significant figures are in the numbers 99 and 100.? (b) If the uncertainty in each number is 1, what is the percent uncertainty in each? (c) Which is a more meaningful way to express the accuracy of these two numbers: significant figures or percent uncertainties?

77. (a) If your speedometer has an uncertainty of 2.0 km/h at a speed of 90 km/h, what is the percent uncertainty? (b) If it has the same percent uncertainty when it reads 60 km/h, what is the range of speeds you could be going?

78. (a) A person's blood pressure is measured to be $120 \pm 2 \text{ mm Hg}$. What is its percent uncertainty? (b) Assuming the same percent uncertainty, what is the uncertainty in a blood pressure measurement of 80 mm Hg?

79. A person measures his or her heart rate by counting the number of beats in 30 s. If 40 ± 1 beats are counted in 30.0 ± 0.5 s, what is the heart rate and its uncertainty in beats per minute?

80. What is the area of a circle 3.102 cm in diameter?

81. Determine the number of significant figures in the following measurements: (a) 0.0009, (b) 15,450.0, (c) 6×10^3, (d) 87.990, and (e) 30.42.

82. Perform the following calculations and express your answer using the correct number of significant digits. (a) A woman has two bags weighing 13.5 lb and one bag with a weight of 10.2 lb. What is the total weight of the bags? (b) The force F on an object is equal to its mass m multiplied by its acceleration a. If a wagon with mass 55 kg accelerates at a rate of 0.0255 m/s^2, what is the force on the wagon? (The unit of force is called the *newton* and it is expressed with the symbol N.)

ADDITIONAL PROBLEMS

83. Consider the equation $y = mt + b$, where the dimension of y is length and the dimension of t is time, and m and b are constants. What are the dimensions and SI units of (a) m and (b) b?

84. Consider the equation $s = s_0 + v_0 t + a_0 t^2/2 + j_0 t^3/6 + S_0 t^4/24 + ct^5/120$, where s is a length and t is a time. What are the dimensions and SI units of (a) s_0, (b) v_0, (c) a_0, (d) j_0, (e) S_0, and (f) c?

85. (a) A car speedometer has a 5% uncertainty. What is the range of possible speeds when it reads 90 km/h? (b) Convert this range to miles per hour. Note 1 km = 0.6214 mi.

86. A marathon runner completes a 42.188-km course in 2 h, 30 min, and 12 s. There is an uncertainty of 25 m in the distance traveled and an uncertainty of 1 s in the elapsed time. (a) Calculate the percent uncertainty in the distance. (b) Calculate the percent uncertainty in the elapsed time. (c) What is the average speed in meters per second? (d) What is the uncertainty in the average speed?

87. The sides of a small rectangular box are measured to be 1.80 ± 0.1 cm, 2.05 ± 0.02 cm, and 3.1 ± 0.1 cm long. Calculate its volume and uncertainty in cubic centimeters.

88. When nonmetric units were used in the United Kingdom, a unit of mass called the pound-mass (lbm) was used, where 1 lbm = 0.4539 kg. (a) If there is an uncertainty of 0.0001 kg in the pound-mass unit, what is its percent uncertainty? (b) Based on that percent uncertainty, what mass in pound-mass has an uncertainty of 1 kg when converted to kilograms?

89. The length and width of a rectangular room are measured to be 3.955 ± 0.005 m and 3.050 ± 0.005 m. Calculate the area of the room and its uncertainty in square meters.

90. A car engine moves a piston with a circular cross-section of 7.500 ± 0.002 cm in diameter a distance of 3.250 ± 0.001 cm to compress the gas in the cylinder. (a) By what amount is the gas decreased in volume in cubic centimeters? (b) Find the uncertainty in this volume.

CHALLENGE PROBLEMS

91. The first atomic bomb was detonated on July 16, 1945, at the Trinity test site about 200 mi south of Los Alamos. In 1947, the U.S. government declassified a film reel of the explosion. From this film reel, British physicist G. I. Taylor was able to determine the rate at which the radius of the fireball from the blast grew. Using dimensional analysis, he was then able to deduce the amount of energy released in the explosion, which was a closely guarded secret at the time. Because of this, Taylor did not publish his results until 1950. This problem challenges you to recreate this famous calculation. (a) Using keen physical insight developed from years of experience, Taylor decided the radius *r* of the fireball should depend only on time since the explosion, *t*, the density of the air, ρ, and the energy of the initial explosion, *E*. Thus, he made the educated guess that $r = kE^a \rho^b t^c$ for some dimensionless constant *k* and some unknown exponents *a*, *b*, and *c*. Given that [E] = ML^2T^{-2}, determine the values of the exponents necessary to make this equation dimensionally consistent. (*Hint*: Notice the equation implies that $k = rE^{-a} \rho^{-b} t^{-c}$ and that $[k] = 1$.) (b) By analyzing data from high-energy conventional explosives, Taylor found the formula he derived seemed to be valid as long as the constant *k* had the value 1.03. From the film reel, he was able to determine many values of *r* and the corresponding values of *t*. For example, he found that after 25.0 ms, the fireball had a radius of 130.0 m. Use these values, along with an average air density of 1.25 kg/m^3, to calculate the initial energy release of the Trinity detonation in joules (J). (*Hint*: To get energy in joules, you need to make sure all the numbers you substitute in are expressed in terms of SI base units.) (c) The energy released in large explosions is often cited in units of "tons of TNT" (abbreviated "t TNT"), where 1 t TNT is about 4.2 GJ. Convert your answer to (b) into kilotons of TNT (that is, kt TNT). Compare your answer with the quick-and-dirty estimate of 10 kt TNT made by physicist Enrico Fermi shortly after witnessing the explosion from what was thought to be a safe distance. (Reportedly, Fermi made his estimate by dropping some shredded bits of paper right before the remnants of the shock wave hit him and looked to see how far they were carried by it.)

92. The purpose of this problem is to show the entire concept of dimensional consistency can be summarized by the old saying "You can't add apples and oranges." If you have studied power series expansions in a calculus course, you know the standard mathematical functions such as trigonometric functions, logarithms, and exponential functions can be expressed as infinite sums of the form $\sum_{n=0}^{\infty} a_n x^n = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots ,$ where the a_n are dimensionless constants for all $n = 0, 1, 2, \cdots$ and *x* is the argument of the function. (If you have not studied power series in calculus yet, just trust us.) Use this fact to explain why the requirement that all terms in an equation have the same dimensions is sufficient as a definition of dimensional consistency. That is, it actually implies the arguments of standard mathematical functions must be dimensionless, so it is not really necessary to make this latter condition a separate requirement of the definition of dimensional consistency as we have done in this section.

2 | THE UNIVERSE AT ITS LIMITS

Chapter Outline

Introduction

In astronomy we deal with distances on a scale you may never have thought about before, with numbers larger than any you may have encountered. We adopt two approaches that make dealing with astronomical numbers a little bit easier. First, we use a system for writing large and small numbers called *scientific notation* (or sometimes *powers-of-ten notation*). This system is very appealing because it eliminates the many zeros that can seem overwhelming to the reader. In scientific notation, if you want to write a number such as 500,000,000, you express it as 5×10^8. The small raised number after the 10, called an *exponent*, keeps track of the number of places we had to move the decimal point to the left to convert 500,000,000 to 5. If you are encountering this system for the first time or would like a refresher, we suggest you look back at **Section 1.3** for more information. The second way we try to keep numbers simple is to use a consistent set of units—the metric International System of Units, or SI (from the French *Système International d'Unités*). The SI system is summarized in **Appendix A**.

Watch this **brief PBS animation (https://openstax.org/l/30scinotation)** that explains how scientific notation works and why it's useful.

Figure 2.1 This beautiful cloud of cosmic raw material (gas and dust from which new stars and planets are being made) called the Orion Nebula is about 1400 light-years away. That's a distance of roughly 1.34×10^{16} kilometers—a pretty big number. The gas and dust in this region are illuminated by the intense light from a few extremely energetic adolescent stars. (credit: NASA, ESA, M. Robberto (Space Telescope Science Institute/ESA) and the Hubble Space Telescope Orion Treasury Project Team)

A common unit astronomers use to describe distances in the universe is a light-year, which is the distance light travels during one year. Because light always travels at the same speed, and because its speed turns out to be the fastest possible speed in the universe, it makes a good standard for keeping track of distances. You might be confused because a "light-year" seems to imply that we are measuring time, but this mix-up of time and distance is common in everyday life as well. For example, when your friend asks where the movie theater is located, you might say "about 20 minutes from downtown."

So, how many kilometers are there in a light-year? Light travels at the amazing pace of 3×10^5 kilometers per second (km/s), which makes a light-year 9.46×10^{12} kilometers. You might think that such a large unit would reach the nearest star easily, but the stars are far more remote than our imaginations might lead us to believe. Even the nearest star is 4.3 light-years away—more than 40 trillion kilometers. Other stars visible to the unaided eye are hundreds to thousands of light-years away (**Figure 2.1**).

Example 2.1.

Scientific Notation

In 2015, the richest human being on our planet had a net worth of \$79.2 billion. Some might say this is an astronomical sum of money. Express this amount in scientific notation.

Solution

\$79.2 billion can be written \$79,200,000,000. Expressed in scientific notation it becomes \$$7.92 \times 10^{10}$.

Example 2.2.

Getting Familiar with a Light-Year

How many kilometers are there in a light-year?

Solution

The speed of light, $c = 3 \times 10^8$ m/s. That is to say, light travels 3×10^8 m in 1 s. So, let's calculate how far it goes in a year. That's the meaning of "one light-year."

We'll use the same "Power of 1" idea from Chapter 1, i.e. that multiplying a quantity by any fraction whose value is precisely 1.0 does not change the quantity.

$$\frac{3\times10^{8}\text{ m}}{1\text{ s}}\times\left(\frac{60\text{ s}}{1\text{ min}}\right)\times\left(\frac{60\text{ min}}{1\text{ hr}}\right)\times\left(\frac{24\text{ hr}}{1\text{ day}}\right)\times\left(\frac{365.24\text{ day}}{1\text{ yr}}\right)=\frac{9.46\times10^{15}\text{ m}}{1\text{ yr}}$$

And, to answer the original question precisely, we will convert that distance into units of km. The final unit conversion gives us:

$$9.46\times10^{15}\text{ m}\times\left(\frac{1\text{ km}}{1000\text{ m}}\right)=9.46\times10^{12}\text{ km}$$

That's almost 10,000,000,000,000 km that light covers in a year. To help you imagine how long this distance is, we'll mention that a string 1 light-year long could fit around the circumference of Earth 236 million times.

2.1 | Consequences of Light Travel Time

Learning Objectives

After completing this section, you should be able to:

- Explain how looking farther out in space is also looking farther back in time

There is another reason the speed of light is such a natural unit of distance for astronomers. Information about the universe comes to us almost exclusively through various forms of light, and all such light travels at the speed of light—that is, 1 light-year every year. This sets a limit on how quickly we can learn about events in the universe. If a star is 100 light-years away, the light we see from it tonight left that star 100 years ago and is just now arriving in our neighborhood. The soonest we can learn about any changes in that star is 100 years after the fact. For a star 500 light-years away, the light we detect tonight left 500 years ago and is carrying 500-year-old news.

Because many of us are accustomed to instant news from the Internet, some might find this frustrating.

"You mean, when I see that star up there," you ask, "I won't know what's actually happening there for another 500 years?"

But this isn't the most helpful way to think about the situation. For astronomers, *now* is when the light reaches us here on Earth. There is no way for us to know anything about that star (or other object) until its light reaches us.

But what at first may seem a great frustration is actually a tremendous benefit in disguise. If astronomers really want to piece together what has happened in the universe since its beginning, they must find evidence about each epoch (or period of time) of the past. Where can we find evidence today about cosmic events that occurred billions of years ago?

The delay in the arrival of light provides an answer to this question. The farther out in space we look, the longer the light has taken to get here, and the longer ago it left its place of origin. By looking billions of light-years out into space, astronomers are actually seeing billions of years into the past. In this way, we can reconstruct the history of the cosmos and get a sense of how it has evolved over time.

This is one reason why astronomers strive to build telescopes that can collect more and more of the faint light in the universe. The more light we collect, the fainter the objects we can observe. On average, fainter objects are farther away and can, therefore, tell us about periods of time even deeper in the past. Instruments such as the Hubble Space Telescope (**Figure 2.2**) and the Very Large Telescope in Chile are giving astronomers views of deep space and deep time better than any we have had before.

Figure 2.2 The Hubble Space Telescope, shown here in orbit around Earth, is one of many astronomical instruments in space. (credit: modification of work by European Space Agency)

2.2 | A Tour of the Universe

Learning Objectives

By the end of this section you will be able to:

- Know the definition of one **astronomical unit** (AU).
- Explain the distance scales involved as we move outward from our planet through the solar system, to other stars in the Milky Way, and on to other galaxies.
- Describe the major bodies that make up our solar system.
- Describe the structure of the Milky Way galaxy.

We can now take a brief introductory tour of the universe as astronomers understand it today to get acquainted with the types of objects and distances you will encounter throughout the text. We begin at home with Earth, a nearly spherical planet about 13,000 kilometers in diameter (**Figure 2.3**). A space traveler entering our planetary system would easily distinguish Earth from the other planets in our solar system by the large amount of liquid water that covers some two thirds of its crust. If the traveler had equipment to receive radio or television signals, or came close enough to see the lights of our cities at night, she would soon find signs that this watery planet has sentient life.

Figure 2.3 This image shows the Western hemisphere as viewed from space 35,400 kilometers (about 22,000 miles) above Earth. Data about the land surface from one satellite was combined with another satellite's data about the clouds to create the image. (credit: modification of work by R. Stockli, A. Nelson, F. Hasler, NASA/ GSFC/ NOAA/ USGS)

Our nearest astronomical neighbor is Earth's satellite, commonly called the *Moon*. **Figure 2.4** shows Earth and the Moon drawn to scale on the same diagram. Notice how small we have to make these bodies to fit them on the page with the right scale. The Moon's distance from Earth is about 30 times Earth's diameter, or approximately 384,000 kilometers, and it takes about a month for the Moon to revolve around Earth. The Moon's diameter is 3476 kilometers, about one fourth the size of Earth.

Figure 2.4 This image shows Earth and the Moon shown to scale for both size and distance. (credit: modification of work by NASA)

Light (or radio waves) takes 1.3 seconds to travel between Earth and the Moon. If you've seen videos of the Apollo flights to the Moon, you may recall that there was a delay of about 3 seconds between the time Mission Control asked a question and the time the astronauts responded. This was not because the astronomers were thinking slowly, but rather because it took the radio waves almost 3 seconds to make the round trip.

Earth revolves around our star, the Sun, which is about 150 million kilometers away—approximately 400 times as far away from us as the Moon. We call the average Earth–Sun distance an *astronomical unit* (AU) because, in the early days of astronomy, it was the most important measuring standard. Light takes slightly more than 8 minutes to travel 1 astronomical unit, which means the latest news we receive from the Sun is always 8 minutes old. The diameter of the Sun is about 1.5 million kilometers; Earth could fit comfortably inside one of the minor eruptions that occurs on the surface of our star. If the Sun were reduced to the size of a basketball, Earth would be a small apple seed about 30 meters from the ball.

It takes Earth 1 year (3×10^7 seconds) to go around the Sun at our distance; to make it around, we must travel at approximately 110,000 kilometers per hour. (If you, like many students, still prefer miles to kilometers, you might find the following trick helpful. To convert kilometers to miles, just multiply kilometers by 0.6. Thus, 110,000 kilometers per hour becomes 66,000 miles per hour.) Because gravity holds us firmly to Earth and there is no resistance to Earth's motion in the vacuum of space, we participate in this extremely fast-moving trip without being aware of it day to day.

Earth is only one of eight planets that revolve around the Sun. These planets, along with their moons and swarms of smaller bodies such as dwarf planets, make up the solar system (**Figure 2.5**). A planet is defined as a body of significant size that orbits a star and does not produce its own light. (If a large body consistently produces its own light, it is then called a *star*.) Later in the book this definition will be modified a bit, but it is perfectly fine for now as you begin your voyage.

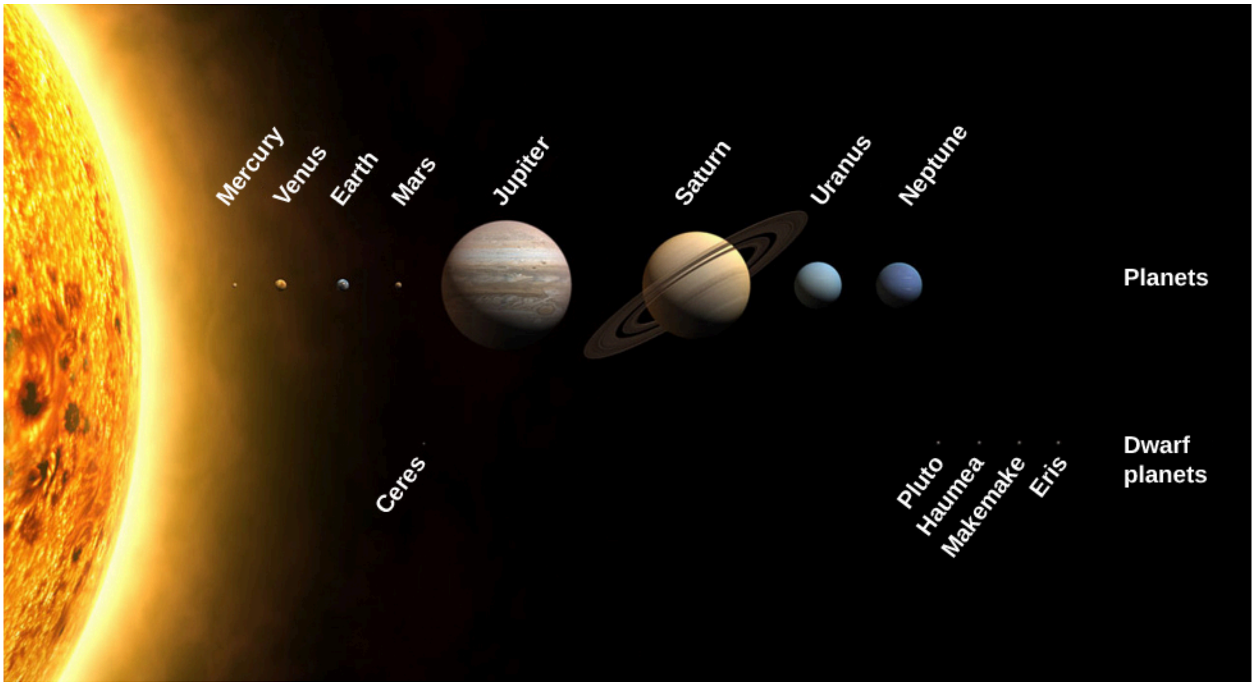


Figure 2.5 The Sun, the planets, and some dwarf planets are shown with their sizes drawn to scale. The orbits of the planets are much more widely separated than shown in this drawing. Notice the size of Earth compared to the giant planets. (credit: modification of work by NASA)

We are able to see the nearby planets in our skies only because they reflect the light of our local star, the Sun. If the planets were much farther away, the tiny amount of light they reflect would usually not be visible to us. The planets we have so far discovered orbiting other stars were found from the pull their gravity exerts on their parent stars, or from the light they block from their stars when they pass in front of them. We can't see most of these planets directly, although a few are now being imaged directly.

The Sun is our local star, and all the other stars are also enormous balls of glowing gas that generate vast amounts of energy by nuclear reactions deep within. We will discuss the processes that cause stars to shine in more detail later in the book. The other stars look faint only because they are so very far away. If we continue our basketball analogy, Proxima Centauri, the nearest star beyond the Sun, which is 4.3 light-years away, would be almost 7000 kilometers from the basketball.

When you look up at a star-filled sky on a clear night, all the stars visible to the unaided eye are part of a single collection of stars we call the *Milky Way Galaxy*, or simply the *Galaxy*. (When referring to the Milky Way, we capitalize *Galaxy*; when talking about other galaxies of stars, we use lowercase *galaxy*.) The Sun is one of hundreds of billions of stars that make up the Galaxy; its extent, as we will see, staggers the human imagination. Within a sphere 10 light-years in radius centered on the Sun, we find roughly ten stars. Within a sphere 100 light-years in radius, there are roughly 10,000 (10^4) stars—far too many to count or name—but we have still traversed only a tiny part of the Milky Way Galaxy. Within a 1000-light-year sphere, we find some ten million (10^7) stars; within a sphere of 100,000 light-years, we finally encompass the entire Milky Way Galaxy.

Our Galaxy looks like a giant disk with a small ball in the middle. If we could move outside our Galaxy and look down on the disk of the Milky Way from above, it would probably resemble the galaxy in **Figure 2.6**, with its spiral structure outlined by the blue light of hot adolescent stars.

Figure 2.6 This galaxy of billions of stars, called by its catalog number NGC 1073, is thought to be similar to our own Milky Way Galaxy. Here we see the giant wheel-shaped system with a bar of stars across its middle. (credit: NASA, ESA)

The Sun is somewhat less than 30,000 light-years from the center of the Galaxy, in a location with nothing much to distinguish it. From our position inside the Milky Way Galaxy, we cannot see through to its far rim (at least not with ordinary light) because the space between the stars is not completely empty. It contains a sparse distribution of gas (mostly the simplest element, hydrogen) intermixed with tiny solid particles that we call *interstellar dust*. This gas and dust collect into enormous clouds in many places in the Galaxy, becoming the raw material for future generations of stars. **Figure 2.7** shows an image of the disk of the Galaxy as seen from our vantage point.

Figure 2.7 Because we are inside the Milky Way Galaxy, we see its disk in cross-section flung across the sky like a great milky white avenue of stars with dark "rifts" of dust. In this dramatic image, part of it is seen above Trona Pinnacles in the California desert. (credit: Ian Norman)

Typically, the interstellar material is so extremely sparse that the space between stars is a much better vacuum than anything we can produce in terrestrial laboratories. Yet, the dust in space, building up over thousands of light-years, can block the light of more distant stars. Like the distant buildings that disappear from our view on a smoggy day in Los Angeles, the more distant regions of the Milky Way cannot be seen behind the layers of interstellar smog. Luckily, astronomers have found that stars and raw material shine with various forms of light, some of which do penetrate the smog, and so we have been able to develop a pretty good map of the Galaxy.

Recent observations, however, have also revealed a rather surprising and disturbing fact. There appears to be more—much more—to the Galaxy than meets the eye (or the telescope). From various investigations, we have evidence that much of our Galaxy is made of material we cannot currently observe directly with our instruments. We therefore call this component of the Galaxy *dark matter*. We know the dark matter is there by the pull its gravity exerts on the stars and raw material we can observe, but what this dark matter is made of and how much of it exists remain a mystery. Furthermore, this dark matter is not confined to our Galaxy; it appears to be an important part of other star groupings as well.

By the way, not all stars live by themselves, as the Sun does. Many are born in double or triple systems with two, three, or more stars revolving about each other. Because the stars influence each other in such close systems, multiple stars allow us to measure characteristics that we cannot discern from observing single stars. In a number of places, enough stars have formed together that we recognized them as star clusters (**Figure 2.8**). Some of the largest of the star clusters that astronomers have cataloged contain hundreds of thousands of stars and take up volumes of space hundreds of light-years across.

Figure 2.8 This large star cluster is known by its catalog number, M9. It contains some 250,000 stars and is seen more clearly from space using the Hubble Space Telescope. It is located roughly 25,000 light-years away. (credit: NASA, ESA)

You may hear stars referred to as "eternal," but in fact no star can last forever. Since the "business" of stars is making energy, and energy production requires some sort of fuel to be used up, eventually all stars run out of fuel. This news should not cause you to panic, though, because our Sun still has at least 5 or 6 billion years to go. Ultimately, the Sun and all stars will die, and it is in their death throes that some of the most intriguing and important processes of the universe are revealed. For example, we now know that many of the atoms in our bodies were once inside stars. These stars exploded at the ends of their lives, recycling their material back into the reservoir of the Galaxy. In this sense, all of us are literally made of recycled "star dust."

2.3 | The Universe on the Large Scale

Learning Objectives

By the end of this section you will be able to:

- Name the galaxies nearest to our own, along with their approximate distances from Earth.
- Explain what is meant by a **galactic cluster**.
- Explain what is meant by a **quasar**.

In a very rough sense, you could think of the solar system as your house or apartment and the Galaxy as your town, made up of many houses and buildings. In the twentieth century, astronomers were able to show that, just as our world is made up of many, many towns, so the universe is made up of enormous numbers of galaxies. (We define the universe to be everything that exists that is accessible to our observations.) Galaxies stretch as far into space as our telescopes can see, many billions of them within the reach of modern instruments. When they were first discovered, some astronomers called galaxies *island universes*, and the term is aptly descriptive; galaxies do look like islands of stars in the vast, dark seas of intergalactic space.

The nearest galaxy, discovered in 1993, is a small one that lies 75,000 light-years from the Sun in the direction of the constellation Sagittarius, where the smog in our own Galaxy makes it especially difficult to discern. (A constellation, we should note, is one of the 88 sections into which astronomers divide the sky, each named after a prominent star pattern within it.) Beyond this Sagittarius dwarf galaxy lie two other small galaxies, about 160,000 light-years away. First recorded by Magellan's crew as he sailed around the world, these are called the *Magellanic Clouds* (**Figure 2.9**). All three of these small galaxies are satellites of the Milky Way Galaxy, interacting with it through the force of gravity. Ultimately, all three may even be swallowed by our much larger Galaxy, as other small galaxies have been over the course of cosmic time.

Figure 2.9 This image shows both the Large Magellanic Cloud and the Small Magellanic Cloud above the telescopes of the Atacama Large Millimeter/Submillimeter Array (ALMA) in the Atacama Desert of northern Chile. (credit: ESO, C. Malin)

The nearest large galaxy is a spiral quite similar to our own, located in the constellation of Andromeda, and is thus called the Andromeda galaxy; it is also known by one of its catalog numbers, M31 (**Figure 2.10**). M31 is a little more than 2 million light-years away and, along with the Milky Way, is part of a small cluster of more than 50 galaxies referred to as the *Local Group*.

Figure 2.10 The Andromeda galaxy (M31) is a spiral-shaped collection of stars similar to our own Milky Way. (credit: Adam Evans)

At distances of 10 to 15 million light-years, we find other small galaxy groups, and then at about 50 million light-years there are more impressive systems with thousands of member galaxies. We have discovered that galaxies occur mostly in clusters, both large and small (**Figure 2.11**).

Figure 2.11 In this image, you can see part of a cluster of galaxies located about 60 million light-years away in the constellation of Fornax. All the objects that are not pinpoints of light in the picture are galaxies of billions of stars. (credit: ESO, J. Emerson, VISTA. Acknowledgment: Cambridge Astronomical Survey Unit)

Some of the clusters themselves form into larger groups called *superclusters*. The Local Group is part of a supercluster of galaxies, called the Virgo Supercluster, which stretches over a diameter of 110 million light-years. We are just beginning to explore the structure of the universe at these enormous scales and are already encountering some unexpected findings.

At even greater distances, where many ordinary galaxies are too dim to see, we find *quasars*. These are brilliant centers of galaxies, glowing with the light of an extraordinarily energetic process. The enormous energy of the quasars is produced by gas that is heated to a temperature of millions of degrees as it falls toward a massive black hole and swirls around it. The brilliance of quasars makes them the most distant beacons we can see in the dark oceans of space. They allow us to probe the universe 10 billion light-years away or more, and thus 10 billion years or more in the past.

With quasars we can see way back close to the Big Bang explosion that marks the beginning of time. Beyond the quasars and the most distant visible galaxies, we have detected the feeble glow of the explosion itself, filling the universe and thus coming to us from all directions in space. The discovery of this "afterglow of creation" is considered to be one of the most significant events in twentieth-century science, and we are still exploring the many things it has to tell us about the earliest times of the universe.

Measurements of the properties of galaxies and quasars in remote locations require large telescopes, sophisticated light-amplifying devices, and painstaking labor. Every clear night, at observatories around the world, astronomers and students are at work on such mysteries as the birth of new stars and the large-scale structure of the universe, fitting their results into the tapestry of our understanding.

2.4 | The Universe of the Very Small

Learning Objectives

By the end of this section, you will be able to:

- Explain what is meant by an **element**.
- Name the most abundant elements in the universe.
- Explain the difference between **atoms** and **molecules**.

The foregoing discussion has likely impressed on you that the universe is extraordinarily large and extraordinarily empty. On average, it is 10,000 times more empty than our Galaxy. Yet, as we have seen, even the Galaxy is mostly empty space. The air we breathe has about 10^{19} atoms in each cubic centimeter—and we usually think of air as empty space. In the interstellar gas of the Galaxy, there is about one atom in every cubic centimeter. Intergalactic space is filled so sparsely that to find one atom, on average, we must search through a cubic meter of space. Most of the universe is fantastically empty; places that are dense, such as the human body, are tremendously rare.

Even our most familiar solids are mostly space. If we could take apart such a solid, piece by piece, we would eventually reach the tiny molecules from which it is formed. Molecules are the smallest particles into which any matter can be divided while still retaining its chemical properties. A molecule of water (H_2O), for example, consists of two hydrogen atoms and one oxygen atom bonded together.

Molecules, in turn, are built of atoms, which are the smallest particles of an element that can still be identified as that element. For example, an atom of gold is the smallest possible piece of gold. Nearly 100 different kinds of atoms (elements) exist in nature. Most of them are rare, and only a handful account for more than 99% of everything with which we come in contact. The most abundant elements in the cosmos today are listed in **Table 2.1**; think of this table as the "greatest hits" of the universe when it comes to elements.

The Cosmically Abundant Elements

Element[1]	Symbol	Number of Atoms per Million Hydrogen Atoms
Hydrogen	H	1,000,000
Helium	He	80,000
Carbon	C	450
Nitrogen	N	92
Oxygen	O	740
Neon	Ne	130
Magnesium	Mg	40
Silicon	Si	37
Sulfur	S	19
Iron	Fe	32

Table 2.1

All atoms consist of a central, positively charged nucleus surrounded by negatively charged electrons. The bulk of the matter in each atom is found in the nucleus, which consists of positive protons and electrically neutral neutrons all bound tightly together in a very small space. Each element is defined by the number of protons in its atoms. Thus, any atom with 6 protons in its nucleus is called *carbon*, any with 50 protons is called *tin*, and any with 70 protons is called *ytterbium*. (For a list of the elements, see **Appendix F**.)

1. This list of elements is arranged in order of the atomic number, which is the number of protons in each nucleus.

The distance from an atomic nucleus to its electrons is typically 100,000 times the size of the nucleus itself. This is why we say that even solid matter is mostly space. The typical atom is far emptier than the solar system out to Neptune. (The distance from Earth to the Sun, for example, is only 100 times the size of the Sun.) This is one reason atoms are not like miniature solar systems.

Remarkably, physicists have discovered that everything that happens in the universe, from the smallest atomic nucleus to the largest superclusters of galaxies, can be explained through the action of only four forces: gravity, electromagnetism (which combines the actions of electricity and magnetism), and two forces that act at the nuclear level. The fact that there are four forces (and not a million, or just one) has puzzled physicists and astronomers for many years and has led to a quest for a unified picture of nature.

To construct an atom, particle by particle, check out this **guided animation (https://openstax.org/l/30buildanatom)** for building an atom.

2.5 | A Conclusion and a Beginning

Learning Objectives

By the end of this section you will be able to:

- Using the analogy of compressing the age of the Universe into one calendar year, state the major events in its history and roughly when each one occurred.

If you are new to astronomy, you have probably reached the end of our brief tour in this chapter with mixed emotions. On the one hand, you may be fascinated by some of the new ideas you've read about and you may be eager to learn more. On the other hand, you may be feeling a bit overwhelmed by the number of topics we have covered, and the number of new words and ideas we have introduced. Learning astronomy is a little like learning a new language: at first it seems there are so many new expressions that you'll never master them all, but with practice, you soon develop facility with them.

At this point you may also feel a bit small and insignificant, dwarfed by the cosmic scales of distance and time. But, there is another way to look at what you have learned from our first glimpses of the cosmos. Let us consider the history of the universe from the Big Bang to today and compress it, for easy reference, into a single year. (We have borrowed this idea from Carl Sagan's 1997 Pulitzer Prize-winning book, *The Dragons of Eden*.)

On this scale, the Big Bang happened at the first moment of January 1, and this moment, when you are reading this chapter would be the end of the very last second of December 31. When did other events in the development of the universe happen in this "cosmic year?" Our solar system formed around September 10, and the oldest rocks we can date on Earth go back to the third week in September (**Figure 2.12**).

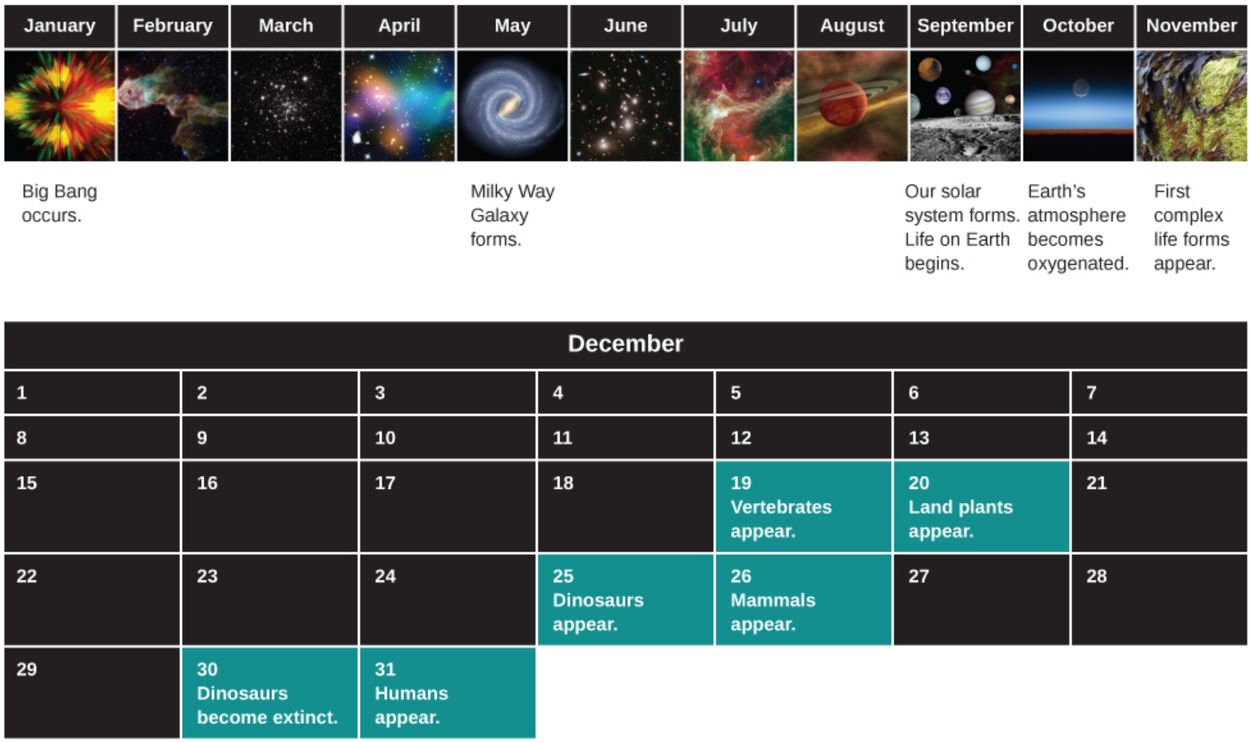


Figure 2.12 On a cosmic calendar, where the time since the Big Bang is compressed into 1 year, creatures we would call human do not emerge on the scene until the evening of December 31. (credit: February: modification of work by NASA, JPL-Caltech, W. Reach (SSC/Caltech); March: modification of work by ESA, Hubble and NASA, Acknowledgement: Giles Chapdelaine; April: modification of work by NASA, ESA, CFHT, CXO, M.J. Jee (University of California, Davis), A. Mahdavi (San Francisco State University); May: modification of work by NASA, JPL-Caltech; June: modification of work by NASA/ESA; July: modification of work by NASA, JPL-Caltech, Harvard-Smithsonian; August: modification of work by NASA, JPL-Caltech, R. Hurt (SSC-Caltech); September: modification of work by NASA; October: modification of work by NASA; November: modification of work by Dénes Emőke)

Where does the origin of human beings fall during the course of this cosmic year? The answer turns out to be the evening of December 31. The invention of the alphabet doesn't occur until the fiftieth second of 11:59 p.m. on December 31. And the beginnings of modern astronomy are a mere fraction of a second before the New Year. Seen in a cosmic context, the amount of time we have had to study the stars is minute, and our success in piecing together as much of the story as we have is remarkable.

Certainly our attempts to understand the universe are not complete. As new technologies and new ideas allow us to gather more and better data about the cosmos, our present picture of astronomy will very likely undergo many changes. Still, as you read our current progress report on the exploration of the universe, take a few minutes every once in a while just to savor how much you have already learned.

For Further Exploration

Websites

If you enjoyed the beautiful images in this chapter (and there are many more fabulous photos to come in other chapters), you may want to know where you can obtain and download such pictures for your own enjoyment. (Many astronomy images are from government-supported instruments or projects, paid for by tax dollars, and therefore are free of copyright laws.) Here are three resources we especially like:

Astronomy Picture of the Day (http://apod.nasa.gov/apod/astropix.html/)

Two space scientists scour the Internet and select one beautiful astronomy image to feature each day. Their archives range widely, from images of planets and nebulae to rockets and space instruments; they also have many photos of the night sky. The search function (see the menu on the bottom of the page) works quite well for finding something specific among the many years' worth of daily images.

Hubble Space Telescope Images (http://www.hubblesite.org/newscenter/archive/browse/images/)

Starting at this page, you can select from among hundreds of Hubble pictures by subject or by date. Note that many of the images have supporting pictures with them, such as diagrams, animations, or comparisons. Excellent captions and background information are provided. Other ways to approach these images are through the more public-oriented Hubble Gallery (www.hubblesite.org/gallery) and the European homepage (www.spacetelescope.org/images).

National Aeronautics and Space Administration's (NASA's) **Planetary Photojournal (http://photojournal.jpl.nasa.gov/)**

This site features thousands of images from planetary exploration, with captions of varied length. You can select images by world, feature name, date, or catalog number, and download images in a number of popular formats. However, only NASA mission images are included. Note the Photojournal Search option on the menu at the top of the homepage to access ways to search their archives.

Videos

Cosmic Voyage. This video presents a portion of Cosmic Voyage, narrated by Morgan Freeman (8:34).

Powers of Ten. This classic short video is a much earlier version of Powers of Ten, narrated by Philip Morrison (9:00).

The Known Universe. This video tour from the American Museum of Natural History has realistic animation, music, and captions (6:30).

Wanderers. This video provides a tour of the solar system, with narrative by Carl Sagan, imagining other worlds with dramatically realistic paintings (3:50).

3 | MOTION ALONG A STRAIGHT LINE

Figure 3.1 A JR Central L0 series five-car maglev (magnetic levitation) train undergoing a test run on the Yamanashi Test Track. The maglev train's motion can be described using kinematics, the subject of this chapter. (credit: modification of work by "Maryland GovPics"/Flickr)

Chapter Outline

Introduction

Our universe is full of objects in motion. From the stars, planets, and galaxies; to the motion of people and animals; down to the microscopic scale of atoms and molecules—everything in our universe is in motion. We can describe motion using the two disciplines of kinematics and dynamics. We will eventually study dynamics, which is concerned with the causes of motion, in **Newton's Synthesis**; but, there is much to be learned about motion without referring to what causes it, and this is the study of kinematics. Kinematics involves describing motion through properties such as position, time, velocity, and acceleration.

A full treatment of **kinematics** considers motion in two and three dimensions. For now, we discuss motion in one dimension, which provides us with the tools necessary to study multidimensional motion. A good example of an object undergoing one-dimensional motion is the maglev (magnetic levitation) train depicted at the beginning of this chapter. As it travels, say, from Tokyo to Kyoto, it is at different positions along the track at various times in its journey, and therefore has displacements, or changes in position. It also has a variety of velocities along its path and it undergoes accelerations (changes in velocity). With the skills learned in this chapter we can calculate these quantities and average velocity. All these quantities can be described using kinematics, without knowing the train's mass or the forces involved.

3.1 | Position, Displacement and Average Velocity

Learning Objectives

By the end of this section, you will be able to:

- Define position, displacement, and distance traveled.
- Calculate the total displacement given the position as a function of time.
- Determine the total distance traveled.
- Calculate the average velocity given the displacement and elapsed time.

When you're in motion, the basic questions to ask are: Where are you? Where are you going? How fast are you getting there? The answers to these questions require that you specify your position, your displacement, and your average velocity—the terms we define in this section.

Position

To describe the motion of an object, you must first be able to describe its **position** (x): *where it is at any particular time*. More precisely, we need to specify its position relative to a convenient frame of reference. A frame of reference is an arbitrary set of axes from which the position and motion of an object are described. Earth is often used as a frame of reference, and we often describe the position of an object as it relates to stationary objects on Earth. For example, a rocket launch could be described in terms of the position of the rocket with respect to Earth as a whole, whereas a cyclist's position could be described in terms of where she is in relation to the buildings she passes Figure 3.2. In other cases, we use reference frames that are not stationary but are in motion relative to Earth. To describe the position of a person in an airplane, for example, we use the airplane, not Earth, as the reference frame. To describe the position of an object undergoing one-dimensional motion, we often use the variable x. Later in the chapter, during the discussion of free fall, we use the variable y.

Figure 3.2 These cyclists in Vietnam can be described by their position relative to buildings or a canal. Their motion can be described by their change in position, or displacement, in a frame of reference. (credit: Suzan Black)

Displacement

If an object moves relative to a frame of reference—for example, if a professor moves to the right relative to a whiteboard Figure 3.3—then the object's position changes. This change in position is called **displacement**. The word *displacement* implies that an object has moved, or has been displaced. Although position is the numerical value of x along a straight line where an object might be located, displacement gives the *change* in position along this line. Since displacement indicates direction, it is a vector and can be either positive or negative, depending on the choice of positive direction. Also, an analysis of motion can have many displacements embedded in it. If right is positive and an object moves 2 m to the right, then 4 m to the left, the individual displacements are 2 m and -4 m, respectively.

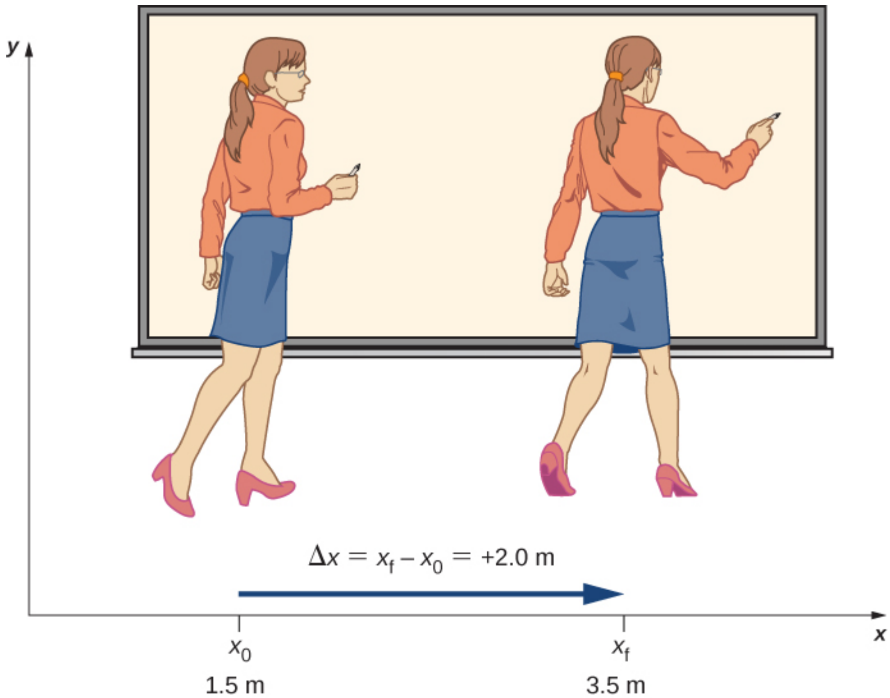


Figure 3.3 A professor paces left and right while lecturing. Her position relative to Earth is given by *x*. The +2.0-m displacement of the professor relative to Earth is represented by an arrow pointing to the right.

Displacement

Displacement Δx is the change in position of an object:

$$\Delta x = x_\text{f} - x_0, \tag{3.1}$$

where Δx is displacement, x_f is the final position, and x_0 is the initial position.

We use the uppercase Greek letter delta (Δ) to mean "change in" whatever quantity follows it; thus, Δx means *change in position* (final position less initial position). We always solve for displacement by subtracting initial position x_0 from final position x_f. Note that the SI unit for displacement is the meter, but sometimes we use kilometers or other units of length. Keep in mind that when units other than meters are used in a problem, you may need to convert them to meters to complete the calculation (see **Appendix B**).

The **displacement** is the first of a number of important physical quantities we will discuss that is a called a **vector**. While we will discuss the mathematics of these quantities in detail in **An Introduction to Vectors**, the meaning of **vector** in the context of this chapter is quite simple. A vector is a quantity that possesses both a magnitude and a direction. In the present discussion, a displacement is either in the positive *x* direction or in the negative *x* direction. The sign of the displacement, Δx, (positive or negative) represents the direction of the displacement vector.

Objects in motion can also have a series of displacements. In the previous example of the pacing professor, the individual displacements are 2 m and -4 m, giving a total displacement of −2 m. We define **total displacement** Δx_Total, as *the sum of the individual displacements*, and express this mathematically with the equation

$$\Delta x_\text{Total} = \sum \Delta x_\text{i}, \tag{3.2}$$

where Δx_i are the individual displacements. In the earlier example,

$$\Delta x_1 = x_1 - x_0 = 2 - 0 = 2 \text{ m}.$$

Similarly,

$$\Delta x_2 = x_2 - x_1 = -2 - (2) = -4 \text{ m}.$$

Thus,

$$\Delta x_{\text{Total}} = \Delta x_1 + \Delta x_2 = 2 - 4 = -2 \text{ m}.$$

The total displacement is 2 − 4 = −2 m to the left, or in the negative direction. It is also useful to calculate the magnitude of the displacement, or its size. The magnitude of the displacement is always positive. This is the absolute value of the displacement, because displacement is a vector and cannot have a negative value of magnitude. In our example, the magnitude of the total displacement is 2 m, whereas the magnitudes of the individual displacements are 2 m and 4 m.

The magnitude of the total displacement should not be confused with the distance traveled. Distance traveled x_{Total} , is the total length of the path traveled between two positions. In the previous problem, the **distance traveled** is the sum of the magnitudes of the individual displacements:

$$x_{\text{Total}} = |\Delta x_1| + |\Delta x_2| = 2 + 4 = 6 \text{ m}.$$

Average Velocity

To calculate the other physical quantities in kinematics we must introduce the time variable. The time variable allows us not only to state where the object is (its position) during its motion, but also how fast it is moving. How fast an object is moving is given by the rate at which the position changes with time.

For each position x_i , we assign a particular time t_i . If the details of the motion at each instant are not important, the rate is usually expressed as the **average velocity** $\bar{v}$. This vector quantity is simply the total displacement between two points divided by the time taken to travel between them. The time taken to travel between two points is called the **elapsed time** Δt .

Average Velocity

If x_1 and x_2 are the positions of an object at times t_1 and t_2 , respectively, then

$$\text{Average velocity} = \bar{v} = \frac{\text{Displacement between two points}}{\text{Elapsed time between two points}}$$
$$\bar{v} = \frac{\Delta x}{\Delta t} = \frac{x_2 - x_1}{t_2 - t_1}. \quad (3.3)$$

It is important to note that the average velocity is a vector and can be negative, depending on positions x_1 and x_2 . The sign of v will be the same as the sign of Δx .

Example 3.1

Delivering Flyers

Jill sets out from her home to deliver flyers for her yard sale, traveling due east along her street lined with houses. At 0.5 km and 9 minutes later she runs out of flyers and has to retrace her steps back to her house to get more. This takes an additional 9 minutes. After picking up more flyers, she sets out again on the same path, continuing where she left off, and ends up 1.0 km from her house. This third leg of her trip takes 15 minutes. At this point she turns back toward her house, heading west. After 1.75 km and 25 minutes she stops to rest.

a. What is Jill's total displacement to the point where she stops to rest?
b. What is the magnitude of the final displacement?
c. What is the average velocity during her entire trip?

d. What is the total distance traveled?

e. Make a graph of position versus time.

A sketch of Jill's movements is shown in **Figure 3.4**.

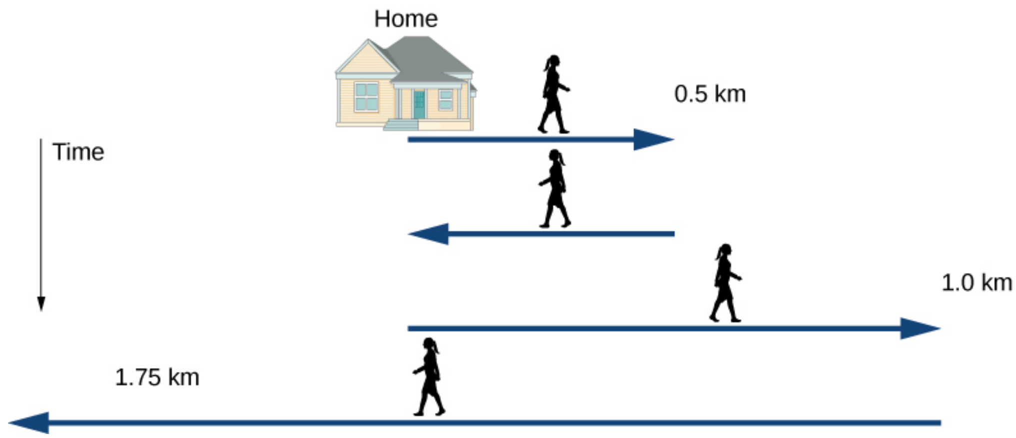


Figure 3.4 Timeline of Jill's movements.

Strategy

The problem contains data on the various legs of Jill's trip, so it would be useful to make a table of the physical quantities. We are given position and time in the wording of the problem so we can calculate the displacements and the elapsed time. We take east to be the positive direction. From this information we can find the total displacement and average velocity. Jill's home is the starting point x_0. The following table gives Jill's time and position in the first two columns, and the displacements are calculated in the third column.

Time t_i (min)	Position x_i (km)	Displacement Δx_i (km)
$t_0 = 0$	$x_0 = 0$	$\Delta x_0 = 0$
$t_1 = 9$	$x_1 = 0.5$	$\Delta x_1 = x_1 - x_0 = 0.5$
$t_2 = 18$	$x_2 = 0$	$\Delta x_2 = x_2 - x_1 = -0.5$
$t_3 = 33$	$x_3 = 1.0$	$\Delta x_3 = x_3 - x_2 = 1.0$
$t_4 = 58$	$x_4 = -0.75$	$\Delta x_4 = x_4 - x_3 = -1.75$

Solution

a. From the above table, the total displacement is

$$\sum \Delta x_i = 0.5 - 0.5 + 1.0 - 1.75 \text{ km} = -0.75 \text{ km}.$$

b. The magnitude of the total displacement is $|-0.75| \text{ km} = 0.75 \text{ km}$.

c. $$\text{Average velocity} = \frac{\text{Total displacement}}{\text{Elapsed time}} = \bar{v} = \frac{-0.75 \text{ km}}{58 \text{ min}} = -0.013 \text{ km/min}$$

d. The total distance traveled (sum of magnitudes of individual displacements) is $x_{\text{Total}} = \sum |\Delta x_i| = 0.5 + 0.5 + 1.0 + 1.75 \text{ km} = 3.75 \text{ km}$.

e. We can graph Jill's position versus time as a useful aid to see the motion; the graph is shown in **Figure 3.5**.

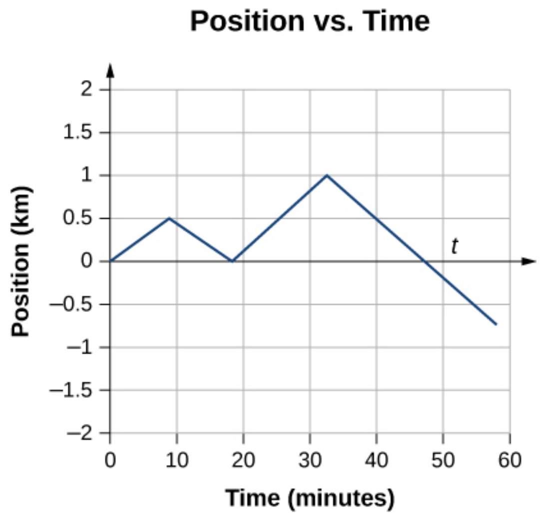
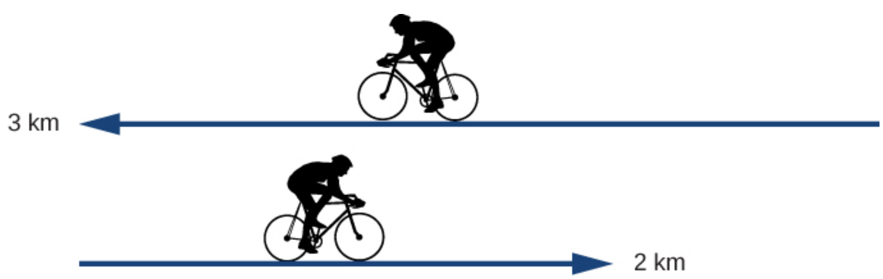


Figure 3.5 This graph depicts Jill's position versus time. The average velocity is the slope of a line connecting the initial and final points.

Significance

Jill's total displacement is −0.75 km, which means at the end of her trip she ends up 0.75 km due west of her home. The average velocity means if someone was to walk due west at 0.013 km/min starting at the same time Jill left her home, they both would arrive at the final stopping point at the same time. Note that if Jill were to end her trip at her house, her total displacement would be zero, as well as her average velocity. The total distance traveled during the 58 minutes of elapsed time for her trip is 3.75 km.

3.1 Check Your Understanding A cyclist rides 3 km west and then turns around and rides 2 km east. (a) What is his displacement? (b) What is the distance traveled? (c) What is the magnitude of his displacement?



3.2 | Instantaneous Velocity and Speed

Learning Objectives

By the end of this section, you will be able to:

- Explain the difference between average velocity and instantaneous velocity.
- Describe the difference between velocity and speed.
- Calculate the instantaneous velocity given the mathematical equation for the velocity.
- Calculate the speed given the instantaneous velocity.

We have now seen how to calculate the average velocity between two positions. However, since objects in the real world move continuously through space and time, we would like to find the velocity of an object at any single point. We can find the velocity of the object anywhere along its path by using some fundamental principles of calculus. This section gives us better insight into the physics of motion and will be useful in later chapters.

Instantaneous Velocity

The quantity that tells us how fast an object is moving anywhere along its path is the **instantaneous velocity**, usually called simply *velocity*. It is the average velocity between two points on the path in the limit that the time (and therefore the displacement) between the two points approaches zero. To illustrate this idea mathematically, we need to express position x as a continuous function of t denoted by $x(t)$. The expression for the average velocity between two points using this notation is $\bar{v} = \frac{x(t_2) - x(t_1)}{t_2 - t_1}$. To find the instantaneous velocity at any position, we let $t_1 = t$ and $t_2 = t + \Delta t$. After inserting these expressions into the equation for the average velocity and taking the limit as $\Delta t \to 0$, we find the expression for the instantaneous velocity:

$$v(t) = \lim_{\Delta t \to 0} \frac{x(t + \Delta t) - x(t)}{\Delta t} = \frac{dx(t)}{dt}.$$

The Calculus of Instantaneous Velocity

The instantaneous velocity of an object is the limit of the average velocity as the elapsed time approaches zero, or the derivative of x with respect to t:

$$v(t) = \frac{d}{dt}x(t). \qquad (3.4)$$

Because many students reading this book may not yet have learned how to take derivatives of functions in their calculus course, we will postpone much use of calculus until later chapters. Nevertheless, it is important to understand the concept of a derivative, as it has been introduced here.

Like average velocity, instantaneous velocity is a vector with dimension of length per time. The instantaneous velocity at a specific time point t_0 is the rate of change of the position function, which is the slope of the position function $x(t)$ at t_0. **Figure 3.6** shows how the average velocity $\bar{v} = \frac{\Delta x}{\Delta t}$ between two times approaches the instantaneous velocity at t_0. The instantaneous velocity is shown at time t_0, which happens to be at the maximum of the position function. The slope of the position graph is zero at this point, and thus the instantaneous velocity is zero. At other times, t_1, t_2, and so on, the instantaneous velocity is not zero because the slope of the position graph would be positive or negative. If the position function had a minimum, the slope of the position graph would also be zero, giving an instantaneous velocity of zero there as well. Thus, the zeros of the velocity function give the minimum and maximum of the position function.

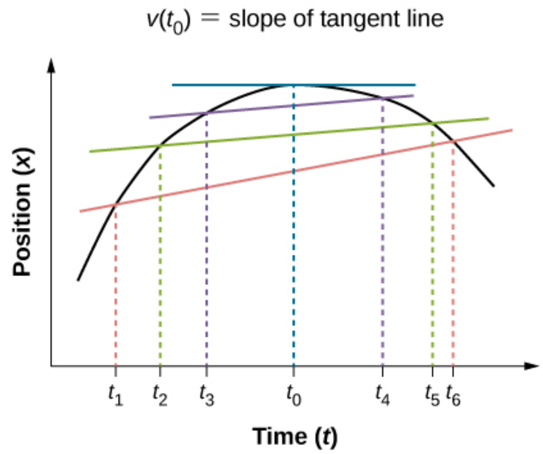


Figure 3.6 In a graph of position versus time, the instantaneous velocity is the slope of the tangent line at a given point. The average velocities $\bar{v} = \frac{\Delta x}{\Delta t} = \frac{x_f - x_i}{t_f - t_i}$ between times $\Delta t = t_6 - t_1$, $\Delta t = t_5 - t_2$, and $\Delta t = t_4 - t_3$ are shown. When $\Delta t \to 0$, the average velocity approaches the instantaneous velocity at $t = t_0$.

Example 3.2

Finding Velocity from a Position-Versus-Time Graph

Given the position-versus-time graph of **Figure 3.7**, find the velocity-versus-time graph.

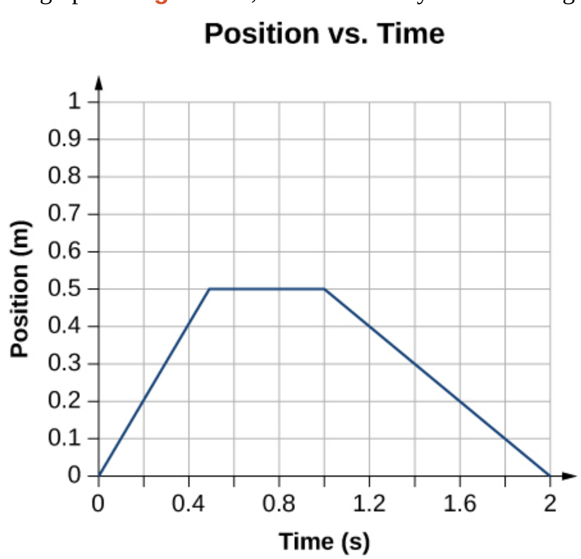


Figure 3.7 The object starts out in the positive direction, stops for a short time, and then reverses direction, heading back toward the origin. Notice that the object comes to rest instantaneously, which would require an infinite force. Thus, the graph is an approximation of motion in the real world. (The concept of force is discussed in **Newton's Synthesis**.)

Strategy

The graph contains three straight lines during three time intervals. We find the velocity during each time interval by taking the slope of the line using the grid.

Solution

Time interval 0 s to 0.5 s: $\bar{v} = \frac{\Delta x}{\Delta t} = \frac{0.5\text{ m} - 0.0\text{ m}}{0.5\text{ s} - 0.0\text{ s}} = 1.0\text{ m/s}$

Time interval 0.5 s to 1.0 s: $\bar{v} = \frac{\Delta x}{\Delta t} = \frac{0.0\text{ m} - 0.0\text{ m}}{1.0\text{ s} - 0.5\text{ s}} = 0.0\text{ m/s}$

Time interval 1.0 s to 2.0 s: $\bar{v} = \frac{\Delta x}{\Delta t} = \frac{0.0\text{ m} - 0.5\text{ m}}{2.0\text{ s} - 1.0\text{ s}} = -0.5\text{ m/s}$

The graph of these values of velocity versus time is shown in **Figure 3.8**.

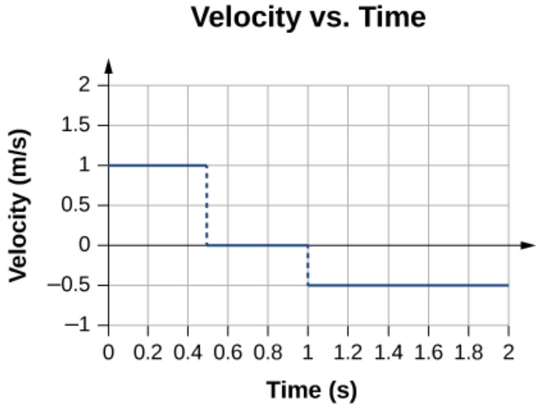


Figure 3.8 The velocity is positive for the first part of the trip, zero when the object is stopped, and negative when the object reverses direction.

Significance

During the time interval between 0 s and 0.5 s, the object's position is moving away from the origin and the position-versus-time curve has a positive slope. At any point along the curve during this time interval, we can find the instantaneous velocity by taking its slope, which is +1 m/s, as shown in **Figure 3.8**. In the subsequent time interval, between 0.5 s and 1.0 s, the position doesn't change and we see the slope is zero. From 1.0 s to 2.0 s, the object is moving back toward the origin and the slope is −0.5 m/s. The object has reversed direction and has a negative velocity.

Speed

In everyday language, most people use the terms *speed* and *velocity* interchangeably. In physics, however, they do not have the same meaning and are distinct concepts. One major difference is that speed has no direction; that is, speed is a scalar.

We can calculate the **average speed** by finding the total distance traveled divided by the elapsed time:

$$\text{Average speed} = \bar{s} = \frac{\text{Total distance}}{\text{Elapsed time}}. \tag{3.5}$$

Average speed is not necessarily the same as the magnitude of the average velocity, which is found by dividing the magnitude of the total displacement by the elapsed time. For example, if a trip starts and ends at the same location, the total displacement is zero, and therefore the average velocity is zero. The average speed, however, is not zero, because the total distance traveled is greater than zero. If we take a road trip of 300 km and need to be at our destination at a certain time, then we would be interested in our average speed.

However, we can calculate the **instantaneous speed** from the magnitude of the instantaneous velocity:

$$\text{Instantaneous speed} = |v(t)|. \quad (3.6)$$

If a particle is moving along the x-axis at +7.0 m/s and another particle is moving along the same axis at −7.0 m/s, they have different velocities, but both have the same speed of 7.0 m/s. Some typical speeds are shown in the following table.

Speed	m/s	mi/h
Continental drift	10^{-7}	2×10^{-7}
Brisk walk	1.7	3.9
Cyclist	4.4	10
Sprint runner	12.2	27
Rural speed limit	24.6	56
Official land speed record	341.1	763
Speed of sound at sea level	343	768
Space shuttle on reentry	7800	17,500
Escape velocity of Earth*	11,200	25,000
Orbital speed of Earth around the Sun	29,783	66,623
Speed of light in a vacuum	299,792,458	670,616,629

Table 3.1 Speeds of Various Objects *Escape velocity is the velocity at which an object must be launched so that it overcomes Earth's gravity and is not pulled back toward Earth.

3.3 | Average and Instantaneous Acceleration

Learning Objectives

By the end of this section, you will be able to:

- Calculate the average acceleration between two points in time.
- Calculate the instantaneous acceleration given the functional form of velocity.
- Explain the vector nature of instantaneous acceleration and velocity.
- Explain the difference between average acceleration and instantaneous acceleration.
- Find instantaneous acceleration at a specified time on a graph of velocity versus time.

The importance of understanding acceleration spans our day-to-day experience, as well as the vast reaches of outer space and the tiny world of subatomic physics. In everyday conversation, to *accelerate* means to speed up; applying the brake pedal causes a vehicle to slow down. We are familiar with the acceleration of our car, for example. The greater the acceleration, the greater the change in velocity over a given time. Acceleration is widely seen in experimental physics. In linear particle accelerator experiments, for example, subatomic particles are accelerated to very high velocities in collision experiments, which tell us information about the structure of the subatomic world as well as the origin of the universe. In space, cosmic rays are subatomic particles that have been accelerated to very high energies in supernovas (exploding massive stars) and active galactic nuclei. It is important to understand the processes that accelerate cosmic rays because these rays contain highly penetrating radiation that can damage electronics flown on spacecraft, for example.

Average Acceleration

The formal definition of acceleration is consistent with these notions just described, but is more inclusive.

Average Acceleration

Average acceleration is the rate at which velocity changes:

$$\bar{a} = \frac{\Delta v}{\Delta t} = \frac{v_f - v_0}{t_f - t_0}, \quad (3.7)$$

where $\bar{a}$ is **average acceleration**, v is velocity, and t is time. (The bar over the a means *average* acceleration.)

Because acceleration is velocity in meters divided by time in seconds, the SI units for acceleration are often abbreviated m/s^2—that is, meters per second squared or meters per second per second. This literally means by how many meters per second the velocity changes every second. Recall that velocity is a vector—it has both magnitude and direction—which means that a change in velocity can be a change in magnitude (or speed), but it can also be a change in direction. For example, if a runner traveling at 10 km/h due east slows to a stop, reverses direction, continues her run at 10 km/h due west, her velocity has changed as a result of the change in direction, although the *magnitude* of the velocity is the same in both directions. Thus, acceleration occurs when velocity changes in magnitude (an increase or decrease in speed) or in direction, or both.

Acceleration as a Vector

Acceleration is a vector in the same direction as the *change* in velocity, Δv. Since velocity is a vector, it can change in magnitude or in direction, or both. Acceleration is, therefore, a change in speed or direction, or both.

Keep in mind that although acceleration is in the direction of the change in velocity, it is not always in the direction of motion. When an object slows down, its acceleration is opposite to the direction of its motion. Although this is commonly referred to as *deceleration* **Figure 3.9**, we say the train is accelerating in a direction opposite to its direction of motion.

Figure 3.9 A subway train in Sao Paulo, Brazil, decelerates as it comes into a station. It is accelerating in a direction opposite to its direction of motion. (credit: Yusuke Kawasaki)

The term *deceleration* can cause confusion in our analysis because it is not a vector and it does not point to a specific direction with respect to a coordinate system, so we do not use it. Acceleration is a vector, so we must choose the appropriate sign for it in our chosen coordinate system. In the case of the train in **Figure 3.9**, acceleration is *in the negative direction in the chosen coordinate system*, so we say the train is undergoing negative acceleration.

If an object in motion has a velocity in the positive direction with respect to a chosen origin and it acquires a constant negative acceleration, the object eventually comes to a rest and reverses direction. If we wait long enough, the object passes through the origin going in the opposite direction. This is illustrated in **Figure 3.10**.

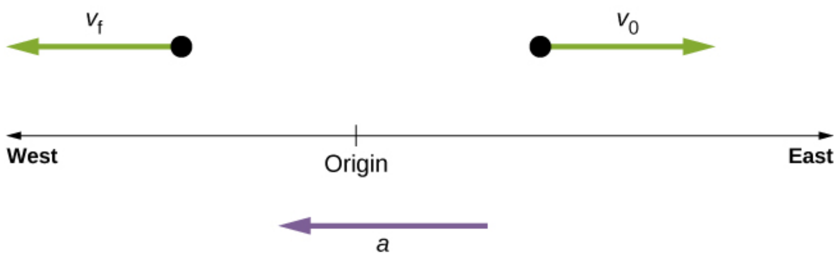


Figure 3.10 An object in motion with a velocity vector toward the east under negative acceleration comes to a rest and reverses direction. It passes the origin going in the opposite direction after a long enough time.

Example 3.3

Calculating Average Acceleration: A Racehorse Leaves the Gate

A racehorse coming out of the gate accelerates from rest to a velocity of 15.0 m/s due west in 1.80 s. What is its average acceleration?

Figure 3.11 Racehorses accelerating out of the gate. (credit: Jon Sullivan)

Strategy

First we draw a sketch and assign a coordinate system to the problem **Figure 3.12**. This is a simple problem, but it always helps to visualize it. Notice that we assign east as positive and west as negative. Thus, in this case, we have negative velocity.

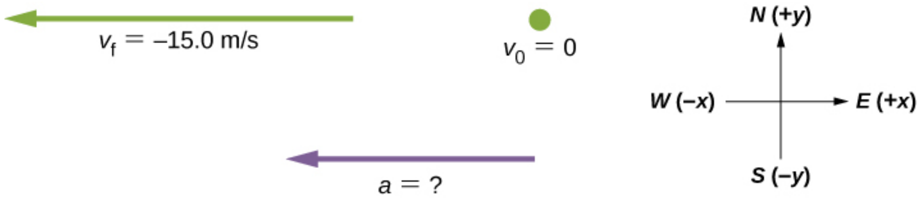


Figure 3.12 Identify the coordinate system, the given information, and what you want to determine.

We can solve this problem by identifying Δv and Δt from the given information, and then calculating the average acceleration directly from the equation $\bar{a} = \frac{\Delta v}{\Delta t} = \frac{v_f - v_0}{t_f - t_0}$.

Solution

First, identify the knowns: $v_0 = 0$, $v_f = -15.0 \text{ m/s}$ (the negative sign indicates direction toward the west), Δt = 1.80 s.

Second, find the change in velocity. Since the horse is going from zero to –15.0 m/s, its change in velocity equals its final velocity:

$$\Delta v = v_f - v_0 = v_f = -15.0 \text{ m/s}.$$

Last, substitute the known values (Δv and Δt) and solve for the unknown $\bar{a}$:

$$\bar{a} = \frac{\Delta v}{\Delta t} = \frac{-15.0 \text{ m/s}}{1.80 \text{ s}} = -8.33 \text{m/s}^2.$$

Significance

The negative sign for acceleration indicates that acceleration is toward the west. An acceleration of 8.33 m/s^2 due west means the horse increases its velocity by 8.33 m/s due west each second; that is, 8.33 meters per second per second, which we write as 8.33 m/s^2. This is truly an average acceleration, because the ride is not smooth. We see later that an acceleration of this magnitude would require the rider to hang on with a force nearly equal to his weight.

3.2 **Check Your Understanding** Protons in a linear accelerator are accelerated from rest to 2.0×10^7 m/s in 10^{-4} s. What is the average acceleration of the protons?

Instantaneous Acceleration

Instantaneous acceleration *a*, or *acceleration at a specific instant in time*, is obtained using the same process discussed for instantaneous velocity. That is, we calculate the average velocity between two points in time separated by Δt and let Δt approach zero. The result is the derivative of the velocity function *v*(*t*), which is **instantaneous acceleration** and is expressed mathematically as

$$a(t) = \frac{d}{dt}v(t). \tag{3.8}$$

Thus, similar to velocity being the derivative of the position function, instantaneous acceleration is the derivative of the velocity function. We can show this graphically in the same way as instantaneous velocity. In **Figure 3.13**, instantaneous acceleration at time t_0 is the slope of the tangent line to the velocity-versus-time graph at time t_0. We see that average acceleration $\bar{a} = \frac{\Delta v}{\Delta t}$ approaches instantaneous acceleration as Δt approaches zero. Also in part (a) of the figure, we see that velocity has a maximum when its slope is zero. This time corresponds to the zero of the acceleration function. In part (b), instantaneous acceleration at the minimum velocity is shown, which is also zero, since the slope of the curve is zero there, too. Thus, for a given velocity function, the zeros of the acceleration function give either the minimum or the maximum velocity.

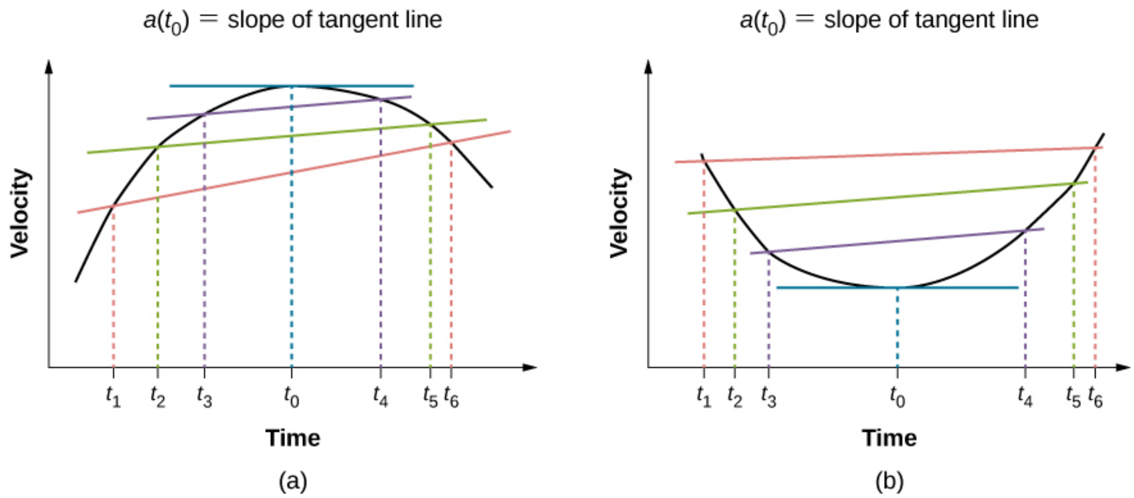


Figure 3.13 In a graph of velocity versus time, instantaneous acceleration is the slope of the tangent line. (a) Shown is average acceleration $\overline{a} = \frac{\Delta v}{\Delta t} = \frac{v_f - v_i}{t_f - t_i}$ between times $\Delta t = t_6 - t_1$, $\Delta t = t_5 - t_2$, and $\Delta t = t_4 - t_3$. When $\Delta t \to 0$, the average acceleration approaches instantaneous acceleration at time *t*0. In view (a), instantaneous acceleration is shown for the point on the velocity curve at maximum velocity. At this point, instantaneous acceleration is the slope of the tangent line, which is zero. At any other time, the slope of the tangent line—and thus instantaneous acceleration—would not be zero. (b) Same as (a) but shown for instantaneous acceleration at minimum velocity.

To illustrate this concept, let's look at two examples. First, a simple example is shown of a velocity-versus-time graph, to find acceleration graphically. This graph is depicted in Figure 3.14(a), which is a straight line. The corresponding graph of acceleration versus time is found from the slope of velocity and is shown in Figure 3.14(b). In this example, the velocity function is a straight line with a constant slope, thus acceleration is a constant.

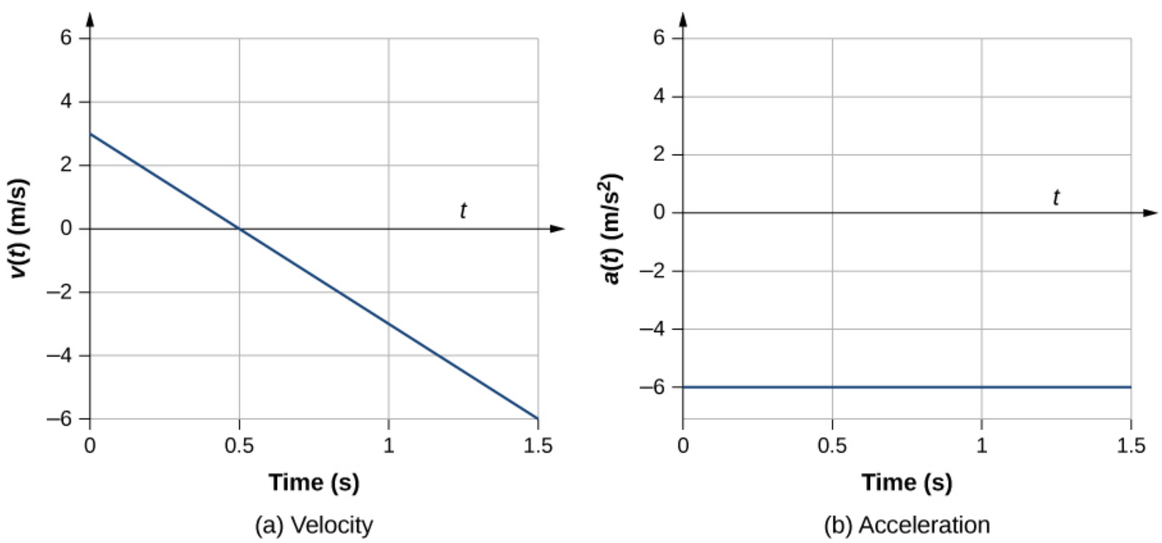


Figure 3.14 (a, b) The velocity-versus-time graph is linear and has a negative constant slope (a) that is equal to acceleration, shown in (b).

3.3 **Check Your Understanding** An airplane lands on a runway traveling east. Describe its acceleration.

Getting a Feel for Acceleration

You are probably used to experiencing acceleration when you step into an elevator, or step on the gas pedal in your car.

However, acceleration is happening to many other objects in our universe with which we don't have direct contact. **Table 3.2** presents the acceleration of various objects. We can see the magnitudes of the accelerations extend over many orders of magnitude.

Acceleration	Value (m/s^2)
High-speed train	0.25
Elevator	2
Cheetah	5
Object in a free fall without air resistance near the surface of Earth	9.8
Space shuttle maximum during launch	29
Parachutist peak during normal opening of parachute	59
F16 aircraft pulling out of a dive	79
Explosive seat ejection from aircraft	147
Sprint missile	982
Fastest rocket sled peak acceleration	1540
Jumping flea	3200
Baseball struck by a bat	30,000
Closing jaws of a trap-jaw ant	1,000,000
Proton in the large Hadron collider	1.9×10^9

Table 3.2 Typical Values of Acceleration (credit: Wikipedia: Orders of Magnitude (acceleration))

In this table, we see that typical accelerations vary widely with different objects and have nothing to do with object size or how massive it is. Acceleration can also vary widely with time during the motion of an object. A drag racer has a large acceleration just after its start, but then it tapers off as the vehicle reaches a constant velocity. Its average acceleration can be quite different from its instantaneous acceleration at a particular time during its motion. **Figure 3.15** compares graphically average acceleration with instantaneous acceleration for two very different motions.

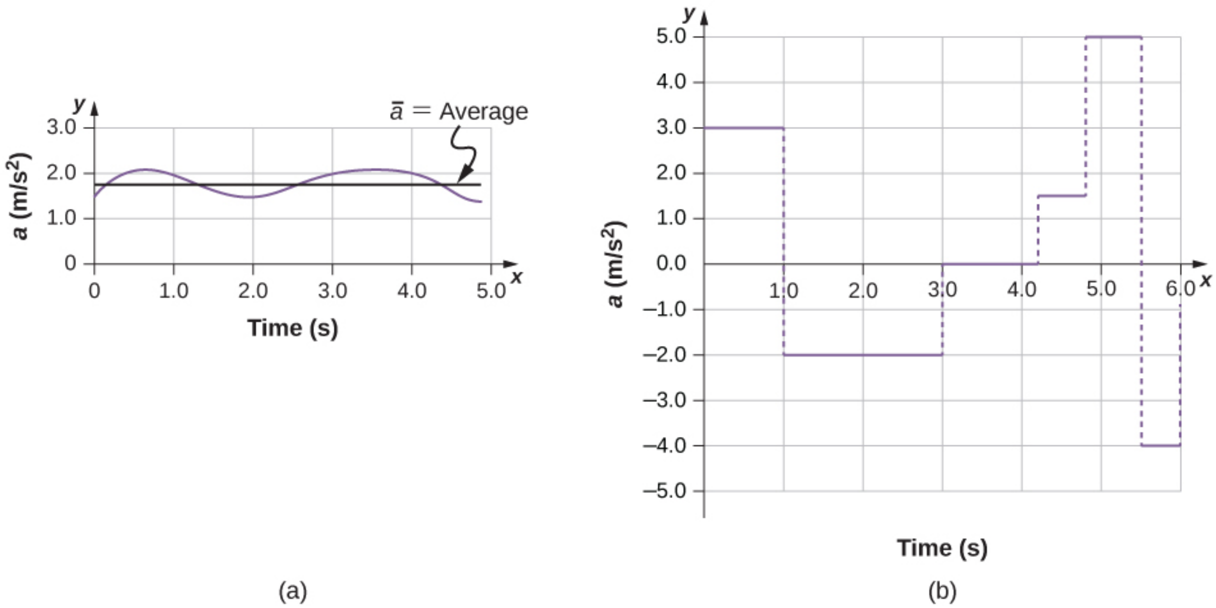


Figure 3.15 Graphs of instantaneous acceleration versus time for two different one-dimensional motions. (a) Acceleration varies only slightly and is always in the same direction, since it is positive. The average over the interval is nearly the same as the acceleration at any given time. (b) Acceleration varies greatly, perhaps representing a package on a post office conveyor belt that is accelerated forward and backward as it bumps along. It is necessary to consider small time intervals (such as from 0–1.0 s) with constant or nearly constant acceleration in such a situation.

Learn about position, velocity, and acceleration graphs. Move the little man back and forth with a mouse and plot his motion. Set the position, velocity, or acceleration and let the simulation move the man for you. Visit **this link (https://openstaxcollege.org/l/21movmansimul)** to use the moving man simulation.

3.4 | Motion with Constant Acceleration

Learning Objectives

By the end of this section, you will be able to:

- Identify which equations of motion are to be used to solve for unknowns.
- Use appropriate equations of motion to solve a two-body pursuit problem.

You might guess that the greater the acceleration of, say, a car moving away from a stop sign, the greater the car's displacement in a given time. But, we have not developed a specific equation that relates acceleration and displacement. In this section, we look at some convenient equations for kinematic relationships, starting from the definitions of displacement, velocity, and acceleration. We first investigate a single object in motion, called single-body motion. Then we investigate the motion of two objects, called **two-body pursuit problems**.

Notation

First, let us make some simplifications in notation. Taking the initial time to be zero, as if time is measured with a stopwatch, is a great simplification. Since elapsed time is $\Delta t = t_{\text{f}} - t_0$, taking $t_0 = 0$ means that $\Delta t = t_{\text{f}}$, the final time on the stopwatch. When initial time is taken to be zero, we use the subscript 0 to denote initial values of position and velocity. That is, x_0 *is the initial position* and v_0 *is the initial velocity*. We put no subscripts on the final values. That is, *t is the final time*, *x is the final position*, and *v is the final velocity*. This gives a simpler expression for elapsed time, $\Delta t = t$. It also simplifies the expression for *x* displacement, which is now $\Delta x = x - x_0$. Also, it simplifies the expression for change in velocity, which is now $\Delta v = v - v_0$. To summarize, using the simplified notation, with the initial time taken to be zero,

$$\Delta t = t$$
$$\Delta x = x - x_0$$
$$\Delta v = v - v_0,$$

where the subscript 0 denotes an initial value and the absence of a subscript denotes a final value in whatever motion is under consideration.

We now make the important assumption that *acceleration is constant*. This assumption allows us to avoid using calculus to find instantaneous acceleration. Since acceleration is constant, the average and instantaneous accelerations are equal—that is,

$$\bar{a} = a = \text{constant}.$$

Thus, we can use the symbol a for acceleration at all times. Assuming acceleration to be constant does not seriously limit the situations we can study nor does it degrade the accuracy of our treatment. For one thing, acceleration *is* constant in a great number of situations. Furthermore, in many other situations we can describe motion accurately by assuming a constant acceleration equal to the average acceleration for that motion. Lastly, for motion during which acceleration changes drastically, such as a car accelerating to top speed and then braking to a stop, motion can be considered in separate parts, each of which has its own constant acceleration.

Displacement and Position from Velocity

To get our first two equations, we start with the definition of average velocity:

$$\bar{v} = \frac{\Delta x}{\Delta t}.$$

Substituting the simplified notation for Δx and Δt yields

$$\bar{v} = \frac{x - x_0}{t}.$$

Solving for x gives us

$$x = x_0 + \bar{v}t, \quad (3.9)$$

where the average velocity is

$$\bar{v} = \frac{v_0 + v}{2}. \quad (3.10)$$

The equation $\bar{v} = \frac{v_0 + v}{2}$ reflects the fact that when acceleration is constant, $\bar{v}$ is just the simple average of the initial and final velocities. **Figure 3.16** illustrates this concept graphically. In part (a) of the figure, acceleration is constant, with velocity increasing at a constant rate. The average velocity during the 1-h interval from 40 km/h to 80 km/h is 60 km/h:

$$\bar{v} = \frac{v_0 + v}{2} = \frac{40\text{ km/h} + 80\text{ km/h}}{2} = 60\text{ km/h}.$$

In part (b), acceleration is not constant. During the 1-h interval, velocity is closer to 80 km/h than 40 km/h. Thus, the average velocity is greater than in part (a).

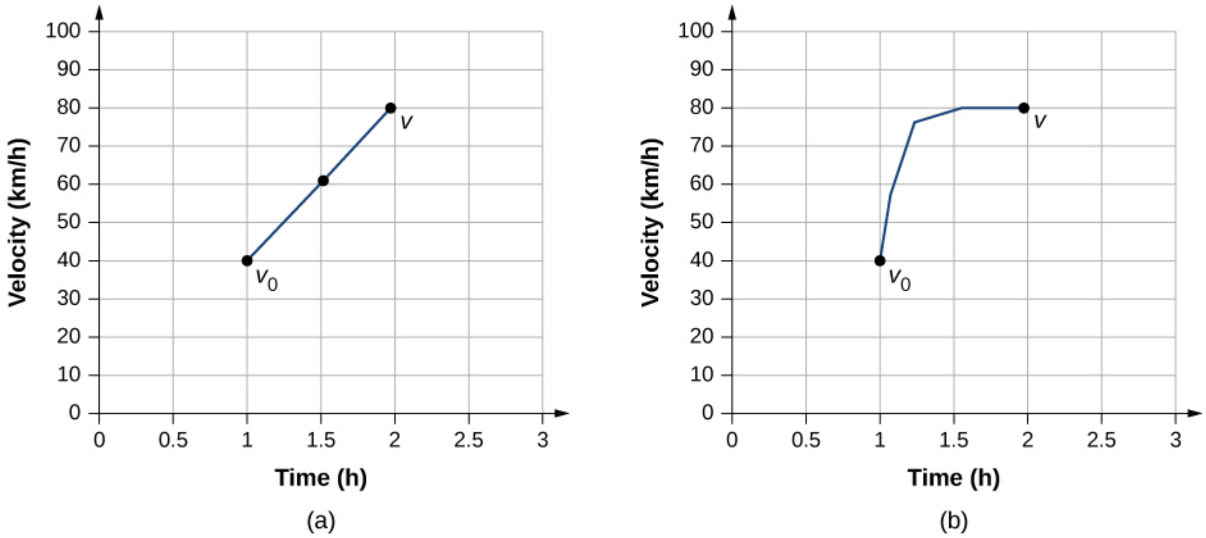


Figure 3.16 (a) Velocity-versus-time graph with constant acceleration showing the initial and final velocities v_0 and v. The average velocity is $\frac{1}{2}(v_0 + v) = 60\ \text{km/h}$. (b) Velocity-versus-time graph with an acceleration that changes with time. The average velocity is not given by $\frac{1}{2}(v_0 + v)$, but is greater than 60 km/h.

Solving for Final Velocity from Acceleration and Time

We can derive another useful equation by manipulating the definition of acceleration:

$$a = \frac{\Delta v}{\Delta t}.$$

Substituting the simplified notation for Δv and Δt gives us

$$a = \frac{v - v_0}{t}\ \text{(constant } a\text{)}.$$

Solving for v yields

$$v = v_0 + at\ \text{(constant } a\text{)}. \quad (3.11)$$

Example 3.4

Calculating Final Velocity

An airplane lands with an initial velocity of 70.0 m/s and then decelerates at 1.50 m/s^2 for 40.0 s. What is its final velocity?

Strategy

First, we identify the knowns: $v_0 = 70\ \text{m/s}, a = -1.50\ \text{m/s}^2, t = 40\ \text{s}$.

Second, we identify the unknown; in this case, it is final velocity v_f.

Last, we determine which equation to use. To do this we figure out which kinematic equation gives the unknown in terms of the knowns. We calculate the final velocity using **Equation 3.11**, $v = v_0 + at$.

Solution

Substitute the known values and solve:

$$v = v_0 + at = 70.0\text{ m/s} + \left(-1.50\text{ m/s}^2\right)(40.0\text{ s}) = 10.0\text{ m/s}.$$

Figure 3.17 is a sketch that shows the acceleration and velocity vectors.

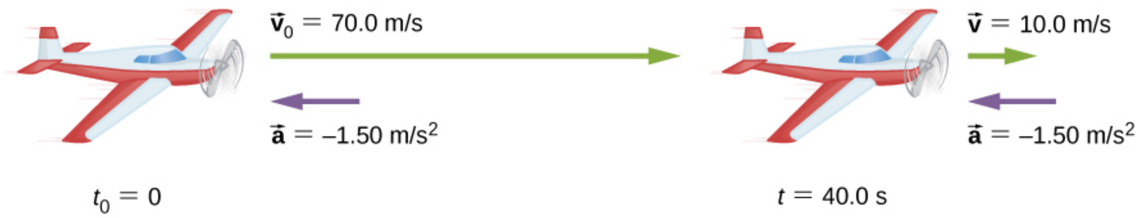


Figure 3.17 The airplane lands with an initial velocity of 70.0 m/s and slows to a final velocity of 10.0 m/s before heading for the terminal. Note the acceleration is negative because its direction is opposite to its velocity, which is positive.

Significance

The final velocity is much less than the initial velocity, as desired when slowing down, but is still positive (see figure). With jet engines, reverse thrust can be maintained long enough to stop the plane and start moving it backward, which is indicated by a negative final velocity, but is not the case here.

In addition to being useful in problem solving, the equation $v = v_0 + at$ gives us insight into the relationships among velocity, acceleration, and time. We can see, for example, that

- Final velocity depends on how large the acceleration is and how long it lasts
- If the acceleration is zero, then the final velocity equals the initial velocity ($v = v_0$), as expected (in other words, velocity is constant)
- If a is negative, then the final velocity is less than the initial velocity

All these observations fit our intuition. Note that it is always useful to examine basic equations in light of our intuition and experience to check that they do indeed describe nature accurately.

Solving for Final Position with Constant Acceleration

We can combine the previous equations to find a third equation that allows us to calculate the final position of an object experiencing constant acceleration. We start with

$$v = v_0 + at.$$

Adding v_0 to each side of this equation and dividing by 2 gives

$$\frac{v_0 + v}{2} = v_0 + \frac{1}{2}at.$$

Since $\frac{v_0 + v}{2} = \bar{v}$ for constant acceleration, we have

$$\bar{v} = v_0 + \frac{1}{2}at.$$

Now we substitute this expression for $\bar{v}$ into the equation for displacement, $x = x_0 + \bar{v}t$, yielding

$$x = x_0 + v_0 t + \frac{1}{2}at^2 \quad (\text{constant } a). \tag{3.12}$$

Example 3.5

Calculating Displacement of an Accelerating Object

Dragsters can achieve an average acceleration of 26.0 m/s^2. Suppose a dragster accelerates from rest at this rate for 5.56 s **Figure 3.18**. How far does it travel in this time?

Figure 3.18 U.S. Army Top Fuel pilot Tony "The Sarge" Schumacher begins a race with a controlled burnout. (credit: Lt. Col. William Thurmond. Photo Courtesy of U.S. Army.)

Strategy

First, let's draw a sketch **Figure 3.19**. We are asked to find displacement, which is *x* if we take x_0 to be zero. (Think about x_0 as the starting line of a race. It can be anywhere, but we call it zero and measure all other positions relative to it.) We can use the equation $x = x_0 + v_0 t + \frac{1}{2}at^2$ when we identify v_0, a, and *t* from the statement of the problem.

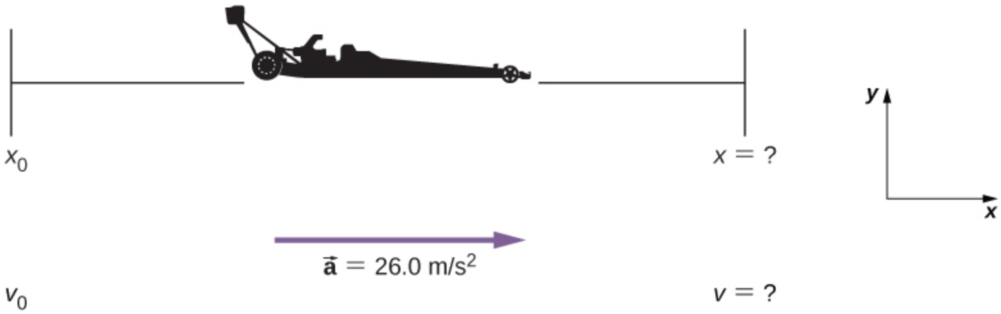


Figure 3.19 Sketch of an accelerating dragster.

Solution

First, we need to identify the knowns. Starting from rest means that $v_0 = 0$, *a* is given as 26.0 m/s^2 and *t* is given as 5.56 s.

Second, we substitute the known values into the equation to solve for the unknown:

$$x = x_0 + v_0 t + \frac{1}{2}at^2.$$

Since the initial position and velocity are both zero, this equation simplifies to

$$x = \frac{1}{2}at^2.$$

Substituting the identified values of a and t gives

$$x = \frac{1}{2}(26.0\ \text{m/s}^2)(5.56\ \text{s})^2 = 402\ \text{m}.$$

Significance

If we convert 402 m to miles, we find that the distance covered is very close to one-quarter of a mile, the standard distance for drag racing. So, our answer is reasonable. This is an impressive displacement to cover in only 5.56 s, but top-notch dragsters can do a quarter mile in even less time than this. If the dragster were given an initial velocity, this would add another term to the distance equation. If the same acceleration and time are used in the equation, the distance covered would be much greater.

What else can we learn by examining the equation $x = x_0 + v_0 t + \frac{1}{2}at^2$? We can see the following relationships:

- Displacement depends on the square of the elapsed time when acceleration is not zero. In **Example 3.5**, the dragster covers only one-fourth of the total distance in the first half of the elapsed time.
- If acceleration is zero, then initial velocity equals average velocity $(v_0 = \bar{v})$, and $x = x_0 + v_0 t + \frac{1}{2}at^2$ becomes $x = x_0 + v_0 t$.

Solving for Final Velocity from Distance and Acceleration

A fourth useful equation can be obtained from another algebraic manipulation of previous equations. If we solve $v = v_0 + at$ for t, we get

$$t = \frac{v - v_0}{a}.$$

Substituting this and $\bar{v} = \frac{v_0 + v}{2}$ into $x = x_0 + \bar{v}t$, we get

$$v^2 = v_0^2 + 2a(x - x_0)\ \ (\text{constant } a). \quad (3.13)$$

Example 3.6

Calculating Final Velocity

Calculate the final velocity of the dragster in **Example 3.5** without using information about time.

Strategy

The equation $v^2 = v_0^2 + 2a(x - x_0)$ is ideally suited to this task because it relates velocities, acceleration, and displacement, and no time information is required.

Solution

First, we identify the known values. We know that $v_0 = 0$, since the dragster starts from rest. We also know that $x - x_0 = 402$ m (this was the answer in **Example 3.5**). The average acceleration was given by $a = 26.0\ \text{m/s}^2$.

Second, we substitute the knowns into the equation $v^2 = v_0^2 + 2a(x - x_0)$ and solve for v:

$$v^2 = 0 + 2\left(26.0\ \text{m/s}^2\right)(402\ \text{m}).$$

Thus,

$$v^2 = 2.09 \times 10^4 \ \text{m}^2/\text{s}^2$$
$$v = \sqrt{2.09 \times 10^4 \ \text{m}^2/\text{s}^2} = 145 \ \text{m/s}.$$

Significance

A velocity of 145 m/s is about 522 km/h, or about 324 mi/h, but even this breakneck speed is short of the record for the quarter mile. Also, note that a square root has two values; we took the positive value to indicate a velocity in the same direction as the acceleration.

An examination of the equation $v^2 = v_0^2 + 2a(x - x_0)$ can produce additional insights into the general relationships among physical quantities:

- The final velocity depends on how large the acceleration is and the distance over which it acts.
- For a fixed acceleration, a car that is going twice as fast doesn't simply stop in twice the distance. It takes much farther to stop. (This is why we have reduced speed zones near schools.)

Putting Equations Together

In the following examples, we continue to explore one-dimensional motion, but in situations requiring slightly more algebraic manipulation. The examples also give insight into problem-solving techniques. The note that follows is provided for easy reference to the equations needed. Be aware that these equations are not independent. In many situations we have two unknowns and need two equations from the set to solve for the unknowns. We need as many equations as there are unknowns to solve a given situation.

Summary of Kinematic Equations (constant acceleration)

$$x = x_0 + \bar{v}t$$
$$\bar{v} = \frac{v_0 + v}{2}$$
$$v = v_0 + at$$
$$x = x_0 + v_0 t + \frac{1}{2}at^2$$
$$v^2 = v_0^2 + 2a(x - x_0)$$

Before we get into the examples, let's look at some of the equations more closely to see the behavior of acceleration at extreme values. Rearranging **Equation 3.11**, we have

$$a = \frac{v - v_0}{t}.$$

From this we see that, for a finite time, if the difference between the initial and final velocities is small, the acceleration is small, approaching zero in the limit that the initial and final velocities are equal. On the contrary, in the limit $t \to 0$ for a finite difference between the initial and final velocities, acceleration becomes infinite.

Similarly, rearranging **Equation 3.13**, we can express acceleration in terms of velocities and displacement:

$$a = \frac{v^2 - v_0^2}{2(x - x_0)}.$$

Thus, for a finite difference between the initial and final velocities acceleration becomes infinite in the limit the displacement approaches zero. Acceleration approaches zero in the limit the difference in initial and final velocities approaches zero for a finite displacement.

Example 3.7

How Far Does a Car Go?

On dry concrete, a car can decelerate at a rate of 7.00 m/s^2, whereas on wet concrete it can decelerate at only 5.00 m/s^2. Find the distances necessary to stop a car moving at 30.0 m/s (about 110 km/h) on (a) dry concrete and (b) wet concrete. (c) Repeat both calculations and find the displacement from the point where the driver sees a traffic light turn red, taking into account his reaction time of 0.500 s to get his foot on the brake.

Strategy

First, we need to draw a sketch **Figure 3.20**. To determine which equations are best to use, we need to list all the known values and identify exactly what we need to solve for.

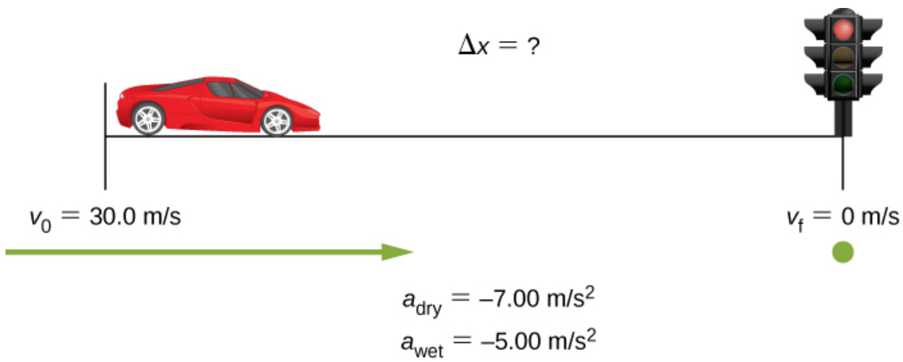


Figure 3.20 Sample sketch to visualize deceleration and stopping distance of a car.

Solution

a. First, we need to identify the knowns and what we want to solve for. We know that v_0 = 30.0 m/s, v = 0, and a = −7.00 m/s^2 (a is negative because it is in a direction opposite to velocity). We take x_0 to be zero. We are looking for displacement Δx, or $x - x_0$.
Second, we identify the equation that will help us solve the problem. The best equation to use is

$$v^2 = v_0^2 + 2a(x - x_0).$$

This equation is best because it includes only one unknown, x. We know the values of all the other variables in this equation. (Other equations would allow us to solve for x, but they require us to know the stopping time, t, which we do not know. We could use them, but it would entail additional calculations.)
Third, we rearrange the equation to solve for x:

$$x - x_0 = \frac{v^2 - v_0^2}{2a}$$

and substitute the known values:

$$x - 0 = \frac{0^2 - (30.0\text{ m/s})^2}{2(-7.00\text{m/s}^2)}.$$

Thus,

$$x = 64.3\text{ m on dry concrete.}$$

b. This part can be solved in exactly the same manner as (a). The only difference is that the acceleration is −5.00 m/s^2. The result is

$$x_{\text{wet}} = 90.0\text{ m on wet concrete.}$$

c. When the driver reacts, the stopping distance is the same as it is in (a) and (b) for dry and wet concrete. So, to answer this question, we need to calculate how far the car travels during the reaction time, and then add that to the stopping time. It is reasonable to assume the velocity remains constant during the driver's reaction time.

To do this, we, again, identify the knowns and what we want to solve for. We know that $\bar{v} = 30.0\,\text{m/s}$, $t_{\text{reaction}} = 0.500\,\text{s}$, and $a_{\text{reaction}} = 0$. We take $x_{0\text{-reaction}}$ to be zero. We are looking for x_{reaction}.

Second, as before, we identify the best equation to use. In this case, $x = x_0 + \bar{v}t$ works well because the only unknown value is *x*, which is what we want to solve for.

Third, we substitute the knowns to solve the equation:

$$x = 0 + (30.0\,\text{m/s})(0.500\,\text{s}) = 15.0\,\text{m}.$$

This means the car travels 15.0 m while the driver reacts, making the total displacements in the two cases of dry and wet concrete 15.0 m greater than if he reacted instantly.

Last, we then add the displacement during the reaction time to the displacement when braking (**Figure 3.21**),

$$x_{\text{braking}} + x_{\text{reaction}} = x_{\text{total}},$$

and find (a) to be 64.3 m + 15.0 m = 79.3 m when dry and (b) to be 90.0 m + 15.0 m = 105 m when wet.

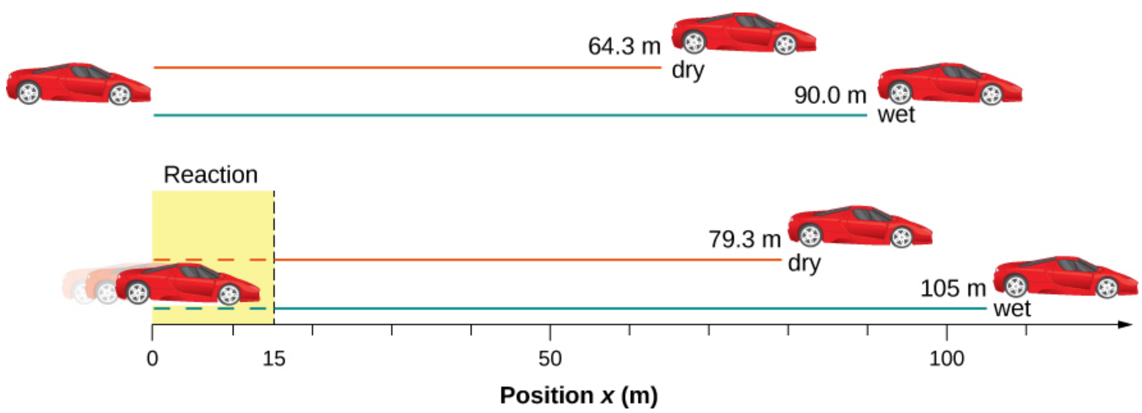


Figure 3.21 The distance necessary to stop a car varies greatly, depending on road conditions and driver reaction time. Shown here are the braking distances for dry and wet pavement, as calculated in this example, for a car traveling initially at 30.0 m/s. Also shown are the total distances traveled from the point when the driver first sees a light turn red, assuming a 0.500-s reaction time.

Significance

The displacements found in this example seem reasonable for stopping a fast-moving car. It should take longer to stop a car on wet pavement than dry. It is interesting that reaction time adds significantly to the displacements, but more important is the general approach to solving problems. We identify the knowns and the quantities to be determined, then find an appropriate equation. If there is more than one unknown, we need as many independent equations as there are unknowns to solve. There is often more than one way to solve a problem. The various parts of this example can, in fact, be solved by other methods, but the solutions presented here are the shortest.

Example 3.8

Calculating Time

Suppose a car merges into freeway traffic on a 200-m-long ramp. If its initial velocity is 10.0 m/s and it accelerates at 2.00 m/s^2, how long does it take the car to travel the 200 m up the ramp? (Such information might be useful to a traffic engineer.)

Strategy

First, we draw a sketch **Figure 3.22**. We are asked to solve for time *t*. As before, we identify the known quantities to choose a convenient physical relationship (that is, an equation with one unknown, *t*.)

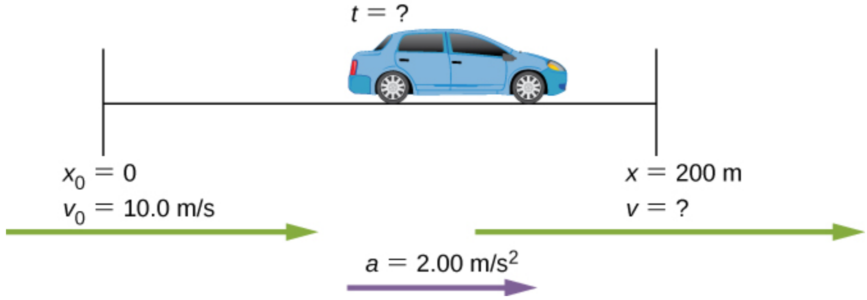


Figure 3.22 Sketch of a car accelerating on a freeway ramp.

Solution

Again, we identify the knowns and what we want to solve for. We know that $x_0 = 0$, $v_0 = 10\text{ m/s}$, $a = 2.00\text{ m/s}^2$, and $x = 200$ m.

We need to solve for *t*. The equation $x = x_0 + v_0 t + \frac{1}{2}at^2$ works best because the only unknown in the equation is the variable *t*, for which we need to solve. From this insight we see that when we input the knowns into the equation, we end up with a quadratic equation.

We need to rearrange the equation to solve for *t*, then substituting the knowns into the equation:

$$200\text{ m} = 0\text{ m} + (10.0\text{ m/s})t + \frac{1}{2}\left(2.00\text{ m/s}^2\right)t^2.$$

We then simplify the equation. The units of meters cancel because they are in each term. We can get the units of seconds to cancel by taking *t* = *t* s, where *t* is the magnitude of time and s is the unit. Doing so leaves

$$200 = 10t + t^2.$$

We then use the quadratic formula to solve for *t*,

$$t^2 + 10t - 200 = 0$$
$$t = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a},$$

which yields two solutions: *t* = 10.0 and *t* = −20.0. A negative value for time is unreasonable, since it would mean the event happened 20 s before the motion began. We can discard that solution. Thus,

$$t = 10.0\text{ s}.$$

Significance

Whenever an equation contains an unknown squared, there are two solutions. In some problems both solutions are meaningful; in others, only one solution is reasonable. The 10.0-s answer seems reasonable for a typical freeway on-ramp.

3.4 Check Your Understanding A manned rocket accelerates at a rate of 20 m/s^2 during launch. How long does it take the rocket to reach a velocity of 400 m/s?

Example 3.9

Acceleration of a Spaceship

A spaceship has left Earth's orbit and is on its way to the Moon. It accelerates at 20 m/s^2 for 2 min and covers a distance of 1000 km. What are the initial and final velocities of the spaceship?

Strategy

We are asked to find the initial and final velocities of the spaceship. Looking at the kinematic equations, we see that one equation will not give the answer. We must use one kinematic equation to solve for one of the velocities and substitute it into another kinematic equation to get the second velocity. Thus, we solve two of the kinematic equations simultaneously.

Solution

First we solve for v_0 using $x = x_0 + v_0 t + \frac{1}{2}at^2$:

$$x - x_0 = v_0 t + \frac{1}{2}at^2$$

$$1.0 \times 10^6 \text{ m} = v_0(120.0 \text{ s}) + \frac{1}{2}(20.0 \text{ m/s}^2)(120.0 \text{ s})^2$$

$$v_0 = 7133.3 \text{ m/s}.$$

Then we substitute v_0 into $v = v_0 + at$ to solve for the final velocity:

$$v = v_0 + at = 7133.3 \text{ m/s} + (20.0 \text{ m/s}^2)(120.0 \text{ s}) = 9533.3 \text{ m/s}.$$

Significance

There are six variables in displacement, time, velocity, and acceleration that describe motion in one dimension. The initial conditions of a given problem can be many combinations of these variables. Because of this diversity, solutions may not be as easy as simple substitutions into one of the equations. This example illustrates that solutions to kinematics may require solving two simultaneous kinematic equations.

With the basics of kinematics established, we can go on to many other interesting examples and applications. In the process of developing kinematics, we have also glimpsed a general approach to problem solving that produces both correct answers and insights into physical relationships. The next level of complexity in our kinematics problems involves the motion of two interrelated bodies, called *two-body pursuit problems*.

Two-Body Pursuit Problems

Up until this point we have looked at examples of motion involving a single body. Even for the problem with two cars and the stopping distances on wet and dry roads, we divided this problem into two separate problems to find the answers. In a **two-body pursuit problem**, the motions of the objects are coupled—meaning, the unknown we seek depends on the motion of both objects. To solve these problems we write the equations of motion for each object and then solve them simultaneously to find the unknown. This is illustrated in **Figure 3.23**.

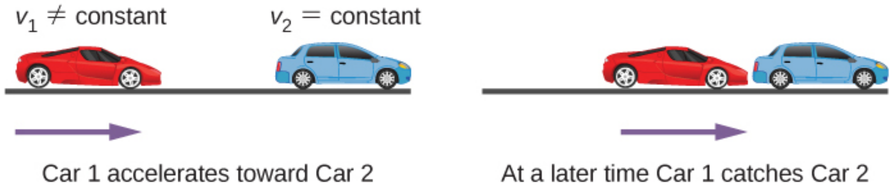


Figure 3.23 A two-body pursuit scenario where car 2 has a constant velocity and car 1 is behind with a constant acceleration. Car 1 catches up with car 2 at a later time.

The time and distance required for car 1 to catch car 2 depends on the initial distance car 1 is from car 2 as well as the velocities of both cars and the acceleration of car 1. The kinematic equations describing the motion of both cars must be solved to find these unknowns.

Consider the following example.

Example 3.10

Cheetah Catching a Gazelle

A cheetah waits in hiding behind a bush. The cheetah spots a gazelle running past at 10 m/s. At the instant the gazelle passes the cheetah, the cheetah accelerates from rest at 4 m/s^2 to catch the gazelle. (a) How long does it take the cheetah to catch the gazelle? (b) What is the displacement of the gazelle and cheetah?

Strategy

We use the set of equations for constant acceleration to solve this problem. Since there are two objects in motion, we have separate equations of motion describing each animal. But what links the equations is a common parameter that has the same value for each animal. If we look at the problem closely, it is clear the common parameter to each animal is their position *x* at a later time *t*. Since they both start at $x_0 = 0$, their displacements are the same at a later time *t*, when the cheetah catches up with the gazelle. If we pick the equation of motion that solves for the displacement for each animal, we can then set the equations equal to each other and solve for the unknown, which is time.

Solution

a. Equation for the gazelle: The gazelle has a constant velocity, which is its average velocity, since it is not accelerating. Therefore, we use **Equation 3.9** with $x_0 = 0$:

$$x = x_0 + \bar{v}t = \bar{v}t.$$

Equation for the cheetah: The cheetah is accelerating from rest, so we use **Equation 3.12** with $x_0 = 0$ and $v_0 = 0$:

$$x = x_0 + v_0 t + \frac{1}{2}at^2 = \frac{1}{2}at^2.$$

Now we have an equation of motion for each animal with a common parameter, which can be eliminated to find the solution. In this case, we solve for *t*:

$$x = \bar{v}t = \frac{1}{2}at^2$$
$$t = \frac{2\bar{v}}{a}.$$

The gazelle has a constant velocity of 10 m/s, which is its average velocity. The acceleration of the cheetah is 4 m/s^2. Evaluating *t*, the time for the cheetah to reach the gazelle, we have

$$t = \frac{2\bar{v}}{a} = \frac{2(10)}{4} = 5\text{ s}.$$

b. To get the displacement, we use either the equation of motion for the cheetah or the gazelle, since they should both give the same answer.
Displacement of the cheetah:

$$x = \frac{1}{2}at^2 = \frac{1}{2}(4)(5)^2 = 50\text{ m}.$$

Displacement of the gazelle:

$$x = \bar{v}t = 10(5) = 50\text{ m}.$$

We see that both displacements are equal, as expected.

Significance

It is important to analyze the motion of each object and to use the appropriate kinematic equations to describe the individual motion. It is also important to have a good visual perspective of the two-body pursuit problem to see the common parameter that links the motion of both objects.

3.5 Check Your Understanding A bicycle has a constant velocity of 10 m/s. A person starts from rest and runs to catch up to the bicycle in 30 s. What is the acceleration of the person?

3.5 | Free Fall

Learning Objectives

By the end of this section, you will be able to:

- Use the kinematic equations with the variables y and g to analyze free-fall motion.
- Describe how the values of the position, velocity, and acceleration change during a free fall.
- Solve for the position, velocity, and acceleration as functions of time when an object is in a free fall.

An interesting application of **Equation 3.4** through **Equation 3.13** is called *free fall*, which describes the motion of an object falling in a gravitational field, such as near the surface of Earth or other celestial objects of planetary size. Let's assume the body is falling in a straight line perpendicular to the surface, so its motion is one-dimensional. For example, we can estimate the depth of a vertical mine shaft by dropping a rock into it and listening for the rock to hit the bottom. But "falling," in the context of free fall, does not necessarily imply the body is moving from a greater height to a lesser height. If a ball is thrown upward, the equations of free fall apply equally to its ascent as well as its descent.

Gravity

The most remarkable and unexpected fact about falling objects is that if air resistance and friction are negligible, then in a given location all objects fall toward the center of Earth with the *same constant acceleration, independent of their mass*. This experimentally determined fact is unexpected because we are so accustomed to the effects of air resistance and friction that we expect light objects to fall slower than heavy ones. Until Galileo Galilei (1564–1642) proved otherwise, people believed that a heavier object has a greater acceleration in a free fall. We now know this is not the case. In the absence of air resistance, heavy objects arrive at the ground at the same time as lighter objects when dropped from the same height **Figure 3.24**.

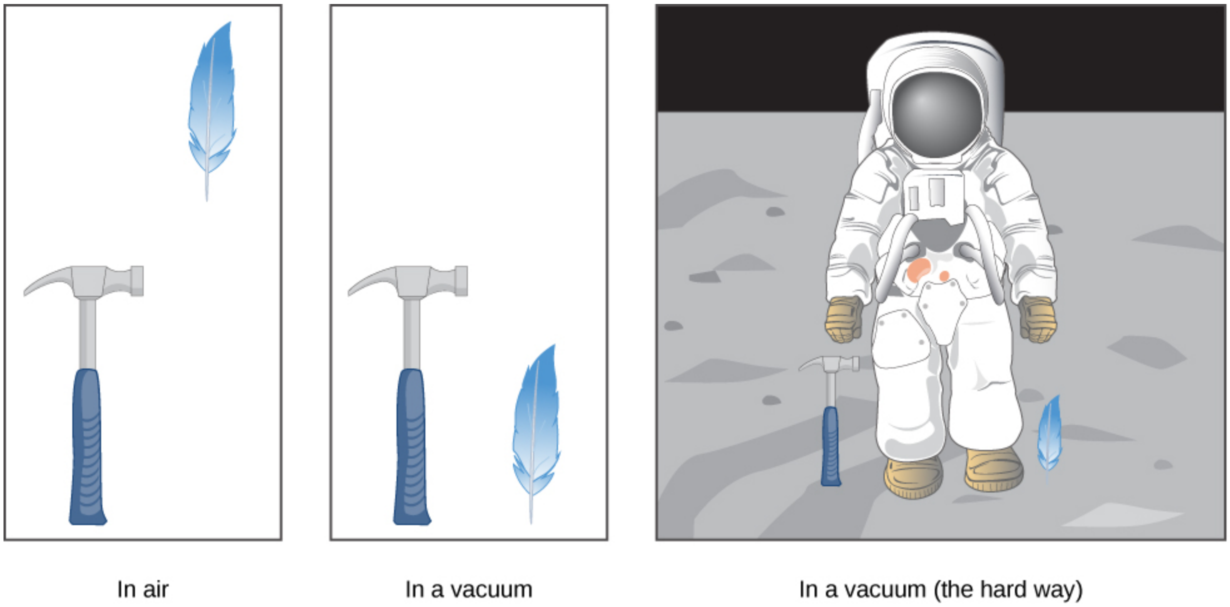


Figure 3.24 A hammer and a feather fall with the same constant acceleration if air resistance is negligible. This is a general characteristic of gravity not unique to Earth, as astronaut David R. Scott demonstrated in 1971 on the Moon, where the acceleration from gravity is only 1.67 m/s2 and there is no atmosphere.

In the real world, air resistance can cause a lighter object to fall slower than a heavier object of the same size. A tennis ball reaches the ground after a baseball dropped at the same time. (It might be difficult to observe the difference if the height is not large.) Air resistance opposes the motion of an object through the air, and friction between objects—such as between clothes and a laundry chute or between a stone and a pool into which it is dropped—also opposes motion between them.

For the ideal situations of these first few chapters, an object *falling without air resistance or friction* is defined to be in **free fall**. The force of gravity causes objects to fall toward the center of Earth. The acceleration of free-falling objects is therefore called **acceleration due to gravity**. Acceleration due to gravity is constant, which means we can apply the kinematic equations to any falling object where air resistance and friction are negligible. This opens to us a broad class of interesting situations.

Acceleration due to gravity is so important that its magnitude is given its own symbol, g. It is constant at any given location on Earth and has the average value

$$g = 9.81 \text{ m/s}^2 \quad (\text{or } 32.2 \text{ ft/s}^2).$$

Although g varies from 9.78 m/s^2 to 9.83 m/s^2, depending on latitude, altitude, underlying geological formations, and local topography, let's use an average value of 9.8 m/s^2 rounded to two significant figures in this text unless specified otherwise. Neglecting these effects on the value of g as a result of position on Earth's surface, as well as effects resulting from Earth's rotation, we take the direction of acceleration due to gravity to be downward (toward the center of Earth). In fact, its direction *defines* what we call vertical. Note that whether acceleration a in the kinematic equations has the value $+g$ or $-g$ depends on how we define our coordinate system. If we define the upward direction as positive, then $a = -g = -9.8 \text{ m/s}^2$, and if we define the downward direction as positive, then $a = g = 9.8 \text{ m/s}^2$.

One-Dimensional Motion Involving Gravity

The best way to see the basic features of motion involving gravity is to start with the simplest situations and then progress toward more complex ones. So, we start by considering straight up-and-down motion with no air resistance or friction. These assumptions mean the velocity (if there is any) is vertical. If an object is dropped, we know the initial velocity is zero when in free fall. When the object has left contact with whatever held or threw it, the object is in free fall. When the object is thrown, it has the same initial speed in free fall as it did before it was released. When the object comes in contact with the ground or any other object, it is no longer in free fall and its acceleration of g is no longer valid. Under these circumstances, the motion is one-dimensional and has constant acceleration of magnitude g. We represent vertical displacement with the symbol y.

Kinematic Equations for Objects in Free Fall

We assume here that acceleration equals $-g$ (with the positive direction upward).

$$v = v_0 - gt \tag{3.14}$$

$$y = y_0 + v_0 t - \frac{1}{2}gt^2 \tag{3.15}$$

$$v^2 = v_0^2 - 2g(y - y_0) \tag{3.16}$$

Problem-Solving Strategy: Free Fall

1. Decide on the sign of the acceleration of gravity. In **Equation 3.14** through **Equation 3.16**, acceleration g is negative, which says the positive direction is upward and the negative direction is downward. In some problems, it may be useful to have acceleration g as positive, indicating the positive direction is downward.
2. Draw a sketch of the problem. This helps visualize the physics involved.
3. Record the knowns and unknowns from the problem description. This helps devise a strategy for selecting the appropriate equations to solve the problem.
4. Decide which of **Equation 3.14** through **Equation 3.16** are to be used to solve for the unknowns.

Example 3.11

Free Fall of a Ball

Figure 3.25 shows the positions of a ball, at 1-s intervals, with an initial velocity of 4.9 m/s downward, that is thrown from the top of a 98-m-high building. (a) How much time elapses before the ball reaches the ground? (b) What is the velocity when it arrives at the ground?

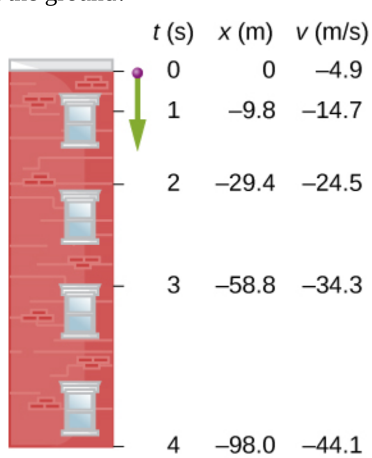


Figure 3.25 The positions and velocities at 1-s intervals of a ball thrown downward from a tall building at 4.9 m/s.

Strategy

Choose the origin at the top of the building with the positive direction upward and the negative direction downward. To find the time when the position is −98 m, we use **Equation 3.15**, with $y_0 = 0$, $v_0 = -4.9\ \text{m/s}$, and $g = 9.8\ \text{m/s}^2$.

Solution

a. Substitute the given values into the equation:

$$y = y_0 + v_0 t - \frac{1}{2}gt^2$$
$$-98.0\text{ m} = 0 - (4.9\text{ m/s})t - \frac{1}{2}(9.8\text{ m/s}^2)t^2.$$

This simplifies to

$$t^2 + t - 20 = 0.$$

This is a quadratic equation with roots $t = -5.0\text{s}$ and $t = 4.0\text{s}$. The positive root is the one we are interested in, since time $t = 0$ is the time when the ball is released at the top of the building. (The time $t = -5.0\text{s}$ represents the fact that a ball thrown upward from the ground would have been in the air for 5.0 s when it passed by the top of the building moving downward at 4.9 m/s.)

b. Using **Equation 3.14**, we have

$$v = v_0 - gt = -4.9\text{ m/s} - (9.8\text{m/s}^2)(4.0\text{ s}) = -44.1\text{ m/s}.$$

Significance

For situations when two roots are obtained from a quadratic equation in the time variable, we must look at the physical significance of both roots to determine which is correct. Since $t = 0$ corresponds to the time when the ball was released, the negative root would correspond to a time before the ball was released, which is not physically meaningful. When the ball hits the ground, its velocity is not immediately zero, but as soon as the ball interacts with the ground, its acceleration is not *g* and it accelerates with a different value over a short time to zero velocity. This problem shows how important it is to establish the correct coordinate system and to keep the signs of *g* in the kinematic equations consistent.

Example 3.12

Vertical Motion of a Baseball

A batter hits a baseball straight upward at home plate and the ball is caught 5.0 s after it is struck **Figure 3.26**. (a) What is the initial velocity of the ball? (b) What is the maximum height the ball reaches? (c) How long does it take to reach the maximum height? (d) What is the acceleration at the top of its path? (e) What is the velocity of the ball when it is caught? Assume the ball is hit and caught at the same location.

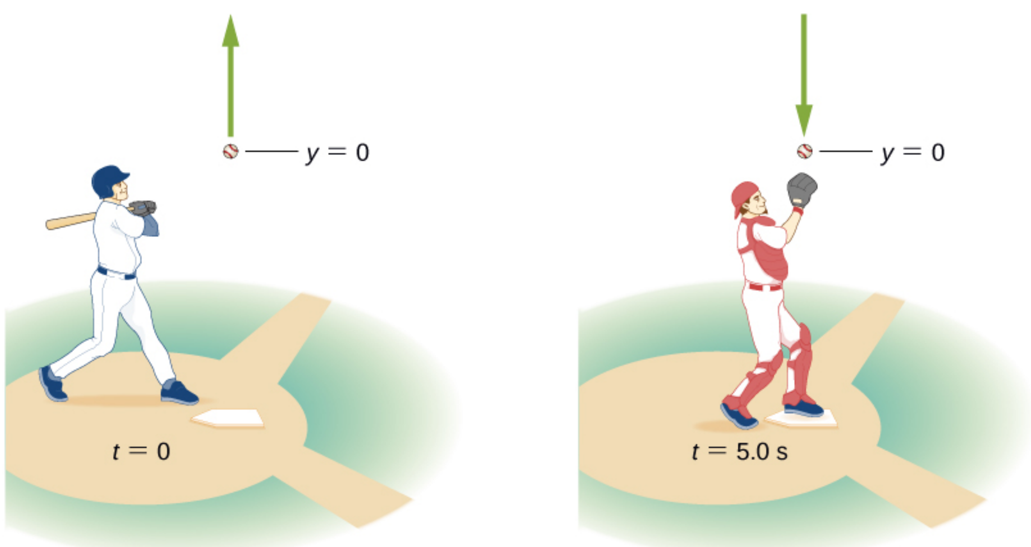


Figure 3.26 A baseball hit straight up is caught by the catcher 5.0 s later.

Strategy

Choose a coordinate system with a positive y-axis that is straight up and with an origin that is at the spot where the ball is hit and caught.

Solution

a. **Equation 3.15** gives

$$y = y_0 + v_0 t - \frac{1}{2} g t^2$$

$$0 = 0 + v_0(5.0 \text{ s}) - \frac{1}{2}\left(9.8 \text{ m/s}^2\right)(5.0 \text{ s})^2,$$

which gives $v_0 = 24.5 \text{ m/sec}$.

b. At the maximum height, $v = 0$. With $v_0 = 24.5 \text{ m/s}$, **Equation 3.16** gives

$$v^2 = v_0^2 - 2g(y - y_0)$$

$$0 = (24.5 \text{ m/s})^2 - 2(9.8 \text{m/s}^2)(y - 0)$$

or

$$y = 30.6 \text{ m}.$$

c. To find the time when $v = 0$, we use **Equation 3.14**:

$$v = v_0 - gt$$

$$0 = 24.5 \text{ m/s} - (9.8 \text{m/s}^2)t.$$

This gives $t = 2.5 \text{ s}$. Since the ball rises for 2.5 s, the time to fall is 2.5 s.

d. The acceleration is 9.8 m/s^2 everywhere, even when the velocity is zero at the top of the path. Although the velocity is zero at the top, it is changing at the rate of 9.8 m/s^2 downward.

e. The velocity at $t = 5.0\text{s}$ can be determined with **Equation 3.14**:

$$\begin{aligned} v &= v_0 - gt \\ &= 24.5 \text{ m/s} - 9.8 \text{m/s}^2(5.0 \text{ s}) \\ &= -24.5 \text{ m/s}. \end{aligned}$$

Significance

The ball returns with the speed it had when it left. This is a general property of free fall for any initial velocity. We used a single equation to go from throw to catch, and did not have to break the motion into two segments, upward and downward. We are used to thinking of the effect of gravity is to create free fall downward toward Earth. It is important to understand, as illustrated in this example, that objects moving upward away from Earth are also in a state of free fall.

3.6 Check Your Understanding A chunk of ice breaks off a glacier and falls 30.0 m before it hits the water. Assuming it falls freely (there is no air resistance), how long does it take to hit the water? Which quantity increases faster, the speed of the ice chunk or its distance traveled?

Example 3.13

Rocket Booster

A small rocket with a booster blasts off and heads straight upward. When at a height of $5.0\,\text{km}$ and velocity of 200.0 m/s, it releases its booster. (a) What is the maximum height the booster attains? (b) What is the velocity of the booster at a height of 6.0 km? Neglect air resistance.

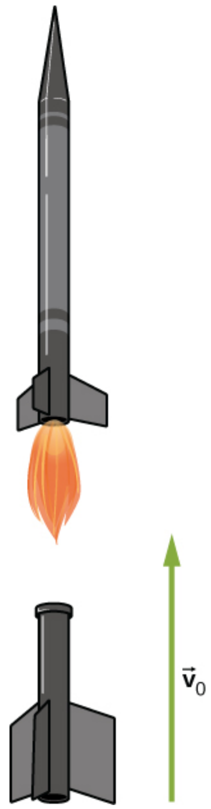


Figure 3.27 A rocket releases its booster at a given height and velocity. How high and how fast does the booster go?

Strategy

We need to select the coordinate system for the acceleration of gravity, which we take as negative downward. We are given the initial velocity of the booster and its height. We consider the point of release as the origin. We know the velocity is zero at the maximum position within the acceleration interval; thus, the velocity of the booster is zero at its maximum height, so we can use this information as well. From these observations, we use **Equation 3.16**, which gives us the maximum height of the booster. We also use **Equation 3.16** to give the velocity at 6.0 km. The initial velocity of the booster is 200.0 m/s.

Solution

a. From **Equation 3.16**, $v^2 = v_0^2 - 2g(y - y_0)$. With $v = 0$ and $y_0 = 0$, we can solve for y:

$$y = \frac{v_0^2}{-2g} = \frac{(2.0 \times 10^2\,\text{m/s})^2}{-2(9.8\,\text{m/s}^2)} = 2040.8\,\text{m}.$$

This solution gives the maximum height of the booster in our coordinate system, which has its origin at the point of release, so the maximum height of the booster is roughly 7.0 km.

b. An altitude of 6.0 km corresponds to $y = 1.0 \times 10^3\,\text{m}$ in the coordinate system we are using. The other initial conditions are $y_0 = 0,$ and $v_0 = 200.0 \text{ m/s}$.

We have, from **Equation 3.16**,

$$v^2 = (200.0 \text{ m/s})^2 - 2(9.8 \text{ m/s}^2)(1.0 \times 10^3 \text{ m}) \Rightarrow v = \pm 142.8 \text{ m/s}.$$

Significance

We have both a positive and negative solution in (b). Since our coordinate system has the positive direction upward, the +142.8 m/s corresponds to a positive upward velocity at 6000 m during the upward leg of the trajectory of the booster. The value $v = -142.8$ m/s corresponds to the velocity at 6000 m on the downward leg. This example is also important in that an object is given an initial velocity at the origin of our coordinate system, but the origin is at an altitude above the surface of Earth, which must be taken into account when forming the solution.

Visit **this site (https://openstaxcollege.org/l/21equatgraph)** to learn about graphing polynomials. The shape of the curve changes as the constants are adjusted. View the curves for the individual terms (for example, $y = bx$) to see how they add to generate the polynomial curve.

CHAPTER 3 REVIEW

KEY TERMS

acceleration due to gravity acceleration of an object as a result of gravity

average acceleration the rate of change in velocity; the change in velocity over time

average speed the total distance traveled divided by elapsed time

average velocity the displacement divided by the time over which displacement occurs

displacement the change in position of an object

distance traveled the total length of the path traveled between two positions

elapsed time the difference between the ending time and the beginning time

free fall the state of movement that results from gravitational force only

instantaneous acceleration acceleration at a specific point in time

instantaneous speed the absolute value of the instantaneous velocity

instantaneous velocity the velocity at a specific instant or time point

kinematics the description of motion through properties such as position, time, velocity, and acceleration

position the location of an object at a particular time

total displacement the sum of individual displacements over a given time period

two-body pursuit problem a kinematics problem in which the unknowns are calculated by solving the kinematic equations simultaneously for two moving objects

KEY EQUATIONS

Displacement	$\Delta x = x_{\text{f}} - x_{\text{i}}$
Total displacement	$\Delta x_{\text{Total}} = \sum \Delta x_{\text{i}}$
Average velocity	$\bar{v} = \frac{\Delta x}{\Delta t} = \frac{x_2 - x_1}{t_2 - t_1}$
Instantaneous velocity	$v(t) = \frac{dx(t)}{dt}$
Average speed	$\text{Average speed} = \bar{s} = \frac{\text{Total distance}}{\text{Elapsed time}}$
Instantaneous speed	$\text{Instantaneous speed} = \lvert v(t) \rvert$
Average acceleration	$\bar{a} = \frac{\Delta v}{\Delta t} = \frac{v_f - v_0}{t_f - t_0}$
Instantaneous acceleration	$a(t) = \frac{dv(t)}{dt}$
Position from average velocity	$x = x_0 + \bar{v}t$
Average velocity	$\bar{v} = \frac{v_0 + v}{2}$
Velocity from acceleration	$v = v_0 + at \;\; (\text{constant } a)$

Position from velocity and acceleration	$x = x_0 + v_0 t + \frac{1}{2}at^2$ (constant a)
Velocity from distance	$v^2 = v_0^2 + 2a(x - x_0)$ (constant a)
Velocity of free fall	$v = v_0 - gt$ (positive upward)
Height of free fall	$y = y_0 + v_0 t - \frac{1}{2}gt^2$
Velocity of free fall from height	$v^2 = v_0^2 - 2g(y - y_0)$

SUMMARY

3.1 Position, Displacement and Average Velocity

- Kinematics is the description of motion without considering its causes. In this chapter, it is limited to motion along a straight line, called one-dimensional motion.
- Displacement is the change in position of an object. The SI unit for displacement is the meter. Displacement has direction as well as magnitude.
- Displacement is a **vector** quantity, meaning that is has both a magnitude and a direction (which is either positive or negative).
- Distance traveled is the total length of the path traveled between two positions.
- Time is measured in terms of change. The time between two position points x_1 and x_2 is $\Delta t = t_2 - t_1$. Elapsed time for an event is $\Delta t = t_f - t_0$, where t_f is the final time and t_0 is the initial time. The initial time is often taken to be zero.
- Average velocity $\bar{v}$ is defined as displacement divided by elapsed time. If x_1, t_1 and x_2, t_2 are two position time points, the average velocity between these points is

$$\bar{v} = \frac{\Delta x}{\Delta t} = \frac{x_2 - x_1}{t_2 - t_1}.$$

- Velocity is also a vector quantity, which gets its direction (i.e. sign) from the sign of the displacement.

3.2 Instantaneous Velocity and Speed

- Instantaneous velocity gives the velocity at any point in time during a particle's motion.
- Instantaneous velocity is a vector and can be negative.
- Instantaneous speed is found by taking the absolute value of instantaneous velocity, and it is always positive.
- Average speed is total distance traveled divided by elapsed time.
- The slope of a position-versus-time graph at a specific time gives instantaneous velocity at that time.

3.3 Average and Instantaneous Acceleration

- Acceleration is the rate at which velocity changes. Acceleration is a vector; it has both a magnitude and direction. The SI unit for acceleration is meters per second squared.
- Acceleration can be caused by a change in the magnitude or the direction of the velocity, or both.
- Instantaneous acceleration $a(t)$ is a continuous function of time and gives the acceleration at any specific time during the motion. It is calculated from the derivative of the velocity function. Instantaneous acceleration is the slope of the velocity-versus-time graph.
- Negative acceleration (sometimes called deceleration) is acceleration in the negative direction in the chosen coordinate system.

3.4 Motion with Constant Acceleration

- When analyzing one-dimensional motion with constant acceleration, identify the known quantities and choose the appropriate equations to solve for the unknowns. Either one or two of the kinematic equations are needed to solve for the unknowns, depending on the known and unknown quantities.
- Two-body pursuit problems always require two equations to be solved simultaneously for the unknowns.

3.5 Free Fall

- An object in free fall experiences constant acceleration if air resistance is negligible.
- On Earth, all free-falling objects have an acceleration g due to gravity, which averages $g = 9.81 \text{ m/s}^2$.
- For objects in free fall, the upward direction is normally taken as positive for displacement, velocity, and acceleration.

CONCEPTUAL QUESTIONS

3.1 Position, Displacement and Average Velocity

1. Give an example in which there are clear distinctions among distance traveled, displacement, and magnitude of displacement. Identify each quantity in your example specifically.

2. Under what circumstances does distance traveled equal magnitude of displacement? What is the only case in which magnitude of displacement and displacement are exactly the same?

3. Bacteria move back and forth using their flagella (structures that look like little tails). Speeds of up to 50 μm/s (50×10^{-6} m/s) have been observed. The total distance traveled by a bacterium is large for its size, whereas its displacement is small. Why is this?

4. Give an example of a device used to measure time and identify what change in that device indicates a change in time.

5. Does a car's odometer measure distance traveled or displacement?

6. During a given time interval the average velocity of an object is zero. What can you say conclude about its displacement over the time interval?

3.2 Instantaneous Velocity and Speed

7. There is a distinction between average speed and the magnitude of average velocity. Give an example that illustrates the difference between these two quantities.

8. Does the speedometer of a car measure speed or velocity?

9. If you divide the total distance traveled on a car trip (as determined by the odometer) by the elapsed time of the trip, are you calculating average speed or magnitude of average velocity? Under what circumstances are these two quantities the same?

10. How are instantaneous velocity and instantaneous speed related to one another? How do they differ?

3.3 Average and Instantaneous Acceleration

11. Is it possible for speed to be constant while acceleration is not zero?

12. Is it possible for velocity to be constant while acceleration is not zero? Explain.

13. Give an example in which velocity is zero yet acceleration is not.

14. If a subway train is moving to the left (has a negative velocity) and then comes to a stop, what is the direction of its acceleration? Is the acceleration positive or negative?

15. Plus and minus signs are used in one-dimensional motion to indicate direction. What is the sign of an acceleration that reduces the magnitude of a negative velocity? Of a positive velocity?

3.4 Motion with Constant Acceleration

16. When analyzing the motion of a single object, what is the required number of known physical variables that are needed to solve for the unknown quantities using the kinematic equations?

17. State two scenarios of the kinematics of single object where three known quantities require two kinematic equations to solve for the unknowns.

3.5 Free Fall

18. What is the acceleration of a rock thrown straight upward on the way up? At the top of its flight? On the way down? Assume there is no air resistance.

19. An object that is thrown straight up falls back to Earth. This is one-dimensional motion. (a) When is its velocity zero? (b) Does its velocity change direction? (c) Does the acceleration have the same sign on the way up as on the way down?

20. Suppose you throw a rock nearly straight up at a coconut in a palm tree and the rock just misses the coconut on the way up but hits the coconut on the way down. Neglecting air resistance and the slight horizontal variation in motion to account for the hit and miss of the coconut, how does the speed of the rock when it hits the coconut on the way down compare with what it would have been if it had hit the coconut on the way up? Is it more likely to dislodge the coconut on the way up or down? Explain.

21. The severity of a fall depends on your speed when you strike the ground. All factors but the acceleration from gravity being the same, how many times higher could a safe fall on the Moon than on Earth (gravitational acceleration on the Moon is about one-sixth that of the Earth)?

22. How many times higher could an astronaut jump on the Moon than on Earth if her takeoff speed is the same in both locations (gravitational acceleration on the Moon is about on-sixth of that on Earth)?

23. When given the acceleration function, what additional information is needed to find the velocity function and position function?

PROBLEMS

3.1 Position, Displacement and Average Velocity

24. Consider a coordinate system in which the positive *x* axis is directed upward vertically. What are the positions of a particle (a) 5.0 m directly above the origin and (b) 2.0 m below the origin?

25. A car is 2.0 km west of a traffic light at $t = 0$ and 5.0 km east of the light at $t = 6.0$ min. Assume the origin of the coordinate system is the light and the positive *x* direction is eastward. (a) What are the car's position vectors at these two times? (b) What is the car's displacement between 0 min and 6.0 min?

26. The Shanghai maglev train connects Longyang Road to Pudong International Airport, a distance of 30 km. The journey takes 8 minutes on average. What is the maglev train's average velocity?

27. The position of a particle moving along the *x*-axis is given by $x(t) = 4.0 - 2.0t$ m. (a) At what time does the particle cross the origin? (b) What is the displacement of the particle between $\text{t} = 3.0\,\text{s}$ and $\text{t} = 6.0\,\text{s}$?

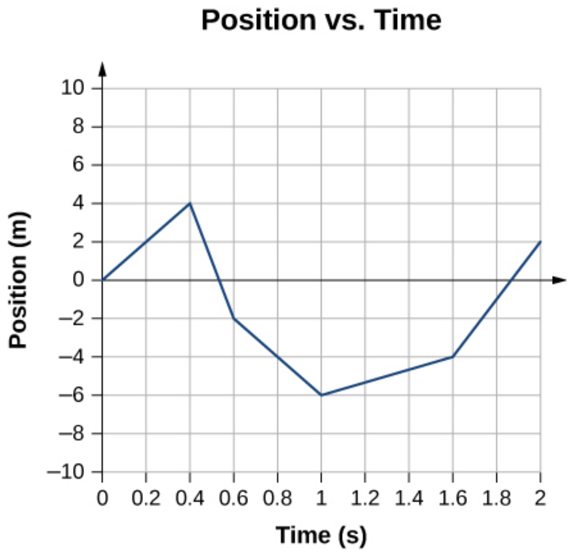
28. A cyclist rides 8.0 km east for 20 minutes, then he turns and heads west for 8 minutes and 3.2 km. Finally, he rides east for 16 km, which takes 40 minutes. (a) What is the final displacement of the cyclist? (b) What is his average velocity?

29. On February 15, 2013, a superbolide meteor (brighter than the Sun) entered Earth's atmosphere over Chelyabinsk, Russia, and exploded at an altitude of 23.5 km. Eyewitnesses could feel the intense heat from the fireball, and the blast wave from the explosion blew out windows in buildings. The blast wave took approximately 2 minutes 30 seconds to reach ground level. (a) What was the average velocity of the blast wave? b) Compare this with the speed of sound, which is 343 m/s at sea level.

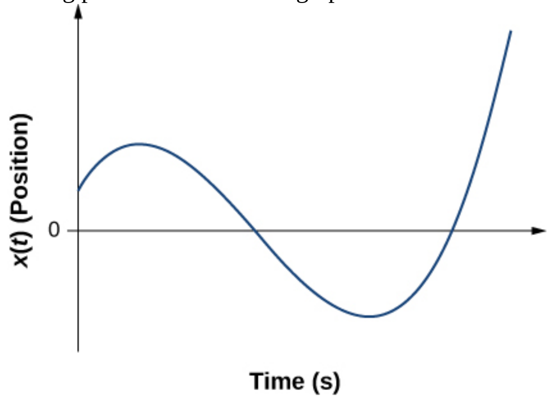
3.2 Instantaneous Velocity and Speed

30. A woodchuck runs 20 m to the right in 5 s, then turns and runs 10 m to the left in 3 s. (a) What is the average velocity of the woodchuck? (b) What is its average speed?

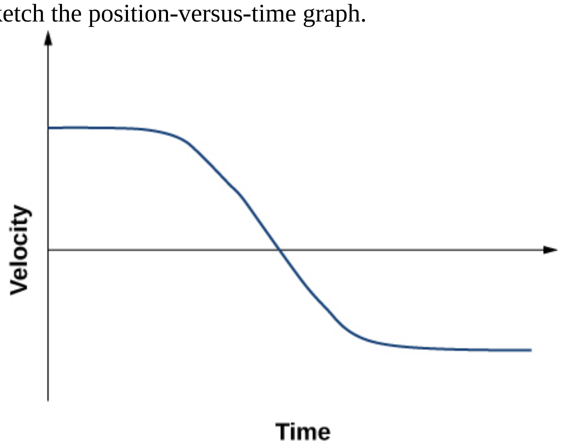
31. Sketch the velocity-versus-time graph from the following position-versus-time graph.



32. Sketch the velocity-versus-time graph from the following position-versus-time graph.



33. Given the following velocity-versus-time graph, sketch the position-versus-time graph.

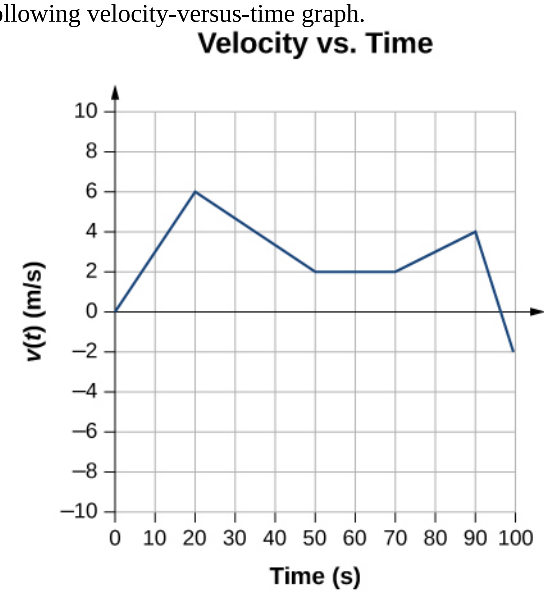


3.3 Average and Instantaneous Acceleration

34. A cheetah can accelerate from rest to a speed of 30.0 m/s in 7.00 s. What is its acceleration?

35. Dr. John Paul Stapp was a U.S. Air Force officer who studied the effects of extreme acceleration on the human body. On December 10, 1954, Stapp rode a rocket sled, accelerating from rest to a top speed of 282 m/s (1015 km/h) in 5.00 s and was brought jarringly back to rest in only 1.40 s. Calculate his (a) acceleration in his direction of motion and (b) acceleration opposite to his direction of motion. Express each in multiples of *g* (9.80 m/s^2) by taking its ratio to the acceleration of gravity.

36. Sketch the acceleration-versus-time graph from the following velocity-versus-time graph.



37. A commuter backs her car out of her garage with an acceleration of 1.40 m/s^2. (a) How long does it take her to reach a speed of 2.00 m/s? (b) If she then brakes to a stop in 0.800 s, what is her acceleration?

38. Assume an intercontinental ballistic missile goes from rest to a suborbital speed of 6.50 km/s in 60.0 s (the actual speed and time are classified). What is its average acceleration in meters per second and in multiples of *g* (9.80 m/s^2)?

39. An airplane, starting from rest, moves down the runway at constant acceleration for 18 s and then takes off at a speed of 60 m/s. What is the average acceleration of the plane?

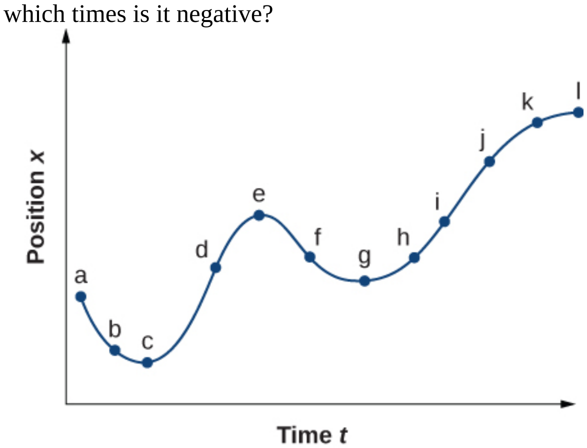
3.4 Motion with Constant Acceleration

40. A particle moves in a straight line at a constant velocity of 30 m/s. What is its displacement between t = 0 and t = 5.0 s?

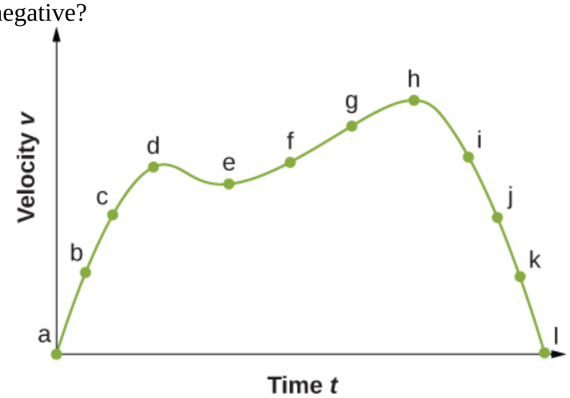
41. A particle moves in a straight line with an initial velocity of 0 m/s and a constant acceleration of 30 m/s^2. If $t = 0 \text{at} x = 0$, what is the particle's position at *t* = 5 s?

42. A particle moves in a straight line with an initial velocity of 30 m/s and constant acceleration 30 m/s^2. (a) What is its displacement at t = 5 s? (b) What is its velocity at this same time?

43. (a) Sketch a graph of velocity versus time corresponding to the graph of displacement versus time given in the following figure. (b) Identify the time or times (t_a, t_b, t_c, etc.) at which the instantaneous velocity has the greatest positive value. (c) At which times is it zero? (d) At which times is it negative?



44. (a) Sketch a graph of acceleration versus time corresponding to the graph of velocity versus time given in the following figure. (b) Identify the time or times (t_a, t_b, t_c, etc.) at which the acceleration has the greatest positive value. (c) At which times is it zero? (d) At which times is it negative?



45. A particle has a constant acceleration of 6.0 m/s^2. (a) If its initial velocity is 2.0 m/s, at what time is its displacement 5.0 m? (b) What is its velocity at that time?

46. At t = 10 s, a particle is moving from left to right with a speed of 5.0 m/s. At t = 20 s, the particle is moving right to left with a speed of 8.0 m/s. Assuming the particle's acceleration is constant, determine (a) its acceleration, (b) its initial velocity, and (c) the instant when its velocity is zero.

47. A well-thrown ball is caught in a well-padded mitt. If the acceleration of the ball is $2.10 \times 10^4 \text{ m/s}^2$, and 1.85 ms $(1 \text{ ms} = 10^{-3} \text{ s})$ elapses from the time the ball first touches the mitt until it stops, what is the initial velocity of the ball?

48. A bullet in a gun is accelerated from the firing chamber to the end of the barrel at an average rate of $6.20 \times 10^5 \text{ m/s}^2$ for $8.10 \times 10^{-4} \text{ s}$. What is its muzzle velocity (that is, its final velocity)?

49. (a) A light-rail commuter train accelerates at a rate of 1.35 m/s^2. How long does it take to reach its top speed of 80.0 km/h, starting from rest? (b) The same train ordinarily decelerates at a rate of 1.65 m/s^2. How long does it take to come to a stop from its top speed? (c) In emergencies, the train can decelerate more rapidly, coming to rest from 80.0 km/h in 8.30 s. What is its emergency acceleration in meters per second squared?

50. While entering a freeway, a car accelerates from rest at a rate of 2.40 m/s^2 for 12.0 s. (a) Draw a sketch of the situation. (b) List the knowns in this problem. (c) How far does the car travel in those 12.0 s? To solve this part, first identify the unknown, then indicate how you chose the appropriate equation to solve for it. After choosing the equation, show your steps in solving for the unknown, check your units, and discuss whether the answer is reasonable. (d) What is the car's final velocity? Solve for this unknown in the same manner as in (c), showing all steps explicitly.

51. Unreasonable results At the end of a race, a runner decelerates from a velocity of 9.00 m/s at a rate of 2.00 m/s^2. (a) How far does she travel in the next 5.00 s? (b) What is her final velocity? (c) Evaluate the result. Does it make sense?

52. Blood is accelerated from rest to 30.0 cm/s in a distance of 1.80 cm by the left ventricle of the heart. (a) Make a sketch of the situation. (b) List the knowns in this problem. (c) How long does the acceleration take? To solve this part, first identify the unknown, then discuss how you chose the appropriate equation to solve for it. After choosing the equation, show your steps in solving for the unknown, checking your units. (d) Is the answer reasonable when compared with the time for a heartbeat?

53. During a slap shot, a hockey player accelerates the puck from a velocity of 8.00 m/s to 40.0 m/s in the same direction. If this shot takes $3.33 \times 10^{-2} \text{ s}$, what is the distance over which the puck accelerates?

54. A powerful motorcycle can accelerate from rest to 26.8 m/s (100 km/h) in only 3.90 s. (a) What is its average acceleration? (b) How far does it travel in that time?

55. Freight trains can produce only relatively small accelerations. (a) What is the final velocity of a freight train that accelerates at a rate of $0.0500\,\text{m/s}^2$ for 8.00 min, starting with an initial velocity of 4.00 m/s? (b) If the train can slow down at a rate of $0.550\,\text{m/s}^2$, how long will it take to come to a stop from this velocity? (c) How far will it travel in each case?

56. A fireworks shell is accelerated from rest to a velocity of 65.0 m/s over a distance of 0.250 m. (a) Calculate the acceleration. (b) How long did the acceleration last?

57. A swan on a lake gets airborne by flapping its wings and running on top of the water. (a) If the swan must reach a velocity of 6.00 m/s to take off and it accelerates from rest at an average rate of $0.35\,\text{m/s}^2$, how far will it travel before becoming airborne? (b) How long does this take?

58. A woodpecker's brain is specially protected from large accelerations by tendon-like attachments inside the skull. While pecking on a tree, the woodpecker's head comes to a stop from an initial velocity of 0.600 m/s in a distance of only 2.00 mm. (a) Find the acceleration in meters per second squared and in multiples of *g*, where *g* = 9.80 m/s^2. (b) Calculate the stopping time. (c) The tendons cradling the brain stretch, making its stopping distance 4.50 mm (greater than the head and, hence, less acceleration of the brain). What is the brain's acceleration, expressed in multiples of *g*?

59. An unwary football player collides with a padded goalpost while running at a velocity of 7.50 m/s and comes to a full stop after compressing the padding and his body 0.350 m. (a) What is his acceleration? (b) How long does the collision last?

60. A care package is dropped out of a cargo plane and lands in the forest. If we assume the care package speed on impact is 54 m/s (123 mph), then what is its acceleration? Assume the trees and snow stops it over a distance of 3.0 m.

61. An express train passes through a station. It enters with an initial velocity of 22.0 m/s and decelerates at a rate of $0.150\,\text{m/s}^2$ as it goes through. The station is 210.0 m long. (a) How fast is it going when the nose leaves the station? (b) How long is the nose of the train in the station? (c) If the train is 130 m long, what is the velocity of the end of the train as it leaves? (d) When does the end of the train leave the station?

62. Unreasonable results Dragsters can actually reach a top speed of 145.0 m/s in only 4.45 s. (a) Calculate the average acceleration for such a dragster. (b) Find the final velocity of this dragster starting from rest and accelerating at the rate found in (a) for 402.0 m (a quarter mile) without using any information on time. (c) Why is the final velocity greater than that used to find the average acceleration? (*Hint*: Consider whether the assumption of constant acceleration is valid for a dragster. If not, discuss whether the acceleration would be greater at the beginning or end of the run and what effect that would have on the final velocity.)

3.5 Free Fall

63. Calculate the displacement and velocity at times of (a) 0.500 s, (b) 1.00 s, (c) 1.50 s, and (d) 2.00 s for a ball thrown straight up with an initial velocity of 15.0 m/s. Take the point of release to be $y_0 = 0$.

64. Calculate the displacement and velocity at times of (a) 0.500 s, (b) 1.00 s, (c) 1.50 s, (d) 2.00 s, and (e) 2.50 s for a rock thrown straight down with an initial velocity of 14.0 m/s from the Verrazano Narrows Bridge in New York City. The roadway of this bridge is 70.0 m above the water.

65. A basketball referee tosses the ball straight up for the starting tip-off. At what velocity must a basketball player leave the ground to rise 1.25 m above the floor in an attempt to get the ball?

66. A rescue helicopter is hovering over a person whose boat has sunk. One of the rescuers throws a life preserver straight down to the victim with an initial velocity of 1.40 m/s and observes that it takes 1.8 s to reach the water. (a) List the knowns in this problem. (b) How high above the water was the preserver released? Note that the downdraft of the helicopter reduces the effects of air resistance on the falling life preserver, so that an acceleration equal to that of gravity is reasonable.

67. Unreasonable results A dolphin in an aquatic show jumps straight up out of the water at a velocity of 15.0 m/s. (a) List the knowns in this problem. (b) How high does his body rise above the water? To solve this part, first note that the final velocity is now a known, and identify its value. Then, identify the unknown and discuss how you chose the appropriate equation to solve for it. After choosing the equation, show your steps in solving for the unknown, checking units, and discuss whether the answer is reasonable. (c) How long a time is the dolphin in the air? Neglect any effects resulting from his size or orientation.

68. A diver bounces straight up from a diving board, avoiding the diving board on the way down, and falls feet first into a pool. She starts with a velocity of 4.00 m/s and her takeoff point is 1.80 m above the pool. (a) What is her highest point above the board? (b) How long a time are her feet in the air? (c) What is her velocity when her feet hit the water?

69. (a) Calculate the height of a cliff if it takes 2.35 s for a rock to hit the ground when it is thrown straight up from the cliff with an initial velocity of 8.00 m/s. (b) How long a time would it take to reach the ground if it is thrown straight down with the same speed?

70. A very strong, but inept, shot putter puts the shot straight up vertically with an initial velocity of 11.0 m/s. How long a time does he have to get out of the way if the shot was released at a height of 2.20 m and he is 1.80 m tall?

71. You throw a ball straight up with an initial velocity of 15.0 m/s. It passes a tree branch on the way up at a height of 7.0 m. How much additional time elapses before the ball passes the tree branch on the way back down?

72. A kangaroo can jump over an object 2.50 m high. (a) Considering just its vertical motion, calculate its vertical speed when it leaves the ground. (b) How long a time is it in the air?

73. Standing at the base of one of the cliffs of Mt. Arapiles in Victoria, Australia, a hiker hears a rock break loose from a height of 105.0 m. He can't see the rock right away, but then does, 1.50 s later. (a) How far above the hiker is the rock when he can see it? (b) How much time does he have to move before the rock hits his head?

74. There is a 250-m-high cliff at Half Dome in Yosemite National Park in California. Suppose a boulder breaks loose from the top of this cliff. (a) How fast will it be going when it strikes the ground? (b) Assuming a reaction time of 0.300 s, how long a time will a tourist at the bottom have to get out of the way after hearing the sound of the rock breaking loose (neglecting the height of the tourist, which would become negligible anyway if hit)? The speed of sound is 335.0 m/s on this day.

ADDITIONAL PROBLEMS

75. Professional baseball player Nolan Ryan could pitch a baseball at approximately 160.0 km/h. At that average velocity, how long did it take a ball thrown by Ryan to reach home plate, which is 18.4 m from the pitcher's mound? Compare this with the average reaction time of a human to a visual stimulus, which is 0.25 s.

76. An airplane leaves Chicago and makes the 3000-km trip to Los Angeles in 5.0 h. A second plane leaves Chicago one-half hour later and arrives in Los Angeles at the same time. Compare the average velocities of the two planes. Ignore the curvature of Earth and the difference in altitude between the two cities.

77. **Unreasonable Results** A cyclist rides 16.0 km east, then 8.0 km west, then 8.0 km east, then 32.0 km west, and finally 11.2 km east. If his average velocity is 24 km/h, how long did it take him to complete the trip? Is this a reasonable time?

78. An object has an acceleration of $+1.2\,\text{cm/s}^2$. At $t = 4.0\,\text{s}$, its velocity is $-3.4\,\text{cm/s}$. Determine the object's velocities at $t = 1.0\,\text{s}$ and $t = 6.0\,\text{s}$.

79. A particle moving at constant acceleration has velocities of $2.0\,\text{m/s}$ at $t = 2.0$ s and $-7.6\,\text{m/s}$ at $t = 5.2$ s. What is the acceleration of the particle?

80. A train is moving up a steep grade at constant velocity (see following figure) when its caboose breaks loose and starts rolling freely along the track. After 5.0 s, the caboose is 30 m behind the train. What is the acceleration of the caboose?

Figure shows a train moving up a hill.

81. An electron is moving in a straight line with a velocity of 4.0×10^5 m/s. It enters a region 5.0 cm long where it undergoes an acceleration of $6.0 \times 10^{12}\ \text{m/s}^2$ along the same straight line. (a) What is the electron's velocity when it emerges from this region? b) How long does the electron take to cross the region?

82. An ambulance driver is rushing a patient to the hospital. While traveling at 72 km/h, she notices the traffic light at the upcoming intersections has turned amber. To reach the intersection before the light turns red, she must travel 50 m in 2.0 s. (a) What minimum acceleration must the ambulance have to reach the intersection before the light turns red? (b) What is the speed of the ambulance when it reaches the intersection?

83. A motorcycle that is slowing down uniformly covers 2.0 successive km in 80 s and 120 s, respectively. Calculate (a) the acceleration of the motorcycle and (b) its velocity at the beginning and end of the 2-km trip.

84. A cyclist travels from point A to point B in 10 min. During the first 2.0 min of her trip, she maintains a uniform acceleration of $0.090\,\text{m/s}^2$. She then travels at constant velocity for the next 5.0 min. Next, she decelerates at a constant rate so that she comes to a rest at point B 3.0 min later. (a) Sketch the velocity-versus-time graph for the trip. (b) What is the acceleration during the last 3 min? (c) How far does the cyclist travel?

85. Two trains are moving at 30 m/s in opposite directions on the same track. The engineers see simultaneously that they are on a collision course and apply the brakes when they are 1000 m apart. Assuming both trains have the same acceleration, what must this acceleration be if the trains are to stop just short of colliding?

86. A 10.0-m-long truck moving with a constant velocity of 97.0 km/h passes a 3.0-m-long car moving with a constant velocity of 80.0 km/h. How much time elapses between the moment the front of the truck is even with the back of the car and the moment the back of the truck is even with the front of the car?
Top drawing shows passenger car with a speed of 80 kilometers per hour in front of the truck with the speed of 97 kilometers per hour. Middle drawing shows passenger car with a speed of 80 kilometers per hour parallel to the truck with the speed of 97 kilometers per hour. Bottom drawing shows passenger car with a speed of 80 kilometers per hour behind the truck with a speed of 97 kilometers per hour.

87. A police car waits in hiding slightly off the highway. A speeding car is spotted by the police car doing 40 m/s. At the instant the speeding car passes the police car, the police car accelerates from rest at 4 m/s^2 to catch the speeding car. How long does it take the police car to catch the speeding car?

88. Pablo is running in a half marathon at a velocity of 3 m/s. Another runner, Jacob, is 50 meters behind Pablo with the same velocity. Jacob begins to accelerate at 0.05 m/s^2. (a) How long does it take Jacob to catch Pablo? (b) What is the distance covered by Jacob? (c) What is the final velocity of the Jacob?

89. Unreasonable results A runner approaches the finish line and is 75 m away; her average speed at this position is 8 m/s. She decelerates at this point at 0.5 m/s^2. How long does it take her to cross the finish line from 75 m away? Is this reasonable?

90. An airplane accelerates at 5.0 m/s^2 for 30.0 s. During this time, it covers a distance of 10.0 km. What are the initial and final velocities of the airplane?

91. Compare the distance traveled of an object that undergoes a change in velocity that is twice its initial velocity with an object that changes its velocity by four times its initial velocity over the same time period. The accelerations of both objects are constant.

92. An object is moving east with a constant velocity and is at position x_0 at time $t_0 = 0$. (a) With what acceleration must the object have for its total displacement to be zero at a later time *t* ? (b) What is the physical interpretation of the solution in the case for $t \to \infty$?

93. A ball is thrown straight up. It passes a 2.00-m-high window 7.50 m off the ground on its path up and takes 1.30 s to go past the window. What was the ball's initial velocity?

94. A coin is dropped from a hot-air balloon that is 300 m above the ground and rising at 10.0 m/s upward. For the coin, find (a) the maximum height reached, (b) its position and velocity 4.00 s after being released, and (c) the time before it hits the ground.

95. A soft tennis ball is dropped onto a hard floor from a height of 1.50 m and rebounds to a height of 1.10 m. (a) Calculate its velocity just before it strikes the floor. (b) Calculate its velocity just after it leaves the floor on its way back up. (c) Calculate its acceleration during contact with the floor if that contact lasts 3.50 ms $(3.50 \times 10^{-3}\ \text{s})$ (d) How much did the ball compress during its collision with the floor, assuming the floor is absolutely rigid?

96. Unreasonable results. A raindrop falls from a cloud 100 m above the ground. Neglect air resistance. What is the speed of the raindrop when it hits the ground? Is this a reasonable number?

97. Compare the time in the air of a basketball player who jumps 1.0 m vertically off the floor with that of a player who jumps 0.3 m vertically.

98. Suppose that a person takes 0.5 s to react and move his hand to catch an object he has dropped. (a) How far does the object fall on Earth, where $g = 9.8\ \text{m/s}^2$? (b) How far does the object fall on the Moon, where the acceleration due to gravity is 1/6 of that on Earth?

99. A hot-air balloon rises from ground level at a constant velocity of 3.0 m/s. One minute after liftoff, a sandbag is dropped accidentally from the balloon. Calculate (a) the time it takes for the sandbag to reach the ground and (b) the velocity of the sandbag when it hits the ground.

100. (a) A world record was set for the men's 100-m dash in the 2008 Olympic Games in Beijing by Usain Bolt of Jamaica. Bolt "coasted" across the finish line with a time of 9.69 s. If we assume that Bolt accelerated for 3.00 s to reach his maximum speed, and maintained that speed for the rest of the race, calculate his maximum speed and his acceleration. (b) During the same Olympics, Bolt also set the world record in the 200-m dash with a time of 19.30 s. Using the same assumptions as for the 100-m dash, what was his maximum speed for this race?

101. An object is dropped from a height of 75.0 m above ground level. (a) Determine the distance traveled during the first second. (b) Determine the final velocity at which the object hits the ground. (c) Determine the distance traveled during the last second of motion before hitting the ground.

102. A steel ball is dropped onto a hard floor from a height of 1.50 m and rebounds to a height of 1.45 m. (a) Calculate its velocity just before it strikes the floor. (b) Calculate its velocity just after it leaves the floor on its way back up. (c) Calculate its acceleration during contact with the floor if that contact lasts 0.0800 ms $(8.00 \times 10^{-5}\ \text{s})$ (d) How much did the ball compress during its collision with the floor, assuming the floor is absolutely rigid?

103. An object is dropped from a roof of a building of height *h*. During the last second of its descent, it drops a distance *h*/3. Calculate the height of the building.

CHALLENGE PROBLEMS

104. In a 100-m race, the winner is timed at 11.2 s. The second-place finisher's time is 11.6 s. How far is the second-place finisher behind the winner when she crosses the finish line? Assume the velocity of each runner is constant throughout the race.

105. A cyclist sprints at the end of a race to clinch a victory. She has an initial velocity of 11.5 m/s and accelerates at a rate of 0.500 m/s^2 for 7.00 s. (a) What is her final velocity? (b) The cyclist continues at this velocity to the finish line. If she is 300 m from the finish line when she starts to accelerate, how much time did she save? (c) The second-place winner was 5.00 m ahead when the winner started to accelerate, but he was unable to accelerate, and traveled at 11.8 m/s until the finish line. What was the difference in finish time in seconds between the winner and runner-up? How far back was the runner-up when the winner crossed the finish line?

106. In 1967, New Zealander Burt Munro set the world record for an Indian motorcycle, on the Bonneville Salt Flats in Utah, of 295.38 km/h. The one-way course was 8.00 km long. Acceleration rates are often described by the time it takes to reach 96.0 km/h from rest. If this time was 4.00 s and Burt accelerated at this rate until he reached his maximum speed, how long did it take Burt to complete the course?

4 | CIRCULAR MOTION AS ONE-DIMENSIONAL MOTION

Figure 4.1 A person moving on a ferris wheel undergoes circular motion. Such motion can also be considered one-dimensional in a mathematical sense.(credit:Tiia Monto [CC BY-SA 3.0 (https://creativecommons.org/licenses/by-sa/3.0)])

Chapter Outline

Introduction

In **Motion Along a Straight Line**, we described motion (kinematics) in one dimension, and introduced important concepts like displacement, velocity and acceleration. The type of motion we considered there is called **translational motion**, because the moving object undergoes a translation from place to place in space. However, we know from everyday life that **rotational motion** is also very important and that many objects that move have both translation and rotation. The wind turbines in our chapter opening image are a prime example of how rotational motion impacts our daily lives, as the market for clean energy sources continues to grow.

As we shall discover in **Section 7.4**, the objects that we study in astronomy often move (translate) in paths that follow the shapes of conic sections (ellipses, parabolas, hyperbolas or circles). Those paths are the mathematical solutions to the equations of motion which arise in the study of objects subjected to the force of gravity. A circular path is the simplest, special case of this kind of motion.

Before considering rotating objects, we will begin by examining the idealization of a single point object moving in a path that is shaped as a perfect circle. Of course, such an object is actually translating in a two dimensional manner. But, as we will see, a mathematical analogy will allow us to describe that object using the kinematic equations of one dimensional motion.

Figure 4.2 shows the orbital paths of some objects in our solar system. The motions of the planets around the Sun, although they actually move along elliptical paths, can (at least initially) be approximated as being circular motion.

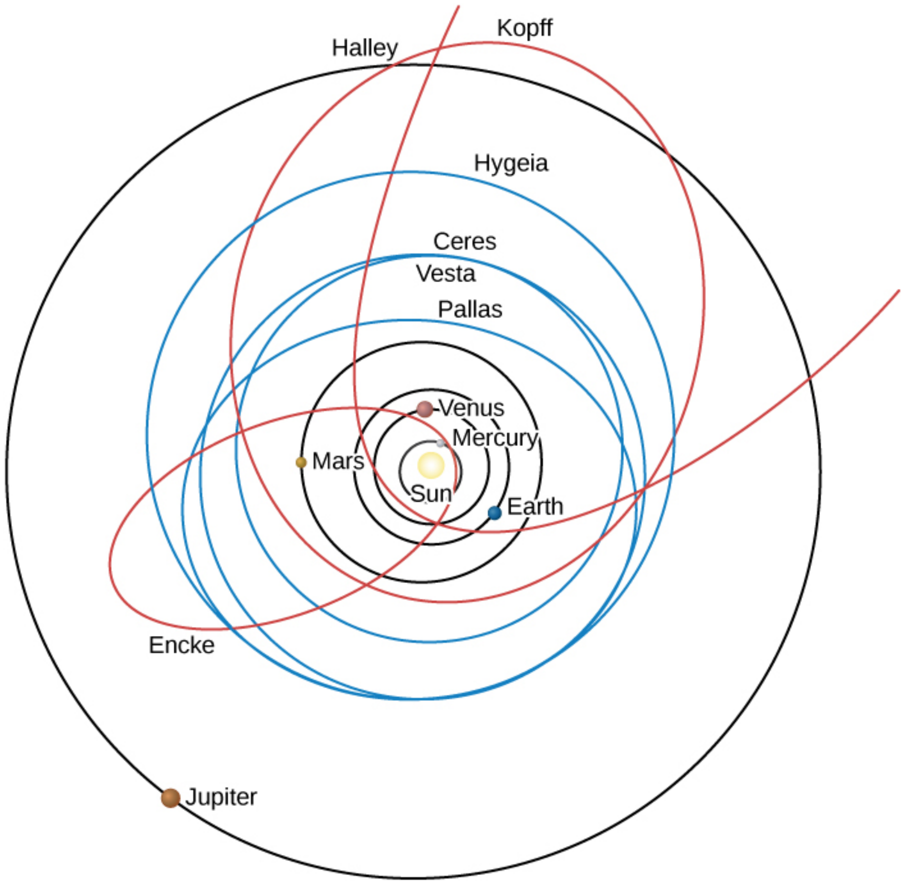


Figure 4.2 We see the orbits of typical comets and asteroids compared with those of the planets Mercury, Venus, Earth, Mars, and Jupiter (shown in black). Shown in red are three comets: Halley, Kopff, and Encke. In blue are the four largest asteroids: Ceres, Pallas, Vesta, and Hygeia. While the orbits of the comets and asteroids are obviously elliptical, those of the planets can be well approximated as circular.

Finally, to conclude the chapter, we will address the fixed-axis rotation of an extended object. Fixed-axis rotation describes the rotation around a fixed axis of a rigid body; that is, an object that does not deform as it moves. We will show how to apply all of the ideas we've developed up to this point about rotational and translational motion to such an object rotating around a fixed axis.

4.1 | Rotational Variables

Learning Objectives

By the end of this section, you will be able to:

- Describe the physical meaning of rotational variables as applied to fixed-axis rotation
- Explain how angular velocity is related to tangential speed
- Calculate the instantaneous angular velocity given the angular position function
- Find the angular velocity and angular acceleration in a rotating system
- Calculate the average angular acceleration when the angular velocity is changing
- Calculate the instantaneous angular acceleration given the angular velocity function

So far in this text, we have only studied translational motion, including the variables that describe it: displacement, velocity, and acceleration. Now we expand our description of motion to rotation—specifically, rotational motion about a fixed axis. We will find that rotational motion is described by a set of related variables similar to those we used in translational motion.

Angular Velocity

Circular motion is, simply, motion in a perfectly circular path. Although this is the simplest case of rotational motion, it is very useful for many situations, and we use it here to introduce rotational variables.

In **Figure 4.3**, we show (in red) a particle moving in a circle. The coordinate system is fixed and serves as a frame of reference to define the particle's position. Its position from the origin of the circle to the particle sweeps out the angle θ, which increases in the counterclockwise direction as the particle moves along its circular path. The angle θ is called the **angular position** of the particle. As the particle moves in its circular path, it also traces an arc length *s*.

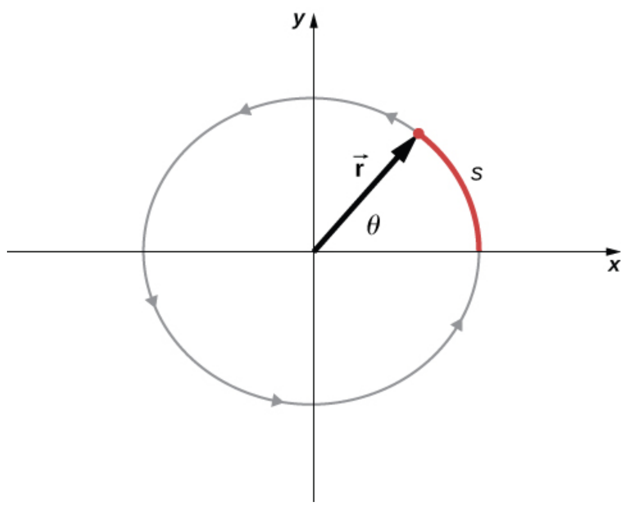


Figure 4.3 A particle follows a circular path. As it moves counterclockwise, it sweeps out a positive angle θ with respect to the *x*-axis and traces out an arc length *s*.

The angle is related to the radius of the circle and the arc length by

$$\theta = \frac{s}{r}. \quad (4.1)$$

The angle θ, the angular position of the particle along its path, has units of radians (rad). There are 2π radians in $360°$. Note that the radian measure is a ratio of length measurements, and therefore is a dimensionless quantity. As the particle moves along its circular path, its angular position changes and it undergoes angular displacements $\Delta\theta$.

Note again that the angle θ is measured counterclockwise starting from the positive *x* axis.

Now, let's note an interesting bit of mathematical simplification. Although this particle is moving in two dimensions, *x* and *y*, we know that those coordinates can be written as:

$$x = r\cos\theta \tag{4.2}$$

$$y = r\sin\theta \tag{4.3}$$

However, the fact that the particle is constrained to move along a circular path, where *r* remains constant, means that its only degree of motional freedom is the angle θ. So, in a mathematical sense, it is only moving in one dimension, and that dimension is represented by the coordinate θ.

We can thus describe its location completely by simply stating the value of the coordinate θ, its angular position. Its motion will be completely described by determining its angular displacements, $\Delta\theta$, as time goes on. This means that we have a system that is mathematically analogous to the one studied in **Motion Along a Straight Line**, the only difference being that, there, the dimension was called *x* (or *y*). Here, the dimension is called θ.

The magnitude of the **angular velocity**, denoted by ω, is the time rate of change of the angle θ as the particle moves in its circular path. The **instantaneous angular velocity** is defined as the limit in which $\Delta t \to 0$ in the average angular velocity $\bar{\omega} = \frac{\Delta\theta}{\Delta t}$:

$$\omega = \lim_{\Delta t \to 0} \frac{\Delta\theta}{\Delta t} = \frac{d\theta}{dt}, \tag{4.4}$$

where θ is the angle of rotation (**Figure 4.3**). The units of angular velocity are radians per second (rad/s). Angular velocity can also be referred to as the rotation rate in radians per second. In many situations, we are given the rotation rate in revolutions/s or cycles/s. To find the angular velocity, we must multiply revolutions/s by 2π, since there are 2π radians in one complete revolution. Since the direction of a positive angle in a circle is counterclockwise, we take counterclockwise rotations as being positive and clockwise rotations as negative. (This choice is analogous to the one we made for horizontal, translational motion. In that situation, moving from left to right was defined as moving in the positive motion, but moving from right to left was in the negative direction.)

Connections to a Rotating Rigid Object

All of the arguments we have just made (and the mathematics we have developed) apply perfectly to the description of the motion of any single point on an object that happens to be rotating about some fixed axis. Take a look at the rotating disk in **Figure 4.4**. It shows two particles, 1 and 2, that are located at different distances (r_1 and r_2) from the axis of rotation.

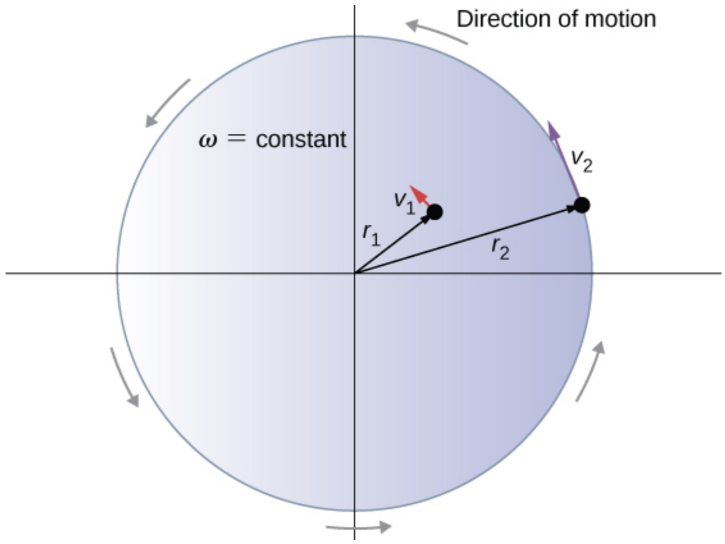


Figure 4.4 Two particles on a rotating disk

Which properties of motion are the same for these two points, and which are different? If you've ever ridden on a merry-go-round, you know that the speed, v_2, of a person located at point r_2 will be greater than the speed, v_1, of a person located at r_1. Your experience tells you that, the farther you are from the axis of the rotating object, the faster you are moving. We call v_1 and v_2 the **tangential speeds**, because they are moving instantaneously in a direction that is tangent to the circle of their motion.

But, how do the angular speeds of the two particles (or people) compare? Since it will take each one exactly the same amount of time to complete one rotation, their angular speed, ω, must be the same.

We can see how angular velocity is related to the tangential speed of the particle by going back to the definition of the tangential displacement or arc length

$$s = r\theta.$$

Noting that the radius r is a constant, we have

$$v_t = \frac{\Delta s}{\Delta t} = \frac{r\Delta\theta}{\Delta t} = r\frac{\Delta\theta}{\Delta t} = r\omega$$

So, simply put, for any point located at a distance r from the axis of rotation, the tangential speed

$$v_{\text{t}} = r\omega. \quad (4.5)$$

That is, the tangential speed of the particle is its angular velocity times the radius of the circle traced out by its motion. From **Equation 4.5**, we see that the tangential speed of the particle increases with its distance from the axis of rotation for a constant angular velocity. This effect is shown in **Figure 4.4**. So it is for our two particles placed at different radii on a rotating disk with a constant angular velocity. As the disk rotates, the tangential speed increases linearly with the radius from the axis of rotation. In **Figure 4.4**, we see that $v_1 = r_1\omega_1$ and $v_2 = r_2\omega_2$. But the disk has a constant angular velocity, so $\omega_1 = \omega_2$. This means $\frac{v_1}{r_1} = \frac{v_2}{r_2}$ or $v_2 = \left(\frac{r_2}{r_1}\right)v_1$. Thus, since $r_2 > r_1$, $v_2 > v_1$.

Example 4.1

Rotation of a Flywheel

A flywheel rotates such that it sweeps out an angle at the rate of $\theta = \omega t = (45.0\,\text{rad/s})t$ radians. The wheel rotates counterclockwise when viewed in the plane of the page. (a) What is the angular velocity of the flywheel? (b) How many radians does the flywheel rotate through in 30 s? (c) What is the tangential speed of a point on the flywheel 10 cm from the axis of rotation?

Strategy

The functional form of the angular position of the flywheel is given in the problem as $\theta(t) = \omega t$, so we can find the angular speed by inspection. It is just 45 rad/s. To find the angular displacement of the flywheel during 30 s, we seek the angular displacement $\Delta\theta$, where the change in angular position is between 0 and 30 s. To find the tangential speed of a point at a distance from the axis of rotation, we multiply its distance times the angular velocity of the flywheel.

Solution

a. $\omega = 45 \text{ rad/s}$. We see that the angular velocity is a constant.

b. $\Delta\theta = \theta(30 \text{ s}) - \theta(0 \text{ s}) = 45.0(30.0 \text{ s}) - 45.0(0 \text{ s}) = 1350.0 \text{ rad}$.

c. $v_\text{t} = r\omega = (0.1 \text{ m})(45.0 \text{ rad/s}) = 4.5 \text{ m/s}$.

Significance

In 30 s, the flywheel has rotated through quite a number of revolutions, about 215 if we divide the angular displacement by 2π. A massive flywheel can be used to store energy in this way, if the losses due to friction are minimal. Recent research has considered superconducting bearings on which the flywheel rests, with zero energy loss due to friction.

Angular Acceleration

We have just discussed angular velocity for circular motion, but not all circular motion is at a uniform velocity. Envision an ice skater spinning with his arms outstretched—when he pulls his arms inward, his angular velocity increases. Or think about a computer's hard disk slowing to a halt as the angular velocity decreases. We will explore these situations later, but we can already see a need to define an **angular acceleration** for describing situations where ω changes. The faster the change in ω, the greater the angular acceleration. We define the **instantaneous angular acceleration** α as the derivative of angular velocity with respect to time:

$$\alpha = \lim_{\Delta t \to 0} \frac{\Delta\omega}{\Delta t} = \frac{d\omega}{dt} = \frac{d^2\theta}{dt^2}, \tag{4.6}$$

where we have taken the limit of the average angular acceleration, $\bar{\alpha} = \frac{\Delta\omega}{\Delta t}$ as $\Delta t \to 0$.

The units of angular acceleration are (rad/s)/s, or rad/s^2.

We can relate the tangential acceleration of a point on a rotating body at a distance from the axis of rotation in the same way that we related the tangential speed to the angular velocity. Again noting that the radius *r* is constant, we obtain

$$a_t = \frac{\Delta a_t}{\Delta t} = \frac{r\Delta\omega}{\Delta t} = r\frac{\Delta\omega}{\Delta t} = r\alpha \tag{4.7}$$

Therefore, for any point located at a distance r from the axis of rotation, the tangential acceleration is the distance from the rotational axis times the angular acceleration.

$$a_\text{t} = r\alpha. \tag{4.8}$$

Let's apply these ideas to the analysis of a few simple fixed-axis rotation scenarios. Before doing so, we present a problem-solving strategy that can be applied to rotational kinematics: the description of rotational motion.

Problem-Solving Strategy: Rotational Kinematics

1. Examine the situation to determine that rotational kinematics (rotational motion) is involved.
2. Identify exactly what needs to be determined in the problem (identify the unknowns). A sketch of the situation is useful.
3. Make a complete list of what is given or can be inferred from the problem as stated (identify the knowns).
4. Solve the appropriate equation or equations for the quantity to be determined (the unknown). It can be useful to think in terms of a translational analog, because by now you are familiar with the equations of translational motion.
5. Substitute the known values along with their units into the appropriate equation and obtain numerical solutions complete with units. Be sure to use units of radians for angles.
6. Check your answer to see if it is reasonable: Does your answer make sense?

Now let's apply this problem-solving strategy to a few specific examples.

Example 4.2

A Spinning Bicycle Wheel

A bicycle mechanic mounts a bicycle on the repair stand and starts the rear wheel spinning from rest to a final angular velocity of 250 rpm in 5.00 s. (a) Calculate the average angular acceleration in rad/s^2. (b) If she now hits the brakes, causing an angular acceleration of $-87.3\ \text{rad/s}^2$, how long does it take the wheel to stop?

Strategy

The average angular acceleration can be found directly from its definition $\bar{\alpha} = \frac{\Delta\omega}{\Delta t}$ because the final angular velocity and time are given. We see that $\Delta\omega = \omega_{\text{final}} - \omega_{\text{initial}} = 250\ \text{rev/min}$ and Δt is 5.00 s. For part (b), we know the angular acceleration and the initial angular velocity. We can find the stopping time by using the definition of average angular acceleration and solving for Δt, yielding

$$\Delta t = \frac{\Delta\omega}{\alpha}.$$

Solution

a. Entering known information into the definition of angular acceleration, we get

$$\bar{\alpha} = \frac{\Delta\omega}{\Delta t} = \frac{250\ \text{rpm}}{5.00\ \text{s}}.$$

Because $\Delta\omega$ is in revolutions per minute (rpm) and we want the standard units of rad/s^2 for angular acceleration, we need to convert from rpm to rad/s:

$$\Delta\omega = 250\frac{\text{rev}}{\text{min}} \cdot \frac{2\pi\ \text{rad}}{\text{rev}} \cdot \frac{1\ \text{min}}{60\ \text{s}} = 26.2\frac{\text{rad}}{\text{s}}.$$

Entering this quantity into the expression for α, we get

$$\alpha = \frac{\Delta\omega}{\Delta t} = \frac{26.2\ \text{rad/s}}{5.00\ \text{s}} = 5.24\ \text{rad/s}^2.$$

b. Here the angular velocity decreases from 26.2 rad/s (250 rpm) to zero, so that $\Delta\omega$ is −26.2 rad/s, and α is given to be $-87.3\ \text{rad/s}^2$. Thus,

$$\Delta t = \frac{-26.2\ \text{rad/s}}{-87.3\ \text{rad/s}^2} = 0.300\ \text{s}.$$

Significance

Note that the angular acceleration as the mechanic spins the wheel is small and positive; it takes 5 s to produce an appreciable angular velocity. When she hits the brake, the angular acceleration is large and negative. The angular velocity quickly goes to zero.

4.1 Check Your Understanding The fan blades on a turbofan jet engine (shown below) accelerate from rest up to a rotation rate of 40.0 rev/s in 20 s. The increase in angular velocity of the fan is constant in time. (The GE90-110B1 turbofan engine mounted on a Boeing 777, as shown, is currently the largest turbofan engine in the world, capable of thrusts of 330–510 kN.)

(a) What is the average angular acceleration?

(b) What is the instantaneous angular acceleration at any time during the first 20 s?

Figure 4.5 (credit: “Bubinator”/ Wikimedia Commons)

We now have a basic vocabulary for discussing fixed-axis rotational kinematics and relationships between rotational variables. We discuss more definitions and connections in the next section.

4.2 | Rotation with Constant Angular Acceleration

Learning Objectives

By the end of this section, you will be able to:

- Select from the kinematic equations for rotational motion with constant angular acceleration the appropriate equations to solve for unknowns in the analysis of systems undergoing fixed-axis rotation

In the preceding section, we defined the rotational variables of angular displacement, angular velocity, and angular acceleration. In this section, we work with these definitions to derive relationships among these variables and use these relationships to analyze rotational motion for a rigid body about a fixed axis under a constant angular acceleration. (By the analogy previously established, these relationships also hold for any point-like object undergoing circular motion.) This analysis forms the basis for rotational kinematics. If the angular acceleration is constant, the equations of rotational kinematics simplify, similar to the equations of linear kinematics discussed in Section 3.4. We can then use this simplified set of equations to describe many applications in physics and engineering where the angular acceleration of the system is constant.

Kinematics of Rotational Motion

Using our intuition, we can begin to see how the rotational quantities θ, ω, α, and t are related to one another. For example, we saw in the preceding section that if a flywheel has an angular acceleration in the same direction as its angular velocity vector, its angular velocity increases with time and its angular displacement also increases. On the contrary, if the angular acceleration is opposite to the angular velocity vector, its angular velocity decreases with time. We can describe these physical situations and many others with a consistent set of rotational kinematic equations under a constant angular acceleration. The method to investigate rotational motion in this way is called **kinematics of rotational motion**.

To begin, we note that if the system is rotating under a constant acceleration, then the average angular velocity follows a simple relation because the angular velocity is increasing linearly with time. The average angular velocity is just half the sum of the initial and final values:

$$\bar{\omega} = \frac{\omega_0 + \omega_f}{2}. \quad (4.9)$$

From the definition of the average angular velocity, we can find an equation that relates the angular position, average angular velocity, and time:

$$\bar{\omega} = \frac{\Delta\theta}{\Delta t}.$$

Solving for θ, we have

$$\theta_f = \theta_0 + \bar{\omega}\, t, \quad (4.10)$$

where we have set $t_0 = 0$. This equation can be very useful if we know the average angular velocity of the system. Then we could find the angular displacement over a given time period. Next, we find an equation relating ω, α, and t. To determine this equation, we start with the definition of angular acceleration:

$$\alpha = \frac{\Delta\omega}{\Delta t}.$$

We rearrange this to get $\alpha\Delta t = \Delta\omega$. In uniform rotational motion, the angular acceleration is constant so this becomes:

$$\alpha(t - t_0) = \omega_f - \omega_0 \quad (4.11)$$

Setting $t_0 = 0$, we have

$$\alpha t = \omega_f - \omega_0.$$

We rearrange this to obtain

$$\omega_f = \omega_0 + \alpha t, \quad (4.12)$$

where ω_0 is the initial angular velocity. Equation 4.12 is the rotational counterpart to the linear kinematics equation $v_f = v_0 + at$. With Equation 4.12, we can find the angular velocity of an object at any specified time t given the initial angular velocity and the angular acceleration.

Again by analogy to the discussion in Section 3.4, where we have set the initial time $t_0 = 0$, we can obtain another equation:

$$\theta_f = \theta_0 + \omega_0 t + \frac{1}{2}\alpha t^2. \quad (4.13)$$

Equation 4.13 is the rotational counterpart to the linear kinematics equation found in **Section 3.4** for position as a function of time. This equation gives us the angular position of a rotating rigid body at any time *t* given the initial conditions (initial angular position and initial angular velocity) and the angular acceleration.

We can find an equation that is independent of time by solving for *t* in **Equation 4.12** and substituting into **Equation 4.13**. **Equation 4.13** becomes

$$\begin{aligned}
\theta_f &= \theta_0 + \omega_0\left(\frac{\omega_f - \omega_0}{\alpha}\right) + \frac{1}{2}\alpha\left(\frac{\omega_f - \omega_0}{\alpha}\right)^2 \\
&= \theta_0 + \frac{\omega_0\omega_f}{\alpha} - \frac{\omega_0^2}{\alpha} + \frac{1}{2}\frac{\omega_f^2}{\alpha} - \frac{\omega_0\omega_f}{\alpha} + \frac{1}{2}\frac{\omega_0^2}{\alpha} \\
&= \theta_0 + \frac{1}{2}\frac{\omega_f^2}{\alpha} - \frac{1}{2}\frac{\omega_0^2}{\alpha}, \\
\theta_f - \theta_0 &= \frac{\omega_f^2 - \omega_0^2}{2\alpha}
\end{aligned}$$

or

$$\omega_f^2 = \omega_0^2 + 2\alpha(\Delta\theta). \quad (4.14)$$

Equation 4.10 through **Equation 4.14** describe fixed-axis rotation for constant acceleration and are summarized in **Table 4.1**.

Angular displacement from average angular velocity	$\theta_f = \theta_0 + \bar{\omega} t$
Angular velocity from angular acceleration	$\omega_f = \omega_0 + \alpha t$
Angular displacement from angular velocity and angular acceleration	$\theta_f = \theta_0 + \omega_0 t + \frac{1}{2}\alpha t^2$
Angular velocity from angular displacement and angular acceleration	$\omega_f^2 = {\omega_0}^2 + 2\alpha(\Delta\theta)$

Table 4.1 Kinematic Equations for Motion with Constant Angular Acceleration

Applying the Equations for Rotational Motion

Now we can apply the key kinematic relations for rotational motion to some simple examples to get a feel for how the equations can be applied to everyday situations.

Example 4.3

Calculating the Acceleration of a Fishing Reel

A deep-sea fisherman hooks a big fish that swims away from the boat, pulling the fishing line from his fishing reel. The whole system is initially at rest, and the fishing line unwinds from the reel at a radius of 4.50 cm from its axis of rotation. The reel is given an angular acceleration of 110 rad/s^2 for 2.00 s (**Figure 4.6**).

(a) What is the final angular velocity of the reel after 2 s?

(b) How many revolutions does the reel make?

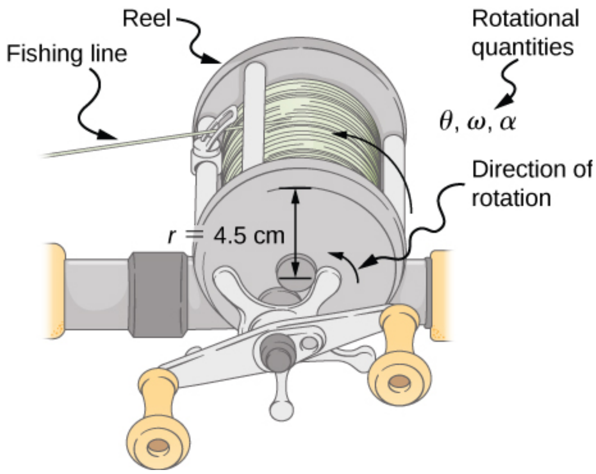


Figure 4.6 Fishing line coming off a rotating reel moves linearly.

Strategy

Identify the knowns and compare with the kinematic equations for constant acceleration. Look for the appropriate equation that can be solved for the unknown, using the knowns given in the problem description.

Solution

a. We are given α and t and want to determine ω. The most straightforward equation to use is $\omega_f = \omega_0 + \alpha t$, since all terms are known besides the unknown variable we are looking for. We are given that $\omega_0 = 0$ (it starts from rest), so

$$\omega_f = 0 + (110\,\text{rad/s}^2)(2.00\,\text{s}) = 220\,\text{rad/s}.$$

b. We are asked to find the number of revolutions. Because $1\,\text{rev} = 2\pi\,\text{rad}$, we can find the number of revolutions by finding θ in radians. We are given α and t, and we know ω_0 is zero, so we can obtain θ by using

$$\begin{aligned}\theta_f &= \theta_i + \omega_i t + \frac{1}{2}\alpha t^2 \\ &= 0 + 0 + (0.500)\left(110\,\text{rad/s}^2\right)(2.00\,\text{s})^2 = 220\,\text{rad}.\end{aligned}$$

Converting radians to revolutions gives

$$\text{Number of rev } = (220\,\text{rad})\frac{1\,\text{rev}}{2\pi\,\text{rad}} = 35.0\,\text{rev}.$$

Significance

This example illustrates that relationships among rotational quantities are highly analogous to those among linear quantities. The answers to the questions are realistic. After unwinding for two seconds, the reel is found to spin at 220 rad/s, which is 2100 rpm. (No wonder reels sometimes make high-pitched sounds.)

In the preceding example, we considered a fishing reel with a positive angular acceleration. Now let us consider what happens with a negative angular acceleration.

Example 4.4

Calculating the Duration When the Fishing Reel Slows Down and Stops

Now the fisherman applies a brake to the spinning reel, achieving an angular acceleration of $-300\,\text{rad/s}^2$. How long does it take the reel to come to a stop?

Strategy

We are asked to find the time *t* for the reel to come to a stop. The initial and final conditions are different from those in the previous problem, which involved the same fishing reel. Now we see that the initial angular velocity is $\omega_0 = 220\text{ rad/s}$ and the final angular velocity ω is zero. The angular acceleration is given as $\alpha = -300\text{ rad/s}^2$. Examining the available equations, we see all quantities but *t* are known in $\omega_f = \omega_0 + \alpha t$, making it easiest to use this equation.

Solution

The equation states

$$\omega_f = \omega_0 + \alpha t.$$

We solve the equation algebraically for *t* and then substitute the known values as usual, yielding

$$t = \frac{\omega_f - \omega_0}{\alpha} = \frac{0 - 220.0\text{ rad/s}}{-300.0\text{ rad/s}^2} = 0.733\text{ s.}$$

Significance

Note that care must be taken with the signs that indicate the directions of various quantities. Also, note that the time to stop the reel is fairly small because the acceleration is rather large. Fishing lines sometimes snap because of the accelerations involved, and fishermen often let the fish swim for a while before applying brakes on the reel. A tired fish is slower, requiring a smaller acceleration.

4.2 Check Your Understanding A centrifuge used in DNA extraction spins at a maximum rate of 7000 rpm, producing a “g-force” on the sample that is 6000 times the force of gravity. If the centrifuge takes 10 seconds to come to rest from the maximum spin rate: (a) What is the angular acceleration of the centrifuge? (b) What is the angular displacement of the centrifuge during this time?

Example 4.5

Angular Acceleration of a Propeller

Figure 4.7 shows a graph of the angular velocity of a propeller on an aircraft as a function of time. Its angular velocity starts at 30 rad/s and drops linearly to 0 rad/s over the course of 5 seconds. (a) Find the angular acceleration of the object. (b) Find the angle through which the propeller rotates during these 5 seconds.

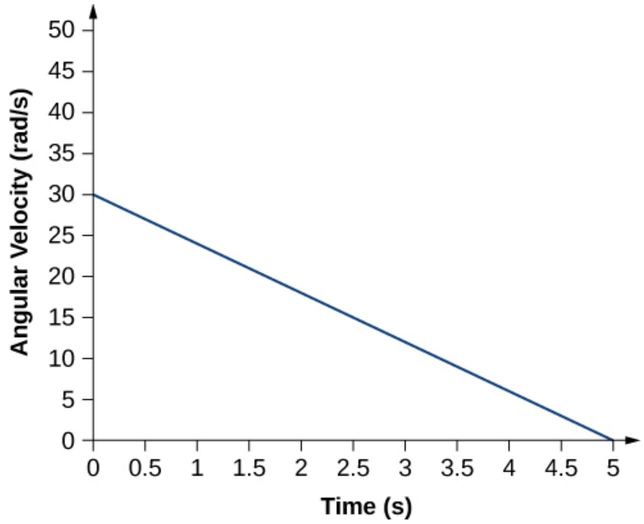


Figure 4.7 A graph of the angular velocity of a propeller versus time.

Strategy

a. Since the angular velocity varies linearly with time, we know that the angular acceleration is constant and does not depend on the time variable. The angular acceleration is the slope of the angular velocity vs. time graph, $\alpha = \frac{d\omega}{dt}$. To calculate the slope, we read directly from **Figure 4.7**, and see that $\omega_0 = 30\,\text{rad/s}$ at $t = 0\,\text{s}$ and $\omega_\text{f} = 0\,\text{rad/s}$ at $t = 5\,\text{s}$.

b. We use the equation $\omega_\text{f}^2 = \omega_0^2 + 2\alpha(\Delta\theta).$ to calculate the angular displacement, $\Delta\theta$.

Solution

a. Calculating the slope, we get

$$\alpha = \frac{\omega - \omega_0}{t - t_0} = \frac{(0 - 30.0)\,\text{rad/s}}{(5.0 - 0)\,\text{s}} = -6.0\,\text{rad/s}^2.$$

b. $$\theta_\text{f} = \theta_0 + \omega_0 t + \frac{1}{2}\alpha t^2.$$

Setting $\theta_0 = 0$, we have

$$\theta_0 = (30.0\,\text{rad/s})(5.0\,\text{s}) + \frac{1}{2}(-6.0\,\text{rad/s}^2)(5.0\,\text{rad/s})^2 = 150.0 - 75.0 = 75.0\,\text{rad}.$$

4.3 | Relating Angular and Translational Quantities

Learning Objectives

By the end of this section, you will be able to:

- Given the linear kinematic equation, write the corresponding rotational kinematic equation
- Calculate the linear distances, velocities, and accelerations of points on a rotating system given the angular velocities and accelerations

In this section, as we consider the motion of some point on a rotating object, we will relate each of the rotational variables to the translational variables defined in **Motion Along a Straight Line**. This will complete our ability to describe rigid-body rotations.

Angular vs. Linear Variables

In **Section 4.1**, we introduced angular kinematic variables. If we compare the angular definitions with the definitions of linear kinematic variables from **Motion Along a Straight Line**, we find that there is a mapping of the linear variables to the rotational ones. Linear position, velocity, and acceleration have their rotational counterparts, as we can see when we write them side by side:

	Linear	Rotational
Position	x	θ
Velocity	$v = \frac{dx}{dt}$	$\omega = \frac{d\theta}{dt}$
Acceleration	$a = \frac{dv}{dt}$	$\alpha = \frac{d\omega}{dt}$

Let's compare the linear and rotational variables individually. The linear variable of position has physical units of meters, whereas the angular position variable has dimensionless units of radians, as can be seen from the definition of $\theta = \frac{s}{r}$, which is the ratio of two lengths. The linear velocity has units of m/s, and its counterpart, the angular velocity, has units of rad/s. In **Section 4.1**, we saw in the case of circular motion that the linear tangential speed of a particle at a radius r from the axis of rotation is related to the angular velocity by the relation $v_t = r\omega$. This could also apply to points on a rigid body rotating about a fixed axis. Here, we consider only circular motion. (In circular motion, both uniform and nonuniform, there exists another acceleration, the centripetal acceleration, which will be discussed in **Section 6.4**.)

Relationships between Rotational and Translational Motion

We can look at two relationships between rotational and translational motion.

1. Generally speaking, the linear kinematic equations have their rotational counterparts. **Table 4.2** lists the four linear kinematic equations and the corresponding rotational counterpart. The two sets of equations look similar to each other, but describe two different physical situations, that is, rotation and translation.

Rotational	Translational
$\theta_f = \theta_0 + \bar{\omega}\, t$	$x = x_0 + \bar{v}\, t$
$\omega_f = \omega_0 + \alpha t$	$v_f = v_0 + at$
$\theta_f = \theta_0 + \omega_0 t + \frac{1}{2}\alpha t^2$	$x_f = x_0 + v_0 t + \frac{1}{2}at^2$
$\omega_f^2 = \omega_0^2 + 2\alpha(\Delta\theta)$	$v_f^2 = v_0^2 + 2a(\Delta x)$

Table 4.2 Rotational and Translational Kinematic Equations

2. The second correspondence has to do with relating linear and rotational variables in the special case of circular motion. Importantly, any object moving in a circular path possesses **both** an angular velocity and a tangential linear velocity. It possesses **both** an angular acceleration and a tangential linear acceleration. This is shown in **Table 4.3**, where in the third column, we have listed the connecting equation that relates the linear variable to the rotational variable for a rotating point located a distance r from its axis of rotation. The rotational variables of angular velocity and acceleration have subscripts that indicate their definition in circular motion.

Rotational	Translational	Relationship (r = radius)
θ	s	$\theta = \frac{s}{r}$
ω	v_t	$\omega = \frac{v_t}{r}$
α	a_t	$\alpha = \frac{a_t}{r}$

Table 4.3 Rotational and Translational Quantities for an Object in Circular Motion

Example 4.6

Linear Acceleration of a Centrifuge

A centrifuge has a radius of 20 cm and accelerates from a maximum rotation rate of 10,000 rpm to rest in 30 seconds under a constant angular acceleration. It is rotating counterclockwise. What is the magnitude of the tangential acceleration of a point at the tip of the centrifuge?

Strategy

With the information given, we can calculate the angular acceleration, which then will allow us to find the tangential acceleration.

Solution

The angular acceleration is

$$\alpha = \frac{\omega - \omega_0}{t} = \frac{0 - (1.0 \times 10^4)2\pi/60.0\,\text{s}(\text{rad/s})}{30.0\,\text{s}} = -34.9\,\text{rad/s}^2.$$

Therefore, the tangential acceleration is

$$a_t = r\alpha = 0.2\,\text{m}(-34.9\,\text{rad/s}^2) = -7.0\,\text{m/s}^2.$$

4.3 Check Your Understanding A boy jumps on a merry-go-round with a radius of 5 m that is at rest. It starts accelerating at a constant rate up to an angular velocity of 5 rad/s in 20 seconds. What is the distance travelled by the boy?

Check out this **PhET simulation (https://openstaxcollege.org/l/21rotatingdisk)** to change the parameters of a rotating disk (the initial angle, angular velocity, and angular acceleration), and place bugs at different radial distances from the axis. The simulation then lets you explore how circular motion relates to the bugs' *xy*-position, velocity, and acceleration using vectors or graphs.

CHAPTER 4 REVIEW

KEY TERMS

angular acceleration time rate of change of angular velocity

angular position angle a body has rotated through in a fixed coordinate system

angular velocity time rate of change of angular position

instantaneous angular acceleration derivative of angular velocity with respect to time

instantaneous angular velocity derivative of angular position with respect to time

kinematics of rotational motion describes the relationships among rotation angle, angular velocity, angular acceleration, and time

KEY EQUATIONS

Average Angular Velocity	$\bar{\omega} = \frac{\Delta\theta}{\Delta t} = \frac{\omega_0 + \omega_f}{2}.$
Angular displacement from average angular velocity	$\theta_f = \theta_0 + \bar{\omega} t$
Angular velocity from angular acceleration	$\omega_f = \omega_0 + \alpha t$
Angular displacement from angular velocity and angular acceleration	$\theta_f = \theta_0 + \omega_0 t + \frac{1}{2}\alpha t^2$
Angular velocity from angular displacement and angular acceleration	$\omega_f^2 = \omega_0{}^2 + 2\alpha(\Delta\theta)$
Arc length (displacement)	$s = r\theta$
Linear velocity from angular velocity	$v = r\omega$
Tangential acceleration from angular acceleration	$a_t = r\alpha$

SUMMARY

4.1 Rotational Variables

- The angular position θ of a rotating body is the angle the body has rotated through in a fixed coordinate system, which serves as a frame of reference.
- The angular velocity of a rotating body about a fixed axis is defined as $\omega(\text{rad/s})$, the rotational rate of the body in radians per second. The instantaneous angular velocity of a rotating body $\omega = \lim_{\Delta t \to 0} \frac{\Delta\omega}{\Delta t} = \frac{d\theta}{dt}$ is the derivative with respect to time of the angular position θ, found by taking the limit $\Delta t \to 0$ in the average angular velocity $\bar{\omega} = \frac{\Delta\theta}{\Delta t}$. The angular velocity relates v_t to the tangential speed of a point on the rotating body through the relation $v_t = r\omega$, where *r* is the radius to the point and v_t is the tangential speed at the given point.
- If the system's angular velocity is not constant, then the system has an angular acceleration. The average angular acceleration over a given time interval is the change in angular velocity over this time interval, $\bar{\alpha} = \frac{\Delta\omega}{\Delta t}$. The instantaneous angular acceleration is the time derivative of angular velocity, $\alpha = \lim_{\Delta t \to 0} \frac{\Delta\omega}{\Delta t} = \frac{d\omega}{dt}$. The sign of the angular acceleration α is found by examining the angular velocity. If a rotation rate of a rotating body is decreasing, the angular acceleration is in the opposite direction to ω. If the rotation rate is increasing, the angular acceleration

is in the same direction as ω.

- The tangential acceleration of a point at a radius from the axis of rotation is the angular acceleration times the radius to the point.

4.2 Rotation with Constant Angular Acceleration

- The kinematics of rotational motion describes the relationships among rotation angle (angular position), angular velocity, angular acceleration, and time.
- For a constant angular acceleration, the angular velocity varies linearly. Therefore, the average angular velocity is 1/2 the initial plus final angular velocity over a given time period:

$$\bar{\omega} = \frac{\omega_0 + \omega_f}{2}.$$

- We derived a set of kinematics equations for rotational motion (with constant angular acceleration) that are mathematically identical to the set of equations from **Section 3.4** for straight-line motion (with constant acceleration). The only differences are the substitution of the rotational-motion variables θ, ω and α for the translational-motion variables x, v and a. The equations are summarized in the following table:

Displacement from average Velocity	$\theta_f = \theta_0 + \bar{\omega} t$	$x_f = x_0 + \bar{v} t$
Velocity from Acceleration	$\omega_f = \omega_0 + \alpha t$	$v_f = v_0 + at$
Displacement from Velocity and Acceleration	$\theta_f = \theta_0 + \omega_0 t + \frac{1}{2}\alpha t^2$	$x_f = x_0 + v_0 t + \frac{1}{2}at^2$
Velocity from Displacement and Acceleration	$\omega_f^2 = \omega_0{}^2 + 2\alpha(\Delta\theta)$	$v_f^2 = v_0{}^2 + 2a(\Delta x)$

Table 4.4 Comparison of the Kinematic Equations for Rotational and Translational Motion

4.3 Relating Angular and Translational Quantities

- The linear kinematic equations have their rotational counterparts such that there is a mapping $x \to \theta, \quad v \to \omega, \quad a \to \alpha$.
- An object undergoing circular motion undergoes both angular and translational displacements, and possesses both angular and tangential velocities and accelerations. Their interrelationships are shown in **Table 4.3**

CONCEPTUAL QUESTIONS

4.1 Rotational Variables

1. A clock is mounted on the wall. As you look at it, what is the direction of the angular velocity vector of the second hand?

2. What is the value of the angular acceleration of the second hand of the clock on the wall?

3. A baseball bat is swung. Do all points on the bat have the same angular velocity? The same tangential speed?

4.2 Rotation with Constant Angular Acceleration

4. If a rigid body has a constant angular acceleration, what is the functional form of the angular velocity in terms of the time variable?

5. If a rigid body has a constant angular acceleration, what is the functional form of the angular position?

6. If the angular acceleration of a rigid body is zero, what is the functional form of the angular velocity?

7. A massless tether with a masses tied to both ends rotates about a fixed axis through the center. Can the total acceleration of the tether/mass combination be zero if the angular velocity is constant?

PROBLEMS

4.1 Rotational Variables

8. Calculate the angular velocity of Earth.

9. A track star runs a 400-m race on a 400-m circular track in 45 s. What is his angular velocity assuming a constant speed?

10. A wheel rotates at a constant rate of 2.0×10^3 rev/min. (a) What is its angular velocity in radians per second? (b) Through what angle does it turn in 10 s? Express the solution in radians and degrees.

11. A particle moves 3.0 m along a circle of radius 1.5 m. (a) Through what angle does it rotate? (b) If the particle makes this trip in 1.0 s at a constant speed, what is its angular velocity?

12. A compact disc rotates at 500 rev/min. If the diameter of the disc is 120 mm, (a) what is the tangential speed of a point at the edge of the disc? (b) At a point halfway to the center of the disc?

13. **Unreasonable results.** The propeller of an aircraft is spinning at 10 rev/s when the pilot shuts off the engine. The propeller reduces its angular velocity at a constant $2.0\,\text{rad/s}^2$ for a time period of 40 s. What is the rotation rate of the propeller in 40 s? Is this a reasonable situation?

14. A gyroscope slows from an initial rate of 32.0 rad/s at a rate of $0.700\,\text{rad/s}^2$. How long does it take to come to rest?

15. On takeoff, the propellers on a UAV (unmanned aerial vehicle) increase their angular velocity from rest at a rate of $\omega = (25.0t)\,\text{rad/s}$ for 3.0 s. (a) What is the instantaneous angular velocity of the propellers at $t = 2.0\,\text{s}$? (b) What is the angular acceleration?

4.2 Rotation with Constant Angular Acceleration

16. A wheel has a constant angular acceleration of $5.0\,\text{rad/s}^2$. Starting from rest, it turns through 300 rad. (a) What is its final angular velocity? (b) How much time elapses while it turns through the 300 radians?

17. During a 6.0-s time interval, a flywheel with a constant angular acceleration turns through 500 radians that acquire an angular velocity of 100 rad/s. (a) What is the angular velocity at the beginning of the 6.0 s? (b) What is the angular acceleration of the flywheel?

18. The angular velocity of a rotating rigid body increases from 500 to 1500 rev/min in 120 s. (a) What is the angular acceleration of the body? (b) Through what angle does it turn in this 120 s?

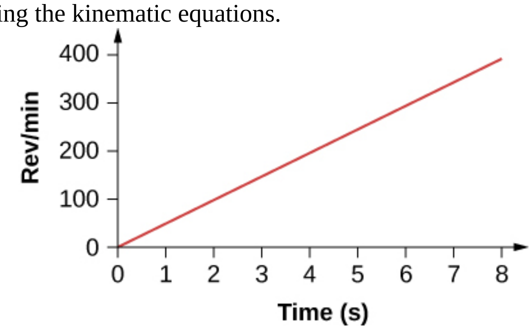
19. A flywheel slows from 600 to 400 rev/min while rotating through 40 revolutions. (a) What is the angular acceleration of the flywheel? (b) How much time elapses during the 40 revolutions?

20. A wheel 1.0 m in diameter rotates with an angular acceleration of $4.0\,\text{rad/s}^2$. (a) If the wheel's initial angular velocity is 2.0 rad/s, what is its angular velocity after 10 s? (b) Through what angle does it rotate in the 10-s interval? (c) What are the tangential speed and acceleration of a point on the rim of the wheel at the end of the 10-s interval?

21. A vertical wheel with a diameter of 50 cm starts from rest and rotates with a constant angular acceleration of $5.0\,\text{rad/s}^2$ around a fixed axis through its center counterclockwise. (a) Where is the point that is initially at the bottom of the wheel at $t = 10\,\text{s}$? (b) What is the point's linear acceleration at this instant?

22. A circular disk of radius 10 cm has a constant angular acceleration of $1.0\,\text{rad/s}^2$; at $t = 0$ its angular velocity is 2.0 rad/s. (a) Determine the disk's angular velocity at $t = 5.0\,\text{s}$. (b) What is the angle it has rotated through during this time? (c) What is the tangential acceleration of a point on the disk at $t = 5.0\,\text{s}$?

23. The angular velocity vs. time for a fan on a hovercraft is shown below. (a) What is the angle through which the fan blades rotate in the first 8 seconds? (b) Verify your result using the kinematic equations.



24. A rod of length 20 cm has two beads attached to its ends. The rod with beads starts rotating from rest. If the beads are to have a tangential speed of 20 m/s in 7 s, what is the angular acceleration of the rod to achieve this?

4.3 Relating Angular and Translational Quantities

25. At its peak, a tornado is 60.0 m in diameter and carries 500 km/h winds. What is its angular velocity in revolutions per second?

26. An ultracentrifuge accelerates from rest to 100,000 rpm in 2.00 min. (a) What is the average angular acceleration in rad/s^2? (b) What is the tangential acceleration of a point 9.50 cm from the axis of rotation? (c) What is the total distance traveled by a point 9.5 cm from the axis of rotation of the ultracentrifuge?

27. A wind turbine is rotating counterclockwise at 0.5 rev/s and slows to a stop in 10 s. Its blades are 20 m in length. (a) What is the initial tangential speed of a point at the tip of one of its blades? (b) What is the angular acceleration of the turbine?

28. What is (a) the angular speed and (b) the linear speed of a point on Earth's surface at latitude $30°$ N. Take the radius of the Earth to be 6309 km. (c) At what latitude would your linear speed be 10 m/s?

29. A bicycle wheel with radius 0.3m rotates from rest to 3 rev/s in 5 s. What is the magnitude of the tangential acceleration of a point on the outside edge of the wheel?

5 | INTRODUCTION TO VECTORS

Figure 5.1 A signpost gives information about distances and directions to towns or to other locations relative to the location of the signpost. Distance is a scalar quantity. Knowing the distance alone is not enough to get to the town; we must also know the direction from the signpost to the town. The direction, together with the distance, is a vector quantity commonly called the displacement vector. A signpost, therefore, gives information about displacement vectors from the signpost to towns. (credit: modification of work by "studio tdes"/Flickr, thedailyenglishshow.com)

Chapter Outline

Introduction

Vectors are essential to physics and engineering. Many fundamental physical quantities are vectors, including displacement, velocity, force, and electric and magnetic vector fields. Scalar products of vectors define other fundamental scalar physical quantities, such as energy. Vector products of vectors define still other fundamental vector physical quantities, such as torque and angular momentum. In other words, vectors are a component part of physics in much the same way as sentences are a component part of literature.

In introductory physics, vectors are Euclidean quantities that have geometric representations as arrows in one dimension (in a line), in two dimensions (in a plane), or in three dimensions (in space). They can be added, subtracted, or multiplied. In this chapter, we explore elements of vector algebra for applications in mechanics and in electricity and magnetism. Vector operations also have numerous generalizations in other branches of physics.

5.1 | Scalars and Vectors

Learning Objectives

By the end of this section, you will be able to:

- Describe the difference between vector and scalar quantities.
- Identify the magnitude and direction of a vector.
- Explain the effect of multiplying a vector quantity by a scalar.
- Describe how one-dimensional vector quantities are added or subtracted.
- Explain the geometric construction for the addition or subtraction of vectors in a plane.
- Distinguish between a vector equation and a scalar equation.

Many familiar physical quantities can be specified completely by giving a single number and the appropriate unit. For example, "a class period lasts 50 min" or "the gas tank in my car holds 65 L" or "the distance between two posts is 100 m." A physical quantity that can be specified completely in this manner is called a **scalar quantity**. Scalar is a synonym of "number." Time, mass, distance, length, volume, temperature, and energy are examples of **scalar** quantities.

Scalar quantities that have the same physical units can be added or subtracted according to the usual rules of algebra for numbers. For example, a class ending 10 min earlier than 50 min lasts $50\text{ min} - 10\text{ min} = 40\text{ min}$. Similarly, a 60-cal serving of corn followed by a 200-cal serving of donuts gives $60\text{ cal} + 200\text{ cal} = 260\text{ cal}$ of energy. When we multiply a scalar quantity by a number, we obtain the same scalar quantity but with a larger (or smaller) value. For example, if yesterday's breakfast had 200 cal of energy and today's breakfast has four times as much energy as it had yesterday, then today's breakfast has $4(200\text{ cal}) = 800\text{ cal}$ of energy. Two scalar quantities can also be multiplied or divided by each other to form a derived scalar quantity. For example, if a train covers a distance of 100 km in 1.0 h, its speed is 100.0 km/1.0 h = 27.8 m/s, where the speed is a derived scalar quantity obtained by dividing distance by time.

Many physical quantities, however, cannot be described completely by just a single number of physical units. For example, when the U.S. Coast Guard dispatches a ship or a helicopter for a rescue mission, the rescue team must know not only the distance to the distress signal, but also the direction from which the signal is coming so they can get to its origin as quickly as possible. Physical quantities specified completely by giving a number of units (magnitude) and a direction are called **vector quantities**. Examples of vector quantities include displacement, velocity, position, force, and torque. In the language of mathematics, physical vector quantities are represented by mathematical objects called **vectors** (**Figure 5.2**). We can add or subtract two vectors, and we can multiply a vector by a scalar or by another vector, but we cannot divide by a vector. The operation of division by a vector is not defined.

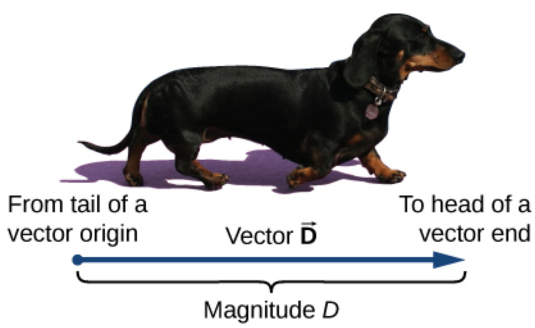


Figure 5.2 We draw a vector from the initial point or origin (called the "tail" of a vector) to the end or terminal point (called the "head" of a vector), marked by an arrowhead. Magnitude is the length of a vector and is always a positive scalar quantity. (credit "photo": modification of work by Cate Sevilla)

Let's examine vector algebra using a graphical method to be aware of basic terms and to develop a qualitative understanding. In practice, however, when it comes to solving physics problems, we use analytical methods, which we'll see in the next section. Analytical methods are more simple computationally and more accurate than graphical methods. From now on, to distinguish between a vector and a scalar quantity, we adopt the common convention that a letter in bold type with an arrow above it denotes a vector, and a letter without an arrow denotes a scalar. For example, a distance of 2.0 km, which is a scalar quantity, is denoted by d = 2.0 km, whereas a displacement of 2.0 km in some direction, which is a vector quantity, is denoted by $\vec{\mathbf{d}}$.

Suppose you tell a friend on a camping trip that you have discovered a terrific fishing hole 6 km from your tent. It is unlikely your friend would be able to find the hole easily unless you also communicate the direction in which it can be found with respect to your campsite. You may say, for example, "Walk about 6 km northeast from my tent." The key concept here is that you have to give not one but *two* pieces of information—namely, the distance or magnitude (6 km) *and* the direction (northeast).

Displacement is a general term used to describe a *change in position*, such as during a trip from the tent to the fishing hole. Displacement is an example of a vector quantity. If you walk from the tent (location A) to the hole (location B), as shown in **Figure 5.3**, the vector $\vec{\mathbf{D}}$, representing your **displacement**, is drawn as the arrow that originates at point A and ends at point B. The arrowhead marks the end of the vector. The direction of the displacement vector $\vec{\mathbf{D}}$ is the direction of the arrow. The length of the arrow represents the **magnitude** D of vector $\vec{\mathbf{D}}$. Here, D = 6 km. Since the magnitude of a vector is its length, which is a positive number, the magnitude is also indicated by placing the absolute value notation around the symbol that denotes the vector; so, we can write equivalently that $D \equiv \left| \vec{\mathbf{D}} \right|$. To solve a vector problem graphically, we need to draw the vector $\vec{\mathbf{D}}$ to scale. For example, if we assume 1 unit of distance (1 km) is represented in the drawing by a line segment of length u = 2 cm, then the total displacement in this example is represented by a vector of length $d = 6u = 6(2\text{ cm}) = 12\text{ cm}$, as shown in **Figure 5.4**. Notice that here, to avoid confusion, we used $D = 6\text{ km}$ to denote the magnitude of the actual displacement and d = 12 cm to denote the length of its representation in the drawing.

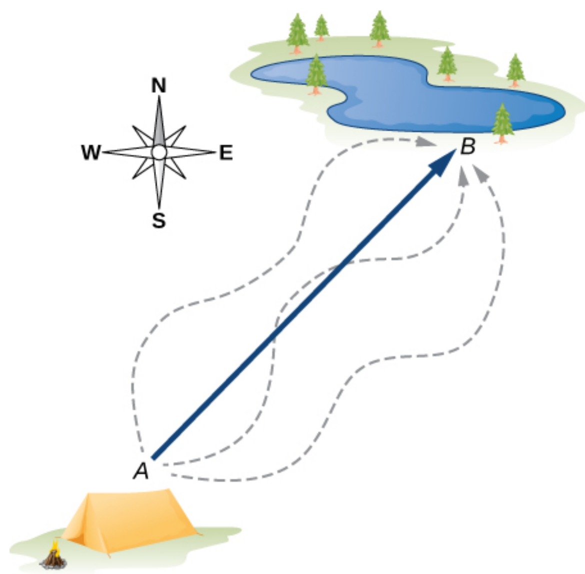


Figure 5.3 The displacement vector from point A (the initial position at the campsite) to point B (the final position at the fishing hole) is indicated by an arrow with origin at point A and end at point B. The displacement is the same for any of the actual paths (dashed curves) that may be taken between points A and B.

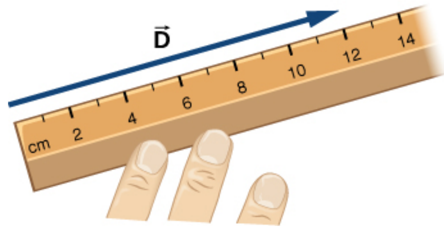


Figure 5.4 A displacement $\vec{\mathbf{D}}$ of magnitude 6 km is drawn to scale as a vector of length 12 cm when the length of 2 cm represents 1 unit of displacement (which in this case is 1 km).

Suppose your friend walks from the campsite at *A* to the fishing pond at *B* and then walks back: from the fishing pond at *B* to the campsite at *A*. The magnitude of the displacement vector $\vec{\mathbf{D}}_{AB}$ from *A* to *B* is the same as the magnitude of the displacement vector $\vec{\mathbf{D}}_{BA}$ from *B* to *A* (it equals 6 km in both cases), so we can write $D_{AB} = D_{BA}$. However, vector $\vec{\mathbf{D}}_{AB}$ is *not* equal to vector $\vec{\mathbf{D}}_{BA}$ because these two vectors have different directions: $\vec{\mathbf{D}}_{AB} \neq \vec{\mathbf{D}}_{BA}$. In **Figure 5.3**, vector $\vec{\mathbf{D}}_{BA}$ would be represented by a vector with an origin at point *B* and an end at point *A*, indicating vector $\vec{\mathbf{D}}_{BA}$ points to the southwest, which is exactly $180°$ opposite to the direction of vector $\vec{\mathbf{D}}_{AB}$. We say that vector $\vec{\mathbf{D}}_{BA}$ is **antiparallel** to vector $\vec{\mathbf{D}}_{AB}$ and write $\vec{\mathbf{D}}_{AB} = -\vec{\mathbf{D}}_{BA}$, where the minus sign indicates the antiparallel direction.

Two vectors that have identical directions are said to be **parallel vectors**—meaning, they are *parallel* to each other. Two parallel vectors $\vec{\mathbf{A}}$ and $\vec{\mathbf{B}}$ are equal, denoted by $\vec{\mathbf{A}} = \vec{\mathbf{B}}$, if and only if they have equal magnitudes $\left|\vec{\mathbf{A}}\right| = \left|\vec{\mathbf{B}}\right|$.

Two vectors with directions perpendicular to each other are said to be **orthogonal vectors**. These relations between vectors are illustrated in **Figure 5.5**.

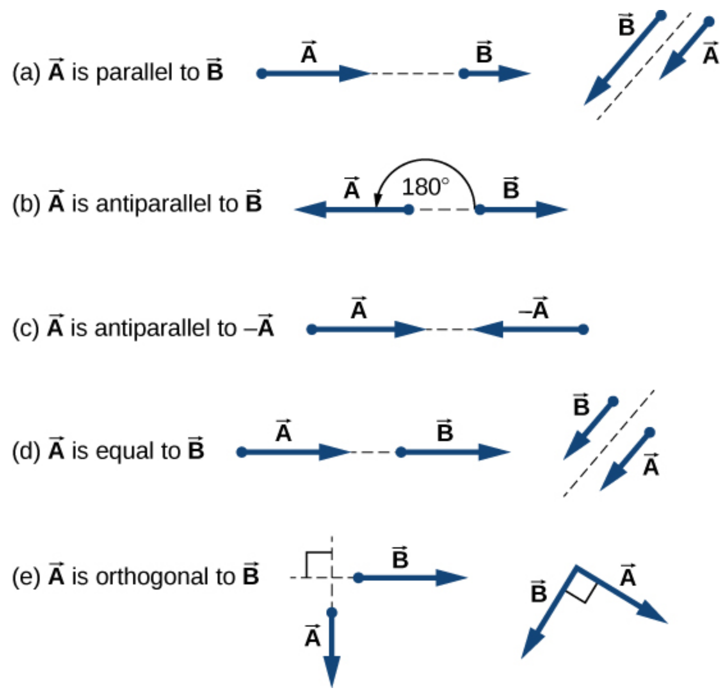


Figure 5.5 Various relations between two vectors $\vec{\mathbf{A}}$ and $\vec{\mathbf{B}}$. (a) $\vec{\mathbf{A}} \neq \vec{\mathbf{B}}$ because $A \neq B$. (b) $\vec{\mathbf{A}} \neq \vec{\mathbf{B}}$ because they are not parallel and $A \neq B$. (c) $\vec{\mathbf{A}} \neq -\vec{\mathbf{A}}$ because they have different directions (even though $\left|\vec{\mathbf{A}}\right| = \left|-\vec{\mathbf{A}}\right| = A$). (d) $\vec{\mathbf{A}} = \vec{\mathbf{B}}$ because they are parallel *and* have identical magnitudes $A = B$. (e) $\vec{\mathbf{A}} \neq \vec{\mathbf{B}}$ because they have different directions (are not parallel); here, their directions differ by $90°$—meaning, they are orthogonal.

5.1 Check Your Understanding Two motorboats named *Alice* and *Bob* are moving on a lake. Given the information about their velocity vectors in each of the following situations, indicate whether their velocity vectors are equal or otherwise. (a) *Alice* moves north at 6 knots and *Bob* moves west at 6 knots. (b) *Alice* moves west at 6 knots and *Bob* moves west at 3 knots. (c) *Alice* moves northeast at 6 knots and *Bob* moves south at 3 knots. (d) *Alice* moves northeast at 6 knots and *Bob* moves southwest at 6 knots. (e) *Alice* moves northeast at 2 knots and *Bob* moves closer to the shore northeast at 2 knots.

Algebra of Vectors in One Dimension

Vectors can be multiplied by scalars, added to other vectors, or subtracted from other vectors. We can illustrate these vector concepts using an example of the fishing trip seen in **Figure 5.6**.

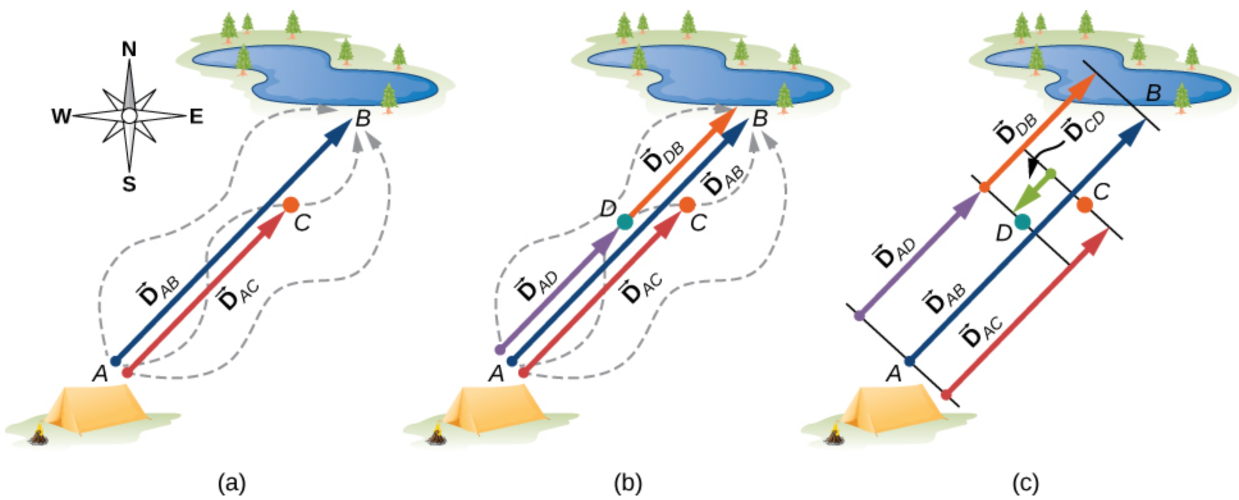


Figure 5.6 Displacement vectors for a fishing trip. (a) Stopping to rest at point *C* while walking from camp (point *A*) to the pond (point *B*). (b) Going back for the dropped tackle box (point *D*). (c) Finishing up at the fishing pond.

Suppose your friend departs from point *A* (the campsite) and walks in the direction to point *B* (the fishing pond), but, along the way, stops to rest at some point *C* located three-quarters of the distance between *A* and *B*, beginning from point *A* (**Figure 5.6**(a)). What is his displacement vector $\vec{\mathbf{D}}_{AC}$ when he reaches point *C*? We know that if he walks all the way to *B*, his displacement vector relative to *A* is $\vec{\mathbf{D}}_{AB}$, which has magnitude $D_{AB} = 6\text{ km}$ and a direction of northeast. If he walks only a 0.75 fraction of the total distance, maintaining the northeasterly direction, at point *C* he must be $0.75D_{AB} = 4.5\text{ km}$ away from the campsite at *A*. So, his displacement vector at the rest point *C* has magnitude $D_{AC} = 4.5\text{ km} = 0.75D_{AB}$ and is parallel to the displacement vector $\vec{\mathbf{D}}_{AB}$. All of this can be stated succinctly in the form of the following **vector equation**:

$$\vec{\mathbf{D}}_{AC} = 0.75\,\vec{\mathbf{D}}_{AB}.$$

In a vector equation, both sides of the equation are vectors. The previous equation is an example of a vector multiplied by a positive scalar (number) $\alpha = 0.75$. The result, $\vec{\mathbf{D}}_{AC}$, of such a multiplication is a new vector with a direction parallel to the direction of the original vector $\vec{\mathbf{D}}_{AB}$.

In general, when a vector $\vec{\mathbf{A}}$ is multiplied by a *positive* scalar α, the result is a new vector $\vec{\mathbf{B}}$ that is *parallel* to $\vec{\mathbf{A}}$:

$$\vec{\mathbf{B}} = \alpha\,\vec{\mathbf{A}}. \tag{5.1}$$

The magnitude $\left|\vec{\mathbf{B}}\right|$ of this new vector is obtained by multiplying the magnitude $\left|\vec{\mathbf{A}}\right|$ of the original vector, as expressed by the **scalar equation**:

$$B = |\alpha|A. \tag{5.2}$$

In a scalar equation, both sides of the equation are numbers. **Equation 5.2** is a scalar equation because the magnitudes of vectors are scalar quantities (and positive numbers). If the scalar α is *negative* in the vector equation **Equation 5.1**, then the magnitude $\left|\vec{\mathbf{B}}\right|$ of the new vector is still given by **Equation 5.2**, but the direction of the new vector $\vec{\mathbf{B}}$ is *antiparallel* to the direction of $\vec{\mathbf{A}}$. These principles are illustrated in **Figure 5.7**(a) by two examples where the length of vector $\vec{\mathbf{A}}$ is 1.5 units. When $\alpha = 2$, the new vector $\vec{\mathbf{B}} = 2\vec{\mathbf{A}}$ has length $B = 2A = 3.0\,\text{units}$ (twice as long as the original vector) and is parallel to the original vector. When $\alpha = -2$, the new vector $\vec{\mathbf{C}} = -2\vec{\mathbf{A}}$ has length $C = |-2|A = 3.0\,\text{units}$ (twice as long as the original vector) and is antiparallel to the original vector.

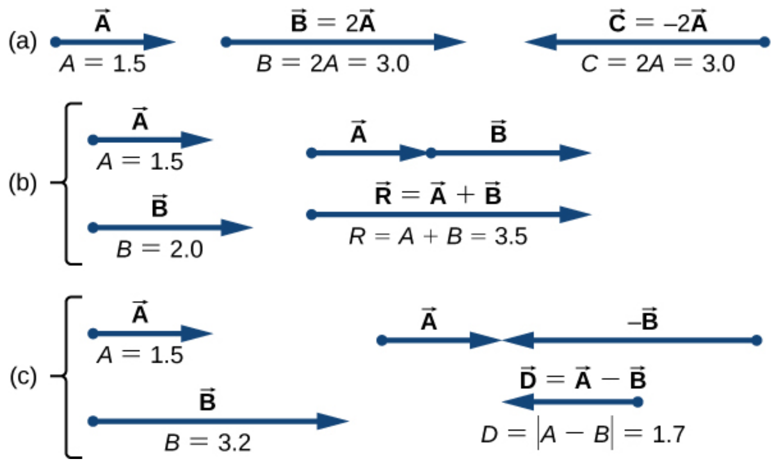


Figure 5.7 Algebra of vectors in one dimension. (a) Multiplication by a scalar. (b) Addition of two vectors ($\vec{\mathbf{R}}$ is called the *resultant* of vectors $\vec{\mathbf{A}}$ and $\vec{\mathbf{B}}$). (c) Subtraction of two vectors ($\vec{\mathbf{D}}$ is the difference of vectors $\vec{\mathbf{A}}$ and $\vec{\mathbf{B}}$).

Now suppose your fishing buddy departs from point *A* (the campsite), walking in the direction to point *B* (the fishing hole), but he realizes he lost his tackle box when he stopped to rest at point *C* (located three-quarters of the distance between *A* and *B*, beginning from point *A*). So, he turns back and retraces his steps in the direction toward the campsite and finds the box lying on the path at some point *D* only 1.2 km away from point *C* (see **Figure 5.6**(b)). What is his displacement vector $\vec{\mathbf{D}}_{AD}$ when he finds the box at point *D*? What is his displacement vector $\vec{\mathbf{D}}_{DB}$ from point *D* to the hole? We have already established that at rest point *C* his displacement vector is $\vec{\mathbf{D}}_{AC} = 0.75\,\vec{\mathbf{D}}_{AB}$. Starting at point *C*, he walks southwest (toward the campsite), which means his new displacement vector $\vec{\mathbf{D}}_{CD}$ from point *C* to point *D* is antiparallel to $\vec{\mathbf{D}}_{AB}$. Its magnitude $\left|\vec{\mathbf{D}}_{CD}\right|$ is $D_{CD} = 1.2\,\text{km} = 0.2D_{AB}$, so his second displacement vector is $\vec{\mathbf{D}}_{CD} = -0.2\,\vec{\mathbf{D}}_{AB}$. His total displacement $\vec{\mathbf{D}}_{AD}$ relative to the campsite is the **vector sum** of the two displacement vectors: vector $\vec{\mathbf{D}}_{AC}$ (from the campsite to the rest point) and vector $\vec{\mathbf{D}}_{CD}$ (from the rest point to the point where he finds his box):

$$\vec{\mathbf{D}}_{AD} = \vec{\mathbf{D}}_{AC} + \vec{\mathbf{D}}_{CD}. \quad (5.3)$$

The vector sum of two (or more) vectors is called the **resultant vector** or, for short, the *resultant*. When the vectors on the right-hand-side of **Equation 5.3** are known, we can find the resultant $\vec{\mathbf{D}}_{AD}$ as follows:

$$\vec{\mathbf{D}}_{AD} = \vec{\mathbf{D}}_{AC} + \vec{\mathbf{D}}_{CD} = 0.75\,\vec{\mathbf{D}}_{AB} - 0.2\,\vec{\mathbf{D}}_{AB} = (0.75 - 0.2)\,\vec{\mathbf{D}}_{AB} = 0.55\,\vec{\mathbf{D}}_{AB}. \quad (5.4)$$

When your friend finally reaches the pond at *B*, his displacement vector $\vec{\mathbf{D}}_{AB}$ from point *A* is the vector sum of his displacement vector $\vec{\mathbf{D}}_{AD}$ from point *A* to point *D* and his displacement vector $\vec{\mathbf{D}}_{DB}$ from point *D* to the fishing hole: $\vec{\mathbf{D}}_{AB} = \vec{\mathbf{D}}_{AD} + \vec{\mathbf{D}}_{DB}$ (see **Figure 5.6**(c)). This means his displacement vector $\vec{\mathbf{D}}_{DB}$ is the **difference of two vectors**:

$$\vec{\mathbf{D}}_{DB} = \vec{\mathbf{D}}_{AB} - \vec{\mathbf{D}}_{AD} = \vec{\mathbf{D}}_{AB} + (-\vec{\mathbf{D}}_{AD}). \tag{5.5}$$

Notice that a difference of two vectors is nothing more than a vector sum of two vectors because the second term in **Equation 5.5** is vector $-\vec{\mathbf{D}}_{AD}$ (which is antiparallel to $\vec{\mathbf{D}}_{AD}$). When we substitute **Equation 5.4** into **Equation 5.5**, we obtain the second displacement vector:

$$\vec{\mathbf{D}}_{DB} = \vec{\mathbf{D}}_{AB} - \vec{\mathbf{D}}_{AD} = \vec{\mathbf{D}}_{AB} - 0.55\,\vec{\mathbf{D}}_{AB} = (1.0 - 0.55)\,\vec{\mathbf{D}}_{AB} = 0.45\,\vec{\mathbf{D}}_{AB}. \tag{5.6}$$

This result means your friend walked $D_{DB} = 0.45D_{AB} = 0.45(6.0\text{ km}) = 2.7\text{ km}$ from the point where he finds his tackle box to the fishing hole.

When vectors $\vec{\mathbf{A}}$ and $\vec{\mathbf{B}}$ lie along a line (that is, in one dimension), such as in the camping example, their resultant $\vec{\mathbf{R}} = \vec{\mathbf{A}} + \vec{\mathbf{B}}$ and their difference $\vec{\mathbf{D}} = \vec{\mathbf{A}} - \vec{\mathbf{B}}$ both lie along the same direction. We can illustrate the addition or subtraction of vectors by drawing the corresponding vectors to scale in one dimension, as shown in **Figure 5.7**.

To illustrate the resultant when $\vec{\mathbf{A}}$ and $\vec{\mathbf{B}}$ are two parallel vectors, we draw them along one line by placing the origin of one vector at the end of the other vector in head-to-tail fashion (see **Figure 5.7**(b)). The magnitude of this resultant is the sum of their magnitudes: $R = A + B$. The direction of the resultant is parallel to both vectors. When vector $\vec{\mathbf{A}}$ is antiparallel to vector $\vec{\mathbf{B}}$, we draw them along one line in either head-to-head fashion (**Figure 5.7**(c)) or tail-to-tail fashion. The magnitude of the vector difference, then, is the *absolute value* $D = |A - B|$ of the difference of their magnitudes. The direction of the difference vector $\vec{\mathbf{D}}$ is parallel to the direction of the longer vector.

In general, in one dimension—as well as in higher dimensions, such as in a plane or in space—we can add any number of vectors and we can do so in any order because the addition of vectors is **commutative**,

$$\vec{\mathbf{A}} + \vec{\mathbf{B}} = \vec{\mathbf{B}} + \vec{\mathbf{A}}, \tag{5.7}$$

and **associative**,

$$(\vec{\mathbf{A}} + \vec{\mathbf{B}}) + \vec{\mathbf{C}} = \vec{\mathbf{A}} + (\vec{\mathbf{B}} + \vec{\mathbf{C}}). \tag{5.8}$$

Moreover, multiplication by a scalar is **distributive**:

$$\alpha_1 \vec{\mathbf{A}} + \alpha_2 \vec{\mathbf{A}} = (\alpha_1 + \alpha_2)\vec{\mathbf{A}}. \tag{5.9}$$

We used the distributive property in **Equation 5.4** and **Equation 5.6**.

When adding many vectors in one dimension, it is convenient to use the concept of a **unit vector**. A unit vector, which is denoted by a letter symbol with a hat, such as $\hat{\mathbf{u}}$, has a magnitude of one and does not have any physical unit so that $|\hat{\mathbf{u}}| \equiv u = 1$. The only role of a unit vector is to specify direction. For example, instead of saying vector $\vec{\mathbf{D}}_{AB}$ has a magnitude of 6.0 km and a direction of northeast, we can introduce a unit vector $\hat{\mathbf{u}}$ that points to the northeast and say succinctly that $\vec{\mathbf{D}}_{AB} = (6.0\text{ km})\hat{\mathbf{u}}$. Then the southwesterly direction is simply given by the unit vector $-\hat{\mathbf{u}}$. In this way, the displacement of 6.0 km in the southwesterly direction is expressed by the vector

$$\vec{\mathbf{D}}_{BA} = (-6.0\text{ km})\hat{\mathbf{u}}.$$

Example 5.1

A Ladybug Walker

A long measuring stick rests against a wall in a physics laboratory with its 200-cm end at the floor. A ladybug lands on the 100-cm mark and crawls randomly along the stick. It first walks 15 cm toward the floor, then it walks 56 cm toward the wall, then it walks 3 cm toward the floor again. Then, after a brief stop, it continues for 25 cm toward the floor and then, again, it crawls up 19 cm toward the wall before coming to a complete rest (**Figure 5.8**). Find the vector of its total displacement and its final resting position on the stick.

Strategy

If we choose the direction along the stick toward the floor as the direction of unit vector $\hat{\mathbf{u}}$, then the direction toward the floor is $+\hat{\mathbf{u}}$ and the direction toward the wall is $-\hat{\mathbf{u}}$. The ladybug makes a total of five displacements:

$$\begin{aligned} \vec{\mathbf{D}}_1 &= (15\text{ cm})(+\hat{\mathbf{u}}), \\ \vec{\mathbf{D}}_2 &= (56\text{ cm})(-\hat{\mathbf{u}}), \\ \vec{\mathbf{D}}_3 &= (3\text{ cm})(+\hat{\mathbf{u}}), \\ \vec{\mathbf{D}}_4 &= (25\text{ cm})(+\hat{\mathbf{u}}), \text{ and} \\ \vec{\mathbf{D}}_5 &= (19\text{ cm})(-\hat{\mathbf{u}}). \end{aligned}$$

The total displacement $\vec{\mathbf{D}}$ is the resultant of all its displacement vectors.

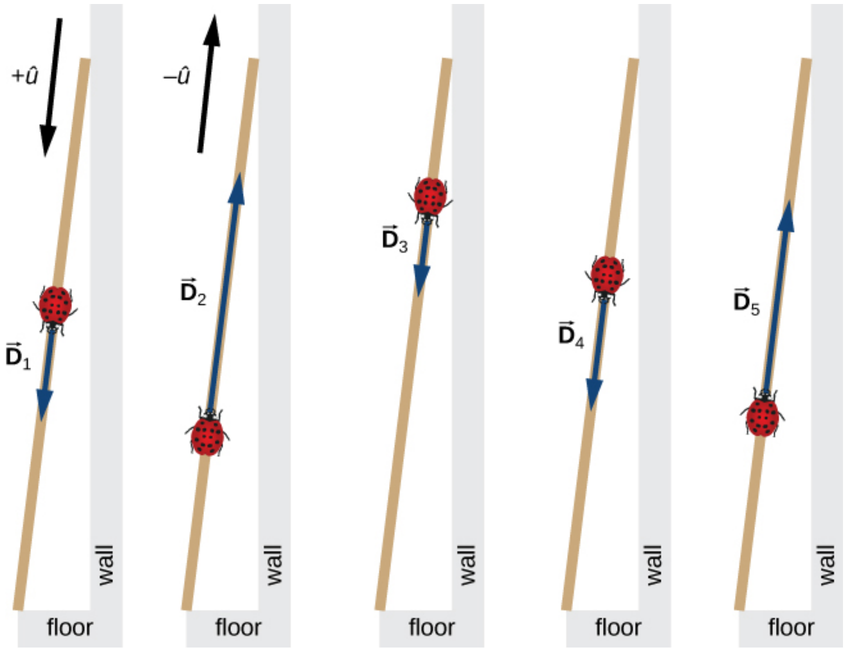


Figure 5.8 Five displacements of the ladybug. Note that in this schematic drawing, magnitudes of displacements are not drawn to scale. (credit "ladybug": modification of work by “Persian Poet Gal”/Wikimedia Commons)

Solution

The resultant of all the displacement vectors is

$$\begin{aligned}\vec{\mathbf{D}} &= \vec{\mathbf{D}}_1 + \vec{\mathbf{D}}_2 + \vec{\mathbf{D}}_3 + \vec{\mathbf{D}}_4 + \vec{\mathbf{D}}_5 \\ &= (15\text{ cm})(+\hat{\mathbf{u}}) + (56\text{ cm})(-\hat{\mathbf{u}}) + (3\text{ cm})(+\hat{\mathbf{u}}) + (25\text{ cm})(+\hat{\mathbf{u}}) + (19\text{ cm})(-\hat{\mathbf{u}}) \\ &= (15 - 56 + 3 + 25 - 19)\text{cm}\,\hat{\mathbf{u}} \\ &= -32\text{ cm}\,\hat{\mathbf{u}}.\end{aligned}$$

In this calculation, we use the distributive law given by **Equation 5.9**. The result reads that the total displacement vector points away from the 100-cm mark (initial landing site) toward the end of the meter stick that touches the wall. The end that touches the wall is marked 0 cm, so the final position of the ladybug is at the (100 – 32)cm = 68-cm mark.

5.2 Check Your Understanding A cave diver enters a long underwater tunnel. When her displacement with respect to the entry point is 20 m, she accidentally drops her camera, but she doesn't notice it missing until she is some 6 m farther into the tunnel. She swims back 10 m but cannot find the camera, so she decides to end the dive. How far from the entry point is she? Taking the positive direction out of the tunnel, what is her displacement vector relative to the entry point?

Algebra of Vectors in Two Dimensions

When vectors lie in a plane—that is, when they are in two dimensions—they can be multiplied by scalars, added to other vectors, or subtracted from other vectors in accordance with the general laws expressed by **Equation 5.1**, **Equation 5.2**, **Equation 5.7**, and **Equation 5.8**. However, the addition rule for two vectors in a plane becomes more complicated than the rule for vector addition in one dimension. We have to use the laws of geometry to construct resultant vectors, followed by trigonometry to find vector magnitudes and directions. This geometric approach is commonly used in navigation (**Figure 5.9**). In this section, we need to have at hand two rulers, a triangle, a protractor, a pencil, and an eraser for drawing vectors to scale by geometric constructions.

Figure 5.9 In navigation, the laws of geometry are used to draw resultant displacements on nautical maps.

For a geometric construction of the sum of two vectors in a plane, we follow **the parallelogram rule**. Suppose two vectors $\vec{\mathbf{A}}$ and $\vec{\mathbf{B}}$ are at the arbitrary positions shown in Figure 5.10. Translate either one of them in parallel to the beginning of the other vector, so that after the translation, both vectors have their origins at the same point. Now, at the end of vector $\vec{\mathbf{A}}$ we draw a line parallel to vector $\vec{\mathbf{B}}$ and at the end of vector $\vec{\mathbf{B}}$ we draw a line parallel to vector $\vec{\mathbf{A}}$ (the dashed lines in Figure 5.10). In this way, we obtain a parallelogram. From the origin of the two vectors we draw a diagonal that is the resultant $\vec{\mathbf{R}}$ of the two vectors: $\vec{\mathbf{R}} = \vec{\mathbf{A}} + \vec{\mathbf{B}}$ (Figure 5.10(a)). The other diagonal of this parallelogram is the vector difference of the two vectors $\vec{\mathbf{D}} = \vec{\mathbf{A}} - \vec{\mathbf{B}}$, as shown in Figure 5.10(b). Notice that the end of the difference vector is placed at the end of vector $\vec{\mathbf{A}}$.

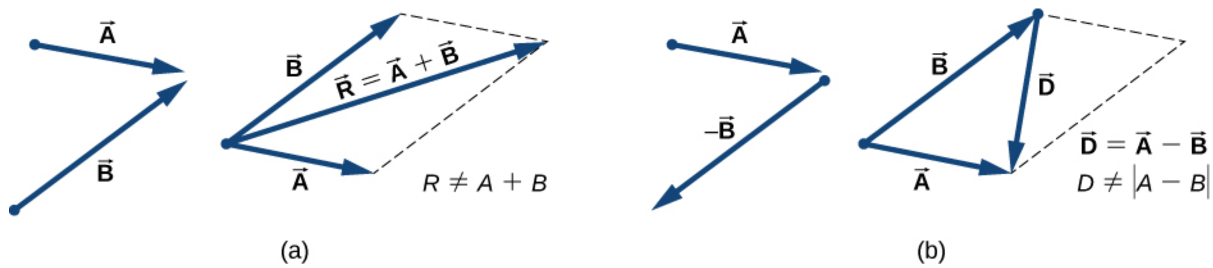


Figure 5.10 The parallelogram rule for the addition of two vectors. Make the parallel translation of each vector to a point where their origins (marked by the dot) coincide and construct a parallelogram with two sides on the vectors and the other two sides (indicated by dashed lines) parallel to the vectors. (a) Draw the resultant vector $\vec{\mathbf{R}}$ along the diagonal of the parallelogram from the common point to the opposite corner. Length R of the resultant vector is *not* equal to the sum of the magnitudes of the two vectors. (b) Draw the difference vector $\vec{\mathbf{D}} = \vec{\mathbf{A}} - \vec{\mathbf{B}}$ along the diagonal connecting the ends of the vectors. Place the origin of vector $\vec{\mathbf{D}}$ at the end of vector $\vec{\mathbf{B}}$ and the end (arrowhead) of vector $\vec{\mathbf{D}}$ at the end of vector $\vec{\mathbf{A}}$. Length D of the difference vector is *not* equal to the difference of magnitudes of the two vectors.

It follows from the parallelogram rule that neither the magnitude of the resultant vector nor the magnitude of the difference vector can be expressed as a simple sum or difference of magnitudes A and B, because the length of a diagonal cannot be expressed as a simple sum of side lengths. When using a geometric construction to find magnitudes $|\vec{\mathbf{R}}|$ and $|\vec{\mathbf{D}}|$, we have to use trigonometry laws for triangles, which may lead to complicated algebra. There are two ways to circumvent this algebraic complexity. One way is to use the method of components, which we examine in the next section. The other way is to draw the vectors to scale, as is done in navigation, and read approximate vector lengths and angles (directions) from the graphs. In this section we examine the second approach.

If we need to add three or more vectors, we repeat the parallelogram rule for the pairs of vectors until we find the resultant of all of the resultants. For three vectors, for example, we first find the resultant of vector 1 and vector 2, and then we find the resultant of this resultant and vector 3. The order in which we select the pairs of vectors does not matter because the operation of vector addition is commutative and associative (see Equation 5.7 and Equation 5.8). Before we state a general rule that follows from repetitive applications of the parallelogram rule, let's look at the following example.

Suppose you plan a vacation trip in Florida. Departing from Tallahassee, the state capital, you plan to visit your uncle Joe in Jacksonville, see your cousin Vinny in Daytona Beach, stop for a little fun in Orlando, see a circus performance in Tampa, and visit the University of Florida in Gainesville. Your route may be represented by five displacement vectors $\vec{\mathbf{A}}$, $\vec{\mathbf{B}}$, $\vec{\mathbf{C}}$, $\vec{\mathbf{D}}$, and $\vec{\mathbf{E}}$, which are indicated by the red vectors in Figure 5.11. What is your total displacement when you reach Gainesville? The total displacement is the vector sum of all five displacement vectors, which may be found by using the parallelogram rule four times. Alternatively, recall that the displacement vector has its beginning at the initial position (Tallahassee) and its end at the final position (Gainesville), so the total displacement vector can be drawn directly as an arrow connecting Tallahassee with Gainesville (see the green vector in Figure 5.11). When we use the parallelogram rule four times, the resultant $\vec{\mathbf{R}}$ we obtain is exactly this green vector connecting Tallahassee with Gainesville: $\vec{\mathbf{R}} = \vec{\mathbf{A}} + \vec{\mathbf{B}} + \vec{\mathbf{C}} + \vec{\mathbf{D}} + \vec{\mathbf{E}}$.

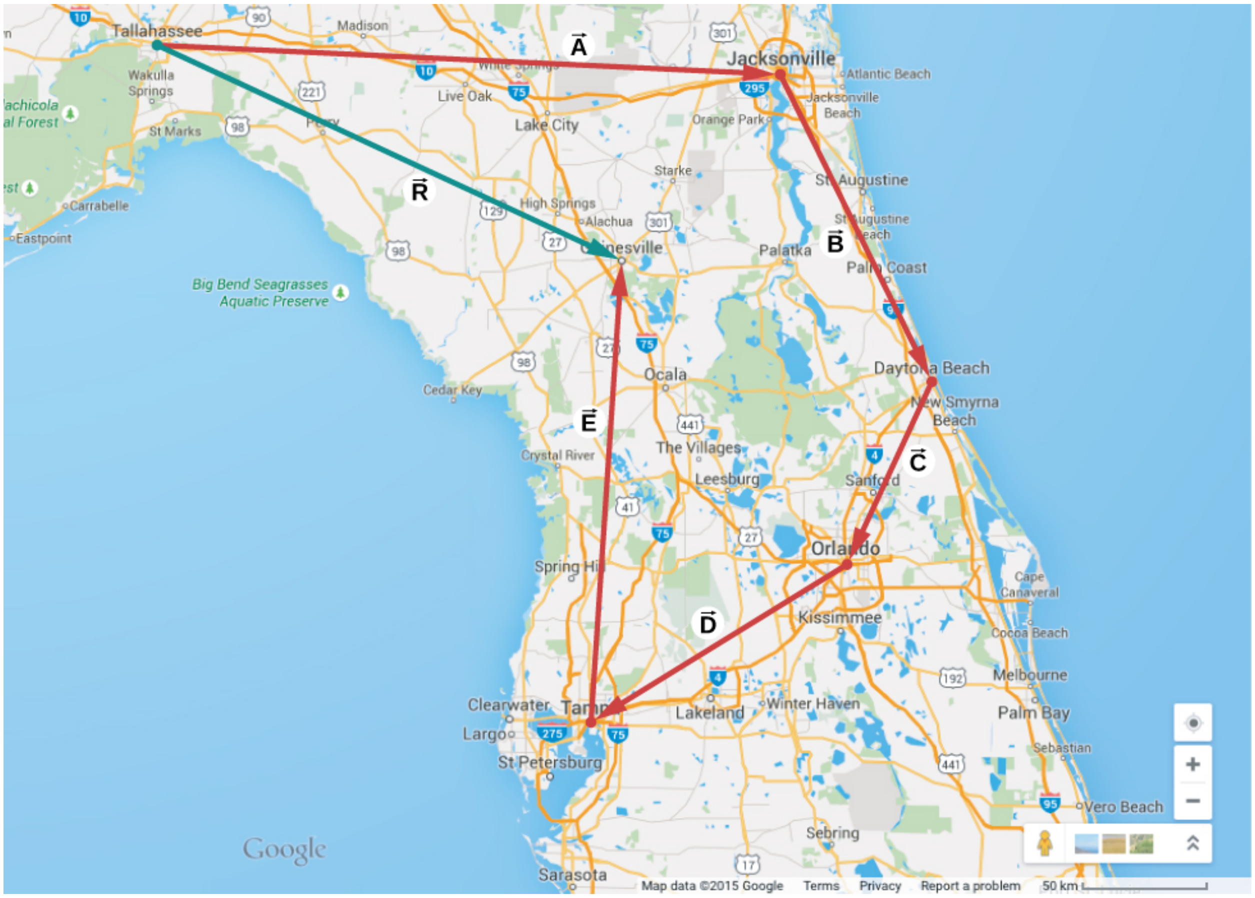


Figure 5.11 When we use the parallelogram rule four times, we obtain the resultant vector $\vec{\mathbf{R}} = \vec{\mathbf{A}} + \vec{\mathbf{B}} + \vec{\mathbf{C}} + \vec{\mathbf{D}} + \vec{\mathbf{E}}$, which is the green vector connecting Tallahassee with Gainesville.

Drawing the resultant vector of many vectors can be generalized by using the following **tail-to-head geometric construction**. Suppose we want to draw the resultant vector $\vec{\mathbf{R}}$ of four vectors $\vec{\mathbf{A}}$, $\vec{\mathbf{B}}$, $\vec{\mathbf{C}}$, and $\vec{\mathbf{D}}$ (**Figure 5.12**(a)). We select any one of the vectors as the first vector and make a parallel translation of a second vector to a position where the origin ("tail") of the second vector coincides with the end ("head") of the first vector. Then, we select a third vector and make a parallel translation of the third vector to a position where the origin of the third vector coincides with the end of the second vector. We repeat this procedure until all the vectors are in a head-to-tail arrangement like the one shown in **Figure 5.12**. We draw the resultant vector $\vec{\mathbf{R}}$ by connecting the origin ("tail") of the first vector with the end ("head") of the last vector. The end of the resultant vector is at the end of the last vector. Because the addition of vectors is associative and commutative, we obtain the same resultant vector regardless of which vector we choose to be first, second, third, or fourth in this construction.

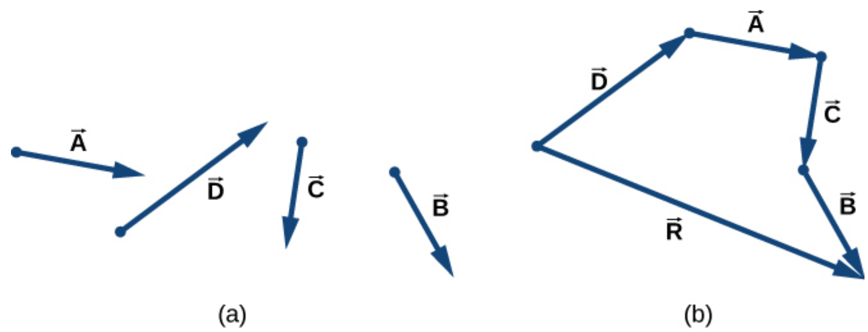


Figure 5.12 Tail-to-head method for drawing the resultant vector $\vec{\mathbf{R}} = \vec{\mathbf{A}} + \vec{\mathbf{B}} + \vec{\mathbf{C}} + \vec{\mathbf{D}}$. (a) Four vectors of different magnitudes and directions. (b) Vectors in (a) are translated to new positions where the origin ("tail") of one vector is at the end ("head") of another vector. The resultant vector is drawn from the origin ("tail") of the first vector to the end ("head") of the last vector in this arrangement.

Example 5.2

Geometric Construction of the Resultant

The three displacement vectors $\vec{\mathbf{A}}$, $\vec{\mathbf{B}}$, and $\vec{\mathbf{C}}$ in **Figure 5.13** are specified by their magnitudes $A = 10.0$, $B = 7.0$, and $C = 8.0$, respectively, and by their respective direction angles with the horizontal direction $\alpha = 35°$, $\beta = -110°$, and $\gamma = 30°$. The physical units of the magnitudes are centimeters. Choose a convenient scale and use a ruler and a protractor to find the following vector sums: (a) $\vec{\mathbf{R}} = \vec{\mathbf{A}} + \vec{\mathbf{B}}$, (b) $\vec{\mathbf{D}} = \vec{\mathbf{A}} - \vec{\mathbf{B}}$, and (c) $\vec{\mathbf{S}} = \vec{\mathbf{A}} - 3\vec{\mathbf{B}} + \vec{\mathbf{C}}$.

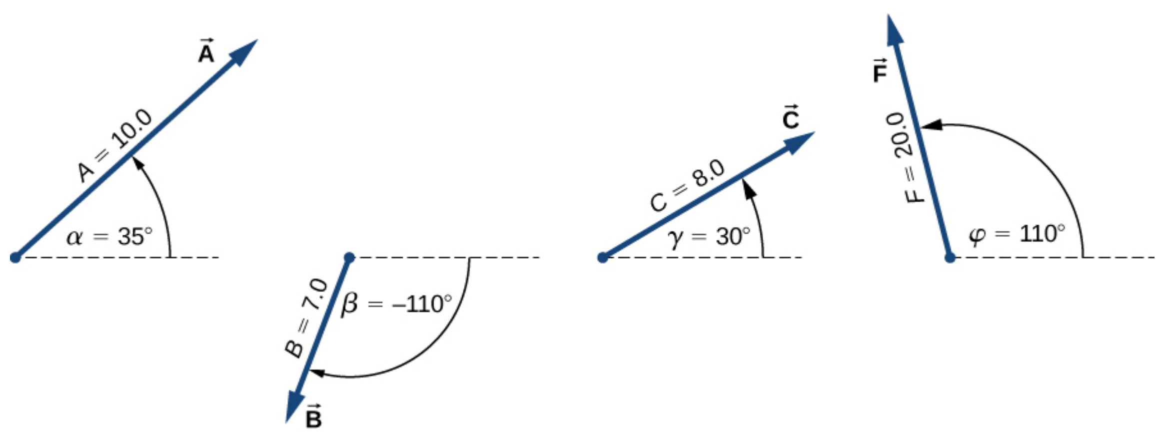


Figure 5.13 Vectors used in **Example 5.2** and in the Check Your Understanding feature that follows.

Strategy

In geometric construction, to find a vector means to find its magnitude and its direction angle with the horizontal direction. The strategy is to draw to scale the vectors that appear on the right-hand side of the equation and construct the resultant vector. Then, use a ruler and a protractor to read the magnitude of the resultant and the direction angle. For parts (a) and (b) we use the parallelogram rule. For (c) we use the tail-to-head method.

Solution

For parts (a) and (b), we attach the origin of vector $\vec{\mathbf{B}}$ to the origin of vector $\vec{\mathbf{A}}$, as shown in **Figure 5.14**, and construct a parallelogram. The shorter diagonal of this parallelogram is the sum $\vec{\mathbf{A}} + \vec{\mathbf{B}}$. The longer of the diagonals is the difference $\vec{\mathbf{A}} - \vec{\mathbf{B}}$. We use a ruler to measure the lengths of the diagonals, and a protractor to measure the angles with the horizontal. For the resultant $\vec{\mathbf{R}}$, we obtain $R = 5.8$ cm and $\theta_R \approx 0°$. For the difference $\vec{\mathbf{D}}$, we obtain $D = 16.2$ cm and $\theta_D = 49.3°$, which are shown in **Figure 5.14**.

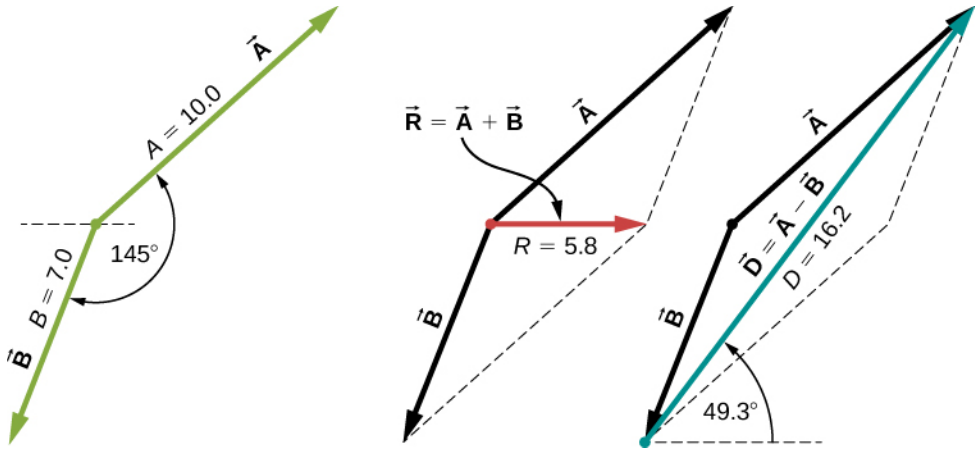


Figure 5.14 Using the parallelogram rule to solve (a) (finding the resultant, red) and (b) (finding the difference, blue).

For (c), we can start with vector $-3\vec{\mathbf{B}}$ and draw the remaining vectors tail-to-head as shown in **Figure 5.15**. In vector addition, the order in which we draw the vectors is unimportant, but drawing the vectors to scale is very important. Next, we draw vector $\vec{\mathbf{S}}$ from the origin of the first vector to the end of the last vector and place the arrowhead at the end of $\vec{\mathbf{S}}$. We use a ruler to measure the length of $\vec{\mathbf{S}}$, and find that its magnitude is $S = 36.9$ cm. We use a protractor and find that its direction angle is $\theta_S = 52.9°$. This solution is shown in **Figure 5.15**.

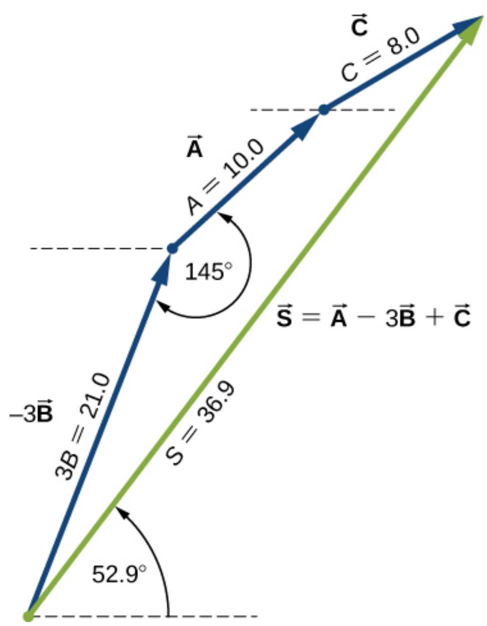


Figure 5.15 Using the tail-to-head method to solve (c) (finding vector $\vec{\mathbf{S}}$, green).

5.3 Check Your Understanding Using the three displacement vectors $\vec{\mathbf{A}}$, $\vec{\mathbf{B}}$, and $\vec{\mathbf{F}}$ in **Figure 5.13**, choose a convenient scale, and use a ruler and a protractor to find vector $\vec{\mathbf{G}}$ given by the vector equation $\vec{\mathbf{G}} = \vec{\mathbf{A}} + 2\vec{\mathbf{B}} - \vec{\mathbf{F}}$.

Observe the addition of vectors in a plane by visiting this **vector calculator (https://openstax.org/l/21compveccalc)** and this **Phet simulation (https://openstax.org/l/21phetvecaddsim)** .

5.2 | Coordinate Systems and Components of a Vector

Learning Objectives

By the end of this section, you will be able to:

- Describe vectors in two and three dimensions in terms of their components, using unit vectors along the axes.
- Distinguish between the vector components of a vector and the scalar components of a vector.
- Explain how the magnitude of a vector is defined in terms of the components of a vector.
- Identify the direction angle of a vector in a plane.
- Explain the connection between polar coordinates and Cartesian coordinates in a plane.

Vectors are usually described in terms of their components in a coordinate system. Even in everyday life we naturally invoke the concept of orthogonal projections in a rectangular coordinate system. For example, if you ask someone for directions to a particular location, you will more likely be told to go 40 km east and 30 km north than 50 km in the direction $37°$ north of east.

In a rectangular (Cartesian) *xy*-coordinate system in a plane, a point in a plane is described by a pair of coordinates (*x*, *y*). In a similar fashion, a vector $\vec{\mathbf{A}}$ in a plane is described by a pair of its *vector* coordinates. The *x*-coordinate of vector $\vec{\mathbf{A}}$ is called its *x*-component and the *y*-coordinate of vector $\vec{\mathbf{A}}$ is called its *y*-component. The vector *x*-component is a vector denoted by $\vec{\mathbf{A}}_x$. The vector *y*-component is a vector denoted by $\vec{\mathbf{A}}_y$. In the Cartesian system, the *x* and *y* **vector components** of a vector are the orthogonal projections of this vector onto the *x*- and *y*-axes, respectively. In this way, following the parallelogram rule for vector addition, each vector on a Cartesian plane can be expressed as the vector sum of its vector components:

$$\vec{\mathbf{A}} = \vec{\mathbf{A}}_x + \vec{\mathbf{A}}_y. \tag{5.10}$$

As illustrated in **Figure 5.16**, vector $\vec{\mathbf{A}}$ is the diagonal of the rectangle where the *x*-component $\vec{\mathbf{A}}_x$ is the side parallel to the *x*-axis and the *y*-component $\vec{\mathbf{A}}_y$ is the side parallel to the *y*-axis. Vector component $\vec{\mathbf{A}}_x$ is orthogonal to vector component $\vec{\mathbf{A}}_y$.

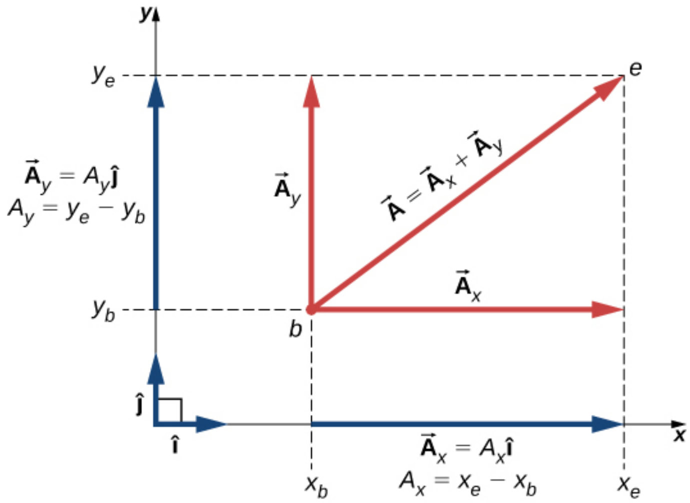


Figure 5.16 Vector $\vec{\mathbf{A}}$ in a plane in the Cartesian coordinate system is the vector sum of its vector *x*- and *y*-components. The *x*-vector component $\vec{\mathbf{A}}_x$ is the orthogonal projection of vector $\vec{\mathbf{A}}$ onto the *x*-axis. The *y*-vector component $\vec{\mathbf{A}}_y$ is the orthogonal projection of vector $\vec{\mathbf{A}}$ onto the *y*-axis. The numbers A_x and A_y that multiply the unit vectors are the scalar components of the vector.

It is customary to denote the positive direction on the *x*-axis by the unit vector $\hat{\mathbf{i}}$ and the positive direction on the *y*-axis by the unit vector $\hat{\mathbf{j}}$. **Unit vectors of the axes**, $\hat{\mathbf{i}}$ and $\hat{\mathbf{j}}$, define two orthogonal directions in the plane. As shown in **Figure 5.16**, the *x*- and *y*- components of a vector can now be written in terms of the unit vectors of the axes:

$$\begin{cases} \vec{\mathbf{A}}_x = A_x \hat{\mathbf{i}} \\ \vec{\mathbf{A}}_y = A_y \hat{\mathbf{j}}. \end{cases} \tag{5.11}$$

The vectors $\vec{\mathbf{A}}_x$ and $\vec{\mathbf{A}}_y$ defined by **Equation 5.11** are the *vector components* of vector $\vec{\mathbf{A}}$. The numbers A_x and A_y that define the vector components in **Equation 5.11** are the **scalar components** of vector $\vec{\mathbf{A}}$. Combining **Equation 5.10** with **Equation 5.11**, we obtain **the component form of a vector**:

$$\vec{\mathbf{A}} = A_x \hat{\mathbf{i}} + A_y \hat{\mathbf{j}}. \tag{5.12}$$

If we know the coordinates $b(x_b, y_b)$ of the origin point of a vector (where *b* stands for “beginning”) and the coordinates $e(x_e, y_e)$ of the end point of a vector (where *e* stands for “end”), we can obtain the scalar components of a vector simply by subtracting the origin point coordinates from the end point coordinates:

$$\begin{cases} A_x = x_e - x_b \\ A_y = y_e - y_b. \end{cases} \tag{5.13}$$

Example 5.3

Displacement of a Mouse Pointer

A mouse pointer on the display monitor of a computer at its initial position is at point (6.0 cm, 1.6 cm) with respect to the lower left-side corner. If you move the pointer to an icon located at point (2.0 cm, 4.5 cm), what is the displacement vector of the pointer?

Strategy

The origin of the *xy*-coordinate system is the lower left-side corner of the computer monitor. Therefore, the unit vector $\hat{\mathbf{i}}$ on the *x*-axis points horizontally to the right and the unit vector $\hat{\mathbf{j}}$ on the *y*-axis points vertically upward. The origin of the displacement vector is located at point *b*(6.0, 1.6) and the end of the displacement vector is located at point *e*(2.0, 4.5). Substitute the coordinates of these points into **Equation 5.13** to find the scalar components D_x and D_y of the displacement vector $\vec{\mathbf{D}}$. Finally, substitute the coordinates into **Equation 5.12** to write the displacement vector in the vector component form.

Solution

We identify $x_b = 6.0$, $x_e = 2.0$, $y_b = 1.6$, and $y_e = 4.5$, where the physical unit is 1 cm. The scalar *x*- and *y*-components of the displacement vector are

$$\begin{aligned} D_x &= x_e - x_b = (2.0 - 6.0)\text{cm} = -4.0\,\text{cm}, \\ D_y &= y_e - y_b = (4.5 - 1.6)\text{cm} = +2.9\,\text{cm}. \end{aligned}$$

The vector component form of the displacement vector is

$$\vec{\mathbf{D}} = D_x\hat{\mathbf{i}} + D_y\hat{\mathbf{j}} = (-4.0\,\text{cm})\hat{\mathbf{i}} + (2.9\,\text{cm})\hat{\mathbf{j}} = (-4.0\hat{\mathbf{i}} + 2.9\hat{\mathbf{j}})\text{cm}. \quad (5.14)$$

This solution is shown in **Figure 5.17**.

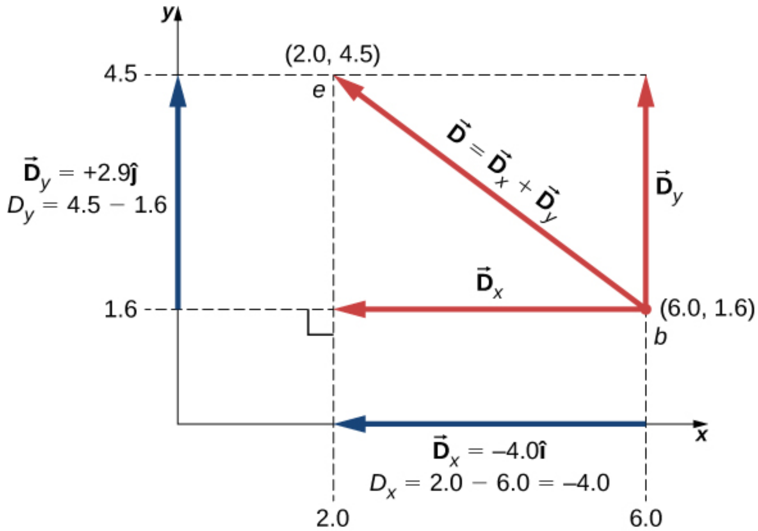


Figure 5.17 The graph of the displacement vector. The vector points from the origin point at *b* to the end point at *e*.

Significance

Notice that the physical unit—here, 1 cm—can be placed either with each component immediately before the unit vector or globally for both components, as in **Equation 5.14**. Often, the latter way is more convenient because it is simpler.

The vector x-component $\vec{\mathbf{D}}_x = -4.0\hat{\mathbf{i}} = 4.0(-\hat{\mathbf{i}})$ of the displacement vector has the magnitude $\left|\vec{\mathbf{D}}_x\right| = \left|-4.0\right|\left|\hat{\mathbf{i}}\right| = 4.0$ because the magnitude of the unit vector is $\left|\hat{\mathbf{i}}\right| = 1$. Notice, too, that the direction of the x-component is $-\hat{\mathbf{i}}$, which is antiparallel to the direction of the +x-axis; hence, the x-component vector $\vec{\mathbf{D}}_x$ points to the left, as shown in **Figure 5.17**. The scalar x-component of vector $\vec{\mathbf{D}}$ is $D_x = -4.0$.

Similarly, the vector y-component $\vec{\mathbf{D}}_y = +2.9\hat{\mathbf{j}}$ of the displacement vector has magnitude $\left|\vec{\mathbf{D}}_y\right| = \left|2.9\right|\left|\hat{\mathbf{j}}\right| = 2.9$ because the magnitude of the unit vector is $\left|\hat{\mathbf{j}}\right| = 1$. The direction of the y-component is $+\hat{\mathbf{j}}$, which is parallel to the direction of the +y-axis. Therefore, the y-component vector $\vec{\mathbf{D}}_y$ points up, as seen in **Figure 5.17**. The scalar y-component of vector $\vec{\mathbf{D}}$ is $D_y = +2.9$. The displacement vector $\vec{\mathbf{D}}$ is the resultant of its two *vector* components.

The vector component form of the displacement vector **Equation 5.14** tells us that the mouse pointer has been moved on the monitor 4.0 cm to the left and 2.9 cm upward from its initial position.

5.4 Check Your Understanding A blue fly lands on a sheet of graph paper at a point located 10.0 cm to the right of its left edge and 8.0 cm above its bottom edge and walks slowly to a point located 5.0 cm from the left edge and 5.0 cm from the bottom edge. Choose the rectangular coordinate system with the origin at the lower left-side corner of the paper and find the displacement vector of the fly. Illustrate your solution by graphing.

When we know the scalar components A_x and A_y of a vector $\vec{\mathbf{A}}$, we can find its magnitude A and its direction angle θ_A. The **direction angle**—or direction, for short—is the angle the vector forms with the positive direction on the x-axis. The angle θ_A is measured in the *counterclockwise direction* from the +x-axis to the vector (**Figure 5.18**). Because the lengths A, A_x, and A_y form a right triangle, they are related by the Pythagorean theorem:

$$A^2 = A_x^2 + A_y^2 \;\Leftrightarrow\; A = \sqrt{A_x^2 + A_y^2}. \tag{5.15}$$

This equation works even if the scalar components of a vector are negative. The direction angle θ_A of a vector is defined via the tangent function of angle θ_A in the triangle shown in **Figure 5.18**:

$$\tan\theta = \frac{A_y}{A_x} \tag{5.16}$$

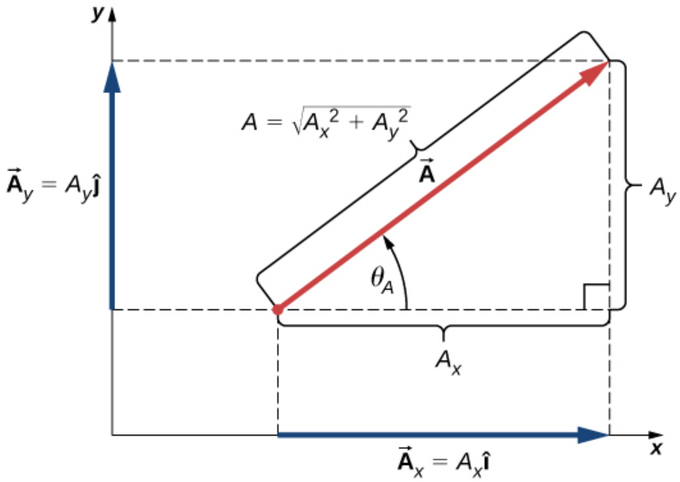


Figure 5.18 When the vector lies either in the first quadrant or in the fourth quadrant, where component A_x is positive (Figure 2.19), the direction angle θ_A in Equation 2.16) is identical to the angle θ .

When the vector lies either in the first quadrant or in the fourth quadrant, where component A_x is positive (**Figure 5.19**), the angle θ in **Equation 5.16** is identical to the direction angle θ_A. For vectors in the fourth quadrant, angle θ is negative, which means that for these vectors, direction angle θ_A is measured *clockwise* from the positive *x*-axis. Similarly, for vectors in the second quadrant, angle θ is negative. When the vector lies in either the second or third quadrant, where component A_x is negative, the direction angle is $\theta_A = \theta + 180°$ (**Figure 5.19**).

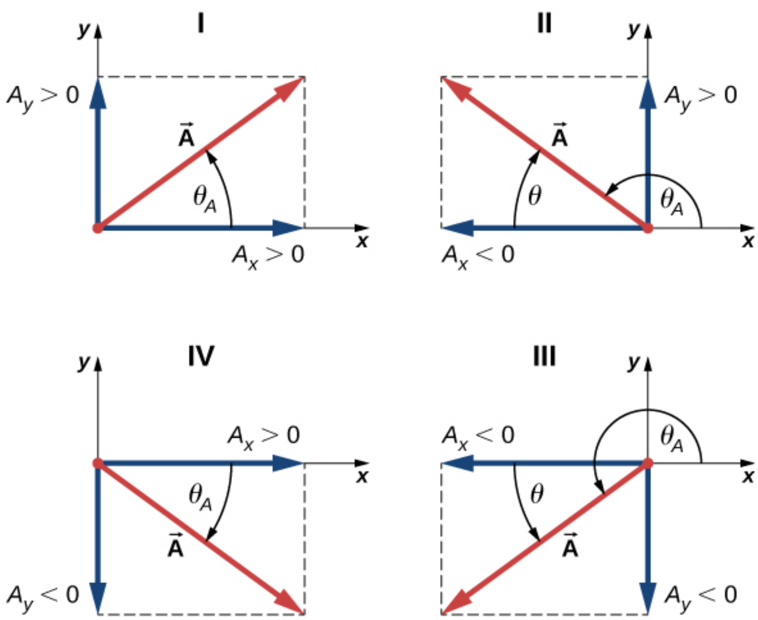


Figure 5.19 Scalar components of a vector may be positive or negative. Vectors in the first quadrant (I) have both scalar components positive and vectors in the third quadrant have both scalar components negative. For vectors in quadrants II and III, the direction angle of a vector is $\theta_A = \theta + 180°$.

Example 5.4

Magnitude and Direction of the Displacement Vector

You move a mouse pointer on the display monitor from its initial position at point (6.0 cm, 1.6 cm) to an icon located at point (2.0 cm, 4.5 cm). What are the magnitude and direction of the displacement vector of the pointer?

Strategy

In **Example 5.3**, we found the displacement vector $\vec{\mathbf{D}}$ of the mouse pointer (see **Equation 5.14**). We identify its scalar components $D_x = -4.0\,\text{cm}$ and $D_y = +2.9\,\text{cm}$ and substitute into **Equation 5.15** and **Equation 5.16** to find the magnitude D and direction θ_D, respectively.

Solution

The magnitude of vector $\vec{\mathbf{D}}$ is

$$D = \sqrt{D_x^2 + D_y^2} = \sqrt{(-4.0\,\text{cm})^2 + (2.9\,\text{cm})^2} = \sqrt{(4.0)^2 + (2.9)^2}\,\text{cm} = 4.9\,\text{cm}.$$

The direction angle is

$$\tan\theta = \frac{D_y}{D_x} = \frac{+2.9\,\text{cm}}{-4.0\,\text{cm}} = -0.725 \quad \Rightarrow \quad \theta = \tan^{-1}(-0.725) = -35.9°.$$

Vector $\vec{\mathbf{D}}$ lies in the second quadrant, so its direction angle is

$$\theta_D = \theta + 180° = -35.9° + 180° = 144.1°.$$

5.5 Check Your Understanding If the displacement vector of a blue fly walking on a sheet of graph paper is $\vec{\mathbf{D}} = (-5.00\,\hat{\mathbf{i}} - 3.00\,\hat{\mathbf{j}})\text{cm}$, find its magnitude and direction.

In many applications, the magnitudes and directions of vector quantities are known and we need to find the resultant of many vectors. For example, imagine 400 cars moving on the Golden Gate Bridge in San Francisco in a strong wind. Each car gives the bridge a different push in various directions and we would like to know how big the resultant push can possibly be. We have already gained some experience with the geometric construction of vector sums, so we know the task of finding the resultant by drawing the vectors and measuring their lengths and angles may become intractable pretty quickly, leading to huge errors. Worries like this do not appear when we use analytical methods. The very first step in an analytical approach is to find vector components when the direction and magnitude of a vector are known.

Let us return to the right triangle in **Figure 5.18**. The quotient of the adjacent side A_x to the hypotenuse A is the cosine function of direction angle θ_A, $A_x/A = \cos\theta_A$, and the quotient of the opposite side A_y to the hypotenuse A is the sine function of θ_A, $A_y/A = \sin\theta_A$. When magnitude A and direction θ_A are known, we can solve these relations for the scalar components:

$$\begin{cases} A_x = A\cos\theta_A \\ A_y = A\sin\theta_A \end{cases}. \quad (5.17)$$

When calculating vector components with **Equation 5.17**, care must be taken with the angle. The direction angle θ_A of a vector is the angle measured *counterclockwise* from the positive direction on the *x*-axis to the vector. The clockwise measurement gives a negative angle.

Example 5.5

Components of Displacement Vectors

A rescue party for a missing child follows a search dog named Trooper. Trooper wanders a lot and makes many trial sniffs along many different paths. Trooper eventually finds the child and the story has a happy ending, but his displacements on various legs seem to be truly convoluted. On one of the legs he walks 200.0 m southeast, then he runs north some 300.0 m. On the third leg, he examines the scents carefully for 50.0 m in the direction $30°$ west of north. On the fourth leg, Trooper goes directly south for 80.0 m, picks up a fresh scent and turns $23°$ west of south for 150.0 m. Find the scalar components of Trooper's displacement vectors and his displacement vectors in vector component form for each leg.

Strategy

Let's adopt a rectangular coordinate system with the positive *x*-axis in the direction of geographic east, with the positive *y*-direction pointed to geographic north. Explicitly, the unit vector $\hat{\mathbf{i}}$ of the *x*-axis points east and the unit vector $\hat{\mathbf{j}}$ of the *y*-axis points north. Trooper makes five legs, so there are five displacement vectors. We start by identifying their magnitudes and direction angles, then we use **Equation 5.17** to find the scalar components of the displacements and **Equation 5.12** for the displacement vectors.

Solution

On the first leg, the displacement magnitude is $L_1 = 200.0\text{ m}$ and the direction is southeast. For direction angle θ_1 we can take either $45°$ measured clockwise from the east direction or $45° + 270°$ measured counterclockwise from the east direction. With the first choice, $\theta_1 = -45°$. With the second choice, $\theta_1 = +315°$. We can use either one of these two angles. The components are

$$L_{1x} = L_1 \cos\theta_1 = (200.0\text{ m})\cos 315° = 141.4\text{ m},$$
$$L_{1y} = L_1 \sin\theta_1 = (200.0\text{ m})\sin 315° = -141.4\text{ m}.$$

The displacement vector of the first leg is

$$\vec{\mathbf{L}}_1 = L_{1x}\hat{\mathbf{i}} + L_{1y}\hat{\mathbf{j}} = (141.4\hat{\mathbf{i}} - 141.4\hat{\mathbf{j}})\text{ m}.$$

On the second leg of Trooper's wanderings, the magnitude of the displacement is $L_2 = 300.0\text{ m}$ and the direction is north. The direction angle is $\theta_2 = +90°$. We obtain the following results:

$$\begin{aligned} L_{2x} &= L_2 \cos\theta_2 = (300.0\text{ m})\cos 90° = 0.0, \\ L_{2y} &= L_2 \sin\theta_2 = (300.0\text{ m})\sin 90° = 300.0\text{ m}, \\ \vec{\mathbf{L}}_2 &= L_{2x}\hat{\mathbf{i}} + L_{2y}\hat{\mathbf{j}} = (300.0\text{ m})\hat{\mathbf{j}}. \end{aligned}$$

On the third leg, the displacement magnitude is $L_3 = 50.0\text{ m}$ and the direction is $30°$ west of north. The direction angle measured counterclockwise from the eastern direction is $\theta_3 = 30° + 90° = +120°$. This gives the following answers:

$$\begin{aligned} L_{3x} &= L_3 \cos\theta_3 = (50.0\text{ m})\cos 120° = -25.0\text{ m}, \\ L_{3y} &= L_3 \sin\theta_3 = (50.0\text{ m})\sin 120° = +43.3\text{ m}, \\ \vec{\mathbf{L}}_3 &= L_{3x}\hat{\mathbf{i}} + L_{3y}\hat{\mathbf{j}} = (-25.0\hat{\mathbf{i}} + 43.3\hat{\mathbf{j}})\text{m}. \end{aligned}$$

On the fourth leg of the excursion, the displacement magnitude is $L_4 = 80.0\text{ m}$ and the direction is south. The direction angle can be taken as either $\theta_4 = -90°$ or $\theta_4 = +270°$. We obtain

$$\begin{aligned} L_{4x} &= L_4 \cos\theta_4 = (80.0\ \text{m})\cos(-90°) = 0\,, \\ L_{4y} &= L_4 \sin\theta_4 = (80.0\ \text{m})\sin(-90°) = -80.0\ \text{m}, \\ \vec{\mathbf{L}}_4 &= L_{4x}\,\hat{\mathbf{i}} + L_{4y}\,\hat{\mathbf{j}} = (-80.0\ \text{m})\,\hat{\mathbf{j}}\,. \end{aligned}$$

On the last leg, the magnitude is $L_5 = 150.0\ \text{m}$ and the angle is $\theta_5 = -23° + 270° = +247°$ $(23°$ west of south), which gives

$$\begin{aligned} L_{5x} &= L_5 \cos\theta_5 = (150.0\ \text{m})\cos 247° = -58.6\ \text{m}, \\ L_{5y} &= L_5 \sin\theta_5 = (150.0\ \text{m})\sin 247° = -138.1\ \text{m}, \\ \vec{\mathbf{L}}_5 &= L_{5x}\,\hat{\mathbf{i}} + L_{5y}\,\hat{\mathbf{j}} = (-58.6\,\hat{\mathbf{i}} - 138.1\,\hat{\mathbf{j}})\text{m}. \end{aligned}$$

5.6 Check Your Understanding If Trooper runs 20 m west before taking a rest, what is his displacement vector?

Polar Coordinates

To describe locations of points or vectors in a plane, we need two orthogonal directions. In the Cartesian coordinate system these directions are given by unit vectors $\hat{\mathbf{i}}$ and $\hat{\mathbf{j}}$ along the *x*-axis and the *y*-axis, respectively. The Cartesian coordinate system is very convenient to use in describing displacements and velocities of objects and the forces acting on them. However, it becomes cumbersome when we need to describe the rotation of objects. When describing rotation, we usually work in the **polar coordinate system**.

In the polar coordinate system, the location of point *P* in a plane is given by two **polar coordinates** (Figure 5.20). The first polar coordinate is the **radial coordinate** *r*, which is the distance of point *P* from the origin. The second polar coordinate is an angle φ that the radial vector makes with some chosen direction, usually the positive *x*-direction. In polar coordinates, angles are measured in radians, or rads. The radial vector is attached at the origin and points away from the origin to point *P*. This radial direction is described by a unit radial vector $\hat{\mathbf{r}}$. The second unit vector $\hat{\mathbf{t}}$ is a vector orthogonal to the radial direction $\hat{\mathbf{r}}$. The positive $+\hat{\mathbf{t}}$ direction indicates how the angle φ changes in the counterclockwise direction. In this way, a point *P* that has coordinates (*x*, *y*) in the rectangular system can be described equivalently in the polar coordinate system by the two polar coordinates (r, φ). Equation 5.17 is valid for any vector, so we can use it to express the *x*- and *y*-coordinates of vector $\vec{\mathbf{r}}$. In this way, we obtain the connection between the polar coordinates and rectangular coordinates of point *P*:

$$\begin{cases} x = r\cos\varphi \\ y = r\sin\varphi \end{cases}. \tag{5.18}$$

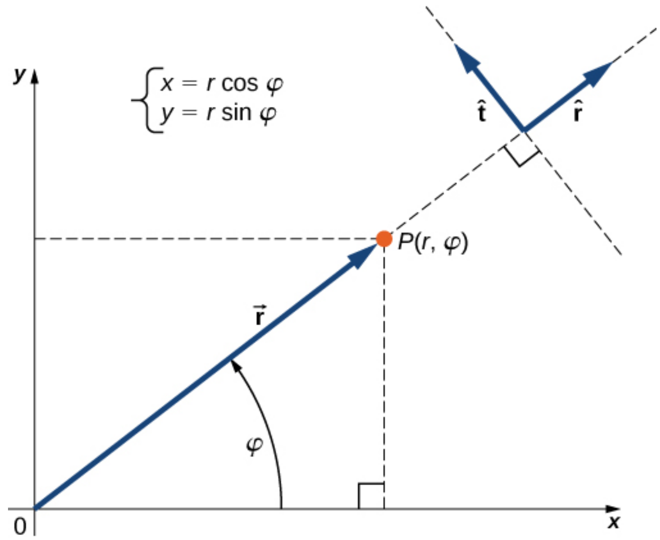


Figure 5.20 Using polar coordinates, the unit vector $\hat{\mathbf{r}}$ defines the positive direction along the radius r (radial direction) and, orthogonal to it, the unit vector $\hat{\mathbf{t}}$ defines the positive direction of rotation by the angle φ .

Example 5.6

Polar Coordinates

A treasure hunter finds one silver coin at a location 20.0 m away from a dry well in the direction $20°$ north of east and finds one gold coin at a location 10.0 m away from the well in the direction $20°$ north of west. What are the polar and rectangular coordinates of these findings with respect to the well?

Strategy

The well marks the origin of the coordinate system and east is the +x-direction. We identify radial distances from the locations to the origin, which are $r_S = 20.0\text{ m}$ (for the silver coin) and $r_G = 10.0\text{ m}$ (for the gold coin). To find the angular coordinates, we convert $20°$ to radians: $20° = \pi 20/180 = \pi/9$. We use **Equation 5.18** to find the x- and y-coordinates of the coins.

Solution

The angular coordinate of the silver coin is $\varphi_S = \pi/9$, whereas the angular coordinate of the gold coin is $\varphi_G = \pi - \pi/9 = 8\pi/9$. Hence, the polar coordinates of the silver coin are $(r_S,\ \varphi_S) = (20.0\text{ m},\ \pi/9)$ and those of the gold coin are $(r_G,\ \varphi_G) = (10.0\text{ m},\ 8\pi/9)$. We substitute these coordinates into **Equation 5.18** to obtain rectangular coordinates. For the gold coin, the coordinates are

$$\begin{cases} x_G = r_G \cos\varphi_G = (10.0\text{ m})\cos 8\pi/9 = -9.4\text{ m} \\ y_G = r_G \sin\varphi_G = (10.0\text{ m})\sin 8\pi/9 = 3.4\text{ m} \end{cases} \Rightarrow\ (x_G,\ y_G) = (-9.4\text{ m},\ 3.4\text{ m}).$$

For the silver coin, the coordinates are

$$\begin{cases} x_S = r_S \cos\varphi_S = (20.0\text{ m})\cos \pi/9 = 18.9\text{ m} \\ y_S = r_S \sin\varphi_S = (20.0\text{ m})\sin \pi/9 = 6.8\text{ m} \end{cases} \Rightarrow\ (x_S,\ y_S) = (18.9\text{ m},\ 6.8\text{ m}).$$

Vectors in Three Dimensions

To specify the location of a point in space, we need three coordinates (x, y, z), where coordinates x and y specify locations in a plane, and coordinate z gives a vertical position above or below the plane. Three-dimensional space has three orthogonal directions, so we need not two but *three* unit vectors to define a three-dimensional coordinate system. In the Cartesian coordinate system, the first two unit vectors are the unit vector of the *x*-axis $\hat{\mathbf{i}}$ and the unit vector of the *y*-axis $\hat{\mathbf{j}}$. The third unit vector $\hat{\mathbf{k}}$ is the direction of the *z*-axis (**Figure 5.21**). The order in which the axes are labeled, which is the order in which the three unit vectors appear, is important because it defines the orientation of the coordinate system. The order *x*-*y*-*z*, which is equivalent to the order $\hat{\mathbf{i}}$ - $\hat{\mathbf{j}}$ - $\hat{\mathbf{k}}$, defines the standard right-handed coordinate system (positive orientation).

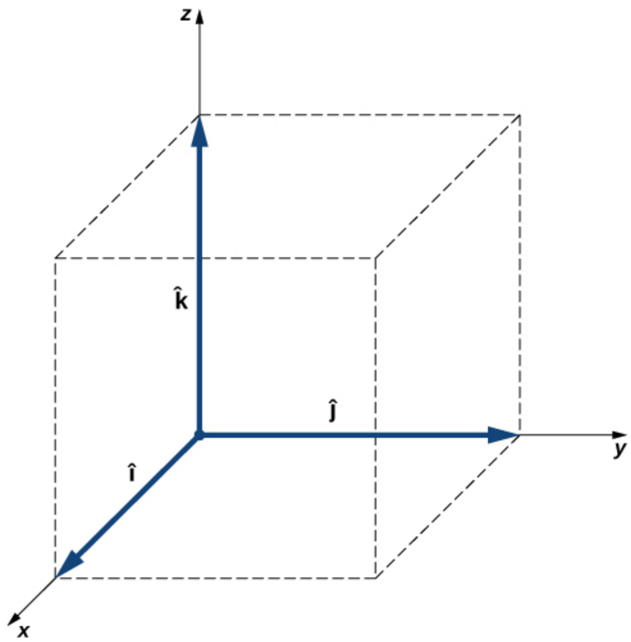


Figure 5.21 Three unit vectors define a Cartesian system in three-dimensional space. The order in which these unit vectors appear defines the orientation of the coordinate system. The order shown here defines the right-handed orientation.

In three-dimensional space, vector $\vec{\mathbf{A}}$ has three vector components: the *x*-component $\vec{\mathbf{A}}_x = A_x \hat{\mathbf{i}}$, which is the part of vector $\vec{\mathbf{A}}$ along the *x*-axis; the *y*-component $\vec{\mathbf{A}}_y = A_y \hat{\mathbf{j}}$, which is the part of $\vec{\mathbf{A}}$ along the *y*-axis; and the *z*-component $\vec{\mathbf{A}}_z = A_z \hat{\mathbf{k}}$, which is the part of the vector along the *z*-axis. A vector in three-dimensional space is the vector sum of its three vector components (**Figure 5.22**):

$$\vec{\mathbf{A}} = A_x \hat{\mathbf{i}} + A_y \hat{\mathbf{j}} + A_z \hat{\mathbf{k}}. \tag{5.19}$$

If we know the coordinates of its origin $b(x_b, y_b, z_b)$ and of its end $e(x_e, y_e, z_e)$, its scalar components are obtained by taking their differences: A_x and A_y are given by **Equation 5.13** and the *z*-component is given by

$$A_z = z_e - z_b. \tag{5.20}$$

Magnitude *A* is obtained by generalizing **Equation 5.15** to three dimensions:

$$A = \sqrt{A_x^2 + A_y^2 + A_z^2}. \tag{5.21}$$

This expression for the vector magnitude comes from applying the Pythagorean theorem twice. As seen in **Figure 5.22**, the diagonal in the *xy*-plane has length $\sqrt{A_x^2 + A_y^2}$ and its square adds to the square A_z^2 to give A^2. Note that when the *z*-component is zero, the vector lies entirely in the *xy*-plane and its description is reduced to two dimensions.

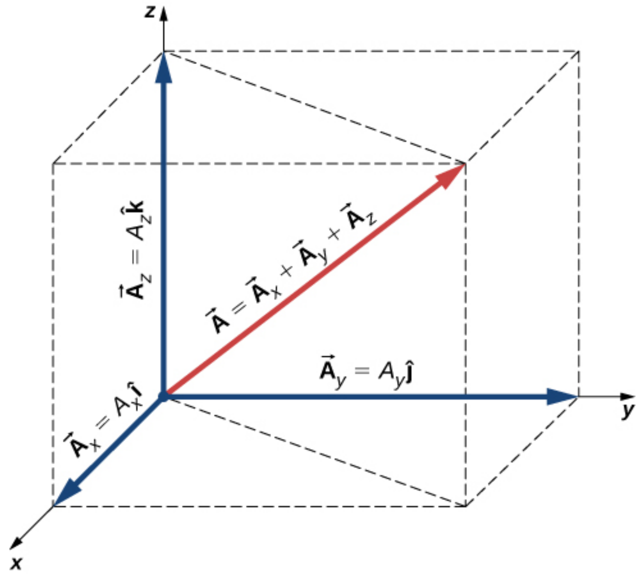


Figure 5.22 A vector in three-dimensional space is the vector sum of its three vector components.

Example 5.7

Takeoff of a Drone

During a takeoff of IAI Heron (**Figure 5.23**), its position with respect to a control tower is 100 m above the ground, 300 m to the east, and 200 m to the north. One minute later, its position is 250 m above the ground, 1200 m to the east, and 2100 m to the north. What is the drone's displacement vector with respect to the control tower? What is the magnitude of its displacement vector?

Figure 5.23 The drone IAI Heron in flight. (credit: SSgt Reynaldo Ramon, USAF)

Strategy

We take the origin of the Cartesian coordinate system as the control tower. The direction of the +*x*-axis is given by unit vector $\hat{\mathbf{i}}$ to the east, the direction of the +*y*-axis is given by unit vector $\hat{\mathbf{j}}$ to the north, and the direction of the +*z*-axis is given by unit vector $\hat{\mathbf{k}}$, which points up from the ground. The drone's first position is the origin (or, equivalently, the beginning) of the displacement vector and its second position is the end of the displacement vector.

Solution

We identify *b*(300.0 m, 200.0 m, 100.0 m) and *e*(1200 m, 2100 m, 250 m), and use **Equation 5.13** and **Equation 5.20** to find the scalar components of the drone's displacement vector:

$$\begin{cases} D_x = x_e - x_b = 1200.0\,\text{m} - 300.0\,\text{m} = 900.0\,\text{m}, \\ D_y = y_e - y_b = 2100.0\,\text{m} - 200.0\,\text{m} = 1900.0\,\text{m}, \\ D_z = z_e - z_b = 250.0\,\text{m} - 100.0\,\text{m} = 150.0\,\text{m}. \end{cases}$$

We substitute these components into **Equation 5.19** to find the displacement vector:

$$\vec{\mathbf{D}} = D_x\hat{\mathbf{i}} + D_y\hat{\mathbf{j}} + D_z\hat{\mathbf{k}} = 900.0\,\text{m}\,\hat{\mathbf{i}} + 1900.0\,\text{m}\,\hat{\mathbf{j}} + 150.0\,\text{m}\,\hat{\mathbf{k}} = (0.90\,\hat{\mathbf{i}} + 1.90\,\hat{\mathbf{j}} + 0.15\,\hat{\mathbf{k}})\,\text{km}.$$

We substitute into **Equation 5.21** to find the magnitude of the displacement:

$$D = \sqrt{D_x^2 + D_y^2 + D_z^2} = \sqrt{(0.90\,\text{km})^2 + (1.90\,\text{km})^2 + (0.15\,\text{km})^2} = 2.11\,\text{km}.$$

5.7 Check Your Understanding If the average velocity vector of the drone in the displacement in **Example 5.7** is $\vec{\mathbf{u}} = (15.0\,\hat{\mathbf{i}} + 31.7\,\hat{\mathbf{j}} + 2.5\,\hat{\mathbf{k}})\text{m/s}$, what is the magnitude of the drone's velocity vector?

5.3 | Algebra of Vectors

Learning Objectives

By the end of this section, you will be able to:

- Apply analytical methods of vector algebra to find resultant vectors and to solve vector equations for unknown vectors.
- Interpret physical situations in terms of vector expressions.

Vectors can be added together and multiplied by scalars. We have already seen that vector addition is associative (**Equation 5.8**) and commutative (**Equation 5.7**), and that vector multiplication by a sum of scalars is distributive (**Equation 5.9**). Also, scalar multiplication by a sum of vectors is distributive:

$$\alpha(\vec{\mathbf{A}} + \vec{\mathbf{B}}) = \alpha\vec{\mathbf{A}} + \alpha\vec{\mathbf{B}}. \tag{5.22}$$

In this equation, α is any number (a scalar). For example, a vector antiparallel to vector $\vec{\mathbf{A}} = A_x\hat{\mathbf{i}} + A_y\hat{\mathbf{j}} + A_z\hat{\mathbf{k}}$ can be expressed simply by multiplying $\vec{\mathbf{A}}$ by the scalar $\alpha = -1$:

$$-\vec{\mathbf{A}} = -A_x\hat{\mathbf{i}} - A_y\hat{\mathbf{j}} - A_z\hat{\mathbf{k}}. \tag{5.23}$$

Example 5.8

Direction of Motion

In a Cartesian coordinate system where $\hat{\mathbf{i}}$ denotes geographic east, $\hat{\mathbf{j}}$ denotes geographic north, and $\hat{\mathbf{k}}$ denotes altitude above sea level, a military convoy advances its position through unknown territory with velocity $\vec{\mathbf{v}} = (4.0\,\hat{\mathbf{i}} + 3.0\,\hat{\mathbf{j}} + 0.1\,\hat{\mathbf{k}})\text{km/h}$. If the convoy had to retreat, in what geographic direction would it be moving?

Solution

The velocity vector has the third component $\vec{\mathbf{v}}_z = (+0.1\text{km/h})\,\hat{\mathbf{k}}$, which says the convoy is climbing at a rate of 100 m/h through mountainous terrain. At the same time, its velocity is 4.0 km/h to the east and 3.0 km/h to the north, so it moves on the ground in direction $\tan^{-1}(3/4) \approx 37°$ north of east. If the convoy had to retreat, its new velocity vector $\vec{\mathbf{u}}$ would have to be antiparallel to $\vec{\mathbf{v}}$ and be in the form $\vec{\mathbf{u}} = -\alpha\,\vec{\mathbf{v}}$, where α is a positive number. Thus, the velocity of the retreat would be $\vec{\mathbf{u}} = \alpha(-4.0\,\hat{\mathbf{i}} - 3.0\,\hat{\mathbf{j}} - 0.1\,\hat{\mathbf{k}})\text{km/h}$. The negative sign of the third component indicates the convoy would be descending. The direction angle of the retreat velocity is $\tan^{-1}(-3\alpha/-4\alpha) \approx 37°$ south of west. Therefore, the convoy would be moving on the ground in direction $37°$ south of west while descending on its way back.

The generalization of the number zero to vector algebra is called the **null vector**, denoted by $\vec{\mathbf{0}}$. All components of the null vector are zero, $\vec{\mathbf{0}} = 0\,\hat{\mathbf{i}} + 0\,\hat{\mathbf{j}} + 0\,\hat{\mathbf{k}}$, so the null vector has no length and no direction.

Two vectors $\vec{\mathbf{A}}$ and $\vec{\mathbf{B}}$ are **equal vectors** if and only if their difference is the null vector:

$$\vec{\mathbf{0}} = \vec{\mathbf{A}} - \vec{\mathbf{B}} = (A_x\,\hat{\mathbf{i}} + A_y\,\hat{\mathbf{j}} + A_z\,\hat{\mathbf{k}}) - (B_x\,\hat{\mathbf{i}} + B_y\,\hat{\mathbf{j}} + B_z\,\hat{\mathbf{k}}) = (A_x - B_x)\,\hat{\mathbf{i}} + (A_y - B_y)\,\hat{\mathbf{j}} + (A_z - B_z)\,\hat{\mathbf{k}}.$$

This vector equation means we must have simultaneously $A_x - B_x = 0$, $A_y - B_y = 0$, and $A_z - B_z = 0$. Hence, we can write $\vec{\mathbf{A}} = \vec{\mathbf{B}}$ if and only if the corresponding components of vectors $\vec{\mathbf{A}}$ and $\vec{\mathbf{B}}$ are equal:

$$\vec{\mathbf{A}} = \vec{\mathbf{B}} \Leftrightarrow \begin{cases} A_x = B_x \\ A_y = B_y. \\ A_z = B_z \end{cases} \tag{5.24}$$

Two vectors are equal when their corresponding scalar components are equal.

Resolving vectors into their scalar components (i.e., finding their scalar components) and expressing them analytically in vector component form (given by **Equation 5.19**) allows us to use vector algebra to find sums or differences of many vectors *analytically* (i.e., without using graphical methods). For example, to find the resultant of two vectors $\vec{\mathbf{A}}$ and $\vec{\mathbf{B}}$, we simply add them component by component, as follows:

$$\vec{\mathbf{R}} = \vec{\mathbf{A}} + \vec{\mathbf{B}} = (A_x\,\hat{\mathbf{i}} + A_y\,\hat{\mathbf{j}} + A_z\,\hat{\mathbf{k}}) + (B_x\,\hat{\mathbf{i}} + B_y\,\hat{\mathbf{j}} + B_z\,\hat{\mathbf{k}}) = (A_x + B_x)\,\hat{\mathbf{i}} + (A_y + B_y)\,\hat{\mathbf{j}} + (A_z + B_z)\,\hat{\mathbf{k}}.$$

In this way, using **Equation 5.24**, scalar components of the resultant vector $\vec{\mathbf{R}} = R_x\,\hat{\mathbf{i}} + R_y\,\hat{\mathbf{j}} + R_z\,\hat{\mathbf{k}}$ are the sums of corresponding scalar components of vectors $\vec{\mathbf{A}}$ and $\vec{\mathbf{B}}$:

$$\begin{cases} R_x = A_x + B_x, \\ R_y = A_y + B_y, \\ R_z = A_z + B_z. \end{cases}$$

Analytical methods can be used to find components of a resultant of many vectors. For example, if we are to sum up N vectors $\vec{\mathbf{F}}_1, \vec{\mathbf{F}}_2, \vec{\mathbf{F}}_3, \ldots, \vec{\mathbf{F}}_N$, where each vector is $\vec{\mathbf{F}}_k = F_{kx}\hat{\mathbf{i}} + F_{ky}\hat{\mathbf{j}} + F_{kz}\hat{\mathbf{k}}$, the resultant vector $\vec{\mathbf{F}}_R$ is

$$\begin{aligned} \vec{\mathbf{F}}_R = \vec{\mathbf{F}}_1 + \vec{\mathbf{F}}_2 + \vec{\mathbf{F}}_3 + \ldots + \vec{\mathbf{F}}_N &= \sum_{k=1}^{N} \vec{\mathbf{F}}_k = \sum_{k=1}^{N} \left(F_{kx}\hat{\mathbf{i}} + F_{ky}\hat{\mathbf{j}} + F_{kz}\hat{\mathbf{k}} \right) \\ &= \left(\sum_{k=1}^{N} F_{kx} \right)\hat{\mathbf{i}} + \left(\sum_{k=1}^{N} F_{ky} \right)\hat{\mathbf{j}} + \left(\sum_{k=1}^{N} F_{kz} \right)\hat{\mathbf{k}}. \end{aligned}$$

Therefore, scalar components of the resultant vector are

$$\begin{cases} F_{Rx} = \sum_{k=1}^{N} F_{kx} = F_{1x} + F_{2x} + \ldots + F_{Nx} \\ F_{Ry} = \sum_{k=1}^{N} F_{ky} = F_{1y} + F_{2y} + \ldots + F_{Ny} \\ F_{Rz} = \sum_{k=1}^{N} F_{kz} = F_{1z} + F_{2z} + \ldots + F_{Nz}. \end{cases} \tag{5.25}$$

Having found the scalar components, we can write the resultant in vector component form:

$$\vec{\mathbf{F}}_R = F_{Rx}\hat{\mathbf{i}} + F_{Ry}\hat{\mathbf{j}} + F_{Rz}\hat{\mathbf{k}}.$$

Analytical methods for finding the resultant and, in general, for solving vector equations are very important in physics because many physical quantities are vectors. For example, we use this method in kinematics to find resultant displacement vectors and resultant velocity vectors, in mechanics to find resultant force vectors and the resultants of many derived vector quantities, and in electricity and magnetism to find resultant electric or magnetic vector fields.

Example 5.9

Analytical Computation of a Resultant

Three displacement vectors $\vec{\mathbf{A}}$, $\vec{\mathbf{B}}$, and $\vec{\mathbf{C}}$ in a plane (**Figure 5.13**) are specified by their magnitudes $A = 10.0$, $B = 7.0$, and $C = 8.0$, respectively, and by their respective direction angles with the horizontal direction $\alpha = 35°$, $\beta = -110°$, and $\gamma = 30°$. The physical units of the magnitudes are centimeters. Resolve the vectors to their scalar components and find the following vector sums: (a) $\vec{\mathbf{R}} = \vec{\mathbf{A}} + \vec{\mathbf{B}} + \vec{\mathbf{C}}$, (b) $\vec{\mathbf{D}} = \vec{\mathbf{A}} - \vec{\mathbf{B}}$, and (c) $\vec{\mathbf{S}} = \vec{\mathbf{A}} - 3\vec{\mathbf{B}} + \vec{\mathbf{C}}$.

Strategy

First, we use **Equation 5.17** to find the scalar components of each vector and then we express each vector in its vector component form given by **Equation 5.12**. Then, we use analytical methods of vector algebra to find the resultants.

Solution

We resolve the given vectors to their scalar components:

$$\begin{cases} A_x = A\cos\alpha = (10.0\text{ cm})\cos 35° = 8.19\text{ cm} \\ A_y = A\sin\alpha = (10.0\text{ cm})\sin 35° = 5.73\text{ cm} \end{cases}$$
$$\begin{cases} B_x = B\cos\beta = (7.0\text{ cm})\cos(-110°) = -2.39\text{ cm} \\ B_y = B\sin\beta = (7.0\text{ cm})\sin(-110°) = -6.58\text{ cm} \end{cases}.$$
$$\begin{cases} C_x = C\cos\gamma = (8.0\text{ cm})\cos 30° = 6.93\text{ cm} \\ C_y = C\sin\gamma = (8.0\text{ cm})\sin 30° = 4.00\text{ cm} \end{cases}$$

For (a) we may substitute directly into **Equation 5.24** to find the scalar components of the resultant:

$$\begin{cases} R_x = A_x + B_x + C_x = 8.19\text{ cm} - 2.39\text{ cm} + 6.93\text{ cm} = 12.73\text{ cm} \\ R_y = A_y + B_y + C_y = 5.73\text{ cm} - 6.58\text{ cm} + 4.00\text{ cm} = 3.15\text{ cm} \end{cases}.$$

Therefore, the resultant vector is $\vec{\mathbf{R}} = R_x\hat{\mathbf{i}} + R_y\hat{\mathbf{j}} = (12.7\hat{\mathbf{i}} + 3.1\hat{\mathbf{j}})\text{cm}$.

For (b), we may want to write the vector difference as

$$\vec{\mathbf{D}} = \vec{\mathbf{A}} - \vec{\mathbf{B}} = (A_x\hat{\mathbf{i}} + A_y\hat{\mathbf{j}}) - (B_x\hat{\mathbf{i}} + B_y\hat{\mathbf{j}}) = (A_x - B_x)\hat{\mathbf{i}} + (A_y - B_y)\hat{\mathbf{j}}.$$

Then, the scalar components of the vector difference are

$$\begin{cases} D_x = A_x - B_x = 8.19\text{ cm} - (-2.39\text{ cm}) = 10.58\text{ cm} \\ D_y = A_y - B_y = 5.73\text{ cm} - (-6.58\text{ cm}) = 12.31\text{ cm} \end{cases}.$$

Hence, the difference vector is $\vec{\mathbf{D}} = D_x\hat{\mathbf{i}} + D_y\hat{\mathbf{j}} = (10.6\hat{\mathbf{i}} + 12.3\hat{\mathbf{j}})\text{cm}$.

For (c), we can write vector $\vec{\mathbf{S}}$ in the following explicit form:

$$\begin{aligned} \vec{\mathbf{S}} &= \vec{\mathbf{A}} - 3\vec{\mathbf{B}} + \vec{\mathbf{C}} = (A_x\hat{\mathbf{i}} + A_y\hat{\mathbf{j}}) - 3(B_x\hat{\mathbf{i}} + B_y\hat{\mathbf{j}}) + (C_x\hat{\mathbf{i}} + C_y\hat{\mathbf{j}}) \\ &= (A_x - 3B_x + C_x)\hat{\mathbf{i}} + (A_y - 3B_y + C_y)\hat{\mathbf{j}}. \end{aligned}$$

Then, the scalar components of $\vec{\mathbf{S}}$ are

$$\begin{cases} S_x = A_x - 3B_x + C_x = 8.19\text{ cm} - 3(-2.39\text{ cm}) + 6.93\text{ cm} = 22.29\text{ cm} \\ S_y = A_y - 3B_y + C_y = 5.73\text{ cm} - 3(-6.58\text{ cm}) + 4.00\text{ cm} = 29.47\text{ cm} \end{cases}.$$

The vector is $\vec{\mathbf{S}} = S_x\hat{\mathbf{i}} + S_y\hat{\mathbf{j}} = (22.3\hat{\mathbf{i}} + 29.5\hat{\mathbf{j}})\text{cm}$.

Significance

Having found the vector components, we can illustrate the vectors by graphing or we can compute magnitudes and direction angles, as shown in **Figure 5.24**. Results for the magnitudes in (b) and (c) can be compared with results for the same problems obtained with the graphical method, shown in **Figure 5.14** and **Figure 5.15**. Notice that the analytical method produces exact results and its accuracy is not limited by the resolution of a ruler or a protractor, as it was with the graphical method used in **Example 5.2** for finding this same resultant.

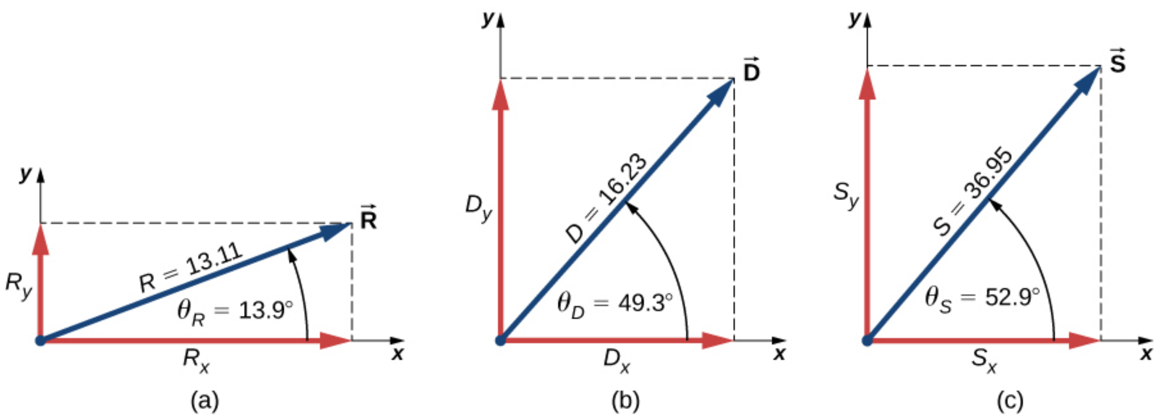


Figure 5.24 Graphical illustration of the solutions obtained analytically in **Example 5.9**.

5.8 Check Your Understanding Three displacement vectors $\vec{\mathbf{A}}$, $\vec{\mathbf{B}}$, and $\vec{\mathbf{F}}$ (**Figure 5.13**) are specified by their magnitudes A = 10.00, B = 7.00, and F = 20.00, respectively, and by their respective direction angles with the horizontal direction $\alpha = 35°$, $\beta = -110°$, and $\varphi = 110°$. The physical units of the magnitudes are centimeters. Use the analytical method to find vector $\vec{\mathbf{G}} = \vec{\mathbf{A}} + 2\vec{\mathbf{B}} - \vec{\mathbf{F}}$. Verify that G = 28.15 cm and that $\theta_G = -68.65°$.

Example 5.10

The Tug-of-War Game

Four dogs named Ang, Bing, Chang, and Dong play a tug-of-war game with a toy (**Figure 5.25**). Ang pulls on the toy in direction $\alpha = 55°$ south of east, Bing pulls in direction $\beta = 60°$ east of north, and Chang pulls in direction $\gamma = 55°$ west of north. Ang pulls strongly with 160.0 units of force (N), which we abbreviate as A = 160.0 N. Bing pulls even stronger than Ang with a force of magnitude B = 200.0 N, and Chang pulls with a force of magnitude C = 140.0 N. When Dong pulls on the toy in such a way that his force balances out the resultant of the other three forces, the toy does not move in any direction. With how big a force and in what direction must Dong pull on the toy for this to happen?

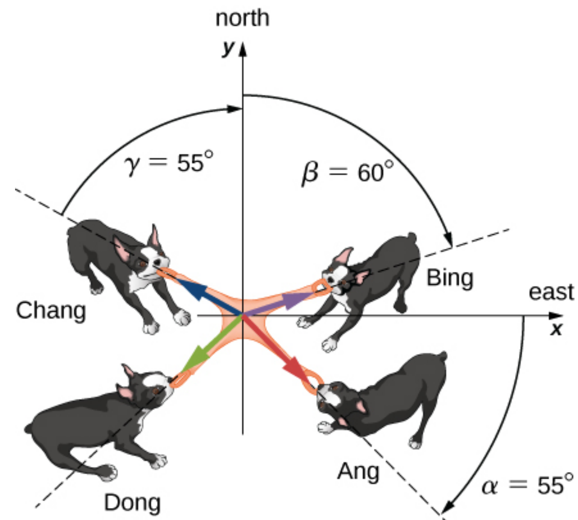


Figure 5.25 Four dogs play a tug-of-war game with a toy.

Strategy

We assume that east is the direction of the positive *x*-axis and north is the direction of the positive *y*-axis. As in **Example 5.9**, we have to resolve the three given forces— $\vec{\mathbf{A}}$ (the pull from Ang), $\vec{\mathbf{B}}$ (the pull from Bing), and $\vec{\mathbf{C}}$ (the pull from Chang)—into their scalar components and then find the scalar components of the resultant vector $\vec{\mathbf{R}} = \vec{\mathbf{A}} + \vec{\mathbf{B}} + \vec{\mathbf{C}}$. When the pulling force $\vec{\mathbf{D}}$ from Dong balances out this resultant, the sum of $\vec{\mathbf{D}}$ and $\vec{\mathbf{R}}$ must give the null vector $\vec{\mathbf{D}} + \vec{\mathbf{R}} = \vec{\mathbf{0}}$. This means that $\vec{\mathbf{D}} = -\vec{\mathbf{R}}$, so the pull from Dong must be antiparallel to $\vec{\mathbf{R}}$.

Solution

The direction angles are $\theta_A = -\alpha = -55°$, $\theta_B = 90° - \beta = 30°$, and $\theta_C = 90° + \gamma = 145°$, and substituting them into **Equation 5.17** gives the scalar components of the three given forces:

$$\begin{cases} A_x = A\cos\theta_A = (160.0\,\text{N})\cos(-55°) = +91.8\,\text{N} \\ A_y = A\sin\theta_A = (160.0\,\text{N})\sin(-55°) = -131.1\,\text{N} \end{cases}$$
$$\begin{cases} B_x = B\cos\theta_B = (200.0\,\text{N})\cos 30° = +173.2\,\text{N} \\ B_y = B\sin\theta_B = (200.0\,\text{N})\sin 30° = +100.0\,\text{N} \end{cases}.$$
$$\begin{cases} C_x = C\cos\theta_C = (140.0\,\text{N})\cos 145° = -114.7\,\text{N} \\ C_y = C\sin\theta_C = (140.0\,\text{N})\sin 145° = +80.3\,\text{N} \end{cases}$$

Now we compute scalar components of the resultant vector $\vec{\mathbf{R}} = \vec{\mathbf{A}} + \vec{\mathbf{B}} + \vec{\mathbf{C}}$:

$$\begin{cases} R_x = A_x + B_x + C_x = +91.8\,\text{N} + 173.2\,\text{N} - 114.7\,\text{N} = +150.3\,\text{N} \\ R_y = A_y + B_y + C_y = -131.1\,\text{N} + 100.0\,\text{N} + 80.3\,\text{N} = +49.2\,\text{N} \end{cases}.$$

The antiparallel vector to the resultant $\vec{\mathbf{R}}$ is

$$\vec{\mathbf{D}} = -\vec{\mathbf{R}} = -R_x\,\hat{\mathbf{i}} - R_y\,\hat{\mathbf{j}} = (-150.3\,\hat{\mathbf{i}} - 49.2\,\hat{\mathbf{j}})\,\text{N}.$$

The magnitude of Dong's pulling force is

$$D = \sqrt{D_x^2 + D_y^2} = \sqrt{(-150.3)^2 + (-49.2)^2}\,\text{N} = 158.1\,\text{N}.$$

The direction of Dong's pulling force is

$$\theta = \tan^{-1}\left(\frac{D_y}{D_x}\right) = \tan^{-1}\left(\frac{-49.2\,\text{N}}{-150.3\,\text{N}}\right) = \tan^{-1}\left(\frac{49.2}{150.3}\right) = 18.1°.$$

Dong pulls in the direction $18.1°$ south of west because both components are negative, which means the pull vector lies in the third quadrant (**Figure 5.19**).

5.9 Check Your Understanding Suppose that Bing in **Example 5.10** leaves the game to attend to more important matters, but Ang, Chang, and Dong continue playing. Ang and Chang's pull on the toy does not change, but Dong runs around and bites on the toy in a different place. With how big a force and in what direction must Dong pull on the toy now to balance out the combined pulls from Chang and Ang? Illustrate this situation by drawing a vector diagram indicating all forces involved.

Example 5.11

Vector Algebra

Find the magnitude of the vector $\vec{\mathbf{C}}$ that satisfies the equation $2\vec{\mathbf{A}} - 6\vec{\mathbf{B}} + 3\vec{\mathbf{C}} = 2\hat{\mathbf{j}}$, where $\vec{\mathbf{A}} = \hat{\mathbf{i}} - 2\hat{\mathbf{k}}$ and $\vec{\mathbf{B}} = -\hat{\mathbf{j}} + \hat{\mathbf{k}}/2$.

Strategy

We first solve the given equation for the unknown vector $\vec{\mathbf{C}}$. Then we substitute $\vec{\mathbf{A}}$ and $\vec{\mathbf{B}}$; group the terms along each of the three directions $\hat{\mathbf{i}}$, $\hat{\mathbf{j}}$, and $\hat{\mathbf{k}}$; and identify the scalar components C_x, C_y, and C_z. Finally, we substitute into **Equation 5.21** to find magnitude *C*.

Solution

$$\begin{aligned}
2\vec{\mathbf{A}} - 6\vec{\mathbf{B}} + 3\vec{\mathbf{C}} &= 2\hat{\mathbf{j}} \\
3\vec{\mathbf{C}} &= 2\hat{\mathbf{j}} - 2\vec{\mathbf{A}} + 6\vec{\mathbf{B}} \\
\vec{\mathbf{C}} &= \frac{2}{3}\hat{\mathbf{j}} - \frac{2}{3}\vec{\mathbf{A}} + 2\vec{\mathbf{B}} \\
&= \frac{2}{3}\hat{\mathbf{j}} - \frac{2}{3}(\hat{\mathbf{i}} - 2\hat{\mathbf{k}}) + 2\left(-\hat{\mathbf{j}} + \frac{\hat{\mathbf{k}}}{2}\right) = \frac{2}{3}\hat{\mathbf{j}} - \frac{2}{3}\hat{\mathbf{i}} + \frac{4}{3}\hat{\mathbf{k}} - 2\hat{\mathbf{j}} + \hat{\mathbf{k}} \\
&= -\frac{2}{3}\hat{\mathbf{i}} + \left(\frac{2}{3} - 2\right)\hat{\mathbf{j}} + \left(\frac{4}{3} + 1\right)\hat{\mathbf{k}} \\
&= -\frac{2}{3}\hat{\mathbf{i}} - \frac{4}{3}\hat{\mathbf{j}} + \frac{7}{3}\hat{\mathbf{k}}.
\end{aligned}$$

The components are $C_x = -2/3$, $C_y = -4/3$, and $C_z = 7/3$, and substituting into **Equation 5.21** gives

$$C = \sqrt{C_x^2 + C_y^2 + C_z^2} = \sqrt{(-2/3)^2 + (-4/3)^2 + (7/3)^2} = \sqrt{23/3}.$$

Example 5.12

Displacement of a Skier

Starting at a ski lodge, a cross-country skier goes 5.0 km north, then 3.0 km west, and finally 4.0 km southwest before taking a rest. Find his total displacement vector relative to the lodge when he is at the rest point. How far and in what direction must he ski from the rest point to return directly to the lodge?

Strategy

We assume a rectangular coordinate system with the origin at the ski lodge and with the unit vector $\hat{\mathbf{i}}$ pointing east and the unit vector $\hat{\mathbf{j}}$ pointing north. There are three displacements: $\vec{\mathbf{D}}_1$, $\vec{\mathbf{D}}_2$, and $\vec{\mathbf{D}}_3$. We identify their magnitudes as $D_1 = 5.0\,\text{km}$, $D_2 = 3.0\,\text{km}$, and $D_3 = 4.0\,\text{km}$. We identify their directions are the angles $\theta_1 = 90°$, $\theta_2 = 180°$, and $\theta_3 = 180° + 45° = 225°$. We resolve each displacement vector to its scalar components and substitute the components into **Equation 5.24** to obtain the scalar components of the resultant displacement $\vec{\mathbf{D}}$ from the lodge to the rest point. On the way back from the rest point to the lodge, the displacement is $\vec{\mathbf{B}} = -\vec{\mathbf{D}}$. Finally, we find the magnitude and direction of $\vec{\mathbf{B}}$.

Solution

Scalar components of the displacement vectors are

$$\begin{cases} D_{1x} = D_1 \cos\theta_1 = (5.0\,\text{km})\cos 90° = 0 \\ D_{1y} = D_1 \sin\theta_1 = (5.0\,\text{km})\sin 90° = 5.0\,\text{km} \end{cases}$$
$$\begin{cases} D_{2x} = D_2 \cos\theta_2 = (3.0\,\text{km})\cos 180° = -3.0\,\text{km} \\ D_{2y} = D_2 \sin\theta_2 = (3.0\,\text{km})\sin 180° = 0 \end{cases}.$$
$$\begin{cases} D_{3x} = D_3 \cos\theta_3 = (4.0\,\text{km})\cos 225° = -2.8\,\text{km} \\ D_{3y} = D_3 \sin\theta_3 = (4.0\,\text{km})\sin 225° = -2.8\,\text{km} \end{cases}$$

Scalar components of the net displacement vector are

$$\begin{cases} D_x = D_{1x} + D_{2x} + D_{3x} = (0 - 3.0 - 2.8)\text{km} = -5.8\,\text{km} \\ D_y = D_{1y} + D_{2y} + D_{3y} = (5.0 + 0 - 2.8)\text{km} = +2.2\,\text{km} \end{cases}.$$

Hence, the skier's net displacement vector is $\vec{\mathbf{D}} = D_x\hat{\mathbf{i}} + D_y\hat{\mathbf{j}} = (-5.8\hat{\mathbf{i}} + 2.2\hat{\mathbf{j}})\text{km}$. On the way back to the lodge, his displacement is $\vec{\mathbf{B}} = -\vec{\mathbf{D}} = -(-5.8\hat{\mathbf{i}} + 2.2\hat{\mathbf{j}})\text{km} = (5.8\hat{\mathbf{i}} - 2.2\hat{\mathbf{j}})\text{km}$. Its magnitude is $B = \sqrt{B_x^2 + B_y^2} = \sqrt{(5.8)^2 + (-2.2)^2}\,\text{km} = 6.2\,\text{km}$ and its direction angle is $\theta = \tan^{-1}(-2.2/5.8) = -20.8°$. Therefore, to return to the lodge, he must go 6.2 km in a direction about $21°$ south of east.

Significance

Notice that no figure is needed to solve this problem by the analytical method. Figures are required when using a graphical method; however, we can check if our solution makes sense by sketching it, which is a useful final step in solving any vector problem.

Example 5.13

Displacement of a Jogger

A jogger runs up a flight of 200 identical steps to the top of a hill and then runs along the top of the hill 50.0 m before he stops at a drinking fountain (**Figure 5.26**). His displacement vector from point *A* at the bottom of the steps to point *B* at the fountain is $\vec{\mathbf{D}}_{AB} = (-90.0\hat{\mathbf{i}} + 30.0\hat{\mathbf{j}})\text{m}$. What is the height and width of each step in the flight? What is the actual distance the jogger covers? If he makes a loop and returns to point *A*, what is his net displacement vector?

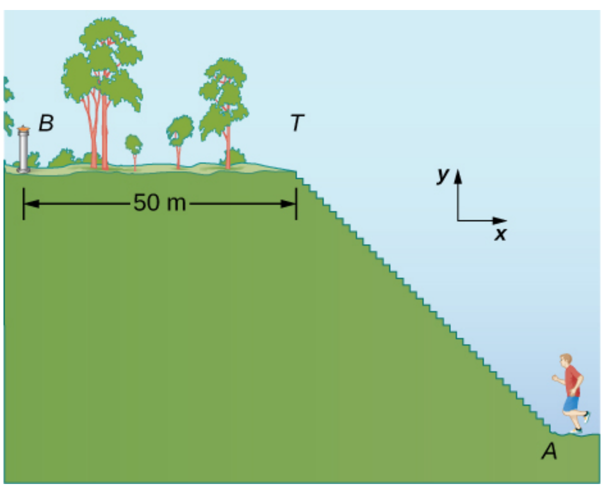


Figure 5.26 A jogger runs up a flight of steps.

Strategy

The displacement vector $\vec{\mathbf{D}}_{AB}$ is the vector sum of the jogger's displacement vector $\vec{\mathbf{D}}_{AT}$ along the stairs (from point *A* at the bottom of the stairs to point *T* at the top of the stairs) and his displacement vector $\vec{\mathbf{D}}_{TB}$ on the top of the hill (from point *T* at the top of the stairs to the fountain at point *B*). We must find the horizontal and the vertical components of $\vec{\mathbf{D}}_{TB}$. If each step has width *w* and height *h*, the horizontal component of $\vec{\mathbf{D}}_{TB}$ must have a length of 200*w* and the vertical component must have a length of 200*h*. The actual distance the jogger covers is the sum of the distance he runs up the stairs and the distance of 50.0 m that he runs along the top of the hill.

Solution

In the coordinate system indicated in **Figure 5.26**, the jogger's displacement vector on the top of the hill is $\vec{\mathbf{D}}_{TB} = (-50.0\,\text{m})\hat{\mathbf{i}}$. His net displacement vector is

$$\vec{\mathbf{D}}_{AB} = \vec{\mathbf{D}}_{AT} + \vec{\mathbf{D}}_{TB}.$$

Therefore, his displacement vector $\vec{\mathbf{D}}_{TB}$ along the stairs is

$$\begin{aligned}\vec{\mathbf{D}}_{AT} &= \vec{\mathbf{D}}_{AB} - \vec{\mathbf{D}}_{TB} = (-90.0\hat{\mathbf{i}} + 30.0\hat{\mathbf{j}})\text{m} - (-50.0\,\text{m})\hat{\mathbf{i}} = [(-90.0 + 50.0)\hat{\mathbf{i}} + 30.0\hat{\mathbf{j}})]\text{m} \\ &= (-40.0\hat{\mathbf{i}} + 30.0\hat{\mathbf{j}})\text{m}.\end{aligned}$$

Its scalar components are $D_{ATx} = -40.0\,\text{m}$ and $D_{ATy} = 30.0\,\text{m}$. Therefore, we must have

$$200w = |-40.0|\text{m and } 200h = 30.0\,\text{m}.$$

Hence, the step width is *w* = 40.0 m/200 = 0.2 m = 20 cm, and the step height is *w* = 30.0 m/200 = 0.15 m = 15 cm. The distance that the jogger covers along the stairs is

$$D_{AT} = \sqrt{D_{ATx}^2 + D_{ATy}^2} = \sqrt{(-40.0)^2 + (30.0)^2}\,\text{m} = 50.0\,\text{m}.$$

Thus, the actual distance he runs is $D_{AT} + D_{TB} = 50.0\,\text{m} + 50.0\,\text{m} = 100.0\,\text{m}$. When he makes a loop and comes back from the fountain to his initial position at point *A*, the total distance he covers is twice this distance, or 200.0 m. However, his net displacement vector is zero, because when his final position is the same as his initial position, the scalar components of his net displacement vector are zero (**Equation 5.13**).

In many physical situations, we often need to know the direction of a vector. For example, we may want to know the direction of a magnetic field vector at some point or the direction of motion of an object. We have already said direction is given by a unit vector, which is a dimensionless entity—that is, it has no physical units associated with it. When the vector in question lies along one of the axes in a Cartesian system of coordinates, the answer is simple, because then its unit vector of direction is either parallel or antiparallel to the direction of the unit vector of an axis. For example, the direction of vector $\vec{\mathbf{d}} = -5\,\text{m}\,\hat{\mathbf{i}}$ is unit vector $\hat{\mathbf{d}} = -\hat{\mathbf{i}}$. The general rule of finding the unit vector $\hat{\mathbf{V}}$ of direction for any vector $\vec{\mathbf{V}}$ is to divide it by its magnitude *V*:

$$\hat{\mathbf{V}} = \frac{\vec{\mathbf{V}}}{V}. \qquad (5.26)$$

We see from this expression that the unit vector of direction is indeed dimensionless because the numerator and the denominator in **Equation 5.26** have the same physical unit. In this way, **Equation 5.26** allows us to express the unit vector of direction in terms of unit vectors of the axes. The following example illustrates this principle.

Example 5.14

The Unit Vector of Direction

If the velocity vector of the military convoy in **Example 5.8** is $\vec{\mathbf{v}} = (4.000\,\hat{\mathbf{i}} + 3.000\,\hat{\mathbf{j}} + 0.100\,\hat{\mathbf{k}})\text{km/h}$, what is the unit vector of its direction of motion?

Strategy

The unit vector of the convoy's direction of motion is the unit vector $\hat{\mathbf{v}}$ that is parallel to the velocity vector. The unit vector is obtained by dividing a vector by its magnitude, in accordance with **Equation 5.26**.

Solution

The magnitude of the vector $\vec{\mathbf{v}}$ is

$$v = \sqrt{v_x^2 + v_y^2 + v_z^2} = \sqrt{4.000^2 + 3.000^2 + 0.100^2}\text{km/h} = 5.001\text{km/h}.$$

To obtain the unit vector $\hat{\mathbf{v}}$, divide $\vec{\mathbf{v}}$ by its magnitude:

$$\begin{aligned}
\hat{\mathbf{v}} &= \frac{\vec{\mathbf{v}}}{v} = \frac{(4.000\,\hat{\mathbf{i}} + 3.000\,\hat{\mathbf{j}} + 0.100\,\hat{\mathbf{k}})\text{km/h}}{5.001\text{km/h}} \\
&= \frac{(4.000\,\hat{\mathbf{i}} + 3.000\,\hat{\mathbf{j}} + 0.100\,\hat{\mathbf{k}})}{5.001} \\
&= \frac{4.000}{5.001}\hat{\mathbf{i}} + \frac{3.000}{5.001}\hat{\mathbf{j}} + \frac{0.100}{5.001}\hat{\mathbf{k}} \\
&= (79.98\,\hat{\mathbf{i}} + 59.99\,\hat{\mathbf{j}} + 2.00\,\hat{\mathbf{k}}) \times 10^{-2}.
\end{aligned}$$

Significance

Note that when using the analytical method with a calculator, it is advisable to carry out your calculations to at least three decimal places and then round off the final answer to the required number of significant figures, which is the way we performed calculations in this example. If you round off your partial answer too early, you risk your final answer having a huge numerical error, and it may be far off from the exact answer or from a value measured in an experiment.

5.10 **Check Your Understanding** Verify that vector $\hat{\mathbf{v}}$ obtained in **Example 5.14** is indeed a unit vector by computing its magnitude. If the convoy in **Example 5.8** was moving across a desert flatland—that is, if the third component of its velocity was zero—what is the unit vector of its direction of motion? Which geographic direction does it represent?

CHAPTER 5 REVIEW

KEY TERMS

antiparallel vectors two vectors with directions that differ by $180°$

associative terms can be grouped in any fashion

commutative operations can be performed in any order

component form of a vector a vector written as the vector sum of its components in terms of unit vectors

difference of two vectors vector sum of the first vector with the vector antiparallel to the second

direction angle in a plane, an angle between the positive direction of the *x*-axis and the vector, measured counterclockwise from the axis to the vector

displacement change in position

distributive multiplication can be distributed over terms in summation

equal vectors two vectors are equal if and only if all their corresponding components are equal; alternately, two parallel vectors of equal magnitudes

magnitude length of a vector

null vector a vector with all its components equal to zero

orthogonal vectors two vectors with directions that differ by exactly $90°$, synonymous with perpendicular vectors

parallel vectors two vectors with exactly the same direction angles

parallelogram rule geometric construction of the vector sum in a plane

polar coordinate system an orthogonal coordinate system where location in a plane is given by polar coordinates

polar coordinates a radial coordinate and an angle

radial coordinate distance to the origin in a polar coordinate system

resultant vector vector sum of two (or more) vectors

scalar a number, synonymous with a scalar quantity in physics

scalar component a number that multiplies a unit vector in a vector component of a vector

scalar equation equation in which the left-hand and right-hand sides are numbers

scalar quantity quantity that can be specified completely by a single number with an appropriate physical unit

tail-to-head geometric construction geometric construction for drawing the resultant vector of many vectors

unit vector vector of a unit magnitude that specifies direction; has no physical unit

unit vectors of the axes unit vectors that define orthogonal directions in a plane or in space

vector mathematical object with magnitude and direction

vector components orthogonal components of a vector; a vector is the vector sum of its vector components.

vector equation equation in which the left-hand and right-hand sides are vectors

vector quantity physical quantity described by a mathematical vector—that is, by specifying both its magnitude and its direction; synonymous with a vector in physics

vector sum resultant of the combination of two (or more) vectors

KEY EQUATIONS

Multiplication by a scalar (vector equation)	$\vec{\mathbf{B}} = \alpha \vec{\mathbf{A}}$

Multiplication by a scalar (scalar equation for magnitudes)	$B = \lvert\alpha\rvert A$
Resultant of two vectors	$\vec{\mathbf{D}}_{AD} = \vec{\mathbf{D}}_{AC} + \vec{\mathbf{D}}_{CD}$
Commutative law	$\vec{\mathbf{A}} + \vec{\mathbf{B}} = \vec{\mathbf{B}} + \vec{\mathbf{A}}$
Associative law	$(\vec{\mathbf{A}} + \vec{\mathbf{B}}) + \vec{\mathbf{C}} = \vec{\mathbf{A}} + (\vec{\mathbf{B}} + \vec{\mathbf{C}})$
Distributive law	$\alpha_1 \vec{\mathbf{A}} + \alpha_2 \vec{\mathbf{A}} = (\alpha_1 + \alpha_2)\vec{\mathbf{A}}$
The component form of a vector in two dimensions	$\vec{\mathbf{A}} = A_x \hat{\mathbf{i}} + A_y \hat{\mathbf{j}}$
Scalar components of a vector in two dimensions	$\begin{cases} A_x = x_e - x_b \\ A_y = y_e - y_b \end{cases}$
Magnitude of a vector in a plane	$A = \sqrt{A_x^2 + A_y^2}$
The direction angle of a vector in a plane	$\theta_A = \tan^{-1}\left(\frac{A_y}{A_x}\right)$
Scalar components of a vector in a plane	$\begin{cases} A_x = A\cos\theta_A \\ A_y = A\sin\theta_A \end{cases}$
Polar coordinates in a plane	$\begin{cases} x = r\cos\varphi \\ y = r\sin\varphi \end{cases}$
The component form of a vector in three dimensions	$\vec{\mathbf{A}} = A_x \hat{\mathbf{i}} + A_y \hat{\mathbf{j}} + A_z \hat{\mathbf{k}}$
The scalar *z*-component of a vector in three dimensions	$A_z = z_e - z_b$
Magnitude of a vector in three dimensions	$A = \sqrt{A_x^2 + A_y^2 + A_z^2}$
Distributive property	$\alpha(\vec{\mathbf{A}} + \vec{\mathbf{B}}) = \alpha\vec{\mathbf{A}} + \alpha\vec{\mathbf{B}}$
Antiparallel vector to $\vec{\mathbf{A}}$	$-\vec{\mathbf{A}} = -A_x \hat{\mathbf{i}} - A_y \hat{\mathbf{j}} - A_z \hat{\mathbf{k}}$
Equal vectors	$\vec{\mathbf{A}} = \vec{\mathbf{B}} \Leftrightarrow \begin{cases} A_x = B_x \\ A_y = B_y \\ A_z = B_z \end{cases}$
Components of the resultant of *N* vectors	$\begin{cases} F_{Rx} = \sum_{k=1}^{N} F_{kx} = F_{1x} + F_{2x} + \ldots + F_{Nx} \\ F_{Ry} = \sum_{k=1}^{N} F_{ky} = F_{1y} + F_{2y} + \ldots + F_{Ny} \\ F_{Rz} = \sum_{k=1}^{N} F_{kz} = F_{1z} + F_{2z} + \ldots + F_{Nz} \end{cases}$
General unit vector	$\hat{\mathbf{V}} = \frac{\vec{\mathbf{V}}}{V}$

SUMMARY

5.1 Scalars and Vectors

- A vector quantity is any quantity that has magnitude and direction, such as displacement or velocity. Vector quantities are represented by mathematical objects called vectors.
- Geometrically, vectors are represented by arrows, with the end marked by an arrowhead. The length of the vector is its magnitude, which is a positive scalar. On a plane, the direction of a vector is given by the angle the vector makes with a reference direction, often an angle with the horizontal. The direction angle of a vector is a scalar.
- Two vectors are equal if and only if they have the same magnitudes and directions. Parallel vectors have the same direction angles but may have different magnitudes. Antiparallel vectors have direction angles that differ by $180°$. Orthogonal vectors have direction angles that differ by $90°$.
- When a vector is multiplied by a scalar, the result is another vector of a different length than the length of the original vector. Multiplication by a positive scalar does not change the original direction; only the magnitude is affected. Multiplication by a negative scalar reverses the original direction. The resulting vector is antiparallel to the original vector. Multiplication by a scalar is distributive. Vectors can be divided by nonzero scalars but cannot be divided by vectors.
- Two or more vectors can be added to form another vector. The vector sum is called the resultant vector. We can add vectors to vectors or scalars to scalars, but we cannot add scalars to vectors. Vector addition is commutative and associative.
- To construct a resultant vector of two vectors in a plane geometrically, we use the parallelogram rule. To construct a resultant vector of many vectors in a plane geometrically, we use the tail-to-head method.

5.2 Coordinate Systems and Components of a Vector

- Vectors are described in terms of their components in a coordinate system. In two dimensions (in a plane), vectors have two components. In three dimensions (in space), vectors have three components.
- A vector component of a vector is its part in an axis direction. The vector component is the product of the unit vector of an axis with its scalar component along this axis. A vector is the resultant of its vector components.
- Scalar components of a vector are differences of coordinates, where coordinates of the origin are subtracted from end point coordinates of a vector. In a rectangular system, the magnitude of a vector is the square root of the sum of the squares of its components.
- In a plane, the direction of a vector is given by an angle the vector has with the positive *x*-axis. This direction angle is measured counterclockwise. The scalar *x*-component of a vector can be expressed as the product of its magnitude with the cosine of its direction angle, and the scalar *y*-component can be expressed as the product of its magnitude with the sine of its direction angle.
- In a plane, there are two equivalent coordinate systems. The Cartesian coordinate system is defined by unit vectors $\hat{\mathbf{i}}$ and $\hat{\mathbf{j}}$ along the *x*-axis and the *y*-axis, respectively. The polar coordinate system is defined by the radial unit vector $\hat{\mathbf{r}}$, which gives the direction from the origin, and a unit vector $\hat{\mathbf{t}}$, which is perpendicular (orthogonal) to the radial direction.

5.3 Algebra of Vectors

- Analytical methods of vector algebra allow us to find resultants of sums or differences of vectors without having to draw them. Analytical methods of vector addition are exact, contrary to graphical methods, which are approximate.
- Analytical methods of vector algebra are used routinely in mechanics, electricity, and magnetism. They are important mathematical tools of physics.

CONCEPTUAL QUESTIONS

5.1 Scalars and Vectors

1. A weather forecast states the temperature is predicted to be $-5\ °C$ the following day. Is this temperature a vector or a scalar quantity? Explain.

2. Which of the following is a vector: a person's height, the altitude on Mt. Everest, the velocity of a fly, the age of Earth, the boiling point of water, the cost of a book, Earth's population, or the acceleration of gravity?

3. Give a specific example of a vector, stating its magnitude, units, and direction.

4. What do vectors and scalars have in common? How do they differ?

5. Suppose you add two vectors $\vec{\mathbf{A}}$ and $\vec{\mathbf{B}}$. What relative direction between them produces the resultant with the greatest magnitude? What is the maximum magnitude? What relative direction between them produces the resultant with the smallest magnitude? What is the minimum magnitude?

6. Is it possible to add a scalar quantity to a vector quantity?

7. Is it possible for two vectors of different magnitudes to add to zero? Is it possible for three vectors of different magnitudes to add to zero? Explain.

8. Does the odometer in an automobile indicate a scalar or a vector quantity?

9. When a 10,000-m runner competing on a 400-m track crosses the finish line, what is the runner's net displacement? Can this displacement be zero? Explain.

10. A vector has zero magnitude. Is it necessary to specify its direction? Explain.

11. Can a magnitude of a vector be negative?

12. Can the magnitude of a particle's displacement be greater that the distance traveled?

13. If two vectors are equal, what can you say about their components? What can you say about their magnitudes? What can you say about their directions?

14. If three vectors sum up to zero, what geometric condition do they satisfy?

5.2 Coordinate Systems and Components of a Vector

15. Give an example of a nonzero vector that has a component of zero.

16. Explain why a vector cannot have a component greater than its own magnitude.

17. If two vectors are equal, what can you say about their components?

18. If vectors $\vec{\mathbf{A}}$ and $\vec{\mathbf{B}}$ are orthogonal, what is the component of $\vec{\mathbf{B}}$ along the direction of $\vec{\mathbf{A}}$? What is the component of $\vec{\mathbf{A}}$ along the direction of $\vec{\mathbf{B}}$?

19. If one of the two components of a vector is not zero, can the magnitude of the other vector component of this vector be zero?

20. If two vectors have the same magnitude, do their components have to be the same?

PROBLEMS

5.1 Scalars and Vectors

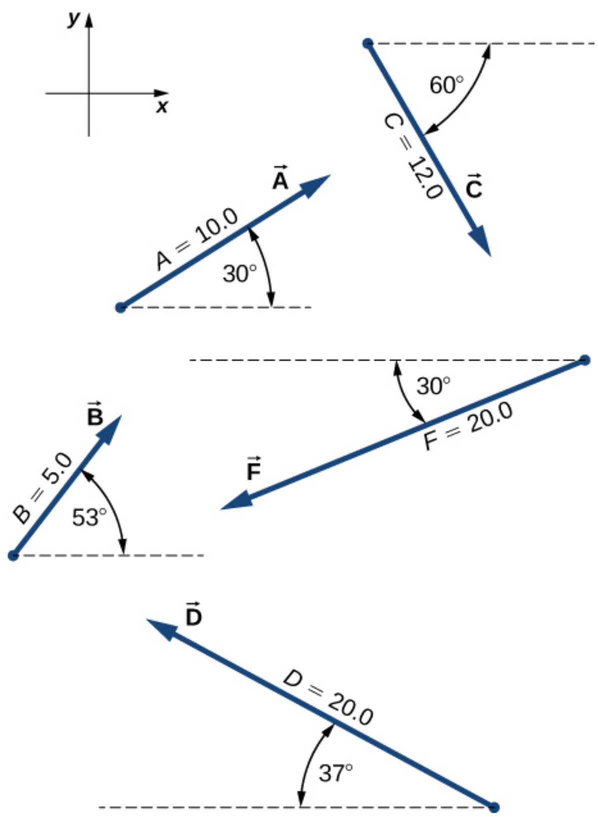
21. A scuba diver makes a slow descent into the depths of the ocean. His vertical position with respect to a boat on the surface changes several times. He makes the first stop 9.0 m from the boat but has a problem with equalizing the pressure, so he ascends 3.0 m and then continues descending for another 12.0 m to the second stop. From there, he ascends 4 m and then descends for 18.0 m, ascends again for 7 m and descends again for 24.0 m, where he makes a stop, waiting for his buddy. Assuming the positive direction up to the surface, express his net vertical displacement vector in terms of the unit vector. What is his distance to the boat?

22. In a tug-of-war game on one campus, 15 students pull on a rope at both ends in an effort to displace the central knot to one side or the other. Two students pull with force

196 N each to the right, four students pull with force 98 N each to the left, five students pull with force 62 N each to the left, three students pull with force 150 N each to the right, and one student pulls with force 250 N to the left. Assuming the positive direction to the right, express the net pull on the knot in terms of the unit vector. How big is the net pull on the knot? In what direction?

23. Suppose you walk 18.0 m straight west and then 25.0 m straight north. How far are you from your starting point and what is the compass direction of a line connecting your starting point to your final position? Use a graphical method.

24. For the vectors given in the following figure, use a graphical method to find the following resultants: (a) $\vec{\mathbf{A}} + \vec{\mathbf{B}}$, (b) $\vec{\mathbf{C}} + \vec{\mathbf{B}}$, (c) $\vec{\mathbf{D}} + \vec{\mathbf{F}}$, (d) $\vec{\mathbf{A}} - \vec{\mathbf{B}}$, (e) $\vec{\mathbf{D}} - \vec{\mathbf{F}}$, (f) $\vec{\mathbf{A}} + 2\vec{\mathbf{F}}$, (g) $\vec{\mathbf{C}} - 2\vec{\mathbf{D}} + 3\vec{\mathbf{F}}$; and (h) $\vec{\mathbf{A}} - 4\vec{\mathbf{D}} + 2\vec{\mathbf{F}}$.



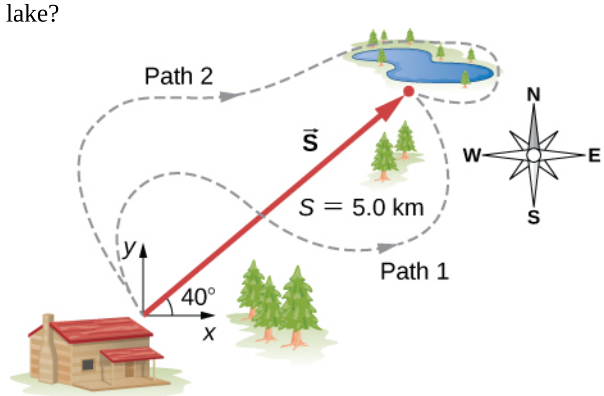
25. A delivery man starts at the post office, drives 40 km north, then 20 km west, then 60 km northeast, and finally 50 km north to stop for lunch. Use a graphical method to find his net displacement vector.

26. An adventurous dog strays from home, runs three blocks east, two blocks north, one block east, one block north, and two blocks west. Assuming that each block is about 100 m, how far from home and in what direction is the dog? Use a graphical method.

27. In an attempt to escape a desert island, a castaway builds a raft and sets out to sea. The wind shifts a great deal during the day and he is blown along the following directions: 2.50 km and $45.0°$ north of west, then 4.70 km and $60.0°$ south of east, then 1.30 km and $25.0°$ south of west, then 5.10 km straight east, then 1.70 km and $5.00°$ east of north, then 7.20 km and $55.0°$ south of west, and finally 2.80 km and $10.0°$ north of east. Use a graphical method to find the castaway's final position relative to the island.

28. A small plane flies 40.0 km in a direction $60°$ north of east and then flies 30.0 km in a direction $15°$ north of east. Use a graphical method to find the total distance the plane covers from the starting point and the direction of the path to the final position.

29. A trapper walks a 5.0-km straight-line distance from his cabin to the lake, as shown in the following figure. Use a graphical method (the parallelogram rule) to determine the trapper's displacement directly to the east and displacement directly to the north that sum up to his resultant displacement vector. If the trapper walked only in directions east and north, zigzagging his way to the lake, how many kilometers would he have to walk to get to the lake?



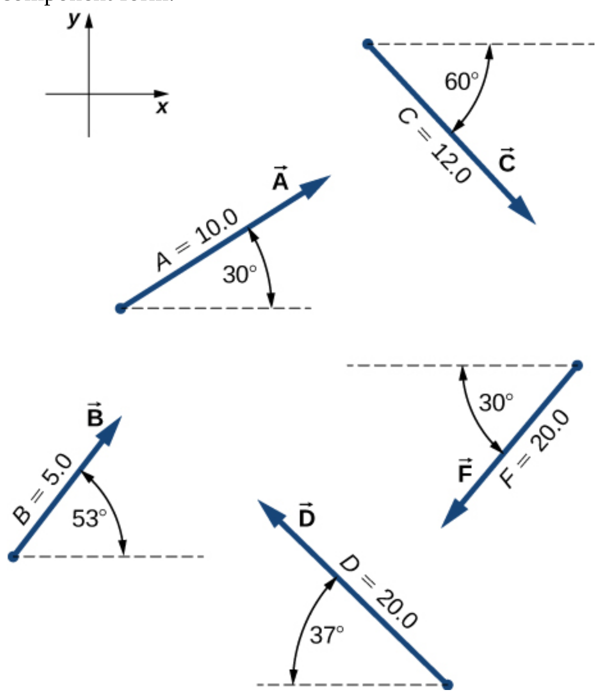
30. A surveyor measures the distance across a river that flows straight north by the following method. Starting directly across from a tree on the opposite bank, the surveyor walks 100 m along the river to establish a baseline. She then sights across to the tree and reads that the angle from the baseline to the tree is $35°$. How wide is the river?

31. A pedestrian walks 6.0 km east and then 13.0 km north. Use a graphical method to find the pedestrian's resultant displacement and geographic direction.

32. The magnitudes of two displacement vectors are A = 20 m and B = 6 m. What are the largest and the smallest values of the magnitude of the resultant $\vec{\mathbf{R}} = \vec{\mathbf{A}} + \vec{\mathbf{B}}$?

5.2 Coordinate Systems and Components of a Vector

33. Assuming the +*x*-axis is horizontal and points to the right, resolve the vectors given in the following figure to their scalar components and express them in vector component form.

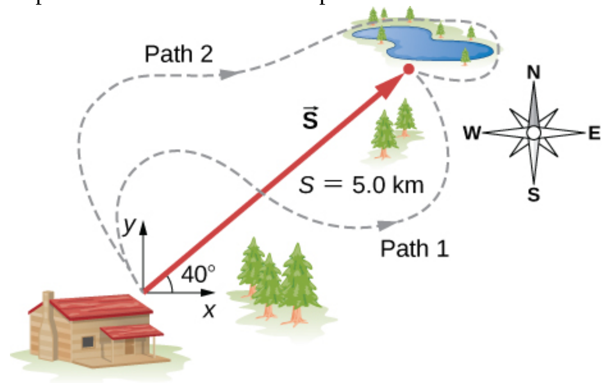


34. Suppose you walk 18.0 m straight west and then 25.0 m straight north. How far are you from your starting point? What is your displacement vector? What is the direction of your displacement? Assume the +*x*-axis is horizontal to the right.

35. You drive 7.50 km in a straight line in a direction $15°$ east of north. (a) Find the distances you would have to drive straight east and then straight north to arrive at the same point. (b) Show that you still arrive at the same point if the east and north legs are reversed in order. Assume the +*x*-axis is to the east.

36. A sledge is being pulled by two horses on a flat terrain. The net force on the sledge can be expressed in the Cartesian coordinate system as vector $\vec{\mathbf{F}} = (-2980.0\,\hat{\mathbf{i}} + 8200.0\,\hat{\mathbf{j}})\text{N}$, where $\hat{\mathbf{i}}$ and $\hat{\mathbf{j}}$ denote directions to the east and north, respectively. Find the magnitude and direction of the pull.

37. A trapper walks a 5.0-km straight-line distance from her cabin to the lake, as shown in the following figure. Determine the east and north components of her displacement vector. How many more kilometers would she have to walk if she walked along the component displacements? What is her displacement vector?



38. The polar coordinates of a point are $4\pi/3$ and 5.50 m. What are its Cartesian coordinates?

39. Two points in a plane have polar coordinates $P_1(2.500\text{ m}, \pi/6)$ and $P_2(3.800\text{ m}, 2\pi/3)$. Determine their Cartesian coordinates and the distance between them in the Cartesian coordinate system. Round the distance to a nearest centimeter.

40. A chameleon is resting quietly on a lanai screen, waiting for an insect to come by. Assume the origin of a Cartesian coordinate system at the lower left-hand corner of the screen and the horizontal direction to the right as the +*x*-direction. If its coordinates are (2.000 m, 1.000 m), (a) how far is it from the corner of the screen? (b) What is its location in polar coordinates?

41. Two points in the Cartesian plane are *A*(2.00 m, −4.00 m) and *B*(−3.00 m, 3.00 m). Find the distance between them and their polar coordinates.

42. A fly enters through an open window and zooms around the room. In a Cartesian coordinate system with three axes along three edges of the room, the fly changes its position from point *b*(4.0 m, 1.5 m, 2.5 m) to point *e*(1.0 m, 4.5 m, 0.5 m). Find the scalar components of the fly's displacement vector and express its displacement vector in vector component form. What is its magnitude?

5.3 Algebra of Vectors

43. For vectors $\vec{\mathbf{B}} = -\hat{\mathbf{i}} - 4\,\hat{\mathbf{j}}$ and $\vec{\mathbf{A}} = -3\,\hat{\mathbf{i}} - 2\,\hat{\mathbf{j}}$, calculate (a) $\vec{\mathbf{A}} + \vec{\mathbf{B}}$ and its magnitude and direction angle, and (b) $\vec{\mathbf{A}} - \vec{\mathbf{B}}$ and its magnitude and direction angle.

44. A particle undergoes three consecutive displacements given by vectors $\vec{\mathbf{D}}_1 = (3.0\,\hat{\mathbf{i}} - 4.0\,\hat{\mathbf{j}} - 2.0\,\hat{\mathbf{k}})\text{mm}$,

$\vec{\mathbf{D}}_2 = (1.0\hat{\mathbf{i}} - 7.0\hat{\mathbf{j}} + 4.0\hat{\mathbf{k}})\text{mm}$, and $\vec{\mathbf{D}}_3 = (-7.0\hat{\mathbf{i}} + 4.0\hat{\mathbf{j}} + 1.0\hat{\mathbf{k}})\text{mm}$. (a) Find the resultant displacement vector of the particle. (b) What is the magnitude of the resultant displacement? (c) If all displacements were along one line, how far would the particle travel?

45. Given two displacement vectors $\vec{\mathbf{A}} = (3.00\hat{\mathbf{i}} - 4.00\hat{\mathbf{j}} + 4.00\hat{\mathbf{k}})\text{m}$ and $\vec{\mathbf{B}} = (2.00\hat{\mathbf{i}} + 3.00\hat{\mathbf{j}} - 7.00\hat{\mathbf{k}})\text{m}$, find the displacements and their magnitudes for (a) $\vec{\mathbf{C}} = \vec{\mathbf{A}} + \vec{\mathbf{B}}$ and (b) $\vec{\mathbf{D}} = 2\vec{\mathbf{A}} - \vec{\mathbf{B}}$.

46. A small plane flies $40.0\,\text{km}$ in a direction $60°$ north of east and then flies $30.0\,\text{km}$ in a direction $15°$ north of east. Use the analytical method to find the total distance the plane covers from the starting point, and the geographic direction of its displacement vector. What is its displacement vector?

47. In an attempt to escape a desert island, a castaway builds a raft and sets out to sea. The wind shifts a great deal during the day, and she is blown along the following straight lines: 2.50 km and $45.0°$ north of west, then 4.70 km and $60.0°$ south of east, then 1.30 km and $25.0°$ south of west, then 5.10 km due east, then 1.70 km and $5.00°$ east of north, then 7.20 km and $55.0°$ south of west, and finally 2.80 km and $10.0°$ north of east. Use the analytical method to find the resultant vector of all her displacement vectors. What is its magnitude and direction?

48. Assuming the +*x*-axis is horizontal to the right for the vectors given in the following figure, use the analytical method to find the following resultants: (a) $\vec{\mathbf{A}} + \vec{\mathbf{B}}$, (b) $\vec{\mathbf{C}} + \vec{\mathbf{B}}$, (c) $\vec{\mathbf{D}} + \vec{\mathbf{F}}$, (d) $\vec{\mathbf{A}} - \vec{\mathbf{B}}$, (e) $\vec{\mathbf{D}} - \vec{\mathbf{F}}$, (f) $\vec{\mathbf{A}} + 2\vec{\mathbf{F}}$, (g) $\vec{\mathbf{C}} - 2\vec{\mathbf{D}} + 3\vec{\mathbf{F}}$, and (h) $\vec{\mathbf{A}} - 4\vec{\mathbf{D}} + 2\vec{\mathbf{F}}$.

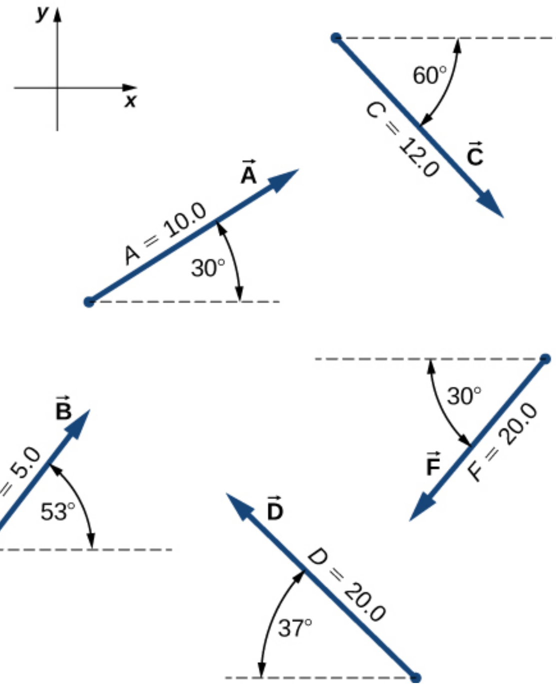


Figure 5.27

49. Given the vectors in the preceding figure, find vector $\vec{\mathbf{R}}$ that solves equations (a) $\vec{\mathbf{D}} + \vec{\mathbf{R}} = \vec{\mathbf{F}}$ and (b) $\vec{\mathbf{C}} - 2\vec{\mathbf{D}} + 5\vec{\mathbf{R}} = 3\vec{\mathbf{F}}$. Assume the +*x*-axis is horizontal to the right.

50. A delivery man starts at the post office, drives 40 km north, then 20 km west, then 60 km northeast, and finally 50 km north to stop for lunch. Use the analytical method to determine the following: (a) Find his net displacement vector. (b) How far is the restaurant from the post office? (c) If he returns directly from the restaurant to the post office, what is his displacement vector on the return trip? (d) What is his compass heading on the return trip? Assume the +*x*-axis is to the east.

51. An adventurous dog strays from home, runs three blocks east, two blocks north, and one block east, one block north, and two blocks west. Assuming that each block is about a 100 yd, use the analytical method to find the dog's net displacement vector, its magnitude, and its direction. Assume the +*x*-axis is to the east. How would your answer be affected if each block was about 100 m?

52. If $\vec{\mathbf{D}} = (6.00\hat{\mathbf{i}} - 8.00\hat{\mathbf{j}})\text{m}$, $\vec{\mathbf{B}} = (-8.00\hat{\mathbf{i}} + 3.00\hat{\mathbf{j}})\text{m}$, and $\vec{\mathbf{A}} = (26.0\hat{\mathbf{i}} + 19.0\hat{\mathbf{j}})\text{m}$, find the unknown constants *a* and *b* such that $a\vec{\mathbf{D}} + b\vec{\mathbf{B}} + \vec{\mathbf{A}} = \vec{\mathbf{0}}$.

53. Given the displacement vector $\vec{\mathbf{D}} = (3\hat{\mathbf{i}} - 4\hat{\mathbf{j}})$m, find the displacement vector $\vec{\mathbf{R}}$ so that $\vec{\mathbf{D}} + \vec{\mathbf{R}} = -4D\hat{\mathbf{j}}$.

54. Find the unit vector of direction for the following vector quantities: (a) Force $\vec{\mathbf{F}} = (3.0\hat{\mathbf{i}} - 2.0\hat{\mathbf{j}})\text{N}$, (b) displacement $\vec{\mathbf{D}} = (-3.0\hat{\mathbf{i}} - 4.0\hat{\mathbf{j}})\text{m}$, and (c) velocity $\vec{\mathbf{v}} = (-5.00\hat{\mathbf{i}} + 4.00\hat{\mathbf{j}})\text{m/s}$.

55. At one point in space, the direction of the electric field vector is given in the Cartesian system by the unit vector $\hat{\mathbf{E}} = 1/\sqrt{5}\,\hat{\mathbf{i}} - 2/\sqrt{5}\,\hat{\mathbf{j}}$. If the magnitude of the electric field vector is E = 400.0 V/m, what are the scalar components E_x, E_y, and E_z of the electric field vector $\vec{E}$ at this point? What is the direction angle θ_E of the electric field vector at this point?

56. A barge is pulled by the two tugboats shown in the following figure. One tugboat pulls on the barge with a force of magnitude 4000 units of force at $15°$ above the line AB (see the figure and the other tugboat pulls on the barge with a force of magnitude 5000 units of force at $12°$ below the line AB. Resolve the pulling forces to their scalar components and find the components of the resultant force pulling on the barge. What is the magnitude of the resultant pull? What is its direction relative to the line AB?

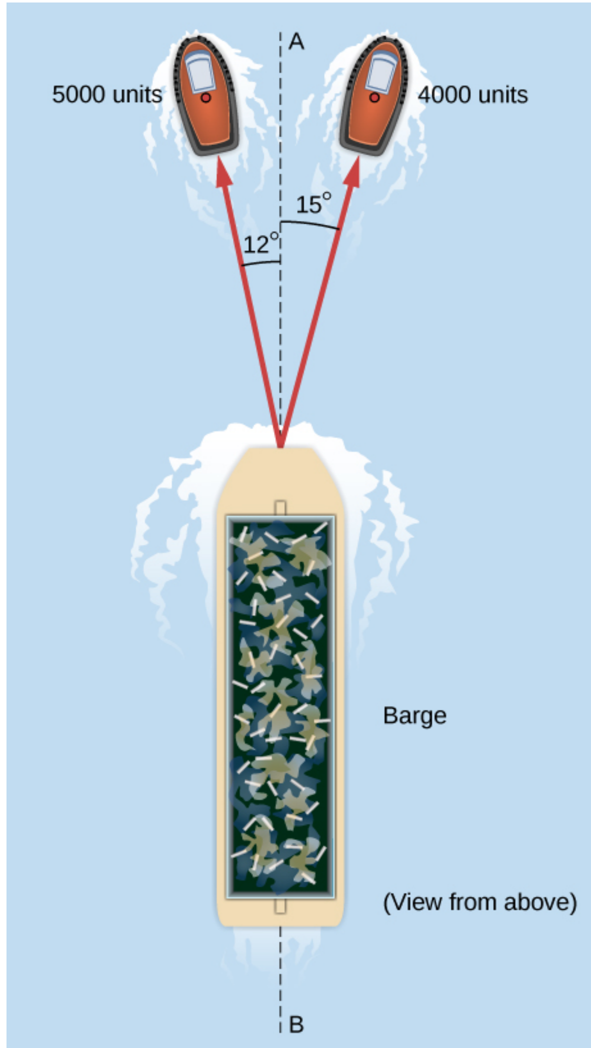


Figure 5.28

57. In the control tower at a regional airport, an air traffic controller monitors two aircraft as their positions change with respect to the control tower. One plane is a cargo carrier Boeing 747 and the other plane is a Douglas DC-3. The Boeing is at an altitude of 2500 m, climbing at $10°$ above the horizontal, and moving $30°$ north of west. The DC-3 is at an altitude of 3000 m, climbing at $5°$ above the horizontal, and cruising directly west. (a) Find the position vectors of the planes relative to the control tower. (b) What is the distance between the planes at the moment the air traffic controller makes a note about their positions?

ADDITIONAL PROBLEMS

58. You fly $32.0\,\text{km}$ in a straight line in still air in the direction $35.0°$ south of west. (a) Find the distances you would have to fly due south and then due west to arrive at the same point. (b) Find the distances you would have to fly first in a direction $45.0°$ south of west and then in a direction $45.0°$ west of north. Note these are the components of the displacement along a different set of axes—namely, the one rotated by $45°$ with respect to the axes in (a).

59. Rectangular coordinates of a point are given by (2, *y*) and its polar coordinates are given by $(r,\ \pi/6)$. Find *y* and *r*.

60. If the polar coordinates of a point are $(r,\ \varphi)$ and its rectangular coordinates are $(x,\ y)$, determine the polar coordinates of the following points: (a) (−*x*, *y*), (b) (−2*x*, −2*y*), and (c) (3*x*, −3*y*).

61. Vectors $\vec{\mathbf{A}}$ and $\vec{\mathbf{B}}$ have identical magnitudes of 5.0 units. Find the angle between them if $\vec{\mathbf{A}} + \vec{\mathbf{B}} = 5\sqrt{2}\,\hat{\mathbf{j}}$.

62. Starting at the island of Moi in an unknown archipelago, a fishing boat makes a round trip with two stops at the islands of Noi and Poi. It sails from Moi for 4.76 nautical miles (nmi) in a direction $37°$ north of east to Noi. From Noi, it sails $69°$ west of north to Poi. On its return leg from Poi, it sails $28°$ east of south. What distance does the boat sail between Noi and Poi? What distance does it sail between Moi and Poi? Express your answer both in nautical miles and in kilometers. Note: 1 nmi = 1852 m.

63. An air traffic controller notices two signals from two planes on the radar monitor. One plane is at altitude 800 m and in a 19.2-km horizontal distance to the tower in a direction $25°$ south of west. The second plane is at altitude 1100 m and its horizontal distance is 17.6 km and $20°$ south of west. What is the distance between these planes?

64. Show that when $\vec{\mathbf{A}} + \vec{\mathbf{B}} = \vec{\mathbf{C}}$, then $C^2 = A^2 + B^2 + 2AB\cos\varphi$, where φ is the angle between vectors $\vec{\mathbf{A}}$ and $\vec{\mathbf{B}}$.

65. Four force vectors each have the same magnitude *f*. What is the largest magnitude the resultant force vector may have when these forces are added? What is the smallest magnitude of the resultant? Make a graph of both situations.

66. A skater glides along a circular path of radius 5.00 m in clockwise direction. When he coasts around one-half of the circle, starting from the west point, find (a) the magnitude of his displacement vector and (b) how far he actually skated. (c) What is the magnitude of his displacement vector when he skates all the way around the circle and comes back to the west point?

67. A stubborn dog is being walked on a leash by its owner. At one point, the dog encounters an interesting scent at some spot on the ground and wants to explore it in detail, but the owner gets impatient and pulls on the leash with force $\vec{\mathbf{F}} = (98.0\,\hat{\mathbf{i}} + 132.0\,\hat{\mathbf{j}} + 32.0\,\hat{\mathbf{k}})\text{N}$ along the leash. (a) What is the magnitude of the pulling force? (b) What angle does the leash make with the vertical?

68. If the velocity vector of a polar bear is $\vec{\mathbf{u}} = (-18.0\,\hat{\mathbf{i}} - 13.0\,\hat{\mathbf{j}})\text{km/h}$, how fast and in what geographic direction is it heading? Here, $\hat{\mathbf{i}}$ and $\hat{\mathbf{j}}$ are directions to geographic east and north, respectively.

69. Find the scalar components of three-dimensional vectors $\vec{\mathbf{G}}$ and $\vec{\mathbf{H}}$ in the following figure and write the vectors in vector component form in terms of the unit vectors of the axes.
Vector G has magnitude 10.0. Its projection in the x y plane is between the positive x and positive y directions, at an angle of 45 degrees from the positive x direction. The angle between vector G and the positive z direction is 60 degrees. Vector H has magnitude 15.0. Its projection in the x y plane is between the negative x and positive y directions, at an angle of 30 degrees from the positive y direction. The angle between vector H and the positive z direction is 450 degrees.

70. A diver explores a shallow reef off the coast of Belize. She initially swims 90.0 m north, makes a turn to the east and continues for 200.0 m, then follows a big grouper for 80.0 m in the direction $30°$ north of east. In the meantime, a local current displaces her by 150.0 m south. Assuming the current is no longer present, in what direction and how far should she now swim to come back to the point where she started?

71. A force vector $\vec{\mathbf{A}}$ has *x*- and *y*-components, respectively, of −8.80 units of force and 15.00 units of force. The *x*- and *y*-components of force vector $\vec{\mathbf{B}}$ are, respectively, 13.20 units of force and −6.60 units of force. Find the components of force vector $\vec{\mathbf{C}}$ that satisfies the vector equation $\vec{\mathbf{A}} - \vec{\mathbf{B}} + 3\vec{\mathbf{C}} = 0$.

6 | MOTION IN TWO AND THREE DIMENSIONS

Figure 6.1 The Red Arrows is the aerobatics display team of Britain's Royal Air Force. Based in Lincolnshire, England, they perform precision flying shows at high speeds, which requires accurate measurement of position, velocity, and acceleration in three dimensions. (credit: modification of work by Phil Long)

Chapter Outline

Introduction

To give a complete description of kinematics, we must explore motion in two and three dimensions. After all, most objects in our universe do not move in straight lines; rather, they follow curved paths. From kicked footballs to the flight paths of birds to the orbital motions of celestial bodies and down to the flow of blood plasma in your veins, most motion follows curved trajectories.

Fortunately, the treatment of motion in one dimension in the previous chapter has given us a foundation on which to build, as the concepts of position, displacement, velocity, and acceleration defined in one dimension can be expanded to two and three dimensions. Consider the Red Arrows, also known as the Royal Air Force Aerobatic team of the United Kingdom. Each jet follows a unique curved trajectory in three-dimensional airspace, as well as has a unique velocity and acceleration. Thus, to describe the motion of any of the jets accurately, we must assign to each jet a unique position vector in three dimensions as well as a unique velocity and acceleration vector. We can apply the same basic equations for displacement, velocity, and acceleration we derived in **Motion Along a Straight Line** to describe the motion of the jets in two and three dimensions, but with some modifications—in particular, the inclusion of vectors.

In this chapter we also explore two special types of motion in two dimensions: projectile motion and circular motion. Last, we conclude with a discussion of relative motion. In the chapter-opening picture, each jet has a relative motion with respect to any other jet in the group or to the people observing the air show on the ground.

6.1 | Displacement and Velocity Vectors

Learning Objectives

By the end of this section, you will be able to:

- Calculate position vectors in a multidimensional displacement problem.
- Solve for the displacement in two or three dimensions.
- Calculate the velocity vector given the position vector as a function of time.
- Calculate the average velocity in multiple dimensions.

Displacement and velocity in two or three dimensions are straightforward extensions of the one-dimensional definitions. However, now they are vector quantities, so calculations with them have to follow the rules of vector algebra, not scalar algebra.

Displacement Vector

To describe motion in two and three dimensions, we must first establish a coordinate system and a convention for the axes. We generally use the coordinates *x*, *y*, and *z* to locate a particle at point *P*(*x*, *y*, *z*) in three dimensions. If the particle is moving, the variables *x*, *y*, and *z* are functions of time (*t*):

$$x = x(t) \quad y = y(t) \quad z = z(t). \tag{6.1}$$

The **position vector** from the origin of the coordinate system to point *P* is $\vec{\mathbf{r}}(t)$. In unit vector notation, introduced in **Coordinate Systems and Components of a Vector**, $\vec{\mathbf{r}}(t)$ is

$$\vec{\mathbf{r}}(t) = x(t)\hat{\mathbf{i}} + y(t)\hat{\mathbf{j}} + z(t)\hat{\mathbf{k}}. \tag{6.2}$$

Figure 6.2 shows the coordinate system and the vector to point *P*, where a particle could be located at a particular time *t*. Note the orientation of the *x*, *y*, and *z* axes. This orientation is called a right-handed coordinate system (**Coordinate Systems and Components of a Vector**) and it is used throughout the chapter.

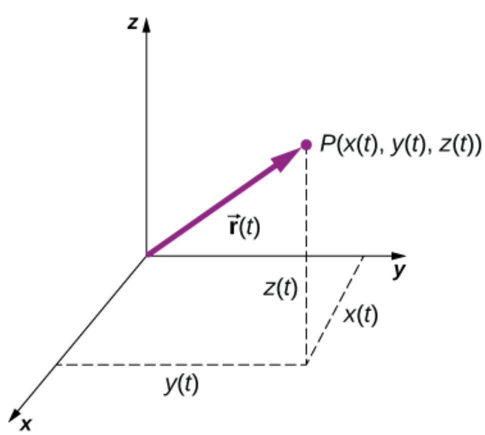


Figure 6.2 A three-dimensional coordinate system with a particle at position *P*(*x*(*t*), *y*(*t*), *z*(*t*)).

With our definition of the position of a particle in three-dimensional space, we can formulate the three-dimensional displacement. **Figure 6.3** shows a particle at time t_1 located at P_1 with position vector $\vec{\mathbf{r}}(t_1)$. At a later time t_2, the particle is located at P_2 with position vector $\vec{\mathbf{r}}(t_2)$. The **displacement vector** $\Delta\vec{\mathbf{r}}$ is found by subtracting $\vec{\mathbf{r}}(t_1)$ from $\vec{\mathbf{r}}(t_2)$:

$$\Delta\vec{\mathbf{r}} = \vec{\mathbf{r}}(t_2) - \vec{\mathbf{r}}(t_1). \tag{6.3}$$

Vector addition is discussed in **Algebra of Vectors**. Note that this is the same operation we did in one dimension, but now the vectors are in three-dimensional space.

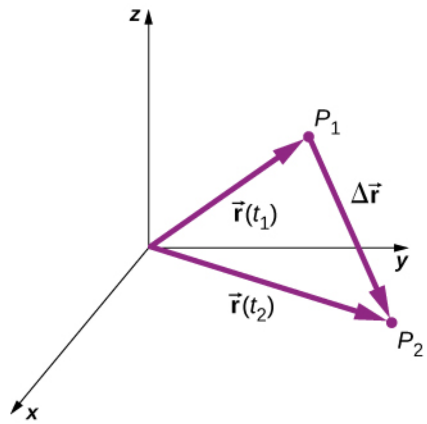


Figure 6.3 The displacement $\Delta\vec{\mathbf{r}} = \vec{\mathbf{r}}(t_2) - \vec{\mathbf{r}}(t_1)$ is the vector from P_1 to P_2.

The following examples illustrate the concept of displacement in multiple dimensions.

Example 6.1

Polar Orbiting Satellite

A satellite is in a circular polar orbit around Earth at an altitude of 400 km—meaning, it passes directly overhead at the North and South Poles. What is the magnitude and direction of the displacement vector from when it is directly over the North Pole to when it is at $-45°$ latitude?

Strategy

We make a picture of the problem to visualize the solution graphically. This will aid in our understanding of the displacement. We then use unit vectors to solve for the displacement.

Solution

Figure 6.4 shows the surface of Earth and a circle that represents the orbit of the satellite. Although satellites are moving in three-dimensional space, they follow trajectories of ellipses, which can be graphed in two dimensions. The position vectors are drawn from the center of Earth, which we take to be the origin of the coordinate system, with the y-axis as north and the x-axis as east. The vector between them is the displacement of the satellite. We take the radius of Earth as 6370 km, so the length of each position vector is 6770 km.

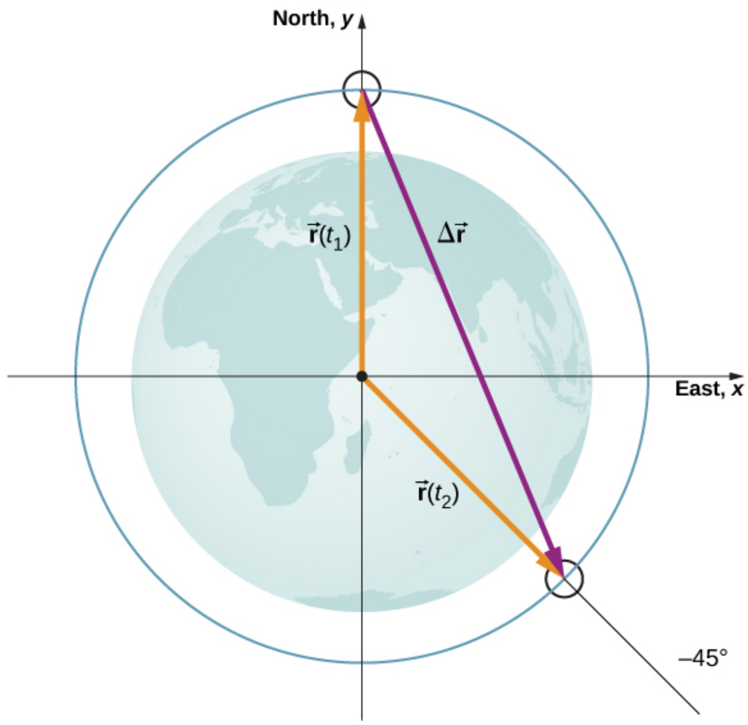


Figure 6.4 Two position vectors are drawn from the center of Earth, which is the origin of the coordinate system, with the *y*-axis as north and the *x*-axis as east. The vector between them is the displacement of the satellite.

In unit vector notation, the position vectors are

$$\vec{\mathbf{r}}(t_1) = 6770.\text{ km}\,\hat{\mathbf{j}}$$
$$\vec{\mathbf{r}}(t_2) = 6770.\text{ km}\,(\cos 45°)\,\hat{\mathbf{i}} + 6770.\text{ km}\,(\sin(-45°))\,\hat{\mathbf{j}}.$$

Evaluating the sine and cosine, we have

$$\begin{aligned} \vec{\mathbf{r}}(t_1) &= 6770.\,\hat{\mathbf{j}} \\ \vec{\mathbf{r}}(t_2) &= 4787\,\hat{\mathbf{i}} - 4787\,\hat{\mathbf{j}}. \end{aligned}$$

Now we can find $\Delta\vec{\mathbf{r}}$, the displacement of the satellite:

$$\Delta\vec{\mathbf{r}} = \vec{\mathbf{r}}(t_2) - \vec{\mathbf{r}}(t_1) = 4787\,\hat{\mathbf{i}} - 11{,}557\,\hat{\mathbf{j}}.$$

The magnitude of the displacement is $|\Delta\vec{\mathbf{r}}| = \sqrt{(4787)^2 + (-11{,}557)^2} = 12{,}509\text{ km}.$ The angle the displacement makes with the *x*-axis is $\theta = \tan^{-1}\left(\frac{-11{,}557}{4787}\right) = -67.5°.$

Significance

Plotting the displacement gives information and meaning to the unit vector solution to the problem. When plotting the displacement, we need to include its components as well as its magnitude and the angle it makes with a chosen axis—in this case, the *x*-axis (**Figure 6.5**).

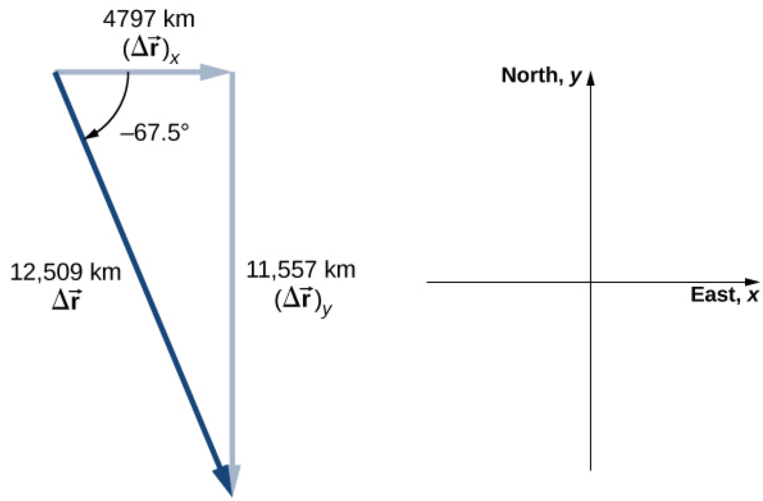


Figure 6.5 Displacement vector with components, angle, and magnitude.

Note that the satellite took a curved path along its circular orbit to get from its initial position to its final position in this example. It also could have traveled 4787 km east, then 11,557 km south to arrive at the same location. Both of these paths are longer than the length of the displacement vector. In fact, the displacement vector gives the shortest path between two points in one, two, or three dimensions.

Many applications in physics can have a series of displacements, as discussed in the previous chapter. The total displacement is the sum of the individual displacements, only this time, we need to be careful, because we are adding vectors. We illustrate this concept with an example of Brownian motion.

Example 6.2

Brownian Motion

Brownian motion is a chaotic random motion of particles suspended in a fluid, resulting from collisions with the molecules of the fluid. This motion is three-dimensional. The displacements in numerical order of a particle undergoing Brownian motion could look like the following, in micrometers (**Figure 6.6**):

$$\begin{aligned}
\Delta\vec{\mathbf{r}}_1 &= 2.0\,\hat{\mathbf{i}} + \hat{\mathbf{j}} + 3.0\,\hat{\mathbf{k}} \\
\Delta\vec{\mathbf{r}}_2 &= -\hat{\mathbf{i}} + 3.0\,\hat{\mathbf{k}} \\
\Delta\vec{\mathbf{r}}_3 &= 4.0\,\hat{\mathbf{i}} - 2.0\,\hat{\mathbf{j}} + \hat{\mathbf{k}} \\
\Delta\vec{\mathbf{r}}_4 &= -3.0\,\hat{\mathbf{i}} + \hat{\mathbf{j}} + 2.0\,\hat{\mathbf{k}}.
\end{aligned}$$

What is the total displacement of the particle from the origin?

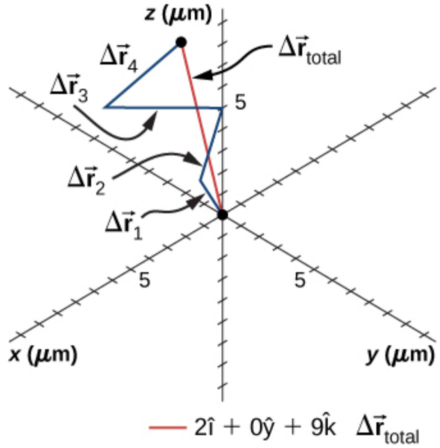


Figure 6.6 Trajectory of a particle undergoing random displacements of Brownian motion. The total displacement is shown in red.

Solution

We form the sum of the displacements and add them as vectors:

$$\begin{aligned}\Delta\vec{\mathbf{r}}_{\text{Total}} &= \sum \Delta\vec{\mathbf{r}}_i = \Delta\vec{\mathbf{r}}_1 + \Delta\vec{\mathbf{r}}_2 + \Delta\vec{\mathbf{r}}_3 + \Delta\vec{\mathbf{r}}_4 \\ &= (2.0 - 1.0 + 4.0 - 3.0)\hat{\mathbf{i}} + (1.0 + 0 - 2.0 + 1.0)\hat{\mathbf{j}} + (3.0 + 3.0 + 1.0 + 2.0)\hat{\mathbf{k}} \\ &= 2.0\hat{\mathbf{i}} + 0\hat{\mathbf{j}} + 9.0\hat{\mathbf{k}}\,\mu\text{m}.\end{aligned}$$

To complete the solution, we express the displacement as a magnitude and direction,

$$\left|\Delta\vec{\mathbf{r}}_{\text{Total}}\right| = \sqrt{2.0^2 + 0^2 + 9.0^2} = 9.2\,\mu\text{m}, \quad \theta = \tan^{-1}\left(\frac{9}{2}\right) = 77°,$$

with respect to the *x*-axis in the *xz*-plane.

Significance

From the figure we can see the magnitude of the total displacement is less than the sum of the magnitudes of the individual displacements.

Velocity Vector

In the previous chapter we found the instantaneous velocity by calculating the derivative of the position function with respect to time. We can do the same operation in two and three dimensions, but we use vectors. The instantaneous **velocity vector** is now

$$\vec{\mathbf{v}}(t) = \lim_{\Delta t \to 0} \frac{\vec{\mathbf{r}}(t + \Delta t) - \vec{\mathbf{r}}(t)}{\Delta t} = \frac{d\vec{\mathbf{r}}}{dt}. \tag{6.4}$$

Let's look at the relative orientation of the position vector and velocity vector graphically. In **Figure 6.7** we show the vectors $\vec{\mathbf{r}}(t)$ and $\vec{\mathbf{r}}(t + \Delta t)$, which give the position of a particle moving along a path represented by the gray line. As Δt goes to zero, the velocity vector, given by **Equation 6.4**, becomes tangent to the path of the particle at time *t*.

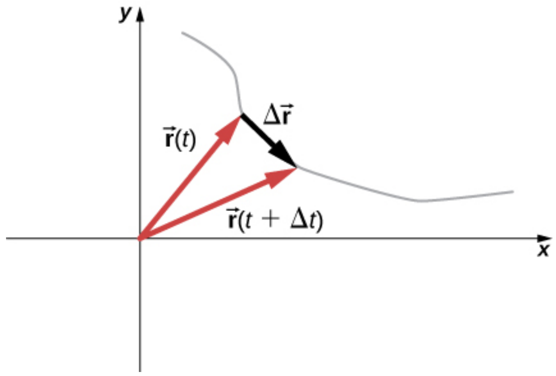


Figure 6.7 A particle moves along a path given by the gray line. In the limit as Δt approaches zero, the velocity vector becomes tangent to the path of the particle.

Equation 6.4 can also be written in terms of the components of $\vec{v}(t)$. Since

$$\vec{r}(t) = x(t)\hat{i} + y(t)\hat{j} + z(t)\hat{k},$$

we can write

$$\vec{v}(t) = v_x(t)\hat{i} + v_y(t)\hat{j} + v_z(t)\hat{k} \tag{6.5}$$

where

$$v_x(t) = \frac{dx(t)}{dt}, \quad v_y(t) = \frac{dy(t)}{dt}, \quad v_z(t) = \frac{dz(t)}{dt}. \tag{6.6}$$

If only the average velocity is of concern, we have the vector equivalent of the one-dimensional average velocity for two and three dimensions:

$$\vec{v}_{\text{avg}} = \frac{\vec{r}(t_2) - \vec{r}(t_1)}{t_2 - t_1}. \tag{6.7}$$

Example 6.3

Calculating the Velocity Vector

The position function of a particle is $\vec{r}(t) = 2.0t^2\hat{i} + (2.0 + 3.0t)\hat{j} + 5.0t\hat{k}$ m. (a) What is the instantaneous velocity and speed at $t = 2.0$ s? (b) What is the average velocity between 1.0 s and 3.0 s?

Solution

Using **Equation 6.5** and **Equation 6.6**, and taking the derivative of the position function with respect to time, we find

(a) $v(t) = \frac{d\mathbf{r}(t)}{dt} = 4.0t\,\hat{\mathbf{i}} + 3.0\,\hat{\mathbf{j}} + 5.0\,\hat{\mathbf{k}}\,\text{m/s}$

$\vec{\mathbf{v}}\,(2.0s) = 8.0\,\hat{\mathbf{i}} + 3.0\,\hat{\mathbf{j}} + 5.0\,\hat{\mathbf{k}}\,\text{m/s}$

Speed $|\vec{\mathbf{v}}\,(2.0\text{ s})| = \sqrt{8^2 + 3^2 + 5^2} = 9.9\text{ m/s}.$

(b) From **Equation 6.7**,

$$\vec{\mathbf{v}}_{\text{avg}} = \frac{\vec{\mathbf{r}}\,(t_2) - \vec{\mathbf{r}}\,(t_1)}{t_2 - t_1} = \frac{\vec{\mathbf{r}}\,(3.0\text{ s}) - \vec{\mathbf{r}}\,(1.0\text{ s})}{3.0\text{ s} - 1.0\text{ s}} = \frac{(18\,\hat{\mathbf{i}} + 11\,\hat{\mathbf{j}} + 15\,\hat{\mathbf{k}})\text{ m} - (2\,\hat{\mathbf{i}} + 5\,\hat{\mathbf{j}} + 5\,\hat{\mathbf{k}})\text{ m}}{2.0\text{ s}}$$
$$= \frac{(16\,\hat{\mathbf{i}} + 6\,\hat{\mathbf{j}} + 10\,\hat{\mathbf{k}})\text{ m}}{2.0\text{ s}} = 8.0\,\hat{\mathbf{i}} + 3.0\,\hat{\mathbf{j}} + 5.0\,\hat{\mathbf{k}}\,\text{m/s}.$$

Significance

We see the average velocity is the same as the instantaneous velocity at $t = 2.0$ s, as a result of the velocity function being linear. This need not be the case in general. In fact, most of the time, instantaneous and average velocities are not the same.

6.1 **Check Your Understanding** The position function of a particle is $\vec{\mathbf{r}}\,(t) = 3.0t^3\,\hat{\mathbf{i}} + 4.0\,\hat{\mathbf{j}}$. (a) What is the instantaneous velocity at $t = 3$ s? (b) Is the average velocity between 2 s and 4 s equal to the instantaneous velocity at $t = 3$ s?

The Independence of Perpendicular Motions

When we look at the three-dimensional equations for position and velocity written in unit vector notation, **Equation 6.2** and **Equation 6.5**, we see the components of these equations are separate and unique functions of time that do not depend on one another. Motion along the x direction has no part of its motion along the y and z directions, and similarly for the other two coordinate axes. Thus, the motion of an object in two or three dimensions can be divided into separate, independent motions along the perpendicular axes of the coordinate system in which the motion takes place.

To illustrate this concept with respect to displacement, consider a woman walking from point A to point B in a city with square blocks. The woman taking the path from A to B may walk east for so many blocks and then north (two perpendicular directions) for another set of blocks to arrive at B. How far she walks east is affected only by her motion eastward. Similarly, how far she walks north is affected only by her motion northward.

Independence of Motion

In the kinematic description of motion, we are able to treat the horizontal and vertical components of motion separately. In many cases, motion in the horizontal direction does not affect motion in the vertical direction, and vice versa.

An example illustrating the independence of vertical and horizontal motions is given by two baseballs. One baseball is dropped from rest. At the same instant, another is thrown horizontally from the same height and it follows a curved path. A stroboscope captures the positions of the balls at fixed time intervals as they fall (**Figure 6.8**).

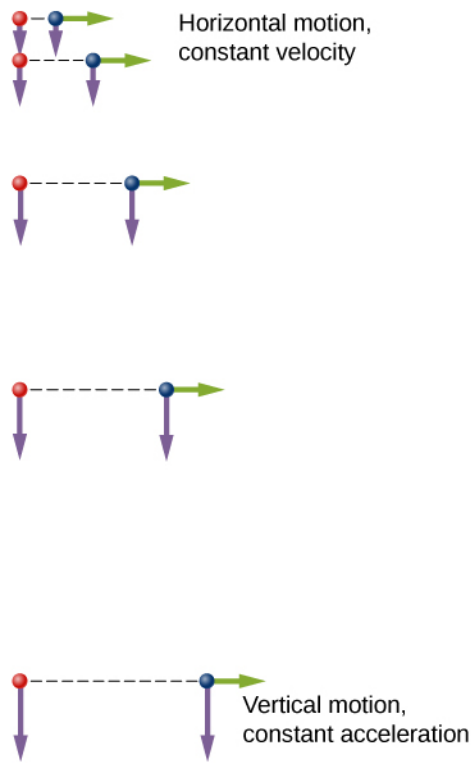


Figure 6.8 A diagram of the motions of two identical balls: one falls from rest and the other has an initial horizontal velocity. Each subsequent position is an equal time interval. Arrows represent the horizontal and vertical velocities at each position. The ball on the right has an initial horizontal velocity whereas the ball on the left has no horizontal velocity. Despite the difference in horizontal velocities, the vertical velocities and positions are identical for both balls, which shows the vertical and horizontal motions are independent.

It is remarkable that for each flash of the strobe, the vertical positions of the two balls are the same. This similarity implies vertical motion is independent of whether the ball is moving horizontally. (Assuming no air resistance, the vertical motion of a falling object is influenced by gravity only, not by any horizontal forces.) Careful examination of the ball thrown horizontally shows it travels the same horizontal distance between flashes. This is because there are no additional forces on the ball in the horizontal direction after it is thrown. This result means horizontal velocity is constant and is affected neither by vertical motion nor by gravity (which is vertical). Note this case is true for ideal conditions only. In the real world, air resistance affects the speed of the balls in both directions.

The two-dimensional curved path of the horizontally thrown ball is composed of two independent one-dimensional motions (horizontal and vertical). The key to analyzing such motion, called *projectile motion*, is to resolve it into motions along perpendicular directions. Resolving two-dimensional motion into perpendicular components is possible because the components are independent.

6.2 | Acceleration Vector

Learning Objectives

By the end of this section, you will be able to:

- Calculate the acceleration vector given the velocity function in unit vector notation.
- Describe the motion of a particle with a constant acceleration in three dimensions.
- Use the one-dimensional motion equations along perpendicular axes to solve a problem in two or three dimensions with a constant acceleration.
- Express the acceleration in unit vector notation.

Instantaneous Acceleration

In addition to obtaining the displacement and velocity vectors of an object in motion, we often want to know its **acceleration vector** at any point in time along its trajectory. This acceleration vector is the instantaneous acceleration and it can be obtained from the derivative with respect to time of the velocity function, as we have seen in a previous chapter. The only difference in two or three dimensions is that these are now vector quantities. Taking the derivative with respect to time $\vec{\mathbf{v}}(t)$, we find

$$\vec{\mathbf{a}}(t) = \lim_{t \to 0} \frac{\vec{\mathbf{v}}(t + \Delta t) - \vec{\mathbf{v}}(t)}{\Delta t} = \frac{d\vec{\mathbf{v}}(t)}{dt}. \tag{6.8}$$

The acceleration in terms of components is

$$\vec{\mathbf{a}}(t) = \frac{dv_x(t)}{dt}\hat{\mathbf{i}} + \frac{dv_y(t)}{dt}\hat{\mathbf{j}} + \frac{dv_z(t)}{dt}\hat{\mathbf{k}}. \tag{6.9}$$

Constant Acceleration

Multidimensional motion with constant acceleration can be treated the same way as shown in the previous chapter for one-dimensional motion. Earlier we showed that three-dimensional motion is equivalent to three one-dimensional motions, each along an axis perpendicular to the others. To develop the relevant equations in each direction, let's consider the two-dimensional problem of a particle moving in the *xy* plane with constant acceleration, ignoring the *z*-component for the moment. The acceleration vector is

$$\vec{\mathbf{a}} = a_{0x}\hat{\mathbf{i}} + a_{0y}\hat{\mathbf{j}}.$$

Each component of the motion has a separate set of equations similar to **Equation 3.9**–**Equation 3.13** of the **Section 3.4**. We show only the equations for position and velocity in the *x*- and *y*-directions. A similar set of kinematic equations could be written for motion in the *z*-direction:

$$x(t) = x_0 + (v_x)_{\text{avg}}\, t \tag{6.10}$$

$$v_x(t) = v_{0x} + a_x t \tag{6.11}$$

$$x(t) = x_0 + v_{0x} t + \frac{1}{2} a_x t^2 \tag{6.12}$$

$$v_x^2(t) = v_{0x}^2 + 2a_x(x - x_0) \tag{6.13}$$

$$y(t) = y_0 + (v_y)_{\text{avg}}\, t \tag{6.14}$$

$$v_y(t) = v_{0y} + a_y t \tag{6.15}$$

$$y(t) = y_0 + v_{0y}t + \frac{1}{2}a_y t^2 \quad \textbf{(6.16)}$$

$$v_y^2(t) = v_{0y}^2 + 2a_y(y - y_0). \quad \textbf{(6.17)}$$

Here the subscript 0 denotes the initial position or velocity. **Equation 6.10** to **Equation 6.17** can be substituted into **Equation 6.2** and **Equation 6.5** without the *z*-component to obtain the position vector and velocity vector as a function of time in two dimensions:

$$\vec{\mathbf{r}}(t) = x(t)\hat{\mathbf{i}} + y(t)\hat{\mathbf{j}} \text{ and } \vec{\mathbf{v}}(t) = v_x(t)\hat{\mathbf{i}} + v_y(t)\hat{\mathbf{j}}.$$

The following example illustrates a practical use of the kinematic equations in two dimensions.

Example 6.4

A Skier

Figure 6.9 shows a skier moving with an acceleration of 2.1 m/s^2 down a slope of $15°$ at $t = 0$. With the origin of the coordinate system at the front of the lodge, her initial position and velocity are

$$\vec{\mathbf{r}}(0) = (75.0\hat{\mathbf{i}} - 50.0\hat{\mathbf{j}})\text{ m}$$

and

$$\vec{\mathbf{v}}(0) = (4.1\hat{\mathbf{i}} - 1.1\hat{\mathbf{j}})\text{ m/s}.$$

(a) What are the *x*- and *y*-components of the skier's position and velocity as functions of time? (b) What are her position and velocity at $t = 10.0$ s?

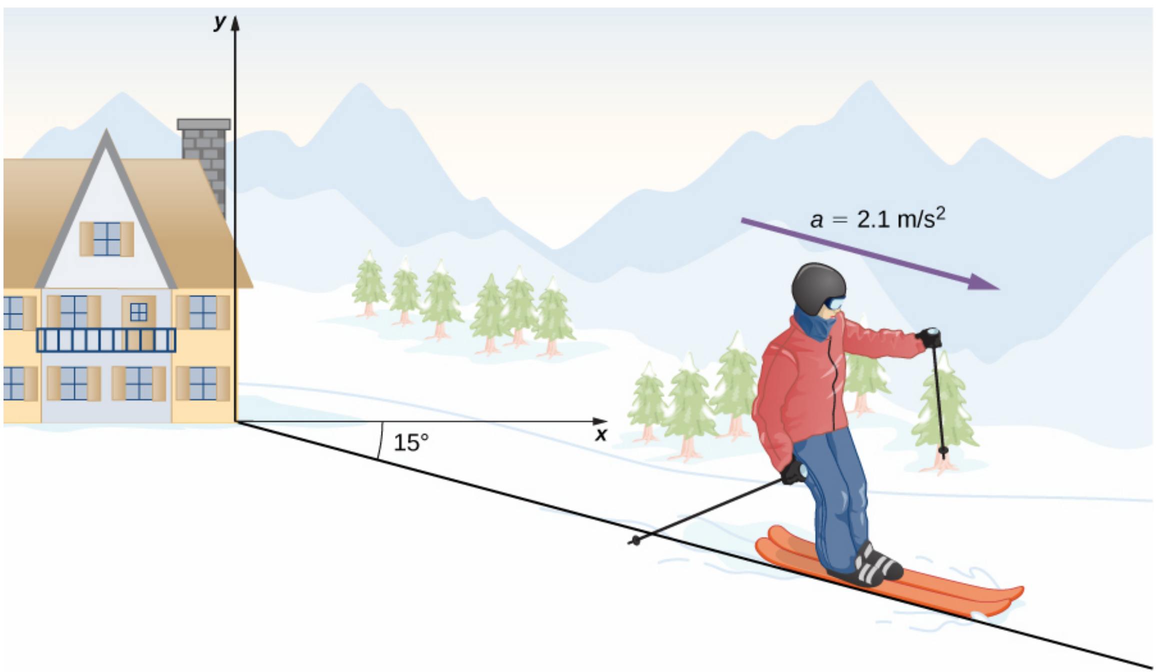


Figure 6.9 A skier has an acceleration of 2.1 m/s^2 down a slope of $15°$. The origin of the coordinate system is at the ski lodge.

Strategy

Since we are evaluating the components of the motion equations in the *x* and *y* directions, we need to find the components of the acceleration and put them into the kinematic equations. The components of the acceleration are found by referring to the coordinate system in **Figure 6.9**. Then, by inserting the components of the initial position and velocity into the motion equations, we can solve for her position and velocity at a later time t.

Solution

(a) The origin of the coordinate system is at the top of the hill with y-axis vertically upward and the x-axis horizontal. By looking at the trajectory of the skier, the x-component of the acceleration is positive and the y-component is negative. Since the angle is $15°$ down the slope, we find

$$a_x = (2.1\text{ m/s}^2)\cos(15°) = 2.0\text{ m/s}^2$$
$$a_y = (-2.1\text{ m/s}^2)\sin 15° = -0.54\text{ m/s}^2.$$

Inserting the initial position and velocity into **Equation 6.11** and **Equation 6.12** for x, we have

$$x(t) = 75.0\text{ m} + (4.1\text{ m/s})t + \frac{1}{2}(2.0\text{ m/s}^2)t^2$$
$$v_x(t) = 4.1\text{ m/s} + (2.0\text{ m/s}^2)t.$$

For y, we have

$$y(t) = -50.0\text{ m} + (-1.1\text{ m/s})t + \frac{1}{2}(-0.54\text{ m/s}^2)t^2$$
$$v_y(t) = -1.1\text{ m/s} + (-0.54\text{ m/s}^2)t.$$

(b) Now that we have the equations of motion for x and y as functions of time, we can evaluate them at t = 10.0 s:

$$x(10.0\text{ s}) = 75.0\text{ m} + (4.1\text{ m/s}^2)(10.0\text{ s}) + \frac{1}{2}(2.0\text{ m/s}^2)(10.0\text{ s})^2 = 216.0\text{ m}$$
$$v_x(10.0\text{ s}) = 4.1\text{ m/s} + (2.0\text{ m/s}^2)(10.0\text{ s}) = 24.1\text{m/s}$$
$$y(10.0\text{ s}) = -50.0\text{ m} + (-1.1\text{ m/s})(10.0\text{ s}) + \frac{1}{2}(-0.54\text{ m/s}^2)(10.0\text{ s})^2 = -88.0\text{ m}$$
$$v_y(10.0\text{ s}) = -1.1\text{ m/s} + (-0.54\text{ m/s}^2)(10.0\text{ s}) = -6.5\text{ m/s}.$$

The position and velocity at t = 10.0 s are, finally,

$$\vec{\mathbf{r}}\,(10.0\text{ s}) = (216.0\,\hat{\mathbf{i}} - 88.0\,\hat{\mathbf{j}})\text{ m}$$
$$\vec{\mathbf{v}}\,(10.0\text{ s}) = (24.1\,\hat{\mathbf{i}} - 6.5\,\hat{\mathbf{j}})\text{m/s}.$$

The magnitude of the velocity of the skier at 10.0 s is 25 m/s, which is 60 mi/h.

Significance

It is useful to know that, given the initial conditions of position, velocity, and acceleration of an object, we can find the position, velocity, and acceleration at any later time.

With **Equation 6.8** through **???** we have completed the set of expressions for the position, velocity, and acceleration of an object moving in two or three dimensions. If the trajectories of the objects look something like the “Red Arrows” in the opening picture for the chapter, then the expressions for the position, velocity, and acceleration can be quite complicated. In the sections to follow we examine two special cases of motion in two and three dimensions by looking at projectile motion and circular motion.

At this **University of Colorado Boulder website (https://openstaxcollege.org/l/21phetmotladyb)** , you can explore the position velocity and acceleration of a ladybug with an interactive simulation that allows you to change these parameters.

6.3 | Projectile Motion

Learning Objectives

By the end of this section, you will be able to:

- Use one-dimensional motion in perpendicular directions to analyze projectile motion.
- Calculate the range, time of flight, and maximum height of a projectile that is launched and impacts a flat, horizontal surface.
- Find the time of flight and impact velocity of a projectile that lands at a different height from that of launch.
- Calculate the trajectory of a projectile.

Projectile motion is the motion of an object thrown or projected into the air, subject only to acceleration as a result of gravity. The applications of projectile motion in physics and engineering are numerous. Some examples include meteors as they enter Earth's atmosphere, fireworks, and the motion of any ball in sports. Such objects are called *projectiles* and their path is called a **trajectory**. The motion of falling objects as discussed in Section 3.5 is a simple one-dimensional type of projectile motion in which there is no horizontal movement. In this section, we consider two-dimensional projectile motion, and our treatment neglects the effects of air resistance.

The most important fact to remember here is that *motions along perpendicular axes are independent* and thus can be analyzed separately. We discussed this fact in Section 6.1, where we saw that vertical and horizontal motions are independent. The key to analyzing two-dimensional projectile motion is to break it into two motions: one along the horizontal axis and the other along the vertical. (This choice of axes is the most sensible because acceleration resulting from gravity is vertical; thus, there is no acceleration along the horizontal axis when air resistance is negligible.) As is customary, we call the horizontal axis the *x*-axis and the vertical axis the *y*-axis. It is not required that we use this choice of axes; it is simply convenient in the case of gravitational acceleration. In other cases we may choose a different set of axes. Figure 6.10 illustrates the notation for displacement, where we define $\vec{s}$ to be the total displacement, and $\vec{x}$ and $\vec{y}$ are its component vectors along the horizontal and vertical axes, respectively. The magnitudes of these vectors are *s*, *x*, and *y*.

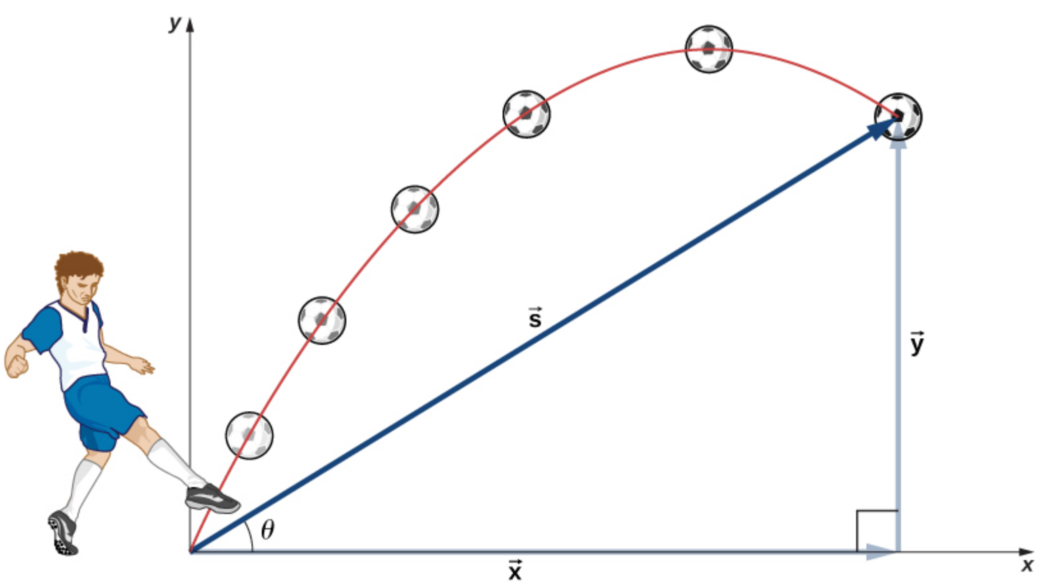


Figure 6.10 The total displacement *s* of a soccer ball at a point along its path. The vector $\vec{s}$ has components $\vec{x}$ and $\vec{y}$ along the horizontal and vertical axes. Its magnitude is *s* and it makes an angle θ with the horizontal.

To describe projectile motion completely, we must include velocity and acceleration, as well as displacement. We must find their components along the *x*- and *y*-axes. Let's assume all forces except gravity (such as air resistance and friction, for example) are negligible. Defining the positive direction to be upward, the components of acceleration are then very simple:

$$a_y = -g = -9.8 \text{ m/s}^2 \ \ (-32 \text{ ft/s}^2).$$

Because gravity is vertical, $a_x = 0.$ If $a_x = 0,$ this means the initial velocity in the *x* direction is equal to the final velocity in the *x* direction, or $v_x = v_{0x}.$ With these conditions on acceleration and velocity, we can write the kinematic **Equation 6.10** through **Equation 6.17** for motion in a uniform gravitational field, including the rest of the kinematic equations for a constant acceleration from **Section 3.4**. The kinematic equations for motion in a uniform gravitational field become kinematic equations with $a_y = -g, \quad a_x = 0$:

Horizontal Motion

$$v_{0x} = v_x, \ \ x = x_0 + v_x t \tag{6.18}$$

Vertical Motion

$$y = y_0 + \frac{1}{2}(v_{0y} + v_y)t \tag{6.19}$$

$$v_y = v_{0y} - gt \tag{6.20}$$

$$y = y_0 + v_{0y} t - \frac{1}{2} g t^2 \tag{6.21}$$

$$v_y^2 = v_{0y}^2 - 2g(y - y_0) \tag{6.22}$$

Using this set of equations, we can analyze projectile motion, keeping in mind some important points.

Problem-Solving Strategy: Projectile Motion

1. Resolve the motion into horizontal and vertical components along the *x*- and *y*-axes. The magnitudes of the components of displacement $\vec{\mathbf{s}}$ along these axes are *x* and *y*. The magnitudes of the components of velocity $\vec{\mathbf{v}}$ are $v_x = v\cos\theta$ and $v_y = v\sin\theta,$ where *v* is the magnitude of the velocity and *θ* is its direction relative to the horizontal, as shown in **Figure 6.11**.
2. Treat the motion as two independent one-dimensional motions: one horizontal and the other vertical. Use the kinematic equations for horizontal and vertical motion presented earlier.
3. Solve for the unknowns in the two separate motions: one horizontal and one vertical. Note that the only common variable between the motions is time *t*. The problem-solving procedures here are the same as those for one-dimensional kinematics and are illustrated in the following solved examples.
4. Recombine quantities in the horizontal and vertical directions to find the total displacement $\vec{\mathbf{s}}$ and velocity $\vec{\mathbf{v}}$. Solve for the magnitude and direction of the displacement and velocity using

$$s = \sqrt{x^2 + y^2}, \quad \theta = \tan^{-1}(y/x), \quad v = \sqrt{v_x^2 + v_y^2},$$

where *θ* is the direction of the displacement $\vec{\mathbf{s}}$.

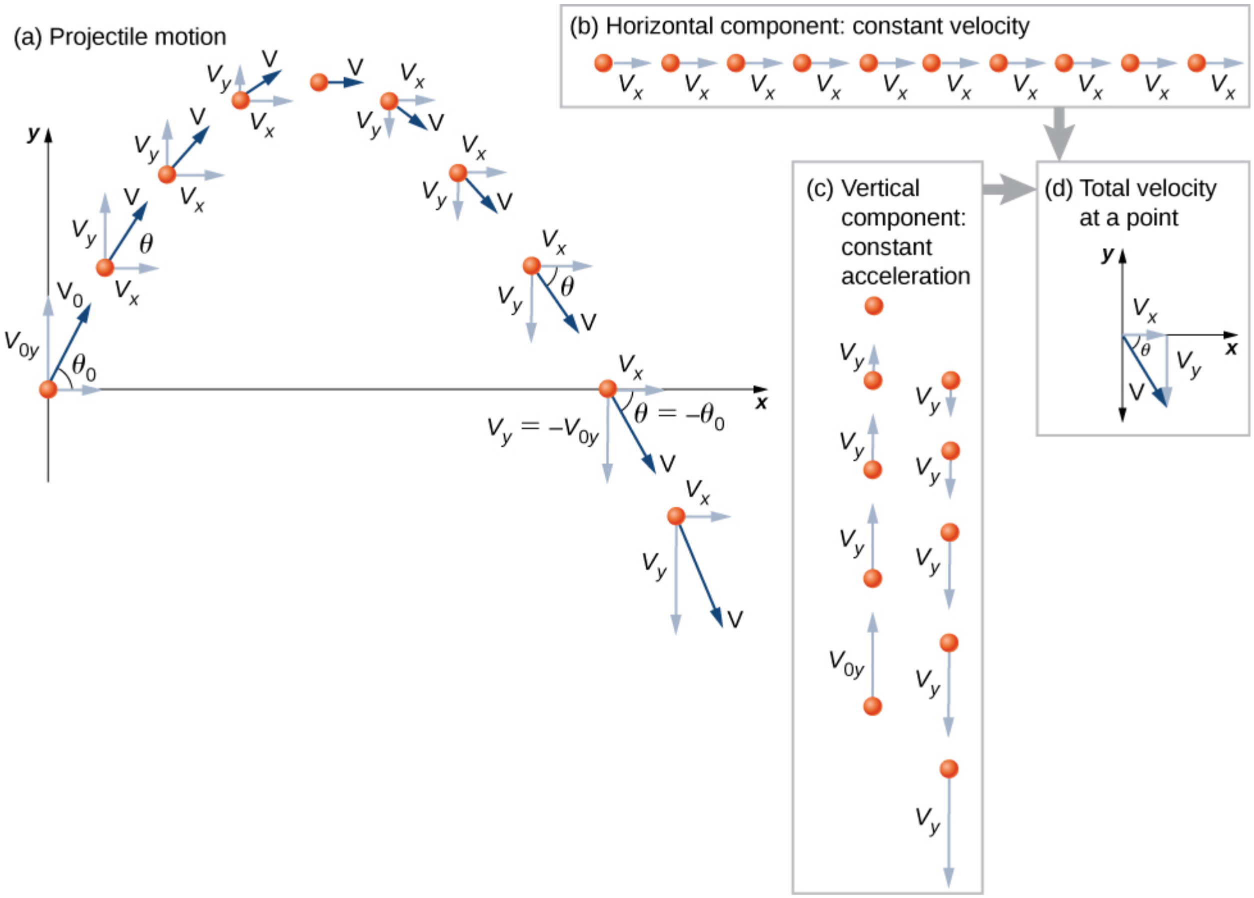


Figure 6.11 (a) We analyze two-dimensional projectile motion by breaking it into two independent one-dimensional motions along the vertical and horizontal axes. (b) The horizontal motion is simple, because $a_x = 0$ and v_x is a constant. (c) The velocity in the vertical direction begins to decrease as the object rises. At its highest point, the vertical velocity is zero. As the object falls toward Earth again, the vertical velocity increases again in magnitude but points in the opposite direction to the initial vertical velocity. (d) The *x* and *y* motions are recombined to give the total velocity at any given point on the trajectory.

Example 6.5

A Fireworks Projectile Explodes High and Away

During a fireworks display, a shell is shot into the air with an initial speed of 70.0 m/s at an angle of $75.0°$ above the horizontal, as illustrated in **Figure 6.12**. The fuse is timed to ignite the shell just as it reaches its highest point above the ground. (a) Calculate the height at which the shell explodes. (b) How much time passes between the launch of the shell and the explosion? (c) What is the horizontal displacement of the shell when it explodes? (d) What is the total displacement from the point of launch to the highest point?

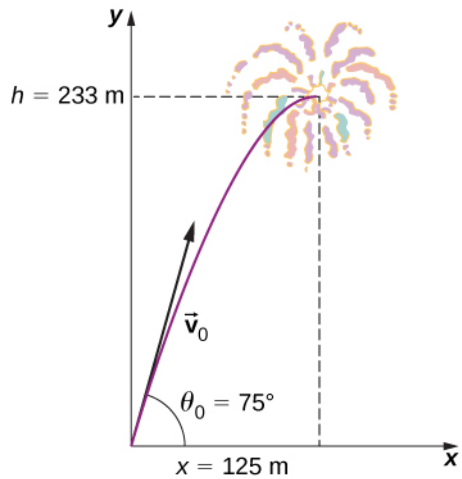


Figure 6.12 The trajectory of a fireworks shell. The fuse is set to explode the shell at the highest point in its trajectory, which is found to be at a height of 233 m and 125 m away horizontally.

Strategy

The motion can be broken into horizontal and vertical motions in which $a_x = 0$ and $a_y = -g$. We can then define x_0 and y_0 to be zero and solve for the desired quantities.

Solution

(a) By "height" we mean the altitude or vertical position *y* above the starting point. The highest point in any trajectory, called the *apex*, is reached when $v_y = 0$. Since we know the initial and final velocities, as well as the initial position, we use the following equation to find *y*:

$$v_y^2 = v_{0y}^2 - 2g(y - y_0).$$

Because y_0 and v_y are both zero, the equation simplifies to

$$0 = v_{0y}^2 - 2gy.$$

Solving for *y* gives

$$y = \frac{v_{0y}^2}{2g}.$$

Now we must find v_{0y}, the component of the initial velocity in the *y* direction. It is given by $v_{0y} = v_0 \sin\theta_0$, where v_0 is the initial velocity of 70.0 m/s and $\theta_0 = 75°$ is the initial angle. Thus,

$$v_{0y} = v_0 \sin\theta = (70.0\text{ m/s})\sin 75° = 67.6\text{ m/s}$$

and *y* is

$$y = \frac{(67.6\text{ m/s})^2}{2(9.80\text{ m/s}^2)}.$$

Thus, we have

$$y = 233\text{ m}.$$

Note that because up is positive, the initial vertical velocity is positive, as is the maximum height, but the acceleration resulting from gravity is negative. Note also that the maximum height depends only on the vertical component of the initial velocity, so that any projectile with a 67.6-m/s initial vertical component of velocity reaches a maximum height of 233 m (neglecting air resistance). The numbers in this example are reasonable for large fireworks displays, the shells of which do reach such heights before exploding. In practice, air resistance is not completely negligible, so the initial velocity would have to be somewhat larger than that given to reach the same height.

(b) As in many physics problems, there is more than one way to solve for the time the projectile reaches its highest point. In this case, the easiest method is to use $v_y = v_{0y} - gt$. Because $v_y = 0$ at the apex, this equation reduces to simply

$$0 = v_{0y} - gt$$

or

$$t = \frac{v_{0y}}{g} = \frac{67.6 \text{ m/s}}{9.80 \text{ m/s}^2} = 6.90\text{s}.$$

This time is also reasonable for large fireworks. If you are able to see the launch of fireworks, notice that several seconds pass before the shell explodes. Another way of finding the time is by using $y = y_0 + \frac{1}{2}(v_{0y} + v_y)t$. This is left for you as an exercise to complete.

(c) Because air resistance is negligible, $a_x = 0$ and the horizontal velocity is constant, as discussed earlier. The horizontal displacement is the horizontal velocity multiplied by time as given by $x = x_0 + v_x t$, where x_0 is equal to zero. Thus,

$$x = v_x t,$$

where v_x is the *x*-component of the velocity, which is given by

$$v_x = v_0 \cos\theta = (70.0 \text{ m/s})\cos 75° = 18.1 \text{ m/s}.$$

Time *t* for both motions is the same, so *x* is

$$x = (18.1 \text{ m/s})6.90 \text{ s} = 125 \text{ m}.$$

Horizontal motion is a constant velocity in the absence of air resistance. The horizontal displacement found here could be useful in keeping the fireworks fragments from falling on spectators. When the shell explodes, air resistance has a major effect, and many fragments land directly below.

(d) The horizontal and vertical components of the displacement were just calculated, so all that is needed here is to find the magnitude and direction of the displacement at the highest point:

$$\vec{\mathbf{s}} = 125\,\hat{\mathbf{i}} + 233\,\hat{\mathbf{j}}$$

$$|\vec{\mathbf{s}}| = \sqrt{125^2 + 233^2} = 264 \text{ m}$$

$$\theta = \tan^{-1}\left(\frac{233}{125}\right) = 61.8°.$$

Note that the angle for the displacement vector is less than the initial angle of launch. To see why this is, review **Figure 6.10**, which shows the curvature of the trajectory toward the ground level.

When solving **Example 6.5**(a), the expression we found for *y* is valid for any projectile motion when air resistance is negligible. Call the maximum height *y* = *h*. Then,

$$h = \frac{v_{0y}^2}{2g}.$$

This equation defines the *maximum height of a projectile above its launch position* and it depends only on the vertical component of the initial velocity.

6.2 Check Your Understanding A rock is thrown horizontally off a cliff 100.0 m high with a velocity of 15.0 m/s. (a) Define the origin of the coordinate system. (b) Which equation describes the horizontal motion? (c) Which equations describe the vertical motion? (d) What is the rock's velocity at the point of impact?

Example 6.6

Calculating Projectile Motion: Tennis Player

A tennis player wins a match at Arthur Ashe stadium and hits a ball into the stands at 30 m/s and at an angle $45°$ above the horizontal (**Figure 6.13**). On its way down, the ball is caught by a spectator 10 m above the point where the ball was hit. (a) Calculate the time it takes the tennis ball to reach the spectator. (b) What are the magnitude and direction of the ball's velocity at impact?

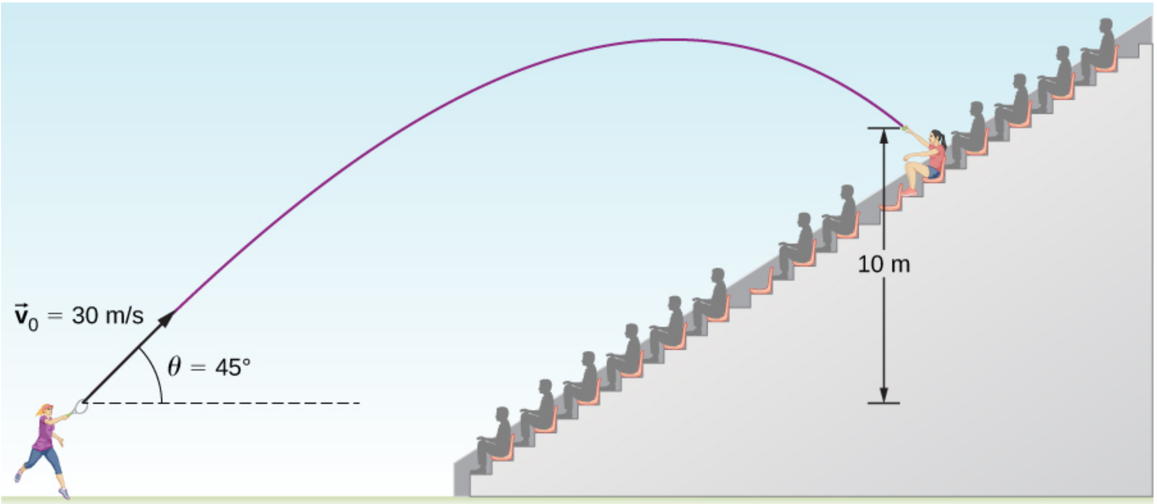


Figure 6.13 The trajectory of a tennis ball hit into the stands.

Strategy

Again, resolving this two-dimensional motion into two independent one-dimensional motions allows us to solve for the desired quantities. The time a projectile is in the air is governed by its vertical motion alone. Thus, we solve for t first. While the ball is rising and falling vertically, the horizontal motion continues at a constant velocity. This example asks for the final velocity. Thus, we recombine the vertical and horizontal results to obtain $\vec{v}$ at final time t, determined in the first part of the example.

Solution

(a) While the ball is in the air, it rises and then falls to a final position 10.0 m higher than its starting altitude. We can find the time for this by using **Equation 6.21**:

$$y = y_0 + v_{0y}t - \frac{1}{2}gt^2.$$

If we take the initial position y_0 to be zero, then the final position is y = 10 m. The initial vertical velocity is the vertical component of the initial velocity:

$$v_{0y} = v_0 \sin\theta_0 = (30.0\text{ m/s})\sin 45° = 21.2\text{ m/s}.$$

Substituting into **Equation 6.21** for y gives us

$$10.0\text{ m} = (21.2\text{ m/s})t - (4.90\text{ m/s}^2)t^2.$$

Rearranging terms gives a quadratic equation in t:

$$(4.90\text{ m/s}^2)t^2 - (21.2\text{ m/s})t + 10.0\text{ m} = 0.$$

Use of the quadratic formula yields t = 3.79 s and t = 0.54 s. Since the ball is at a height of 10 m at two times during its trajectory—once on the way up and once on the way down—we take the longer solution for the time it takes the ball to reach the spectator:

$$t = 3.79 \text{ s}.$$

The time for projectile motion is determined completely by the vertical motion. Thus, any projectile that has an initial vertical velocity of 21.2 m/s and lands 10.0 m below its starting altitude spends 3.79 s in the air.

(b) We can find the final horizontal and vertical velocities v_x and v_y with the use of the result from (a). Then, we can combine them to find the magnitude of the total velocity vector $\vec{\mathbf{v}}$ and the angle θ it makes with the horizontal. Since v_x is constant, we can solve for it at any horizontal location. We choose the starting point because we know both the initial velocity and the initial angle. Therefore,

$$v_x = v_0 \cos\theta_0 = (30 \text{ m/s})\cos 45° = 21.2 \text{ m/s}.$$

The final vertical velocity is given by **Equation 6.20**:

$$v_y = v_{0y} - gt.$$

Since v_{0y} was found in part (a) to be 21.2 m/s, we have

$$v_y = 21.2 \text{ m/s} - 9.8 \text{ m/s}^2(3.79 \text{ s}) = -15.9 \text{ m/s}.$$

The magnitude of the final velocity $\vec{\mathbf{v}}$ is

$$v = \sqrt{v_x^2 + v_y^2} = \sqrt{(21.2 \text{ m/s})^2 + (-15.9 \text{ m/s})^2} = 26.5 \text{ m/s}.$$

The direction θ_v is found using the inverse tangent:

$$\theta_v = \tan^{-1}\left(\frac{v_y}{v_x}\right) = \tan^{-1}\left(\frac{21.2}{-15.9}\right) = -53.1°.$$

Significance

(a) As mentioned earlier, the time for projectile motion is determined completely by the vertical motion. Thus, any projectile that has an initial vertical velocity of 21.2 m/s and lands 10.0 m below its starting altitude spends 3.79 s in the air. (b) The negative angle means the velocity is $53.1°$ below the horizontal at the point of impact. This result is consistent with the fact that the ball is impacting at a point on the other side of the apex of the trajectory and therefore has a negative y component of the velocity. The magnitude of the velocity is less than the magnitude of the initial velocity we expect since it is impacting 10.0 m above the launch elevation.

Time of Flight, Trajectory, and Range for Symmetric Projectile Motion

Of interest are the time of flight, trajectory, and range for a projectile launched on a flat horizontal surface and impacting on the same surface. In this case, kinematic equations give useful expressions for these quantities, which are derived in the following sections. As a word of caution, however, these equations are **not** valid unless the motion has the aforementioned symmetry - namely that the starting and ending heights are the same.

Time of flight

We can solve for the time of flight of a projectile that is both launched and impacts on a flat horizontal surface by performing some manipulations of the kinematic equations. We note the position and displacement in y must be zero at launch and at impact on an even surface. Thus, we set the displacement in y equal to zero and find

$$y - y_0 = v_{0y}t - \frac{1}{2}gt^2 = (v_0 \sin\theta_0)t - \frac{1}{2}gt^2 = 0.$$

Factoring, we have

$$t\left(v_0 \sin\theta_0 - \frac{gt}{2}\right) = 0.$$

Solving for t gives us

$$T_{\text{tof}} = \frac{2(v_0 \sin\theta_0)}{g}. \quad (6.23)$$

This is the **time of flight** for a projectile both launched and impacting on a flat horizontal surface. Equation 6.23 does not apply when the projectile lands at a different elevation than it was launched, as we saw in Example 6.6 of the tennis player hitting the ball into the stands. The other solution, $t = 0$, corresponds to the time at launch. The time of flight is linearly proportional to the initial velocity in the y direction and inversely proportional to g. Thus, on the Moon, where gravity is one-sixth that of Earth, a projectile launched with the same velocity as on Earth would be airborne six times as long.

Trajectory

The trajectory of a projectile can be found by eliminating the time variable t from the kinematic equations for arbitrary t and solving for $y(x)$. We take $x_0 = y_0 = 0$ so the projectile is launched from the origin. The kinematic equation for x gives

$$x = v_{0x}t \Rightarrow t = \frac{x}{v_{0x}} = \frac{x}{v_0 \cos\theta_0}.$$

Substituting the expression for t into the equation for the position $y = (v_0 \sin\theta_0)t - \frac{1}{2}gt^2$ gives

$$y = (v_0 \sin\theta_0)\left(\frac{x}{v_0 \cos\theta_0}\right) - \frac{1}{2}g\left(\frac{x}{v_0 \cos\theta_0}\right)^2.$$

Rearranging terms, we have

$$y = (\tan\theta_0)x - \left[\frac{g}{2(v_0 \cos\theta_0)^2}\right]x^2. \quad (6.24)$$

This trajectory equation is of the form $y = ax + bx^2,$ which is an equation of a parabola with coefficients

$$a = \tan\theta_0, \quad b = -\frac{g}{2(v_0 \cos\theta_0)^2}.$$

Range

From the trajectory equation we can also find the **range**, or the horizontal distance traveled by the projectile. Factoring Equation 6.24, we have

$$y = x\left[\tan\theta_0 - \frac{g}{2(v_0 \cos\theta_0)^2}x\right].$$

The position y is zero for both the launch point and the impact point, since we are again considering only a flat horizontal surface. Setting $y = 0$ in this equation gives solutions $x = 0$, corresponding to the launch point, and

$$x = \frac{2v_0^2 \sin\theta_0 \cos\theta_0}{g},$$

corresponding to the impact point. Using the trigonometric identity $2\sin\theta\cos\theta = \sin 2\theta$ and setting $x = R$ for range, we find

$$R = \frac{v_0^2 \sin 2\theta_0}{g}. \quad (6.25)$$

Note particularly that **Equation 6.25** is valid only for launch and impact on a horizontal surface. We see the range is directly proportional to the square of the initial speed v_0 and $\sin 2\theta_0$, and it is inversely proportional to the acceleration of gravity. Thus, on the Moon, the range would be six times greater than on Earth for the same initial velocity. Furthermore, we see from the factor $\sin 2\theta_0$ that the range is maximum at $45°$. These results are shown in **Figure 6.14**. In (a) we see that the greater the initial velocity, the greater the range. In (b), we see that the range is maximum at $45°$. This is true only for conditions neglecting air resistance. If air resistance is considered, the maximum angle is somewhat smaller. It is interesting that the same range is found for two initial launch angles that sum to $90°$. The projectile launched with the smaller angle has a lower apex than the higher angle, but they both have the same range.

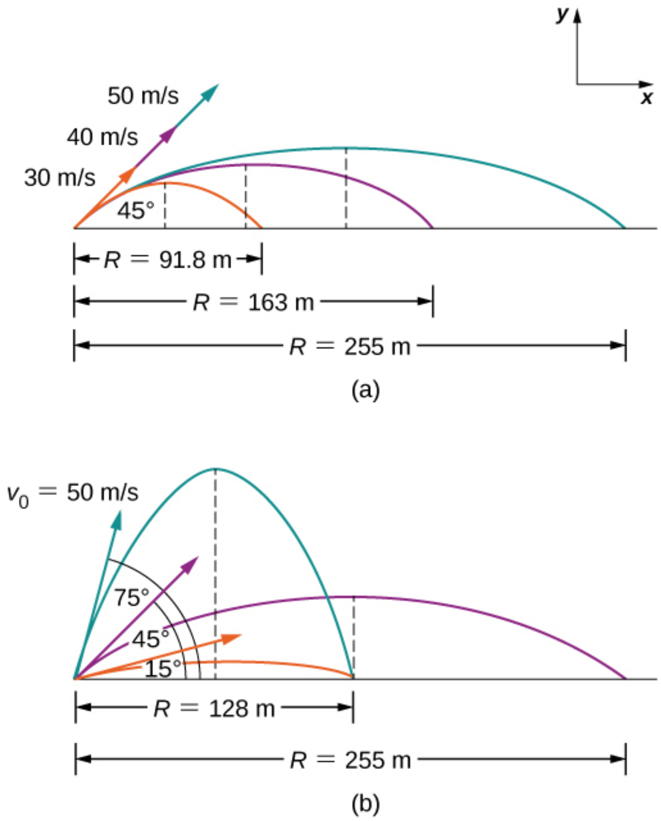


Figure 6.14 Trajectories of projectiles on level ground. (a) The greater the initial speed v_0, the greater the range for a given initial angle. (b) The effect of initial angle θ_0 on the range of a projectile with a given initial speed. Note that the range is the same for initial angles of $15°$ and $75°$, although the maximum heights of those paths are different.

Example 6.7

Comparing Golf Shots

A golfer finds himself in two different situations on different holes. On the second hole he is 120 m from the green and wants to hit the ball 90 m and let it run onto the green. He angles the shot low to the ground at $30°$ to the horizontal to let the ball roll after impact. On the fourth hole he is 90 m from the green and wants to let the ball drop with a minimum amount of rolling after impact. Here, he angles the shot at $70°$ to the horizontal to minimize rolling after impact. Both shots are hit and impacted on a level surface.

(a) What is the initial speed of the ball at the second hole?

(b) What is the initial speed of the ball at the fourth hole?

(c) Write the trajectory equation for both cases.

(d) Graph the trajectories.

Strategy

We see that the range equation has the initial speed and angle, so we can solve for the initial speed for both (a) and (b). When we have the initial speed, we can use this value to write the trajectory equation.

Solution

(a) $R = \frac{v_0^2 \sin 2\theta_0}{g} \Rightarrow v_0 = \sqrt{\frac{Rg}{\sin 2\theta_0}} = \sqrt{\frac{90.0\text{ m}(9.8\text{ m/s}^2)}{\sin(2(70°))}} = 37.0\text{ m/s}$

(b) $R = \frac{v_0^2 \sin 2\theta_0}{g} \Rightarrow v_0 = \sqrt{\frac{Rg}{\sin 2\theta_0}} = \sqrt{\frac{90.0\text{ m}(9.8\text{ m/s}^2)}{\sin(2(30°))}} = 31.9\text{ m/s}$

(c)

$$y = x\left[\tan\theta_0 - \frac{g}{2(v_0\cos\theta_0)^2}x\right]$$

Second hole: $y = x\left[\tan 70° - \frac{9.8\text{ m/s}^2}{2[(37.0\text{ m/s})(\cos 70°)]^2}x\right] = 2.75x - 0.0306x^2$

Fourth hole: $y = x\left[\tan 30° - \frac{9.8\text{ m/s}^2}{2[(31.9\text{ m/s})(\cos 30°)]^2}x\right] = 0.58x - 0.0064x^2$

(d) Using a graphing utility, we can compare the two trajectories, which are shown in **Figure 6.15**.

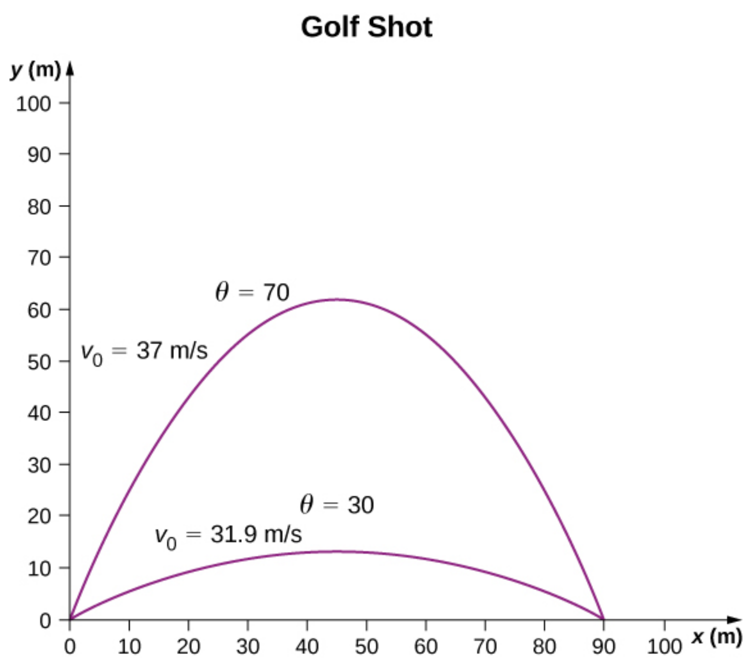


Figure 6.15 Two trajectories of a golf ball with a range of 90 m. The impact points of both are at the same level as the launch point.

Significance

The initial speed for the shot at $70°$ is greater than the initial speed of the shot at $30°$. Note from **Figure 6.15** that two projectiles launched at the same speed but at different angles have the same range if the launch angles add to $90°$. The launch angles in this example add to give a number greater than $90°$. Thus, the shot at $70°$ has to have a greater launch speed to reach 90 m, otherwise it would land at a shorter distance.

6.3 Check Your Understanding If the two golf shots in Example 6.7 were launched at the same speed, which shot would have the greatest range?

When we speak of the range of a projectile on level ground, we assume *R* is very small compared with the circumference of Earth. If, however, the range is large, Earth curves away below the projectile and the acceleration resulting from gravity changes direction along the path. The range is larger than predicted by the range equation given earlier because the projectile has farther to fall than it would on level ground, as shown in Figure 6.16, which is based on a drawing in Newton's *Principia* . If the initial speed is great enough, the projectile goes into orbit. Earth's surface drops 5 m every 8000 m. In 1 s an object falls 5 m without air resistance. Thus, if an object is given a horizontal velocity of 8000 m/s (or $18{,}000 \text{mi/hr}$) near Earth's surface, it will go into orbit around the planet because the surface continuously falls away from the object. This is roughly the speed of the Space Shuttle in a low Earth orbit when it was operational, or any satellite in a low Earth orbit. These and other aspects of orbital motion, such as Earth's rotation, are covered in greater depth in Newton's Synthesis.

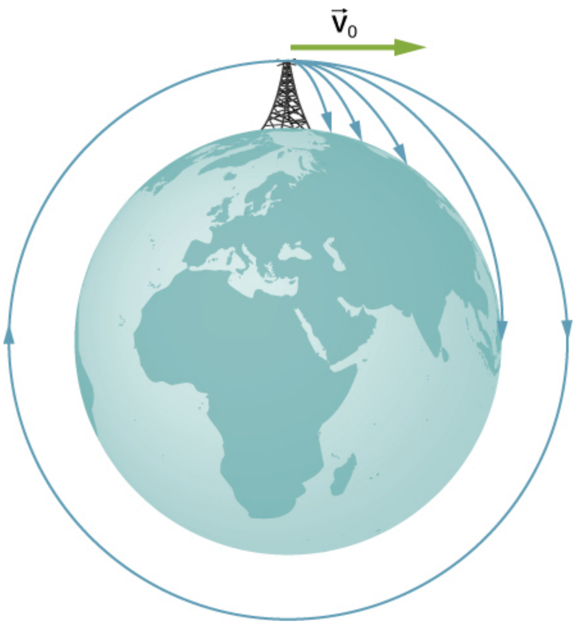


Figure 6.16 Projectile to satellite. In each case shown here, a projectile is launched from a very high tower to avoid air resistance. With increasing initial speed, the range increases and becomes longer than it would be on level ground because Earth curves away beneath its path. With a speed of 8000 m/s, orbit is achieved.

At PhET Explorations: Projectile Motion (https://openstaxcollege.org/l/21phetpromot) , learn about projectile motion in terms of the launch angle and initial velocity.

6.4 | Uniform Circular Motion

Learning Objectives

By the end of this section, you will be able to:

- Solve for the centripetal acceleration of an object moving on a circular path.
- Use the equations of circular motion to find the position, velocity, and acceleration of a particle executing circular motion.
- Explain the differences between centripetal acceleration and tangential acceleration resulting from nonuniform circular motion.
- Evaluate centripetal and tangential acceleration in nonuniform circular motion, and find the total acceleration vector.

Uniform Circular Motion Revisited

Uniform circular motion is a specific type of motion in which an object travels in a circle with a constant speed. For example, any point on a propeller spinning at a constant rate is executing uniform circular motion. Other examples are the second, minute, and hour hands of a watch. In **Rotation with Constant Angular Acceleration** we discussed the tangential acceleration of such a point. But, even in the absence of any tangential acceleration, when the motion is taking place at a constant angular speed, the points on these rotating objects are actually accelerating. To see this, we must analyze the circular motion in terms of vectors.

Centripetal Acceleration

In one-dimensional kinematics, objects with a constant speed have zero acceleration. However, in two- and three-dimensional kinematics, even if the speed is a constant, a particle can have acceleration if it moves along a curved trajectory such as a circle. In this case the velocity vector is changing, or $d\vec{\mathbf{v}}/dt \neq 0.$ This is shown in **Figure 6.17**. As the particle moves counterclockwise in time Δt on the circular path, its position vector moves from $\vec{\mathbf{r}}(t)$ to $\vec{\mathbf{r}}(t+\Delta t).$ The velocity vector has constant magnitude and is tangent to the path as it changes from $\vec{\mathbf{v}}(t)$ to $\vec{\mathbf{v}}(t+\Delta t),$ changing its direction only. Since the velocity vector $\vec{\mathbf{v}}(t)$ is perpendicular to the position vector $\vec{\mathbf{r}}(t),$ the triangles formed by the position vectors and $\Delta\vec{\mathbf{r}},$ and the velocity vectors and $\Delta\vec{\mathbf{v}}$ are similar. Furthermore, since $|\vec{\mathbf{r}}(t)| = |\vec{\mathbf{r}}(t+\Delta t)|$ and $|\vec{\mathbf{v}}(t)| = |\vec{\mathbf{v}}(t+\Delta t)|,$ the two triangles are isosceles. From these facts we can make the assertion

$$\frac{\Delta v}{v} = \frac{\Delta r}{r} \text{ or } \Delta v = \frac{v}{r}\Delta r.$$

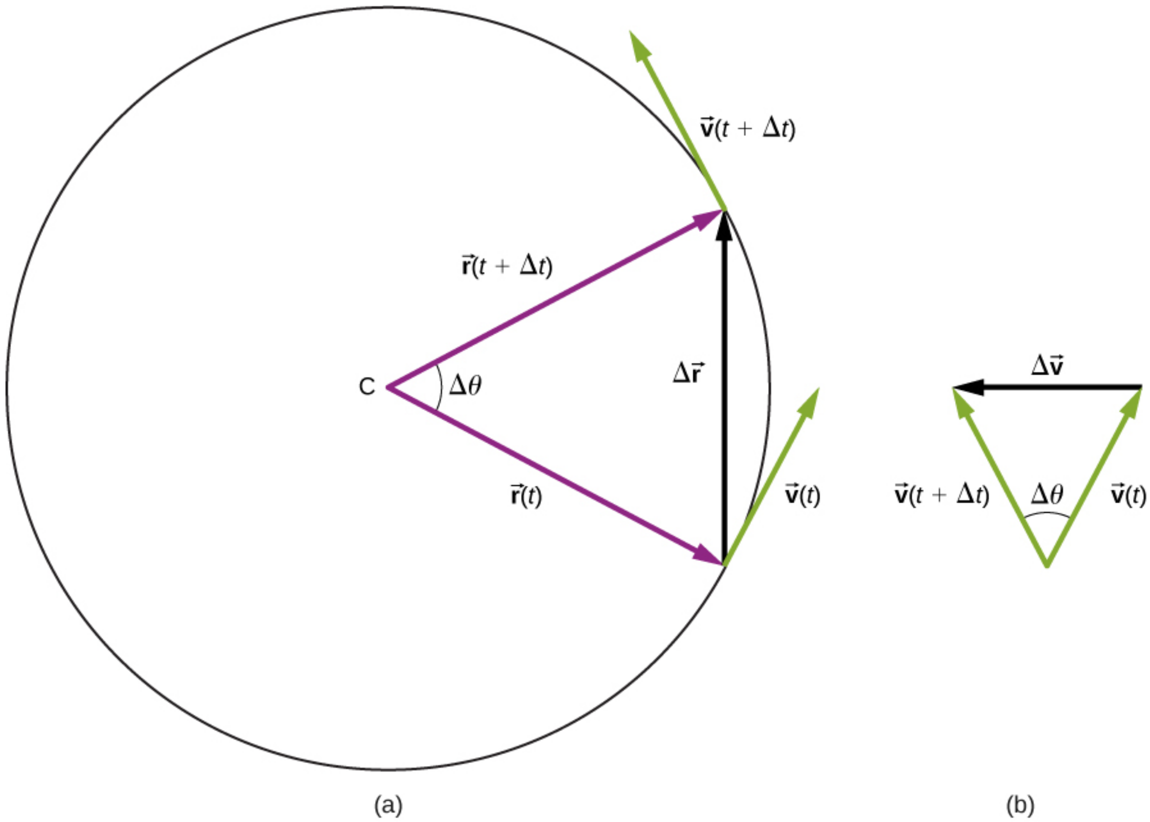


Figure 6.17 (a) A particle is moving in a circle at a constant speed, with position and velocity vectors at times t and $t + \Delta t$. (b) Velocity vectors forming a triangle. The two triangles in the figure are similar. The vector $\Delta \vec{v}$ points toward the center of the circle in the limit $\Delta t \rightarrow 0$.

We can find the magnitude of the acceleration from

$$a = \lim_{\Delta t \rightarrow 0}\left(\frac{\Delta v}{\Delta t}\right) = \frac{v}{r}\left(\lim_{\Delta t \rightarrow 0}\frac{\Delta r}{\Delta t}\right) = \frac{v^2}{r}.$$

The direction of the acceleration can also be found by noting that as Δt and therefore $\Delta\theta$ approach zero, the vector $\Delta \vec{v}$ approaches a direction perpendicular to $\vec{v}$. In the limit $\Delta t \rightarrow 0,$ $\Delta \vec{v}$ is perpendicular to $\vec{v}$. Since $\vec{v}$ is tangent to the circle, the acceleration $d\vec{v}/dt$ points toward the center of the circle. Summarizing, a particle moving in a circle at a constant speed has an acceleration with magnitude

$$a_{\text{C}} = \frac{v^2}{r}. \quad (6.26)$$

The direction of the acceleration vector is toward the center of the circle (**Figure 6.18**). This is a radial acceleration and is called the **centripetal acceleration**, which is why we give it the subscript c. The word *centripetal* comes from the Latin words *centrum* (meaning "center") and *petere* (meaning to seek"), and thus takes the meaning "center seeking."

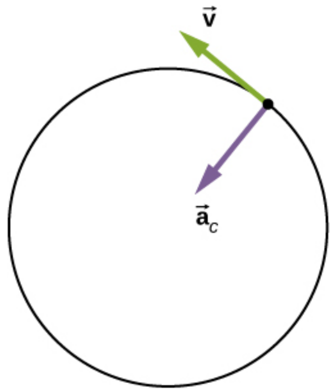


Figure 6.18 The centripetal acceleration vector points toward the center of the circular path of motion and is an acceleration in the radial direction. The velocity vector is also shown and is tangent to the circle.

Let's investigate some examples that illustrate the relative magnitudes of the velocity, radius, and centripetal acceleration.

Example 6.8

Creating an Acceleration of 1 *g*

A jet is flying at 134.1 m/s along a straight line and makes a turn along a circular path level with the ground. What does the radius of the circle have to be to produce a centripetal acceleration of 1 *g* on the pilot and jet toward the center of the circular trajectory?

Strategy

Given the speed of the jet, we can solve for the radius of the circle in the expression for the centripetal acceleration.

Solution

Set the centripetal acceleration equal to the acceleration of gravity: $9.8\text{ m/s}^2 = v^2/r$.

Solving for the radius, we find

$$r = \frac{(134.1\text{ m/s})^2}{9.8\text{ m/s}^2} = 1835\text{ m} = 1.835\text{ km}.$$

Significance

To create a greater acceleration than *g* on the pilot, the jet would either have to decrease the radius of its circular trajectory or increase its speed on its existing trajectory or both.

6.4 Check Your Understanding A flywheel has a radius of 20.0 cm. What is the speed of a point on the edge of the flywheel if it experiences a centripetal acceleration of 900.0 cm/s^2?

Centripetal acceleration can have a wide range of values, depending on the speed and radius of curvature of the circular path. Typical centripetal accelerations are given in the following table.

Object	Centripetal Acceleration (m/s^2 or factors of *g*)
Earth around the Sun	5.93×10^{-3}

Table 6.1 Typical Centripetal Accelerations

Object	Centripetal Acceleration (m/s^2 or factors of *g*)
Moon around the Earth	2.73×10^{-3}
Satellite in geosynchronous orbit	0.233
Outer edge of a CD when playing	5.78
Jet in a barrel roll	(2–3 *g*)
Roller coaster	(5 *g*)
Electron orbiting a proton in a simple Bohr model of the atom	9.0×10^{22}

Table 6.1 Typical Centripetal Accelerations

Equations of Motion for Uniform Circular Motion

A particle executing circular motion can be described by its position vector $\vec{\mathbf{r}}(t)$. **Figure 6.19** shows a particle executing circular motion in a counterclockwise direction. As the particle moves on the circle, its position vector sweeps out the angle θ with the *x*-axis. Vector $\vec{\mathbf{r}}(t)$ making an angle θ with the *x*-axis is shown with its components along the *x*- and *y*-axes. The magnitude of the position vector is $A = \left|\vec{\mathbf{r}}(t)\right|$ and is also the radius of the circle, so that in terms of its components,

$$\vec{\mathbf{r}}(t) = A\cos\omega t\,\hat{\mathbf{i}} + A\sin\omega t\,\hat{\mathbf{j}}. \tag{6.27}$$

Here, ω is a constant called the **angular frequency** of the particle. The angular frequency has units of radians (rad) per second and is simply the number of radians of angular measure through which the particle passes per second. The angle θ that the position vector has at any particular time is ωt.

If *T* is the period of motion, or the time to complete one revolution (2π rad), then

$$\omega = \frac{2\pi}{T}.$$

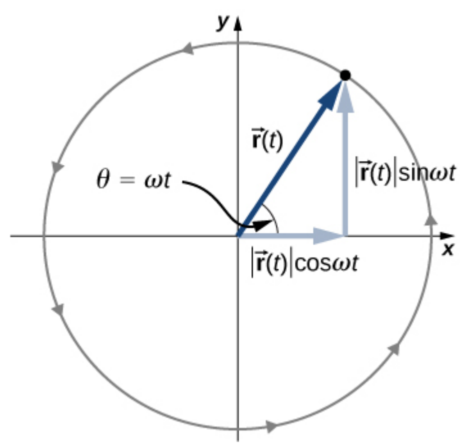


Figure 6.19 The position vector for a particle in circular motion with its components along the *x*- and *y*-axes. The particle moves counterclockwise. Angle θ is the angular frequency ω in radians per second multiplied by *t*.

Velocity and acceleration can be obtained from the position function by differentiation:

$$\vec{\mathbf{v}}(t) = \frac{d\vec{\mathbf{r}}(t)}{dt} = -A\omega \sin \omega t \hat{\mathbf{i}} + A\omega \cos \omega t \hat{\mathbf{j}}. \quad (6.28)$$

It can be shown from **Figure 6.19** that the velocity vector is tangential to the circle at the location of the particle, with magnitude $A\omega$. Similarly, the acceleration vector is found by differentiating the velocity:

$$\vec{\mathbf{a}}(t) = \frac{d\vec{\mathbf{v}}(t)}{dt} = -A\omega^2 \cos \omega t \hat{\mathbf{i}} - A\omega^2 \sin \omega t \hat{\mathbf{j}}. \quad (6.29)$$

From this equation we see that the acceleration vector has magnitude $A\omega^2$ and is directed opposite the position vector, toward the origin, because $\vec{\mathbf{a}}(t) = -\omega^2 \vec{\mathbf{r}}(t)$.

Example 6.9

Circular Motion of a Proton

A proton has speed 5×10^6 m/s and is moving in a circle in the *xy* plane of radius *r* = 0.175 m. What is its position in the *xy* plane at time $t = 2.0 \times 10^{-7}$ s = 200 ns? At *t* = 0, the position of the proton is $0.175 \text{ m} \hat{\mathbf{i}}$ and it circles counterclockwise. Sketch the trajectory.

Solution

From the given data, the proton has period and angular frequency:

$$T = \frac{2\pi r}{v} = \frac{2\pi(0.175 \text{ m})}{5.0 \times 10^6 \text{ m/s}} = 2.20 \times 10^{-7} \text{ s}$$

$$\omega = \frac{2\pi}{T} = \frac{2\pi}{2.20 \times 10^{-7} \text{ s}} = 2.856 \times 10^7 \text{ rad/s}.$$

The position of the particle at $t = 2.0 \times 10^{-7}$ s with *A* = 0.175 m is

$$\begin{aligned}
\vec{\mathbf{r}}(2.0 \times 10^{-7} \text{s}) &= A \cos \omega(2.0 \times 10^{-7} \text{ s}) \hat{\mathbf{i}} + A \sin \omega(2.0 \times 10^{-7} \text{ s}) \hat{\mathbf{j}} \text{ m} \\
&= 0.175\cos[(2.856 \times 10^7 \text{ rad/s})(2.0 \times 10^{-7} \text{ s})] \hat{\mathbf{i}} \\
&\qquad + 0.175\sin[(2.856 \times 10^7 \text{ rad/s})(2.0 \times 10^{-7} \text{ s})] \hat{\mathbf{j}} \text{ m} \\
&= 0.175\cos(5.712 \text{ rad}) \hat{\mathbf{i}} + 0.175\sin(5.712 \text{ rad}) \hat{\mathbf{j}} = 0.147 \hat{\mathbf{i}} - 0.095 \hat{\mathbf{j}} \text{ m}.
\end{aligned}$$

From this result we see that the proton is located slightly below the *x*-axis. This is shown in **Figure 6.20**.

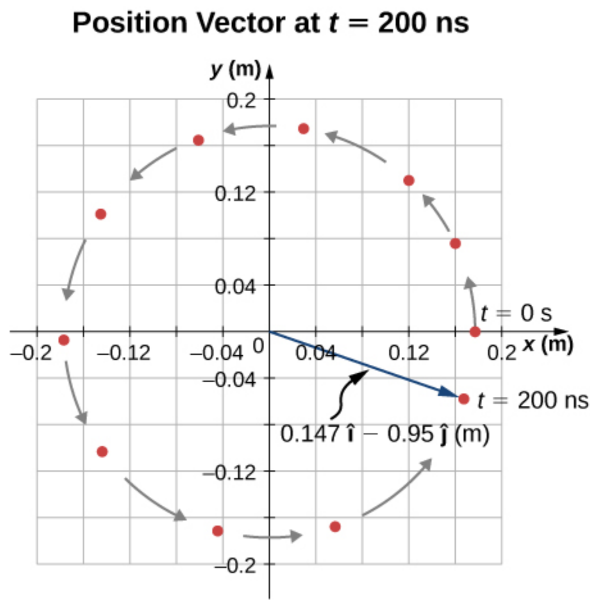


Figure 6.20 Position vector of the proton at $t = 2.0 \times 10^{-7}\,\text{s} = 200\,\text{ns}.$ The trajectory of the proton is shown. The angle through which the proton travels along the circle is 5.712 rad, which a little less than one complete revolution.

Significance

We picked the initial position of the particle to be on the *x*-axis. This was completely arbitrary. If a different starting position were given, we would have a different final position at *t* = 200 ns.

Nonuniform Circular Motion

Circular motion does not have to be at a constant speed. A particle can travel in a circle and speed up or slow down, showing an acceleration in the direction of the motion.

In uniform circular motion, the particle executing circular motion has a constant speed and the circle is at a fixed radius. If the speed of the particle is changing as well, then we introduce an additional acceleration in the direction tangential to the circle. Such accelerations occur at a point on a top that is changing its spin rate, or any accelerating rotor. In Section 6.1 we showed that centripetal acceleration is the time rate of change of the direction of the velocity vector. If the speed of the particle is changing, then it has a **tangential acceleration** that is the time rate of change of the magnitude of the velocity:

$$a_{\text{T}} = \frac{d|\vec{\mathbf{v}}|}{dt}. \quad (6.30)$$

The direction of tangential acceleration is tangent to the circle whereas the direction of centripetal acceleration is radially inward toward the center of the circle. Thus, a particle in circular motion with a tangential acceleration has a **total acceleration** that is the vector sum of the centripetal and tangential accelerations:

$$\vec{\mathbf{a}} = \vec{\mathbf{a}}_{\text{C}} + \vec{\mathbf{a}}_{\text{T}}. \quad (6.31)$$

The acceleration vectors are shown in **Figure 6.21**. Note that the two acceleration vectors $\vec{\mathbf{a}}_C$ and $\vec{\mathbf{a}}_T$ are perpendicular to each other, with $\vec{\mathbf{a}}_C$ in the radial direction and $\vec{\mathbf{a}}_T$ in the tangential direction. The total acceleration $\vec{\mathbf{a}}$ points at an angle between $\vec{\mathbf{a}}_C$ and $\vec{\mathbf{a}}_T$.

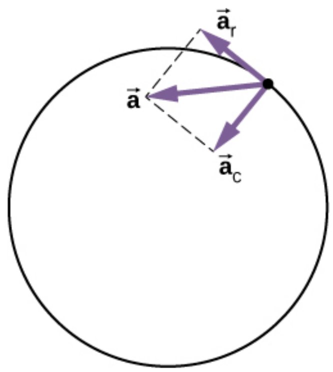


Figure 6.21 The centripetal acceleration points toward the center of the circle. The tangential acceleration is tangential to the circle at the particle's position. The total acceleration is the vector sum of the tangential and centripetal accelerations, which are perpendicular.

Example 6.10

Total Acceleration during Circular Motion

A particle moves in a circle of radius r = 2.0 m. During the time interval from t = 1.5 s to t = 4.0 s its speed varies with time according to

$$v(t) = c_1 - \frac{c_2}{t^2},\quad c_1 = 4.0 \text{ m/s},\ c_2 = 6.0 \text{ m} \cdot \text{s}.$$

What is the total acceleration of the particle at t = 2.0 s?

Strategy

We are given the speed of the particle and the radius of the circle, so we can calculate centripetal acceleration easily. The direction of the centripetal acceleration is toward the center of the circle. We find the magnitude of the tangential acceleration by taking the derivative with respect to time of $|v(t)|$ using **Equation 6.30** and evaluating it at t = 2.0 s. We use this and the magnitude of the centripetal acceleration to find the total acceleration.

Solution

Centripetal acceleration is

$$v(2.0\text{s}) = \left(4.0 - \frac{6.0}{(2.0)^2}\right)\text{m/s} = 2.5 \text{ m/s}$$

$$a_C = \frac{v^2}{r} = \frac{(2.5 \text{ m/s})^2}{2.0 \text{ m}} = 3.1 \text{ m/s}^2$$

directed toward the center of the circle. Tangential acceleration is

$$a_T = \left|\frac{d\vec{\mathbf{v}}}{dt}\right| = \frac{2c_2}{t^3} = \frac{12.0}{(2.0)^3}\text{m/s}^2 = 1.5 \text{ m/s}^2.$$

Total acceleration is

$$|\vec{\mathbf{a}}| = \sqrt{3.1^2 + 1.5^2}\text{m/s}^2 = 3.44 \text{ m/s}^2$$

and $\theta = \tan^{-1}\frac{3.1}{1.5} = 64°$ from the tangent to the circle. See **Figure 6.22**.

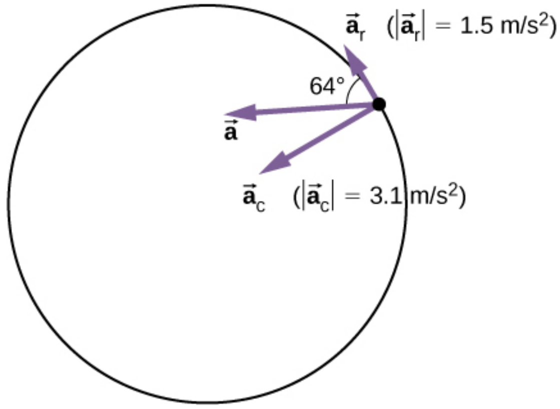


Figure 6.22 The tangential and centripetal acceleration vectors. The net acceleration $\vec{a}$ is the vector sum of the two accelerations.

Significance

The directions of centripetal and tangential accelerations can be described more conveniently in terms of a polar coordinate system, with unit vectors in the radial and tangential directions. This coordinate system, which is used for motion along curved paths, is discussed in detail later in the book.

6.5 | Relative Motion in One and Two Dimensions

Learning Objectives

By the end of this section, you will be able to:

- Explain the concept of reference frames.
- Write the position and velocity vector equations for relative motion.
- Draw the position and velocity vectors for relative motion.
- Analyze one-dimensional and two-dimensional relative motion problems using the position and velocity vector equations.

The discussion in this section will assume that the velocities of all the objects involved are much, much less than the speed of light, so that the physics of Einstein's theory of special relativity may be ignored. To explain in a nutshell what we are about to discuss, it is important in kinematics to ask the question: "Who is the observer?"

Motion Relative to What?

Motion does not happen in isolation. If you're riding in a train moving at 10 m/s east, this velocity is measured relative to the ground on which you're traveling. However, if another train passes you at 15 m/s east, your velocity relative to this other train is different from your velocity relative to the ground. Your velocity relative to the other train is 5 m/s west. To explore this idea further, we first need to establish some terminology.

Reference Frames

To discuss relative motion in one or more dimensions, we return to the concept of **reference frames**. When we say an object has a certain velocity, we must state it has a velocity with respect to a given reference frame. In most examples we have examined so far, this reference frame has been Earth. If you say a person is sitting in a train moving at 10 m/s east, then you imply the person on the train is moving relative to the surface of Earth at this velocity, and Earth is the reference frame. We can expand our view of the motion of the person on the train and say Earth is spinning in its orbit around the Sun, in which case the motion becomes more complicated. In this case, the solar system is the reference frame. In summary, all discussion of relative motion must define the reference frames involved. We now develop a method to refer to reference frames in relative motion.

Relative Motion in One Dimension

We introduce relative motion in one dimension first, because the velocity vectors simplify to having only two possible directions. Take the example of the person sitting in a train moving east. If we choose east as the positive direction and Earth as the reference frame, then we can write the velocity of the train with respect to the Earth as $\vec{\mathbf{v}}_{\text{TE}} = 10\text{ m/s }\hat{\mathbf{i}}$ east, where the subscripts TE refer to train and Earth. Let's now say the person gets up out of /her seat and walks toward the back of the train at 2 m/s. This tells us she has a velocity relative to the reference frame of the train. Since the person is walking west, in the negative direction, we write her velocity with respect to the train as $\vec{\mathbf{v}}_{\text{PT}} = -2\text{ m/s }\hat{\mathbf{i}}$. We can add the two velocity vectors to find the velocity of the person with respect to Earth. This **relative velocity** is written as

$$\vec{\mathbf{v}}_{\text{PE}} = \vec{\mathbf{v}}_{\text{PT}} + \vec{\mathbf{v}}_{\text{TE}}. \qquad (6.32)$$

Note the ordering of the subscripts for the various reference frames in Equation 6.32. The subscripts for the coupling reference frame, which is the train, appear consecutively in the right-hand side of the equation. Figure 6.23 shows the correct order of subscripts when forming the vector equation.

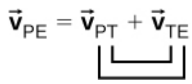


Figure 6.23 When constructing the vector equation, the subscripts for the coupling reference frame appear consecutively on the inside. The subscripts on the left-hand side of the equation are the same as the two outside subscripts on the right-hand side of the equation.

Adding the vectors, we find $\vec{\mathbf{v}}_{\text{PE}} = 8\text{ m/s }\hat{\mathbf{i}}$, so the person is moving 8 m/s east with respect to Earth. Graphically, this is shown in Figure 6.24.

Figure 6.24 Velocity vectors of the train with respect to Earth, person with respect to the train, and person with respect to Earth.

Relative Velocity in Two Dimensions

We can now apply these concepts to describing motion in two dimensions. Consider a particle P and reference frames S and S', as shown in Figure 6.25. The position of the origin of S' as measured in S is $\vec{\mathbf{r}}_{S'S}$, the position of P as measured in S' is $\vec{\mathbf{r}}_{PS'}$, and the position of P as measured in S is $\vec{\mathbf{r}}_{PS}$.

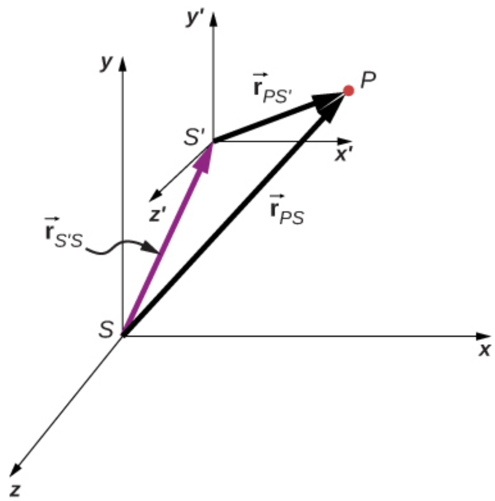


Figure 6.25 The positions of particle P relative to frames S and S' are $\vec{r}_{PS}$ and $\vec{r}_{PS'}$, respectively.

From **Figure 6.25** we see that

$$\vec{r}_{PS} = \vec{r}_{PS'} + \vec{r}_{S'S}. \tag{6.33}$$

The relative velocities are the time derivatives of the position vectors. Therefore,

$$\vec{v}_{PS} = \vec{v}_{PS'} + \vec{v}_{S'S}. \tag{6.34}$$

The velocity of a particle relative to S is equal to its velocity relative to S' plus the velocity of S' relative to S.

We can extend **Equation 6.34** to any number of reference frames. For particle P with velocities $\vec{v}_{PA}$, $\vec{v}_{PB}$, and $\vec{v}_{PC}$ in frames A, B, and C,

$$\vec{v}_{PC} = \vec{v}_{PA} + \vec{v}_{AB} + \vec{v}_{BC}. \tag{6.35}$$

We can also see how the accelerations are related as observed in two reference frames by differentiating **Equation 6.34**:

$$\vec{a}_{PS} = \vec{a}_{PS'} + \vec{a}_{S'S}. \tag{6.36}$$

We see that if the velocity of S' relative to S is a constant, then $\vec{a}_{S'S} = 0$ and

$$\vec{a}_{PS} = \vec{a}_{PS'}. \tag{6.37}$$

This says the acceleration of a particle is the same as measured by two observers moving at a constant velocity relative to each other.

Example 6.11

Motion of a Car Relative to a Truck

A truck is traveling south at a speed of 70 km/h toward an intersection. A car is traveling east toward the intersection at a speed of 80 km/h (Figure 6.26). What is the velocity of the car relative to the truck?

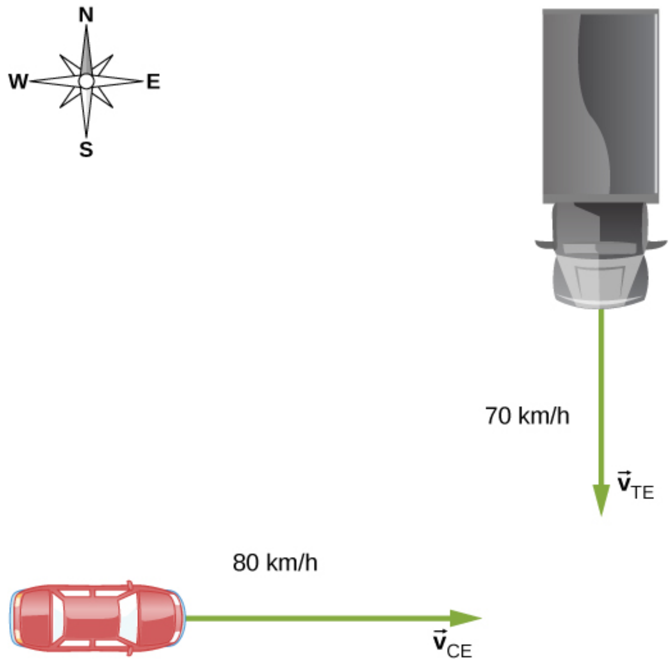


Figure 6.26 A car travels east toward an intersection while a truck travels south toward the same intersection.

Strategy

First, we must establish the reference frame common to both vehicles, which is Earth. Then, we write the velocities of each with respect to the reference frame of Earth, which enables us to form a vector equation that links the car, the truck, and Earth to solve for the velocity of the car with respect to the truck.

Solution

The velocity of the car with respect to Earth is $\vec{v}_{CE} = 80\,\text{km/h}\,\hat{i}$. The velocity of the truck with respect to Earth is $\vec{v}_{TE} = -70\,\text{km/h}\,\hat{j}$. Using the velocity addition rule, the relative motion equation we are seeking is

$$\vec{v}_{CT} = \vec{v}_{CE} + \vec{v}_{ET}.$$

Here, $\vec{v}_{CT}$ is the velocity of the car with respect to the truck, and Earth is the connecting reference frame. Since we have the velocity of the truck with respect to Earth, the negative of this vector is the velocity of Earth with respect to the truck: $\vec{v}_{ET} = -\vec{v}_{TE}$. The vector diagram of this equation is shown in Figure 6.27.

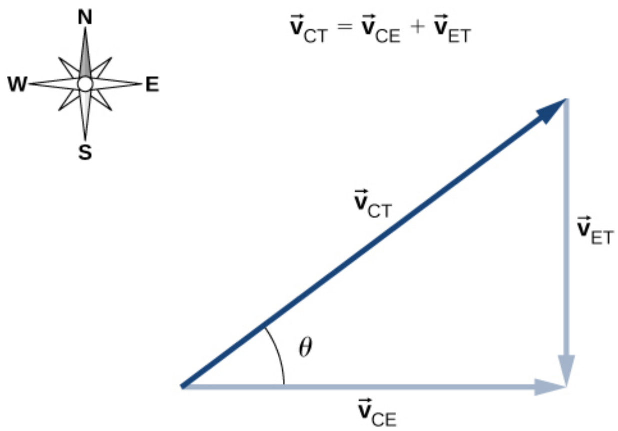


Figure 6.27 Vector diagram of the vector equation $\vec{v}_{CT} = \vec{v}_{CE} + \vec{v}_{ET}$.

We can now solve for the velocity of the car with respect to the truck:

$$|\vec{v}_{CT}| = \sqrt{(80.0\text{ km/h})^2 + (70.0\text{ km/h})^2} = 106.\text{ km/h}$$

and

$$\theta = \tan^{-1}\left(\frac{70.0}{80.0}\right) = 41.2° \text{ north of east.}$$

Significance

Drawing a vector diagram showing the velocity vectors can help in understanding the relative velocity of the two objects.

6.5 Check Your Understanding A boat heads north in still water at 4.5 m/s directly across a river that is running east at 3.0 m/s. What is the velocity of the boat with respect to Earth?

Example 6.12

Flying a Plane in a Wind

A pilot must fly his plane due north to reach his destination. The plane can fly at 300 km/h in still air. A wind is blowing out of the northeast at 90 km/h. (a) What is the speed of the plane relative to the ground? (b) In what direction must the pilot head her plane to fly due north?

Strategy

The pilot must point her plane somewhat east of north to compensate for the wind velocity. We need to construct a vector equation that contains the velocity of the plane with respect to the ground, the velocity of the plane with respect to the air, and the velocity of the air with respect to the ground. Since these last two quantities are known, we can solve for the velocity of the plane with respect to the ground. We can graph the vectors and use this diagram to evaluate the magnitude of the plane's velocity with respect to the ground. The diagram will also tell us the angle the plane's velocity makes with north with respect to the air, which is the direction the pilot must head her plane.

Solution

The vector equation is $\vec{v}_{PG} = \vec{v}_{PA} + \vec{v}_{AG}$, where P = plane, A = air, and G = ground. From the geometry in Figure 6.28, we can solve easily for the magnitude of the velocity of the plane with respect to the ground and the angle of the plane's heading, θ.

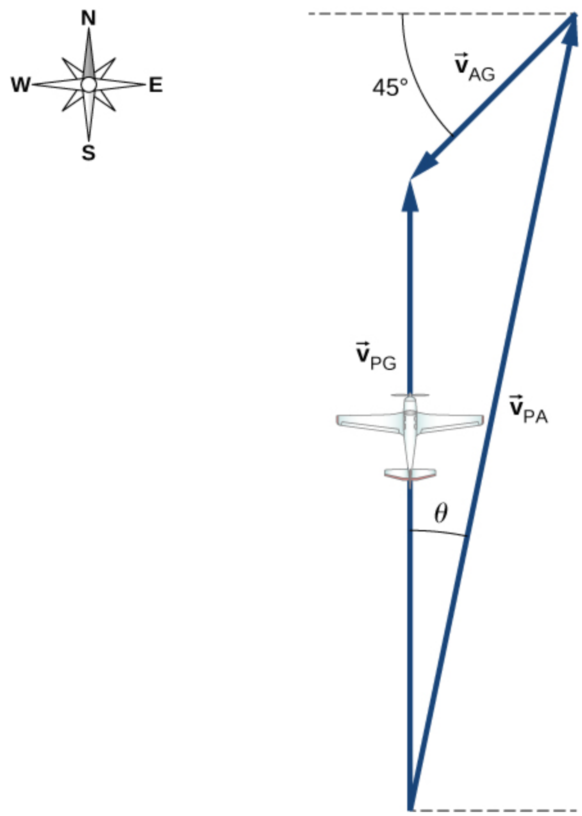


Figure 6.28 Vector diagram for **Equation 6.33** showing the vectors $\vec{v}_{PA}$, $\vec{v}_{AG}$, and $\vec{v}_{PG}$.

(a) Known quantities:

$$|\vec{v}_{PA}| = 300\ \text{km/h}$$
$$|\vec{v}_{AG}| = 90\ \text{km/h}$$

Substituting into the equation of motion, we obtain $|\vec{v}_{PG}| = 230\ \text{km/h}$.

(b) The angle $\theta = \tan^{-1}\frac{63.64}{300} = 12°$ east of north.

CHAPTER 6 REVIEW

KEY TERMS

acceleration vector instantaneous acceleration found by taking the derivative of the velocity function with respect to time in unit vector notation

angular frequency ω, rate of change of an angle with which an object that is moving on a circular path

centripetal acceleration component of acceleration of an object moving in a circle that is directed radially inward toward the center of the circle

displacement vector vector from the initial position to a final position on a trajectory of a particle

position vector vector from the origin of a chosen coordinate system to the position of a particle in two- or three-dimensional space

projectile motion motion of an object subject only to the acceleration of gravity

range maximum horizontal distance a projectile travels

reference frame coordinate system in which the position, velocity, and acceleration of an object at rest or moving is measured

relative velocity velocity of an object as observed from a particular reference frame, or the velocity of one reference frame with respect to another reference frame

tangential acceleration magnitude of which is the time rate of change of speed. Its direction is tangent to the circle.

time of flight elapsed time a projectile is in the air

total acceleration vector sum of centripetal and tangential accelerations

trajectory path of a projectile through the air

velocity vector vector that gives the instantaneous speed and direction of a particle; tangent to the trajectory

KEY EQUATIONS

Position vector	$\vec{\mathbf{r}}(t) = x(t)\hat{\mathbf{i}} + y(t)\hat{\mathbf{j}} + z(t)\hat{\mathbf{k}}$
Displacement vector	$\Delta\vec{\mathbf{r}} = \vec{\mathbf{r}}(t_2) - \vec{\mathbf{r}}(t_1)$
Velocity vector	$\vec{\mathbf{v}}(t) = \lim_{\Delta t \to 0} \frac{\vec{\mathbf{r}}(t+\Delta t) - \vec{\mathbf{r}}(t)}{\Delta t} = \frac{d\vec{\mathbf{r}}}{dt}$
Velocity in terms of components	$\vec{\mathbf{v}}(t) = v_x(t)\hat{\mathbf{i}} + v_y(t)\hat{\mathbf{j}} + v_z(t)\hat{\mathbf{k}}$
Velocity components	$v_x(t) = \frac{dx(t)}{dt} \quad v_y(t) = \frac{dy(t)}{dt} \quad v_z(t) = \frac{dz(t)}{dt}$
Average velocity	$\vec{\mathbf{v}}_{\text{avg}} = \frac{\vec{\mathbf{r}}(t_2) - \vec{\mathbf{r}}(t_1)}{t_2 - t_1}$
Instantaneous acceleration	$\vec{\mathbf{a}}(t) = \lim_{t \to 0} \frac{\vec{\mathbf{v}}(t+\Delta t) - \vec{\mathbf{v}}(t)}{\Delta t} = \frac{d\vec{\mathbf{v}}(t)}{dt}$
Instantaneous acceleration, component form	$\vec{\mathbf{a}}(t) = \frac{dv_x(t)}{dt}\hat{\mathbf{i}} + \frac{dv_y(t)}{dt}\hat{\mathbf{j}} + \frac{dv_z(t)}{dt}\hat{\mathbf{k}}$
Instantaneous acceleration as second derivatives of position	$\vec{\mathbf{a}}(t) = \frac{d^2x(t)}{dt^2}\hat{\mathbf{i}} + \frac{d^2y(t)}{dt^2}\hat{\mathbf{j}} + \frac{d^2z(t)}{dt^2}\hat{\mathbf{k}}$

Time of flight (symmetric motion)	$T_{\text{tof}} = \frac{2(v_0 \sin\theta)}{g}$
Trajectory (symmetric motion)	$y = (\tan\theta_0)x - \left[\frac{g}{2(v_0 \cos\theta_0)^2}\right]x^2$
Range (symmetric motion)	$R = \frac{v_0^2 \sin 2\theta_0}{g}$
Centripetal acceleration	$a_\text{C} = \frac{v^2}{r}$
Position vector, uniform circular motion	$\vec{\mathbf{r}}(t) = A\cos\omega t\,\hat{\mathbf{i}} + A\sin\omega t\,\hat{\mathbf{j}}$
Velocity vector, uniform circular motion	$\vec{\mathbf{v}}(t) = \frac{d\vec{\mathbf{r}}(t)}{dt} = -A\omega\sin\omega t\,\hat{\mathbf{i}} + A\omega\cos\omega t\,\hat{\mathbf{j}}$
Acceleration vector, uniform circular motion	$\vec{\mathbf{a}}(t) = \frac{d\vec{\mathbf{v}}(t)}{dt} = -A\omega^2\cos\omega t\,\hat{\mathbf{i}} - A\omega^2\sin\omega t\,\hat{\mathbf{j}}$
Tangential acceleration	$a_\text{T} = \frac{d\lvert\vec{\mathbf{v}}\rvert}{dt}$
Total acceleration	$\vec{\mathbf{a}} = \vec{\mathbf{a}}_\text{C} + \vec{\mathbf{a}}_\text{T}$
Position vector in frame S is the position vector in frame S' plus the vector from the origin of S to the origin of S'	$\vec{\mathbf{r}}_{PS} = \vec{\mathbf{r}}_{PS'} + \vec{\mathbf{r}}_{S'S}$
Relative velocity equation connecting two reference frames	$\vec{\mathbf{v}}_{PS} = \vec{\mathbf{v}}_{PS'} + \vec{\mathbf{v}}_{S'S}$
Relative velocity equation connecting more than two reference frames	$\vec{\mathbf{v}}_{PC} = \vec{\mathbf{v}}_{PA} + \vec{\mathbf{v}}_{AB} + \vec{\mathbf{v}}_{BC}$
Relative acceleration equation	$\vec{\mathbf{a}}_{PS} = \vec{\mathbf{a}}_{PS'} + \vec{\mathbf{a}}_{S'S}$

SUMMARY

6.1 Displacement and Velocity Vectors

- The position function $\vec{\mathbf{r}}(t)$ gives the position as a function of time of a particle moving in two or three dimensions. Graphically, it is a vector from the origin of a chosen coordinate system to the point where the particle is located at a specific time.
- The displacement vector $\Delta\vec{\mathbf{r}}$ gives the shortest distance between any two points on the trajectory of a particle in two or three dimensions.
- Instantaneous velocity gives the speed and direction of a particle at a specific time on its trajectory in two or three dimensions, and is a vector in two and three dimensions.
- The velocity vector is tangent to the trajectory of the particle.
- Displacement $\vec{\mathbf{r}}(t)$ can be written as a vector sum of the one-dimensional displacements $\vec{x}(t)$, $\vec{y}(t)$, $\vec{z}(t)$ along the *x*, *y*, and *z* directions.
- Velocity $\vec{\mathbf{v}}(t)$ can be written as a vector sum of the one-dimensional velocities $v_x(t)$, $v_y(t)$, $v_z(t)$ along the *x*, *y*, and *z* directions.

- Motion in any given direction is independent of motion in a perpendicular direction.

6.2 Acceleration Vector

- In two and three dimensions, the acceleration vector can have an arbitrary direction and does not necessarily point along a given component of the velocity.
- The instantaneous acceleration is produced by a change in velocity taken over a very short (infinitesimal) time period. Instantaneous acceleration is a vector in two or three dimensions. It is found by taking the derivative of the velocity function with respect to time.
- In three dimensions, acceleration $\vec{\mathbf{a}}(t)$ can be written as a vector sum of the one-dimensional accelerations $a_x(t)$, $a_y(t)$, and $a_z(t)$ along the *x*-, *y*-, and *z*-axes.
- The kinematic equations for constant acceleration can be written as the vector sum of the constant acceleration equations in the *x*, *y*, and *z* directions.

6.3 Projectile Motion

- Projectile motion is the motion of an object subject only to the acceleration of gravity, where the acceleration is constant, as near the surface of Earth.
- To solve projectile motion problems, we analyze the motion of the projectile in the horizontal and vertical directions using the one-dimensional kinematic equations for *x* and *y*.
- The time of flight of a projectile launched with initial vertical velocity v_{0y} on an even surface is given by

$$T_{tof} = \frac{2(v_0 \sin\theta)}{g}.$$

  This equation is valid only when the projectile lands at the same elevation from which it was launched.
- The maximum horizontal distance traveled by a projectile is called the range. Again, the equation for range is valid only when the projectile lands at the same elevation from which it was launched.

6.4 Uniform Circular Motion

- Uniform circular motion is motion in a circle at constant speed.
- Centripetal acceleration $\vec{\mathbf{a}}_C$ is the acceleration a particle must have to follow a circular path. Centripetal acceleration always points toward the center of rotation and has magnitude $a_C = v^2/r$.
- Nonuniform circular motion occurs when there is tangential acceleration of an object executing circular motion such that the speed of the object is changing. This acceleration is called tangential acceleration $\vec{\mathbf{a}}_T$. The magnitude of tangential acceleration is the time rate of change of the magnitude of the velocity. The tangential acceleration vector is tangential to the circle, whereas the centripetal acceleration vector points radially inward toward the center of the circle. The total acceleration is the vector sum of tangential and centripetal accelerations.
- An object executing uniform circular motion can be described with equations of motion. The position vector of the object is $\vec{\mathbf{r}}(t) = A\cos\omega t\,\hat{\mathbf{i}} + A\sin\omega t\,\hat{\mathbf{j}}$, where *A* is the magnitude $|\vec{\mathbf{r}}(t)|$, which is also the radius of the circle, and ω is the angular frequency.

6.5 Relative Motion in One and Two Dimensions

- When analyzing motion of an object, the reference frame in terms of position, velocity, and acceleration needs to be specified.
- Relative velocity is the velocity of an object as observed from a particular reference frame, and it varies with the choice of reference frame.

- If S and S' are two reference frames moving relative to each other at a constant velocity, then the velocity of an object relative to S is equal to its velocity relative to S' plus the velocity of S' relative to S.
- If two reference frames are moving relative to each other at a constant velocity, then the accelerations of an object as observed in both reference frames are equal.

CONCEPTUAL QUESTIONS

6.1 Displacement and Velocity Vectors

1. What form does the trajectory of a particle have if the distance from any point *A* to point *B* is equal to the magnitude of the displacement from *A* to *B*?

2. Give an example of a trajectory in two or three dimensions caused by independent perpendicular motions.

3. If the instantaneous velocity is zero, what can be said about the slope of the position function?

6.2 Acceleration Vector

4. If the position function of a particle is a linear function of time, what can be said about its acceleration?

5. If an object has a constant *x*-component of the velocity and suddenly experiences an acceleration in the *y* direction, does the *x*-component of its velocity change?

6. If an object has a constant *x*-component of velocity and suddenly experiences an acceleration at an angle of $70°$ in the *x* direction, does the *x*-component of velocity change?

6.3 Projectile Motion

7. Answer the following questions for projectile motion on level ground assuming negligible air resistance, with the initial angle being neither $0°$ nor $90°$: (a) Is the velocity ever zero? (b) When is the velocity a minimum? A maximum? (c) Can the velocity ever be the same as the initial velocity at a time other than at $t = 0$? (d) Can the speed ever be the same as the initial speed at a time other than at $t = 0$?

8. Answer the following questions for projectile motion on level ground assuming negligible air resistance, with the initial angle being neither $0°$ nor $90°$: (a) Is the acceleration ever zero? (b) Is the acceleration ever in the same direction as a component of velocity? (c) Is the acceleration ever opposite in direction to a component of velocity?

9. A dime is placed at the edge of a table so it hangs over slightly. A quarter is slid horizontally on the table surface perpendicular to the edge and hits the dime head on. Which coin hits the ground first?

6.4 Uniform Circular Motion

10. Can centripetal acceleration change the speed of a particle undergoing circular motion?

11. Can tangential acceleration change the speed of a particle undergoing circular motion?

6.5 Relative Motion in One and Two Dimensions

12. What frame or frames of reference do you use instinctively when driving a car? When flying in a commercial jet?

13. A basketball player dribbling down the court usually keeps his eyes fixed on the players around him. He is moving fast. Why doesn't he need to keep his eyes on the ball?

14. If someone is riding in the back of a pickup truck and throws a softball straight backward, is it possible for the ball to fall straight down as viewed by a person standing at the side of the road? Under what condition would this occur? How would the motion of the ball appear to the person who threw it?

15. The hat of a jogger running at constant velocity falls off the back of his head. Draw a sketch showing the path of the hat in the jogger's frame of reference. Draw its path as viewed by a stationary observer. Neglect air resistance.

16. A clod of dirt falls from the bed of a moving truck. It strikes the ground directly below the end of the truck. (a) What is the direction of its velocity relative to the truck just before it hits? (b) Is this the same as the direction of its velocity relative to ground just before it hits? Explain your answers.

PROBLEMS

6.1 Displacement and Velocity Vectors

17. The coordinates of a particle in a rectangular coordinate system are (1.0, –4.0, 6.0). What is the position vector of the particle?

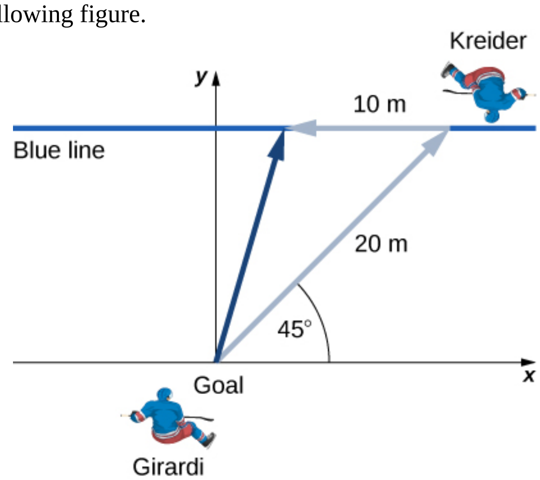
18. The position of a particle changes from $\vec{\mathbf{r}}_1 = (2.0\,\hat{\mathbf{i}} + 3.0\,\hat{\mathbf{j}})\text{cm}$ to $\vec{\mathbf{r}}_2 = (-4.0\,\hat{\mathbf{i}} + 3.0\,\hat{\mathbf{j}})\,\text{cm}$. What is the particle's displacement?

19. The 18th hole at Pebble Beach Golf Course is a dogleg to the left of length 496.0 m. The fairway off the tee is taken to be the *x* direction. A golfer hits his tee shot a distance of 300.0 m, corresponding to a displacement $\Delta\vec{\mathbf{r}}_1 = 300.0\,\text{m}\,\hat{\mathbf{i}}$, and hits his second shot 189.0 m with a displacement $\Delta\vec{\mathbf{r}}_2 = 172.0\,\text{m}\,\hat{\mathbf{i}} + 80.3\,\text{m}\,\hat{\mathbf{j}}$. What is the final displacement of the golf ball from the tee?

20. A bird flies straight northeast a distance of 95.0 km for 3.0 h. With the *x*-axis due east and the *y*-axis due north, what is the displacement in unit vector notation for the bird? What is the average velocity for the trip?

21. A cyclist rides 5.0 km due east, then 10.0 km $20°$ west of north. From this point she rides 8.0 km due west. What is the final displacement from where the cyclist started?

22. New York Rangers defenseman Daniel Girardi stands at the goal and passes a hockey puck 20 m and $45°$ from straight down the ice to left wing Chris Kreider waiting at the blue line. Kreider waits for Girardi to reach the blue line and passes the puck directly across the ice to him 10 m away. What is the final displacement of the puck? See the following figure.



23. The position of a particle is $\vec{\mathbf{r}}(t) = 4.0t^2\,\hat{\mathbf{i}} - 3.0\,\hat{\mathbf{j}} + 2.0t^3\,\hat{\mathbf{k}}\,\text{m}$. (a) What is the velocity of the particle at 0 s and at 1.0 s? (b) What is the average velocity between 0 s and 1.0 s?

24. Clay Matthews, a linebacker for the Green Bay Packers, can reach a speed of 10.0 m/s. At the start of a play, Matthews runs downfield at $45°$ with respect to the 50-yard line and covers 8.0 m in 1 s. He then runs straight down the field at $90°$ with respect to the 50-yard line for 12 m, with an elapsed time of 1.2 s. (a) What is Matthews' final displacement from the start of the play? (b) What is his average velocity?

25. The F-35B Lighting II is a short-takeoff and vertical landing fighter jet. If it does a vertical takeoff to 20.00-m height above the ground and then follows a flight path angled at $30°$ with respect to the ground for 20.00 km, what is the final displacement?

6.2 Acceleration Vector

26. A particle's acceleration is $(4.0\,\hat{\mathbf{i}} + 3.0\,\hat{\mathbf{j}})\text{m/s}^2$. At *t* = 0, its position and velocity are zero. (a) What are the particle's position and velocity as functions of time? (b) Find the equation of the path of the particle. Draw the *x*- and *y*-axes and sketch the trajectory of the particle.

27. A boat leaves the dock at *t* = 0 and heads out into a lake with an acceleration of $2.0\,\text{m/s}^2\,\hat{\mathbf{i}}$. A strong wind is pushing the boat, giving it an additional velocity of $2.0\,\text{m/s}\,\hat{\mathbf{i}} + 1.0\,\text{m/s}\,\hat{\mathbf{j}}$. (a) What is the velocity of the boat at *t* = 10 s? (b) What is the position of the boat at *t* = 10s? Draw a sketch of the boat's trajectory and position at *t* = 10 s, showing the *x*- and *y*-axes.

28. The acceleration of a particle is a constant. At *t* = 0 the velocity of the particle is $(10\,\hat{\mathbf{i}} + 20\,\hat{\mathbf{j}})\text{m/s}$. At *t* = 4 s the velocity is $10\,\hat{\mathbf{j}}\,\text{m/s}$. (a) What is the particle's acceleration? (b) How do the position and velocity vary with time? Assume the particle is initially at the origin.

29. A Lockheed Martin F-35 II Lighting jet takes off from an aircraft carrier with a runway length of 90 m and a takeoff speed 70 m/s at the end of the runway. Jets are catapulted into airspace from the deck of an aircraft carrier with two sources of propulsion: the jet propulsion and the catapult. At the point of leaving the deck of the aircraft carrier, the F-35's acceleration decreases to a constant acceleration of $5.0\,\text{m/s}^2$ at $30°$ with respect to the horizontal. (a) What is the initial acceleration of the F-35 on the deck of the aircraft carrier to make it airborne? (b) Write the position and velocity of the F-35 in unit vector notation from the point it leaves the deck of the aircraft carrier. (c) At what altitude is the fighter 5.0 s after it leaves the deck of the aircraft carrier? (d) What is its velocity and speed at this time? (e) How far has it traveled horizontally?

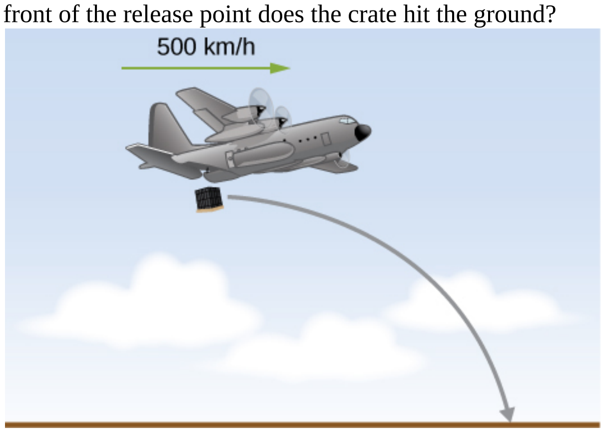
6.3 Projectile Motion

30. A bullet is shot horizontally from shoulder height (1.5 m) with and initial speed 200 m/s. (a) How much time elapses before the bullet hits the ground? (b) How far does the bullet travel horizontally?

31. A marble rolls off a tabletop 1.0 m high and hits the floor at a point 3.0 m away from the table's edge in the horizontal direction. (a) How long is the marble in the air? (b) What is the speed of the marble when it leaves the table's edge? (c) What is its speed when it hits the floor?

32. A dart is thrown horizontally at a speed of 10 m/s at the bull's-eye of a dartboard 2.4 m away, as in the following figure. (a) How far below the intended target does the dart hit? (b) What does your answer tell you about how proficient dart players throw their darts?

33. An airplane flying horizontally with a speed of 500 km/h at a height of 800 m drops a crate of supplies (see the following figure). If the parachute fails to open, how far in front of the release point does the crate hit the ground?



34. Suppose the airplane in the preceding problem fires a projectile horizontally in its direction of motion at a speed of 300 m/s relative to the plane. (a) How far in front of the release point does the projectile hit the ground? (b) What is its speed when it hits the ground?

35. A fastball pitcher can throw a baseball at a speed of 40 m/s (90 mi/h). (a) Assuming the pitcher can release the ball 16.7 m from home plate so the ball is moving horizontally, how long does it take the ball to reach home plate? (b) How far does the ball drop between the pitcher's hand and home plate?

36. A projectile is launched at an angle of $30°$ and lands 20 s later at the same height as it was launched. (a) What is the initial speed of the projectile? (b) What is the maximum altitude? (c) What is the range? (d) Calculate the displacement from the point of launch to the position on its trajectory at 15 s.

37. A basketball player shoots toward a basket 6.1 m away and 3.0 m above the floor. If the ball is released 1.8 m above the floor at an angle of $60°$ above the horizontal, what must the initial speed be if it were to go through the basket?

38. At a particular instant, a hot air balloon is 100 m in the air and descending at a constant speed of 2.0 m/s. At this exact instant, a girl throws a ball horizontally, relative to herself, with an initial speed of 20 m/s. When she lands, where will she find the ball? Ignore air resistance.

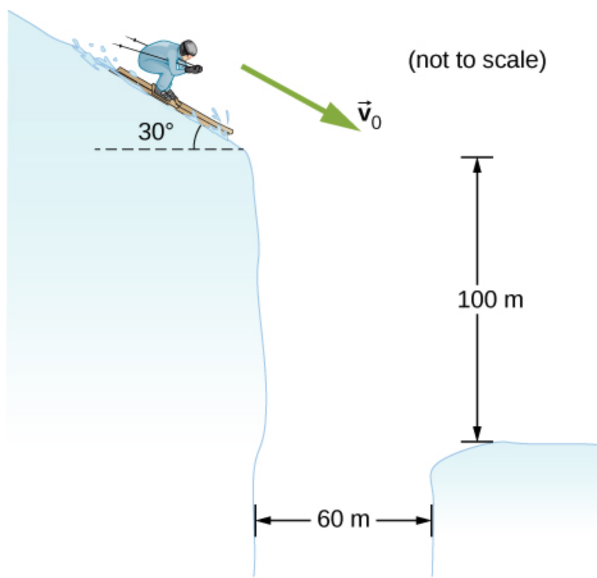
39. A man on a motorcycle traveling at a uniform speed of 10 m/s throws an empty can straight upward relative to himself with an initial speed of 3.0 m/s. Find the equation of the trajectory as seen by a police officer on the side of the road. Assume the initial position of the can is the point where it is thrown. Ignore air resistance.

40. An athlete can jump a distance of 8.0 m in the broad jump. What is the maximum distance the athlete can jump on the Moon, where the gravitational acceleration is one-sixth that of Earth?

41. The maximum horizontal distance a boy can throw a ball is 50 m. Assume he can throw with the same initial speed at all angles. How high does he throw the ball when he throws it straight upward?

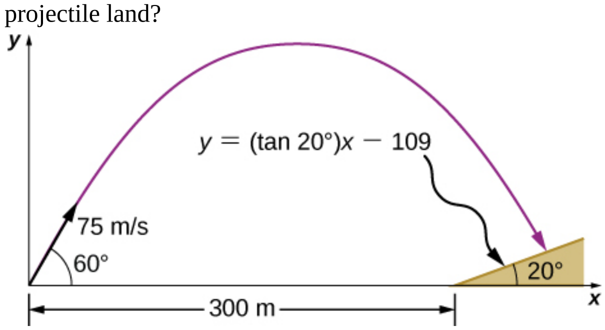
42. A rock is thrown off a cliff at an angle of $53°$ with respect to the horizontal. The cliff is 100 m high. The initial speed of the rock is 30 m/s. (a) How high above the edge of the cliff does the rock rise? (b) How far has it moved horizontally when it is at maximum altitude? (c) How long after the release does it hit the ground? (d) What is the range of the rock? (e) What are the horizontal and vertical positions of the rock relative to the edge of the cliff at t = 2.0 s, t = 4.0 s, and t = 6.0 s?

43. Trying to escape his pursuers, a secret agent skis off a slope inclined at $30°$ below the horizontal at 60 km/h. To survive and land on the snow 100 m below, he must clear a gorge 60 m wide. Does he make it? Ignore air resistance.



44. A golfer on a fairway is 70 m away from the green, which sits below the level of the fairway by 20 m. If the golfer hits the ball at an angle of $40°$ with an initial speed of 20 m/s, how close to the green does she come?

45. A projectile is shot at a hill, the base of which is 300 m away. The projectile is shot at $60°$ above the horizontal with an initial speed of 75 m/s. The hill can be approximated by a plane sloped at $20°$ to the horizontal. Relative to the coordinate system shown in the following figure, the equation of this straight line is $y = (\tan 20°)x - 109.$ Where on the hill does the projectile land?



46. An astronaut on Mars kicks a soccer ball at an angle of $45°$ with an initial velocity of 15 m/s. If the acceleration of gravity on Mars is 3.7 m/s, (a) what is the range of the soccer kick on a flat surface? (b) What would be the range of the same kick on the Moon, where gravity is one-sixth that of Earth?

47. Mike Powell holds the record for the long jump of 8.95 m, established in 1991. If he left the ground at an angle of $15°,$ what was his initial speed?

48. MIT's robot cheetah can jump over obstacles 46 cm high and has speed of 12.0 km/h. (a) If the robot launches itself at an angle of $60°$ at this speed, what is its maximum height? (b) What would the launch angle have to be to reach a height of 46 cm?

49. Mt. Asama, Japan, is an active volcano. In 2009, an eruption threw solid volcanic rocks that landed 1 km horizontally from the crater. If the volcanic rocks were launched at an angle of $40°$ with respect to the horizontal and landed 900 m below the crater, (a) what would be their initial velocity and (b) what is their time of flight?

50. Drew Brees of the New Orleans Saints can throw a football 23.0 m/s (50 mph). If he angles the throw at $10°$ from the horizontal, what distance does it go if it is to be caught at the same elevation as it was thrown?

51. The Lunar Roving Vehicle used in NASA's late *Apollo* missions reached an unofficial lunar land speed of 5.0 m/s by astronaut Eugene Cernan. If the rover was moving at this speed on a flat lunar surface and hit a small bump that projected it off the surface at an angle of $20°,$ how long would it be "airborne" on the Moon?

52. A soccer goal is 2.44 m high. A player kicks the ball at a distance 10 m from the goal at an angle of $25°.$ What is the initial speed of the soccer ball?

53. Olympus Mons on Mars is the largest volcano in the solar system, at a height of 25 km and with a radius of 312 km. If you are standing on the summit, with what initial velocity would you have to fire a projectile from a cannon horizontally to clear the volcano and land on the surface of Mars? Note that Mars has an acceleration of gravity of $3.7 \text{ m/s}^2.$

54. In 1999, Robbie Knievel was the first to jump the Grand Canyon on a motorcycle. At a narrow part of the canyon (69.0 m wide) and traveling 35.8 m/s off the takeoff ramp, he reached the other side. What was his launch angle?

55. You throw a baseball at an initial speed of 15.0 m/s at an angle of $30°$ with respect to the horizontal. What would the ball's initial speed have to be at $30°$ on a planet that has twice the acceleration of gravity as Earth to achieve the same range? Consider launch and impact on a horizontal surface.

56. Aaron Rogers throws a football at 20.0 m/s to his wide receiver, who runs straight down the field at 9.4 m/s for 20.0 m. If Aaron throws the football when the wide receiver has reached 10.0 m, what angle does Aaron have to launch the ball so the receiver catches it at the 20.0 m mark?

6.4 Uniform Circular Motion

57. A flywheel is rotating at 30 rev/s. What is the total angle, in radians, through which a point on the flywheel rotates in 40 s?

58. A particle travels in a circle of radius 10 m at a constant speed of 20 m/s. What is the magnitude of the acceleration?

59. Cam Newton of the Carolina Panthers throws a perfect football spiral at 8.0 rev/s. The radius of a pro football is 8.5 cm at the middle of the short side. What is the centripetal acceleration of the laces on the football?

60. A fairground ride spins its occupants inside a flying saucer-shaped container. If the horizontal circular path the riders follow has an 8.00-m radius, at how many revolutions per minute are the riders subjected to a centripetal acceleration equal to that of gravity?

61. A runner taking part in the 200-m dash must run around the end of a track that has a circular arc with a radius of curvature of 30.0 m. The runner starts the race at a constant speed. If she completes the 200-m dash in 23.2 s and runs at constant speed throughout the race, what is her centripetal acceleration as she runs the curved portion of the track?

62. What is the acceleration of Venus toward the Sun, assuming a circular orbit?

63. An experimental jet rocket travels around Earth along its equator just above its surface. At what speed must the jet travel if the magnitude of its acceleration is *g*?

64. A fan is rotating at a constant 360.0 rev/min. What is the magnitude of the acceleration of a point on one of its blades 10.0 cm from the axis of rotation?

65. A point located on the second hand of a large clock has a radial acceleration of 0.1cm/s^2. How far is the point from the axis of rotation of the second hand?

6.5 Relative Motion in One and Two Dimensions

66. The coordinate axes of the reference frame S' remain parallel to those of *S*, as S' moves away from *S* at a constant velocity $\vec{\mathbf{v}}_{S'} = (4.0\,\hat{\mathbf{i}} + 3.0\,\hat{\mathbf{j}} + 5.0\,\hat{\mathbf{k}})\text{ m/s}$. (a) If at time *t* = 0 the origins coincide, what is the position of the origin O' in the *S* frame as a function of time? (b) How is particle position for $\vec{\mathbf{r}}(t)$ and $\vec{\mathbf{r}}\,'(t)$, as measured in *S* and S', respectively, related? (c) What is the relationship between particle velocities $\vec{\mathbf{v}}(t)$ and $\vec{\mathbf{v}}\,'(t)$? (d) How are accelerations $\vec{\mathbf{a}}(\text{t})$ and $\vec{\mathbf{a}}\,'(\text{t})$ related?

67. The coordinate axes of the reference frame S' remain parallel to those of *S*, as S' moves away from *S* at a constant velocity $\vec{\mathbf{v}}_{S'S} = (1.0\,\hat{\mathbf{i}} + 2.0\,\hat{\mathbf{j}} + 3.0\,\hat{\mathbf{k}})t\text{ m/s}$. (a) If at time *t* = 0 the origins coincide, what is the position of origin O' in the *S* frame as a function of time? (b) How is particle position for $\vec{\mathbf{r}}(t)$ and $\vec{\mathbf{r}}\,'(t)$, as measured in *S* and S', respectively, related? (c) What is the relationship between particle velocities $\vec{\mathbf{v}}(t)$ and $\vec{\mathbf{v}}\,'(t)$? (d) How are accelerations $\vec{\mathbf{a}}(\text{t})$ and $\vec{\mathbf{a}}\,'(\text{t})$ related?

68. The velocity of a particle in reference frame *A* is $(2.0\,\hat{\mathbf{i}} + 3.0\,\hat{\mathbf{j}})\text{ m/s}$. The velocity of reference frame *A* with respect to reference frame *B* is $4.0\,\hat{\mathbf{k}}\text{ m/s}$, and the velocity of reference frame *B* with respect to *C* is $2.0\,\hat{\mathbf{j}}\text{ m/s}$. What is the velocity of the particle in reference frame *C*?

69. Raindrops fall vertically at 4.5 m/s relative to the earth. What does an observer in a car moving at 22.0 m/s in a straight line measure as the velocity of the raindrops?

70. A seagull can fly at a velocity of 9.00 m/s in still air. (a) If it takes the bird 20.0 min to travel 6.00 km straight into an oncoming wind, what is the velocity of the wind? (b) If the bird turns around and flies with the wind, how long will it take the bird to return 6.00 km?

71. A ship sets sail from Rotterdam, heading due north at 7.00 m/s relative to the water. The local ocean current is 1.50 m/s in a direction $40.0°$ north of east. What is the velocity of the ship relative to Earth?

72. A boat can be rowed at 8.0 km/h in still water. (a) How much time is required to row 1.5 km downstream in a river moving 3.0 km/h relative to the shore? (b) How much time is required for the return trip? (c) In what direction must the boat be aimed to row straight across the river? (d) Suppose the river is 0.8 km wide. What is the velocity of the boat with respect to Earth and how much time is required to get to the opposite shore? (e) Suppose, instead, the boat is aimed straight across the river. How much time is required to get across and how far downstream is the boat when it reaches the opposite shore?

73. A small plane flies at 200 km/h in still air. If the wind blows directly out of the west at 50 km/h, (a) in what direction must the pilot head her plane to move directly north across land and (b) how long does it take her to reach a point 300 km directly north of her starting point?

74. A cyclist traveling southeast along a road at 15 km/h feels a wind blowing from the southwest at 25 km/h. To a stationary observer, what are the speed and direction of the wind?

75. A river is moving east at 4 m/s. A boat starts from the dock heading $30°$ north of west at 7 m/s. If the river is 1800 m wide, (a) what is the velocity of the boat with respect to Earth and (b) how long does it take the boat to cross the river?

ADDITIONAL PROBLEMS

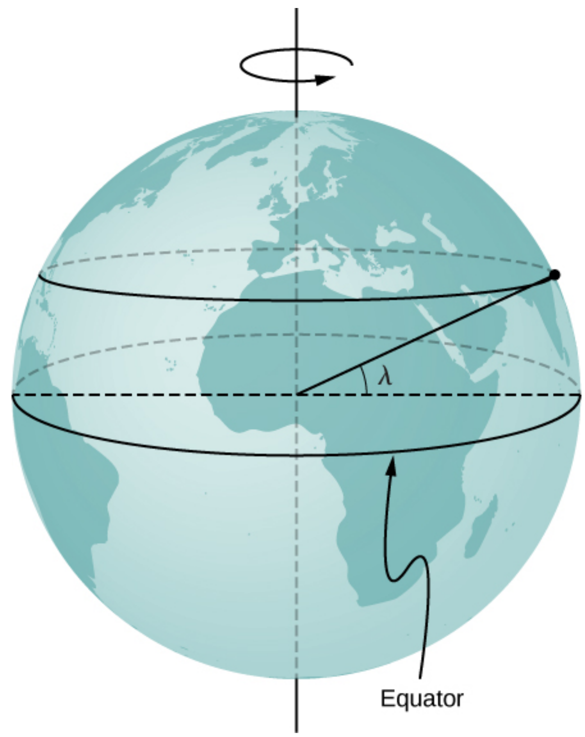
76. A Formula One race car is traveling at 89.0 m/s along a straight track enters a turn on the race track with radius of curvature of 200.0 m. What centripetal acceleration must the car have to stay on the track?

77. A particle travels in a circular orbit of radius 10 m. Its speed is changing at a rate of $15.0\,\text{m/s}^2$ at an instant when its speed is 40.0 m/s. What is the magnitude of the acceleration of the particle?

78. The driver of a car moving at 90.0 km/h presses down on the brake as the car enters a circular curve of radius 150.0 m. If the speed of the car is decreasing at a rate of 9.0 km/h each second, what is the magnitude of the acceleration of the car at the instant its speed is 60.0 km/h?

79. A race car entering the curved part of the track at the Daytona 500 drops its speed from 85.0 m/s to 80.0 m/s in 2.0 s. If the radius of the curved part of the track is 316.0 m, calculate the total acceleration of the race car at the beginning and ending of reduction of speed.

80. An elephant is located on Earth's surface at a latitude λ. Calculate the centripetal acceleration of the elephant resulting from the rotation of Earth around its polar axis. Express your answer in terms of λ, the radius R_E of Earth, and time T for one rotation of Earth. Compare your answer with g for $\lambda = 40°$.



81. A proton in a synchrotron is moving in a circle of radius 1 km and increasing its speed by $v(t) = c_1 + c_2 t^2$, where $c_1 = 2.0 \times 10^5\,\text{m/s}$, $c_2 = 10^5\,\text{m/s}^3$. (a) What is the proton's total acceleration at t = 5.0 s? (b) At what time does the expression for the velocity become unphysical?

82. A propeller blade at rest starts to rotate from t = 0 s to t = 5.0 s with a tangential acceleration of the tip of the blade at $3.00\,\text{m/s}^2$. The tip of the blade is 1.5 m from the axis of rotation. At t = 5.0 s, what is the total acceleration of the tip of the blade?

83. A particle is executing circular motion with a constant angular frequency of $\omega = 4.00\,\text{rad/s}$. If time t = 0 corresponds to the position of the particle being located at y = 0 m and x = 5 m, (a) what is the position of the particle at t = 10 s? (b) What is its velocity at this time? (c) What is its acceleration?

84. A particle's centripetal acceleration is $a_C = 4.0 \text{ m/s}^2$ at $t = 0$ s. It is executing uniform circular motion about an axis at a distance of 5.0 m. What is its velocity at $t = 10$ s?

85. A rod 3.0 m in length is rotating at 2.0 rev/s about an axis at one end. Compare the centripetal accelerations at radii of (a) 1.0 m, (b) 2.0 m, and (c) 3.0 m.

86. A particle located initially at $(1.5\hat{\mathbf{j}} + 4.0\hat{\mathbf{k}})$m undergoes a displacement of $(2.5\hat{\mathbf{i}} + 3.2\hat{\mathbf{j}} - 1.2\hat{\mathbf{k}})$ m. What is the final position of the particle?

87. The position of a particle is given by $\vec{\mathbf{r}}(t) = (50 \text{ m/s})t\hat{\mathbf{i}} - (4.9 \text{ m/s}^2)t^2\hat{\mathbf{j}}$. (a) What are the particle's velocity and acceleration as functions of time? (b) What are the initial conditions to produce the motion?

88. A spaceship is traveling at a constant velocity of $\vec{\mathbf{v}}(t) = 250.0\hat{\mathbf{i}}$ m/s when its rockets fire, giving it an acceleration of $\vec{\mathbf{a}}(t) = (3.0\hat{\mathbf{i}} + 4.0\hat{\mathbf{k}})\text{m/s}^2$. What is its velocity 5 s after the rockets fire?

89. A crossbow is aimed horizontally at a target 40 m away. The arrow hits 30 cm below the spot at which it was aimed. What is the initial velocity of the arrow?

90. A long jumper can jump a distance of 8.0 m when he takes off at an angle of $45°$ with respect to the horizontal. Assuming he can jump with the same initial speed at all angles, how much distance does he lose by taking off at $30°$?

91. On planet Arcon, the maximum horizontal range of a projectile launched at 10 m/s is 20 m. What is the acceleration of gravity on this planet?

92. A mountain biker encounters a jump on a race course that sends him into the air at $60°$ to the horizontal. If he lands at a horizontal distance of 45.0 m and 20 m below his launch point, what is his initial speed?

93. Which has the greater centripetal acceleration, a car with a speed of 15.0 m/s along a circular track of radius 100.0 m or a car with a speed of 12.0 m/s along a circular track of radius 75.0 m?

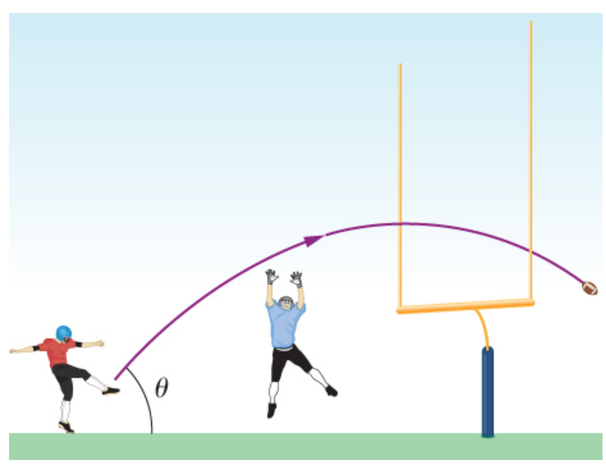
94. A geosynchronous satellite orbits Earth at a distance of 42,250.0 km and has a period of 1 day. What is the centripetal acceleration of the satellite?

95. Two speedboats are traveling at the same speed relative to the water in opposite directions in a moving river. An observer on the riverbank sees the boats moving at 4.0 m/s and 5.0 m/s. (a) What is the speed of the boats relative to the river? (b) How fast is the river moving relative to the shore?

CHALLENGE PROBLEMS

96. World's Longest Par 3. The tee of the world's longest par 3 sits atop South Africa's Hanglip Mountain at 400.0 m above the green and can only be reached by helicopter. The horizontal distance to the green is 359.0 m. Neglect air resistance and answer the following questions. (a) If a golfer launches a shot that is $40°$ with respect to the horizontal, what initial velocity must she give the ball? (b) What is the time to reach the green?

97. When a field goal kicker kicks a football as hard as he can at $45°$ to the horizontal, the ball just clears the 3-m-high crossbar of the goalposts 45.7 m away. (a) What is the maximum speed the kicker can impart to the football? (b) In addition to clearing the crossbar, the football must be high enough in the air early during its flight to clear the reach of the onrushing defensive lineman. If the lineman is 4.6 m away and has a vertical reach of 2.5 m, can he block the 45.7-m field goal attempt? (c) What if the lineman is 1.0 m away?



98. A truck is traveling east at 80 km/h. At an intersection 32 km ahead, a car is traveling north at 50 km/h. (a) How long after this moment will the vehicles be closest to each other? (b) How far apart will they be at that point?

7 | OVERVIEW OF THE SOLAR SYSTEM

Chapter Outline

Introduction

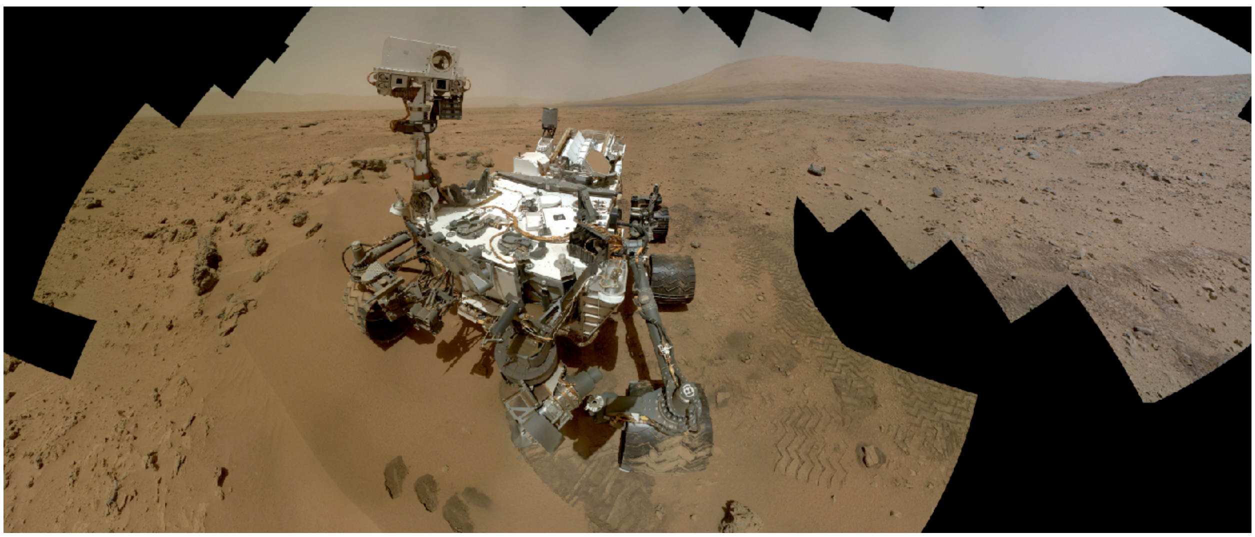


Figure 7.1 This picture was taken by the *Curiosity* Rover on Mars in 2012. The image is reconstructed digitally from 55 different images taken by a camera on the rover's extended mast, so that the many positions of the mast (which acted like a selfie stick) are edited out. (credit: modification of work by NASA/JPL-Caltech/MSSS)

Surrounding the Sun is a complex system of worlds with a wide range of conditions: eight major planets, many dwarf planets, hundreds of moons, and countless smaller objects. Thanks largely to visits by spacecraft, we can now envision the members of the solar system as other worlds like our own, each with its own chemical and geological history, and unique sights that interplanetary tourists may someday visit. Some have called these past few decades the "golden age of planetary exploration," comparable to the golden age of exploration in the fifteenth century, when great sailing ships plied Earth's oceans and humanity became familiar with our own planet's surface.

In this chapter, we discuss our planetary system and introduce the idea of comparative planetology—studying how the planets work by comparing them with one another. We want to get to know the planets not only for what we can learn about them, but also to see what they can tell us about the origin and evolution of the entire solar system. In the upcoming chapters, we describe the better-known members of the solar system and begin to compare them to the thousands of planets that have been discovered recently, orbiting other stars.

7.1 | Overview of Our Planetary System

Learning Objectives

By the end of this section you will be able to:

- Describe how the objects in our solar system are identified, explored, and characterized
- Describe the types of small bodies in our solar system, their locations, and how they formed
- Model the solar system with distances from everyday life to better comprehend distances in space

The solar system[1] consists of the Sun and many smaller objects: the planets, their moons and rings, and such “debris” as asteroids, comets, and dust. Decades of observation and spacecraft exploration have revealed that most of these objects formed together with the Sun about 4.5 billion years ago. They represent clumps of material that condensed from an enormous cloud of gas and dust. The central part of this cloud became the Sun, and a small fraction of the material in the outer parts eventually formed the other objects.

During the past 50 years, we have learned more about the solar system than anyone imagined before the space age. In addition to gathering information with powerful new telescopes, we have sent spacecraft directly to many members of the planetary system. (Planetary astronomy is the only branch of our science in which we can, at least vicariously, travel to the objects we want to study.) With evocative names such as Voyager, Pioneer, *Curiosity*, and Pathfinder, our robot explorers have flown past, orbited, or landed on every planet, returning images and data that have dazzled both astronomers and the public. In the process, we have also investigated two dwarf planets, hundreds of fascinating moons, four ring systems, a dozen asteroids, and several comets (smaller members of our solar system that we will discuss later).

Our probes have penetrated the atmosphere of Jupiter and landed on the surfaces of Venus, Mars, our Moon, Saturn’s moon Titan, the asteroids Eros and Itokawa, and the Comet Churyumov-Gerasimenko (usually referred to as 67P). Humans have set foot on the Moon and returned samples of its surface soil for laboratory analysis (**Figure 7.2**). We have even discovered other places in our solar system that might be able to support some kind of life.

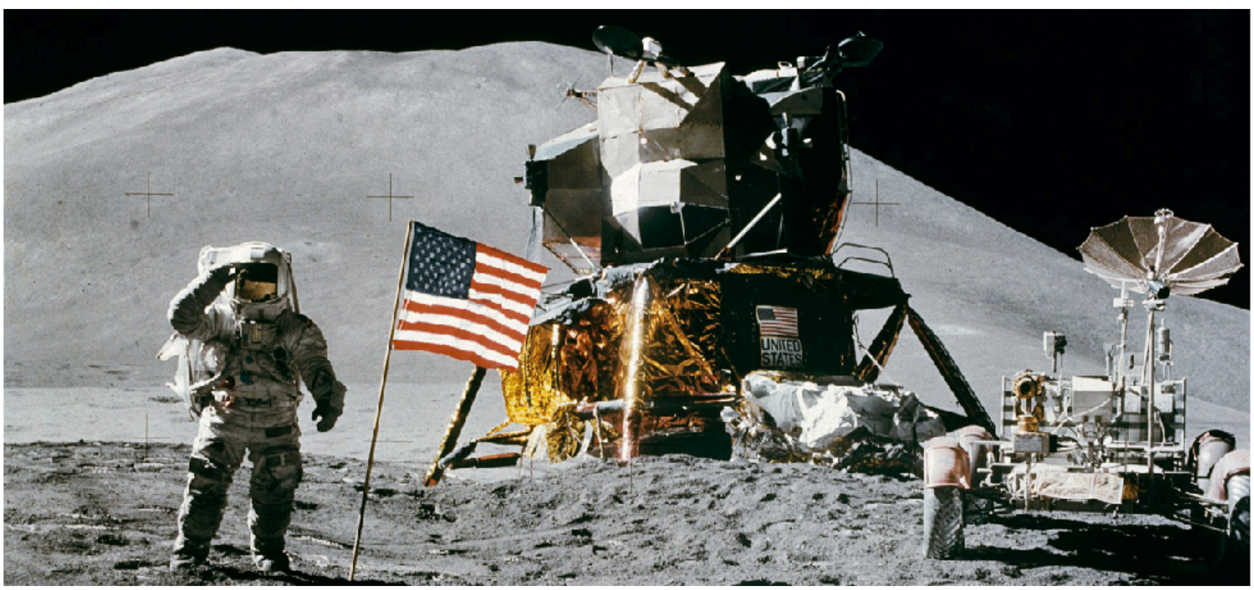


Figure 7.2 The lunar lander and surface rover from the Apollo 15 mission are seen in this view of the one place beyond Earth that has been explored directly by humans. (credit: modification of work by David R. Scott, NASA)

View this gallery of **NASA images (https://openstaxcollege.org/l/30projapolloarc)** that trace the history of the Apollo mission.

1. The generic term for a group of planets and other bodies circling a star is *planetary system*. Ours is called the *solar system* because our Sun is sometimes called *Sol*. Strictly speaking, then, there is only one solar system; planets orbiting other stars are in planetary systems.

An Inventory

The Sun, a star that is brighter than about 80% of the stars in the Galaxy, is by far the most massive member of the solar system, as shown in **Table 7.1**. It is an enormous ball about 1.4 million kilometers in diameter, with surface layers of incandescent gas and an interior temperature of millions of degrees. The Sun will be discussed in later chapters as our first, and best-studied, example of a star.

Mass of Members of the Solar System

Object	Percentage of Total Mass of Solar System
Sun	99.80
Jupiter	0.10
Comets	0.0005–0.03 (estimate)
All other planets and dwarf planets	0.04
Moons and rings	0.00005
Asteroids	0.000002 (estimate)
Cosmic dust	0.0000001 (estimate)

Table 7.1

Table 7.1 also shows that most of the material of the planets is actually concentrated in the largest one, Jupiter, which is more massive than all the rest of the planets combined. Astronomers were able to determine the masses of the planets centuries ago using Kepler's laws of planetary motion and Newton's law of gravity to measure the planets' gravitational effects on one another or on moons that orbit them (see **Newtons' Law of Universal Gravitation**). Today, we make even more precise measurements of their masses by tracking their gravitational effects on the motion of spacecraft that pass near them.

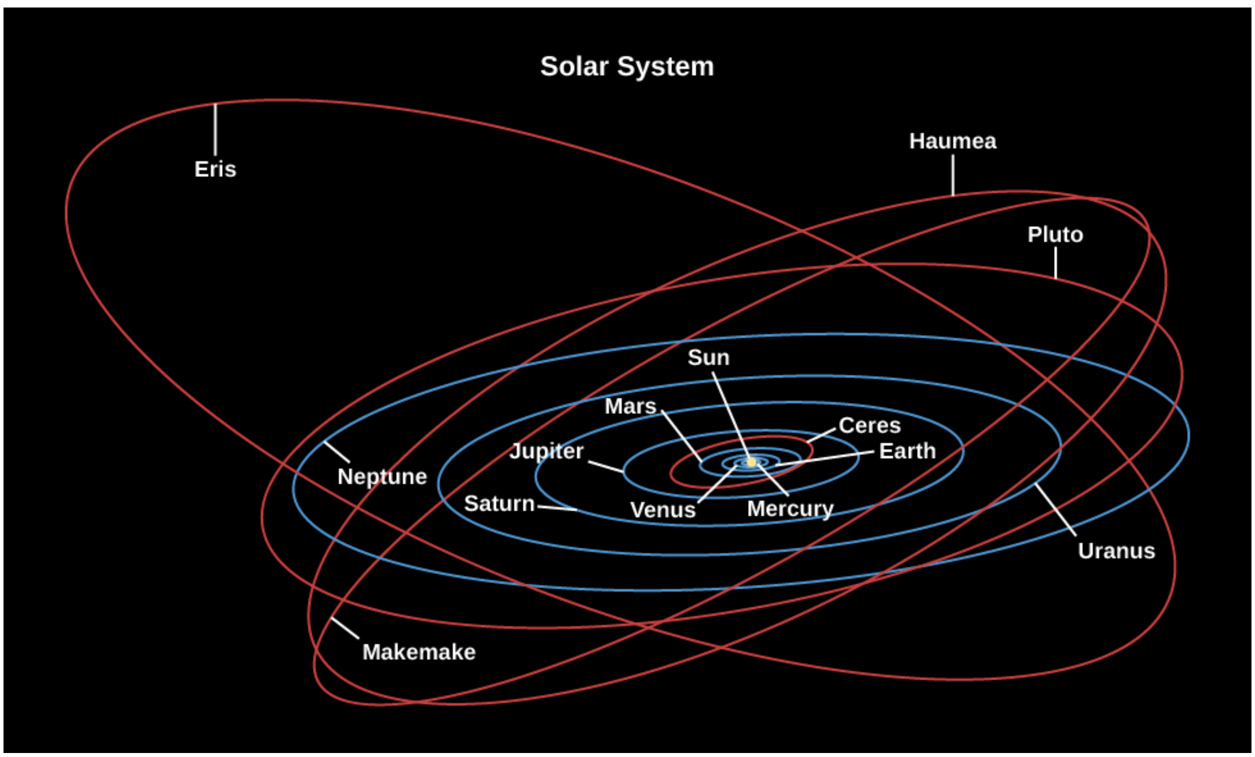


Figure 7.3 All eight major planets orbit the Sun in roughly the same plane. The five currently known dwarf planets are also shown: Eris, Haumea, Pluto, Ceres, and Makemake. Note that Pluto's orbit is not in the plane of the planets.

Beside Earth, five other planets were known to the ancients—Mercury, Venus, Mars, Jupiter, and Saturn—and two were discovered after the invention of the telescope: Uranus and Neptune. The eight planets all revolve in the same direction around the Sun. They orbit in approximately the same plane, like cars traveling on concentric tracks on a giant, flat racecourse. Each planet stays in its own “traffic lane,” following a nearly circular orbit about the Sun and obeying the “traffic” laws discovered by Galileo, Kepler, and Newton. Besides these planets, we have also been discovering smaller worlds beyond Neptune that are called trans-Neptunian objects or TNOs (see **Figure 7.3**). The first to be found, in 1930, was Pluto, but others have been discovered during the twenty-first century. One of them, Eris, is about the same size as Pluto and has at least one moon (Pluto has five known moons.) The largest TNOs are also classed as *dwarf planets*, as is the largest asteroid, Ceres. To date, more than 1750 of these TNOs have been discovered.

Each of the planets and dwarf planets also rotates (spins) about an axis running through it, and in most cases the direction of rotation is the same as the direction of revolution about the Sun. The exceptions are Venus, which rotates backward very slowly (that is, in a retrograde direction), and Uranus and Pluto, which also have strange rotations, each spinning about an axis tipped nearly on its side. We do not yet know the spin orientations of Eris, Haumea, and Makemake.

The four planets closest to the Sun (Mercury through Mars) are called the inner or **terrestrial planets**. Often, the Moon is also discussed as a part of this group, bringing the total of terrestrial objects to five. (We generally call Earth’s satellite “the Moon,” with a capital M, and the other satellites “moons,” with lowercase m’s.) The terrestrial planets are relatively small worlds, composed primarily of rock and metal. All of them have solid surfaces that bear the records of their geological history in the forms of craters, mountains, and volcanoes (**Figure 7.4**).

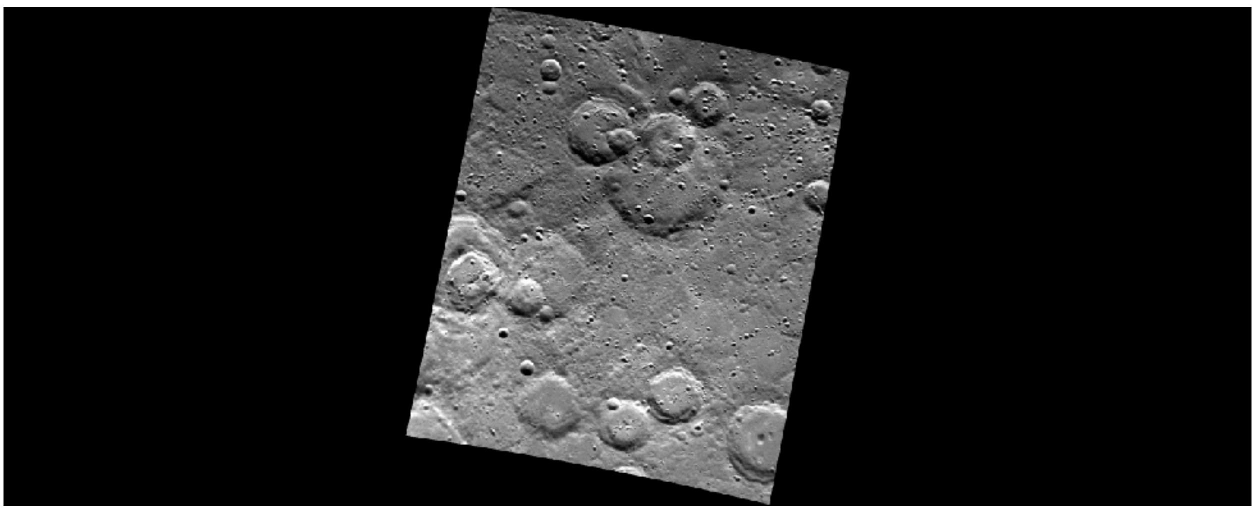


Figure 7.4 The pockmarked face of the terrestrial world of Mercury is more typical of the inner planets than the watery surface of Earth. This black-and-white image, taken with the Mariner 10 spacecraft, shows a region more than 400 kilometers wide. (credit: modification of work by NASA/John Hopkins University Applied Physics Laboratory/Carnegie Institution of Washington)

The next four planets (Jupiter through Neptune) are much larger and are composed primarily of lighter ices, liquids, and gases. We call these four the jovian planets (after “Jove,” another name for Jupiter in mythology) or **giant planets**—a name they richly deserve (**Figure 7.5**). More than 1400 Earths could fit inside Jupiter, for example. These planets do not have solid surfaces on which future explorers might land. They are more like vast, spherical oceans with much smaller, dense cores.

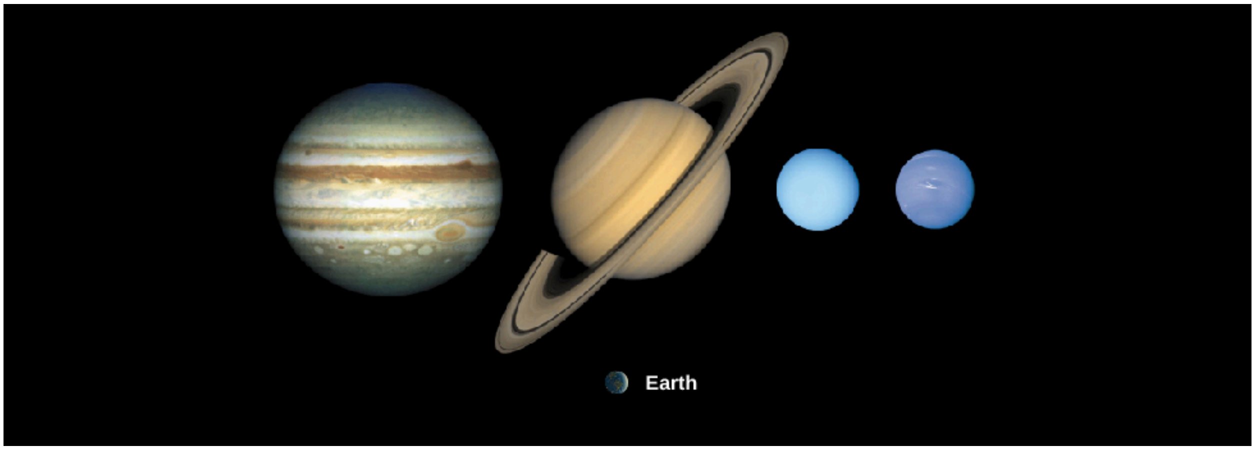


Figure 7.5 This montage shows the four giant planets: Jupiter, Saturn, Uranus, and Neptune. Below them, Earth is shown to scale. (credit: modification of work by NASA, Solar System Exploration)

Near the outer edge of the system lies Pluto, which was the first of the distant icy worlds to be discovered beyond Neptune (Pluto was visited by a spacecraft, the NASA New Horizons mission, in 2015 [see **Figure 7.6**]). **Table 7.2** summarizes some of the main facts about the planets.

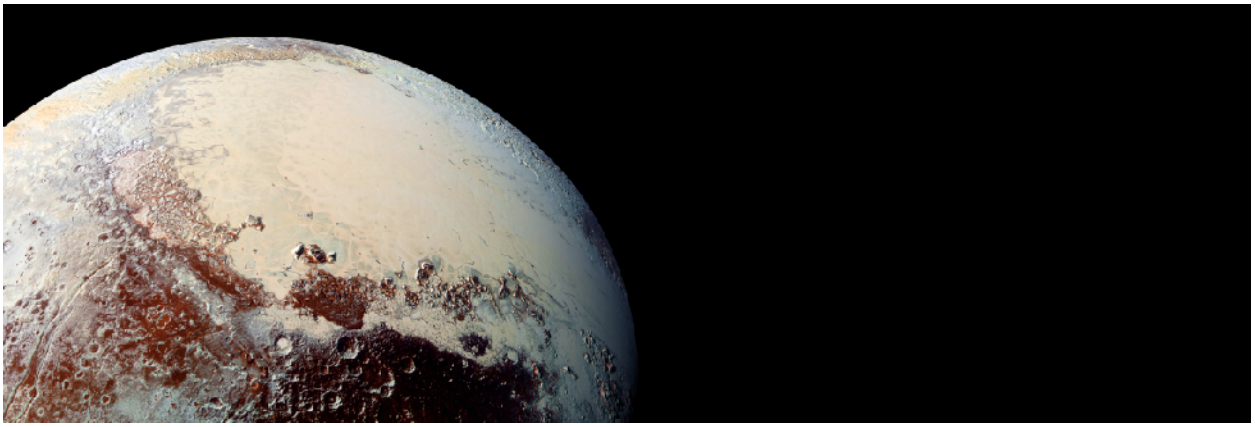


Figure 7.6 This intriguing image from the New Horizons spacecraft, taken when it flew by the dwarf planet in July 2015, shows some of its complex surface features. The rounded white area is temporarily being called the Sputnik Plain, after humanity's first spacecraft. (credit: modification of work by NASA/Johns Hopkins University Applied Physics Laboratory/ Southwest Research Institute)

The Planets

Name	Distance from Sun (AU)[2]	Revolution Period (y)	Diameter (km)	Mass (10^{23} kg)	Density (g/cm^3)[3]
Mercury	0.39	0.24	4,878	3.3	5.4
Venus	0.72	0.62	12,120	48.7	5.2
Earth	1.00	1.00	12,756	59.8	5.5
Mars	1.52	1.88	6,787	6.4	3.9
Jupiter	5.20	11.86	142,984	18,991	1.3

Table 7.2

2. An AU (or astronomical unit) is the distance from Earth to the Sun.
3. We give densities in units where the density of water is 1 g/cm^3. To get densities in units of kg/m^3, multiply the given value by 1000.

The Planets

Name	Distance from Sun (AU)	Revolution Period (y)	Diameter (km)	Mass (10^{23} kg)	Density (g/cm^3)
Saturn	9.54	29.46	120,536	5686	0.7
Uranus	19.18	84.07	51,118	866	1.3
Neptune	30.06	164.82	49,660	1030	1.6

Table 7.2

Example 7.1

Comparing Densities

Let's compare the densities of several members of the solar system. The density of an object equals its mass divided by its volume. The volume (*V*) of a sphere (like a planet) is calculated using the equation

$$V = \frac{4}{3}\pi R^3$$

where π (the Greek letter pi) has a value of approximately 3.14. Although planets are not perfect spheres, this equation works well enough. The masses and diameters of the planets are given in **Table 7.2**. For data on selected moons, see **Section D.2**. Let's use Saturn's moon Mimas as our example, with a mass of 4×10^{19} kg and a diameter of approximately 400 km (radius, 200 km = 2×10^5 m).

Solution

The volume of Mimas is

$$\frac{4}{3} \times 3.14 \times \left(2 \times 10^5 \text{ m}\right)^3 = 3.3 \times 10^{16} \text{ m}^3.$$

Density is mass divided by volume:

$$\frac{4 \times 10^{19} \text{ kg}}{3.3 \times 10^{16} \text{ m}^3} = 1.2 \times 10^3 \text{ kg/m}^3.$$

Note that the density of water in these units is 1000 kg/m^3, so Mimas must be made mainly of ice, not rock. (Note that the density of Mimas given in **Section D.2** is 1.2, but the units used there are different. In that table, we give density in units of g/cm^3, for which the density of water equals 1. Can you show, by converting units, that 1 g/cm^3 is the same as 1000 kg/m^3?)

7.1 Calculate the average density of our own planet, Earth. Show your work. How does it compare to the density of an ice moon like Mimas? See **Table 7.2** for data.

Learn more about NASA's **mission to Pluto (https://openstaxcollege.org/l/30NASAmisspluto)** and see high-resolution images of Pluto's moon Charon.

Smaller Members of the Solar System

Most of the planets are accompanied by one or more moons; only Mercury and Venus move through space alone. There are more than 180 known moons orbiting planets and dwarf planets (see **Section D.2** for a listing of the larger ones), and undoubtedly many other small ones remain undiscovered. The largest of the moons are as big as small planets and just as interesting. In addition to our Moon, they include the four largest moons of Jupiter (called the Galilean moons, after their discoverer) and the largest moons of Saturn and Neptune (confusingly named Titan and Triton).

Each of the giant planets also has rings made up of countless small bodies ranging in size from mountains to mere grains of dust, all in orbit about the equator of the planet. The bright rings of Saturn are, by far, the easiest to see. They are among the most beautiful sights in the solar system (**Figure 7.7**). But, all four ring systems are interesting to scientists because of their complicated forms, influenced by the pull of the moons that also orbit these giant planets.

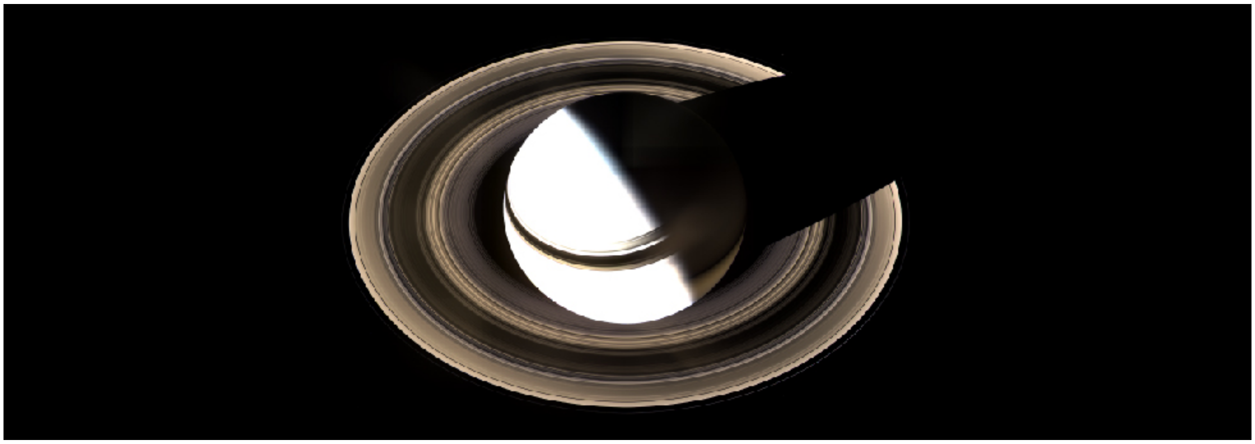


Figure 7.7 This 2007 Cassini image shows Saturn and its complex system of rings, taken from a distance of about 1.2 million kilometers. This natural-color image is a composite of 36 images taken over the course of 2.5 hours. (credit: modification of work by NASA/JPL/Space Science Institute)

The solar system has many other less-conspicuous members. Another group is the **asteroids**, rocky bodies that orbit the Sun like miniature planets, mostly in the space between Mars and Jupiter (although some do cross the orbits of planets like Earth—see **Figure 7.8**). Most asteroids are remnants of the initial population of the solar system that existed before the planets themselves formed. Some of the smallest moons of the planets, such as the moons of Mars, are very likely captured asteroids.

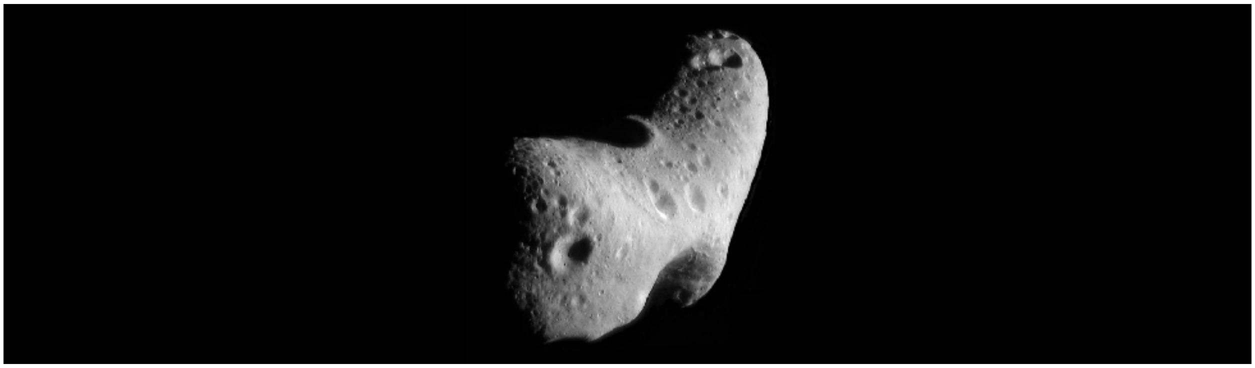


Figure 7.8 This small Earth-crossing asteroid image was taken by the NEAR-Shoemaker spacecraft from an altitude of about 100 kilometers. This view of the heavily cratered surface is about 10 kilometers wide. The spacecraft orbited Eros for a year before landing gently on its surface. (credit: modification of work by NASA/JHUAPL)

Another class of small bodies is composed mostly of ice, made of frozen gases such as water, carbon dioxide, and carbon monoxide; these objects are called **comets** (see **Figure 7.9**). Comets also are remnants from the formation of the solar system, but they were formed and continue (with rare exceptions) to orbit the Sun in distant, cooler regions—stored in a sort of cosmic deep freeze. This is also the realm of the larger icy worlds, called dwarf planets.

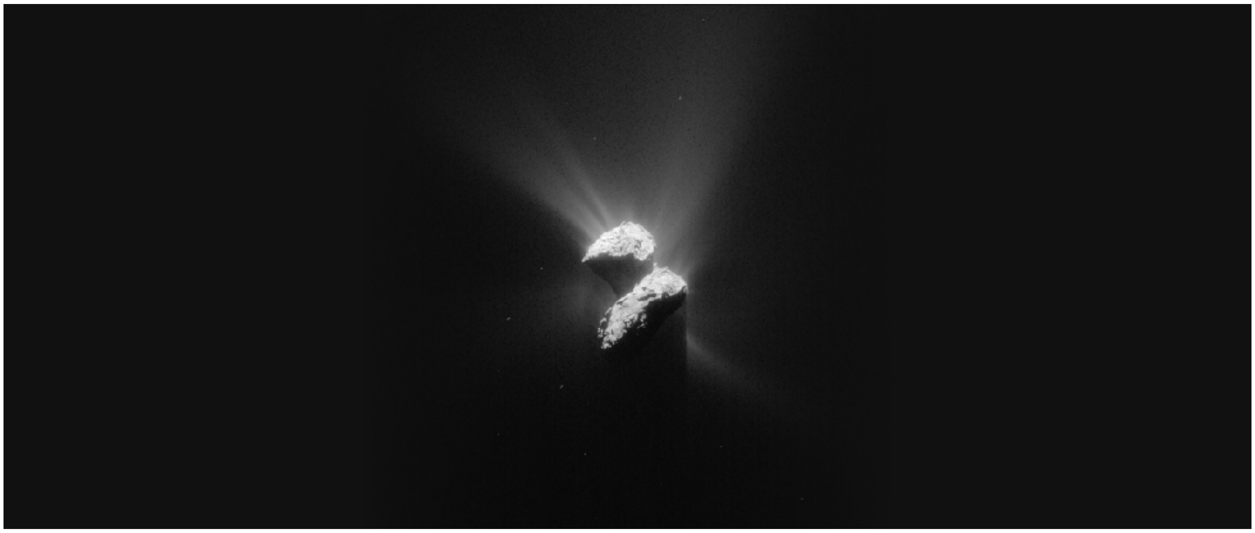


Figure 7.9 This image shows Comet Churyumov-Gerasimenko, also known as 67P, near its closest approach to the Sun in 2015, as seen from the *Rosetta* spacecraft. Note the jets of gas escaping from the solid surface. (credit: modification of work by ESA/Rosetta/NAVACAM, CC BY-SA IGO 3.0 (http://creativecommons.org/licenses/by-sa/3.0/igo/))

Finally, there are countless grains of broken rock, which we call cosmic dust, scattered throughout the solar system. When these particles enter Earth's atmosphere (as millions do each day) they burn up, producing a brief flash of light in the night sky known as a **meteor** (meteors are often referred to as shooting stars). Occasionally, some larger chunk of rocky or metallic material survives its passage through the atmosphere and lands on Earth. Any piece that strikes the ground is known as a **meteorite**. (You can see meteorites on display in many natural history museums and can sometimes even purchase pieces of them from gem and mineral dealers.)

Carl Sagan: Solar System Advocate

The best-known astronomer in the world during the 1970s and 1980s, Carl Sagan devoted most of his professional career to studying the planets and considerable energy to raising public awareness of what we can learn from exploring the solar system (see Figure 7.10). Born in Brooklyn, New York, in 1934, Sagan became interested in astronomy as a youngster; he also credits science fiction stories for sustaining his fascination with what's out in the universe.

Figure 7.10 Sagan was Tyson's inspiration to become a scientist. (credit "Sagan": modification of work by NASA, JPL; credit "Tyson": modification of work by Bruce F. Press)

In the early 1960s, when many scientists still thought Venus might turn out to be a hospitable place, Sagan calculated that the thick atmosphere of Venus could act like a giant greenhouse, keeping the heat in and raising the temperature enormously. He showed that the seasonal changes astronomers had seen on Mars were caused, not by vegetation, but by wind-blown dust. He was a member of the scientific teams for many of the robotic missions that explored the solar system and was instrumental in getting NASA to put a message-bearing plaque aboard the Pioneer spacecraft, as well as audio-video records on the Voyager spacecraft—all of them destined to leave our solar system entirely and send these little bits of Earth technology out among the stars.

To encourage public interest and public support of planetary exploration, Sagan helped found The Planetary Society, now the largest space-interest organization in the world. He was a tireless and eloquent advocate of the need to study the solar system close-up and the value of learning about other worlds in order to take better care of our own.

Sagan simulated conditions on early Earth to demonstrate how some of life's fundamental building blocks might have formed from the "primordial soup" of natural compounds on our planet. In addition, he and his colleagues developed computer models showing the consequences of nuclear war for Earth would be even more devastating than anyone had thought (this is now called the nuclear winter hypothesis) and demonstrating some of the serious consequences of continued pollution of our atmosphere.

Sagan was perhaps best known, however, as a brilliant popularizer of astronomy and the author of many books on science, including the best-selling *Cosmos*, and several evocative tributes to solar system exploration such as *The Cosmic Connection* and *Pale Blue Dot*. His book *The Demon Haunted World*, completed just before his death in 1996, is perhaps the best antidote to fuzzy thinking about pseudo-science and irrationality in print today. An intriguing science fiction novel he wrote, titled *Contact*, which became a successful film as well, is still recommended by many science instructors as a scenario for making contact with life elsewhere that is much more reasonable than most science fiction.

Sagan was a master, too, of the television medium. His 13-part public television series, *Cosmos*, was seen by an estimated 500 million people in 60 countries and has become one of the most-watched series in the history of public broadcasting. A few astronomers scoffed at a scientist who spent so much time in the public eye, but it is probably fair to say that Sagan's enthusiasm and skill as an explainer won more friends for the science of astronomy than anyone or anything else in the second half of the twentieth century.

In the two decades since Sagan's death, no other scientist has achieved the same level of public recognition. Perhaps closest is the director of the Hayden Planetarium, Neil deGrasse Tyson, who followed in Sagan's footsteps by making an updated version of the *Cosmos* program in 2014. Tyson is quick to point out that Sagan was his inspiration to become a scientist, telling how Sagan invited him to visit for a day at Cornell when he was a high school student looking for a career. However, the media environment has fragmented a great deal since Sagan's time. It is interesting to speculate whether Sagan could have adapted his communication style to the world of cable television, Twitter, Facebook, and podcasts.

Two imaginative videos provide a tour of the solar system objects we have been discussing. Shane Gellert's **I Need Some Space (https://openstaxcollege.org/l/30needsomespace)** uses NASA photography and models to show the various worlds with which we share our system. In the more science fiction-oriented **Wanderers (https://openstaxcollege.org/l/30wanderers)** video, we see some of the planets and moons as tourist destinations for future explorers, with commentary taken from recordings by Carl Sagan.

A Scale Model of the Solar System

Astronomy often deals with dimensions and distances that far exceed our ordinary experience. What does 1.4 billion kilometers—the distance from the Sun to Saturn—really mean to anyone? It can be helpful to visualize such large systems in terms of a scale model.

In our imaginations, let us build a scale model of the solar system, adopting a scale factor of 1 billion (10^9)—that is, reducing the actual solar system by dividing every dimension by a factor of 10^9. Earth, then, has a diameter of 1.3 centimeters, about the size of a grape. The Moon is a pea orbiting this at a distance of 40 centimeters, or a little more than a foot away. The Earth-Moon system fits into a standard backpack.

In this model, the Sun is nearly 1.5 meters in diameter, about the average height of an adult, and our Earth is at a distance of 150 meters—about one city block—from the Sun. Jupiter is five blocks away from the Sun, and its diameter is 15 centimeters, about the size of a very large grapefruit. Saturn is 10 blocks from the Sun; Uranus, 20 blocks; and Neptune, 30 blocks. Pluto, with a distance that varies quite a bit during its 249-year orbit, is currently just beyond 30 blocks and getting farther with time. Most of the moons of the outer solar system are the sizes of various kinds of seeds orbiting the grapefruit, oranges, and lemons that represent the outer planets.

In our scale model, a human is reduced to the dimensions of a single atom, and cars and spacecraft to the size of molecules. Sending the Voyager spacecraft to Neptune involves navigating a single molecule from the Earth–grape toward a lemon 5 kilometers away with an accuracy equivalent to the width of a thread in a spider's web.

If that model represents the solar system, where would the nearest stars be? If we keep the same scale, the closest stars would be tens of thousands of kilometers away. If you built this scale model in the city where you live, you would have to place the representations of these stars on the other side of Earth or beyond.

By the way, model solar systems like the one we just presented have been built in cities throughout the world. In Sweden, for example, Stockholm's huge Globe Arena has become a model for the Sun, and Pluto is represented by a 12-centimeter sculpture in the small town of Delsbo, 300 kilometers away. Another model solar system is in Washington on the Mall between the White House and Congress (perhaps proving they are worlds apart?).

Names in the Solar System

We humans just don't feel comfortable until something has a name. Types of butterflies, new elements, and the mountains of Venus all need names for us to feel we are acquainted with them. How do we give names to objects and features in the solar system?

Planets and moons are named after gods and heroes in Greek and Roman mythology (with a few exceptions among the moons of Uranus, which have names drawn from English literature). When William Herschel, a German immigrant to England, first discovered the planet we now call Uranus, he wanted to name it Georgium Sidus (George's star) after King George III of his adopted country. This caused such an outcry among astronomers in other nations, however, that the classic tradition was upheld—and has been maintained ever since. Luckily, there were a lot of minor gods in the ancient pantheon, so plenty of names are left for the many small moons we are discovering around the giant planets. (**Section D.2** lists the larger moons).

Comets are often named after their discoverers (offering an extra incentive to comet hunters). Asteroids are named by their discoverers after just about anyone or anything they want. Recently, asteroid names have been used to recognize people who have made significant contributions to astronomy, including the three original authors of this book.

That was pretty much all the naming that was needed while our study of the solar system was confined to Earth. But now, our spacecraft have surveyed and photographed many worlds in great detail, and each world has a host of features that also need names. To make sure that naming things in space remains multinational, rational, and somewhat dignified, astronomers have given the responsibility of approving names to a special committee of the International Astronomical Union (IAU), the body that includes scientists from every country that does astronomy.

This IAU committee has developed a set of rules for naming features on other worlds. For example, craters on Venus are named for women who have made significant contributions to human knowledge and welfare. Volcanic features on Jupiter's moon Io, which is in a constant state of volcanic activity, are named after gods of fire and thunder from the mythologies of many cultures. Craters on Mercury commemorate famous novelists, playwrights, artists, and composers. On Saturn's moon Tethys, all the features are named after characters and places in Homer's great epic poem, *The Odyssey*. As we explore further, it may well turn out that more places in the solar system need names than Earth history can provide. Perhaps by then, explorers and settlers on these worlds will be ready to develop their own names for the places they may (if but for a while) call home.

You may be surprised to know that the meaning of the word *planet* has recently become controversial because we have discovered many other planetary systems that don't look very much like our own. Even within our solar system, the planets differ greatly in size and chemical properties. The biggest dispute concerns Pluto, which is much smaller than the other eight major planets. The category of dwarf planet was invented to include Pluto and similar icy objects beyond Neptune. But is a dwarf planet also a planet? Logically, it should be, but even this simple issue of grammar has been the subject of heated debate among both astronomers and the general public.

7.2 | Composition and Structure of Planets

Learning Objectives

By the end of this section you will be able to:

- Describe the characteristics of the giant planets, terrestrial planets, and small bodies in the solar system
- Explain what influences the temperature of a planet's surface
- Explain why there is geological activity on some planets and not on others

The fact that there are two distinct kinds of planets—the rocky terrestrial planets and the gas-rich jovian planets—leads us to believe that they formed under different conditions. Certainly their compositions are dominated by different elements. Let us look at each type in more detail.

The Giant Planets

The two largest planets, Jupiter and Saturn, have nearly the same chemical makeup as the Sun; they are composed primarily of the two elements hydrogen and helium, with 75% of their mass being hydrogen and 25% helium. On Earth, both hydrogen and helium are gases, so Jupiter and Saturn are sometimes called gas planets. But, this name is misleading. Jupiter and Saturn are so large that the gas is compressed in their interior until the hydrogen becomes a liquid. Because the bulk of both planets consists of compressed, liquefied hydrogen, we should really call them liquid planets.

Under the force of gravity, the heavier elements sink toward the inner parts of a liquid or gaseous planet. Both Jupiter and Saturn, therefore, have cores composed of heavier rock, metal, and ice, but we cannot see these regions directly. In fact, when we look down from above, all we see is the atmosphere with its swirling clouds (**Figure 7.11**). We must infer the existence of the denser core inside these planets from studies of each planet's gravity.

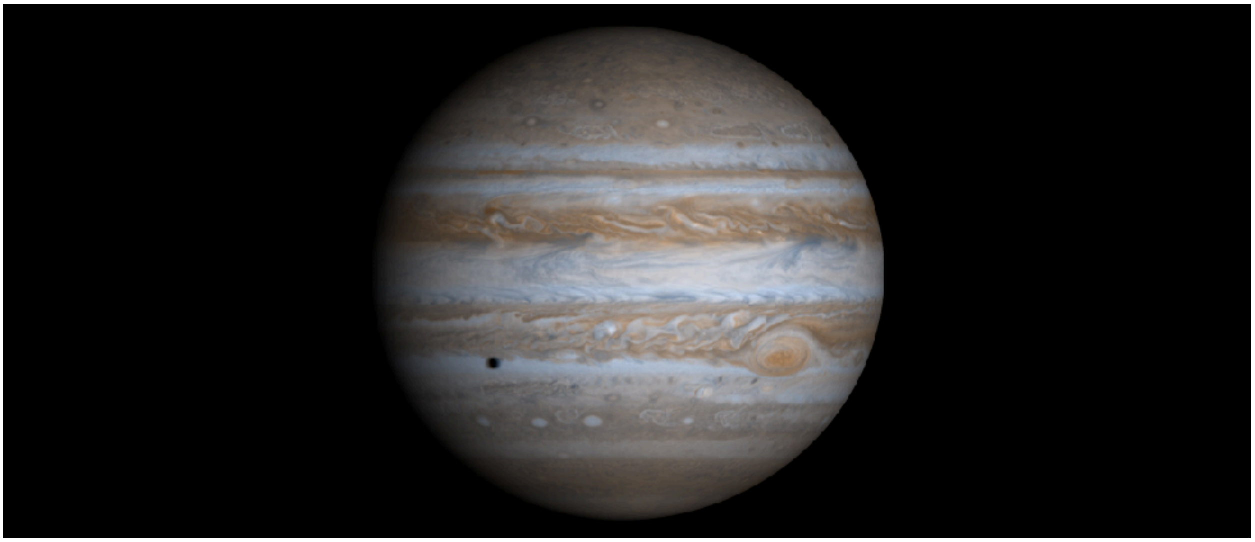


Figure 7.11 This true-color image of Jupiter was taken from the Cassini spacecraft in 2000. (credit: modification of work by NASA/JPL/University of Arizona)

Uranus and Neptune are much smaller than Jupiter and Saturn, but each also has a core of rock, metal, and ice. Uranus and Neptune were less efficient at attracting hydrogen and helium gas, so they have much smaller atmospheres in proportion to their cores.

Chemically, each giant planet is dominated by hydrogen and its many compounds. Nearly all the oxygen present is combined chemically with hydrogen to form water (H_2O). Chemists call such a hydrogen-dominated composition *reduced*. Throughout the outer solar system, we find abundant water (mostly in the form of ice) and reducing chemistry.

The Terrestrial Planets

The terrestrial planets are quite different from the giants. In addition to being much smaller, they are composed primarily of rocks and metals. These, in turn, are made of elements that are less common in the universe as a whole. The most abundant rocks, called silicates, are made of silicon and oxygen, and the most common metal is iron. We can tell from their densities (see **Table 34.2**) that Mercury has the greatest proportion of metals (which are denser) and the Moon has the lowest. Earth, Venus, and Mars all have roughly similar bulk compositions: about one third of their mass consists of iron-nickel or iron-sulfur combinations; two thirds is made of silicates. Because these planets are largely composed of oxygen compounds (such as the silicate minerals of their crusts), their chemistry is said to be *oxidized*.

When we look at the internal structure of each of the terrestrial planets, we find that the densest metals are in a central core, with the lighter silicates near the surface. If these planets were liquid, like the giant planets, we could understand this effect as the result the sinking of heavier elements due to the pull of gravity. This leads us to conclude that, although the terrestrial planets are solid today, at one time they must have been hot enough to melt.

Differentiation is the process by which gravity helps separate a planet's interior into layers of different compositions and densities. The heavier metals sink to form a core, while the lightest minerals float to the surface to form a crust. Later, when the planet cools, this layered structure is preserved. In order for a rocky planet to differentiate, it must be heated to the melting point of rocks, which is typically more than 1300 K.

Moons, Asteroids, and Comets

Chemically and structurally, Earth's Moon is like the terrestrial planets, but most moons are in the outer solar system, and they have compositions similar to the cores of the giant planets around which they orbit. The three largest moons—Ganymede and Callisto in the jovian system, and Titan in the saturnian system—are composed half of frozen water, and half of rocks and metals. Most of these moons differentiated during formation, and today they have cores of rock and metal, with upper layers and crusts of very cold and—thus very hard—ice (**Figure 7.12**).

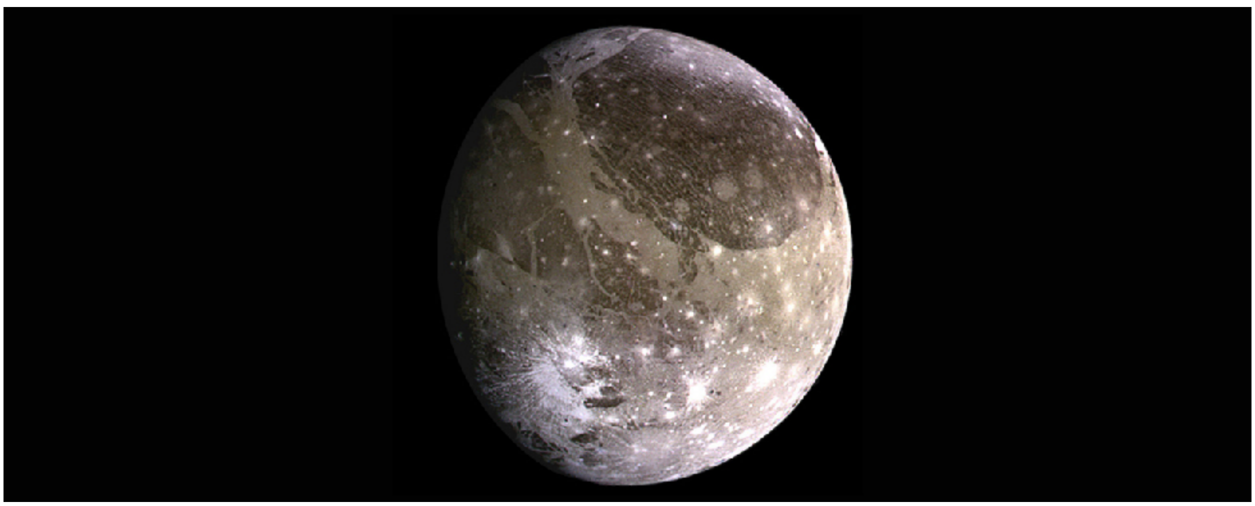


Figure 7.12 This view of Jupiter's moon Ganymede was taken in June 1996 by the Galileo spacecraft. The brownish gray color of the surface indicates a dusty mixture of rocky material and ice. The bright spots are places where recent impacts have uncovered fresh ice from underneath. (credit: modification of work by NASA/JPL)

Most of the asteroids and comets, as well as the smallest moons, were probably never heated to the melting point. However, some of the largest asteroids, such as Vesta, appear to be differentiated; others are fragments from differentiated bodies. Because most asteroids and comets retain their original composition, they represent relatively unmodified material dating back to the time of the formation of the solar system. In a sense, they act as chemical fossils, helping us to learn about a time long ago whose traces have been erased on larger worlds.

Temperatures: Going to Extremes

Generally speaking, the farther a planet or moon is from the Sun, the cooler its surface. The planets are heated by the radiant energy of the Sun, which gets weaker with the square of the distance. You know how rapidly the heating effect of a fireplace or an outdoor radiant heater diminishes as you walk away from it; the same effect applies to the Sun. Mercury, the closest planet to the Sun, has a blistering surface temperature that ranges from 280–430 °C on its sunlit side, whereas the surface temperature on Pluto is only about –220 °C, colder than liquid air.

Mathematically, the temperatures decrease approximately in proportion to the square root of the distance from the Sun. Pluto is about 30 AU at its closest to the Sun (or 100 times the distance of Mercury) and about 49 AU at its farthest from the Sun. Thus, Pluto's temperature is less than that of Mercury by the square root of 100, or a factor of 10: from 500 K to 50 K.

In addition to its distance from the Sun, the surface temperature of a planet can be influenced strongly by its atmosphere. Without our atmospheric insulation (the greenhouse effect, which keeps the heat in), the oceans of Earth would be permanently frozen. Conversely, if Mars once had a larger atmosphere in the past, it could have supported a more temperate climate than it has today. Venus is an even more extreme example, where its thick atmosphere of carbon dioxide acts as insulation, reducing the escape of heat built up at the surface, resulting in temperatures greater than those on Mercury. Today, Earth is the only planet where surface temperatures generally lie between the freezing and boiling points of water. As far as we know, Earth is the only planet to support life.

There's No Place Like Home

In the classic film *The Wizard of Oz*, Dorothy, the heroine, concludes after her many adventures in "alien" environments that "there's no place like home." The same can be said of the other worlds in our solar system. There are many fascinating places, large and small, that we might like to visit, but humans could not survive on any without a great deal of artificial assistance.

A thick carbon dioxide atmosphere keeps the surface temperature on our neighbor Venus at a sizzling 700 K (near 900 °F). Mars, on the other hand, has temperatures generally below freezing, with air (also mostly carbon dioxide) so thin that it resembles that found at an altitude of 30 kilometers (100,000 feet) in Earth's atmosphere. And the red planet is so dry that it has not had any rain for billions of years.

The outer layers of the jovian planets are neither warm enough nor solid enough for human habitation. Any bases we build in the systems of the giant planets may well have to be in space or one of their moons—none of which is particularly hospitable to a luxury hotel with a swimming pool and palm trees. Perhaps we will find warmer havens deep inside the clouds of Jupiter or in the ocean under the frozen ice of its moon Europa.

All of this suggests that we had better take good care of Earth because it is the only site where life as we know it could survive. Recent human activity may be reducing the habitability of our planet by adding pollutants to the atmosphere, especially the potent greenhouse gas carbon dioxide. Human civilization is changing our planet dramatically, and these changes are not necessarily for the better. In a solar system that seems unready to receive us, making Earth less hospitable to life may be a grave mistake.

Geological Activity

The crusts of all of the terrestrial planets, as well as of the larger moons, have been modified over their histories by both internal and external forces. Externally, each has been battered by a slow rain of projectiles from space, leaving their surfaces pockmarked by impact craters of all sizes (see **Figure 7.4**). We have good evidence that this bombardment was far greater in the early history of the solar system, but it certainly continues to this day, even if at a lower rate. The collision of more than 20 large pieces of Comet Shoemaker–Levy 9 with Jupiter in the summer of 1994 (see **Figure 7.13**) is one dramatic example of this process.

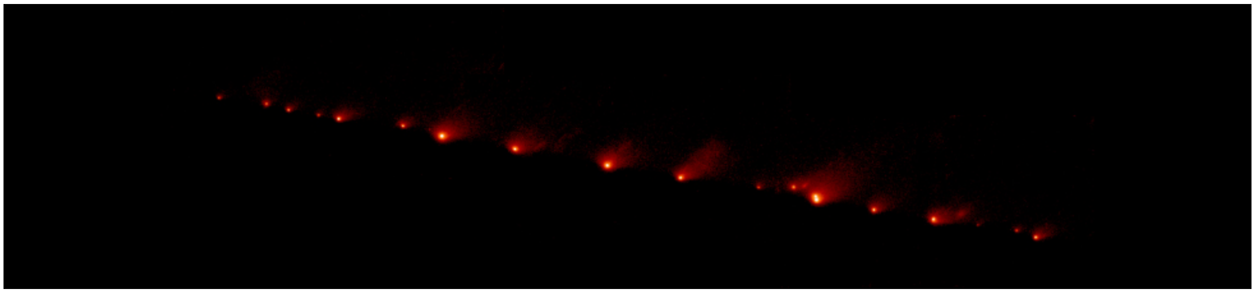


Figure 7.13 In this image of Comet Shoemaker–Levy 9 taken on May 17, 1994, by NASA's Hubble Space Telescope, you can see about 20 icy fragments into which the comet broke. The comet was approximately 660 million kilometers from Earth, heading on a collision course with Jupiter. (credit: modification of work by NASA, ESA, H. Weaver (STScl), E. Smith (STScl))

Figure 7.14 shows the aftermath of these collisions, when debris clouds larger than Earth could be seen in Jupiter's atmosphere.

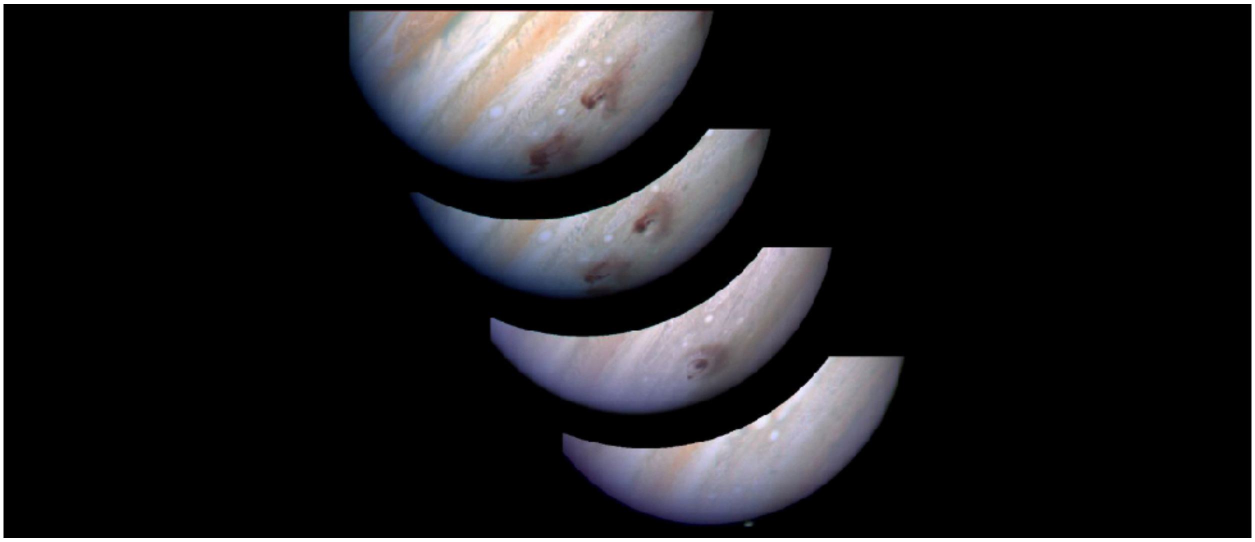


Figure 7.14 The Hubble Space Telescope took this sequence of images of Jupiter in summer 1994, when fragments of Comet Shoemaker–Levy 9 collided with the giant planet. Here we see the site hit by fragment G, from five minutes to five days after impact. Several of the dust clouds generated by the collisions became larger than Earth. (credit: modification of work by H. Hammel, NASA)

During the time all the planets have been subject to such impacts, internal forces on the terrestrial planets have buckled and twisted their crusts, built up mountain ranges, erupted as volcanoes, and generally reshaped the surfaces in what we call geological activity. (The prefix *geo* means "Earth," so this is a bit of an "Earth-chauvinist" term, but it is so widely used that we bow to tradition.) Among the terrestrial planets, Earth and Venus have experienced the most geological activity over their histories, although some of the moons in the outer solar system are also surprisingly active. In contrast, our own Moon is a dead world where geological activity ceased billions of years ago.

Geological activity on a planet is the result of a hot interior. The forces of volcanism and mountain building are driven by heat escaping from the interiors of planets. As we will see, each of the planets was heated at the time of its birth, and this primordial heat initially powered extensive volcanic activity, even on our Moon. But, small objects such as the Moon soon cooled off. The larger the planet or moon, the longer it retains its internal heat, and therefore the more we expect to see surface evidence of continuing geological activity. The effect is similar to our own experience with a hot baked potato: the larger the potato, the more slowly it cools. If we want a potato to cool quickly, we cut it into small pieces.

For the most part, the history of volcanic activity on the terrestrial planets conforms to the predictions of this simple theory. The Moon, the smallest of these objects, is a geologically dead world. Although we know less about Mercury, it seems likely that this planet, too, ceased most volcanic activity about the same time the Moon did. Mars represents an intermediate case. It has been much more active than the Moon, but less so than Earth. Earth and Venus, the largest terrestrial planets, still have molten interiors even today, some 4.5 billion years after their birth.

7.3 | Origin of the Solar System

Learning Objectives

By the end of this section you will be able to:

- Describe the characteristics of planets that are used to create formation models of the solar system.
- Describe how the characteristics of extrasolar systems help us to model our own solar system.
- Explain the importance of collisions in the formation of the solar system.

Much of astronomy is motivated by a desire to understand the origin of things: to find at least partial answers to age-old questions of where the universe, the Sun, Earth, and we ourselves came from. Each planet and moon is a fascinating place that may stimulate our imagination as we try to picture what it would be like to visit. Taken together, the members of the solar system preserve patterns that can tell us about the formation of the entire system. As we begin our exploration of the planets, we want to introduce our modern picture of how the solar system formed.

The recent discovery of hundreds of planets in orbit around other stars has shown astronomers that many exoplanetary systems can be quite different from our own solar system. For example, it is common for these systems to include planets intermediate in size between our terrestrial and giant planets. These are often called *superearths*. Some exoplanet systems even have giant planets close to the star, reversing the order we see in our system. In **Exoplanets**, we will look at these systems. But for now, let us focus on theories of how our own particular system has formed and evolved.

Looking for Patterns

One way to approach our question of origin is to look for regularities among the planets. We found, for example, that all the planets lie in nearly the same plane and revolve in the same direction around the Sun. The Sun also spins in the same direction about its own axis. Astronomers interpret this pattern as evidence that the Sun and planets formed together from a spinning cloud of gas and dust that we call the **solar nebula** (**Figure 7.15**).

Figure 7.15 This artist's conception of the solar nebula shows the flattened cloud of gas and dust from which our planetary system formed. Icy and rocky planetesimals (precursors of the planets) can be seen in the foreground. The bright center is where the Sun is forming. (credit: William K. Hartmann, Planetary Science Institute)

The composition of the planets gives another clue about origins. Spectroscopic analysis allows us to determine which elements are present in the Sun and the planets. The Sun has the same hydrogen-dominated composition as Jupiter and Saturn, and therefore appears to have been formed from the same reservoir of material. In comparison, the terrestrial planets and our Moon are relatively deficient in the light gases and the various ices that form from the common elements oxygen, carbon, and nitrogen. Instead, on Earth and its neighbors, we see mostly the rarer heavy elements such as iron and silicon. This pattern suggests that the processes that led to planet formation in the inner solar system must somehow have excluded much of the lighter materials that are common elsewhere. These lighter materials must have escaped, leaving a residue of heavy stuff.

The reason for this is not hard to guess, bearing in mind the heat of the Sun. The inner planets and most of the asteroids are made of rock and metal, which can survive heat, but they contain very little ice or gas, which evaporate when temperatures are high. (To see what we mean, just compare how long a rock and an ice cube survive when they are placed in the sunlight.) In the outer solar system, where it has always been cooler, the planets and their moons, as well as icy dwarf planets and comets, are composed mostly of ice and gas.

The Evidence from Far Away

A second approach to understanding the origins of the solar system is to look outward for evidence that other systems of planets are forming elsewhere. We cannot look back in time to the formation of our own system, but many stars in space are much younger than the Sun. In these systems, the processes of planet formation might still be accessible to direct observation. We observe that there are many other "solar nebulas" or *circumstellar disks*—flattened, spinning clouds of gas and dust surrounding young stars. These disks resemble our own solar system's initial stages of formation billions of years ago (**Figure 7.16**).

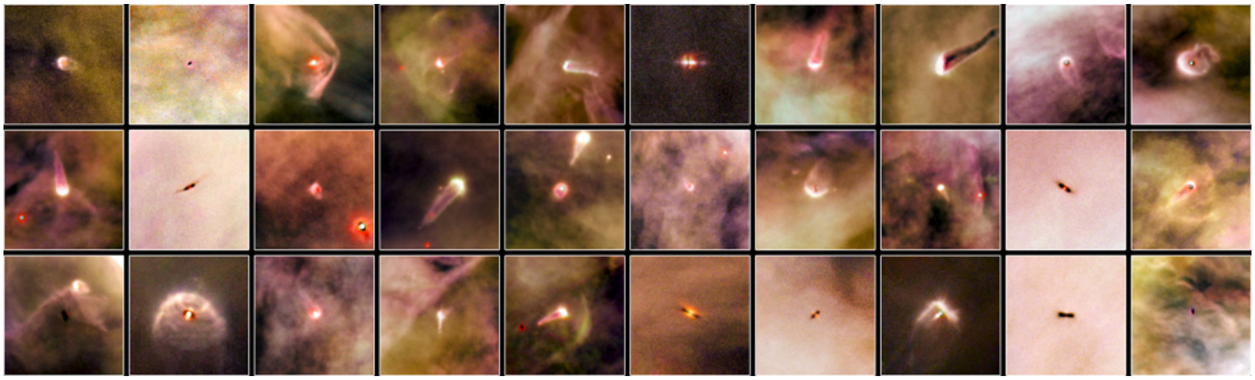


Figure 7.16 These Hubble Space Telescope photos show sections of the Orion Nebula, a relatively close-by region where stars are currently forming. Each image shows an embedded circumstellar disk orbiting a very young star. Seen from different angles, some are energized to glow by the light of a nearby star while others are dark and seen in silhouette against the bright glowing gas of the Orion Nebula. Each is a contemporary analog of our own solar nebula—a location where planets are probably being formed today. (credit: modification of work by NASA/ESA, L. Ricci (ESO))

Building Planets

Circumstellar disks are a common occurrence around very young stars, suggesting that disks and stars form together. Astronomers can use theoretical calculations to see how solid bodies might form from the gas and dust in these disks as they cool. These models show that material begins to coalesce first by forming smaller objects, precursors of the planets, which we call **planetesimals**.

Today's fast computers can simulate the way millions of planetesimals, probably no larger than 100 kilometers in diameter, might gather together under their mutual gravity to form the planets we see today. We are beginning to understand that this process was a violent one, with planetesimals crashing into each other and sometimes even disrupting the growing planets themselves. As a consequence of those violent impacts (and the heat from radioactive elements in them), all the planets were heated until they were liquid and gas, and therefore differentiated, which helps explain their present internal structures.

The process of impacts and collisions in the early solar system was complex and, apparently, often random. The solar nebula model can explain many of the regularities we find in the solar system, but the random collisions of massive collections of planetesimals could be the reason for some exceptions to the "rules" of solar system behavior. For example, why do the planets Uranus and Pluto spin on their sides? Why does Venus spin slowly and in the opposite direction from the other planets? Why does the composition of the Moon resemble Earth in many ways and yet exhibit substantial differences? The answers to such questions probably lie in enormous collisions that took place in the solar system long before life on Earth began.

Today, some 4.5 billion years after its origin, the solar system is—thank goodness—a much less violent place. As we will see, however, some planetesimals have continued to interact and collide, and their fragments move about the solar system as roving "transients" that can make trouble for the established members of the Sun's family, such as our own Earth.

A great variety of **infographics (https://openstaxcollege.org/l/30worldsinsolar)** at space.com let you explore what it would be like to live on various worlds in the solar system.

7.4 | Kepler's Laws of Planetary Motion

Learning Objectives

By the end of this section, you will be able to:

- Describe the conic sections and how they relate to orbital motion.
- Describe how orbital velocity is related to orbital distance.
- Determine the period of an elliptical orbit from its semimajor axis and vice versa.

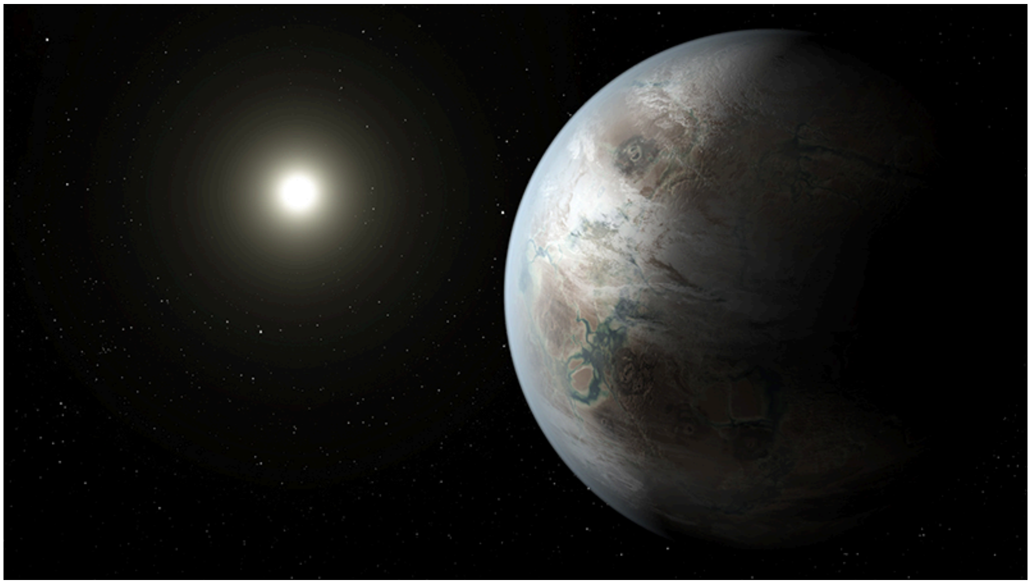


Figure 7.17 This artist's concept depicts one possible appearance of the planet Kepler-452b, the first near-Earth-size world to be found in the habitable zone of a star that is similar to our Sun. (credit: NASA)

How would you detect the existence (and determine the properties) of planets that orbit around other stars in our galaxy? That is the question facing astronomers in the 21st century. The picture of an Earth-sized planet shown in **Figure 7.17** is from the NASA Kepler Mission, named after the 16th-century scientist whose work gave us the laws of planetary kinematics.

How would you find a new planet at the outskirts of our solar system that is too dim to be seen with the unaided eye and is so far away that it moves very slowly among the stars? This was the problem confronting astronomers during the nineteenth century as they tried to pin down a full inventory of our solar system.

If we could look down on the solar system from somewhere out in space, interpreting planetary motions would be much simpler. But the fact is, we must observe the positions of all the other planets from our own moving planet. Scientists of the Renaissance did not know the details of Earth's motions any better than the motions of the other planets. Their problem was that they had to deduce the nature of all planetary motion using only their earthbound observations of the other planets' positions in the sky. To solve this complex problem more fully, better observations and better models of the planetary system were needed.

At about the time that Galileo was beginning his experiments with falling bodies, the efforts of two other scientists dramatically advanced our understanding of the motions of the planets. These two astronomers were the observer Tycho Brahe and the mathematician Johannes Kepler. Together, they placed the speculations of Copernicus on a sound mathematical basis and paved the way for the work of Isaac Newton in the next century.

Tycho Brahe's Observatory

Three years after the publication of Copernicus' *De Revolutionibus*, Tycho Brahe was born to a family of Danish nobility. He developed an early interest in astronomy and, as a young man, made significant astronomical observations. Among these was a careful study of what we now know was an exploding star that flared up to great brilliance in the night sky. His growing reputation gained him the patronage of the Danish King Frederick II, and at the age of 30, Brahe was able to establish a fine astronomical observatory on the North Sea island of Hven (**Figure 7.18**). Brahe was the last and greatest of the pre-telescopic observers in Europe.

(a) (b)

Figure 7.18 (a) A stylized engraving shows Tycho Brahe using his instruments to measure the altitude of celestial objects above the horizon. The large curved instrument in the foreground allowed him to measure precise angles in the sky. Note that the scene includes hints of the grandeur of Brahe's observatory at Hven. (b) Kepler was a German mathematician and astronomer. His discovery of the basic laws that describe planetary motion placed the heliocentric cosmology of Copernicus on a firm mathematical basis.

At Hven, Brahe made a continuous record of the positions of the Sun, Moon, and planets for almost 20 years. His extensive and precise observations enabled him to note that the positions of the planets varied from those given in published tables, which were based on the work of Ptolemy. These data were extremely valuable, but Brahe didn't have the ability to analyze them and develop a better model than what Ptolemy had published. He was further inhibited because he was an extravagant and cantankerous fellow, and he accumulated enemies among government officials. When his patron, Frederick II, died in 1597, Brahe lost his political base and decided to leave Denmark. He took up residence in Prague, where he became court astronomer to Emperor Rudolf of Bohemia. There, in the year before his death, Brahe found a most able young mathematician, Johannes Kepler, to assist him in analyzing his extensive planetary data.

Johannes Kepler

Johannes Kepler was born into a poor family in the German province of Württemberg and lived much of his life amid the turmoil of the Thirty Years' War (see **Figure 7.18**). He attended university at Tubingen and studied for a theological career. There, he learned the principles of the Copernican system and became converted to the heliocentric hypothesis. Eventually, Kepler went to Prague to serve as an assistant to Brahe, who set him to work trying to find a satisfactory theory of planetary motion—one that was compatible with the long series of observations made at Hven. Brahe was reluctant to provide Kepler with much material at any one time for fear that Kepler would discover the secrets of the universal motion by himself, thereby robbing Brahe of some of the glory. Only after Brahe's death in 1601 did Kepler get full possession of the priceless records. Their study occupied most of Kepler's time for more than 20 years.

Through his analysis of the motions of the planets, Kepler developed a series of principles, now known as *Kepler's three laws,* which described the behavior of planets based on their paths through space. The first two laws of planetary motion were published in 1609 in *The New Astronomy*. Their discovery was a profound step in the development of modern science.

The Laws of Planetary Motion

The path of an object through space is called its **orbit**. Kepler initially assumed that the orbits of planets were circles, but doing so did not allow him to find orbits that were consistent with Brahe's observations. Working with the data for Mars, he eventually discovered that the orbit of that planet had the shape of a somewhat flattened circle, or **ellipse**. Next to the circle, the ellipse is the simplest kind of closed curve, belonging to a family of curves known as *conic sections* (Figure 7.19).

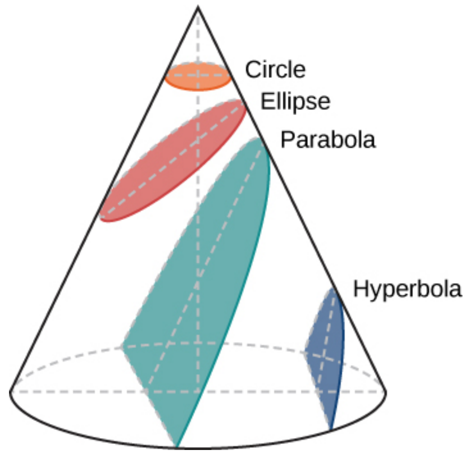


Figure 7.19 The circle, ellipse, parabola, and hyperbola are all formed by the intersection of a plane with a cone. This is why such curves are called conic sections.

Using the precise data collected by Tycho Brahe, Johannes Kepler carefully analyzed the positions in the sky of all the known planets and the Moon, plotting their positions at regular intervals of time. From this analysis, he formulated three laws, which we address in this section.

Kepler's First Law

The prevailing view during the time of Kepler was that all planetary orbits were circular. The data for Mars presented the greatest challenge to this view and that eventually encouraged Kepler to give up the popular idea. **Kepler's first law** states that every planet moves along an ellipse, with the Sun located at a focus of the ellipse. An ellipse is defined as the set of all points such that the sum of the distance from each point to two foci is a constant. Figure 7.20 shows an ellipse and describes a simple way to create it.

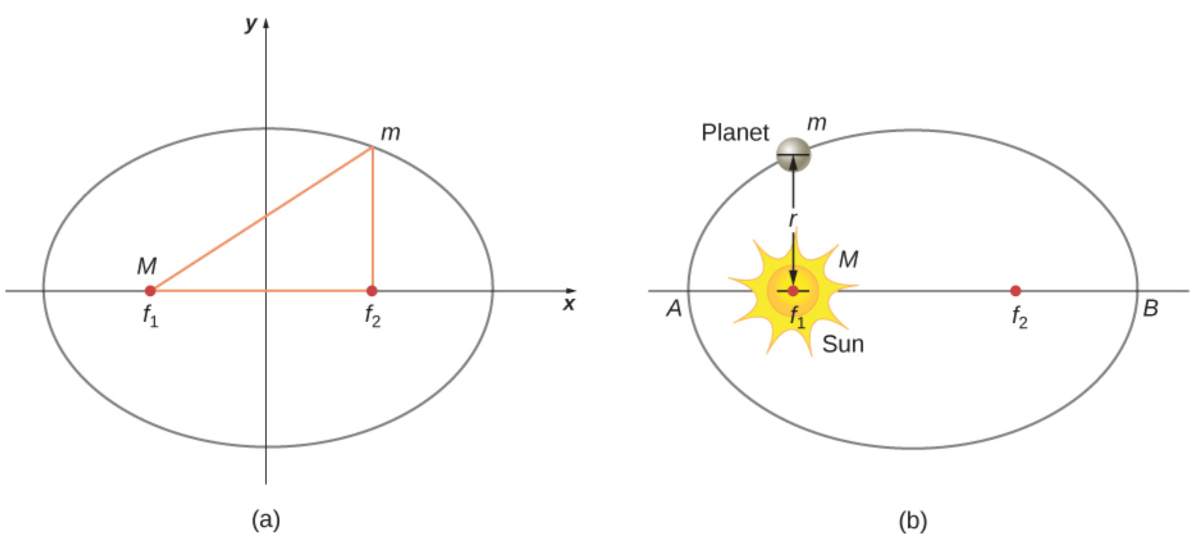


Figure 7.20 (a) An ellipse is a curve in which the sum of the distances from a point on the curve to two foci $(f_1 \text{ and } f_2)$ is a constant. From this definition, you can see that an ellipse can be created in the following way. Place a pin at each focus, then place a loop of string around a pencil and the pins. Keeping the string taught, move the pencil around in a complete circuit. If the two foci occupy the same place, the result is a circle—a special case of an ellipse. (b) For an elliptical orbit, if $m \ll M$, then *m* follows an elliptical path with *M* at one focus. More exactly, both *m* and *M* move in their own ellipse about the common center of mass.

For elliptical orbits, the point of closest approach of a planet to the Sun is called the **perihelion**. It is labeled point *A* in

Figure 7.20. The farthest point is the **aphelion** and is labeled point *B* in the figure. For the Moon's orbit about Earth, those points are called the perigee and apogee, respectively.

An ellipse has several mathematical forms, but all are a specific case of the more general equation for conic sections. There are four different conic sections, all given by the equation

$$\frac{\alpha}{r} = 1 + e\cos\theta. \tag{7.1}$$

The variables r and θ are shown in Figure 7.21 in the case of an ellipse. The values of α and e determine which of the four conic sections (hyperbola, parabola, ellipse or circle) represents the path of the satellite. For an ellipse, $0 \le e < 1$, with an eccentricity of zero meaning a circular orbit.

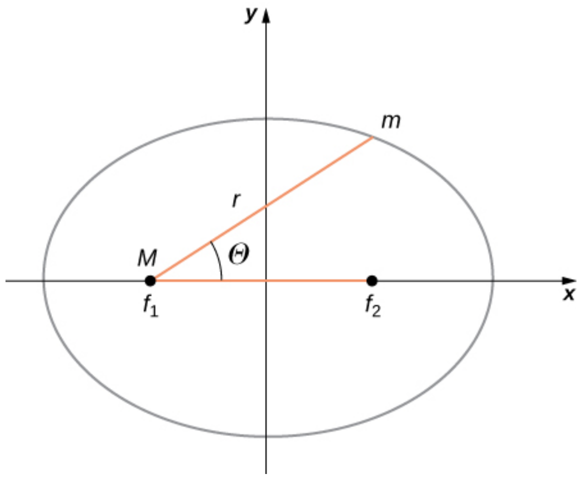


Figure 7.21 As before, the distance between the planet and the Sun is *r*, and the angle measured from the *x*-axis, which is along the major axis of the ellipse, is θ.

You can see an animation of two interacting objects at the *My Solar System* page at Phet (https://openstaxcollege.org/l/21mysolarsys) . Choose the Sun and Planet preset option. You can also view the more complicated multiple body problems as well. You may find the actual path of the Moon quite surprising, yet is obeying Newton's simple laws of motion.

Kepler's Second Law

Kepler's second law states that a planet sweeps out equal areas in equal times, that is, the area divided by time, called the areal velocity, is constant. Consider Figure 7.22. The time it takes a planet to move from position *A* to *B*, sweeping out area A_1, is exactly the time taken to move from position *C* to *D*, sweeping area A_2, and to move from *E* to *F*, sweeping out area A_3. These areas are the same: $A_1 = A_2 = A_3$.

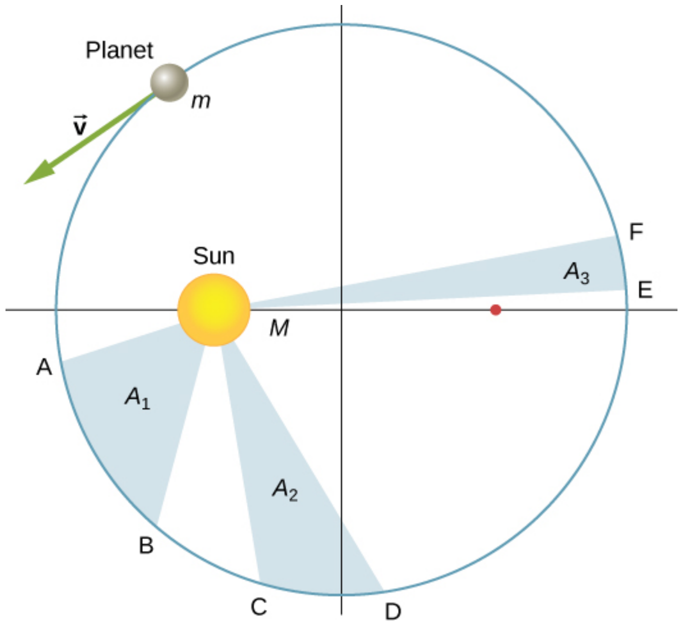


Figure 7.22 The shaded regions shown have equal areas and represent the same time interval.

Comparing the areas in the figure and the distance traveled along the ellipse in each case, we can see that in order for the areas to be equal, the planet must speed up as it gets closer to the Sun and slow down as it moves away.

Now consider **Figure 7.23**. A small triangular area ΔA is swept out in time Δt. The velocity is along the path and it makes an angle θ with the radial direction. Hence, the perpendicular velocity is given by $v_{\text{perp}} = v\sin\theta$. The planet moves a distance $\Delta s = v\Delta t\sin\theta$ projected along the direction perpendicular to *r*. Since the area of a triangle is one-half the base (r) times the height (Δs), for a small displacement, the area is given by $\Delta A = \frac{1}{2}r\Delta s$.

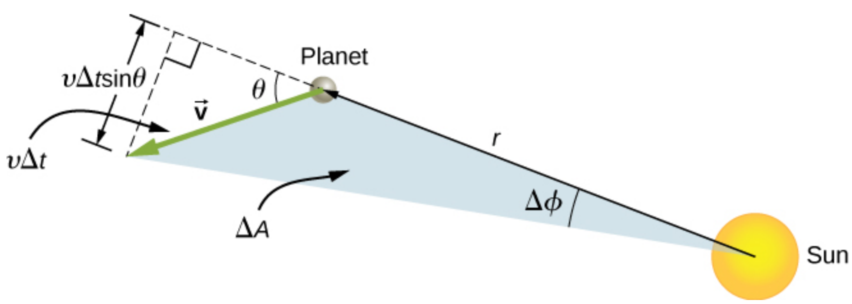


Figure 7.23 The element of area ΔA swept out in time Δt as the planet moves through angle $\Delta\phi$. The angle between the radial direction and $\vec{v}$ is θ.

The areal velocity is simply the rate of change of area with time, so we have

$$\text{areal velocity} = \frac{\Delta A}{\Delta t} = \frac{\frac{1}{2}r\,\Delta s}{\Delta t} = \frac{1}{2}r\,v \tag{7.2}$$

The fact that the areal velocity remains constant, then, implies that the product of the planet's distance from the Sun (r) and its instantaneous speed (v) is a constant of the motion. Hence, the closer it is to the Sun, the faster it moves, and vice versa.

You can view an **animated version (https://openstaxcollege.org/l/21animationgrav)** of **Figure 7.22**, and many other interesting animations as well, at the School of Physics (University of New South Wales) site.

Kepler's Third Law

Kepler's first two laws of planetary motion describe the shape of a planet's orbit and allow us to calculate the speed of its motion at any point in the orbit. Kepler was pleased to have discovered such fundamental rules, but they did not satisfy his quest to fully understand planetary motions. He wanted to know why the orbits of the planets were spaced as they are and to find a mathematical pattern in their movements—a "harmony of the spheres" as he called it. For many years he worked to discover mathematical relationships governing planetary spacing and the time each planet took to go around the Sun.

In 1619, Kepler discovered a basic relationship to relate the planets' orbits to their relative distances from the Sun. We define a planet's **orbital period**, (***T***), as the time it takes a planet to travel once around the Sun. Also, recall that a planet's semimajor axis, *a*, is equal to its average distance from the Sun. The relationship, now known as *Kepler's third law*, says that a planet's orbital period squared is proportional to the semimajor axis of its orbit cubed, or

$$T^2 \propto a^3$$

When *T* (the orbital period) is measured in years, and *a* is expressed in a quantity known as an **astronomical unit (AU)**, the two sides of the formula are not only proportional but equal. One AU is the average distance between Earth and the Sun and is approximately equal to 1.5×10^8 kilometers. In these units,

Kepler's Third Law in its Original Form

$$T^2 = a^3 \tag{7.3}$$

Kepler's third law applies to all objects orbiting the Sun, including Earth, and provides a means for calculating their relative distances from the Sun from the time they take to orbit. Let's look at a specific example to illustrate how useful Kepler's third law is.

For instance, suppose you time how long Mars takes to go around the Sun (in Earth years). Kepler's third law can then be used to calculate Mars' average distance from the Sun. Mars' orbital period (1.88 Earth years) squared, or T^2, is $1.88^2 = 3.53$, and according to the equation for Kepler's third law, this equals the cube of its semimajor axis, or a^3. So what number must be cubed to give 3.53? The answer is 1.52 (since 1.52 × 1.52 × 1.52 = 3.53). Thus, Mars' semimajor axis in astronomical units must be 1.52 AU. In other words, to go around the Sun in a little less than two years, Mars must be about 50% (half again) as far from the Sun as Earth is.

Example 7.2

Calculating Periods

Imagine an object is traveling around the Sun. What would be the orbital period of the object if its orbit has a semimajor axis of 50 AU?

Solution

From Kepler's third law, we know that (when we use units of years and AU)

$$T^2 = a^3$$

If the object's orbit has a semimajor axis of 50 AU (*a* = 50), we can cube 50 and then take the square root of the result to get T:

$$\begin{aligned} T &= \sqrt{a^3} \\ T &= \sqrt{50 \times 50 \times 50} = \sqrt{125{,}000} = 353.6 \text{ years} \end{aligned}$$

Therefore, the orbital period of the object is about 350 years. This would place our hypothetical object beyond the orbit of Pluto.

Kepler's third law states that the square of the period is proportional to the cube of the semi-major axis of the orbit. When written in the form of Equation 7.3, it is simple and easy to use for objects orbiting our Sun. The distances are measured in AU and the periods in years. These units came naturally out of Kepler's work on the solar system. But, in fact, Kepler's Third Law is much more general, and can be applied to bodies orbiting any large object. To do so, it makes the most sense to measure quantities in SI units. Before we do so, however, we must realize that Kepler's Third Law in the form of Equation 7.3 does in fact contain a missing constant of proportionality, which is the the inverse of the mass of the Sun. The reason that it does not appear in Equation 7.3 is that it is measured in units of "solar masses", which for the Sun has a value of exactly 1. Nevertheless, the full equation for Kepler's Third Law should read:

Kepler's Third Law in Solar Units

$$T^2 = \frac{a^3}{M} \quad (7.4)$$

where M is the mass of the large body (around which the satellite is in orbit) measured in units of solar masses. (Again, T is measured in years, while a is measured in AU.)

Finally, to express all quantities in SI units, we can rewrite the equation in what is often referred to as "Newton's version":

Newton's Version of Kepler's Third Law

$$T^2 = \frac{4\pi^2}{GM}a^3 \quad (7.5)$$

In this equation, the period T is measured in seconds, the distance a in meters, and the mass M in kg. The constant $G = 6.67{\times}10^{-11}\ \text{N} \cdot \text{m}^2/\text{kg}^2$ is Newton's universal gravitational constant, which we will encounter in Chapter 9.

Example 7.3

Orbit of Halley's Comet

Determine the semi-major axis of the orbit of Halley's comet, given that it arrives at perihelion every 75.3 years. If the perihelion is 0.586 AU, what is the aphelion?

Strategy

We are given the period, so we can rearrange **Equation 7.5**, solving for the semi-major axis. Since we know the value for the perihelion, we can use the definition of the semi-major axis, given earlier in this section, to find the aphelion. We note that 1 Astronomical Unit (AU) is the average radius of Earth's orbit and is defined to be $1\ \text{AU} = 1.50 \times 10^{11}\ \text{m}$.

Solution

Rearranging **Equation 7.5** and inserting the values of the period of Halley's comet and the mass of the Sun, we have

$$\begin{aligned} a &= \left(\frac{GM}{4\pi^2}T^2\right)^{1/3} \\ &= \left(\frac{(6.67 \times 10^{-11}\ \text{N} \cdot \text{m}^2/\text{kg}^2)(2.00 \times 10^{30}\ \text{kg})}{4\pi^2}(75.3\ \text{yr} \times 365\ \text{days/yr} \times 24\ \text{hr/day} \times 3600\ \text{s/hr})^2\right)^{1/3}. \end{aligned}$$

This yields a value of $2.67 \times 10^{12}\ \text{m}$ or 17.8 AU for the semi-major axis.

The semi-major axis is one-half the sum of the aphelion and perihelion, so we have

$$\begin{aligned} a &= \frac{1}{2}(\text{aphelion} + \text{perihelion}) \\ \text{aphelion} &= 2a - \text{perihelion}. \end{aligned}$$

Substituting for the values, we found for the semi-major axis and the value given for the perihelion, we find the value of the aphelion to be 35.0 AU.

Significance

Edmond Halley, a contemporary of Newton, first suspected that three comets, reported in 1531, 1607, and 1682, were actually the same comet. Before Tycho Brahe made measurements of comets, it was believed that they were one-time events, perhaps disturbances in the atmosphere, and that they were not affected by the Sun. Halley used Newton's new mechanics to predict his namesake comet's return in 1758.

7.2 Check Your Understanding The nearly circular orbit of Saturn has an average radius of about 9.5 AU and has a period of 30 years, whereas Uranus averages about 19 AU and has a period of 84 years. Is this consistent with our results for Halley's comet?

For Further Exploration

Websites

Gazetteer of Planetary Nomenclature: **http://planetarynames.wr.usgs.gov/ (http://planetarynames.wr.usgs.gov/)** . Outlines the rules for naming bodies and features in the solar system.

Planetary Photojournal: **http://photojournal.jpl.nasa.gov/index.html (http://photojournal.jpl.nasa.gov/index.html)** . This NASA site features thousands of the best images from planetary exploration, with detailed captions and excellent indexing. You can find images by world, feature name, or mission, and download them in a number of formats. And the images are copyright-free because your tax dollars paid for them.

The following sites present introductory information and pictures about each of the worlds of our solar system:

- **NASA/JPL Solar System Exploration pages: http://solarsystem.nasa.gov/index.cfm (http://solarsystem.nasa.gov/index.cfm) .**
- **National Space Science Data Center Lunar and Planetary Science pages: http://nssdc.gsfc.nasa.gov/planetary/ (http://nssdc.gsfc.nasa.gov/planetary/) .**
- **Nine [now 8] Planets Solar System Tour: http://www.nineplanets.org/ (http://www.nineplanets.org/) .**
- **Planetary Society solar system pages: http://www.planetary.org/explore/space-topics/compare/ (http://www.planetary.org/explore/space-topics/compare/) .**
- **Views of the Solar System by Calvin J. Hamilton: http://www.solarviews.com/eng/homepage.htm (http://www.solarviews.com/eng/homepage.htm) .**

Videos

Brown Dwarfs and Free Floating Planets: When You Are Just Too Small to Be a Star: **https://www.youtube.com/watch?v=zXCDsb4n4KU (https://www.youtube.com/watch?v=zXCDsb4n4KU)** . A nontechnical talk by Gibor Basri of the University of California at Berkeley, discussing some of the controversies about the meaning of the word "planet" (1:32:52).

In the Land of Enchantment: The Epic Story of the Cassini Mission to Saturn: **https://www.youtube.com/watch?v=Vx135n8VFxY (https://www.youtube.com/watch?v=Vx135n8VFxY)** . A public lecture by Dr. Carolyn Porco that focuses mainly on the exploration of Saturn and its moons, but also presents an eloquent explanation of why we explore the solar system (1:37:52).

Origins of the Solar System: **http://www.pbs.org/wgbh/nova/space/origins-solar-system.html (http://www.pbs.org/wgbh/nova/space/origins-solar-system.html)** . A video from PBS that focuses on the evidence from meteorites, narrated by Neil deGrasse Tyson (13:02).

To Scale: The Solar System: **https://www.youtube.com/watch?t=84&v=zR3Igc3Rhfg (https://www.youtube.com/watch?t=84&v=zR3Igc3Rhfg)** . Constructing a scale model of the solar system in the Nevada desert (7:06).

CHAPTER 7 REVIEW

KEY TERMS

aphelion farthest point from the Sun of an orbiting body; the corresponding term for the Moon's farthest point from Earth is the apogee

asteroid a stony or metallic object orbiting the Sun that is smaller than a major planet but that shows no evidence of an atmosphere or of other types of activity associated with comets

astronomical unit (AU) the unit of length defined as the average distance between Earth and the Sun; this distance is about 1.5×10^8 kilometers

comet a small body of icy and dusty matter that revolves about the Sun; when a comet comes near the Sun, some of its material vaporizes, forming a large head of tenuous gas and often a tail

differentiation gravitational separation of materials of different density into layers in the interior of a planet or moon

eccentricity in an ellipse, the ratio of the distance between the foci to the major axis

ellipse a closed curve for which the sum of the distances from any point on the ellipse to two points inside (called the foci) is always the same

focus (plural: foci) one of two fixed points inside an ellipse from which the sum of the distances to any point on the ellipse is constant

giant planet any of the planets Jupiter, Saturn, Uranus, and Neptune in our solar system, or planets of roughly that mass and composition in other planetary systems

Kepler's first law each planet moves around the Sun in an orbit that is an ellipse, with the Sun at one focus of the ellipse

Kepler's second law the straight line joining a planet and the Sun sweeps out equal areas in space in equal intervals of time

Kepler's third law the square of a planet's orbital period is directly proportional to the cube of the semimajor axis of its orbit

major axis the maximum diameter of an ellipse

meteor a small piece of solid matter that enters Earth's atmosphere and burns up, popularly called a *shooting star* because it is seen as a small flash of light

meteorite a portion of a meteor that survives passage through an atmosphere and strikes the ground

orbit the path of an object that is in revolution about another object or point

orbital period (*T*) the time it takes an object to travel once around the Sun

orbital speed the speed at which an object (usually a planet) orbits around the mass of another object; in the case of a planet, the speed at which each planet moves along its ellipse

perihelion point of closest approach to the Sun of an orbiting body; the corresponding term for the Moon's closest approach to Earth is the perigee

planetesimals objects, from tens to hundreds of kilometers in diameter, that formed in the solar nebula as an intermediate step between tiny grains and the larger planetary objects we see today; the comets and some asteroids may be leftover planetesimals

semimajor axis half of the major axis of a conic section, such as an ellipse

solar nebula the cloud of gas and dust from which the solar system formed

terrestrial planet any of the planets Mercury, Venus, Earth, or Mars; sometimes the Moon is included in the list

KEY EQUATIONS

Kepler's Second Law	$\text{areal velocity} = \frac{\Delta A}{\Delta t} = \frac{1}{2} r\, v = \text{constant}$

Kepler's Third Law in Solar Units $T^2 = \frac{a^3}{M}$

Kepler's Third Law in SI Units $T^2 = \frac{4\pi^2}{GM}a^3$

SUMMARY

7.1 Overview of Our Planetary System

- Our solar system currently consists of the Sun, eight planets, five dwarf planets, nearly 200 known moons, and a host of smaller objects.
- The planets can be divided into two groups: the inner terrestrial planets and the outer giant planets.
- Pluto, Eris, Haumea, and Makemake do not fit into either category; as icy dwarf planets, they exist in an ice realm on the fringes of the main planetary system.
- The giant planets are composed mostly of liquids and gases.
- Smaller members of the solar system include asteroids (including the dwarf planet Ceres), which are rocky and metallic objects found mostly between Mars and Jupiter; comets, which are made mostly of frozen gases and generally orbit far from the Sun; and countless smaller grains of cosmic dust.
- When a meteor survives its passage through our atmosphere and falls to Earth, we call it a meteorite.

7.2 Composition and Structure of Planets

- The giant planets have dense cores roughly 10 times the mass of Earth, surrounded by layers of hydrogen and helium.
- The terrestrial planets consist mostly of rocks and metals. They were once molten, which allowed their structures to differentiate (that is, their denser materials sank to the center).
- The Moon resembles the terrestrial planets in composition, but most of the other moons—which orbit the giant planets—have larger quantities of frozen ice within them.
- In general, worlds closer to the Sun have higher surface temperatures.
- The surfaces of terrestrial planets have been modified by impacts from space and by varying degrees of geological activity.

7.3 Origin of the Solar System

- Regularities among the planets have led astronomers to hypothesize that the Sun and the planets formed together in a giant, spinning cloud of gas and dust called the solar nebula.
- Astronomical observations show tantalizingly similar circumstellar disks around other stars.
- Within the solar nebula, material first coalesced into planetesimals; many of these gathered together to make the planets and moons.
- The remainder can still be seen as comets and asteroids.
- Probably all planetary systems have formed in similar ways, but many exoplanet systems have evolved along quite different paths, as we will see in **New Perspectives on Planet Formation**.

7.4 Kepler's Laws of Planetary Motion

- All orbital motion follows the path of a conic section. Bound or closed orbits are either a circle or an ellipse; unbounded or open orbits are either a parabola or a hyperbola.
- The areal velocity of any orbit is constant, meaning that the product of the body's speed and its distance from the Sun is constant.
- The square of the period of an elliptical orbit is proportional to the cube of the semi-major axis of that orbit.

CONCEPTUAL QUESTIONS

7.3 Origin of the Solar System

1. Venus rotates backward and Uranus and Pluto spin about an axis tipped nearly on its side. Based on what you learned about the motion of small bodies in the solar system and the surfaces of the planets, what might be the cause of these strange rotations?

2. What is the difference between a differentiated body and an undifferentiated body, and how might that influence a body's ability to retain heat for the age of the solar system?

3. What does a planet need in order to retain an atmosphere? How does an atmosphere affect the surface of a planet and the ability of life to exist?

4. Which type of planets have the most moons? Where did these moons likely originate?

5. What is the difference between a meteor and a meteorite?

6. Explain our ideas about why the terrestrial planets are rocky and have less gas than the giant planets.

7. Do all planetary systems look the same as our own?

8. What is comparative planetology and why is it useful to astronomers?

9. What changed in our understanding of the Moon and Moon-Earth system as a result of humans landing on the Moon's surface?

10. If Earth was to be hit by an extraterrestrial object, where in the solar system could it come from and how would we know its source region?

11. List some reasons that the study of the planets has progressed more in the past few decades than any other branch of astronomy.

12. Imagine you are a travel agent in the next century. An eccentric billionaire asks you to arrange a "Guinness Book of Solar System Records" kind of tour. Where would you direct him to find the following (use this chapter and **Appendix D**:

A. the least-dense planet
B. the densest planet
C. the largest moon in the solar system
D. excluding the jovian planets, the planet where you would weigh the most on its surface (Hint: Weight is directly proportional to surface gravity.)
E. the smallest planet
F. the planet that takes the longest time to rotate
G. the planet that takes the shortest time to rotate
H. the planet with a diameter closest to Earth's
I. the moon with the thickest atmosphere
J. the densest moon
K. the most massive moon

13. What characteristics do the worlds in our solar system have in common that lead astronomers to believe that they all formed from the same "mother cloud" (solar nebula)?

14. How do terrestrial and giant planets differ? List as many ways as you can think of.

15. Why are there so many craters on the Moon and so few on Earth?

16. How do asteroids and comets differ?

17. How and why is Earth's Moon different from the larger moons of the giant planets?

18. Where would you look for some "original" planetesimals left over from the formation of our solar system?

19. What was the solar nebula like? Why did the Sun form at its center?

20. What can we learn about the formation of our solar system by studying other stars? Explain.

21. Earlier in this chapter, we modeled the solar system with Earth at a distance of about one city block from the Sun. If you were to make a model of the distances in the solar system to match your height, with the Sun at the top of your head and Pluto at your feet, which planet would be near your waist? How far down would the zone of the terrestrial planets reach?

22. Seasons are a result of the inclination of a planet's axial tilt being inclined from the normal of the planet's orbital plane. For example, Earth has an axis tilt of 23.4° (**Table 34.2**). Using information about just the inclination alone, which planets might you expect to have seasonal cycles similar to Earth, although different in duration because orbital periods around the Sun are different?

23. Again using Appendix D, which planet(s) might you expect not to have significant seasonal activity? Why?

24. Again using Appendix D, which planets might you expect to have extreme seasons? Why?

25. Using some of the astronomical resources in your college library or the Internet, find five names of features on each of three other worlds that are named after real people. In a sentence or two, describe each of these people and what contributions they made to the progress of science or human thought.

26. Explain why the planet Venus is differentiated, but asteroid Fraknoi, a very boring and small member of the asteroid belt, is not.

27. Would you expect as many impact craters per unit area on the surface of Venus as on the surface of Mars? Why or why not?

28. Interview a sample of 20 people who are not taking an astronomy class and ask them if they can name a living astronomer. What percentage of those interviewed were able to name one? Typically, the two living astronomers the public knows these days are Stephen Hawking and Neil deGrasse Tyson. Why are they better known than most astronomers? How would your result have differed if you had asked the same people to name a movie star or a professional basketball player?

29. Using Appendix D, complete the following table that describes the characteristics of the Galilean moons of Jupiter, starting from Jupiter and moving outward in distance.

Moon	Semimajor Axis (km^3)	Diameter	Density (g/cm^3)
Io			
Europa			
Ganymede			
Callisto			

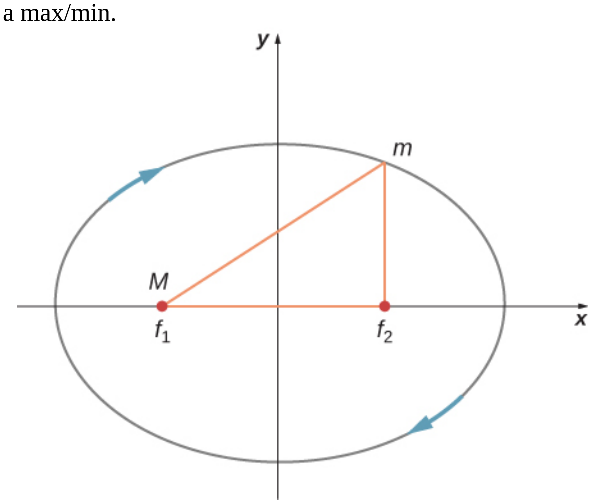
Table A

This system has often been described as a mini solar system. Why might this be so? If Jupiter were to represent the Sun and the Galilean moons represented planets, which moons could be considered more terrestrial in nature and which ones more like gas/ice giants? Why? (Hint: Use the values in your table to help explain your categorization.)

7.4 Kepler's Laws of Planetary Motion

30. Are Kepler's laws purely descriptive, or do they contain causal information?

31. In the diagram below for a satellite in an elliptical orbit about a much larger mass, indicate where its speed is the greatest and where it is the least. What conservation law dictates this behavior? Indicate the directions of the force, acceleration, and velocity at these points. Draw vectors for these same three quantities at the two points where the *y*-axis intersects (along the semi-minor axis) and from this determine whether the speed is increasing decreasing, or at a max/min.



PROBLEMS

7.3 Origin of the Solar System

32. Calculate the density of Jupiter. Show your work. Is it more or less dense than Earth? Why?

33. Calculate the density of Saturn. Show your work. How does it compare with the density of water? Explain how this can be.

34. What is the density of Jupiter's moon Europa (see Appendix D for data on moons)? Show your work.

35. Look at Appendix D and indicate the moon with a diameter that is the largest fraction of the diameter of the planet or dwarf planet it orbits.

36. Barnard's Star, the second closest star to us, is about

56 trillion (5.6×10^{12}) km away. Calculate how far it would be using the scale model of the solar system given in **Section 7.1**.

7.4 Kepler's Laws of Planetary Motion

37. Calculate the mass of the Sun based on data for average Earth's orbit and compare the value obtained with the Sun's commonly listed value of 1.989×10^{30} kg.

38. Io orbits Jupiter with an average radius of 421,700 km and a period of 1.769 days. Based upon these data, what is the mass of Jupiter?

39. The "mean" orbital radius listed for astronomical objects orbiting the Sun is typically not an integrated average but is calculated such that it gives the correct period when applied to the equation for circular orbits. Given that, what is the mean orbital radius in terms of aphelion and perihelion?

40. The perihelion of Halley's comet is 0.586 AU and the aphelion is 17.8 AU. Given that its speed at perihelion is 55 km/s, what is the speed at aphelion ($1 \text{ AU} = 1.496 \times 10^{11} \text{ m}$)? (*Hint:* You may use either conservation of energy or angular momentum, but the latter is much easier.)

41. The perihelion of the comet Lagerkvist is 2.61 AU and it has a period of 7.36 years. Show that the aphelion for this comet is 4.95 AU.

42. What is the ratio of the speed at perihelion to that at aphelion for the comet Lagerkvist in the previous problem?

43. Eros has an elliptical orbit about the Sun, with a perihelion distance of 1.13 AU and aphelion distance of 1.78 AU. What is the period of its orbit?

8 | NEWTON'S SYNTHESIS

Chapter Outline

Introduction

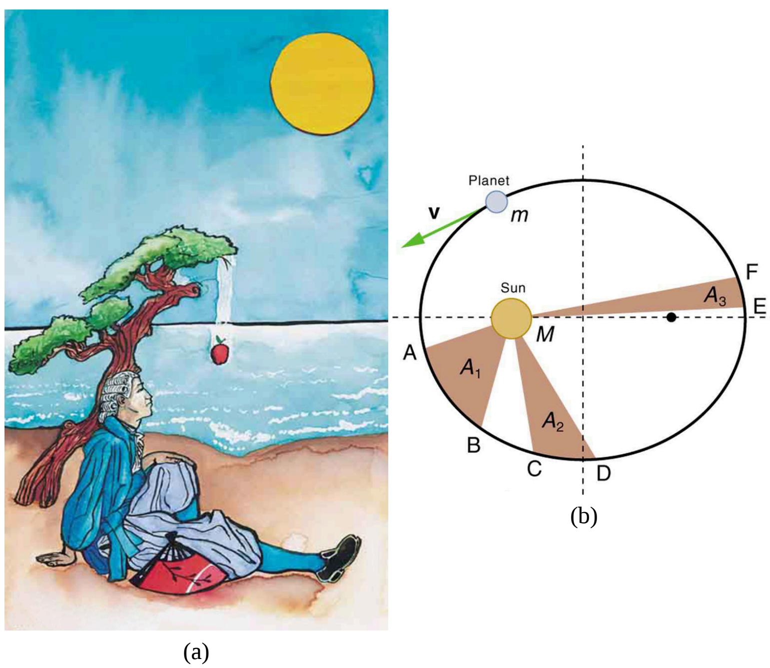


Figure 8.1 (a) We have long known that gravity causes the apple to fall straight down with an acceleration of g. (b) We have also known that the planets orbit the Sun in elliptical (almost circular) paths obeying Kepler's laws. (credits: OpenStax College Physics)

We have developed a description of motions, both translational and rotational, but now we want to take things a step further. The question we seek to answer is, "**Why** do things move as they do?" In particular, we have quantitatively studied two very important motions:

A. Free fall of objects near Earth's surface; and
B. Orbits of the planets around the Sun.

The first is an example of **terrestrial motion** - it takes place here on Earth. The second is an example of **celestial motion** - it takes place in the realm of the planets and stars. You will recall that our studies in kinematics have revealed that these two motions are described, respectively, by:

A. Motion with a constant downward acceleration of $g = 9.8\ \text{m/s}^2$

B. Kepler's Three Laws:

  1. The planets follow elliptical paths;
  2. The product of a planet's speed v times its distance from the Sun r remains constant; and
  3. The square of a planet's orbital period is proportional to the cube of its average distance from the Sun $T^2 = a^3$ (if we choose the appropriate units for the quantities involved)

For centuries, terrestrial motion and celestial motion were considered to be separate subjects, because they referred to actions that seemed to occur in two completely different realms. The successful explanations of both of these types of motion was achieved, in a single stroke, by perhaps the single most influential physicist of all time - Sir Isaac Newton. His study of the dynamics of gravity brought about a synthesis of the ideas about terrestrial motion and celestial motion. And it made a profound change in that it demonstrated that there are not two, separate realms of reality in the physical world. Both terrestrial and celestial motions are attributable to the same set of physical principles, called Newton's Laws.

The Newtonian synthesis was achieved by, first, laying out the principles that describe how objects in the Universe interact with one another, and secondly by explaining the nature of gravity itself. The gravitational force affects objects on Earth and the motion of the Universe itself. Gravity is the first force to be postulated as an action-at-a-distance force, that is, objects exert a gravitational force on one another without physical contact and that force falls to zero only at an infinite distance. Earth exerts a gravitational force on you, but so do our Sun, the Milky Way galaxy, and the billions of galaxies, like those shown above, which are so distant that we cannot see them with the naked eye.

The road to an explanation begins with the concept of **force**. Forces affect every moment of your life. Your body is held to Earth by force and held together by the forces of charged particles. When you open a door, walk down a street, lift your fork, or touch a baby’s face, you are applying forces. Zooming in deeper, your body’s atoms are held together by electrical forces, and the core of the atom, called the nucleus, is held together by the strongest force we know—strong nuclear force.

8.1 | Forces

Learning Objectives

By the end of the section, you will be able to:

- Distinguish between kinematics and dynamics
- Understand the definition of force
- Identify simple free-body diagrams
- Define the SI unit of force, the newton
- Describe force as a vector

Kinematics only describes the way objects move—their velocity and their acceleration. **Dynamics** is the study of how forces affect the motion of objects and systems. It considers the causes of motion of objects and systems of interest, where a system is anything being analyzed. The foundation of dynamics are the laws of motion stated by Isaac Newton (1642–1727). These laws provide an example of the breadth and simplicity of principles under which nature functions. They are also universal laws in that they apply to situations on Earth and in space.

Newton’s laws of motion were just one part of the monumental work that has made him legendary (**Figure 8.2**). The development of Newton’s laws marks the transition from the Renaissance to the modern era. Not until the advent of modern physics was it discovered that Newton’s laws produce a good description of motion only when the objects are moving at speeds much less than the speed of light and when those objects are larger than the size of most molecules (about 10^{-9} m in diameter). These constraints define the realm of Newtonian mechanics. At the beginning of the twentieth century, Albert Einstein (1879–1955) developed the theory of relativity and, along with many other scientists, quantum mechanics. Quantum mechanics does not have the constraints present in Newtonian physics. All of the situations we consider in this chapter are in the realm of Newtonian physics.

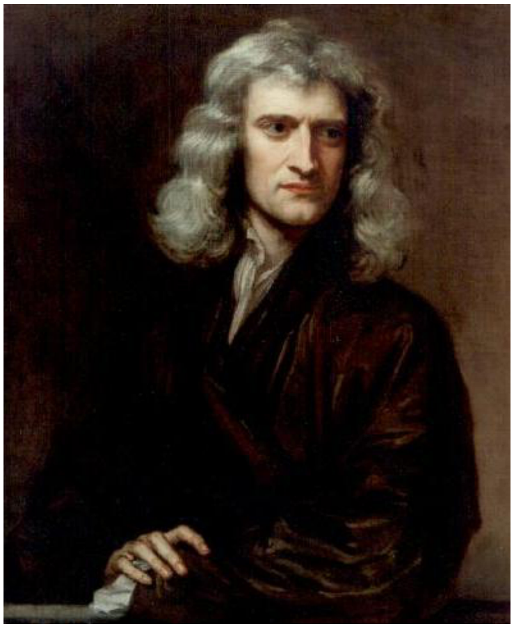


Figure 8.2 Isaac Newton (1642–1727) published his amazing work, *Philosophiae Naturalis Principia Mathematica*, in 1687. It proposed scientific laws that still apply today to describe the motion of objects (the laws of motion). Newton also discovered the law of gravity, invented calculus, and made great contributions to the theories of light and color.

Working Definition of Force

Dynamics is the study of the forces that cause objects and systems to move. To understand this, we need a working definition of force. An intuitive definition of **force**—that is, a push or a pull—is a good place to start. We know that a push or a pull has both magnitude and direction (therefore, it is a vector quantity), so we can define force as the push or pull on an object with a specific magnitude and direction. Force can be represented by vectors or expressed as a multiple of a standard force.

The push or pull on an object can vary considerably in either magnitude or direction. For example, a cannon exerts a strong force on a cannonball that is launched into the air. In contrast, Earth exerts only a tiny downward pull on a flea. Our everyday experiences also give us a good idea of how multiple forces add. If two people push in different directions on a third person, as illustrated in **Figure 8.3**, we might expect the total force to be in the direction shown. Since force is a vector, it adds just like other vectors. Forces, like other vectors, are represented by arrows and can be added using the familiar head-to-tail method or trigonometric methods. These ideas were developed in **An Introduction to Vectors**.

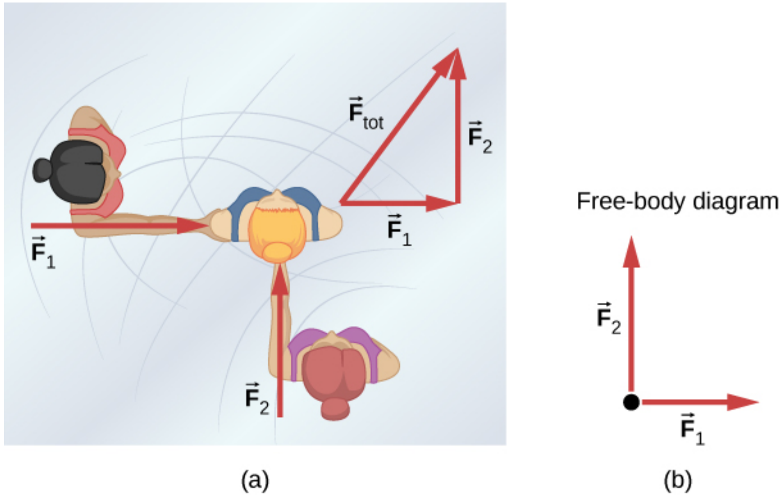


Figure 8.3 (a) An overhead view of two ice skaters pushing on a third skater. Forces are vectors and add like other vectors, so the total force on the third skater is in the direction shown. (b) A free-body diagram representing the forces acting on the third skater.

Figure 8.3(b) is our first example of a **free-body diagram**, which is a sketch showing all external forces acting on an object or system. The object or system is represented by a single isolated point (or free body), and only those forces acting *on* it that originate outside of the object or system—that is, **external forces**—are shown. (These forces are the only ones shown because only external forces acting on the free body affect its motion. We can ignore any internal forces within the body.) The forces are represented by vectors extending outward from the free body.

Free-body diagrams are useful in analyzing forces acting on an object or system, and are employed extensively in the study and application of Newton's laws of motion. You will see them throughout this text and in all your studies of physics. The following steps briefly explain how a free-body diagram is created; we examine this strategy in more detail in **Drawing Free-Body Diagrams**.

Problem-Solving Strategy: Drawing Free-Body Diagrams

1. Draw the object under consideration. If you are treating the object as a particle, represent the object as a point. Place this point at the origin of an *xy*-coordinate system.
2. Include all forces that act on the object, representing these forces as vectors. However, do not include the net force on the object or the forces that the object exerts on its environment.
3. Resolve all force vectors into *x*- and *y*-components.
4. Draw a separate free-body diagram for each object in the problem.

We illustrate this strategy with two examples of free-body diagrams (**Figure 8.4**). The terms used in this figure are explained in more detail later in the chapter.

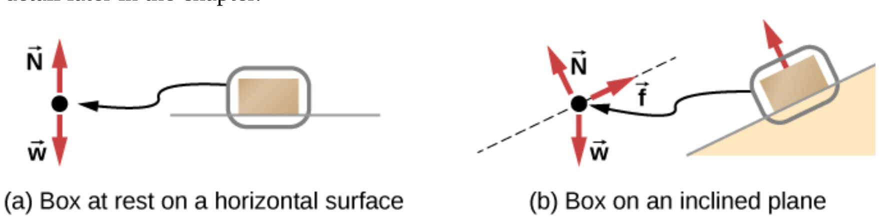


Figure 8.4 In these free-body diagrams, $\vec{\mathbf{N}}$ is the normal force, $\vec{\mathbf{w}}$ is the weight of the object, and $\vec{\mathbf{f}}$ is the friction.

The steps given here are sufficient to guide you in this important problem-solving strategy. The final section of this chapter explains in more detail how to draw free-body diagrams when working with the ideas presented in this chapter.

Development of the Force Concept

Forces are inherently vector quantities. They always have both a magnitude and, importantly, a specific direction in which they act on some body. Because of this, we will use vector notation when specifying any particular force. However, the discussion and examples in this chapter will focus on situations where we are concerned only with multiple forces that act in the same (or the opposite) direction. And, while the magnitudes and/or direction of a force may vary in time, we will limit our initial discussion to forces that are constant in time. Such constant-force, one-dimensional examples are an easy way to begin our study of forces, and **Figure 8.5** shows one such example.

Let's analyze force more deeply. Suppose a physics student sits at a table, working diligently on his homework (**Figure 8.5**). What external forces act on him? Can we determine the origin of these forces?

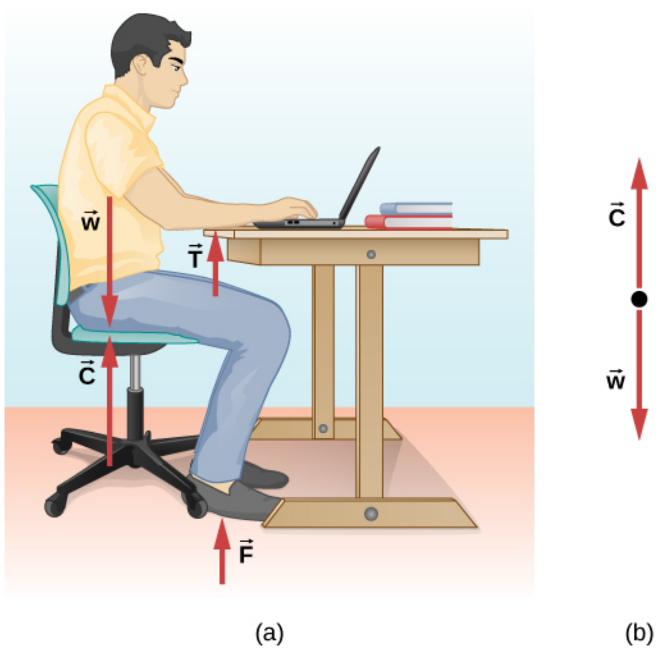


Figure 8.5 (a) The forces acting on the student are due to the chair, the table, the floor, and Earth's gravitational attraction. (b) In solving a problem involving the student, we may want to consider the forces acting along the line running through his torso. A free-body diagram for this situation is shown.

In most situations, forces are grouped into two categories: *contact forces* and *field forces*. As you might guess, contact forces are due to direct physical contact between objects. For example, the student in **Figure 8.5** experiences the contact forces $\vec{\mathbf{C}}$, $\vec{\mathbf{F}}$, and $\vec{\mathbf{T}}$, which are exerted by the chair on his posterior, the floor on his feet, and the table on his forearms, respectively. Field forces, however, act without the necessity of physical contact between objects. They depend on the presence of a "field" in the region of space surrounding the body under consideration. Since the student is in Earth's gravitational field, he feels a gravitational force $\vec{\mathbf{w}}$; in other words, he has weight.

You can think of a field as a property of space that is detectable by the forces it exerts. Scientists think there are only four fundamental force fields in nature. These are the gravitational, electromagnetic, strong nuclear, and weak fields (we consider these four forces in nature later in this text). As noted for $\vec{\mathbf{w}}$ in **Figure 8.5**, the gravitational field is responsible for the weight of a body. The forces of the electromagnetic field include those of static electricity and magnetism; they are also responsible for the attraction among atoms in bulk matter. Both the strong nuclear and the weak force fields are effective only over distances roughly equal to a length of scale no larger than an atomic nucleus (10^{-15} m). Their range is so small that neither field has influence in the macroscopic world of Newtonian mechanics.

Contact forces are fundamentally electromagnetic. While the elbow of the student in **Figure 8.5** is in contact with the tabletop, the atomic charges in his skin interact electromagnetically with the charges in the surface of the table. The net (total) result is the force $\vec{\mathbf{T}}$. Similarly, when adhesive tape sticks to a piece of paper, the atoms of the tape are intermingled with those of the paper to cause a net electromagnetic force between the two objects. However, in the context of Newtonian mechanics, the electromagnetic origin of contact forces is not an important concern.

Vector Notation for Force

As previously discussed, force is a vector; it has both magnitude and direction. The SI unit of force is called the **newton** (abbreviated N), and 1 N is the force needed to accelerate an object with a mass of 1 kg at a rate of $1\ \text{m/s}^2$: $1\ \text{N} = 1\ \text{kg} \cdot \text{m/s}^2$. An easy way to remember the size of a newton is to imagine holding a small apple; it has a weight of about 1 N.

We can thus describe a two-dimensional force in the form $\vec{\mathbf{F}} = a\hat{\mathbf{i}} + b\hat{\mathbf{j}}$ (the unit vectors $\hat{\mathbf{i}}$ and $\hat{\mathbf{j}}$ indicate the direction of these forces along the *x*-axis and the *y*-axis, respectively) and a three-dimensional force in the form $\vec{\mathbf{F}} = a\hat{\mathbf{i}} + b\hat{\mathbf{j}} + c\hat{\mathbf{k}}$. In **Figure 8.3**, let's suppose that ice skater 1, on the left side of the figure, pushes horizontally with a force of 30.0 N to the right; we represent this as $\vec{\mathbf{F}}_1 = 30.0\hat{\mathbf{i}}\ \text{N}$. Similarly, if ice skater 2 pushes with a force of 40.0 N in the positive vertical direction shown, we would write $\vec{\mathbf{F}}_2 = 40.0\hat{\mathbf{j}}\ \text{N}$. The resultant of the two forces causes a mass to accelerate—in this case, the third ice skater. This resultant is called the **net external force** $\vec{\mathbf{F}}_{\text{net}}$ and is found by taking the vector sum of all external forces acting on an object or system (thus, we can also represent net external force as $\sum \vec{\mathbf{F}}$):

$$\vec{\mathbf{F}}_{\text{net}} = \sum \vec{\mathbf{F}} = \vec{\mathbf{F}}_1 + \vec{\mathbf{F}}_2 + \cdots \tag{8.1}$$

This equation can be extended to any number of forces.

In this example, we have $\vec{\mathbf{F}}_{\text{net}} = \sum \vec{\mathbf{F}} = \vec{\mathbf{F}}_1 + \vec{\mathbf{F}}_2 = 30.0\hat{\mathbf{i}} + 40.0\hat{\mathbf{j}}\ \text{N}$. The hypotenuse of the triangle shown in **Figure 8.3** is the resultant force, or net force. It is a vector. To find its magnitude (the size of the vector, without regard to direction), we use the rule given in **Components of a Vector**, taking the square root of the sum of the squares of the components:

$$F_{\text{net}} = \sqrt{(30.0\ \text{N})^2 + (40.0\ \text{N})^2} = 50.0\ \text{N}.$$

The direction is given by

$$\theta = \tan^{-1}\left(\frac{F_2}{F_1}\right) = \tan^{-1}\left(\frac{40.0}{30.0}\right) = 53.1°,$$

measured from the positive *x*-axis, as shown in the free-body diagram in **Figure 8.3**(b).

Let's suppose the ice skaters now push the third ice skater with $\vec{\mathbf{F}}_1 = 3.0\hat{\mathbf{i}} + 8.0\hat{\mathbf{j}}\ \text{N}$ and $\vec{\mathbf{F}}_2 = 5.0\hat{\mathbf{i}} + 4.0\hat{\mathbf{j}}\ \text{N}$. What is the resultant of these two forces? We must recognize that force is a vector; therefore, we must add using the rules for vector addition:

$$\vec{\mathbf{F}}_{\text{net}} = \vec{\mathbf{F}}_1 + \vec{\mathbf{F}}_2 = \left(3.0\hat{\mathbf{i}} + 8.0\hat{\mathbf{j}}\right) + \left(5.0\hat{\mathbf{i}} + 4.0\hat{\mathbf{j}}\right) = 8.0\hat{\mathbf{i}} + 12\hat{\mathbf{j}}\ \text{N}$$

8.1 **Check Your Understanding** Find the magnitude and direction of the net force in the ice skater example just given.

View this **interactive simulation (https://openstaxcollege.org/l/21addvectors)** to learn how to add vectors. Drag vectors onto a graph, change their length and angle, and sum them together. The magnitude, angle, and components of each vector can be displayed in several formats.

8.2 | Newton's First Law

Learning Objectives

By the end of the section, you will be able to:

- Describe Newton's first law of motion
- Recognize friction as an external force
- Define inertia
- Identify inertial reference frames
- Calculate equilibrium for a system

Experience suggests that an object at rest remains at rest if left alone and that an object in motion tends to slow down and stop unless some effort is made to keep it moving. However, **Newton's first law** gives a deeper explanation of this observation.

Newton's First Law of Motion

A body at rest remains at rest or, if in motion, remains in motion at constant velocity unless acted on by a net external force.

Note the repeated use of the verb "remains." We can think of this law as preserving the status quo of motion. Also note the expression "constant velocity;" this means that the object maintains a path along a straight line, since neither the magnitude nor the direction of the velocity vector changes. We can use **Figure 8.6** to consider the two parts of Newton's first law.

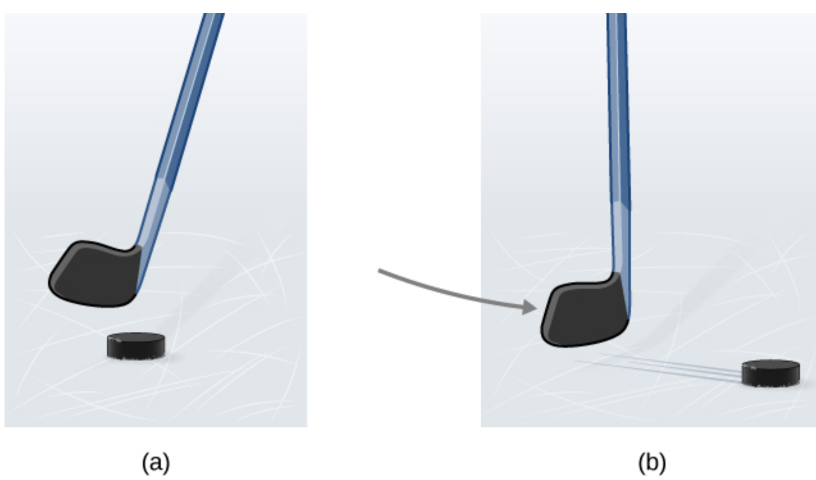


Figure 8.6 (a) A hockey puck is shown at rest; it remains at rest until an outside force such as a hockey stick changes its state of rest; (b) a hockey puck is shown in motion; it continues in motion in a straight line until an outside force causes it to change its state of motion. Although it is slick, an ice surface provides some friction that slows the puck.

Rather than contradicting our experience, Newton's first law says that there must be a cause for any change in velocity (a change in either magnitude or direction) to occur. This cause is a net external force, which we defined earlier in the chapter. An object sliding across a table or floor slows down due to the net force of friction acting on the object. If friction disappears, will the object still slow down?

The idea of cause and effect is crucial in accurately describing what happens in various situations. For example, consider what happens to an object sliding along a rough horizontal surface. The object quickly grinds to a halt. If we spray the surface with talcum powder to make the surface smoother, the object slides farther. If we make the surface even smoother by rubbing lubricating oil on it, the object slides farther yet. Extrapolating to a frictionless surface and ignoring air resistance, we can imagine the object sliding in a straight line indefinitely. Friction is thus the cause of slowing (consistent with Newton’s first law). The object would not slow down if friction were eliminated.

Consider an air hockey table (**Figure 8.7**). When the air is turned off, the puck slides only a short distance before friction slows it to a stop. However, when the air is turned on, it creates a nearly frictionless surface, and the puck glides long distances without slowing down. Additionally, if we know enough about the friction, we can accurately predict how quickly the object slows down.

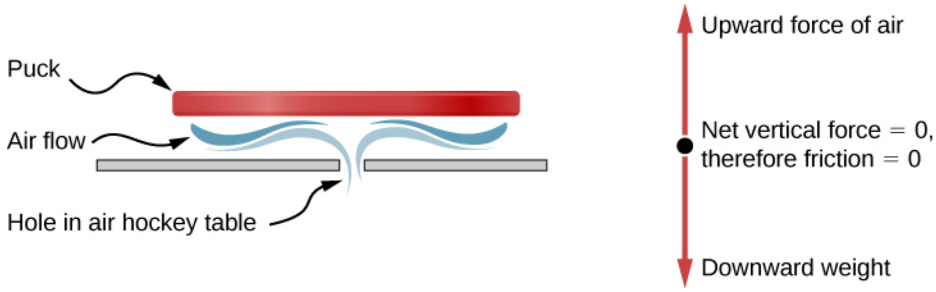


Figure 8.7 An air hockey table is useful in illustrating Newton’s laws. When the air is off, friction quickly slows the puck; but when the air is on, it minimizes contact between the puck and the hockey table, and the puck glides far down the table.

Newton’s first law is general and can be applied to anything from an object sliding on a table to a satellite in orbit to blood pumped from the heart. Experiments have verified that any change in velocity (speed or direction) must be caused by an external force. The idea of *generally applicable or universal laws* is important—it is a basic feature of all laws of physics. Identifying these laws is like recognizing patterns in nature from which further patterns can be discovered. The genius of Galileo, who first developed the idea for the first law of motion, and Newton, who clarified it, was to ask the fundamental question: “What is the cause?” Thinking in terms of cause and effect is fundamentally different from the typical ancient Greek approach, when questions such as “Why does a tiger have stripes?” would have been answered in Aristotelian fashion, such as “That is the nature of the beast.” The ability to think in terms of cause and effect is the ability to make a connection between an observed behavior and the surrounding world.

Gravitation and Inertia

Regardless of the scale of an object, whether a molecule or a subatomic particle, two properties remain valid and thus of interest to physics: gravitation and inertia. Both are connected to mass. Roughly speaking, *mass* is a measure of the amount of matter in something. *Gravitation* is the attraction of one mass to another, such as the attraction between yourself and Earth that holds your feet to the floor. The magnitude of this attraction is your weight, and it is a force.

Mass is also related to **inertia**, the ability of an object to resist changes in its motion—in other words, to resist acceleration. Newton’s first law is often called the **law of inertia**. As we know from experience, some objects have more inertia than others. It is more difficult to change the motion of a large boulder than that of a basketball, for example, because the boulder has more mass than the basketball. In other words, the inertia of an object is measured by its mass. The relationship between mass and weight is explored later in this chapter.

Inertial Reference Frames

In **Section 6.5** we mentioned that the measurements that are made of the kinematic quantities (position, speed, etc.) for some moving body depend upon the reference frame of the observer who makes those measurements. Newton's First Law gives us the basis to properly define one special type of reference frame - an **inertial reference frame**.

Earlier, we stated Newton’s first law as “A body at rest remains at rest or, if in motion, remains in motion at constant velocity unless acted on by a net external force.” It can also be stated as “Every body remains in its state of uniform motion in a straight line unless it is compelled to change that state by forces acting on it.” To Newton, “uniform motion in a straight line” meant constant velocity, which includes the case of zero velocity, or rest. Therefore, the first law says that the velocity of an object remains constant if the net force on it is zero.

In principle, we can make the net force on a body zero. If its velocity relative to a given frame is constant, then that frame is said to be inertial. So by definition, an inertial reference frame is a reference frame in which Newton’s first law is valid.

Newton’s first law applies to objects with constant velocity. From this fact, we can infer the following statement.

Inertial Reference Frame

A reference frame moving at constant velocity relative to an inertial frame is also inertial. A reference frame accelerating relative to an inertial frame is not inertial.

Are inertial frames common in nature? It turns out that well within experimental error, a reference frame at rest relative to the most distant, or “fixed,” stars is inertial. All frames moving uniformly with respect to this fixed-star frame are also inertial. For example, a nonrotating reference frame attached to the Sun is, for all practical purposes, inertial, because its velocity relative to the fixed stars does not vary by more than one part in 10^{10}. Earth accelerates relative to the fixed stars because it rotates on its axis and revolves around the Sun; hence, a reference frame attached to its surface is not inertial. For most problems, however, such a frame serves as a sufficiently accurate approximation to an inertial frame, because the acceleration of a point on Earth’s surface relative to the fixed stars is rather small ($< 3.4 \times 10^{-2}\ \text{m/s}^2$). Thus, unless indicated otherwise, we consider reference frames fixed on Earth to be inertial.

Finally, no particular inertial frame is more special than any other. As far as the laws of nature are concerned, all inertial frames are equivalent. In analyzing a problem, we choose one inertial frame over another simply on the basis of convenience.

Newton’s First Law and Equilibrium

Newton’s first law tells us about the equilibrium of a system, which is the state in which the forces on the system are balanced. Returning to **Forces** and the ice skaters in **Figure 8.3**, we know that the forces $\vec{\mathbf{F}}_1$ and $\vec{\mathbf{F}}_2$ combine to form a resultant force, or the net external force: $\vec{\mathbf{F}}_R = \vec{\mathbf{F}}_{net} = \vec{\mathbf{F}}_1 + \vec{\mathbf{F}}_2$. To create equilibrium, we require a balancing force that will produce a net force of zero. This force must be equal in magnitude but opposite in direction to $\vec{\mathbf{F}}_R$, which means the vector must be $-\vec{\mathbf{F}}_R$. Referring to the ice skaters, for which we found $\vec{\mathbf{F}}_R$ to be $30.0\,\hat{\mathbf{i}} + 40.0\,\hat{\mathbf{j}}\ \text{N}$, we can determine the balancing force by simply finding $-\vec{\mathbf{F}}_R = -30.0\,\hat{\mathbf{i}} - 40.0\,\hat{\mathbf{j}}\ \text{N}$. See the free-body diagram in **Figure 8.3**(b).

We can give Newton’s first law in vector form:

$$\vec{\mathbf{v}} = \text{constant when } \vec{\mathbf{F}}_{net} = \vec{\mathbf{0}}\ \text{N}. \tag{8.2}$$

This equation says that a net force of zero implies that the velocity $\vec{\mathbf{v}}$ of the object is constant. (The word “constant” can indicate zero velocity.)

Newton’s first law is deceptively simple. If a car is at rest, the only forces acting on the car are weight and the contact force of the pavement pushing up on the car (**Figure 8.8**). It is easy to understand that a nonzero net force is required to change the state of motion of the car. However, if the car is in motion with constant velocity, a common misconception is that the engine force propelling the car forward is larger in magnitude than the friction force that opposes forward motion. In fact, the two forces have identical magnitude.

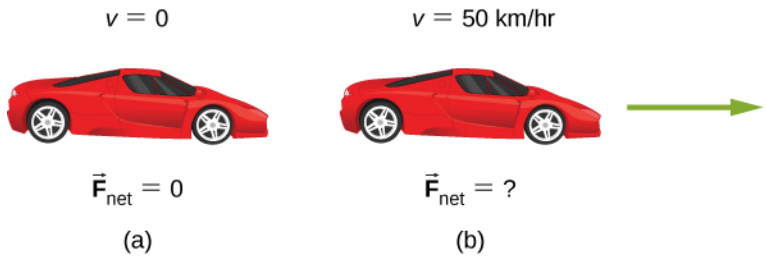


Figure 8.8 A car is shown (a) parked and (b) moving at constant velocity. How do Newton's laws apply to the parked car? What does the knowledge that the car is moving at constant velocity tell us about the net horizontal force on the car?

Example 8.1

When Does Newton's First Law Apply to Your Car?

Newton's laws can be applied to all physical processes involving force and motion, including something as mundane as driving a car.

(a) Your car is parked outside your house. Does Newton's first law apply in this situation? Why or why not?

(b) Your car moves at constant velocity down the street. Does Newton's first law apply in this situation? Why or why not?

Strategy

In (a), we are considering the first part of Newton's first law, dealing with a body at rest; in (b), we look at the second part of Newton's first law for a body in motion.

Solution

a. When your car is parked, all forces on the car must be balanced; the vector sum is 0 N. Thus, the net force is zero, and Newton's first law applies. The acceleration of the car is zero, and in this case, the velocity is also zero.

b. When your car is moving at constant velocity down the street, the net force must also be zero according to Newton's first law. The car's engine produces a forward force; friction, a force between the road and the tires of the car that opposes forward motion, has exactly the same magnitude as the engine force, producing the net force of zero. The body continues in its state of constant velocity until the net force becomes nonzero. Realize that *a net force of zero means that an object is either at rest or moving with constant velocity, that is, it is not accelerating.* What do you suppose happens when the car accelerates? We explore this idea in the next section.

Significance

As this example shows, there are two kinds of equilibrium. In (a), the car is at rest; we say it is in *static equilibrium*. In (b), the forces on the car are balanced, but the car is moving; we say that it is in *dynamic equilibrium*. Again, it is possible for two (or more) forces to act on an object and yet for the object to not move. In addition, a net force of zero cannot produce acceleration.

8.2 Check Your Understanding A skydiver opens his parachute, and shortly thereafter, he is moving at constant velocity. (a) What forces are acting on him? (b) Which force is bigger?

Engage this **simulation (https://openstaxcollege.org/l/21forcemotion)** to predict, qualitatively, how an external force will affect the speed and direction of an object's motion. Explain the effects with the help of a free-body diagram. Use free-body diagrams to draw position, velocity, acceleration, and force graphs, and vice versa. Explain how the graphs relate to one another. Given a scenario or a graph, sketch all four graphs.

8.3 | Newton's Second Law

Learning Objectives

By the end of the section, you will be able to:

- Distinguish between external and internal forces
- Describe Newton's second law of motion
- Explain the dependence of acceleration on net force and mass

Newton’s second law is closely related to his first law. It mathematically gives the cause-and-effect relationship between force and changes in motion. Newton’s second law is quantitative and is used extensively to calculate what happens in situations involving a force. Before we can write down Newton’s second law as a simple equation that gives the exact relationship of force, mass, and acceleration, we need to sharpen some ideas we mentioned earlier.

Force and Acceleration

First, what do we mean by a change in motion? The answer is that a change in motion is equivalent to a change in velocity. A change in velocity means, by definition, that there is acceleration. Newton’s first law says that a net external force causes a change in motion; thus, we see that a *net external force causes nonzero acceleration*.

We defined external force in **Forces** as force acting on an object or system that originates outside of the object or system. Let’s consider this concept further. An intuitive notion of *external* is correct—it is outside the system of interest. For example, in **Figure 8.9**(a), the system of interest is the car plus the person within it. The two forces exerted by the two students are external forces. In contrast, an internal force acts between elements of the system. Thus, the force the person in the car exerts to hang on to the steering wheel is an internal force between elements of the system of interest. Only external forces affect the motion of a system, according to Newton’s first law. (The internal forces cancel each other out, as explained in the next section.) Therefore, we must define the boundaries of the system before we can determine which forces are external. Sometimes, the system is obvious, whereas at other times, identifying the boundaries of a system is more subtle. The concept of a system is fundamental to many areas of physics, as is the correct application of Newton’s laws. This concept is revisited many times in the study of physics.

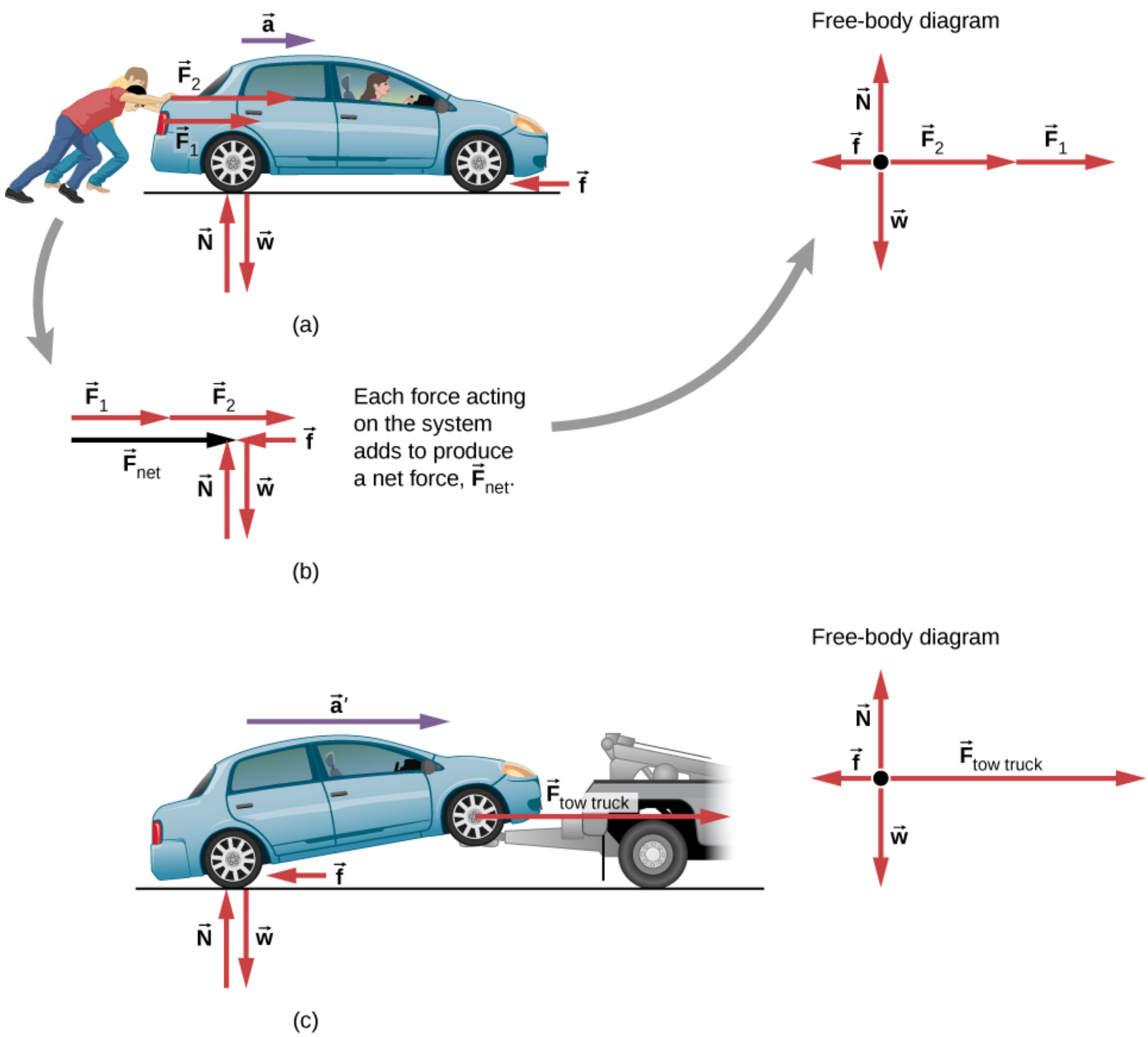


Figure 8.9 Different forces exerted on the same mass produce different accelerations. (a) Two students push a stalled car. All external forces acting on the car are shown. (b) The forces acting on the car are transferred to a coordinate plane (free-body diagram) for simpler analysis. (c) The tow truck can produce greater external force on the same mass, and thus greater acceleration.

From this example, you can see that different forces exerted on the same mass produce different accelerations. In **Figure 8.9**(a), the two students push a car with a driver in it. Arrows representing all external forces are shown. The system of interest is the car and its driver. The weight $\vec{w}$ of the system and the support of the ground $\vec{N}$ are also shown for completeness and are assumed to cancel (because there was no vertical motion and no imbalance of forces in the vertical direction to create a change in motion). The vector $\vec{f}$ represents the friction acting on the car, and it acts to the left, opposing the motion of the car. (We discuss friction in more detail in the next chapter.) In **Figure 8.9**(b), all external forces acting on the system add together to produce the net force $\vec{F}_{net}$. The free-body diagram shows all of the forces acting on the system of interest. The dot represents the center of mass of the system. Each force vector extends from this dot. Because there are two forces acting to the right, the vectors are shown collinearly. Finally, in **Figure 8.9**(c), a larger net external force produces a larger acceleration ($\vec{a}' > \vec{a}$) when the tow truck pulls the car.

It seems reasonable that acceleration would be directly proportional to and in the same direction as the net external force acting on a system. This assumption has been verified experimentally and is illustrated in **Figure 8.9**. To obtain an equation for Newton's second law, we first write the relationship of acceleration $\vec{a}$ and net external force $\vec{F}_{net}$ as the proportionality

$$\vec{a} \propto \vec{F}_{net}$$

where the symbol $\propto$ means “proportional to.” (Recall from **Forces** that the net external force is the vector sum of all external forces and is sometimes indicated as $\sum \vec{F}$.) This proportionality shows what we have said in words—acceleration is directly proportional to net external force. Once the system of interest is chosen, identify the external forces and ignore the internal ones. It is a tremendous simplification to disregard the numerous internal forces acting between objects within the system, such as muscular forces within the students’ bodies, let alone the myriad forces between the atoms in the objects. Still, this simplification helps us solve some complex problems.

It also seems reasonable that acceleration should be inversely proportional to the mass of the system. In other words, the larger the mass (the inertia), the smaller the acceleration produced by a given force. As illustrated in **Figure 8.10**, the same net external force applied to a basketball produces a much smaller acceleration when it is applied to an SUV. The proportionality is written as

$$a \propto \frac{1}{m},$$

where m is the mass of the system and a is the magnitude of the acceleration. Experiments have shown that acceleration is exactly inversely proportional to mass, just as it is directly proportional to net external force.

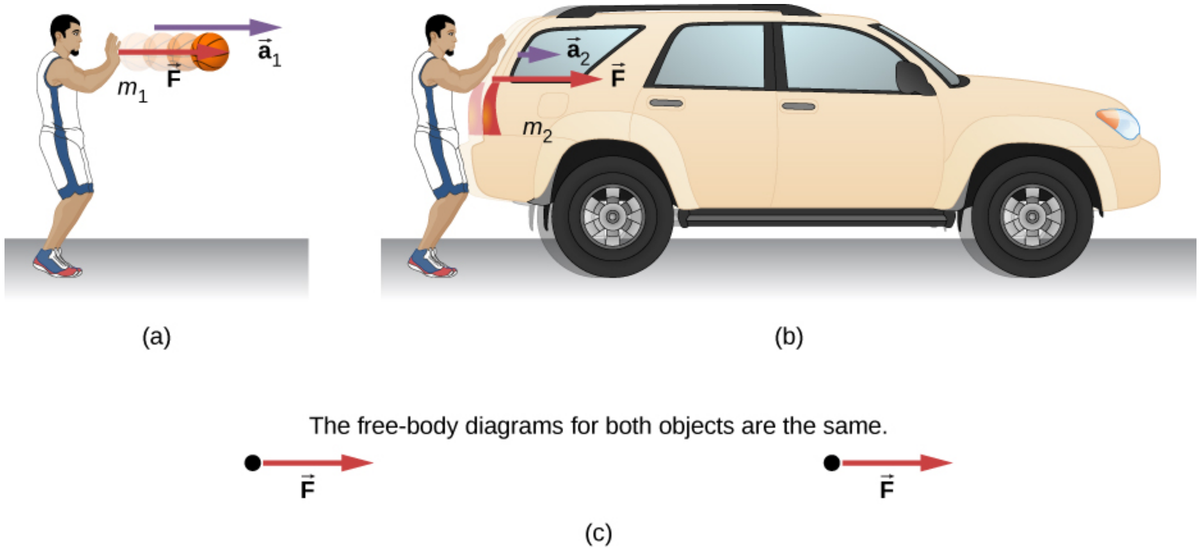


Figure 8.10 The same force exerted on systems of different masses produces different accelerations. (a) A basketball player pushes on a basketball to make a pass. (Ignore the effect of gravity on the ball.) (b) The same player exerts an identical force on a stalled SUV and produces far less acceleration. (c) The free-body diagrams are identical, permitting direct comparison of the two situations. A series of patterns for free-body diagrams will emerge as you do more problems and learn how to draw them in **Drawing Free-Body Diagrams**.

It has been found that the acceleration of an object depends only on the net external force and the mass of the object. Combining the two proportionalities just given yields **Newton’s second law**.

Newton’s Second Law of Motion

The acceleration of a system is directly proportional to and in the same direction as the net external force acting on the system and is inversely proportion to its mass. In equation form, Newton’s second law is

$$\vec{a} = \frac{\vec{F}_{net}}{m},$$

where $\vec{a}$ is the acceleration, $\vec{F}_{net}$ is the net force, and m is the mass. This is often written in the more familiar form

$$\vec{\mathbf{F}}_{\text{net}} = \sum \vec{\mathbf{F}} = m\vec{\mathbf{a}}, \tag{8.3}$$

but the first equation gives more insight into what Newton's second law means. When only the magnitude of force and acceleration are considered, this equation can be written in the simpler scalar form:

$$F_{\text{net}} = ma. \tag{8.4}$$

The law is a cause-and-effect relationship among three quantities that is not simply based on their definitions. The validity of the second law is based on experimental verification. The free-body diagram, which you will learn to draw in **Drawing Free-Body Diagrams**, is the basis for writing Newton's second law.

Example 8.2

What Acceleration Can a Person Produce When Pushing a Lawn Mower?

Suppose that the net external force (push minus friction) exerted on a lawn mower is 51 N (about 11 lb.) parallel to the ground (**Figure 8.11**). The mass of the mower is 24 kg. What is its acceleration?

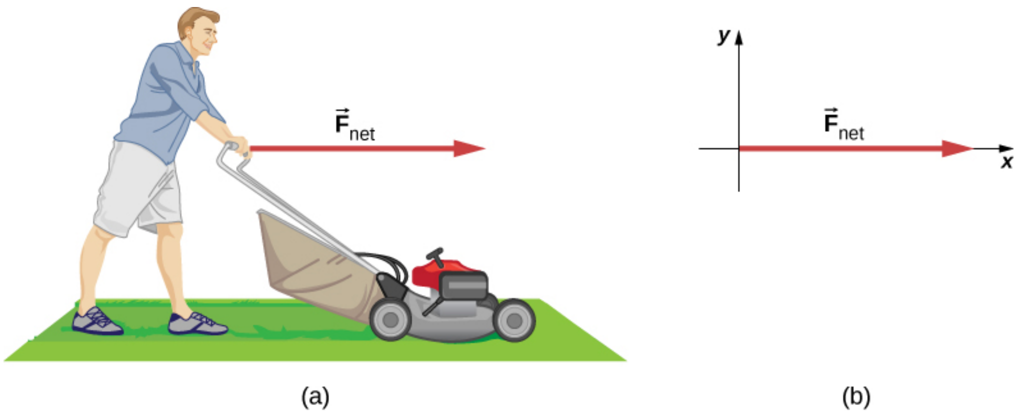


Figure 8.11 (a) The net force on a lawn mower is 51 N to the right. At what rate does the lawn mower accelerate to the right? (b) The free-body diagram for this problem is shown.

Strategy

This problem involves only motion in the horizontal direction; we are also given the net force, indicated by the single vector, but we can suppress the vector nature and concentrate on applying Newton's second law. Since F_{net} and m are given, the acceleration can be calculated directly from Newton's second law as $F_{\text{net}} = ma$.

Solution

The magnitude of the acceleration a is $a = F_{\text{net}}/m$. Entering known values gives

$$a = \frac{51\ \text{N}}{24\ \text{kg}}.$$

Substituting the unit of kilograms times meters per square second for newtons yields

$$a = \frac{51\ \text{kg}\cdot\text{m/s}^2}{24\ \text{kg}} = 2.1\ \text{m/s}^2.$$

Significance

The direction of the acceleration is the same direction as that of the net force, which is parallel to the ground. This is a result of the vector relationship expressed in Newton's second law, that is, the vector representing net force is the scalar multiple of the acceleration vector. There is no information given in this example about the individual external forces acting on the system, but we can say something about their relative magnitudes. For example, the force exerted by the person pushing the mower must be greater than the friction opposing the motion (since we know the mower moved forward), and the vertical forces must cancel because no acceleration occurs in the vertical direction (the mower is moving only horizontally). The acceleration found is small enough to be reasonable for a person pushing a mower. Such an effort would not last too long, because the person's top speed would soon be reached.

8.3 Check Your Understanding At the time of its launch, the HMS *Titanic* was the most massive mobile object ever built, with a mass of 6.0×10^7 kg . If a force of 6 MN $(6 \times 10^6 \text{ N})$ was applied to the ship, what acceleration would it experience?

In the preceding example, we dealt with net force only for simplicity. However, several forces act on the lawn mower. The weight $\vec{\mathbf{w}}$ (discussed in detail in **Mass and Weight**) pulls down on the mower, toward the center of Earth; this produces a contact force on the ground. The ground must exert an upward force on the lawn mower, known as the normal force $\vec{\mathbf{N}}$, which we define in **Common Forces**. These forces are balanced and therefore do not produce vertical acceleration. In the next example, we show both of these forces. As you continue to solve problems using Newton's second law, be sure to show multiple forces.

Example 8.3

Which Force Is Bigger?

(a) The car shown in **Figure 8.12** is moving at a constant speed. Which force is bigger, $\vec{\mathbf{F}}_{\text{engine}}$ or $\vec{\mathbf{F}}_{\text{friction}}$? Explain.

(b) The same car is now accelerating to the right. Which force is bigger, $\vec{\mathbf{F}}_{\text{engine}}$ or $\vec{\mathbf{F}}_{\text{friction}}$? Explain.

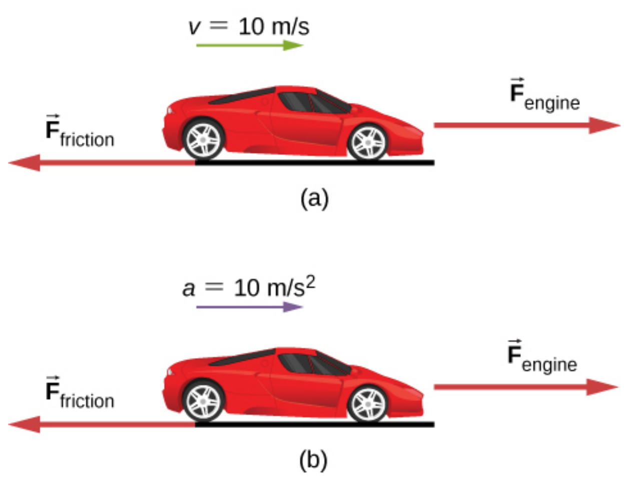


Figure 8.12 A car is shown (a) moving at constant speed and (b) accelerating. How do the forces acting on the car compare in each case? (a) What does the knowledge that the car is moving at constant velocity tell us about the net horizontal force on the car compared to the friction force? (b) What does the knowledge that the car is accelerating tell us about the horizontal force on the car compared to the friction force?

Strategy

We must consider Newton's first and second laws to analyze the situation. We need to decide which law applies; this, in turn, will tell us about the relationship between the forces.

Solution

a. The forces are equal. According to Newton's first law, if the net force is zero, the velocity is constant.

b. In this case, $\vec{\mathbf{F}}_{\text{engine}}$ must be larger than $\vec{\mathbf{F}}_{\text{friction}}$. According to Newton's second law, a net force is required to cause acceleration.

Significance

These questions may seem trivial, but they are commonly answered incorrectly. For a car or any other object to move, it must be accelerated from rest to the desired speed; this requires that the engine force be greater than the friction force. Once the car is moving at constant velocity, the net force must be zero; otherwise, the car will accelerate (gain speed). To solve problems involving Newton's laws, we must understand whether to apply Newton's first law (where $\sum \vec{\mathbf{F}} = \vec{\mathbf{0}}$) or Newton's second law (where $\sum \vec{\mathbf{F}}$ is not zero). This will be apparent as you see more examples and attempt to solve problems on your own.

Example 8.4

What Rocket Thrust Accelerates This Sled?

Before manned space flights, rocket sleds were used to test aircraft, missile equipment, and physiological effects on human subjects at high speeds. They consisted of a platform that was mounted on one or two rails and propelled by several rockets.

Calculate the magnitude of force exerted by each rocket, called its thrust *T*, for the four-rocket propulsion system shown in **Figure 8.13**. The sled's initial acceleration is 49 m/s^2, the mass of the system is 2100 kg, and the force of friction opposing the motion is 650 N.

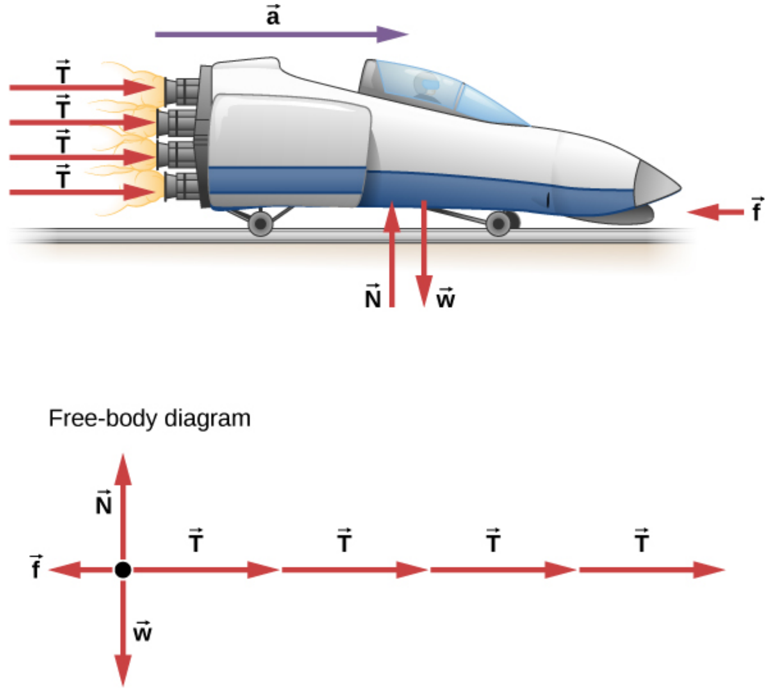


Figure 8.13 A sled experiences a rocket thrust that accelerates it to the right. Each rocket creates an identical thrust *T*. The system here is the sled, its rockets, and its rider, so none of the forces between these objects are considered. The arrow representing friction ($\vec{\mathbf{f}}$) is drawn larger than scale.

Strategy

Although forces are acting both vertically and horizontally, we assume the vertical forces cancel because there is no vertical acceleration. This leaves us with only horizontal forces and a simpler one-dimensional problem. Directions are indicated with plus or minus signs, with right taken as the positive direction. See the free-body diagram in **Figure 8.13**.

Solution

Since acceleration, mass, and the force of friction are given, we start with Newton's second law and look for ways to find the thrust of the engines. We have defined the direction of the force and acceleration as acting "to the right," so we need to consider only the magnitudes of these quantities in the calculations. Hence we begin with

$$F_{\text{net}} = ma$$

where F_{net} is the net force along the horizontal direction. We can see from the figure that the engine thrusts add, whereas friction opposes the thrust. In equation form, the net external force is

$$F_{\text{net}} = 4T - f.$$

Substituting this into Newton's second law gives us

$$F_{\text{net}} = ma = 4T - f.$$

Using a little algebra, we solve for the total thrust 4*T*:

$$4T = ma + f.$$

Substituting known values yields

$$4T = ma + f = (2100\ \text{kg})(49\ \text{m/s}^2) + 650\ \text{N}.$$

Therefore, the total thrust is

$$4T = 1.0 \times 10^5\ \text{N},$$

and the individual thrusts are

$$T = \frac{1.0 \times 10^5\ \text{N}}{4} = 2.5 \times 10^4\ \text{N}.$$

Significance

The numbers are quite large, so the result might surprise you. Experiments such as this were performed in the early 1960s to test the limits of human endurance, and the setup was designed to protect human subjects in jet fighter emergency ejections. Speeds of 1000 km/h were obtained, with accelerations of 45 *g*'s. (Recall that *g*, acceleration due to gravity, is $9.80\ \text{m/s}^2$. When we say that acceleration is 45 *g*'s, it is $45 \times 9.8\ \text{m/s}^2$, which is approximately $440\ \text{m/s}^2$.) Although living subjects are not used anymore, land speeds of 10,000 km/h have been obtained with a rocket sled.

In this example, as in the preceding one, the system of interest is obvious. We see in later examples that choosing the system of interest is crucial—and the choice is not always obvious.

Newton's second law is more than a definition; it is a relationship among acceleration, force, and mass. It can help us make predictions. Each of those physical quantities can be defined independently, so the second law tells us something basic and universal about nature.

8.4 Check Your Understanding A 550-kg sports car collides with a 2200-kg truck, and during the collision, the net force on each vehicle is the force exerted by the other. If the magnitude of the truck's acceleration is $10\ \text{m/s}^2$, what is the magnitude of the sports car's acceleration?

8.4 | Mass and Weight

Learning Objectives

By the end of the section, you will be able to:

- Explain the difference between mass and weight
- Explain why falling objects on Earth are never truly in free fall
- Describe the concept of weightlessness

Mass and weight are often used interchangeably in everyday conversation. For example, our medical records often show our weight in kilograms but never in the correct units of newtons. In physics, however, there is an important distinction. Weight is the pull of Earth on an object. It depends on the distance from the center of Earth. Unlike weight, mass does not vary with location. The mass of an object is the same on Earth, in orbit, or on the surface of the Moon.

Units of Force

The equation $F_{\text{net}} = ma$ is used to define net force in terms of mass, length, and time. As explained earlier, the SI unit of force is the newton. Since $F_{\text{net}} = ma$,

$$1\text{ N} = 1\text{ kg}\cdot\text{m/s}^2.$$

Although almost the entire world uses the newton for the unit of force, in the United States, the most familiar unit of force is the pound (lb), where 1 N = 0.225 lb. Thus, a 225-lb person weighs 1000 N.

Weight and Gravitational Force

When an object is dropped, it accelerates toward the center of Earth. Newton's second law says that a net force on an object is responsible for its acceleration. If air resistance is negligible, the net force on a falling object is the gravitational force, commonly called its **weight** $\vec{\mathbf{w}}$, or its force due to gravity acting on an object of mass *m*. Weight can be denoted as a vector because it has a direction; *down* is, by definition, the direction of gravity, and hence, weight is a downward force. The magnitude of weight is denoted as *w*. Galileo was instrumental in showing that, in the absence of air resistance, all objects fall with the same acceleration *g*. Using Galileo's result and Newton's second law, we can derive an equation for weight.

Consider an object with mass *m* falling toward Earth. It experiences only the downward force of gravity, which is the weight $\vec{\mathbf{w}}$. Newton's second law says that the magnitude of the net external force on an object is $\vec{\mathbf{F}}_{\text{net}} = m\vec{\mathbf{a}}$. We know that the acceleration of an object due to gravity is $\vec{\mathbf{g}}$, or $\vec{\mathbf{a}} = \vec{\mathbf{g}}$. Substituting these into Newton's second law gives us the following equations.

Weight

The gravitational force on a mass is its weight. We can write this in vector form, where $\vec{\mathbf{w}}$ is weight and *m* is mass, as

$$\vec{\mathbf{w}} = m\vec{\mathbf{g}}. \tag{8.5}$$

In scalar form, we can write

$$w = mg. \tag{8.6}$$

Since $g = 9.80\text{ m/s}^2$ on Earth, the weight of a 1.00-kg object on Earth is 9.80 N:

$$w = mg = (1.00\text{ kg})(9.80\text{ m/s}^2) = 9.80\text{ N}.$$

When the net external force on an object is its weight, we say that it is in **free fall**, that is, the only force acting on the object is gravity. However, when objects on Earth fall downward, they are never truly in free fall because there is always some upward resistance force from the air acting on the object.

Acceleration due to gravity g varies slightly over the surface of Earth, so the weight of an object depends on its location and is not an intrinsic property of the object. Weight varies dramatically if we leave Earth's surface. On the Moon, for example, acceleration due to gravity is only 1.67 m/s^2. A 1.0-kg mass thus has a weight of 9.8 N on Earth and only about 1.7 N on the Moon.

The broadest definition of weight in this sense is that the weight of an object is the gravitational force on it from the nearest large body, such as Earth, the Moon, or the Sun. This is the most common and useful definition of weight in physics. It differs dramatically, however, from the definition of weight used by NASA and the popular media in relation to space travel and exploration. When they speak of "weightlessness" and "microgravity," they are referring to the phenomenon we call "free fall" in physics. We use the preceding definition of weight, force $\vec{\mathbf{w}}$ due to gravity acting on an object of mass m, and we make careful distinctions between free fall and actual weightlessness.

Be aware that weight and mass are different physical quantities, although they are closely related. Mass is an intrinsic property of an object: It is a quantity of matter. The quantity or amount of matter of an object is determined by the numbers of atoms and molecules of various types it contains. Because these numbers do not vary, in Newtonian physics, mass does not vary; therefore, its response to an applied force does not vary. In contrast, weight is the gravitational force acting on an object, so it does vary depending on gravity. For example, a person closer to the center of Earth, at a low elevation such as New Orleans, weighs slightly more than a person who is located in the higher elevation of Denver, even though they may have the same mass.

It is tempting to equate mass to weight, because most of our examples take place on Earth, where the weight of an object varies only a little with the location of the object. In addition, it is difficult to count and identify all of the atoms and molecules in an object, so mass is rarely determined in this manner. If we consider situations in which $\vec{\mathbf{g}}$ is a constant on Earth, we see that weight $\vec{\mathbf{w}}$ is directly proportional to mass m, since $\vec{\mathbf{w}} = m\vec{\mathbf{g}}$, that is, the more massive an object is, the more it weighs. Operationally, the masses of objects are determined by comparison with the standard kilogram, as we discussed in **Section 1.1**. But by comparing an object on Earth with one on the Moon, we can easily see a variation in weight but not in mass. For instance, on Earth, a 5.0-kg object weighs 49 N; on the Moon, where g is 1.67 m/s^2, the object weighs 8.4 N. However, the mass of the object is still 5.0 kg on the Moon.

Example 8.5

Clearing a Field

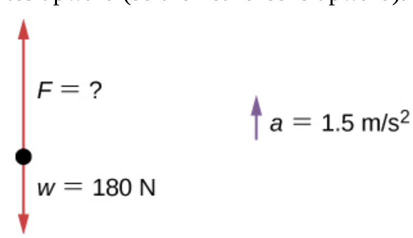
A farmer is lifting some moderately heavy rocks from a field to plant crops. He lifts a stone that weighs 40.0 lb. (about 180 N). What force does he apply if the stone accelerates at a rate of 1.5 m/s^2?

Strategy

We were given the weight of the stone, which we use in finding the net force on the stone. However, we also need to know its mass to apply Newton's second law, so we must apply the equation for weight, $w = mg$, to determine the mass.

Solution

No forces act in the horizontal direction, so we can concentrate on vertical forces, as shown in the following free-body diagram. We label the acceleration to the side; technically, it is not part of the free-body diagram, but it helps to remind us that the object accelerates upward (so the net force is upward).



$$\begin{aligned} w &= mg \\ m &= \frac{w}{g} = \frac{180\text{ N}}{9.8\text{ m/s}^2} = 18\text{ kg} \\ \sum F &= ma \\ F - w &= ma \\ F - 180\text{ N} &= (18\text{ kg})(1.5\text{ m/s}^2) \\ F - 180\text{ N} &= 27\text{ N} \\ F &= 207\text{ N} = 210\text{ N to two significant figures} \end{aligned}$$

Significance

To apply Newton's second law as the primary equation in solving a problem, we sometimes have to rely on other equations, such as the one for weight or one of the kinematic equations, to complete the solution.

8.5 Check Your Understanding For **Example 8.5**, find the acceleration when the farmer's applied force is 230.0 N.

Can you avoid the boulder field and land safely just before your fuel runs out, as Neil Armstrong did in 1969? This **version of the classic video game (https://openstaxcollege.org/l/21lunarlander)** accurately simulates the real motion of the lunar lander, with the correct mass, thrust, fuel consumption rate, and lunar gravity. The real lunar lander is hard to control.

Use this **interactive simulation (https://openstaxcollege.org/l/21gravityorbits)** to move the Sun, Earth, Moon, and space station to see the effects on their gravitational forces and orbital paths. Visualize the sizes and distances between different heavenly bodies, and turn off gravity to see what would happen without it.

8.5 | Newton's Third Law

Learning Objectives

By the end of the section, you will be able to:

- State Newton's third law of motion
- Identify the action and reaction forces in different situations
- Apply Newton's third law to define systems and solve problems of motion

We have thus far considered force as a push or a pull; however, if you think about it, you realize that no push or pull ever occurs by itself. When you push on a wall, the wall pushes back on you. This brings us to **Newton's third law**.

Newton's Third Law of Motion

Whenever one body exerts a force on a second body, the first body experiences a force that is equal in magnitude and opposite in direction to the force that it exerts. Mathematically, if a body *A* exerts a force $\vec{\mathbf{F}}$ on body *B*, then *B* simultaneously exerts a force $-\vec{\mathbf{F}}$ on *A*, or in vector equation form,

$$\vec{\mathbf{F}}_{AB} = -\vec{\mathbf{F}}_{BA}. \tag{8.7}$$

Newton's third law represents a certain symmetry in nature: Forces always occur in pairs, and one body cannot exert a force on another without experiencing a force itself. We sometimes refer to this law loosely as "action-reaction," where the force exerted is the action and the force experienced as a consequence is the reaction. Newton's third law has practical uses in analyzing the origin of forces and understanding which forces are external to a system.

We can readily see Newton's third law at work by taking a look at how people move about. Consider a swimmer pushing off the side of a pool (**Figure 8.14**). She pushes against the wall of the pool with her feet and accelerates in the direction opposite that of her push. The wall has exerted an equal and opposite force on the swimmer. You might think that two equal and opposite forces would cancel, but they do not *because they act on different systems*. In this case, there are two systems that we could investigate: the swimmer and the wall. If we select the swimmer to be the system of interest, as in the figure, then $F_{\text{wall on feet}}$ is an external force on this system and affects its motion. The swimmer moves in the direction of this force. In contrast, the force $F_{\text{feet on wall}}$ acts on the wall, not on our system of interest. Thus, $F_{\text{feet on wall}}$ does not directly affect the motion of the system and does not cancel $F_{\text{wall on feet}}.$ The swimmer pushes in the direction opposite that in which she wishes to move. The reaction to her push is thus in the desired direction. In a free-body diagram, such as the one shown in **Figure 8.14**, we never include both forces of an action-reaction pair; in this case, we only use $F_{\text{wall on feet}}$, not $F_{\text{feet on wall}}$.

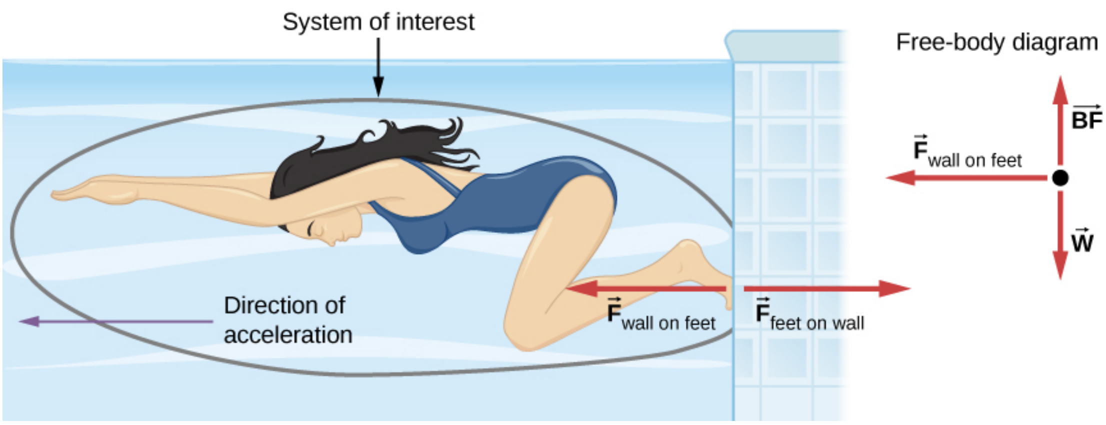


Figure 8.14 When the swimmer exerts a force on the wall, she accelerates in the opposite direction; in other words, the net external force on her is in the direction opposite of $F_{\text{feet on wall}}.$ This opposition occurs because, in accordance with Newton's third law, the wall exerts a force $F_{\text{wall on feet}}$ on the swimmer that is equal in magnitude but in the direction opposite to the one she exerts on it. The line around the swimmer indicates the system of interest. Thus, the free-body diagram shows only $F_{\text{wall on feet}}$, w (the gravitational force), and *BF*, which is the buoyant force of the water supporting the swimmer's weight. The vertical forces *w* and *BF* cancel because there is no vertical acceleration.

Other examples of Newton's third law are easy to find:

- As a professor paces in front of a whiteboard, he exerts a force backward on the floor. The floor exerts a reaction force forward on the professor that causes him to accelerate forward.
- A car accelerates forward because the ground pushes forward on the drive wheels, in reaction to the drive wheels pushing backward on the ground. You can see evidence of the wheels pushing backward when tires spin on a gravel road and throw the rocks backward.
- Rockets move forward by expelling gas backward at high velocity. This means the rocket exerts a large backward force on the gas in the rocket combustion chamber; therefore, the gas exerts a large reaction force forward on the rocket. This reaction force, which pushes a body forward in response to a backward force, is called **thrust**. It is a common misconception that rockets propel themselves by pushing on the ground or on the air behind them. They actually work better in a vacuum, where they can more readily expel the exhaust gases.
- Helicopters create lift by pushing air down, thereby experiencing an upward reaction force.
- Birds and airplanes also fly by exerting force on the air in a direction opposite that of whatever force they need. For example, the wings of a bird force air downward and backward to get lift and move forward.
- An octopus propels itself in the water by ejecting water through a funnel from its body, similar to a jet ski.
- When a person pulls down on a vertical rope, the rope pulls up on the person (**Figure 8.15**).

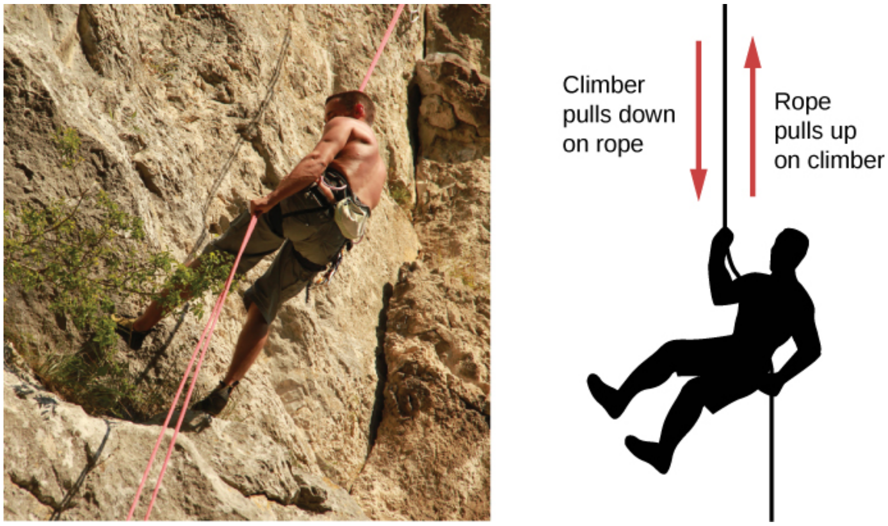


Figure 8.15 When the mountain climber pulls down on the rope, the rope pulls up on the mountain climber.

There are two important features of Newton's third law. First, the forces exerted (the action and reaction) are always equal in magnitude but opposite in direction. Second, these forces are acting on different bodies or systems: *A*'s force acts on *B* and *B*'s force acts on *A*. In other words, the two forces are distinct forces that do not act on the same body. Thus, they do not cancel each other.

For the situation shown in **Figure 8.5**, the third law indicates that because the chair is pushing upward on the boy with force $\vec{\mathbf{C}}$, he is pushing downward on the chair with force $-\vec{\mathbf{C}}$. Similarly, he is pushing downward with forces $-\vec{\mathbf{F}}$ and $-\vec{\mathbf{T}}$ on the floor and table, respectively. Finally, since Earth pulls downward on the boy with force $\vec{\mathbf{w}}$, he pulls upward on Earth with force $-\vec{\mathbf{w}}$. If that student were to angrily pound the table in frustration, he would quickly learn the painful lesson (avoidable by studying Newton's laws) that the table hits back just as hard.

A person who is walking or running applies Newton's third law instinctively. For example, the runner in **Figure 8.16** pushes backward on the ground so that it pushes him forward.

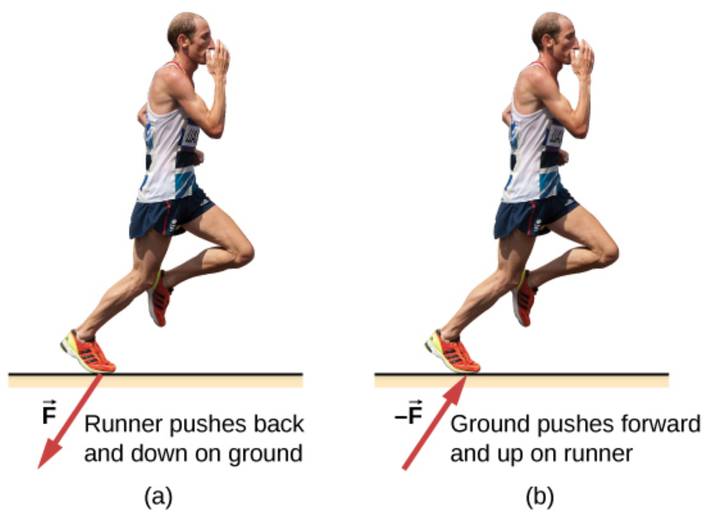


Figure 8.16 The runner experiences Newton's third law. (a) A force is exerted by the runner on the ground. (b) The reaction force of the ground on the runner pushes him forward.

Example 8.6

Forces on a Stationary Object

The package in Figure 8.17 is sitting on a scale. The forces on the package are $\vec{\mathbf{S}}$, which is due to the scale, and $-\vec{\mathbf{w}}$, which is due to Earth's gravitational field. The reaction forces that the package exerts are $-\vec{\mathbf{S}}$ on the scale and $\vec{\mathbf{w}}$ on Earth. Because the package is not accelerating, application of the second law yields

$$\vec{\mathbf{S}} - \vec{\mathbf{w}} = m\vec{\mathbf{a}} = \vec{\mathbf{0}},$$

so

$$\vec{\mathbf{S}} = \vec{\mathbf{w}}.$$

Thus, the scale reading gives the magnitude of the package's weight. However, the scale does not measure the weight of the package; it measures the force $-\vec{\mathbf{S}}$ on its surface. If the system is accelerating, $\vec{\mathbf{S}}$ and $-\vec{\mathbf{w}}$ would not be equal.

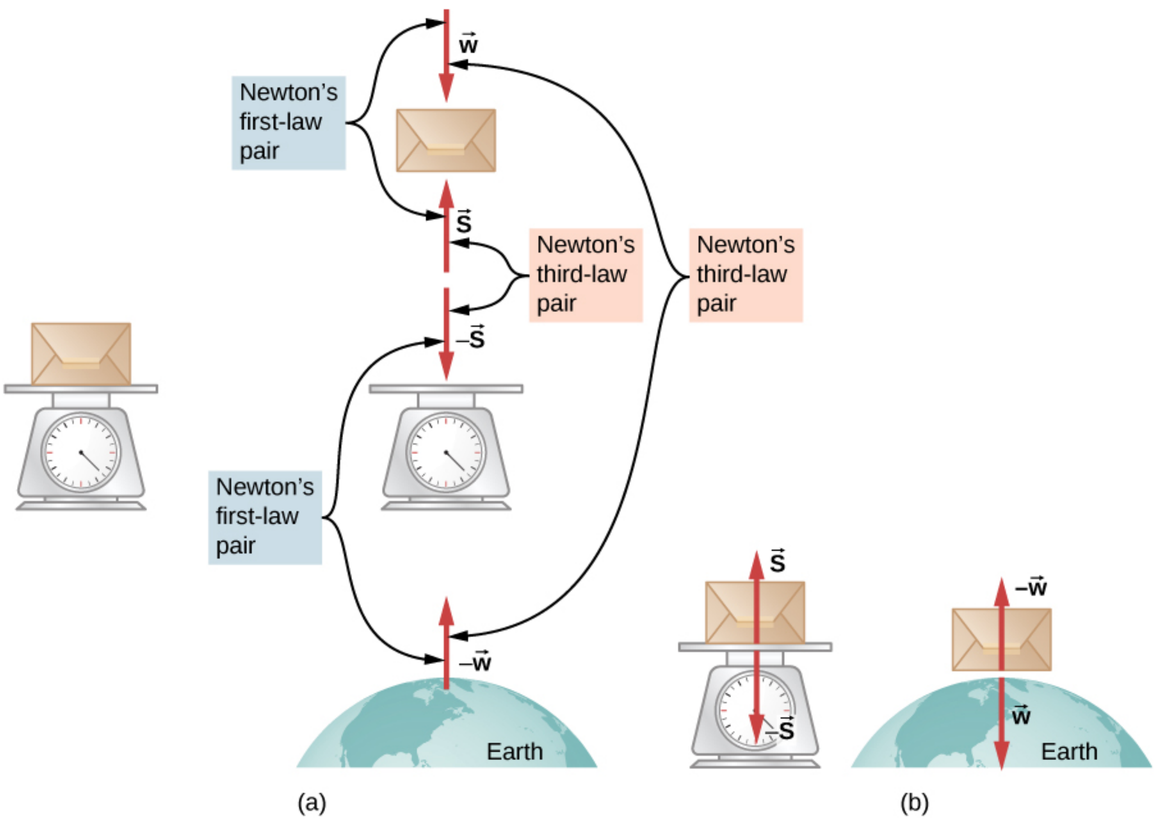


Figure 8.17 (a) The forces on a package sitting on a scale, along with their reaction forces. The force $\vec{\mathbf{w}}$ is the weight of the package (the force due to Earth's gravity) and $\vec{\mathbf{S}}$ is the force of the scale on the package. (b) Isolation of the package-scale system and the package-Earth system makes the action and reaction pairs clear.

Example 8.7

Getting Up to Speed: Choosing the Correct System

A physics professor pushes a cart of demonstration equipment to a lecture hall (Figure 8.18). Her mass is 65.0 kg, the cart's mass is 12.0 kg, and the equipment's mass is 7.0 kg. Calculate the acceleration produced when the professor exerts a backward force of 150 N on the floor. All forces opposing the motion, such as friction on the cart's wheels and air resistance, total 24.0 N.

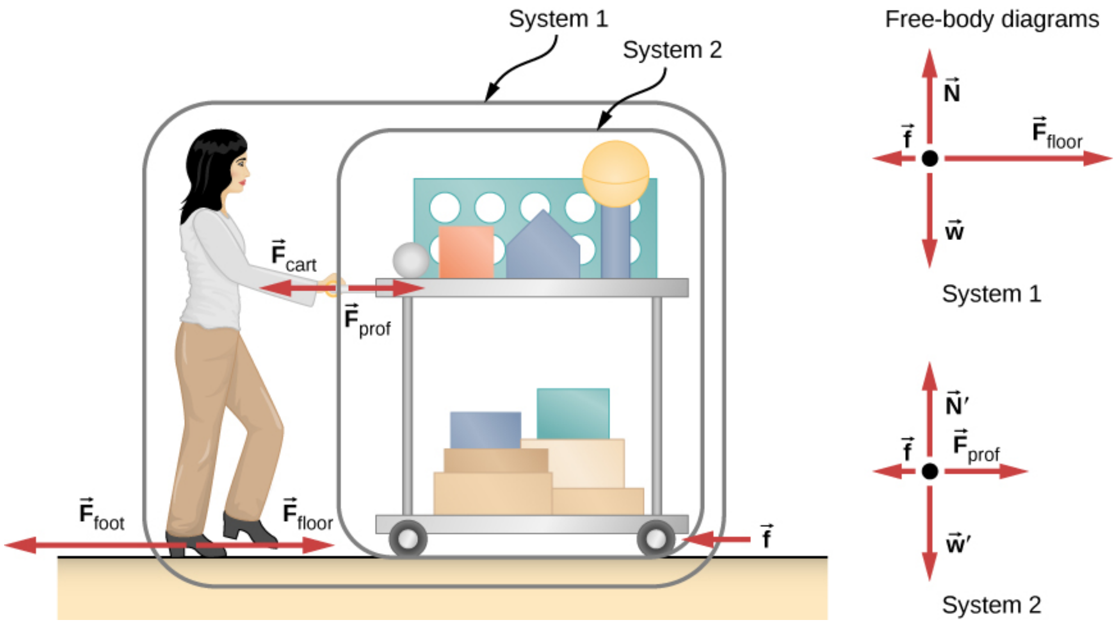


Figure 8.18 A professor pushes the cart with her demonstration equipment. The lengths of the arrows are proportional to the magnitudes of the forces (except for $\vec{f}$, because it is too small to drawn to scale). System 1 is appropriate for this example, because it asks for the acceleration of the entire group of objects. Only $\vec{F}_{floor}$ and $\vec{f}$ are external forces acting on System 1 along the line of motion. All other forces either cancel or act on the outside world. System 2 is chosen for the next example so that $\vec{F}_{prof}$ is an external force and enters into Newton's second law. The free-body diagrams, which serve as the basis for Newton's second law, vary with the system chosen.

Strategy

Since they accelerate as a unit, we define the system to be the professor, cart, and equipment. This is System 1 in **Figure 8.18**. The professor pushes backward with a force F_{foot} of 150 N. According to Newton's third law, the floor exerts a forward reaction force F_{floor} of 150 N on System 1. Because all motion is horizontal, we can assume there is no net force in the vertical direction. Therefore, the problem is one-dimensional along the horizontal direction. As noted, friction f opposes the motion and is thus in the opposite direction of F_{floor}. We do not include the forces F_{prof} or F_{cart} because these are internal forces, and we do not include F_{foot} because it acts on the floor, not on the system. There are no other significant forces acting on System 1. If the net external force can be found from all this information, we can use Newton's second law to find the acceleration as requested. See the free-body diagram in the figure.

Solution

Newton's second law is given by

$$a = \frac{F_{net}}{m}.$$

The net external force on System 1 is deduced from **Figure 8.18** and the preceding discussion to be

$$F_{net} = F_{floor} - f = 150\,N - 24.0\,N = 126\,N.$$

The mass of System 1 is

$$m = (65.0 + 12.0 + 7.0)\ \text{kg} = 84\ \text{kg}.$$

These values of F_{net} and m produce an acceleration of

$$a = \frac{F_{\text{net}}}{m} = \frac{126\ \text{N}}{84\ \text{kg}} = 1.5\ \text{m/s}^2.$$

Significance

None of the forces between components of System 1, such as between the professor's hands and the cart, contribute to the net external force because they are internal to System 1. Another way to look at this is that forces between components of a system cancel because they are equal in magnitude and opposite in direction. For example, the force exerted by the professor on the cart results in an equal and opposite force back on the professor. In this case, both forces act on the same system and therefore cancel. Thus, internal forces (between components of a system) cancel. Choosing System 1 was crucial to solving this problem.

Example 8.8

Force on the Cart: Choosing a New System

Calculate the force the professor exerts on the cart in **Figure 8.18**, using data from the previous example if needed.

Strategy

If we define the system of interest as the cart plus the equipment (System 2 in **Figure 8.18**), then the net external force on System 2 is the force the professor exerts on the cart minus friction. The force she exerts on the cart, F_{prof}, is an external force acting on System 2. F_{prof} was internal to System 1, but it is external to System 2 and thus enters Newton's second law for this system.

Solution

Newton's second law can be used to find F_{prof}. We start with

$$a = \frac{F_{\text{net}}}{m}.$$

The magnitude of the net external force on System 2 is

$$F_{\text{net}} = F_{\text{prof}} - f.$$

We solve for F_{prof}, the desired quantity:

$$F_{\text{prof}} = F_{\text{net}} + f.$$

The value of f is given, so we must calculate net F_{net}. That can be done because both the acceleration and the mass of System 2 are known. Using Newton's second law, we see that

$$F_{\text{net}} = ma,$$

where the mass of System 2 is 19.0 kg ($m = 12.0\ \text{kg} + 7.0\ \text{kg}$) and its acceleration was found to be $a = 1.5\ \text{m/s}^2$ in the previous example. Thus,

$$F_{\text{net}} = ma = (19.0\ \text{kg})\left(1.5\ \text{m/s}^2\right) = 29\ \text{N}.$$

Now we can find the desired force:

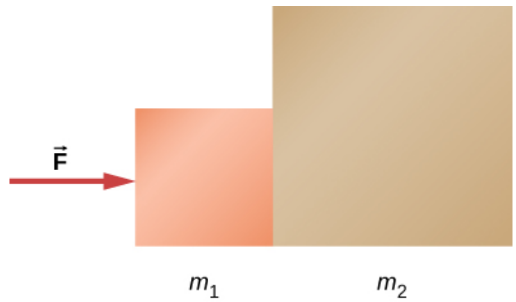
$$F_{\text{prof}} = F_{\text{net}} + f = 29\ \text{N} + 24.0\ \text{N} = 53\ \text{N}.$$

Significance

This force is significantly less than the 150-N force the professor exerted backward on the floor. Not all of that

150-N force is transmitted to the cart; some of it accelerates the professor. The choice of a system is an important analytical step both in solving problems and in thoroughly understanding the physics of the situation (which are not necessarily the same things).

8.6 Check Your Understanding Two blocks are at rest and in contact on a frictionless surface as shown below, with $m_1 = 2.0\,\text{kg}$, $m_2 = 6.0\,\text{kg}$, and applied force 24 N. (a) Find the acceleration of the system of blocks. (b) Suppose that the blocks are later separated. What force will give the second block, with the mass of 6.0 kg, the same acceleration as the system of blocks?



View this **video (https://openstaxcollege.org/l/21actionreact)** to watch examples of action and reaction.

View this **video (https://openstaxcollege.org/l/21NewtonsLaws)** to watch examples of Newton's laws and internal and external forces.

8.6 | Common Forces

Learning Objectives

By the end of the section, you will be able to:

- Define normal and tension forces
- Distinguish between real and fictitious forces
- Apply Newton's laws of motion to solve problems involving a variety of forces

Forces are given many names, such as push, pull, thrust, and weight. Traditionally, forces have been grouped into several categories and given names relating to their source, how they are transmitted, or their effects. Several of these categories are discussed in this section, together with some interesting applications. Further examples of forces are discussed later in this text.

A Catalog of Forces: Normal, Tension, and Other Examples of Forces

A catalog of forces will be useful for reference as we solve various problems involving force and motion. These forces include normal force, tension, and friction.

Normal force

Weight (also called the force of gravity) is a pervasive force that acts at all times and must be counteracted to keep an object from falling. You must support the weight of a heavy object by pushing up on it when you hold it stationary, as illustrated in **Figure 8.19**(a). But how do inanimate objects like a table support the weight of a mass placed on them, such as shown in **Figure 8.19**(b)? When the bag of dog food is placed on the table, the table sags slightly under the load. This would be noticeable if the load were placed on a card table, but even a sturdy oak table deforms when a force is applied to it. Unless an object is deformed beyond its limit, it will exert a restoring force much like a deformed spring (or a trampoline or diving board). The greater the deformation, the greater the restoring force. Thus, when the load is placed on the table, the table sags until the restoring force becomes as large as the weight of the load. At this point, the net external force on the load is zero. That is the situation when the load is stationary on the table. The table sags quickly and the sag is slight, so we do not notice it. But it is similar to the sagging of a trampoline when you climb onto it.

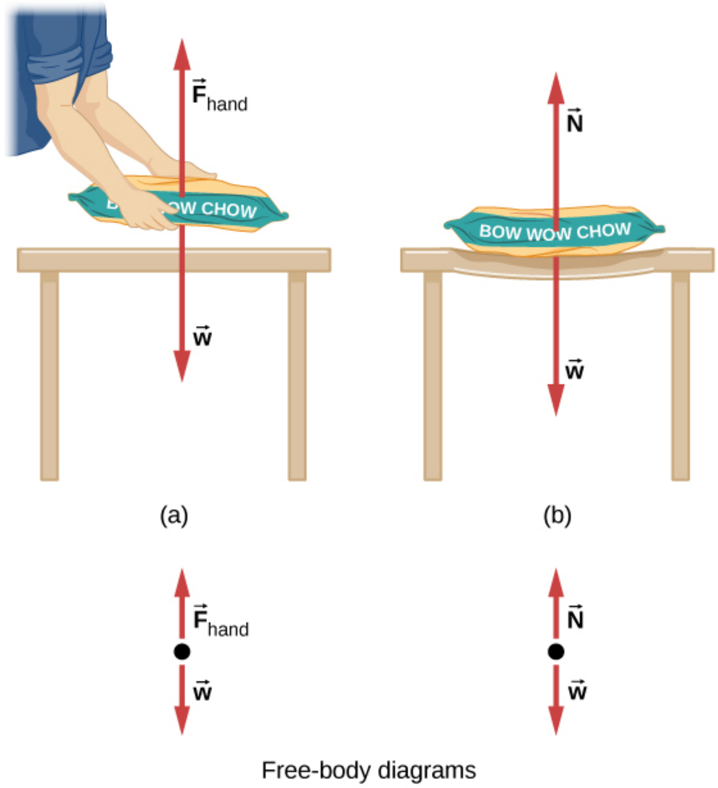


Figure 8.19 (a) The person holding the bag of dog food must supply an upward force $\vec{F}_{hand}$ equal in magnitude and opposite in direction to the weight of the food $\vec{w}$ so that it doesn't drop to the ground. (b) The card table sags when the dog food is placed on it, much like a stiff trampoline. Elastic restoring forces in the table grow as it sags until they supply a force $\vec{N}$ equal in magnitude and opposite in direction to the weight of the load.

We must conclude that whatever supports a load, be it animate or not, must supply an upward force equal to the weight of the load, as we assumed in a few of the previous examples. If the force supporting the weight of an object, or a load, is perpendicular to the surface of contact between the load and its support, this force is defined as a **normal force** and here is given by the symbol $\vec{N}$. (This is not the newton unit for force, or N.) The word *normal* means perpendicular to a surface. This means that the normal force experienced by an object resting on a horizontal surface can be expressed in vector form as follows:

$$\vec{N} = -m\vec{g}. \tag{8.8}$$

In scalar form, this becomes

$$N = mg. \tag{8.9}$$

The normal force can be less than the object’s weight if the object is on an incline.

When an object rests on an incline that makes an angle θ with the horizontal, the force of gravity acting on the object is divided into two components: a force acting perpendicular to the plane, w_y, and a force acting parallel to the plane, w_x (**Figure 8.20**). The normal force $\vec{\mathbf{N}}$ is typically equal in magnitude and opposite in direction to the perpendicular component of the weight w_y. The force acting parallel to the plane, w_x, causes the object to accelerate down the incline.

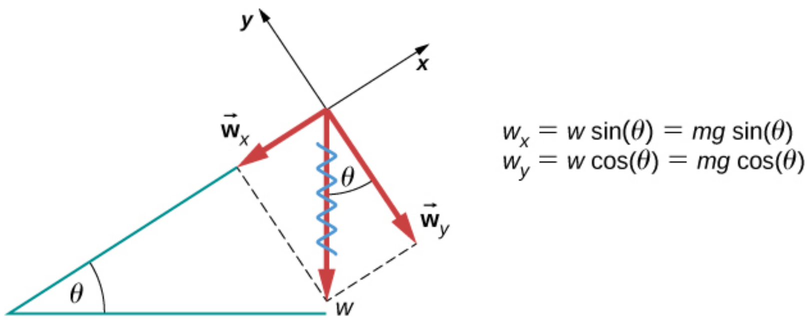


Figure 8.20 An object rests on an incline that makes an angle θ with the horizontal.

Be careful when resolving the weight of the object into components. If the incline is at an angle θ to the horizontal, then the magnitudes of the weight components are

$$w_x = w\sin\theta = mg\sin\theta$$

and

$$w_y = w\cos\theta = mg\cos\theta.$$

We use the second equation to write the normal force experienced by an object resting on an inclined plane:

$$N = mg\cos\theta. \qquad (8.10)$$

Instead of memorizing these equations, it is helpful to be able to determine them from reason. To do this, we draw the right angle formed by the three weight vectors. The angle θ of the incline is the same as the angle formed between *w* and w_y.

Knowing this property, we can use trigonometry to determine the magnitude of the weight components:

$$\cos\theta = \frac{w_y}{w}, \quad w_y = w\cos\theta = mg\sin\theta$$

$$\sin\theta = \frac{w_x}{w}, \quad w_x = w\sin\theta = mg\sin\theta.$$

8.7 **Check Your Understanding** A force of 1150 N acts parallel to a ramp to push a 250-kg gun safe into a moving van. The ramp is frictionless and inclined at $17°$. (a) What is the acceleration of the safe up the ramp? (b) If we consider friction in this problem, with a friction force of 120 N, what is the acceleration of the safe?

Tension

A **tension** is a force along the length of a medium; in particular, it is a pulling force that acts along a stretched flexible connector, such as a rope or cable. The word “tension” comes from a Latin word meaning “to stretch.” Not coincidentally, the flexible cords that carry muscle forces to other parts of the body are called *tendons*.

Any flexible connector, such as a string, rope, chain, wire, or cable, can only exert a pull parallel to its length; thus, a force carried by a flexible connector is a tension with a direction parallel to the connector. Tension is a pull in a connector. Consider the phrase: “You can’t push a rope.” Instead, tension force pulls outward along the two ends of a rope.

Consider a person holding a mass on a rope, as shown in **Figure 8.21**. If the 5.00-kg mass in the figure is stationary, then its acceleration is zero and the net force is zero. The only external forces acting on the mass are its weight and the tension supplied by the rope. Thus,

$$F_{\text{net}} = T - w = 0,$$

where T and w are the magnitudes of the tension and weight, respectively, and their signs indicate direction, with up being positive. As we proved using Newton's second law, the tension equals the weight of the supported mass:

$$T = w = mg. \tag{8.11}$$

Thus, for a 5.00-kg mass (neglecting the mass of the rope), we see that

$$T = mg = (5.00\ \text{kg})\left(9.80\ \text{m/s}^2\right) = 49.0\ \text{N}.$$

If we cut the rope and insert a spring, the spring would extend a length corresponding to a force of 49.0 N, providing a direct observation and measure of the tension force in the rope.

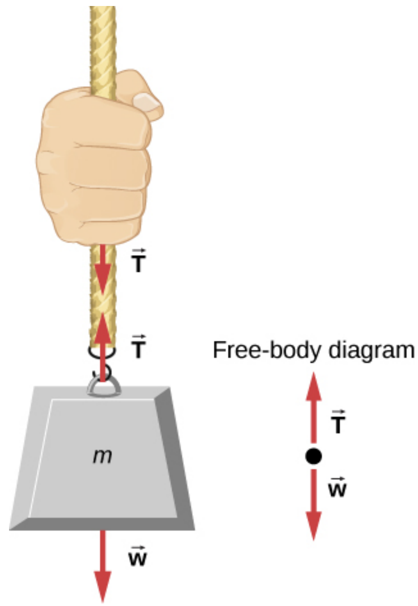


Figure 8.21 When a perfectly flexible connector (one requiring no force to bend it) such as this rope transmits a force $\vec{\mathbf{T}}$, that force must be parallel to the length of the rope, as shown. By Newton's third law, the rope pulls with equal force but in opposite directions on the hand and the supported mass (neglecting the weight of the rope). The rope is the medium that carries the equal and opposite forces between the two objects. The tension anywhere in the rope between the hand and the mass is equal. Once you have determined the tension in one location, you have determined the tension at all locations along the rope.

Flexible connectors are often used to transmit forces around corners, such as in a hospital traction system, a tendon, or a bicycle brake cable. If there is no friction, the tension transmission is undiminished; only its direction changes, and it is always parallel to the flexible connector, as shown in **Figure 8.22**.

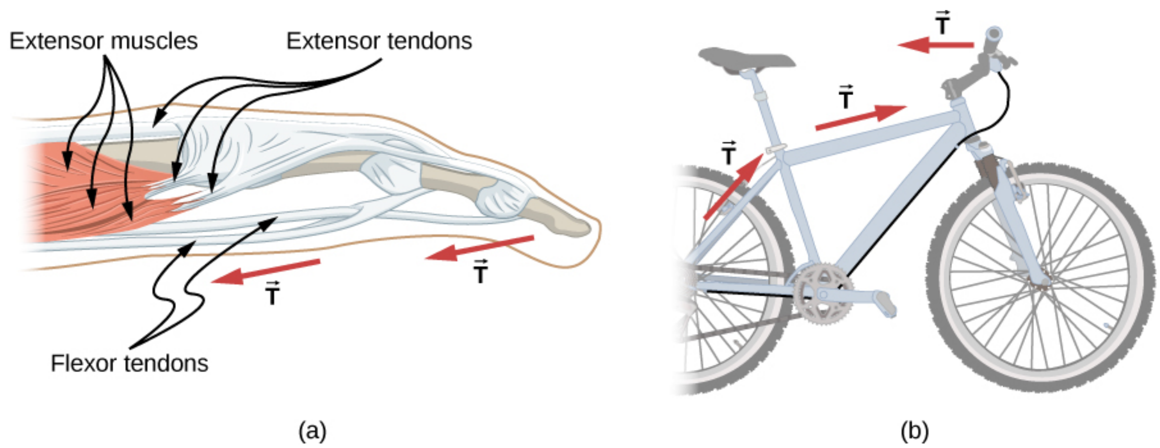


Figure 8.22 (a) Tendons in the finger carry force T from the muscles to other parts of the finger, usually changing the force's direction but not its magnitude (the tendons are relatively friction free). (b) The brake cable on a bicycle carries the tension T from the brake lever on the handlebars to the brake mechanism. Again, the direction but not the magnitude of T is changed.

Friction

Friction is a resistive force opposing motion or its tendency. Imagine an object at rest on a horizontal surface. The net force acting on the object must be zero, leading to equality of the weight and the normal force, which act in opposite directions. If the surface is tilted, the normal force balances the component of the weight perpendicular to the surface. If the object does not slide downward, the component of the weight parallel to the inclined plane is balanced by friction. This is discussed in greater detail in section on **Friction**.

8.7 | Drawing Free-Body Diagrams

Learning Objectives

By the end of the section, you will be able to:

- Explain the rules for drawing a free-body diagram
- Construct free-body diagrams for different situations

The first step in describing and analyzing most phenomena in physics involves the careful drawing of a free-body diagram. Free-body diagrams have been used in examples throughout this chapter. Remember that a free-body diagram must only include the external forces acting on the body of interest. Once we have drawn an accurate free-body diagram, we can apply Newton's first law if the body is in equilibrium (balanced forces; that is, $F_{\text{net}} = 0$) or Newton's second law if the body is accelerating (unbalanced force; that is, $F_{\text{net}} \neq 0$).

In **Forces**, we gave a brief problem-solving strategy to help you understand free-body diagrams. Here, we add some details to the strategy that will help you in constructing these diagrams.

Problem-Solving Strategy: Constructing Free-Body Diagrams

Observe the following rules when constructing a free-body diagram:

1. Draw the object under consideration; it does not have to be artistic. At first, you may want to draw a circle around the object of interest to be sure you focus on labeling the forces acting on the object. If you are treating the object as a particle (no size or shape and no rotation), represent the object as a point. We often place this point at the origin of an *xy*-coordinate system.
2. Include all forces that act on the object, representing these forces as vectors. Consider the types of forces described in **Common Forces**—normal force, friction, and tension—as well as weight and applied force. Do not include the net force on the object. With the exception of gravity, all of the forces we have discussed require direct contact with the object. However, forces that the object exerts on its environment must not be

included. We never include both forces of an action-reaction pair.

3. Convert the free-body diagram into a more detailed diagram showing the *x*- and *y*-components of a given force (this is often helpful when solving a problem using Newton's first or second law). In this case, place a squiggly line through the original vector to show that it is no longer in play—it has been replaced by its *x*- and *y*-components.
4. If there are two or more objects, or bodies, in the problem, draw a separate free-body diagram for each object.

Note: If there is acceleration, we do not directly include it in the free-body diagram; however, it may help to indicate acceleration outside the free-body diagram. You can label it in a different color to indicate that it is separate from the free-body diagram.

Let's apply the problem-solving strategy in drawing a free-body diagram for a sled. In **Figure 8.23**(a), a sled is pulled by force **P** at an angle of $30°$. In part (b), we show a free-body diagram for this situation, as described by steps 1 and 2 of the problem-solving strategy. In part (c), we show all forces in terms of their *x*- and *y*-components, in keeping with step 3.

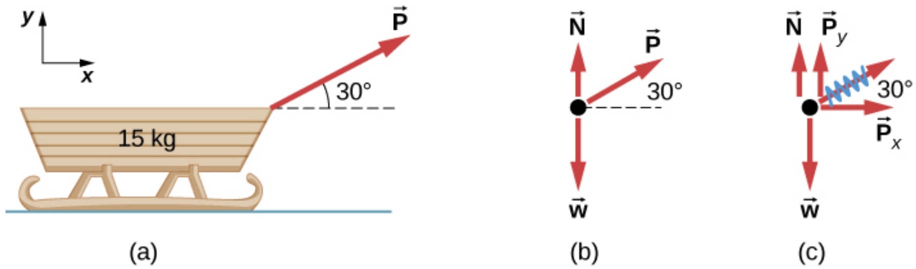


Figure 8.23 (a) A moving sled is shown as (b) a free-body diagram and (c) a free-body diagram with force components.

Example 8.9

Two Blocks on an Inclined Plane

Construct the free-body diagram for object A and object B in **Figure 8.24**.

Strategy

We follow the four steps listed in the problem-solving strategy.

Solution

We start by creating a diagram for the first object of interest. In **Figure 8.24**(a), object A is isolated (circled) and represented by a dot.

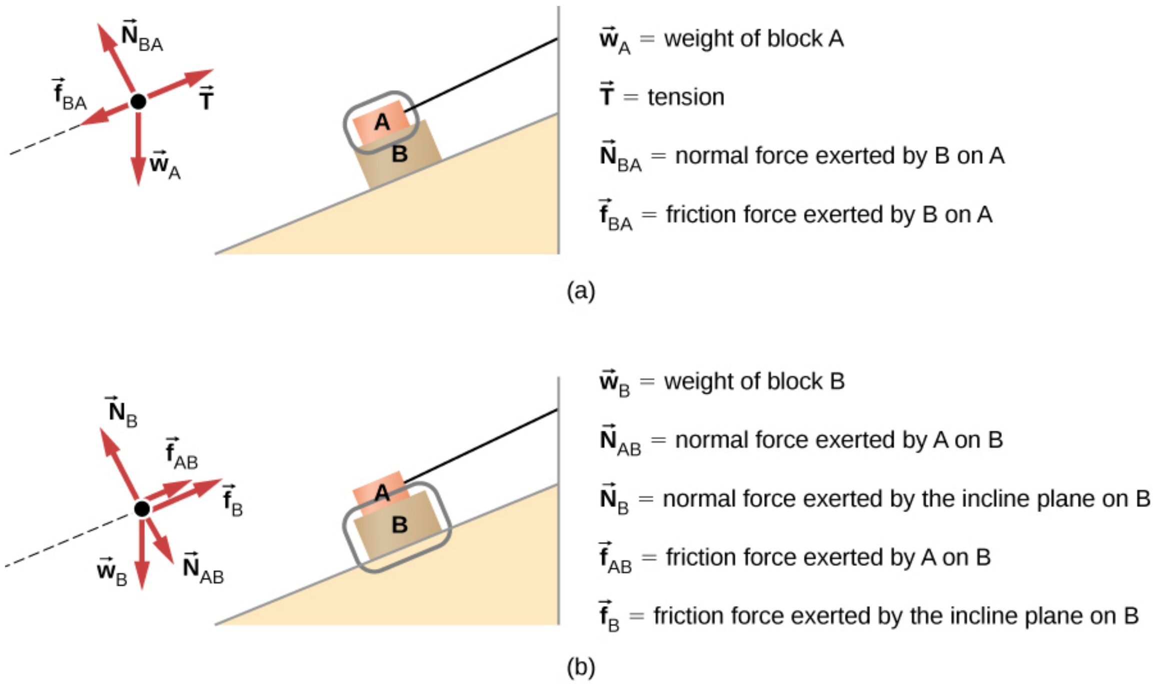


Figure 8.24 (a) The free-body diagram for isolated object A. (b) The free-body diagram for isolated object B. Comparing the two drawings, we see that friction acts in the opposite direction in the two figures. Because object A experiences a force that tends to pull it to the right, friction must act to the left. Because object B experiences a component of its weight that pulls it to the left, down the incline, the friction force must oppose it and act up the ramp. Friction always acts opposite the intended direction of motion.

We now include any force that acts on the body. Here, no applied force is present. The weight of the object acts as a force pointing vertically downward, and the presence of the cord indicates a force of tension pointing away from the object. Object A has one interface and hence experiences a normal force, directed away from the interface. The source of this force is object B, and this normal force is labeled accordingly. Since object B has a tendency to slide down, object A has a tendency to slide up with respect to the interface, so the friction f_{BA} is directed downward parallel to the inclined plane.

As noted in step 4 of the problem-solving strategy, we then construct the free-body diagram in **Figure 8.24**(b) using the same approach. Object B experiences two normal forces and two friction forces due to the presence of two contact surfaces. The interface with the inclined plane exerts external forces of N_B and f_B, and the interface with object B exerts the normal force N_{AB} and friction f_{AB}; N_{AB} is directed away from object B, and f_{AB} is opposing the tendency of the relative motion of object B with respect to object A.

Significance

The object under consideration in each part of this problem was circled in gray. When you are first learning how to draw free-body diagrams, you will find it helpful to circle the object before deciding what forces are acting on that particular object. This focuses your attention, preventing you from considering forces that are not acting on the body.

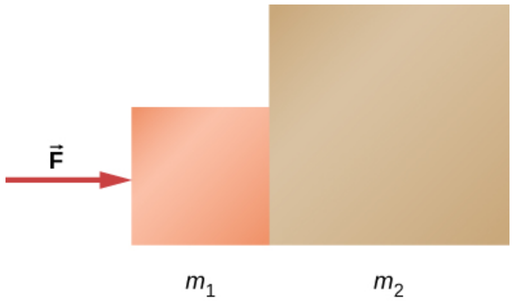
Example 8.10

Two Blocks in Contact

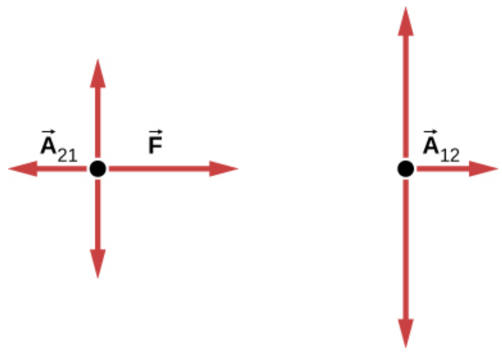
A force is applied to two blocks in contact, as shown.

Strategy

Draw a free-body diagram for each block. Be sure to consider Newton’s third law at the interface where the two blocks touch.



Solution



Significance

$\vec{A}_{21}$ is the action force of block 2 on block 1. $\vec{A}_{12}$ is the reaction force of block 1 on block 2.

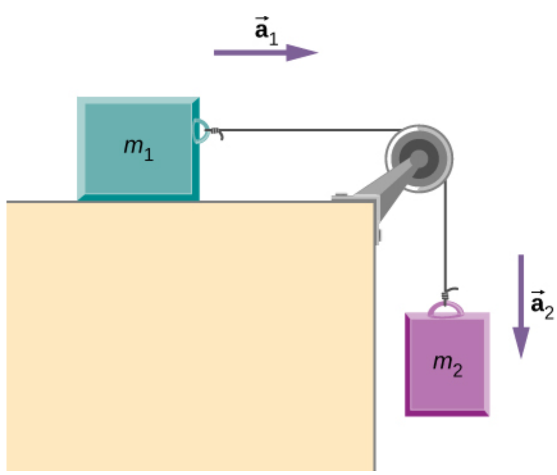
Example 8.11

Block on the Table (Coupled Blocks)

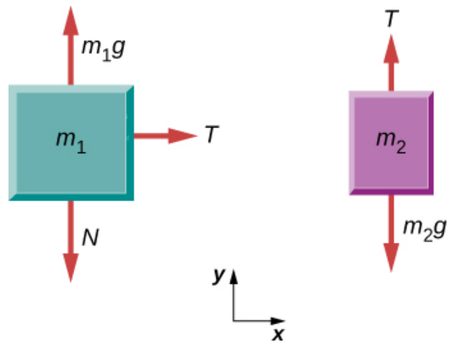
A block rests on the table, as shown. A light rope is attached to it and runs over a pulley. The other end of the rope is attached to a second block. The two blocks are said to be coupled. Block m_2 exerts a force due to its weight, which causes the system (two blocks and a string) to accelerate.

Strategy

We assume that the string has no mass so that we do not have to consider it as a separate object. Draw a free-body diagram for each block.



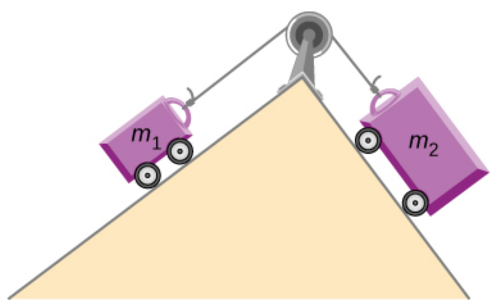
Solution



Significance

Each block accelerates (notice the labels shown for $\vec{a}_1$ and $\vec{a}_2$); however, assuming the string remains taut, they accelerate at the same rate. Thus, we have $a_1 = a_2$. If we were to continue solving the problem, we could simply call the magnitude of the acceleration a. Also, we use two free-body diagrams because we are usually finding tension *T*, which may require us to use a system of two equations in this type of problem. The tension is the same on both m_1 and m_2.

8.8 Check Your Understanding (a) Draw the free-body diagram for the situation shown. (b) Redraw it showing components; use *x*-axes parallel to the two ramps.



View this **simulation (https://openstaxcollege.org/l/21forcemotion)** to predict, qualitatively, how an external force will affect the speed and direction of an object's motion. Explain the effects with the help of a free-body diagram. Use free-body diagrams to draw position, velocity, acceleration, and force graphs, and vice versa. Explain how the graphs relate to one another. Given a scenario or a graph, sketch all four graphs.

8.8 | Friction

Learning Objectives

By the end of the section, you will be able to:

- Describe the general characteristics of friction
- List the various types of friction
- Calculate the magnitude of static and kinetic friction, and use these in problems involving Newton's laws of motion

When a body is in motion, it has resistance because the body interacts with its surroundings. This resistance is a force of friction. Friction opposes relative motion between systems in contact but also allows us to move, a concept that becomes obvious if you try to walk on ice. Friction is a common yet complex force, and its behavior still not completely understood.

Still, it is possible to understand the circumstances in which it behaves.

Static and Kinetic Friction

The basic definition of **friction** is relatively simple to state.

Friction

Friction is a force that opposes relative motion between systems in contact.

There are several forms of friction. One of the simpler characteristics of sliding friction is that it is parallel to the contact surfaces between systems and is always in a direction that opposes motion or attempted motion of the systems relative to each other. If two systems are in contact and moving relative to one another, then the friction between them is called kinetic friction. For example, friction slows a hockey puck sliding on ice. When objects are stationary, static friction can act between them; the static friction is usually greater than the kinetic friction between two objects.

Static and Kinetic Friction

If two systems are in contact and stationary relative to one another, then the friction between them is called **static friction**. If two systems are in contact and moving relative to one another, then the friction between them is called **kinetic friction**.

Imagine, for example, trying to slide a heavy crate across a concrete floor—you might push very hard on the crate and not move it at all. This means that the static friction responds to what you do—it increases to be equal to and in the opposite direction of your push. If you finally push hard enough, the crate seems to slip suddenly and starts to move. Now static friction gives way to kinetic friction. Once in motion, it is easier to keep it in motion than it was to get it started, indicating that the kinetic frictional force is less than the static frictional force. If you add mass to the crate, say by placing a box on top of it, you need to push even harder to get it started and also to keep it moving. Furthermore, if you oiled the concrete you would find it easier to get the crate started and keep it going (as you might expect).

Figure 8.25 is a crude pictorial representation of how friction occurs at the interface between two objects. Close-up inspection of these surfaces shows them to be rough. Thus, when you push to get an object moving (in this case, a crate), you must raise the object until it can skip along with just the tips of the surface hitting, breaking off the points, or both. A considerable force can be resisted by friction with no apparent motion. The harder the surfaces are pushed together (such as if another box is placed on the crate), the more force is needed to move them. Part of the friction is due to adhesive forces between the surface molecules of the two objects, which explains the dependence of friction on the nature of the substances. For example, rubber-soled shoes slip less than those with leather soles. Adhesion varies with substances in contact and is a complicated aspect of surface physics. Once an object is moving, there are fewer points of contact (fewer molecules adhering), so less force is required to keep the object moving. At small but nonzero speeds, friction is nearly independent of speed.

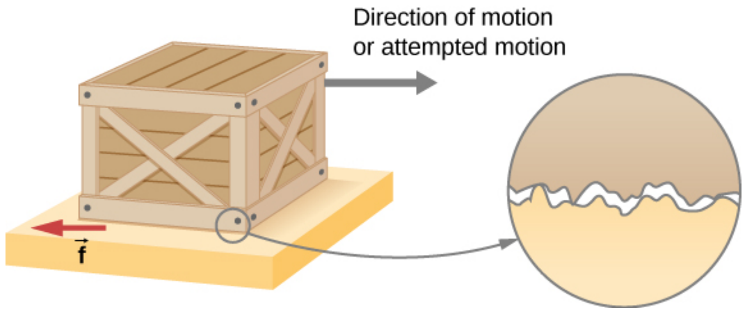


Figure 8.25 Frictional forces, such as $\vec{\mathbf{f}}$, always oppose motion or attempted motion between objects in contact. Friction arises in part because of the roughness of the surfaces in contact, as seen in the expanded view. For the object to move, it must rise to where the peaks of the top surface can skip along the bottom surface. Thus, a force is required just to set the object in motion. Some of the peaks will be broken off, also requiring a force to maintain motion. Much of the friction is actually due to attractive forces between molecules making up the two objects, so that even perfectly smooth surfaces are not friction-free. (In fact, perfectly smooth, clean surfaces of similar materials would adhere, forming a bond called a "cold weld.")

The magnitude of the frictional force has two forms: one for static situations (static friction), the other for situations involving motion (kinetic friction). What follows is an approximate empirical (experimentally determined) model only. These equations for static and kinetic friction are not vector equations.

Magnitude of Static Friction

The magnitude of static friction f_s is

$$f_\text{s} \le \mu_\text{s} N, \tag{8.12}$$

where μ_s is the coefficient of static friction and N is the magnitude of the normal force.

The symbol $\le$ means *less than or equal to*, implying that static friction can have a maximum value of $\mu_\text{s} N$. Static friction is a responsive force that increases to be equal and opposite to whatever force is exerted, up to its maximum limit. Once the applied force exceeds
$f_\text{s}(\text{max})$, the object moves. Thus,

$$f_\text{s}(\text{max}) = \mu_\text{s} N.$$

Magnitude of Kinetic Friction

The magnitude of kinetic friction f_k is given by

$$f_\text{k} = \mu_\text{k} N, \tag{8.13}$$

where μ_k is the coefficient of kinetic friction.

A system in which $f_\text{k} = \mu_\text{k} N$ is described as a system in which *friction behaves simply*. The transition from static friction to kinetic friction is illustrated in **Figure 8.26**.

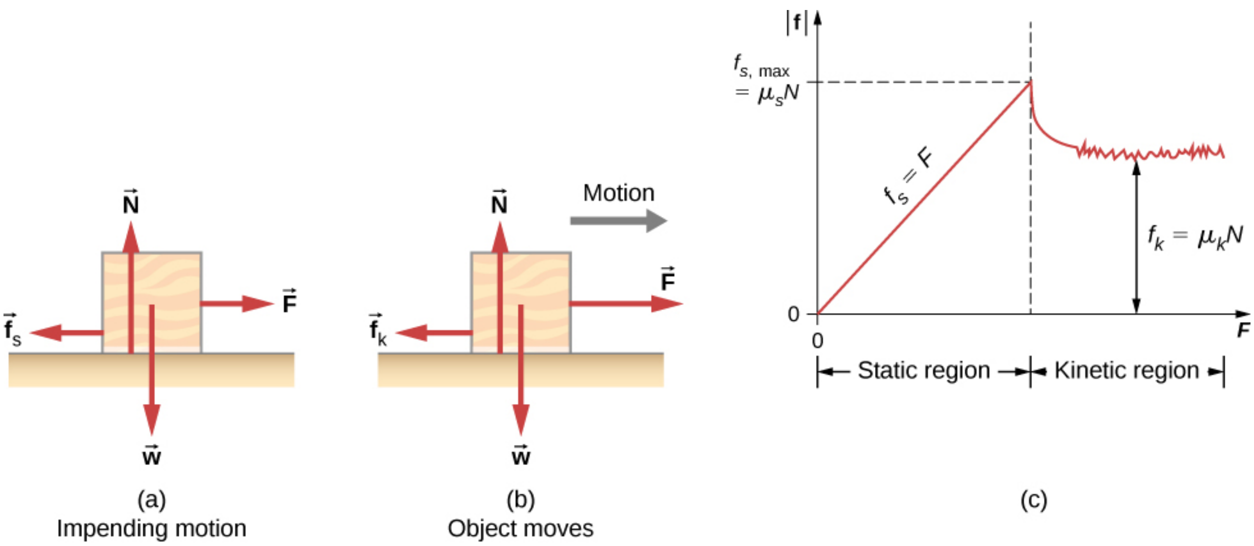


Figure 8.26 (a) The force of friction $\vec{\mathbf{f}}$ between the block and the rough surface opposes the direction of the applied force $\vec{\mathbf{F}}$. The magnitude of the static friction balances that of the applied force. This is shown in the left side of the graph in (c). (b) At some point, the magnitude of the applied force is greater than the force of kinetic friction, and the block moves to the right. This is shown in the right side of the graph. (c) The graph of the frictional force versus the applied force; note that $f_s(\text{max}) > f_k$. This means that $\mu_s > \mu_k$.

As you can see in **Table 8.1**, the coefficients of kinetic friction are less than their static counterparts. The approximate values of μ are stated to only one or two digits to indicate the approximate description of friction given by the preceding two equations.

System	Static Friction μ_s	Kinetic Friction μ_k
Rubber on dry concrete	1.0	0.7
Rubber on wet concrete	0.5-0.7	0.3-0.5
Wood on wood	0.5	0.3
Waxed wood on wet snow	0.14	0.1
Metal on wood	0.5	0.3
Steel on steel (dry)	0.6	0.3
Steel on steel (oiled)	0.05	0.03
Teflon on steel	0.04	0.04
Bone lubricated by synovial fluid	0.016	0.015
Shoes on wood	0.9	0.7
Shoes on ice	0.1	0.05
Ice on ice	0.1	0.03
Steel on ice	0.4	0.02

Table 8.1 Approximate Coefficients of Static and Kinetic Friction

Equation 8.12 and **Equation 8.13** include the dependence of friction on materials and the normal force. The direction of friction is always opposite that of motion, parallel to the surface between objects, and perpendicular to the normal force. For example, if the crate you try to push (with a force parallel to the floor) has a mass of 100 kg, then the normal force is equal to its weight,

$$w = mg = (100\ \text{kg})\left(9.80\ \text{m/s}^2\right) = 980\ \text{N},$$

perpendicular to the floor. If the coefficient of static friction is 0.45, you would have to exert a force parallel to the floor greater than

$$f_\text{s}(\text{max}) = \mu_\text{s}N = (0.45)(980\ \text{N}) = 440\ \text{N}$$

to move the crate. Once there is motion, friction is less and the coefficient of kinetic friction might be 0.30, so that a force of only

$$f_\text{k} = \mu_\text{k}N = (0.30)(980\ \text{N}) = 290\ \text{N}$$

keeps it moving at a constant speed. If the floor is lubricated, both coefficients are considerably less than they would be without lubrication. Coefficient of friction is a unitless quantity with a magnitude usually between 0 and 1.0. The actual value depends on the two surfaces that are in contact.

Many people have experienced the slipperiness of walking on ice. However, many parts of the body, especially the joints, have much smaller coefficients of friction—often three or four times less than ice. A joint is formed by the ends of two bones, which are connected by thick tissues. The knee joint is formed by the lower leg bone (the tibia) and the thighbone (the femur). The hip is a ball (at the end of the femur) and socket (part of the pelvis) joint. The ends of the bones in the joint are covered by cartilage, which provides a smooth, almost-glassy surface. The joints also produce a fluid (synovial fluid) that reduces friction and wear. A damaged or arthritic joint can be replaced by an artificial joint (**Figure 8.27**). These replacements can be made of metals (stainless steel or titanium) or plastic (polyethylene), also with very small coefficients of friction.

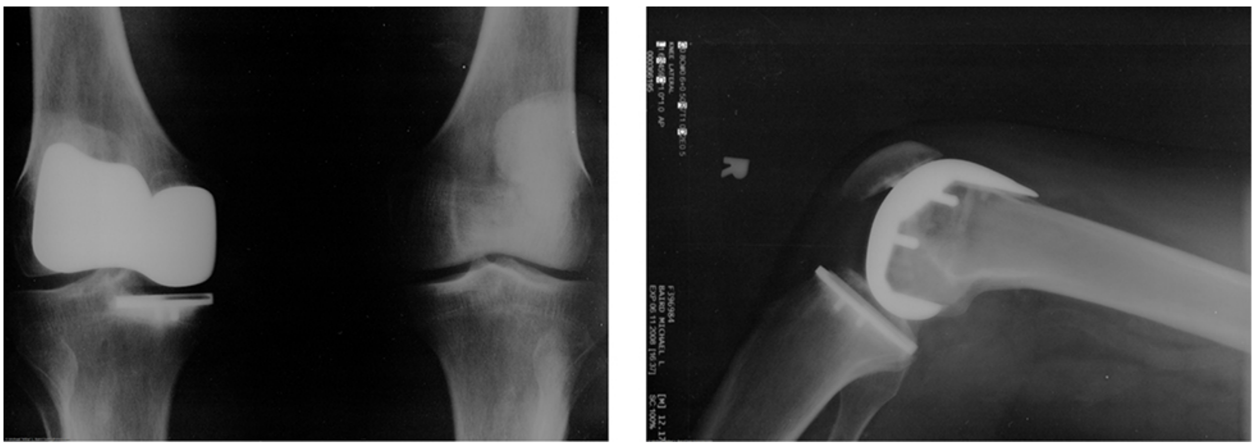


Figure 8.27 Artificial knee replacement is a procedure that has been performed for more than 20 years. These post-operative X-rays show a right knee joint replacement. (credit: Mike Baird)

Natural lubricants include saliva produced in our mouths to aid in the swallowing process, and the slippery mucus found between organs in the body, allowing them to move freely past each other during heartbeats, during breathing, and when a person moves. Hospitals and doctor's clinics commonly use artificial lubricants, such as gels, to reduce friction.

The equations given for static and kinetic friction are empirical laws that describe the behavior of the forces of friction. While these formulas are very useful for practical purposes, they do not have the status of mathematical statements that represent general principles (e.g., Newton's second law). In fact, there are cases for which these equations are not even good approximations. For instance, neither formula is accurate for lubricated surfaces or for two surfaces siding across each other at high speeds. Unless specified, we will not be concerned with these exceptions.

Example 8.12

Static and Kinetic Friction

A 20.0-kg crate is at rest on a floor as shown in **Figure 8.28**. The coefficient of static friction between the crate and floor is 0.700 and the coefficient of kinetic friction is 0.600. A horizontal force $\vec{\mathbf{P}}$ is applied to the crate. Find the force of friction if (a) $\vec{\mathbf{P}} = 20.0\ \text{N}$, (b) $\vec{\mathbf{P}} = 30.0\ \text{N}$, (c) $\vec{\mathbf{P}} = 120.0\ \text{N}$, and (d) $\vec{\mathbf{P}} = 180.0\ \text{N}$.

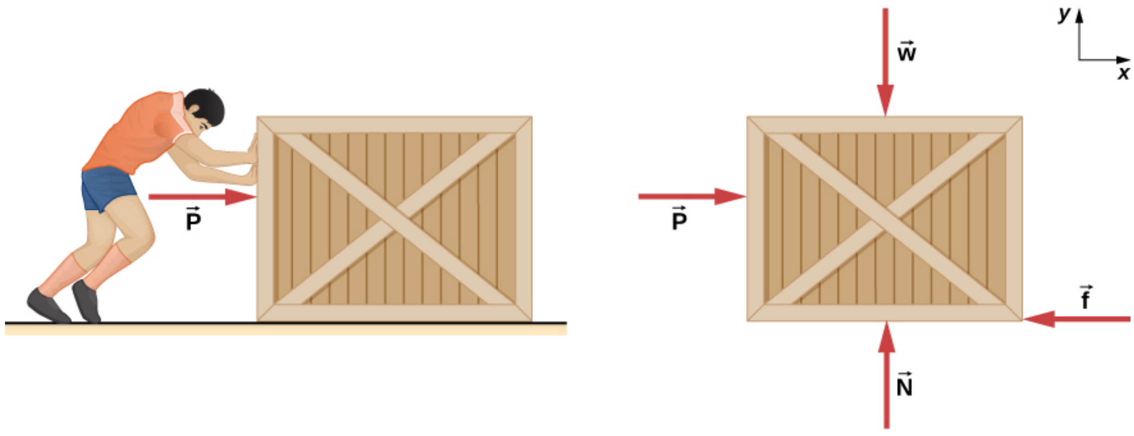


Figure 8.28 (a) A crate on a horizontal surface is pushed with a force $\vec{\mathbf{P}}$. (b) The forces on the crate. Here, $\vec{\mathbf{f}}$ may represent either the static or the kinetic frictional force.

Strategy

The free-body diagram of the crate is shown in **Figure 8.28**(b). We apply Newton's second law in the horizontal and vertical directions, including the friction force in opposition to the direction of motion of the box.

Solution

Newton's second law gives

$$\begin{aligned} \sum F_x &= ma_x & \sum F_y &= ma_y \\ P - f &= ma_x & N - w &= 0. \end{aligned}$$

Here we are using the symbol f to represent the frictional force since we have not yet determined whether the crate is subject to station friction or kinetic friction. We do this whenever we are unsure what type of friction is acting. Now the weight of the crate is

$$w = (20.0\,\text{kg})(9.80\,\text{m/s}^2) = 196\,\text{N},$$

which is also equal to N. The maximum force of static friction is therefore $(0.700)(196\,\text{N}) = 137\,\text{N}$. As long as $\vec{\mathbf{P}}$ is less than 137 N, the force of static friction keeps the crate stationary and $f_s = \vec{\mathbf{P}}$. Thus, (a) $f_s = 20.0\,\text{N}$, (b) $f_s = 30.0\,\text{N}$, and (c) $f_s = 120.0\,\text{N}$.

(d) If $\vec{\mathbf{P}} = 180.0\,\text{N}$, the applied force is greater than the maximum force of static friction (137 N), so the crate can no longer remain at rest. Once the crate is in motion, kinetic friction acts. Then

$$f_k = \mu_k N = (0.600)(196\,\text{N}) = 118\,\text{N},$$

and the acceleration is

$$a_x = \frac{\vec{\mathbf{P}} - f_k}{m} = \frac{180.0\,\text{N} - 118\,\text{N}}{20.0\,\text{kg}} = 3.10\,\text{m/s}^2.$$

Significance

This example illustrates how we consider friction in a dynamics problem. Notice that static friction has a value that matches the applied force, until we reach the maximum value of static friction. Also, no motion can occur until the applied force equals the force of static friction, but the force of kinetic friction will then become smaller.

8.9 Check Your Understanding A block of mass 1.0 kg rests on a horizontal surface. The frictional coefficients for the block and surface are $\mu_s = 0.50$ and $\mu_k = 0.40.$ (a) What is the minimum horizontal force required to move the block? (b) What is the block's acceleration when this force is applied?

Friction and the Inclined Plane

One situation where friction plays an obvious role is that of an object on a slope. It might be a crate being pushed up a ramp to a loading dock or a skateboarder coasting down a mountain, but the basic physics is the same. We usually generalize the sloping surface and call it an inclined plane but then pretend that the surface is flat. Let's look at an example of analyzing motion on an inclined plane with friction.

Example 8.13

Downhill Skier

A skier with a mass of 62 kg is sliding down a snowy slope at a constant velocity. Find the coefficient of kinetic friction for the skier if friction is known to be 45.0 N.

Strategy

The magnitude of kinetic friction is given as 45.0 N. Kinetic friction is related to the normal force N by $f_\text{k} = \mu_\text{k} N$; thus, we can find the coefficient of kinetic friction if we can find the normal force on the skier. The normal force is always perpendicular to the surface, and since there is no motion perpendicular to the surface, the normal force should equal the component of the skier's weight perpendicular to the slope. (See Figure 8.29, which repeats a figure from the chapter on Newton's laws of motion.)

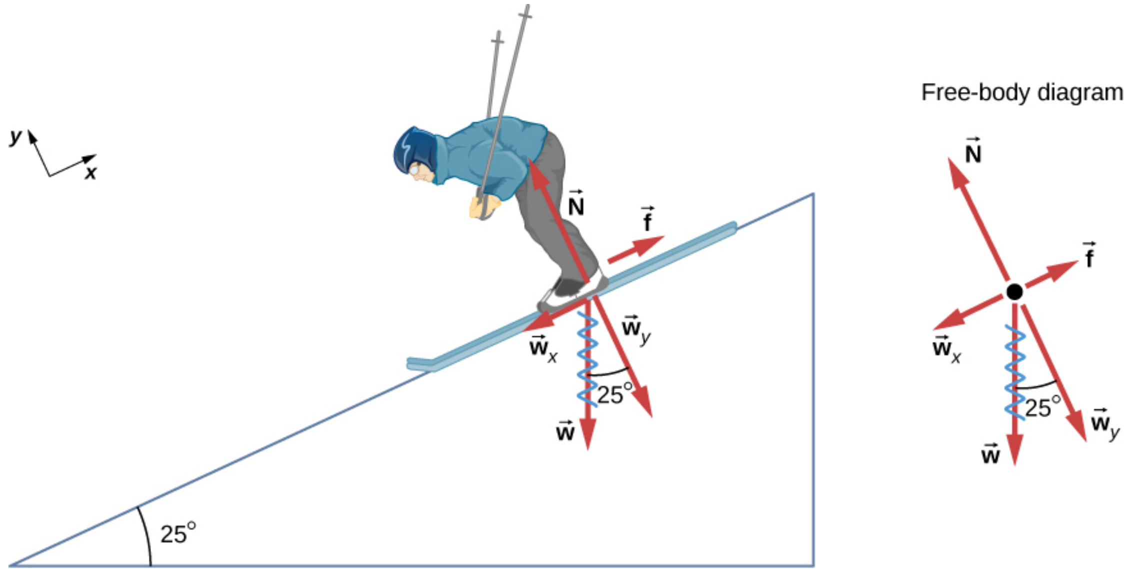


Figure 8.29 The motion of the skier and friction are parallel to the slope, so it is most convenient to project all forces onto a coordinate system where one axis is parallel to the slope and the other is perpendicular (axes shown to left of skier). The normal force $\vec{N}$ is perpendicular to the slope, and friction $\vec{f}$ is parallel to the slope, but the skier's weight $\vec{w}$ has components along both axes, namely $\vec{w}_y$ and $\vec{w}_x$. The normal force $\vec{N}$ is equal in magnitude to $\vec{w}_y$, so there is no motion perpendicular to the slope. However, $\vec{f}$ is less than $\vec{w}_x$ in magnitude, so there is acceleration down the slope (along the *x*-axis).

We have

$$N = w_y = w \cos 25° = mg \cos 25°.$$

Substituting this into our expression for kinetic friction, we obtain

$$f_{\text{k}} = \mu_{\text{k}} mg \cos 25°,$$

which can now be solved for the coefficient of kinetic friction μ_{k}.

Solution

Solving for μ_{k} gives

$$\mu_{\text{k}} = \frac{f_{\text{k}}}{N} = \frac{f_{\text{k}}}{w \cos 25°} = \frac{f_{\text{k}}}{mg \cos 25°}.$$

Substituting known values on the right-hand side of the equation,

$$\mu_{\text{k}} = \frac{45.0 \text{ N}}{(62 \text{ kg})(9.80 \text{ m/s}^2)(0.906)} = 0.082.$$

Significance

This result is a little smaller than the coefficient listed in **Table 8.1** for waxed wood on snow, but it is still reasonable since values of the coefficients of friction can vary greatly. In situations like this, where an object of mass m slides down a slope that makes an angle θ with the horizontal, friction is given by $f_{\text{k}} = \mu_{\text{k}} mg \cos \theta$. All objects slide down a slope with constant acceleration under these circumstances.

We have discussed that when an object rests on a horizontal surface, the normal force supporting it is equal in magnitude to its weight. Furthermore, simple friction is always proportional to the normal force. When an object is not on a horizontal surface, as with the inclined plane, we must find the force acting on the object that is directed perpendicular to the surface; it is a component of the weight.

We now derive a useful relationship for calculating coefficient of friction on an inclined plane. Notice that the result applies only for situations in which the object slides at constant speed down the ramp.

An object slides down an inclined plane at a constant velocity if the net force on the object is zero. We can use this fact to measure the coefficient of kinetic friction between two objects. As shown in **Example 8.12**, the kinetic friction on a slope is $f_k = \mu_k mg \cos \theta$. The component of the weight down the slope is equal to $mg \sin \theta$ (see the free-body diagram in **Figure 8.29**). These forces act in opposite directions, so when they have equal magnitude, the acceleration is zero. Writing these out,

$$\mu_{\text{k}} mg \cos \theta = mg \sin \theta.$$

Solving for μ_{k}, we find that

$$\mu_{\text{k}} = \frac{mg \sin \theta}{mg \cos \theta} = \tan \theta.$$

Put a coin on a book and tilt it until the coin slides at a constant velocity down the book. You might need to tap the book lightly to get the coin to move. Measure the angle of tilt relative to the horizontal and find μ_{k}. Note that the coin does not start to slide at all until an angle greater than θ is attained, since the coefficient of static friction is larger than the coefficient of kinetic friction. Think about how this may affect the value for μ_{k} and its uncertainty.

Atomic-Scale Explanations of Friction

The simpler aspects of friction dealt with so far are its macroscopic (large-scale) characteristics. Great strides have been made in the atomic-scale explanation of friction during the past several decades. Researchers are finding that the atomic nature of friction seems to have several fundamental characteristics. These characteristics not only explain some of the simpler aspects of friction—they also hold the potential for the development of nearly friction-free environments that could save hundreds of billions of dollars in energy which is currently being converted (unnecessarily) into heat.

Figure 8.30 illustrates one macroscopic characteristic of friction that is explained by microscopic (small-scale) research. We have noted that friction is proportional to the normal force, but not to the amount of area in contact, a somewhat counterintuitive notion. When two rough surfaces are in contact, the actual contact area is a tiny fraction of the total area because only high spots touch. When a greater normal force is exerted, the actual contact area increases, and we find that the friction is proportional to this area.

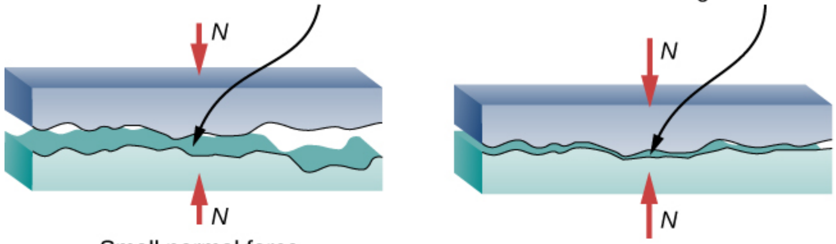


Figure 8.30 Two rough surfaces in contact have a much smaller area of actual contact than their total area. When the normal force is larger as a result of a larger applied force, the area of actual contact increases, as does friction.

However, the atomic-scale view promises to explain far more than the simpler features of friction. The mechanism for how heat is generated is now being determined. In other words, why do surfaces get warmer when rubbed? Essentially, atoms are linked with one another to form lattices. When surfaces rub, the surface atoms adhere and cause atomic lattices to vibrate—essentially creating sound waves that penetrate the material. The sound waves diminish with distance, and their energy is converted into heat. Chemical reactions that are related to frictional wear can also occur between atoms and molecules on the surfaces. **Figure 8.31** shows how the tip of a probe drawn across another material is deformed by atomic-scale friction. The force needed to drag the tip can be measured.

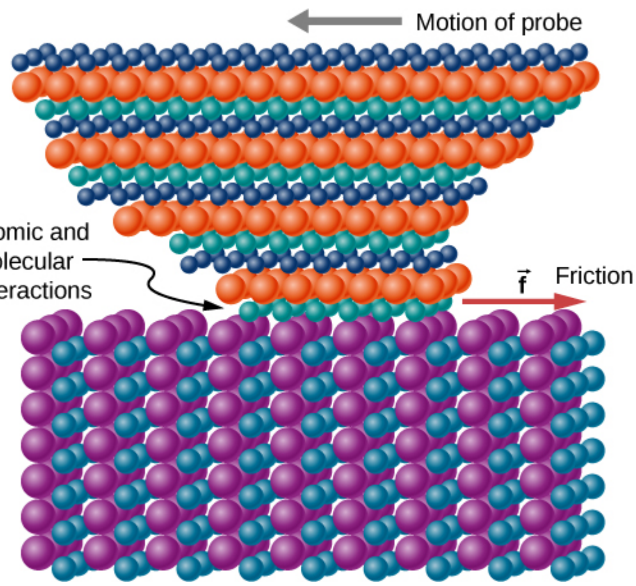


Figure 8.31 The tip of a probe is deformed sideways by frictional force as the probe is dragged across a surface. Measurements of how the force varies for different materials are yielding fundamental insights into the atomic nature of friction.

Check Your Understanding

You can view a **simulation model for friction (https://openstaxcollege.org/l/21friction)** . After viewing it, you should be able to describe friction on a molecular level. Do so. Describe matter in terms of molecular motion. The description should include diagrams to support the description; how the temperature affects the image; what are the differences and similarities between solid, liquid, and gas particle motion; and how the size and speed of gas molecules relate to everyday objects.

Example 8.14

Sliding Blocks

The two blocks of **Figure 8.32** are attached to each other by a massless string that is wrapped around a frictionless pulley. When the bottom 4.00-kg block is pulled to the left by the constant force $\vec{\mathbf{P}}$, the top 2.00-kg block slides across it to the right. Find the magnitude of the force necessary to move the blocks at constant speed. Assume that the coefficient of kinetic friction between all surfaces is 0.400.

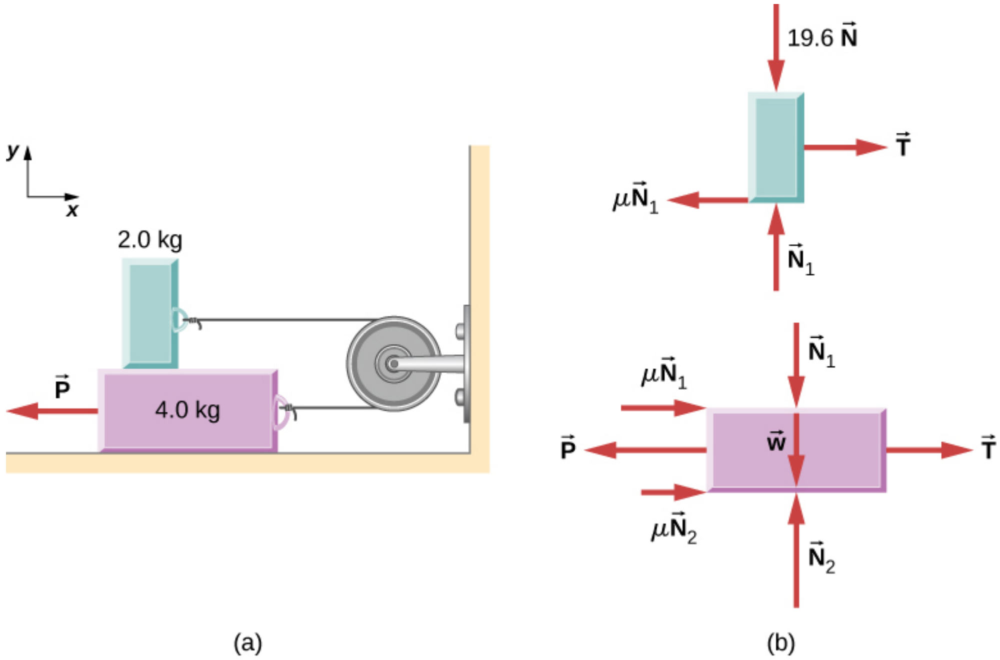


Figure 8.32 (a) Each block moves at constant velocity. (b) Free-body diagrams for the blocks.

Strategy

We analyze the motions of the two blocks separately. The top block is subjected to a contact force exerted by the bottom block. The components of this force are the normal force N_1 and the frictional force $-0.400N_1$. Other forces on the top block are the tension $T\mathrm{i}$ in the string and the weight of the top block itself, 19.6 N. The bottom block is subjected to contact forces due to the top block and due to the floor. The first contact force has components $-N_1$ and $0.400N_1$, which are simply reaction forces to the contact forces that the bottom block exerts on the top block. The components of the contact force of the floor are N_2 and $0.400N_2$. Other forces on this block are $-P$, the tension $T\mathrm{i}$, and the weight –39.2 N.

Solution

Since the top block is moving horizontally to the right at constant velocity, its acceleration is zero in both the horizontal and the vertical directions. From Newton's second law,

$$\begin{aligned} \sum F_x &= m_1 a_x & \qquad \sum F_y &= m_1 a_y \\ T - 0.400N_1 &= 0 & \qquad N_1 - 19.6\,\mathrm{N} &= 0. \end{aligned}$$

Solving for the two unknowns, we obtain $N_1 = 19.6\,\mathrm{N}$ and $T = 0.40N_1 = 7.84\,\mathrm{N}$. The bottom block is also not accelerating, so the application of Newton's second law to this block gives

$$\begin{aligned} &\sum F_x = m_2 a_x & \qquad &\sum F_y = m_2 a_y \\ &T - P + 0.400\,N_1 + 0.400\,N_2 = 0 & \qquad &N_2 - 39.2\,\mathrm{N} - N_1 = 0. \end{aligned}$$

The values of N_1 and T were found with the first set of equations. When these values are substituted into the second set of equations, we can determine N_2 and P. They are

$$N_2 = 58.8\text{ N} \text{ and } P = 39.2\text{ N}.$$

Significance

Understanding what direction in which to draw the friction force is often troublesome. Notice that each friction force labeled in **Figure 8.32** acts in the direction opposite the motion of its corresponding block.

Example 8.15

A Crate on an Accelerating Truck

A 50.0-kg crate rests on the bed of a truck as shown in **Figure 8.33**. The coefficients of friction between the surfaces are $\mu_k = 0.300$ and $\mu_s = 0.400$. Find the frictional force on the crate when the truck is accelerating forward relative to the ground at (a) 2.00 m/s^2, and (b) 5.00 m/s^2.

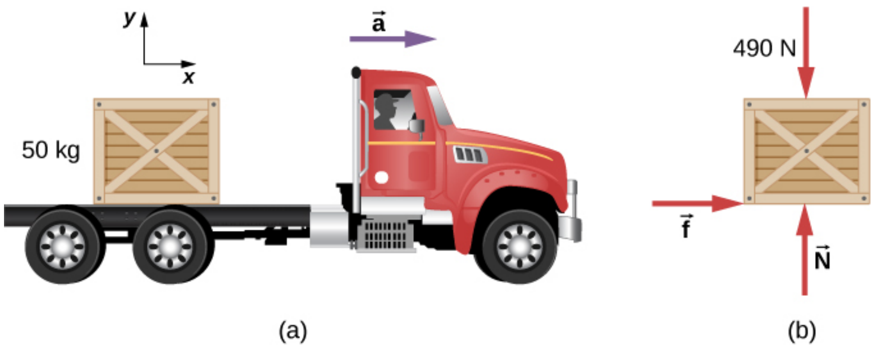


Figure 8.33 (a) A crate rests on the bed of the truck that is accelerating forward. (b) The free-body diagram of the crate.

Strategy

The forces on the crate are its weight and the normal and frictional forces due to contact with the truck bed. We start by *assuming* that the crate is not slipping. In this case, the static frictional force f_s acts on the crate. Furthermore, the accelerations of the crate and the truck are equal.

Solution

a. Application of Newton's second law to the crate, using the reference frame attached to the ground, yields

$$\begin{aligned} \sum F_x &= ma_x & \sum F_y &= ma_y \\ f_s &= (50.0\text{ kg})(2.00\text{ m/s}^2) & N - 4.90 \times 10^2\text{ N} &= (50.0\text{ kg})(0) \\ &= 1.00 \times 10^2\text{ N} & N &= 4.90 \times 10^2\text{ N}. \end{aligned}$$

We can now check the validity of our no-slip assumption. The maximum value of the force of static friction is

$$\mu_s N = (0.400)(4.90 \times 10^2\text{ N}) = 196\text{ N},$$

whereas the *actual* force of static friction that acts when the truck accelerates forward at 2.00 m/s^2 is only $1.00 \times 10^2\text{ N}$. Thus, the assumption of no slipping is valid.

b. If the crate is to move with the truck when it accelerates at 5.0 m/s^2, the force of static friction must be

$$f_s = ma_x = (50.0\text{ kg})(5.00\text{ m/s}^2) = 250\text{ N}.$$

Since this exceeds the maximum of 196 N, the crate must slip. The frictional force is therefore kinetic and is

$$f_k = \mu_k N = (0.300)(4.90 \times 10^2\text{ N}) = 147\text{ N}.$$

The horizontal acceleration of the crate relative to the ground is now found from

$$\begin{aligned} \sum F_x &= ma_x \\ 147\text{ N} &= (50.0\text{ kg})a_x, \\ \text{so } a_x &= 2.94\text{ m/s}^2. \end{aligned}$$

Significance

Relative to the ground, the truck is accelerating forward at 5.0 m/s^2 and the crate is accelerating forward at 2.94 m/s^2. Hence the crate is sliding backward relative to the bed of the truck with an acceleration $2.94\text{ m/s}^2 - 5.00\text{ m/s}^2 = -2.06\text{ m/s}^2$.

Example 8.16

Snowboarding

Earlier, we analyzed the situation of a downhill skier moving at constant velocity to determine the coefficient of kinetic friction. Now let's do a similar analysis to determine acceleration. The snowboarder of **Figure 8.34** glides down a slope that is inclined at $\theta = 13^0$ to the horizontal. The coefficient of kinetic friction between the board and the snow is $\mu_k = 0.20$. What is the acceleration of the snowboarder?

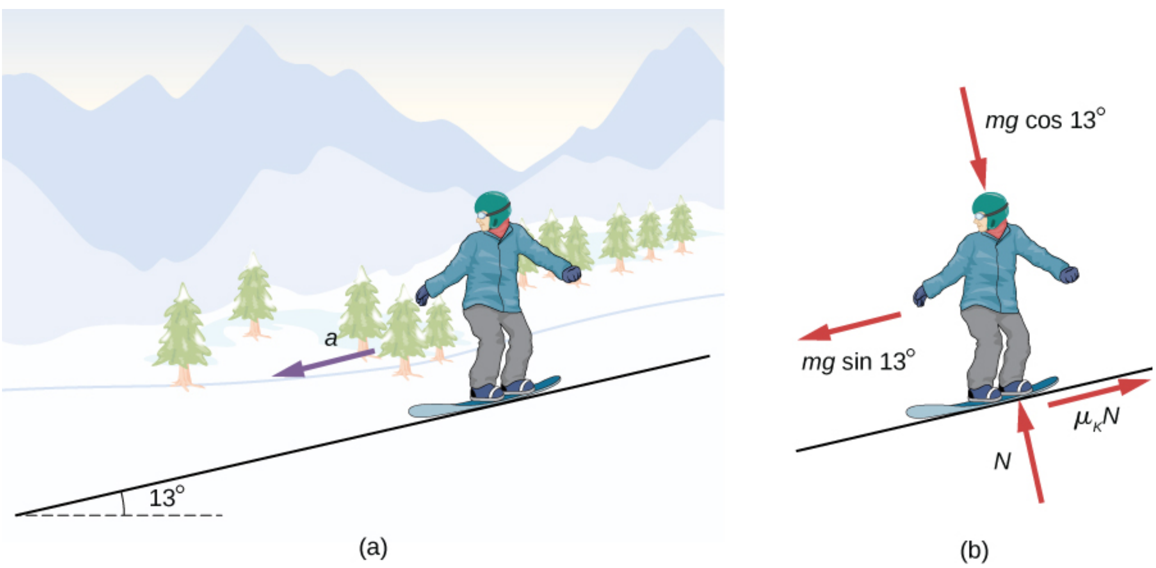


Figure 8.34 (a) A snowboarder glides down a slope inclined at 13° to the horizontal. (b) The free-body diagram of the snowboarder.

Strategy

The forces acting on the snowboarder are her weight and the contact force of the slope, which has a component normal to the incline and a component along the incline (force of kinetic friction). Because she moves along the slope, the most convenient reference frame for analyzing her motion is one with the *x*-axis along and the *y*-axis perpendicular to the incline. In this frame, both the normal and the frictional forces lie along coordinate axes, the components of the weight are $mg \sin \theta$ along the slope and $mg \cos \theta$ at right angles into the slope, and the only acceleration is along the *x*-axis $(a_y = 0)$.

Solution

We can now apply Newton's second law to the snowboarder:

$$\begin{aligned} \sum F_x &= ma_x & \sum F_y &= ma_y \\ mg \sin \theta - \mu_k N &= ma_x & N - mg \cos \theta &= m(0). \end{aligned}$$

From the second equation, $N = mg \cos \theta$. Upon substituting this into the first equation, we find

$$\begin{aligned} a_x &= g(\sin \theta - \mu_\text{k} \cos \theta) \\ &= g(\sin 13° - 0.20 \cos 13°) = 0.29 \text{ m/s}^2. \end{aligned}$$

Significance

Notice from this equation that if θ is small enough or μ_k is large enough, a_x is negative, that is, the snowboarder slows down.

8.10 Check Your Understanding The snowboarder is now moving down a hill with incline $10.0°$. What is the skier's acceleration?

8.9 | Centripetal Force

Learning Objectives

By the end of the section, you will be able to:

- Explain the equation for centripetal acceleration
- Apply Newton's second law to develop the equation for centripetal force
- Use circular motion concepts in solving problems involving Newton's laws of motion

In **Motion in Two and Three Dimensions**, we examined the basic concepts of circular motion. An object undergoing circular motion, like a race car going around a corner, must be accelerating because it is changing the direction of its velocity. We proved that this centrally directed acceleration, called centripetal acceleration, is given by the formula

$$a_\text{c} = \frac{v^2}{r}$$

where *v* is the velocity of the object, directed along a tangent line to the curve at any instant. If we know the angular velocity ω, then we can use

$$a_\text{c} = r\omega^2.$$

Angular velocity gives the rate at which the object is turning through the curve, in units of rad/s. This acceleration acts along the radius of the curved path and is thus also referred to as a radial acceleration.

An acceleration must be produced by a force. Any force or combination of forces can cause a centripetal or radial acceleration. Just a few examples are the tension in the rope on a tether ball, the force of Earth's gravity on the Moon, friction between roller skates and a rink floor, a banked roadway's force on a car, and forces on the tube of a spinning centrifuge. Any net force causing uniform circular motion is called a **centripetal force**. The direction of a centripetal force is toward the center of curvature, the same as the direction of centripetal acceleration. According to Newton's second law of motion, net force is mass times acceleration: $F_\text{net} = ma$. For uniform circular motion, the acceleration is the centripetal acceleration: $a = a_c$. Thus, the magnitude of centripetal force F_c is

$$F_c = ma_c.$$

By substituting the expressions for centripetal acceleration a_c $(a_c = \frac{v^2}{r}; a_c = r\omega^2)$, we get two expressions for the centripetal force F_c in terms of mass, velocity, angular velocity, and radius of curvature:

$$F_c = m\frac{v^2}{r}; \quad F_c = mr\omega^2. \tag{8.14}$$

You may use whichever expression for centripetal force is more convenient. Centripetal force $\vec{\mathbf{F}}_c$ is always perpendicular to the path and points to the center of curvature, because $\vec{\mathbf{a}}_c$ is perpendicular to the velocity and points to the center of curvature. Note that if you solve the first expression for *r*, you get

$$r = \frac{mv^2}{F_c}.$$

This implies that for a given mass and velocity, a large centripetal force causes a small radius of curvature—that is, a tight curve, as in **Figure 8.35**.

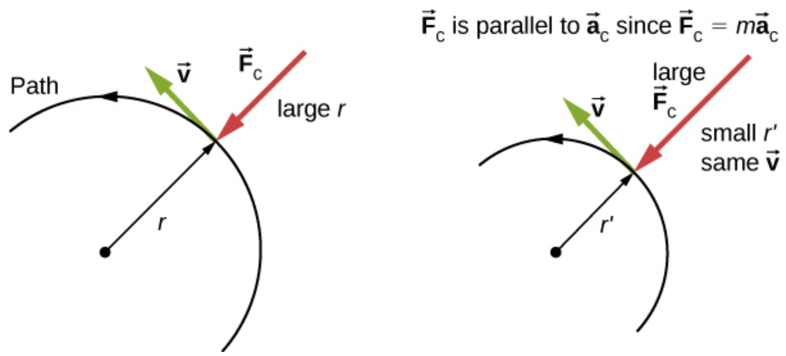


Figure 8.35 The frictional force supplies the centripetal force and is numerically equal to it. Centripetal force is perpendicular to velocity and causes uniform circular motion. The larger the F_c, the smaller the radius of curvature *r* and the sharper the curve. The second curve has the same *v*, but a larger F_c produces a smaller *r'*.

Example 8.17

What Coefficient of Friction Do Cars Need on a Flat Curve?

(a) Calculate the centripetal force exerted on a 900.0-kg car that negotiates a 500.0-m radius curve at 25.00 m/s.
(b) Assuming an unbanked curve, find the minimum static coefficient of friction between the tires and the road, static friction being the reason that keeps the car from slipping (**Figure 8.36**).

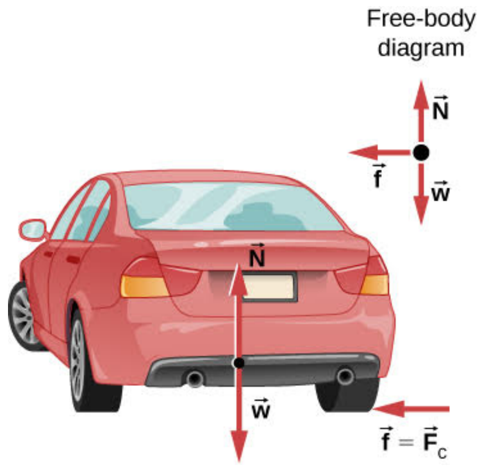


Figure 8.36 This car on level ground is moving away and turning to the left. The centripetal force causing the car to turn in a circular path is due to friction between the tires and the road. A minimum coefficient of friction is needed, or the car will move in a larger-radius curve and leave the roadway.

Strategy

a. We know that $F_{\text{c}} = \frac{mv^2}{r}$. Thus,

$$F_{\text{c}} = \frac{mv^2}{r} = \frac{(900.0\text{ kg})(25.00\text{ m/s})^2}{(500.0\text{ m})} = 1125\text{ N}.$$

b. **Figure 8.36** shows the forces acting on the car on an unbanked (level ground) curve. Friction is to the left, keeping the car from slipping, and because it is the only horizontal force acting on the car, the friction is the centripetal force in this case. We know that the maximum static friction (at which the tires roll but do not slip) is $\mu_{\text{s}}N$, where μ_{s} is the static coefficient of friction and N is the normal force. The normal force equals the car's weight on level ground, so $N = mg$. Thus the centripetal force in this situation is

$$F_{\text{c}} \equiv f = \mu_{\text{s}}N = \mu_{\text{s}}mg.$$

Now we have a relationship between centripetal force and the coefficient of friction. Using the equation

$$F_{\text{c}} = m\frac{v^2}{r},$$

we obtain

$$m\frac{v^2}{r} = \mu_{\text{s}}mg.$$

We solve this for μ_{s}, noting that mass cancels, and obtain

$$\mu_{\text{s}} = \frac{v^2}{rg}.$$

Substituting the knowns,

$$\mu_{\text{s}} = \frac{(25.00\text{ m/s})^2}{(500.0\text{ m})(9.80\text{ m/s}^2)} = 0.13.$$

(Because coefficients of friction are approximate, the answer is given to only two digits.)

Significance

The coefficient of friction found in **Figure 8.36**(b) is much smaller than is typically found between tires and roads. The car still negotiates the curve if the coefficient is greater than 0.13, because static friction is a responsive force, able to assume a value less than but no more than $\mu_s N$. A higher coefficient would also allow the car to negotiate the curve at a higher speed, but if the coefficient of friction is less, the safe speed would be less than 25 m/s. Note that mass cancels, implying that, in this example, it does not matter how heavily loaded the car is to negotiate the turn. Mass cancels because friction is assumed proportional to the normal force, which in turn is proportional to mass. If the surface of the road were banked, the normal force would be less, as discussed next.

8.11 Check Your Understanding A car moving at 96.8 km/h travels around a circular curve of radius 182.9 m on a flat country road. What must be the minimum coefficient of static friction to keep the car from slipping?

8.10 | Newton's Law of Universal Gravitation

Learning Objectives

By the end of this section, you will be able to:

- List the significant milestones in the history of gravitation
- Calculate the gravitational force between two point masses
- Estimate the gravitational force between collections of mass

Newton's three laws provide a fundamental basis for explaining all motion in the Universe. And, with the addition of a fourth law, his system answered the two most salient questions of his time. Namely,

1. Why do objects near the Earth's surface all accelerate downward at a rate of 9.8 m/s^2?
2. Why do the planets move around the Sun according to Kepler's three laws?

Before he could apply his system to answer these questions, he needed to understand, and explain, the force of gravity.

The History of Gravitation

The earliest philosophers wondered why objects naturally tend to fall toward the ground. Aristotle (384–322 BCE) believed that it was the nature of rocks to seek Earth and the nature of fire to seek the Heavens. Brahmagupta (598~665 CE) postulated that Earth was a sphere and that objects possessed a natural affinity for it, falling toward the center from wherever they were located.

The motions of the Sun, our Moon, and the planets have been studied for thousands of years as well. These motions were described with amazing accuracy by Ptolemy (90–168 CE), whose method of epicycles described the paths of the planets as circles within circles. However, there is little evidence that anyone connected the motion of astronomical bodies with the motion of objects falling to Earth—until the seventeenth century.

Nicolaus Copernicus (1473–1543) is generally credited as being the first to challenge Ptolemy's geocentric (Earth-centered) system and suggest a heliocentric system, in which the Sun is at the center of the solar system. This idea was supported by the incredibly precise naked-eye measurements of planetary motions by Tycho Brahe and their analysis by Johannes Kepler and Galileo Galilei. Kepler showed that the motion of each planet is an ellipse (the first of his three laws, discussed in **Kepler's Laws of Planetary Motion**), and Robert Hooke intuitively suggested that these motions are due to the planets being attracted to the Sun. However, it was Isaac Newton who connected the acceleration of objects near Earth's surface with the centripetal acceleration of the Moon in its orbit about Earth.

Newton's Law of Universal Gravitation

Newton noted that objects at Earth's surface (hence at a distance of R_E from the center of Earth) have an acceleration of *g*, but the Moon, at a distance of about $60\,R_E$, has a centripetal acceleration about $(60)^2$ times smaller than *g*. He could explain this by postulating that a force exists between any two objects, whose magnitude is given by the product of the two masses divided by the square of the distance between them. We now know that this inverse square law is ubiquitous in nature, a function of geometry for point sources. The strength of any source at a distance *r* is spread over the surface of a sphere centered about the mass. The surface area of that sphere is proportional to r^2. In later chapters, we see this same form in the electromagnetic force.

Newton's Law of Gravitation

Newton's law of gravitation can be expressed as

$$\vec{\mathbf{F}}_{12} = G\frac{m_1 m_2}{r^2}\hat{\mathbf{r}}_{12} \tag{8.15}$$

where $\vec{\mathbf{F}}_{12}$ is the force on object 1 exerted by object 2 and $\hat{\mathbf{r}}_{12}$ is a unit vector that points from object 1 toward object 2.

As shown in **Figure 8.37**, the $\vec{\mathbf{F}}_{12}$ vector points from object 1 toward object 2, and hence represents an attractive force between the objects. The equal but opposite force $\vec{\mathbf{F}}_{21}$ is the force on object 2 exerted by object 1.

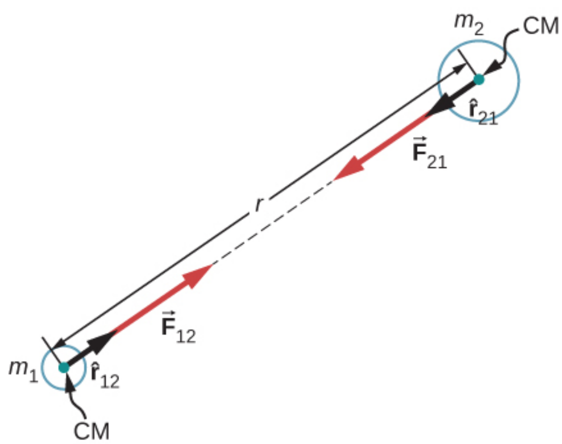


Figure 8.37 Gravitational force acts along a line joining the centers of mass of two objects.

These equal but opposite forces reflect Newton's third law, which we discussed earlier. Note that strictly speaking, **Equation 8.15** applies to point masses—all the mass is located at one point. But it applies equally to any spherically symmetric objects, where *r* is the distance between the centers of mass of those objects. In many cases, it works reasonably well for nonsymmetrical objects, if their separation is large compared to their size, and we take *r* to be the distance between the center of mass of each body.

The Cavendish Experiment

A century after Newton published his law of universal gravitation, Henry Cavendish determined the proportionality constant *G* by performing a painstaking experiment. He constructed a device similar to that shown in **Figure 8.38**, in which small masses are suspended from a wire. Once in equilibrium, two fixed, larger masses are placed symmetrically near the smaller ones. The gravitational attraction creates a torsion (twisting) in the supporting wire that can be measured.

The constant *G* is called the **universal gravitational constant** and Cavendish determined it to be $G = 6.67 \times 10^{-11}\ \text{N} \cdot \text{m}^2/\text{kg}^2$. The word 'universal' indicates that scientists think that this constant applies to masses of any composition and that it is the same throughout the Universe. The value of *G* is an incredibly small number, showing that the force of gravity is very weak. The attraction between masses as small as our bodies, or even objects the size of skyscrapers, is incredibly small. For example, two 1.0-kg masses located 1.0 meter apart exert a force of 6.7×10^{-11} N on each other. This is the weight of a typical grain of pollen.

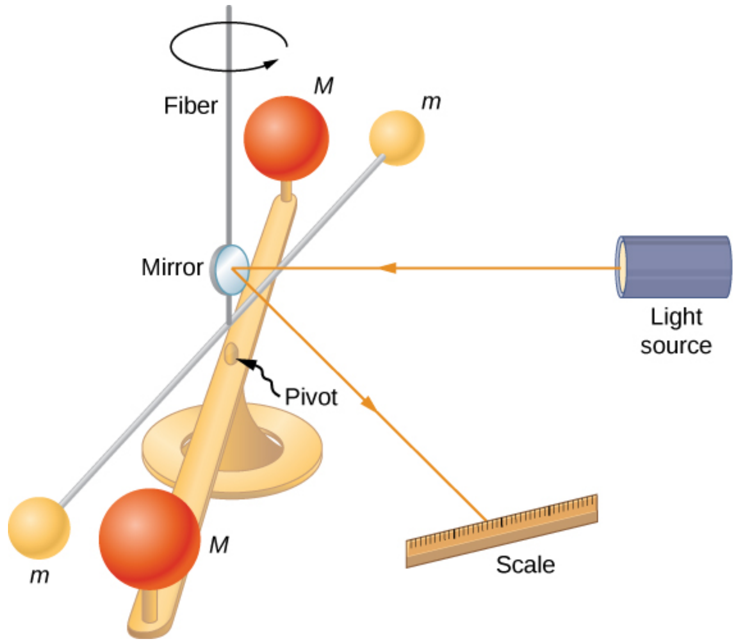


Figure 8.38 Cavendish used an apparatus similar to this to measure the gravitational attraction between two spheres (m) suspended from a wire and two stationary spheres (M). This is a common experiment performed in undergraduate laboratories, but it is quite challenging. Passing trucks outside the laboratory can create vibrations that overwhelm the gravitational forces.

Although gravity is the weakest of the four fundamental forces of nature, its attractive nature is what holds us to Earth, causes the planets to orbit the Sun and the Sun to orbit our galaxy, and binds galaxies into clusters, ranging from a few to millions. Gravity is the force that forms the Universe.

Problem-Solving Strategy: Newton's Law of Gravitation

To determine the motion caused by the gravitational force, follow these steps:

1. Identify the two masses, one or both, for which you wish to find the gravitational force.
2. Draw a free-body diagram, sketching the force acting on each mass and indicating the distance between their centers of mass.
3. Apply Newton's second law of motion to each mass to determine how it will move.

Example 8.18

A Collision in Orbit

Consider two nearly spherical *Soyuz* payload vehicles, in orbit about Earth, each with mass 9000 kg and diameter 4.0 m. They are initially at rest relative to each other, 10.0 m from center to center. (As we saw in **Kepler's Laws of Planetary Motion**, both orbit Earth at the same speed and interact nearly the same as if they were isolated in deep space.) Determine the gravitational force between them and their initial acceleration. Estimate how long it takes for them to drift together, and how fast they are moving upon impact.

Strategy

We use Newton's law of gravitation to determine the force between them and then use Newton's second law to find the acceleration of each. For the *estimate*, we assume this acceleration is constant, and we use the constant-acceleration equations from **Motion along a Straight Line** to find the time and speed of the collision.

Solution

The magnitude of the force is

$$\left|\vec{\mathbf{F}}_{12}\right| = F_{12} = G\frac{m_1 m_2}{r^2} = 6.67 \times 10^{-11}\ \text{N} \cdot \text{m}^2/\text{kg}^2 \frac{(9000\ \text{kg})(9000\ \text{kg})}{(10\ \text{m})^2} = 5.4 \times 10^{-5}\,\text{N}.$$

The initial acceleration of each payload is

$$a = \frac{F}{m} = \frac{5.4 \times 10^{-5}\ \text{N}}{9000\ \text{kg}} = 6.0 \times 10^{-9}\,\text{m/s}^2.$$

The vehicles are 4.0 m in diameter, so the vehicles move from 10.0 m to 4.0 m apart, or a distance of 3.0 m each. A similar calculation to that above, for when the vehicles are 4.0 m apart, yields an acceleration of $3.8 \times 10^{-8}\ \text{m/s}^2$, and the average of these two values is $2.2 \times 10^{-8}\ \text{m/s}^2$. If we assume a constant acceleration of this value and they start from rest, then the vehicles collide with speed given by

$$v^2 = v_0^2 + 2a(x - x_0), \ \text{where}\ v_0 = 0,$$

so

$$v = \sqrt{2(2.2 \times 10^{-9}\ \text{N})(3.0\ \text{m})} = 3.6 \times 10^{-4}\ \ \text{m/s}.$$

We use $v = v_0 + at$ to find $t = v/a = 1.7 \times 10^4\ \text{s}$ or about 4.6 hours.

Significance

These calculations—including the initial force—are only estimates, as the vehicles are probably not spherically symmetrical. But you can see that the force is incredibly small. Astronauts must tether themselves when doing work outside even the massive International Space Station (ISS), as in **Figure 8.39**, because the gravitational attraction cannot save them from even the smallest push away from the station.

Figure 8.39 This photo shows Ed White tethered to the Space Shuttle during a spacewalk. (credit: NASA)

8.12 Check Your Understanding What happens to force and acceleration as the vehicles fall together? What will our estimate of the velocity at a collision higher or lower than the speed actually be? And finally, what would happen if the masses were not identical? Would the force on each be the same or different? How about their accelerations?

The effect of gravity between two objects with masses on the order of these space vehicles is indeed small. Yet, the effect of gravity on you from Earth is significant enough that a fall into Earth of only a few feet can be dangerous. We examine the force of gravity near Earth's surface in the next section.

Example 8.19

Attraction between Galaxies

Find the acceleration of our galaxy, the Milky Way, due to the nearest comparably sized galaxy, the Andromeda galaxy (**Figure 8.40**). The approximate mass of each galaxy is 800 billion solar masses (a solar mass is the mass of our Sun), and they are separated by 2.5 million light-years. (Note that the mass of Andromeda is not so well known but is believed to be slightly larger than our galaxy.) Each galaxy has a diameter of roughly 100,000 light-years $(1 \text{ light-year} = 9.5 \times 10^{15}\ \text{m})$.

Figure 8.40 Galaxies interact gravitationally over immense distances. The Andromeda galaxy is the nearest spiral galaxy to the Milky Way, and they will eventually collide. (credit: Boris Štromar)

Strategy

As in the preceding example, we use Newton's law of gravitation to determine the force between them and then use Newton's second law to find the acceleration of the Milky Way. We can consider the galaxies to be point masses, since their sizes are about 25 times smaller than their separation. The mass of the Sun (see **Appendix D**) is $2.0 \times 10^{30}\ \text{kg}$ and a light-year is the distance light travels in one year, $9.5 \times 10^{15}\ \text{m}$.

Solution

The magnitude of the force is

$$F_{12} = G\frac{m_1 m_2}{r^2} = (6.67 \times 10^{-11}\ \text{N} \cdot \text{m}^2/\text{kg}^2)\frac{[(800 \times 10^9)(2.0 \times 10^{30}\ \text{kg})]^2}{[(2.5 \times 10^6)(9.5 \times 10^{15}\ \text{m})]^2} = 3.0 \times 10^{29}\ \text{N}.$$

The acceleration of the Milky Way is

$$a = \frac{F}{m} = \frac{3.0 \times 10^{29}\ \text{N}}{(800 \times 10^9)(2.0 \times 10^{30}\ \text{kg})} = 1.9 \times 10^{-13}\ \text{m/s}^2.$$

Significance

Does this value of acceleration seem astoundingly small? If they start from rest, then they would accelerate directly toward each other, "colliding" at their center of mass. Let's estimate the time for this to happen. The initial acceleration is $\sim 10^{-13}\ \text{m/s}^2$, so using $v = at$, we see that it would take $\sim 10^{13}\ \text{s}$ for each galaxy to reach a speed of 1.0 m/s, and they would be only $\sim 0.5 \times 10^{13}\ \text{m}$ closer. That is nine orders of magnitude smaller than the initial distance between them. In reality, such motions are rarely simple. These two galaxies, along with about 50 other smaller galaxies, are all gravitationally bound into our local cluster. Our local cluster is gravitationally bound to other clusters in what is called a supercluster. All of this is part of the great cosmic dance that results from gravitation, as shown in **Figure 8.41**.

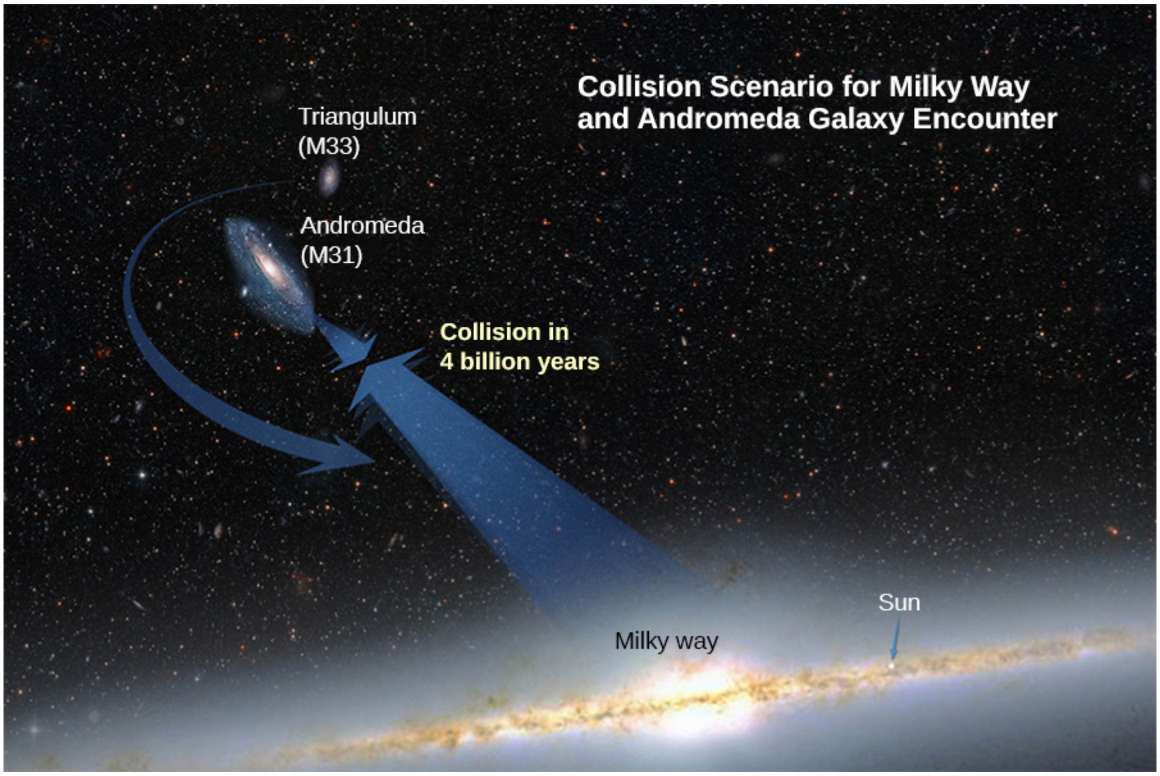


Figure 8.41 Based on the results of this example, plus what astronomers have observed elsewhere in the Universe, our galaxy will collide with the Andromeda Galaxy in about 4 billion years. (credit: NASA)

8.11 | The Newtonian Synthesis

Learning Objectives

By the end of this section, you will be able to:

- Explain the connection between the constants G and g
- Determine the mass of an astronomical body from free-fall acceleration at its surface
- Describe how the value of g varies due to location and Earth's rotation

We are ready to put all of the pieces together, as Newton did, in his grand synthesis. First, we will observe how Newton's law of gravitation applies at the surface of a planet and how it connects with what we learned earlier about free fall. Then, we will see how applying Newton's laws to the orbital motion of satellites yields Kepler's laws of planetary motion.

Weight - Gravitation Near the Earth's Surface

Recall that the acceleration of a free-falling object near Earth's surface is approximately $g = 9.80\ \text{m/s}^2$. The force causing this acceleration is called the weight of the object, and from Newton's second law, it has the value *mg*. This weight is present regardless of whether the object is in free fall. We now know that this force is the gravitational force between the object and Earth. If we substitute *mg* for the magnitude of $\vec{\mathbf{F}}_{12}$ in Newton's law of universal gravitation, *m* for m_1, and M_E for m_2, we obtain the scalar equation

$$mg = G\frac{mM_E}{r^2}$$

where *r* is the distance between the centers of mass of the object and Earth. For objects within a few kilometers of Earth's surface, we can take $r = R_E$ (see **Figure 8.420**). The mass *m* of the object cancels, and if we use the values from **Appendix D**, this gives us

$$g = G\frac{M_E}{R_E^2} = \frac{\left(6.674\times10^{-11}\,\frac{\text{N}\cdot\text{m}^2}{\text{kg}^2}\right)\left(5.974\times10^{24}\,\text{kg}\right)}{\left(6.378\times10^{6}\,\text{m}\right)^2} = 9.801\frac{\text{N}}{\text{kg}} = 9.801\frac{\text{m}}{\text{s}^2} \tag{8.16}$$

This explains why all masses free fall with the same acceleration. We have ignored the fact that Earth also accelerates toward the falling object, but that is acceptable as long as the mass of Earth is much larger than that of the object. So, here is the first part of the Newtonian synthesis - now we know where the 9.8 m/s^2 comes from.

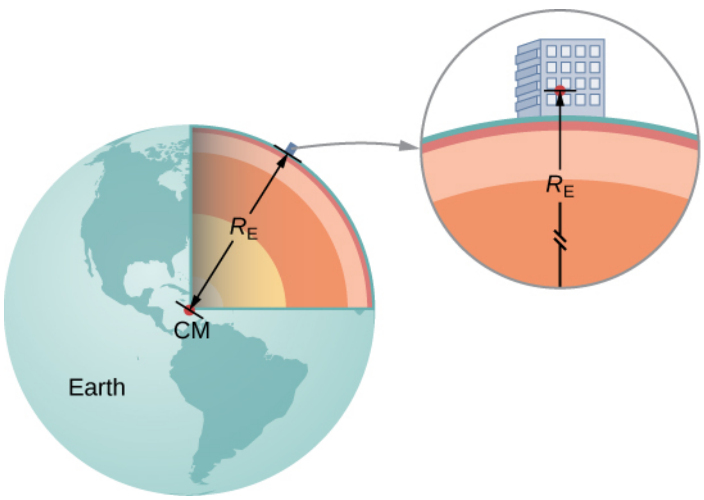


Figure 8.42 We can take the distance between the centers of mass of Earth and an object on its surface to be the radius of Earth, provided that its size is much less than the radius of Earth.

Example 8.20

Masses of Earth and Moon

Have you ever wondered how we know the mass of Earth? We certainly can't place it on a scale. The values of *g* and the radius of Earth were measured with reasonable accuracy centuries ago.

a. Use the standard values of *g*, R_E, and **Equation 8.16** to find the mass of Earth.

b. Estimate the value of *g* on the Moon. Use the fact that the Moon has a radius of about 1700 km (a value of this accuracy was determined many centuries ago) and assume it has the same average density as Earth,

$5500\,\text{kg/m}^3$.

Strategy

With the known values of g and R_E, we can use **Equation 8.16** to find M_E. For the Moon, we use the assumption of equal average density to determine the mass from a ratio of the volumes of Earth and the Moon.

Solution

a. Rearranging **Equation 8.16**, we have

$$M_E = \frac{gR_E^2}{G} = \frac{9.80\,\text{m/s}^2(6.37 \times 10^6\ \text{m})^2}{6.67 \times 10^{-11}\ \text{N} \cdot \text{m}^2/\text{kg}^2} = 5.95 \times 10^{24}\ \text{kg}.$$

b. The volume of a sphere is proportional to the radius cubed, so a simple ratio gives us

$$\frac{M_M}{M_E} = \frac{R_M^3}{R_E^3} \rightarrow M_M = \left(\frac{(1.7 \times 10^6\ \text{m})^3}{(6.37 \times 10^6\ \text{m})^3}\right)(5.95 \times 10^{24}\ \text{kg}) = 1.1 \times 10^{23}\ \text{kg}.$$

We now use **Equation 8.16**.

$$g_M = G\frac{M_M}{r_M^2} = (6.67 \times 10^{-11}\ \text{N} \cdot \text{m}^2/\text{kg}^2)\frac{(1.1 \times 10^{23}\ \text{kg})}{(1.7 \times 10^6\ \text{m})^2} = 2.5\ \text{m/s}^2$$

Significance

As soon as Cavendish determined the value of G in 1798, the mass of Earth could be calculated. (In fact, that was the ultimate purpose of Cavendish's experiment in the first place.) The value we calculated for g of the Moon is incorrect. The average density of the Moon is actually only $3340\,\text{kg/m}^3$ and $g = 1.6\,\text{m/s}^2$ at the surface. Newton attempted to measure the mass of the Moon by comparing the effect of the Sun on Earth's ocean tides compared to that of the Moon. His value was a factor of two too small. The most accurate values for g and the mass of the Moon come from tracking the motion of spacecraft that have orbited the Moon. But the mass of the Moon can actually be determined accurately without going to the Moon. Earth and the Moon orbit about a common center of mass, and careful astronomical measurements can determine that location. The ratio of the Moon's mass to Earth's is the ratio of [the distance from the common center of mass to the Moon's center] to [the distance from the common center of mass to Earth's center].

Later in this chapter, we will see that the mass of other astronomical bodies also can be determined by the period of small satellites orbiting them. But until Cavendish determined the value of G, the masses of all these bodies were unknown.

Example 8.21

Gravity above Earth's Surface

What is the value of g 400 km above Earth's surface, where the International Space Station is in orbit?

Strategy

Using the value of M_E and noting the radius is $r = R_E + 400\,\text{km}$, we use **Equation 8.16** to find g.

From **Equation 8.16** we have

$$g = G\frac{M_E}{r^2} = 6.67 \times 10^{-11}\ \text{N} \cdot \text{m}^2/\text{kg}^2 \frac{5.96 \times 10^{24}\ \text{kg}}{(6.37 \times 10^6 + 400 \times 10^3\ \text{m})^2} = 8.67\ \text{m/s}^2.$$

Significance

We often see video of astronauts in space stations, apparently weightless. But clearly, the force of gravity is acting on them. Comparing the value of g we just calculated to that on Earth $(9.80\,\text{m/s}^2)$, we see that the astronauts in the International Space Station still have 88% of their weight. They only appear to be weightless because they are in free fall.

8.13 Check Your Understanding How does your weight at the top of a tall building compare with that on the first floor? Do you think engineers need to take into account the change in the value of g when designing structural support for a very tall building?

Planetary Orbits - Gravitation in Outer Space

The Moon orbits Earth. In turn, Earth and the other planets orbit the Sun. The space directly above our atmosphere is filled with artificial satellites in orbit. We examine the simplest of these orbits, the circular orbit, to understand the relationship between the speed and period of planets and satellites in relation to their positions and the bodies that they orbit.

As noted at the beginning of this chapter, Nicolaus Copernicus first suggested that Earth and all other planets orbit the Sun in circles. He further noted that orbital periods increased with distance from the Sun. Later analysis by Kepler showed that these orbits are actually ellipses, but the orbits of most planets in the solar system are nearly circular. Earth's orbital distance from the Sun varies a mere 2%. The exception is the eccentric orbit of Mercury, whose orbital distance varies nearly 40%.

Determining the **orbital speed** and **orbital period** of a satellite is much easier for circular orbits, so we make that assumption in the derivation that follows. We focus on objects orbiting Earth, but our results can be generalized for other cases.

Consider a satellite of mass m in a circular orbit about Earth at distance r from the center of Earth (Figure 8.43). It has centripetal acceleration directed toward the center of Earth. Earth's gravity is the only force acting, so Newton's second law gives

$$\frac{GmM_E}{r^2} = ma_c = \frac{mv_{orbit}^2}{r}.$$

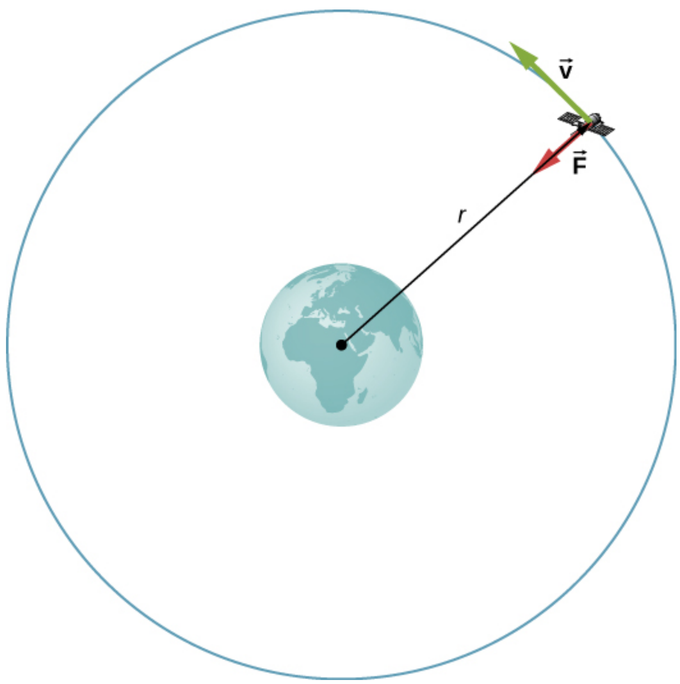


Figure 8.43 A satellite of mass m orbiting at radius r from the center of Earth. The gravitational force supplies the centripetal acceleration.

We solve for the speed of the orbit, noting that m cancels, to get the orbital speed

$$v_{\text{orbit}} = \sqrt{\frac{GM_E}{r}}. \quad (8.17)$$

Note that the value of *m* does not appear in **Equation 8.17**. The value of the orbital velocity depends only upon the distance from the center of the planet, and *not* upon the mass of the object being acted upon.

To find the period of a circular orbit, we note that the satellite travels the circumference of the orbit $2\pi r$ in one period *T*. Using the definition of speed, we have $v_{\text{orbit}} = 2\pi r/T$. We substitute this into **Equation 8.17** and rearrange to get

$$T = 2\pi\sqrt{\frac{r^3}{GM_E}}. \quad (8.18)$$

Referring to **Section 7.4** this is Kepler's third law for the case of circular orbits. It also confirms Copernicus's observation that the period of a planet increases with increasing distance from the Sun. We need only replace M_E with M_{Sun} in **Equation 8.18**.

We conclude this section by returning to our earlier discussion about astronauts in orbit appearing to be weightless, as if they were free-falling towards Earth. In fact, they are in free fall. Consider the trajectories shown in **Figure 8.44**. (This figure is based on a drawing by Newton in his *Principia* and also appeared earlier in **Motion in Two and Three Dimensions**.) All the trajectories shown that hit the surface of Earth have less than orbital velocity. The astronauts would accelerate toward Earth along the noncircular paths shown and feel weightless. (Astronauts actually train for life in orbit by riding in airplanes that free fall for 30 seconds at a time.) But with the correct orbital velocity, Earth's surface curves away from them at exactly the same rate as they fall toward Earth. Of course, staying the same distance from the surface is the point of a circular orbit.

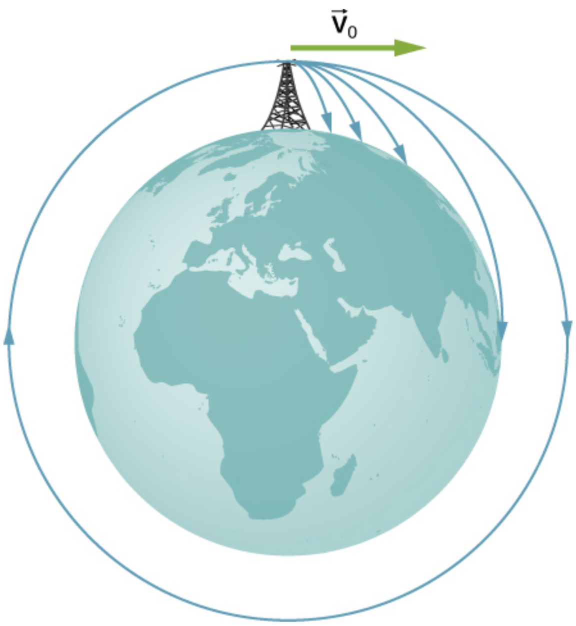


Figure 8.44 A circular orbit is the result of choosing a tangential velocity such that Earth's surface curves away at the same rate as the object falls toward Earth.

Example 8.22

The International Space Station

Determine the orbital speed and period for the International Space Station (ISS).

Strategy

Since the ISS orbits $4.00 \times 10^2\,\text{km}$ above Earth's surface, the radius at which it orbits is $R_\text{E} + 4.00 \times 10^2\,\text{km}$. We use **Equation 8.17** and **Equation 8.18** to find the orbital speed and period, respectively.

Solution

Using **Equation 8.17**, the orbital velocity is

$$v_\text{orbit} = \sqrt{\frac{GM_\text{E}}{r}} = \sqrt{\frac{6.67 \times 10^{-11}\ \text{N} \cdot \text{m}^2/\text{kg}^2(5.96 \times 10^{24}\ \text{kg})}{(6.36 \times 10^6 + 4.00 \times 10^5\ \text{m})}} = 7.67 \times 10^3\ \text{m/s}$$

which is about 17,000 mph. Using **Equation 8.18**, the period is

$$T = 2\pi\sqrt{\frac{r^3}{GM_\text{E}}} = 2\pi\sqrt{\frac{(6.37 \times 10^6 + 4.00 \times 10^5\ \text{m})^3}{(6.67 \times 10^{-11}\ \text{N} \cdot \text{m}^2/\text{kg}^2)(5.96 \times 10^{24}\ \text{kg})}} = 5.55 \times 10^3\ \text{s}$$

which is just over 90 minutes.

Significance

The ISS is considered to be in low Earth orbit (LEO). Nearly all satellites are in LEO, including most weather satellites. GPS satellites, at about 20,000 km, are considered medium Earth orbit. The higher the orbit, the more energy is required to put it there and the more energy is needed to reach it for repairs. Of particular interest are the satellites in geosynchronous orbit. All fixed satellite dishes on the ground pointing toward the sky, such as TV reception dishes, are pointed toward geosynchronous satellites. These satellites are placed at the exact distance, and just above the equator, such that their period of orbit is 1 day. They remain in a fixed position relative to Earth's surface.

8.14 Check Your Understanding By what factor must the radius change to reduce the orbital velocity of a satellite by one-half? By what factor would this change the period?

Example 8.23

Determining the Mass of Earth

Determine the mass of Earth from the orbit of the Moon.

Strategy

We use **Equation 8.18**, solve for M_E, and substitute for the period and radius of the orbit. The radius and period of the Moon's orbit was measured with reasonable accuracy thousands of years ago. From the astronomical data in **Appendix D**, the period of the Moon is 27.3 days $= 2.36 \times 10^6\ \text{s}$, and the *average* distance between the centers of Earth and the Moon is 384,000 km.

Solution

Solving for M_E,

$$T = 2\pi\sqrt{\frac{r^3}{GM_\text{E}}}$$

$$M_\text{E} = \frac{4\pi^2 r^3}{GT^2} = \frac{4\pi^2(3.84 \times 10^8\ \text{m})^3}{(6.67 \times 10^{-11}\ \text{N} \cdot \text{m}^2/\text{kg}^2)(2.36 \times 10^6\ \text{m})^2} = 6.01 \times 10^{24}\ \text{kg}.$$

Significance

Compare this to the value of $5.95 \times 10^{24}\ \text{kg}$ that we obtained in **Equation 8.16**, using the value of g at the surface of Earth. Although these values are very close (~0.8%), both calculations use average values. The value of g varies from the equator to the poles by approximately 0.5%. But the Moon has an elliptical orbit in which the value of r varies just over 10%. (The apparent size of the full Moon actually varies by about this amount, but it is difficult to notice through casual observation as the time from one extreme to the other is many months.)

8.15 Check Your Understanding There is another consideration to this last calculation of M_E. We derived **Equation 8.18** assuming that the satellite orbits around the center of the astronomical body at the same radius used in the expression for the gravitational force between them. What assumption is made to justify this? Earth is about 81 times more massive than the Moon. Does the Moon orbit about the exact center of Earth?

Example 8.24

Galactic Speed and Period

Let's revisit **Example 8.19**. Assume that the Milky Way and Andromeda galaxies are in a circular orbit about each other. What would be the velocity of each and how long would their orbital period be? Assume the mass of each is 800 billion solar masses and their centers are separated by 2.5 million light years.

Strategy

We cannot use **Equation 8.17** and **Equation 8.18** directly because they were derived assuming that the object of mass m orbited about the center of a much larger planet of mass M. We determined the gravitational force in **Example 8.19** using Newton's law of universal gravitation. We can use Newton's second law, applied to the centripetal acceleration of either galaxy, to determine their tangential speed. From that result we can determine the period of the orbit.

Solution

In **Example 8.19**, we found the force between the galaxies to be

$$F_{12} = G\frac{m_1 m_2}{r^2} = (6.67 \times 10^{-11}\ \text{N} \cdot \text{m}^2/\text{kg}^2)\frac{[(800 \times 10^9)(2.0 \times 10^{30}\ \text{kg})]^2}{[(2.5 \times 10^6)(9.5 \times 10^{15}\ \text{m})]^2} = 3.0 \times 10^{29}\ \text{N}$$

and that the acceleration of each galaxy is

$$a = \frac{F}{m} = \frac{3.0 \times 10^{29}\ \text{N}}{(800 \times 10^9)(2.0 \times 10^{30}\ \text{kg})} = 1.9 \times 10^{-13}\ \text{m/s}^2.$$

Since the galaxies are in a circular orbit, they have centripetal acceleration. If we ignore the effect of other galaxies, then the centers of mass of the two galaxies remain fixed. Hence, the galaxies must orbit about this common center of mass. For equal masses, the center of mass is exactly half way between them. So the radius of the orbit, r_{orbit}, is not the same as the distance between the galaxies, but one-half that value, or 1.25 million light-years. These two different values are shown in **Figure 8.45**.

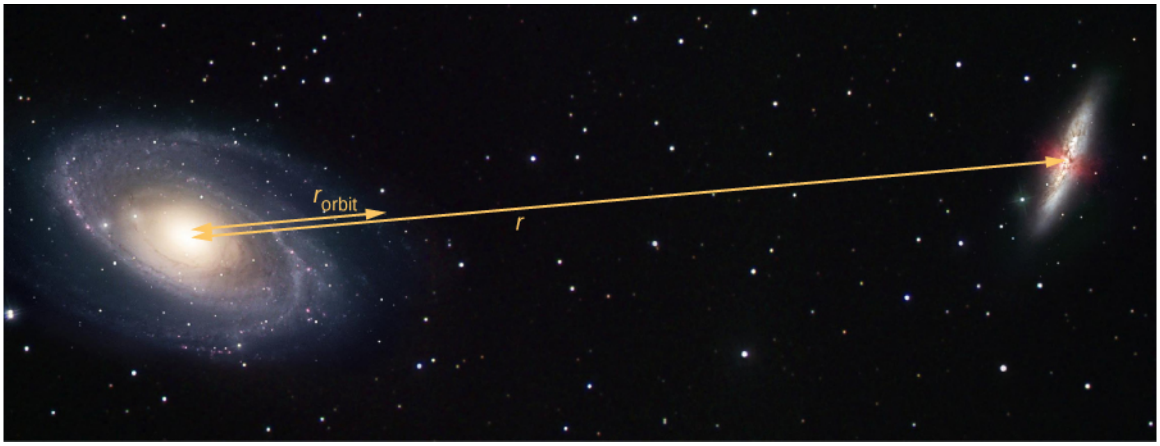


Figure 8.45 The distance between two galaxies, which determines the gravitational force between them, is *r*, and is different from r_{orbit}, which is the radius of orbit for each. For equal masses, $r_{\text{orbit}} = 1/2r$. (credit: modification of work by Marc Van Norden)

Using the expression for centripetal acceleration, we have

$$\begin{aligned} a_c &= \frac{v_{\text{orbit}}^2}{r_{\text{orbit}}} \\ 1.9 \times 10^{-13}\,\text{m/s}^2 &= \frac{v_{\text{orbit}}^2}{(1.25 \times 10^6)(9.5 \times 10^{15}\ \text{m})}. \end{aligned}$$

Solving for the orbit velocity, we have $v_{\text{orbit}} = 47\ \text{km/s}$. Finally, we can determine the period of the orbit directly from $T = 2\pi r/v_{\text{orbit}}$, to find that the period is $T = 1.6 \times 10^{18}\ \text{s}$, about 50 billion years.

Significance

The orbital speed of 47 km/s might seem high at first. But this speed is comparable to the escape speed from the Sun, which we calculated in an earlier example. To give even more perspective, this period is nearly four times longer than the time that the Universe has been in existence.

In fact, the present relative motion of these two galaxies is such that they are expected to collide in about 4 billion years. Although the density of stars in each galaxy makes a direct collision of any two stars unlikely, such a collision will have a dramatic effect on the shape of the galaxies. Examples of such collisions are well known in astronomy.

8.16 Check Your Understanding Galaxies are not single objects. How does the gravitational force of one galaxy exerted on the "closer" stars of the other galaxy compare to those farther away? What effect would this have on the shape of the galaxies themselves?

See the **Sloan Digital Sky Survey page (https://openstaxcollege.org/l/21sloandigskysu)** for more information on colliding galaxies.

CHAPTER 8 REVIEW

KEY TERMS

centripetal force any net force causing uniform circular motion

dynamics study of how forces affect the motion of objects and systems

external force force acting on an object or system that originates outside of the object or system

force push or pull on an object with a specific magnitude and direction; can be represented by vectors or expressed as a multiple of a standard force

free fall situation in which the only force acting on an object is gravity

free-body diagram sketch showing all external forces acting on an object or system; the system is represented by a single isolated point, and the forces are represented by vectors extending outward from that point

friction force that opposes relative motion or attempts at motion between systems in contact

inertia ability of an object to resist changes in its motion

inertial reference frame reference frame moving at constant velocity relative to an inertial frame is also inertial; a reference frame accelerating relative to an inertial frame is not inertial

kinetic friction force that opposes the motion of two systems that are in contact and moving relative to each other

law of inertia see Newton's first law of motion

net external force vector sum of all external forces acting on an object or system; causes a mass to accelerate

newton SI unit of force; 1 N is the force needed to accelerate an object with a mass of 1 kg at a rate of 1 m/s^2

Newton's first law of motion body at rest remains at rest or, if in motion, remains in motion at constant velocity unless acted on by a net external force; also known as the law of inertia

Newton's law of gravitation every mass attracts every other mass with a force proportional to the product of their masses, inversely proportional to the square of the distance between them, and with direction along the line connecting the center of mass of each

Newton's second law of motion acceleration of a system is directly proportional to and in the same direction as the net external force acting on the system and is inversely proportional to its mass

Newton's third law of motion whenever one body exerts a force on a second body, the first body experiences a force that is equal in magnitude and opposite in direction to the force that it exerts

normal force force supporting the weight of an object, or a load, that is perpendicular to the surface of contact between the load and its support; the surface applies this force to an object to support the weight of the object

static friction force that opposes the motion of two systems that are in contact and are not moving relative to each other

tension pulling force that acts along a stretched flexible connector, such as a rope or cable

thrust reaction force that pushes a body forward in response to a backward force

universal gravitational constant constant representing the strength of the gravitational force, that is believed to be the same throughout the universe

weight force $\vec{\mathbf{w}}$ due to gravity acting on an object of mass *m*

KEY EQUATIONS

Net external force	$\vec{\mathbf{F}}_{\text{net}} = \sum \vec{\mathbf{F}} = \vec{\mathbf{F}}_1 + \vec{\mathbf{F}}_2 + \cdots$
Newton's first law	$\vec{\mathbf{v}} = \text{constant when } \vec{\mathbf{F}}_{\text{net}} = \vec{\mathbf{0}}\text{ N}$

Newton’s second law, vector form	$\vec{\mathbf{F}}_{\text{net}} = \sum \vec{\mathbf{F}} = m\vec{\mathbf{a}}$
Newton’s second law, scalar form	$F_{\text{net}} = ma$
Newton’s second law, component form	$\sum F_x = ma_x, \ \sum F_y = ma_y, \ \sum F_z = ma_z,$
Definition of weight, vector form	$\vec{\mathbf{w}} = m\vec{\mathbf{g}}$
Definition of weight, scalar form	$w = mg$
Newton’s third law	$\vec{\mathbf{F}}_{\text{AB}} = -\vec{\mathbf{F}}_{\text{BA}}$
Normal force on an object resting on a horizontal surface, vector form	$\vec{\mathbf{N}} = -m\vec{\mathbf{g}}$
Normal force on an object resting on a horizontal surface, scalar form	$N = mg$
Normal force on an object resting on an inclined plane, scalar form	$N = mg\cos\theta$
Tension in a cable supporting an object of mass *m* at rest, scalar form	$T = w = mg$
Static friction	$f_{\text{s}} \leq \mu_{\text{s}} N,$
Kinetic friction	$f_{\text{k}} = \mu_{\text{k}} N,$
Centripetal force in terms of linear velocity	$F_{\text{c}} = m\frac{v^2}{r}$
Centripetal force in terms of angular velocity	$F_{\text{c}} = mr\omega^2.$
Newton's law of universal gravitation	$\vec{\mathbf{F}}_{12} = G\frac{m_1 m_2}{r^2}\hat{\mathbf{r}}_{12}$
Gravitational acceleration near Earth's surface	$g = G\frac{M_{\text{E}}}{R_{\text{E}}^2}$
Orbital speed of an object in a circular orbit around earth	$v_{\text{orbit}} = \sqrt{\frac{GM_{\text{E}}}{r}}.$
Orbital period of an object in circular orbit around Earth	$T = 2\pi\sqrt{\frac{r^3}{GM_{\text{E}}}}.$

SUMMARY

8.1 Forces

- Dynamics is the study of how forces affect the motion of objects, whereas kinematics simply describes the way objects move.
- Force is a push or pull that can be defined in terms of various standards, and it is a vector that has both magnitude and direction.
- External forces are any outside forces that act on a body. A free-body diagram is a drawing of all external forces acting on a body.
- The SI unit of force is the newton (N).

8.2 Newton's First Law

- According to Newton’s first law, there must be a cause for any change in velocity (a change in either magnitude or direction) to occur. This law is also known as the law of inertia.
- Friction is an external force that causes an object to slow down.
- Inertia is the tendency of an object to remain at rest or remain in motion. Inertia is related to an object’s mass.
- If an object’s velocity relative to a given frame is constant, then the frame is inertial. This means that for an inertial reference frame, Newton’s first law is valid.
- Equilibrium is achieved when the forces on a system are balanced.
- A net force of zero means that an object is either at rest or moving with constant velocity; that is, it is not accelerating.

8.3 Newton's Second Law

- An external force acts on a system from outside the system, as opposed to internal forces, which act between components within the system.
- Newton’s second law of motion says that the net external force on an object with a certain mass is directly proportional to and in the same direction as the acceleration of the object.
- The motion of an object in two or three dimensions can be analyzed by applying Newton's second law in each dimension independently of the others.

8.4 Mass and Weight

- Mass is the quantity of matter in a substance.
- The weight of an object is the net force on a falling object, or its gravitational force. The object experiences acceleration due to gravity.
- Some upward resistance force from the air acts on all falling objects on Earth, so they can never truly be in free fall.
- Careful distinctions must be made between free fall and weightlessness using the definition of weight as force due to gravity acting on an object of a certain mass.

8.5 Newton’s Third Law

- Newton’s third law of motion represents a basic symmetry in nature, with an experienced force equal in magnitude and opposite in direction to an exerted force.
- Two equal and opposite forces do not cancel because they act on different systems.
- Action-reaction pairs include a swimmer pushing off a wall, helicopters creating lift by pushing air down, and an octopus propelling itself forward by ejecting water from its body. Rockets, airplanes, and cars are pushed forward by a thrust reaction force.
- Choosing a system is an important analytical step in understanding the physics of a problem and solving it.

8.6 Common Forces

- When an object rests on a surface, the surface applies a force to the object that supports the weight of the object. This supporting force acts perpendicular to and away from the surface. It is called a normal force.
- When an object rests on a nonaccelerating horizontal surface, the magnitude of the normal force is equal to the weight of the object.
- When an object rests on an inclined plane that makes an angle θ with the horizontal surface, the weight of the object can be resolved into components that act perpendicular and parallel to the surface of the plane.
- The pulling force that acts along a stretched flexible connector, such as a rope or cable, is called tension. When a rope supports the weight of an object at rest, the tension in the rope is equal to the weight of the object. If the object is accelerating, tension is greater than weight, and if it is decelerating, tension is less than weight.

- The force of friction is a force experienced by a moving object (or an object that has a tendency to move) parallel to the interface opposing the motion (or its tendency).

8.7 Drawing Free-Body Diagrams

- To draw a free-body diagram, we draw the object of interest, draw all forces acting on that object, and resolve all force vectors into *x*- and *y*-components. We must draw a separate free-body diagram for each object in the problem.
- A free-body diagram is a useful means of describing and analyzing all the forces that act on a body to determine equilibrium according to Newton’s first law or acceleration according to Newton’s second law.

8.8 Friction

- Friction is a contact force that opposes the motion or attempted motion between two systems. Simple friction is proportional to the normal force N supporting the two systems.
- The magnitude of static friction force between two materials stationary relative to each other is determined using the coefficient of static friction, which depends on both materials.
- The kinetic friction force between two materials moving relative to each other is determined using the coefficient of kinetic friction, which also depends on both materials and is always less than the coefficient of static friction.

8.10 Newton's Law of Universal Gravitation

- All masses attract one another with a gravitational force proportional to their masses and inversely proportional to the square of the distance between them.
- Spherically symmetrical masses can be treated as if all their mass were located at the center.
- Nonsymmetrical objects can be treated as if their mass were concentrated at their center of mass, provided their distance from other masses is large compared to their size.

8.11 The Newtonian Synthesis

- The weight of an object is the gravitational attraction between Earth and the object.
- The gravitational acceleration of 9.8 m/s^2 for any object near Earth's surface comes directly from the calculation of the gravitational force of the Earth on that object.
- Orbital velocities are determined by the mass of the body being orbited and the distance from the center of that body, and not by the mass of a much smaller orbiting object.
- The period of the orbit is likewise independent of the orbiting object’s mass.
- Bodies of comparable masses orbit about their common center of mass and their velocities and periods should be determined from Newton’s second law and law of gravitation.

CONCEPTUAL QUESTIONS

8.1 Forces

1. What properties do forces have that allow us to classify them as vectors?

8.2 Newton's First Law

2. Taking a frame attached to Earth as inertial, which of the following objects cannot have inertial frames attached to them, and which are inertial reference frames?

(a) A car moving at constant velocity

(b) A car that is accelerating

(c) An elevator in free fall

(d) A space capsule orbiting Earth

(e) An elevator descending uniformly

3. A woman was transporting an open box of cupcakes to a school party. The car in front of her stopped suddenly; she applied her brakes immediately. She was wearing her seat belt and suffered no physical harm (just a great deal of embarrassment), but the cupcakes flew into the dashboard and became “smushcakes.” Explain what happened.

8.3 Newton's Second Law

4. Why can we neglect forces such as those holding a body together when we apply Newton’s second law?

5. A rock is thrown straight up. At the top of the trajectory, the velocity is momentarily zero. Does this imply that the force acting on the object is zero? Explain your answer.

8.4 Mass and Weight

6. What is the relationship between weight and mass? Which is an intrinsic, unchanging property of a body?

7. How much does a 70-kg astronaut weight in space, far from any celestial body? What is her mass at this location?

8. Which of the following statements is accurate?

(a) Mass and weight are the same thing expressed in different units.

(b) If an object has no weight, it must have no mass.

(c) If the weight of an object varies, so must the mass.

(d) Mass and inertia are different concepts.

(e) Weight is always proportional to mass.

9. When you stand on Earth, your feet push against it with a force equal to your weight. Why doesn't Earth accelerate away from you?

10. How would you give the value of $\vec{\mathbf{g}}$ in vector form?

8.5 Newton's Third Law

11. Identify the action and reaction forces in the following situations: (a) Earth attracts the Moon, (b) a boy kicks a football, (c) a rocket accelerates upward, (d) a car accelerates forward, (e) a high jumper leaps, and (f) a bullet is shot from a gun.

12. Suppose that you are holding a cup of coffee in your hand. Identify all forces on the cup and the reaction to each force.

13. (a) Why does an ordinary rifle recoil (kick backward) when fired? (b) The barrel of a recoilless rifle is open at both ends. Describe how Newton's third law applies when one is fired. (c) Can you safely stand close behind one when it is fired?

8.6 Common Forces

14. A table is placed on a rug. Then a book is placed on the table. What does the floor exert a normal force on?

15. A particle is moving to the right. (a) Can the force on it to be acting to the left? If yes, what would happen? (b) Can that force be acting downward? If yes, why?

8.7 Drawing Free-Body Diagrams

16. In completing the solution for a problem involving forces, what do we do after constructing the free-body diagram? That is, what do we apply?

17. If a book is located on a table, how many forces should be shown in a free-body diagram of the book? Describe them.

18. If the book in the previous question is in free fall, how many forces should be shown in a free-body diagram of the book? Describe them.

8.8 Friction

19. The glue on a piece of tape can exert forces. Can these forces be a type of simple friction? Explain, considering especially that tape can stick to vertical walls and even to ceilings.

20. When you learn to drive, you discover that you need to let up slightly on the brake pedal as you come to a stop or the car will stop with a jerk. Explain this in terms of the relationship between static and kinetic friction.

21. When you push a piece of chalk across a chalkboard, it sometimes screeches because it rapidly alternates between slipping and sticking to the board. Describe this process in more detail, in particular, explaining how it is related to the fact that kinetic friction is less than static friction. (The same slip-grab process occurs when tires screech on pavement.)

22. A physics major is cooking breakfast when she notices that the frictional force between her steel spatula and Teflon frying pan is only 0.200 N. Knowing the coefficient of kinetic friction between the two materials, she quickly calculates the normal force. What is it?

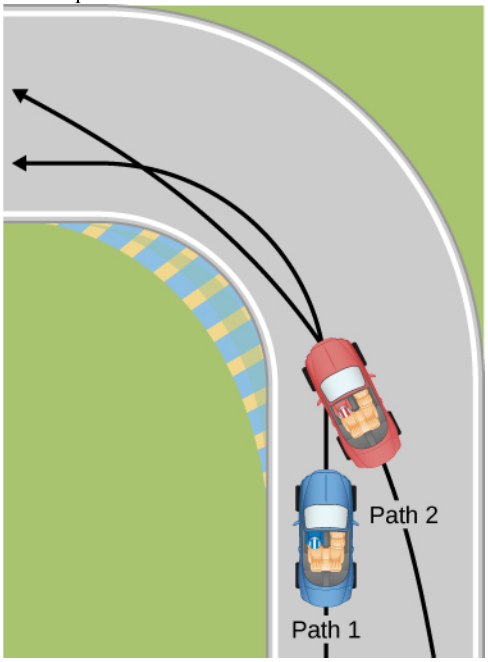
8.9 Centripetal Force

23. If you wish to reduce the stress (which is related to centripetal force) on high-speed tires, would you use large- or small-diameter tires? Explain.

24. Define centripetal force. Can any type of force (for example, tension, gravitational force, friction, and so on) be a centripetal force? Can any combination of forces be a centripetal force?

25. If centripetal force is directed toward the center, why do you feel that you are 'thrown' away from the center as a car goes around a curve? Explain.

26. Race car drivers routinely cut corners, as shown below (Path 2). Explain how this allows the curve to be taken at the greatest speed.



27. Many amusement parks have rides that make vertical loops like the one shown below. For safety, the cars are attached to the rails in such a way that they cannot fall off. If the car goes over the top at just the right speed, gravity alone will supply the centripetal force. What other force acts and what is its direction if:

(a) The car goes over the top at faster than this speed?

(b) The car goes over the top at slower than this speed?

28. What causes water to be removed from clothes in a spin-dryer?

29. As a skater forms a circle, what force is responsible for making his turn? Use a free-body diagram in your answer.

30. A car rounds a curve and encounters a patch of ice with a very low coefficient of kinetic fiction. The car slides off the road. Describe the path of the car as it leaves the road.

31. Two friends are having a conversation. Anna says a satellite in orbit is in free fall because the satellite keeps falling toward Earth. Tom says a satellite in orbit is not in free fall because the acceleration due to gravity is not $9.80\ \text{m/s}^2$. Who do you agree with and why?

8.10 Newton's Law of Universal Gravitation

32. Action at a distance, such as is the case for gravity, was once thought to be illogical and therefore untrue. What is the ultimate determinant of the truth in science, and why was this action at a distance ultimately accepted?

33. In the law of universal gravitation, Newton assumed that the force was proportional to the product of the two masses ($\sim m_1 m_2$). While all scientific conjectures must be experimentally verified, can you provide arguments as to why this must be? (You may wish to consider simple examples in which any other form would lead to contradictory results.)

8.11 The Newtonian Synthesis

34. One student argues that a satellite in orbit is in free fall because the satellite keeps falling toward Earth. Another says a satellite in orbit is not in free fall because the acceleration due to gravity is not $9.80\ \text{m/s}^2$. With whom do you agree with and why?

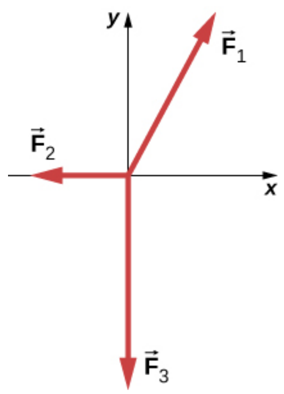
35. Many satellites are placed in geosynchronous orbits. What is special about these orbits? For a global communication network, how many of these satellites would be needed?

PROBLEMS

8.1 Forces

36. Two ropes are attached to a tree, and forces of $\vec{F}_1 = 2.0\hat{i} + 4.0\hat{j}$ N and $\vec{F}_2 = 3.0\hat{i} + 6.0\hat{j}$ N are applied. The forces are coplanar (in the same plane). (a) What is the resultant (net force) of these two force vectors? (b) Find the magnitude and direction of this net force.

37. A telephone pole has three cables pulling as shown from above, with $\vec{F}_1 = \left(300.0\hat{i} + 500.0\hat{j}\right)$, $\vec{F}_2 = -200.0\hat{i}$, and $\vec{F}_3 = -800.0\hat{j}$. (a) Find the net force on the telephone pole in component form. (b) Find the magnitude and direction of this net force.



38. Two teenagers are pulling on ropes attached to a tree. The angle between the ropes is $30.0°$. David pulls with a force of 400.0 N and Stephanie pulls with a force of 300.0 N. (a) Find the component form of the net force. (b) Find the magnitude of the resultant (net) force on the tree and the angle it makes with David's rope.

8.2 Newton's First Law

39. Two forces of $\vec{F}_1 = \frac{75.0}{\sqrt{2}}\left(\hat{i} - \hat{j}\right)$N and $\vec{F}_2 = \frac{150.0}{\sqrt{2}}\left(\hat{i} - \hat{j}\right)$N act on an object. Find the third force $\vec{F}_3$ that is needed to balance the first two forces.

40. While sliding a couch across a floor, Andrea and Jennifer exert forces $\vec{F}_A$ and $\vec{F}_J$ on the couch. Andrea's force is due north with a magnitude of 130.0 N and Jennifer's force is $32°$ east of north with a magnitude of 180.0 N. (a) Find the net force in component form. (b) Find the magnitude and direction of the net force. (c) If Andrea and Jennifer's housemates, David and Stephanie, disagree with the move and want to prevent its relocation, with what combined force $\vec{F}_{DS}$ should they push so that the couch does not move?

8.3 Newton's Second Law

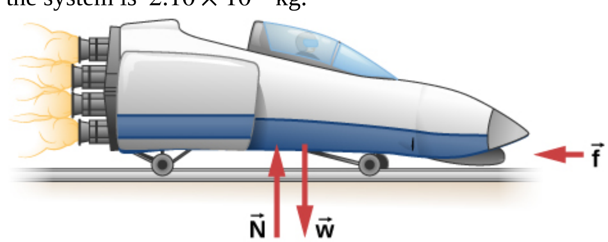
41. Andrea, a 63.0-kg sprinter, starts a race with an acceleration of $4.200\ \text{m/s}^2$. What is the net external force on her?

42. If the sprinter from the previous problem accelerates at that rate for 20.00 m and then maintains that velocity for the remainder of a 100.00-m dash, what will her time be for the race?

43. A cleaner pushes a 4.50-kg laundry cart in such a way that the net external force on it is 60.0 N. Calculate the magnitude of his cart's acceleration.

44. Astronauts in orbit are apparently weightless. This means that a clever method of measuring the mass of astronauts is needed to monitor their mass gains or losses, and adjust their diet. One way to do this is to exert a known force on an astronaut and measure the acceleration produced. Suppose a net external force of 50.0 N is exerted, and an astronaut's acceleration is measured to be $0.893\ \text{m/s}^2$. (a) Calculate her mass. (b) By exerting a force on the astronaut, the vehicle in which she orbits experiences an equal and opposite force. Use this knowledge to find an equation for the acceleration of the system (astronaut and spaceship) that would be measured by a nearby observer. (c) Discuss how this would affect the measurement of the astronaut's acceleration. Propose a method by which recoil of the vehicle is avoided.

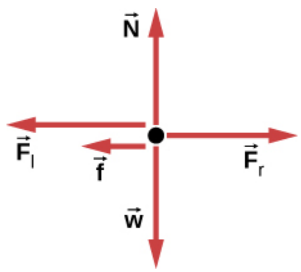
45. The rocket sled shown below decelerates at a rate of $196\ \text{m/s}^2$. What force is necessary to produce this deceleration? Assume that the rockets are off. The mass of the system is 2.10×10^3 kg.



46. If the rocket sled shown in the previous problem starts with only one rocket burning, what is the magnitude of this acceleration? Assume that the mass of the system is 2.10×10^3 kg, the thrust T is 2.40×10^4 N, and the force of friction opposing the motion is 650.0 N. (b) Why is the acceleration not one-fourth of what it is with all rockets burning?

47. What is the deceleration of the rocket sled if it comes to rest in 1.10 s from a speed of 1000.0 km/h? (Such deceleration caused one test subject to black out and have temporary blindness.)

48. Suppose two children push horizontally, but in exactly opposite directions, on a third child in a wagon. The first child exerts a force of 75.0 N, the second exerts a force of 90.0 N, friction is 12.0 N, and the mass of the third child plus wagon is 23.0 kg. (a) What is the system of interest if the acceleration of the child in the wagon is to be calculated? (See the free-body diagram.) (b) Calculate the acceleration. (c) What would the acceleration be if friction were 15.0 N?



49. A powerful motorcycle can produce an acceleration of $3.50\ \text{m/s}^2$ while traveling at 90.0 km/h. At that speed, the forces resisting motion, including friction and air resistance, total 400.0 N. (Air resistance is analogous to air friction. It always opposes the motion of an object.) What is the magnitude of the force that motorcycle exerts backward on the ground to produce its acceleration if the mass of the motorcycle with rider is 245 kg?

50. A car with a mass of 1000.0 kg accelerates from 0 to 90.0 km/h in 10.0 s. (a) What is its acceleration? (b) What is the net force on the car?

51. The driver in the previous problem applies the brakes when the car is moving at 90.0 km/h, and the car comes to rest after traveling 40.0 m. What is the net force on the car during its deceleration?

52. An 80.0-kg passenger in an SUV traveling at 1.00×10^2 km/h is wearing a seat belt. The driver slams on the brakes and the SUV stops in 45.0 m. Find the force of the seat belt on the passenger.

53. A particle of mass 2.0 kg is acted on by a single force $\vec{\mathbf{F}}_1 = 18\,\hat{\mathbf{i}}$ N. (a) What is the particle's acceleration? (b) If the particle starts at rest, how far does it travel in the first 5.0 s?

8.4 Mass and Weight

54. The weight of an astronaut plus his space suit on the Moon is only 250 N. (a) How much does the suited astronaut weigh on Earth? (b) What is the mass on the Moon? On Earth?

55. Suppose the mass of a fully loaded module in which astronauts take off from the Moon is 1.00×10^4 kg. The thrust of its engines is 3.00×10^4 N. (a) Calculate the module's magnitude of acceleration in a vertical takeoff from the Moon. (b) Could it lift off from Earth? If not, why not? If it could, calculate the magnitude of its acceleration.

56. A rocket sled accelerates at a rate of $49.0\ \text{m/s}^2$. Its passenger has a mass of 75.0 kg. (a) Calculate the horizontal component of the force the seat exerts against his body. Compare this with his weight using a ratio. (b) Calculate the direction and magnitude of the total force the seat exerts against his body.

57. Repeat the previous problem for a situation in which the rocket sled decelerates at a rate of $201\ \text{m/s}^2$. In this problem, the forces are exerted by the seat and the seat belt.

58. A body of mass 2.00 kg is pushed straight upward by a 25.0 N vertical force. What is its acceleration?

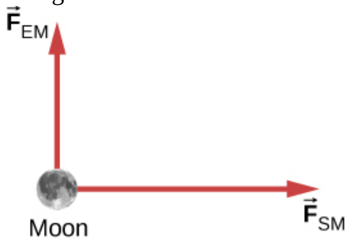
59. A car weighing 12,500 N starts from rest and accelerates to 83.0 km/h in 5.00 s. The friction force is 1350 N. Find the applied force produced by the engine.

60. A body with a mass of 10.0 kg is assumed to be in Earth's gravitational field with $g = 9.80\ \text{m/s}^2$. What is its acceleration?

61. A fireman has mass *m*; he hears the fire alarm and slides down the pole with acceleration *a* (which is less than *g* in magnitude). (a) Write an equation giving the vertical force he must apply to the pole. (b) If his mass is 90.0 kg and he accelerates at $5.00\ \text{m/s}^2$, what is the magnitude of his applied force?

62. A baseball catcher is performing a stunt for a television commercial. He will catch a baseball (mass 145 g) dropped from a height of 60.0 m above his glove. His glove stops the ball in 0.0100 s. What is the force exerted by his glove on the ball?

63. When the Moon is directly overhead at sunset, the force by Earth on the Moon, F_{EM}, is essentially at $90°$ to the force by the Sun on the Moon, F_{SM}, as shown below. Given that $F_{\text{EM}} = 1.98 \times 10^{20}\ \text{N}$ and $F_{\text{SM}} = 4.36 \times 10^{20}\ \text{N}$, all other forces on the Moon are negligible, and the mass of the Moon is $7.35 \times 10^{22}\ \text{kg}$, determine the magnitude of the Moon's acceleration.

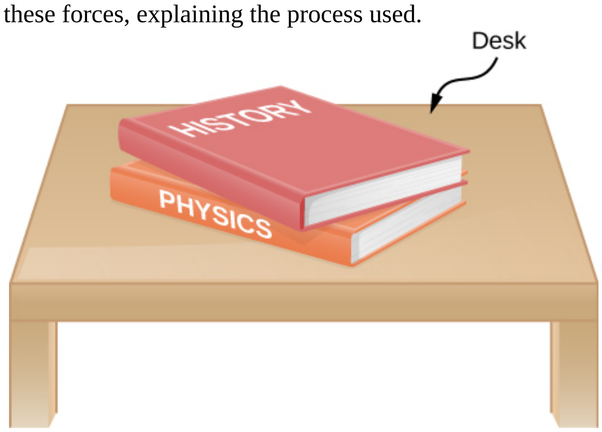


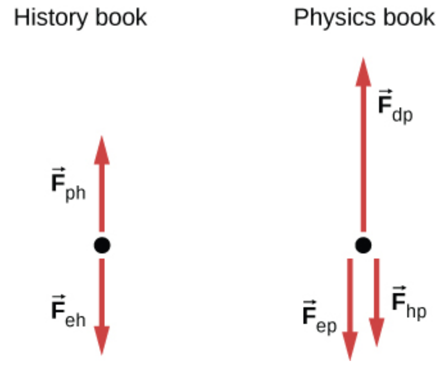
8.5 Newton's Third Law

64. (a) What net external force is exerted on a 1100.0-kg artillery shell fired from a battleship if the shell is accelerated at $2.40 \times 10^4\ \text{m/s}^2$? (b) What is the magnitude of the force exerted on the ship by the artillery shell, and why?

65. A brave but inadequate rugby player is being pushed backward by an opposing player who is exerting a force of 800.0 N on him. The mass of the losing player plus equipment is 90.0 kg, and he is accelerating backward at $1.20\ \text{m/s}^2$. (a) What is the force of friction between the losing player's feet and the grass? (b) What force does the winning player exert on the ground to move forward if his mass plus equipment is 110.0 kg?

66. A history book is lying on top of a physics book on a desk, as shown below; a free-body diagram is also shown. The history and physics books weigh 14 N and 18 N, respectively. Identify each force on each book with a double subscript notation (for instance, the contact force of the history book pressing against physics book can be described as $\vec{\mathbf{F}}_{\text{HP}}$), and determine the value of each of these forces, explaining the process used.





67. A truck collides with a car, and during the collision, the net force on each vehicle is essentially the force exerted by the other. Suppose the mass of the car is 550 kg, the mass of the truck is 2200 kg, and the magnitude of the truck's acceleration is $10\ \text{m/s}^2$. Find the magnitude of the car's acceleration.

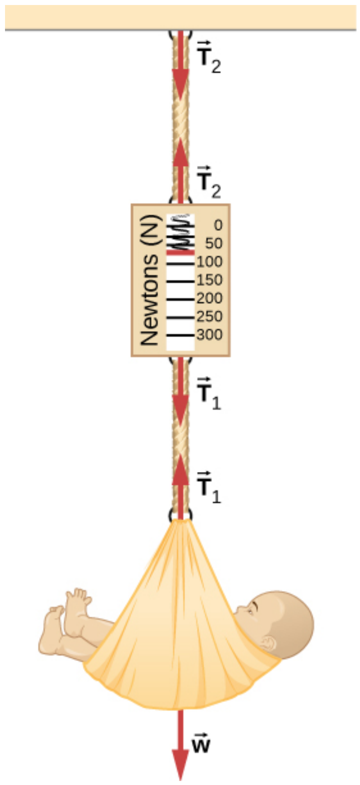
8.6 Common Forces

68. What force does a trampoline have to apply to Jennifer, a 45.0-kg gymnast, to accelerate her straight up at $7.50\ \text{m/s}^2$? The answer is independent of the velocity of the gymnast—she can be moving up or down or can be instantly stationary.

69. Calculate the tension in a vertical strand of spider web if a spider of mass $2.00 \times 10^{-5}\ \text{kg}$ hangs motionless on it.

70. Suppose Kevin, a 60.0-kg gymnast, climbs a rope. (a) What is the tension in the rope if he climbs at a constant speed? (b) What is the tension in the rope if he accelerates upward at a rate of $1.50\ \text{m/s}^2$?

71. Consider the baby being weighed in the following figure. (a) What is the mass of the infant and basket if a scale reading of 55 N is observed? (b) What is tension T_1 in the cord attaching the baby to the scale? (c) What is tension T_2 in the cord attaching the scale to the ceiling, if the scale has a mass of 0.500 kg? (d) Sketch the situation, indicating the system of interest used to solve each part. The masses of the cords are negligible.



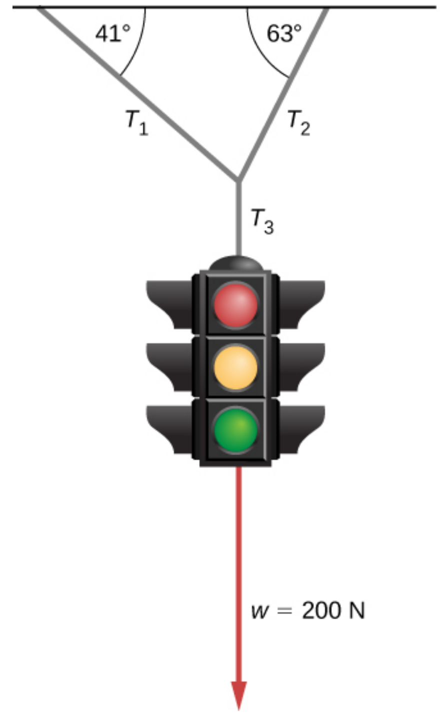
8.7 Drawing Free-Body Diagrams

72. A ball of mass *m* hangs at rest, suspended by a string. (a) Sketch all forces. (b) Draw the free-body diagram for the ball.

73. A car moves along a horizontal road. Draw a free-body diagram; be sure to include the friction of the road that opposes the forward motion of the car.

74. A runner pushes against the track, as shown. (a) Provide a free-body diagram showing all the forces on the runner. (*Hint:* Place all forces at the center of his body, and include his weight.) (b) Give a revised diagram showing the *xy*-component form.

75. The traffic light hangs from the cables as shown. Draw a free-body diagram on a coordinate plane for this situation.



8.8 Friction

76. (a) When rebuilding his car's engine, a physics major must exert 3.00×10^2 N of force to insert a dry steel piston into a steel cylinder. What is the normal force between the piston and cylinder? (b) What force would he have to exert if the steel parts were oiled?

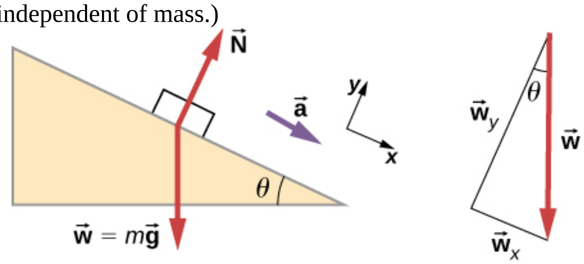
77. (a) What is the maximum frictional force in the knee joint of a person who supports 66.0 kg of her mass on that knee? (b) During strenuous exercise, it is possible to exert forces to the joints that are easily 10 times greater than the weight being supported. What is the maximum force of friction under such conditions? The frictional forces in joints are relatively small in all circumstances except when the joints deteriorate, such as from injury or arthritis. Increased frictional forces can cause further damage and pain.

78. Suppose you have a 120-kg wooden crate resting on a wood floor, with coefficient of static friction 0.500 between these wood surfaces. (a) What maximum force can you exert horizontally on the crate without moving it? (b) If you continue to exert this force once the crate starts to slip, what will its acceleration then be? The coefficient of sliding friction is known to be 0.300 for this situation.

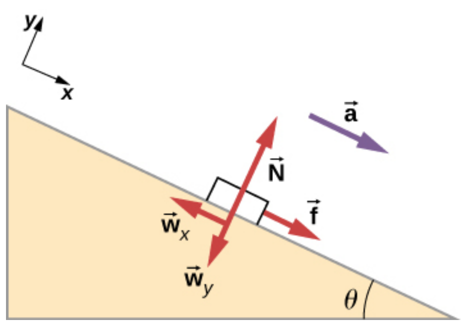
79. (a) If half of the weight of a small 1.00×10^3-kg utility truck is supported by its two drive wheels, what is the maximum acceleration it can achieve on dry concrete? (b) Will a metal cabinet lying on the wooden bed of the truck slip if it accelerates at this rate? (c) Solve both problems assuming the truck has four-wheel drive.

80. A team of eight dogs pulls a sled with waxed wood runners on wet snow (mush!). The dogs have average masses of 19.0 kg, and the loaded sled with its rider has a mass of 210 kg. (a) Calculate the acceleration of the dogs starting from rest if each dog exerts an average force of 185 N backward on the snow. (b) Calculate the force in the coupling between the dogs and the sled.

81. Show that the acceleration of any object down a frictionless incline that makes an angle θ with the horizontal is $a = g\sin\theta$. (Note that this acceleration is independent of mass.)



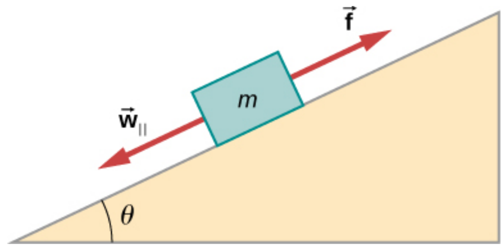
82. Show that the acceleration of any object down an incline where friction behaves simply (that is, where $f_\text{k} = \mu_\text{k} N$) is $a = g(\sin\theta - \mu_\text{k}\cos\theta)$. Note that the acceleration is independent of mass and reduces to the expression found in the previous problem when friction becomes negligibly small $(\mu_\text{k} = 0)$.



83. Calculate the deceleration of a snow boarder going up a $5.00°$ slope, assuming the coefficient of friction for waxed wood on wet snow. The result of the preceding problem may be useful, but be careful to consider the fact that the snow boarder is going uphill.

84. A machine at a post office sends packages out a chute and down a ramp to be loaded into delivery vehicles. (a) Calculate the acceleration of a box heading down a $10.0°$ slope, assuming the coefficient of friction for a parcel on waxed wood is 0.100. (b) Find the angle of the slope down which this box could move at a constant velocity. You can neglect air resistance in both parts.

85. If an object is to rest on an incline without slipping, then friction must equal the component of the weight of the object parallel to the incline. This requires greater and greater friction for steeper slopes. Show that the maximum angle of an incline above the horizontal for which an object will not slide down is $\theta = \tan^{-1}\mu_\text{s}$. You may use the result of the previous problem. Assume that $a = 0$ and that static friction has reached its maximum value.



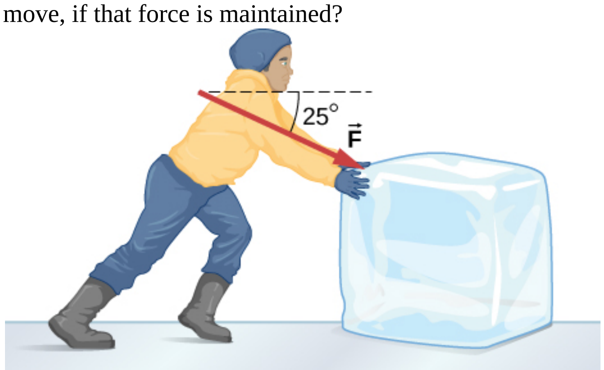
86. Calculate the maximum acceleration of a car that is heading down a $6.00°$ slope (one that makes an angle of $6.00°$ with the horizontal) under the following road conditions. You may assume that the weight of the car is evenly distributed on all four tires and that the coefficient of static friction is involved—that is, the tires are not allowed to slip during the deceleration. (Ignore rolling.) Calculate for a car: (a) On dry concrete. (b) On wet concrete. (c) On ice, assuming that $\mu_\text{s} = 0.100$, the same as for shoes on ice.

87. Calculate the maximum acceleration of a car that is heading up a $4.00°$ slope (one that makes an angle of $4.00°$ with the horizontal) under the following road conditions. Assume that only half the weight of the car is supported by the two drive wheels and that the coefficient of static friction is involved—that is, the tires are not allowed to slip during the acceleration. (Ignore rolling.) (a) On dry concrete. (b) On wet concrete. (c) On ice, assuming that $\mu_\text{s} = 0.100$, the same as for shoes on ice.

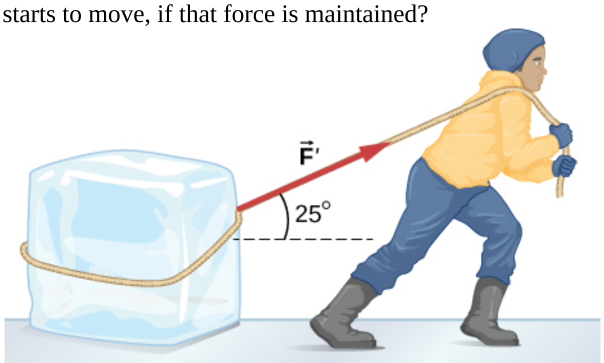
88. Repeat the preceding problem for a car with four-wheel drive.

89. A freight train consists of two 8.00×10^5-kg engines and 45 cars with average masses of 5.50×10^5 kg. (a) What force must each engine exert backward on the track to accelerate the train at a rate of $5.00 \times 10^{-2}\,\text{m/s}^2$ if the force of friction is $7.50 \times 10^5\,\text{N}$, assuming the engines exert identical forces? This is not a large frictional force for such a massive system. Rolling friction for trains is small, and consequently, trains are very energy-efficient transportation systems. (b) What is the force in the coupling between the 37th and 38th cars (this is the force each exerts on the other), assuming all cars have the same mass and that friction is evenly distributed among all of the cars and engines?

90. A contestant in a winter sporting event pushes a 45.0-kg block of ice across a frozen lake as shown below. (a) Calculate the minimum force *F* he must exert to get the block moving. (b) What is its acceleration once it starts to move, if that force is maintained?



91. The contestant now pulls the block of ice with a rope over his shoulder at the same angle above the horizontal as shown below. Calculate the minimum force *F* he must exert to get the block moving. (b) What is its acceleration once it starts to move, if that force is maintained?



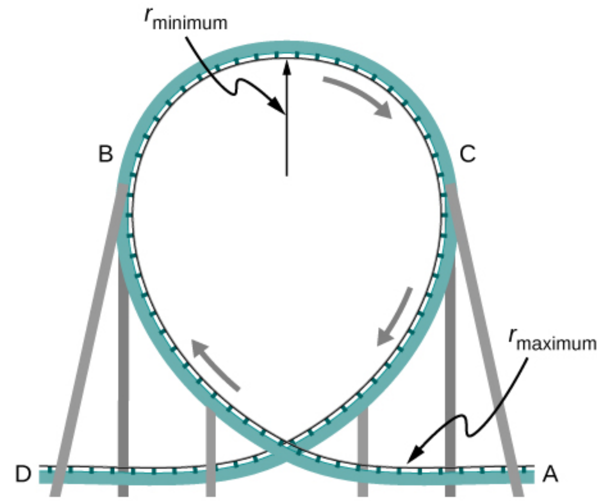
92. At a post office, a parcel that is a 20.0-kg box slides down a ramp inclined at $30.0°$ with the horizontal. The coefficient of kinetic friction between the box and plane is 0.0300. (a) Find the acceleration of the box. (b) Find the velocity of the box as it reaches the end of the plane, if the length of the plane is 2 m and the box starts at rest.

8.9 Centripetal Force

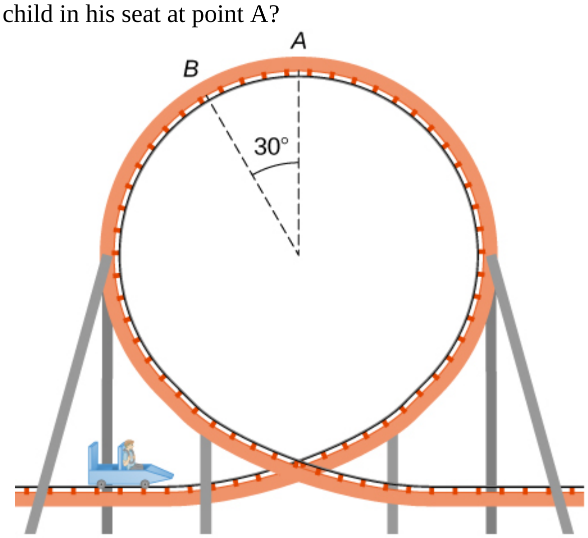
93. (a) A 22.0-kg child is riding a playground merry-go-round that is rotating at 40.0 rev/min. What centripetal force is exerted if he is 1.25 m from its center? (b) What centripetal force is exerted if the merry-go-round rotates at 3.00 rev/min and he is 8.00 m from its center? (c) Compare each force with his weight.

94. Calculate the centripetal force on the end of a 100-m (radius) wind turbine blade that is rotating at 0.5 rev/s. Assume the mass is 4 kg.

95. Modern roller coasters have vertical loops like the one shown here. The radius of curvature is smaller at the top than on the sides so that the downward centripetal acceleration at the top will be greater than the acceleration due to gravity, keeping the passengers pressed firmly into their seats. (a) What is the speed of the roller coaster at the top of the loop if the radius of curvature there is 15.0 m and the downward acceleration of the car is 1.50 *g*? (b) How high above the top of the loop must the roller coaster start from rest, assuming negligible friction? (c) If it actually starts 5.00 m higher than your answer to (b), how much energy did it lose to friction? Its mass is 1.50×10^3 kg .



96. A child of mass 40.0 kg is in a roller coaster car that travels in a loop of radius 7.00 m. At point A the speed of the car is 10.0 m/s, and at point B, the speed is 10.5 m/s. Assume the child is not holding on and does not wear a seat belt. (a) What is the force of the car seat on the child at point A? (b) What is the force of the car seat on the child at point B? (c) What minimum speed is required to keep the child in his seat at point A?



97. In the simple Bohr model of the ground state of the hydrogen atom, the electron travels in a circular orbit around a fixed proton. The radius of the orbit is 5.28×10^{-11} m, and the speed of the electron is

2.18×10^6 m/s. The mass of an electron is 9.11×10^{-31} kg. What is the force on the electron?

98. The CERN particle accelerator is circular with a circumference of 7.0 km. (a) What is the acceleration of the protons $(m = 1.67 \times 10^{-27}\text{ kg})$ that move around the accelerator at 5% of the speed of light? (The speed of light is $v = 3.00 \times 10^8$ m/s.) (b) What is the force on the protons?

99. A car rounds an unbanked curve of radius 65 m. If the coefficient of static friction between the road and car is 0.70, what is the maximum speed at which the car traverse the curve without slipping?

8.10 Newton's Law of Universal Gravitation

100. Evaluate the magnitude of gravitational force between two 5-kg spherical steel balls separated by a center-to-center distance of 15 cm.

101. Estimate the gravitational force between two sumo wrestlers, with masses 220 kg and 240 kg, when they are embraced and their centers are 1.2 m apart.

102. Astrology makes much of the position of the planets at the moment of one's birth. The only known force a planet exerts on Earth is gravitational. (a) Calculate the gravitational force exerted on a 4.20-kg baby by a 100-kg father 0.200 m away at birth (he is assisting, so he is close to the child). (b) Calculate the force on the baby due to Jupiter if it is at its closest distance to Earth, some 6.29×10^{11} m away. How does the force of Jupiter on the baby compare to the force of the father on the baby? Other objects in the room and the hospital building also exert similar gravitational forces. (Of course, there could be an unknown force acting, but scientists first need to be convinced that there is even an effect, much less that an unknown force causes it.)

103. A mountain 10.0 km from a person exerts a gravitational force on him equal to 2.00% of his weight. (a) Calculate the mass of the mountain. (b) Compare the mountain's mass with that of Earth. (c) What is unreasonable about these results? (d) Which premises are unreasonable or inconsistent? (Note that accurate gravitational measurements can easily detect the effect of nearby mountains and variations in local geology.)

104. The International Space Station has a mass of approximately 370,000 kg. (a) What is the force on a 150-kg suited astronaut if she is 20 m from the center of mass of the station? (b) How accurate do you think your answer would be?

Figure 8.46 (credit: ©ESA–David Ducros)

105. Asteroid Toutatis passed near Earth in 2006 at four times the distance to our Moon. This was the closest approach we will have until 2060. If it has mass of 5.0×10^{13} kg, what force did it exert on Earth at its closest approach?

106. (a) What was the acceleration of Earth caused by asteroid Toutatis (see previous problem) at its closest approach? (b) What was the acceleration of Toutatis at this point?

8.11 The Newtonian Synthesis

107. (a) Calculate Earth's mass given the acceleration due to gravity at the North Pole is measured to be 9.832 m/s^2 and the radius of the Earth at the pole is 6356 km. (b) Compare this with the NASA's Earth Fact Sheet value of 5.9726×10^{24} kg.

108. (a) What is the acceleration due to gravity on the surface of the Moon? (b) On the surface of Mars? The mass of Mars is 6.418×10^{23} kg and its radius is 3.38×10^6 m.

109. (a) Calculate the acceleration due to gravity on the surface of the Sun. (b) By what factor would your weight increase if you could stand on the Sun? (Never mind that you cannot.)

110. The mass of a particle is 15 kg. (a) What is its weight on Earth? (b) What is its weight on the Moon? (c) What is its mass on the Moon? (d) What is its weight in outer space far from any celestial body? (e) What is its mass at this point?

111. On a planet whose radius is 1.2×10^7 m, the acceleration due to gravity is 18 m/s^2. What is the mass of the planet?

112. The mean diameter of the planet Saturn is $1.2 \times 10^8 \text{ m}$, and its mean mass density is 0.69 g/cm^3. Find the acceleration due to gravity at Saturn's surface.

113. The mean diameter of the planet Mercury is $4.88 \times 10^6 \text{ m}$, and the acceleration due to gravity at its surface is 3.78 m/s^2. Estimate the mass of this planet.

114. The acceleration due to gravity on the surface of a planet is three times as large as it is on the surface of Earth. The mass density of the planet is known to be twice that of Earth. What is the radius of this planet in terms of Earth's radius?

115. A body on the surface of a planet with the same radius as Earth's weighs 10 times more than it does on Earth. What is the mass of this planet in terms of Earth's mass?

116. If a planet with 1.5 times the mass of Earth was traveling in Earth's orbit, what would its period be?

117. Two planets in circular orbits around a star have speeds of *v* and 2*v*. (a) What is the ratio of the orbital radii of the planets? (b) What is the ratio of their periods?

118. Using the average distance of Earth from the Sun, and the orbital period of Earth, (a) find the centripetal acceleration of Earth in its motion about the Sun. (b) Compare this value to that of the centripetal acceleration at the equator due to Earth's rotation.

119. What is the orbital radius of an Earth satellite having a period of 1.00 h? (b) What is unreasonable about this result?

120. Calculate the mass of the Sun based on data for Earth's orbit and compare the value obtained with the Sun's actual mass.

121. Find the mass of Jupiter based on the fact that Io, its innermost moon, has an average orbital radius of 421,700 km and a period of 1.77 days.

122. Astronomical observations of our Milky Way galaxy indicate that it has a mass of about 8.0×10^{11} solar masses. A star orbiting on the galaxy's periphery is about 6.0×10^4 light-years from its center. (a) What should the orbital period of that star be? (b) If its period is 6.0×10^7 years instead, what is the mass of the galaxy? Such calculations are used to imply the existence of other matter, such as a very massive black hole at the center of the Milky Way.

123. (a) In order to keep a small satellite from drifting into a nearby asteroid, it is placed in orbit with a period of 3.02 hours and radius of 2.0 km. What is the mass of the asteroid? (b) Does this mass seem reasonable for the size of the orbit?

124. The Moon and Earth rotate about their common center of mass, which is located about 4700 km from the center of Earth. (This is 1690 km below the surface.) (a) Calculate the acceleration due to the Moon's gravity at that point. (b) Calculate the centripetal acceleration of the center of Earth as it rotates about that point once each lunar month (about 27.3 d) and compare it with the acceleration found in part (a). Comment on whether or not they are equal and why they should or should not be.

125. The Sun orbits the Milky Way galaxy once each $2.60 \times 10^8 \text{ years}$, with a roughly circular orbit averaging a radius of 3.00×10^4 light-years. (A light-year is the distance traveled by light in 1 year.) Calculate the centripetal acceleration of the Sun in its galactic orbit. Does your result support the contention that a nearly inertial frame of reference can be located at the Sun? (b) Calculate the average speed of the Sun in its galactic orbit. Does the answer surprise you?

126. A geosynchronous Earth satellite is one that has an orbital period of precisely 1 day. Such orbits are useful for communication and weather observation because the satellite remains above the same point on Earth (provided it orbits in the equatorial plane in the same direction as Earth's rotation). Calculate the radius of such an orbit based on the data for Earth in Appendix D.

ADDITIONAL PROBLEMS

127. Draw a free-body diagram of a diver who has entered the water, moved downward, and is acted on by an upward force due to the water which balances the weight (that is, the diver is suspended).

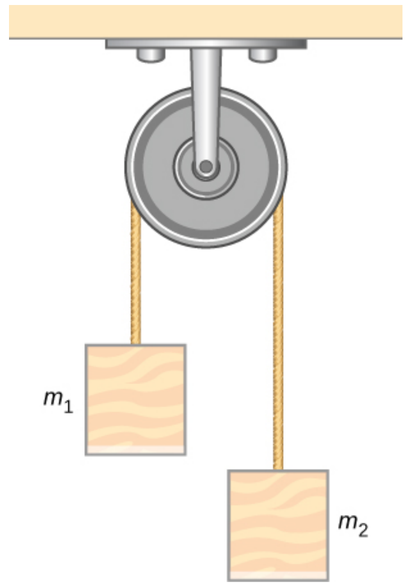
128. For a swimmer who has just jumped off a diving board, assume air resistance is negligible. The swimmer has a mass of 80.0 kg and jumps off a board 10.0 m above the water. Three seconds after entering the water, her downward motion is stopped. What average upward force did the water exert on her?

129. (a) Find an equation to determine the magnitude of the net force required to stop a car of mass *m*, given that the initial speed of the car is v_0 and the stopping distance is *x*.

(b) Find the magnitude of the net force if the mass of the car is 1050 kg, the initial speed is 40.0 km/h, and the stopping distance is 25.0 m.

130. Two forces are applied to a 5.0-kg object, and it accelerates at a rate of 2.0 m/s^2 in the positive *y*-direction. If one of the forces acts in the positive *x*-direction with magnitude 12.0 N, find the magnitude of the other force.

131. The block on the right shown below has more mass than the block on the left ($m_2 > m_1$). Draw free-body diagrams for each block.



CHALLENGE PROBLEMS

132. On June 25, 1983, shot-putter Udo Beyer of East Germany threw the 7.26-kg shot 22.22 m, which at that time was a world record. (a) If the shot was released at a height of 2.20 m with a projection angle of $45.0°$, what was its initial velocity? (b) If while in Beyer's hand the shot was accelerated uniformly over a distance of 1.20 m, what was the net force on it?

133. A body of mass *m* has initial velocity v_0 in the positive *x*-direction. It is acted on by a constant force *F* for time *t* until the velocity becomes zero; the force continues to act on the body until its velocity becomes $-v_0$ in the same amount of time. Write an expression for the total distance the body travels in terms of the variables indicated.

134. A bullet shot from a rifle has mass of 10.0 g and travels to the right at 350 m/s. It strikes a target, a large bag of sand, penetrating it a distance of 34.0 cm. Find the magnitude and direction of the retarding force that slows and stops the bullet.

135. In a particle accelerator, a proton has mass 1.67×10^{-27} kg and an initial speed of 2.00×10^5 m/s. It moves in a straight line, and its speed increases to 9.00×10^5 m/s in a distance of 10.0 cm. Assume that the acceleration is constant. Find the magnitude of the force exerted on the proton.

9 | NEWTON'S LAWS FOR ROTATIONS

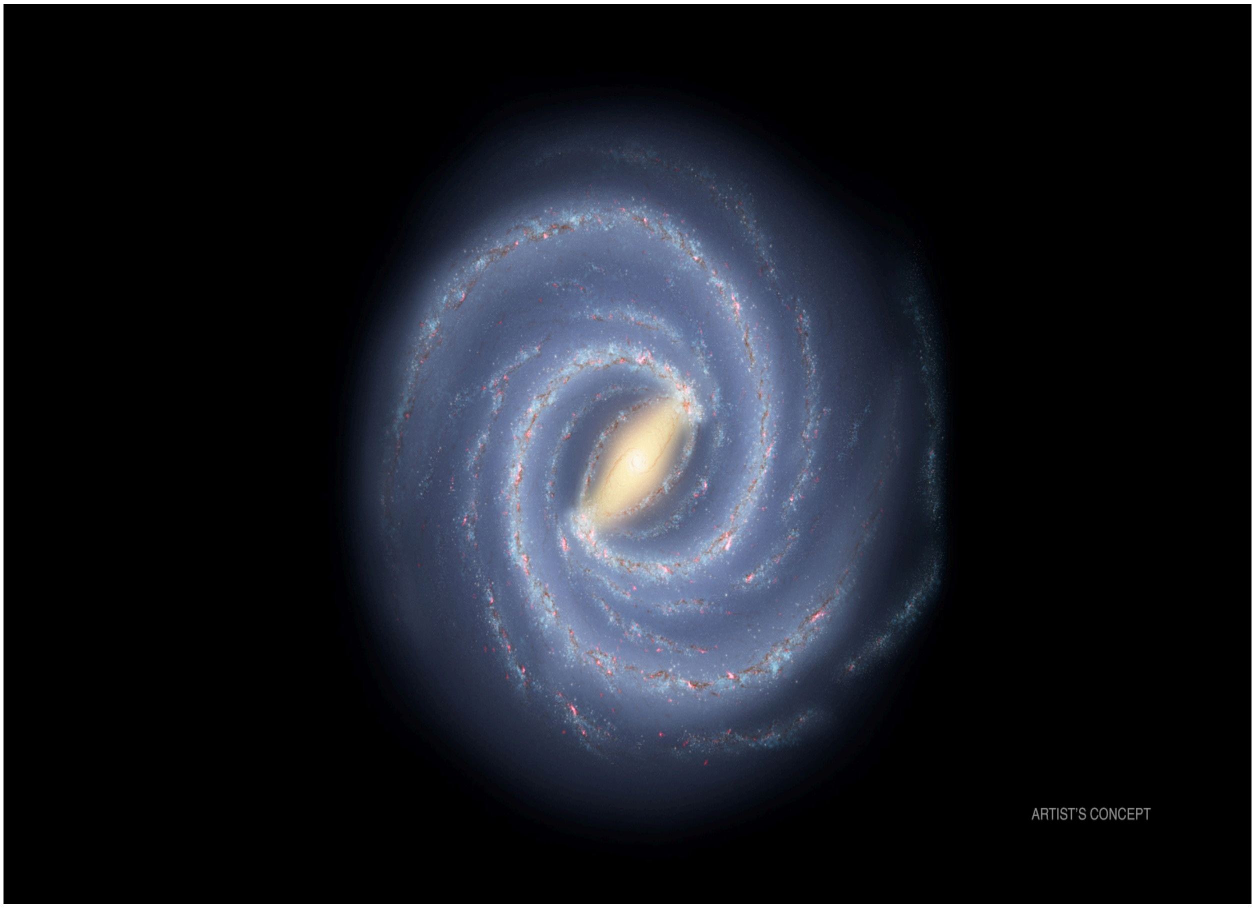


Figure 9.1 An artist's conception of our own Milky Way galaxy. (Image credit: NASA/JPL-Caltech/R. Hurt (SSC/Caltech))

Chapter Outline

Introduction

In previous chapters, we have described the dynamics of objects that translate in one, two or three dimensions. The connection between how they move (kinematics) and why they move as they do (dynamics) was Newton's Laws. We have also described the kinematics of objects that move in circular paths (like the stars orbiting in a spiral galaxy) and rigid objects that rotate about a fixed axis. In this chapter, we seek to extend Newton's Second Law to those systems.

9.1 | Moment of Inertia

Learning Objectives

By the end of this section, you will be able to:

- Define the physical concept of moment of inertia in terms of the mass distribution from the rotational axis
- Apply the parallel axis theorem to find the moment of inertia about any axis parallel to one already known
- Calculate the moment of inertia for compound objects

In **Section 4.**, we introduced rotational kinematics: the description of motion for a rotating rigid body with a fixed axis of rotation. We need to define two new quantities that will be helpful for analyzing the properties of rotating objects: **moment of inertia** and **torque**. With these properties defined, we will have two important tools we need for analyzing rotational dynamics. We will eventually formulate a version of Newton's Second Law for rotations.

Moment of Inertia

In translational motion, if we think about the role played by the quantity we call **mass**, it represents the **inertia** of an object. That is to say, the greater the mass of an object, the more it resists any changes to its motion. So, in Newton's Second Law, the acceleration of an object is inversely proportional to its mass:

$$a = F/m \tag{9.1}$$

For rotating objects, however, things are not quite so simple. The resistance of a rigid object to a change in its rotational motion depends not only upon its mass, but upon how that mass is distributed relative to the axis of rotation. This suggests that we have a new rotational variable to add to our list of our analogies between rotational and translational variables.

Imagine trying to spin two wheels, which have the same radius and total mass. However, they are constructed differently. The first wheel has most of its mass in a hub near its center, but the second wheel has most of its mass around the outside edge as far from the hub as possible. Suppose that, initially, both wheels are at rest. Which wheel would be harder for you to spin into rotation? It turns out that the second wheel, with more mass farther away from the axis of rotation, is harder to get spinning. It has greater resistance to a change in its rotational motion, i.e. greater rotational inertia.

The quantity which represents the rotational inertia of an object, and is therefore the counterpart for mass in the equations for rotational motion, is called the **moment of inertia** *I*, with units of $\text{kg} \cdot \text{m}^2$. It is found by summing, for each piece of mass in an object, the product of that mass times the square of its distance from the rotation axis.

$$I = \sum_j m_j r_j^2. \tag{9.2}$$

For now, we leave the expression in summation form, representing the moment of inertia of a system of point particles rotating about a fixed axis. We note that the moment of inertia of a single point particle about a fixed axis is simply mr^2, with *r* being the distance from the point particle to the axis of rotation. (Integral calculus can be used to calculate the moment of inertia of some regular-shaped rigid bodies whose mass is distributed continuously over a volume of some particular shape.)

The moment of inertia is the quantitative measure of rotational inertia, just as in translational motion, and mass is the quantitative measure of linear inertia—that is, the more massive an object is, the more inertia it has, and the greater is its resistance to change in linear velocity. Similarly, the greater the moment of inertia of a rigid body or system of particles, the greater is its resistance to change in angular velocity about a fixed axis of rotation. It is interesting to see how the moment of inertia varies with *r*, the distance to the axis of rotation of the mass particles in **Equation 9.2**. Rigid bodies and systems of particles with more mass concentrated at a greater distance from the axis of rotation have greater moments of inertia than bodies and systems of the same mass, but concentrated near the axis of rotation. In this way, we can see that a hollow cylinder has more rotational inertia than a solid cylinder of the same mass when rotating about an axis through the center.

Calculating Moment of Inertia

We defined the moment of inertia I of an object to be $I = \sum_i m_i r_i^2$ for all the point masses that make up the object. Because r is the distance to the axis of rotation from each piece of mass that makes up the object, the moment of inertia for any object depends on the chosen axis. To see this, let's take a simple example of two masses at the end of a massless (negligibly small mass) rod (**Figure 9.2**) and calculate the moment of inertia about two different axes. In this case, the summation over the masses is simple because the two masses at the end of the barbell can be approximated as point masses, and the sum therefore has only two terms.

In the case with the axis in the center of the barbell, each of the two masses m is a distance R away from the axis, giving a moment of inertia of

$$I_1 = mR^2 + mR^2 = 2mR^2.$$

In the case with the axis at the end of the barbell—passing through one of the masses—the moment of inertia is

$$I_2 = m(0)^2 + m(2R)^2 = 4mR^2.$$

From this result, we can conclude that it is twice as hard to rotate the barbell about the end than about its center.

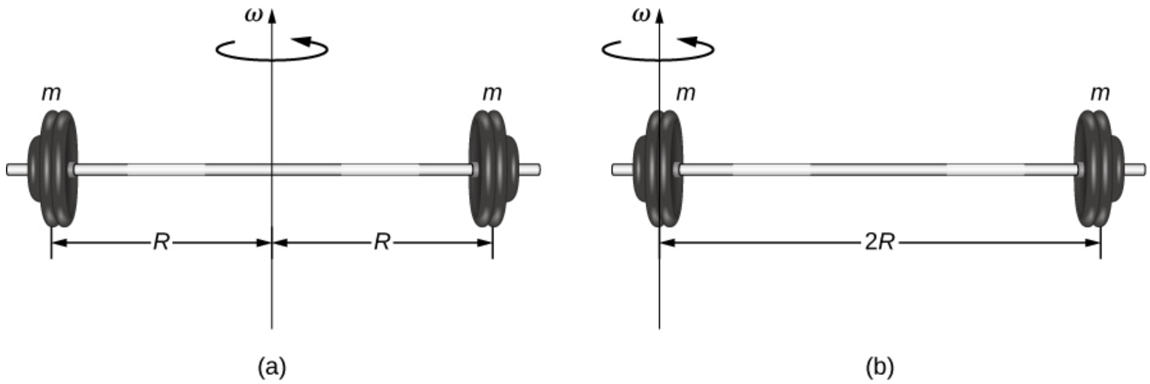


Figure 9.2 (a) A barbell with an axis of rotation through its center; (b) a barbell with an axis of rotation through one end.

Example 9.1

Moment of Inertia of a System of Particles

Six small washers are spaced 10 cm apart on a rod of negligible mass and 0.5 m in length. The mass of each washer is 20 g. The rod rotates about an axis located at 25 cm, as shown in **Figure 9.3**. (a) What is the moment of inertia of the system? (b) If the two washers closest to the axis are removed, what is the moment of inertia of the remaining four washers?

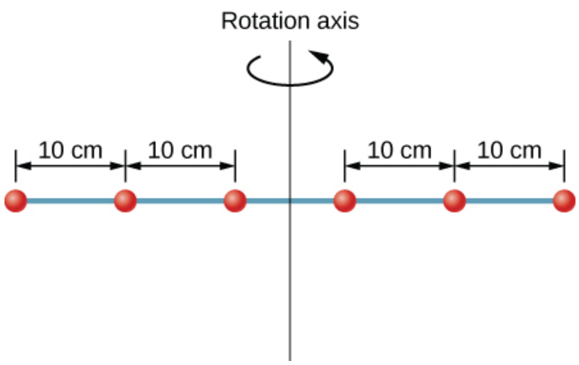


Figure 9.3 Six washers are spaced 10 cm apart on a rod of negligible mass and rotating about a vertical axis.

Strategy

a. We use the definition for moment of inertia for a system of particles and perform the summation to evaluate this quantity. The masses are all the same so we can pull that quantity in front of the summation symbol.

b. We do a similar calculation.

Solution

a. $I = \sum_j m_j r_j^2 = (0.02\ \text{kg})(2 \times (0.25\ \text{m})^2 + 2 \times (0.15\ \text{m})^2 + 2 \times (0.05\ \text{m})^2) = 0.0035\ \text{kg} \cdot \text{m}^2$.

b. $I = \sum_j m_j r_j^2 = (0.02\ \text{kg})(2 \times (0.25\ \text{m})^2 + 2 \times (0.15\ \text{m})^2) = 0.0034\ \text{kg} \cdot \text{m}^2$.

Significance

We can see the individual contributions to the moment of inertia. The masses close to the axis of rotation have a very small contribution. When we removed them, it had a very small effect on the moment of inertia.

As mentioned previously, the summation formula given in **Equation 9.2** can be turned (through the use of integral calculus) into a method to calculate moments of inertia for rigid bodies. For the present discussion, it is sufficient that **Figure 9.4** below gives the expressions for rotational inertia for common object shapes around specified axes.

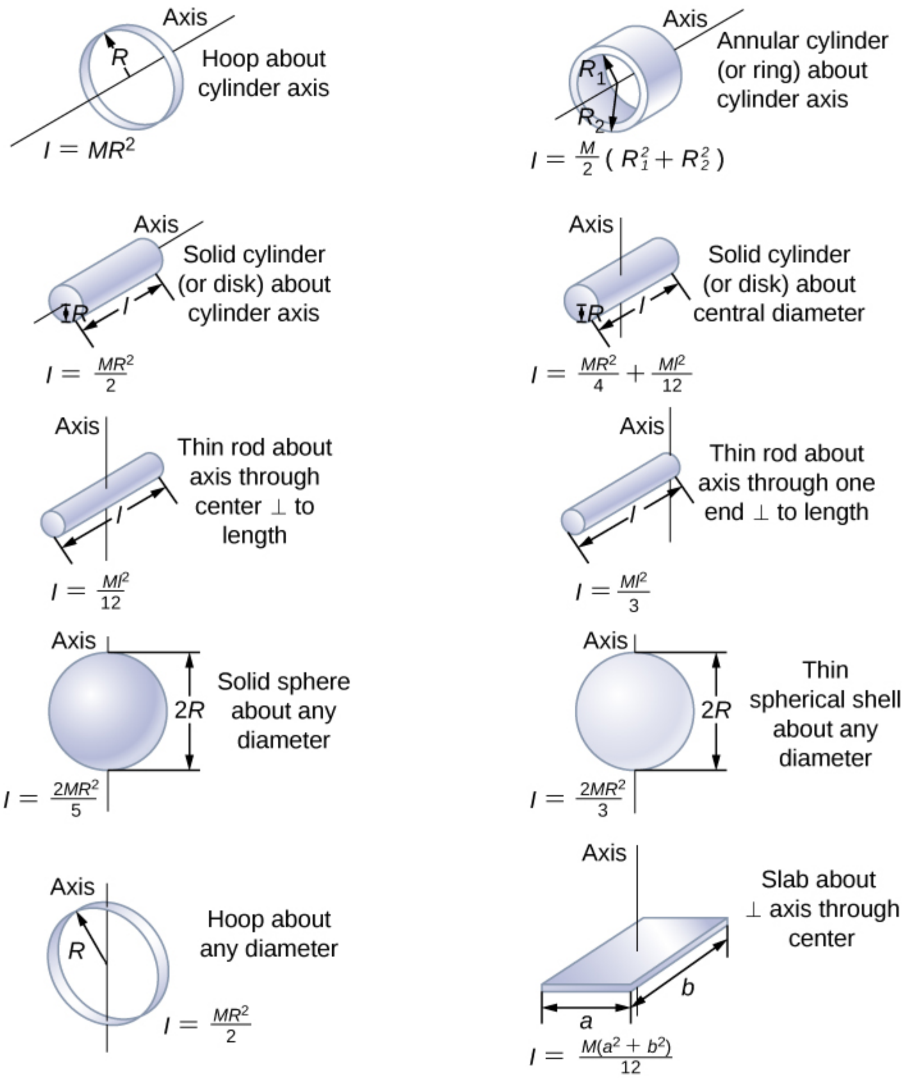


Figure 9.4 Values of rotational inertia for common shapes of objects.

Parallel Axis Theorem

Look at the two examples in **Figure 9.4** of the uniform thin rod. The difference between the moment of inertia of a rod about an axis through its middle ($\frac{1}{12}mL^2$) and about an axis through its end ($\frac{1}{3}mL^2$) is striking. The fact that I is smaller for an axis through the center than for an axis about one end of the rod in just an indication that, when rotating about the center, more of the mass is closer to the axis of rotation. This suggests that there might be a simpler method for determining the moment of inertia for a rod about any axis parallel to the axis through the center of mass. Such an axis is called a **parallel axis**. There is a theorem for this, called the **parallel-axis theorem**, which we state here but do not derive in this text.

Parallel-Axis Theorem

Let *m* be the mass of an object and let *d* be the distance from an axis through the object's center of mass to a new axis. Then we have

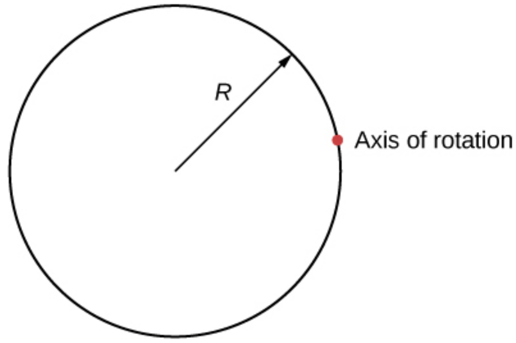
$$I_{\text{parallel-axis}} = I_{\text{center of mass}} + md^2. \tag{9.3}$$

Let's apply this to the rod examples solved above:

$$I_{\text{end}} = I_{\text{center of mass}} + md^2 = \frac{1}{12}mL^2 + m\left(\frac{L}{2}\right)^2 = \left(\frac{1}{12} + \frac{1}{4}\right)mL^2 = \frac{1}{3}mL^2.$$

This result agrees with the results shown in **Figure 9.4**. This is a useful equation that we apply in some of the examples and problems.

9.1 What is the moment of inertia of a cylinder of radius R and mass m about an axis through a point on the surface, as shown below?



Calculating the moment of inertia for compound objects

Now consider a compound object such as that in **Figure 9.5**, which depicts a thin disk at the end of a thin rod. This cannot be easily integrated to find the moment of inertia because it is not a uniformly shaped object. However, if we go back to the initial definition of moment of inertia as a summation, we can reason that a compound object's moment of inertia can be found from the sum of each part of the object:

$$I_{\text{total}} = \sum_i I_i. \tag{9.4}$$

It is important to note that the moments of inertia of the objects in **Equation 9.4** are *about a common axis*. In the case of this object, that would be a rod of length L rotating about its end, and a thin disk of radius R rotating about an axis shifted off of the center by a distance $L + R$, where R is the radius of the disk. Let's define the mass of the rod to be m_{r} and the mass of the disk to be m_{d}.

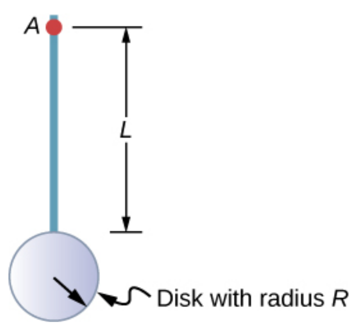


Figure 9.5 Compound object consisting of a disk at the end of a rod. The axis of rotation is located at A.

The moment of inertia of the rod is simply $\frac{1}{3}m_{\text{r}}L^2$, but we have to use the parallel-axis theorem to find the moment of inertia of the disk about the axis shown. The moment of inertia of the disk about its center is $\frac{1}{2}m_{\text{d}}R^2$ and we apply the parallel-axis theorem $I_{\text{parallel-axis}} = I_{\text{center of mass}} + md^2$ to find

$$I_{\text{parallel-axis}} = \frac{1}{2}m_{\text{d}}R^2 + m_{\text{d}}(L + R)^2.$$

Adding the moment of inertia of the rod plus the moment of inertia of the disk with a shifted axis of rotation, we find the moment of inertia for the compound object to be

$$I_{total} = \frac{1}{3}m_r L^2 + \frac{1}{2}m_d R^2 + m_d (L + R)^2.$$

Applying moment of inertia calculations to solve problems

Now let's examine some practical applications of moment of inertia calculations.

Example 9.2

Person on a Merry-Go-Round

A 25-kg child stands at a distance $r = 1.0\,\text{m}$ from the axis of a rotating merry-go-round (**Figure 9.6**). The merry-go-round can be approximated as a uniform solid disk with a mass of 500 kg and a radius of 2.0 m. Find the moment of inertia of this system.

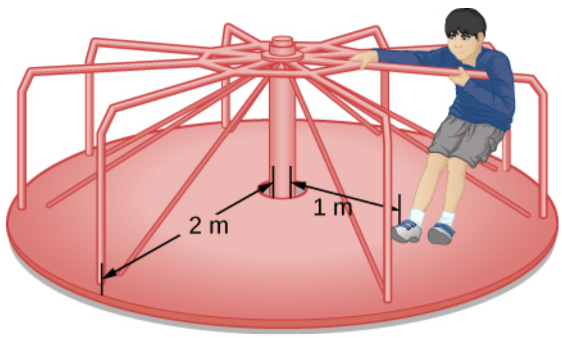


Figure 9.6 Calculating the moment of inertia for a child on a merry-go-round.

Strategy

This problem involves the calculation of a moment of inertia. We are given the mass and distance to the axis of rotation of the child as well as the mass and radius of the merry-go-round. Since the mass and size of the child are much smaller than the merry-go-round, we can approximate the child as a point mass. The notation we use is $m_c = 25\,\text{kg},\ r_c = 1.0\,\text{m},\ m_m = 500\,\text{kg},\ r_m = 2.0\,\text{m}$.

Our goal is to find $I_{total} = \sum_i I_i$.

Solution

For the child, $I_c = m_c r^2$, and for the merry-go-round, $I_m = \frac{1}{2}m_m r^2$. Therefore

$$I_{total} = 25(1)^2 + \frac{1}{2}(500)(2)^2 = 25 + 1000 = 1025\,\text{kg} \cdot \text{m}^2.$$

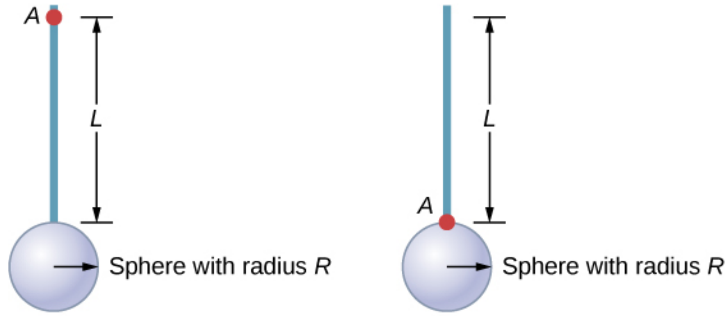
Significance

The value should be close to the moment of inertia of the merry-go-round by itself because it has much more mass distributed away from the axis than the child does.

Example 9.3

Rod and Solid Sphere

Find the moment of inertia of the rod and solid sphere combination about the two axes as shown below. The rod has length 0.5 m and mass 2.0 kg. The radius of the sphere is 20.0 cm and has mass 1.0 kg.



Strategy

Since we have a compound object in both cases, we can use the parallel-axis theorem to find the moment of inertia about each axis. In (a), the center of mass of the sphere is located at a distance $L+R$ from the axis of rotation. In (b), the center of mass of the sphere is located a distance R from the axis of rotation. In both cases, the moment of inertia of the rod is about an axis at one end. Refer to **Figure 9.4** for the moments of inertia for the individual objects.

a. $I_{\text{total}} = \sum_i I_i = I_{\text{Rod}} + I_{\text{Sphere}};$

$$I_{\text{Sphere}} = I_{\text{center of mass}} + m_{\text{Sphere}}(L+R)^2 = \frac{2}{5}m_{\text{Sphere}}R^2 + m_{\text{Sphere}}(L+R)^2;$$

$$I_{\text{total}} = I_{\text{Rod}} + I_{\text{Sphere}} = \frac{1}{3}m_{\text{Rod}}L^2 + \frac{2}{5}m_{\text{Sphere}}R^2 + m_{\text{Sphere}}(L+R)^2;$$

$$I_{\text{total}} = \frac{1}{3}(2.0\,\text{kg})(0.5\,\text{m})^2 + \frac{2}{5}(1.0\,\text{kg})(0.2\,\text{m})^2 + (1.0\,\text{kg})(0.5\,\text{m} + 0.2\,\text{m})^2;$$

$$I_{\text{total}} = (0.167 + 0.016 + 0.490)\,\text{kg}\cdot\text{m}^2 = 0.673\,\text{kg}\cdot\text{m}^2.$$

b. $I_{\text{Sphere}} = \frac{2}{5}m_{\text{Sphere}}R^2 + m_{\text{Sphere}}R^2;$

$$I_{\text{total}} = I_{\text{Rod}} + I_{\text{Sphere}} = \frac{1}{3}m_{\text{Rod}}L^2 + \frac{2}{5}m_{\text{Sphere}}R^2 + m_{\text{Sphere}}R^2;$$

$$I_{\text{total}} = \frac{1}{3}(2.0\,\text{kg})(0.5\,\text{m})^2 + \frac{2}{5}(1.0\,\text{kg})(0.2\,\text{m})^2 + (1.0\,\text{kg})(0.2\,\text{m})^2;$$

$$I_{\text{total}} = (0.167 + 0.016 + 0.04)\,\text{kg}\cdot\text{m}^2 = 0.223\,\text{kg}\cdot\text{m}^2.$$

Significance

Using the parallel-axis theorem eases the computation of the moment of inertia of compound objects. We see that the moment of inertia is greater in (a) than (b). This is because the axis of rotation is closer to the center of mass of the system in (b). The simple analogy is that of a rod. The moment of inertia about one end is $\frac{1}{3}mL^2$, but the moment of inertia through the center of mass along its length is $\frac{1}{12}mL^2$.

9.2 | Torque

Learning Objectives

By the end of this section, you will be able to:

- Describe how the magnitude of a torque depends on the magnitude of the lever arm and the angle the force vector makes with the lever arm
- Determine the sign (positive or negative) of a torque from the direction of the rotation it would induce
- Calculate individual torques about a common axis and sum them to find the net torque

An important quantity for describing the dynamics of a rotating rigid body is **torque**. We see the application of torque in many ways in our world. We all have an intuition about torque, as when we use a large wrench to unscrew a stubborn bolt. Torque is at work in unseen ways, as when we press on the accelerator in a car, causing the engine to put additional torque on the drive train. Or every time we move our bodies from a standing position, we apply a torque to our limbs. In this section, we define torque and make an argument for the equation for calculating torque for a rigid body with fixed-axis rotation.

Defining Torque

So far we have defined many variables that are rotational equivalents to their translational counterparts. Let's consider what the counterpart to force must be. Since forces change the translational motion of objects, the rotational counterpart must be related to changing the rotational motion of an object about an axis. We call this rotational counterpart **torque**.

In everyday life, we rotate objects about an axis all the time, so intuitively we already know much about torque. Consider, for example, how we rotate a door to open it. First, we know that a door opens slowly if we push too close to its hinges; it is more efficient to rotate a door open if we push far from the hinges. Second, we know that we should push perpendicular to the plane of the door; if we push parallel to the plane of the door, we are not able to rotate it. Third, the larger the force, the more effective it is in opening the door; the harder you push, the more rapidly the door opens. The first point implies that the farther the force is applied from the axis of rotation, the greater the angular acceleration; the second implies that the effectiveness depends on the angle at which the force is applied; the third implies that the magnitude of the force must also be part of the equation. Note that for rotation in a plane, torque has two possible directions. Torque is either clockwise or counterclockwise relative to the chosen pivot point. **Figure 9.7** shows counterclockwise rotations.

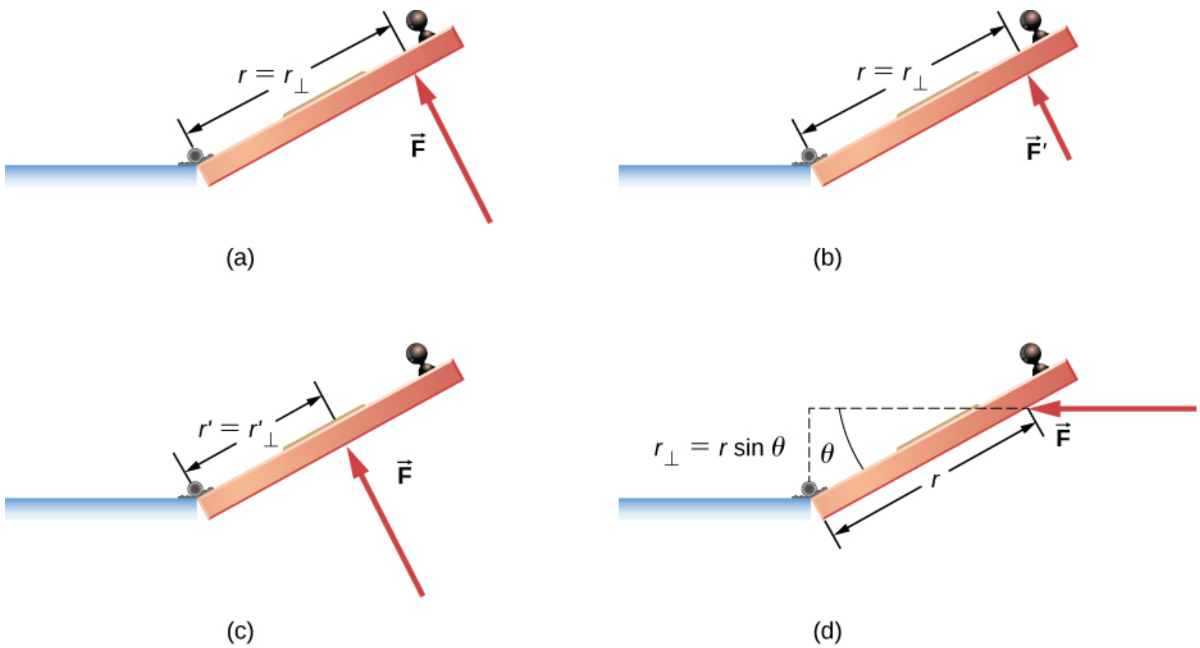


Figure 9.7 Torque is the turning or twisting effectiveness of a force, illustrated here for door rotation on its hinges (as viewed from overhead). Torque has both magnitude and direction. (a) A counterclockwise torque is produced by a force $\vec{\mathbf{F}}$ acting at a distance *r* from the hinges (the pivot point). (b) A smaller counterclockwise torque is produced when a smaller force $\vec{\mathbf{F}}\,'$ acts at the same distance *r* from the hinges. (c) The same force as in (a) produces a smaller counterclockwise torque when applied at a smaller distance from the hinges. (d) A smaller counterclockwise torque is produced by the same magnitude force as (a) acting at the same distance as (a) but at an angle θ that is less than $90°$.

Now let's consider how to define torques mathematically.

Torque

When a force $\vec{\mathbf{F}}$ is applied to a point *P* whose position is $\vec{\mathbf{r}}$ relative to *O* (**Figure 9.8**), the magnitude of the torque τ around *O* is:

$$\left|\vec{\boldsymbol{\tau}}\right| == rF\sin\theta,$$

where θ is the angle between the vectors $\vec{\mathbf{r}}$ and $\vec{\mathbf{F}}$. The SI unit of torque is newtons times meters, usually written as $\text{N}\cdot\text{m}$. The quantity $r_{\perp} = r\sin\theta$ is the perpendicular distance from *O* to the line determined by the vector $\vec{\mathbf{F}}$ and is called the **lever arm**. Note that the greater the lever arm, the greater the magnitude of the torque. In terms of the lever arm, the magnitude of the torque is

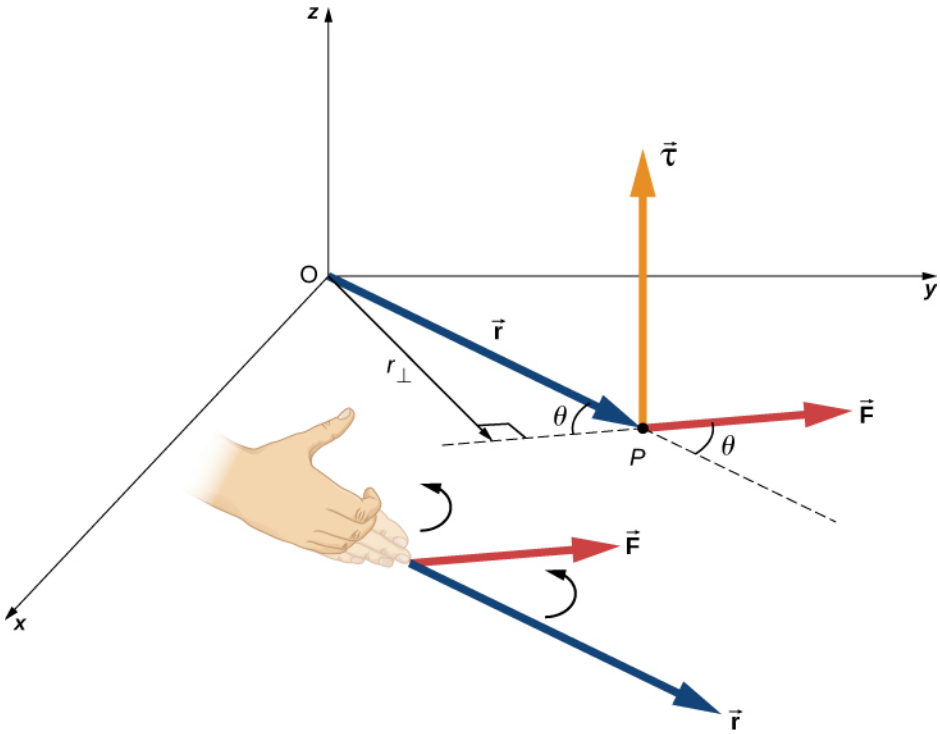


Figure 9.8 The torque is perpendicular to the plane defined by $\vec{\mathbf{r}}$ and $\vec{\mathbf{F}}$ and its direction is determined by the direction of the rotation it would cause, as viewed from above. Counterclockwise torques are positive; clockwise torques are negative.

$$|\vec{\boldsymbol{\tau}}| = r_\perp F. \quad (9.5)$$

Torque is a vector quantity, and in the case of one-dimensional rotations, its direction is indicated by a positive or negative sign. The sign of the torque vector is taken to be positive if it is a counterclockwise torque, and the sign is negative if it is a clockwise torque.

If we consider a disk that is free to rotate about an axis through the center, as shown in **Figure 9.9**, we can see how the angle between the radius $\vec{\mathbf{r}}$ and the force $\vec{\mathbf{F}}$ affects the magnitude of the torque. If the angle is zero, the torque is zero; if the angle is $90°$, the torque is maximum. The torque in **Figure 9.9** is positive because the disk rotates counterclockwise due to the torque, in the same direction as a positive angular acceleration.

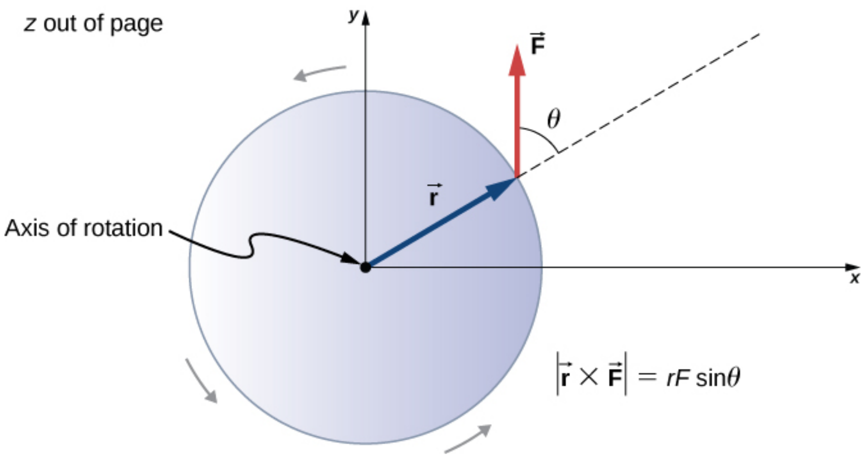


Figure 9.9 A disk is free to rotate about its axis through the center. The magnitude of the torque on the disk is $rF\sin\theta$. When $\theta = 0°$, the torque is zero and the disk does not rotate. When $\theta = 90°$, the torque is maximum and the disk rotates with maximum angular acceleration.

Any number of torques can be calculated about a given axis. The individual torques add to produce a net torque about the axis. When the appropriate sign (positive or negative) is assigned to the magnitudes of individual torques about a specified axis, the net torque about the axis is the sum of the individual torques:

$$\vec{\boldsymbol{\tau}}_{\text{net}} = \sum_i \vec{\boldsymbol{\tau}}_i. \tag{9.6}$$

Calculating Net Torque for Rigid Bodies on a Fixed Axis

In the following examples, we calculate the torque both abstractly and as applied to a rigid body.

We first introduce a problem-solving strategy.

Problem-Solving Strategy: Finding Net Torque

1. Choose a coordinate system with the pivot point or axis of rotation as the origin of the selected coordinate system.
2. Determine the angle between the lever arm $\vec{\mathbf{r}}$ and the force vector.
3. Take the cross product of $\vec{\mathbf{r}}$ and $\vec{\mathbf{F}}$ to determine if the torque is positive or negative about the pivot point or axis.
4. Evaluate the magnitude of the torque using $r_\perp F$.
5. Assign the appropriate sign, positive or negative, to the magnitude.
6. Sum the torques to find the net torque.

Example 9.4

Calculating Torque

Four forces are shown in **Figure 9.10** at particular locations and orientations with respect to a given *xy*-coordinate system. Find the torque due to each force about the origin, then use your results to find the net torque about the origin.

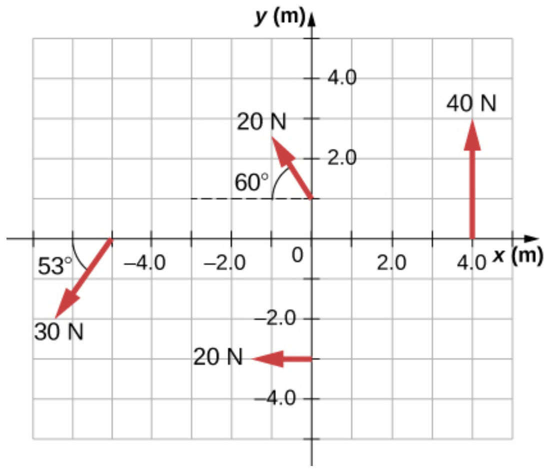


Figure 9.10 Four forces producing torques.

Strategy

This problem requires calculating torque. All known quantities—forces with directions and lever arms—are given in the figure. The goal is to find each individual torque and the net torque by summing the individual torques. Be careful to assign the correct sign to each torque by using the cross product of $\vec{\mathbf{r}}$ and the force vector $\vec{\mathbf{F}}$.

Solution

Use $|\vec{\tau}| = r_{\perp} F = rF\sin\theta$ to find the magnitude and then determine the sign of the torque by the direction of rotation it would cause.

The torque from force 40 N in the first quadrant is given by $(4)(40)\sin 90° = 160\,\text{N}\cdot\text{m}$.

The torque from this force would tend to cause a counter-clockwise rotation about the origin, so the torque is positive.

The torque from force 20 N in the third quadrant is given by $-(3)(20)\sin 90° = -60\,\text{N}\cdot\text{m}$.

The torque from this force would tend to cause a clockwise rotation about the origin, so the torque is negative.

The torque from force 30 N in the third quadrant is given by $(5)(30)\sin 53° = 120\,\text{N}\cdot\text{m}$.

The torque from this force would tend to cause a counter-clockwise rotation about the origin, so the torque is positive.

The torque from force 20 N in the second quadrant is given by $(1)(20)\sin 30° = 10\,\text{N}\cdot\text{m}$.

The torque from this force would tend to cause a counter-clockwise rotation about the origin, so the torque is positive.

The net torque is therefore $\tau_{\text{net}} = \sum_i |\boldsymbol{\tau}_i| = 160 - 60 + 120 + 10 = 230\,\text{N}\cdot\text{m}.$

Significance

Note that each force that acts in the counterclockwise direction has a positive torque, whereas each force that acts in the clockwise direction has a negative torque. The torque is greater when the distance, force, or perpendicular components are greater.

Example 9.5

Calculating Torque on a rigid body

Figure 9.11 shows several forces acting at different locations and angles on a flywheel. We have $|\vec{F}_1| = 20\,\text{N}$, $|\vec{F}_2| = 30\,\text{N}$, $|\vec{F}_3| = 30\,\text{N}$, and $r = 0.5\,\text{m}$. Find the net torque on the flywheel about an axis through the center.

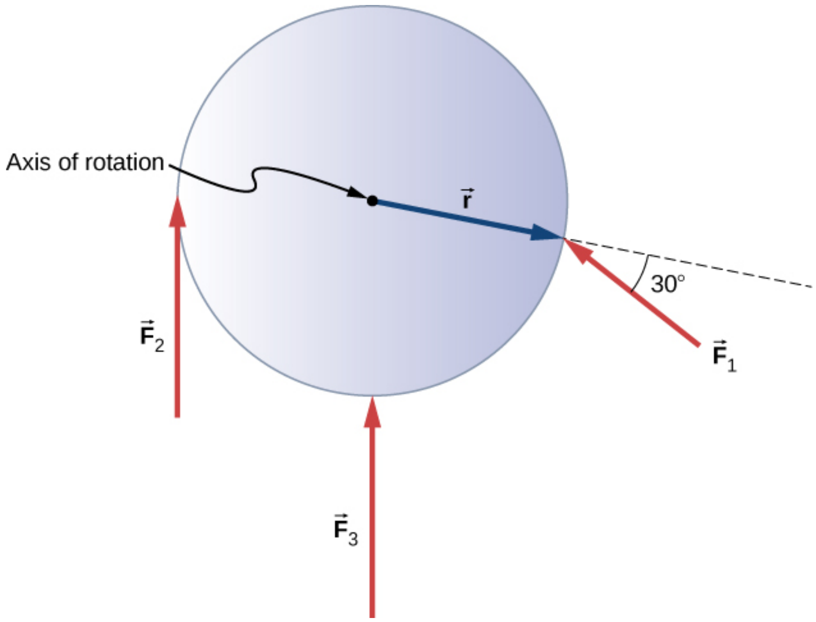


Figure 9.11 Three forces acting on a flywheel.

Strategy

We calculate each torque individually, using the cross product, and determine the sign of the torque. Then we sum the torques to find the net torque.

Solution

We start with $\vec{F}_1$. If we look at Figure 9.11, we see that $\vec{F}_1$ makes an angle of $90° + 60°$ with the radius vector $\vec{r}$. Taking the cross product, we see that it is out of the page and so is positive. We also see this from calculating its magnitude:

$$|\vec{\tau}_1| = rF_1 \sin 150° = 0.5\,\text{m}(20\,\text{N})(0.5) = 5.0\,\text{N} \cdot \text{m}.$$

Next we look at $\vec{F}_2$. The angle between $\vec{F}_2$ and $\vec{r}$ is $90°$ and the cross product is into the page so the torque is negative. Its value is

$$|\vec{\tau}_2| = -rF_2 \sin 90° = -0.5\,\text{m}(30\,\text{N}) = -15.0\,\text{N} \cdot \text{m}.$$

When we evaluate the torque due to $\vec{F}_3$, we see that the angle it makes with $\vec{r}$ is zero so $\vec{r} \times \vec{F}_3 = 0.$ Therefore, $\vec{F}_3$ does not produce any torque on the flywheel.

We evaluate the sum of the torques:

$$\tau_{\text{net}} = \sum_i |\boldsymbol{\tau}_i| = 5 - 15 = -10\,\text{N} \cdot \text{m}.$$

Significance

The axis of rotation is at the center of mass of the flywheel. Since the flywheel is on a fixed axis, it is not free to translate. If it were on a frictionless surface and not fixed in place, $\vec{\mathbf{F}}_3$ would cause the flywheel to translate, as well as $\vec{\mathbf{F}}_1$. Its motion would be a combination of translation and rotation.

9.2 Check Your Understanding A large ocean-going ship runs aground near the coastline, similar to the fate of the *Costa Concordia*, and lies at an angle as shown below. Salvage crews must apply a torque to right the ship in order to float the vessel for transport. A force of 5.0×10^5 N acting at point *A* must be applied to right the ship. What is the torque about the point of contact of the ship with the ground (**Figure 9.12**)?

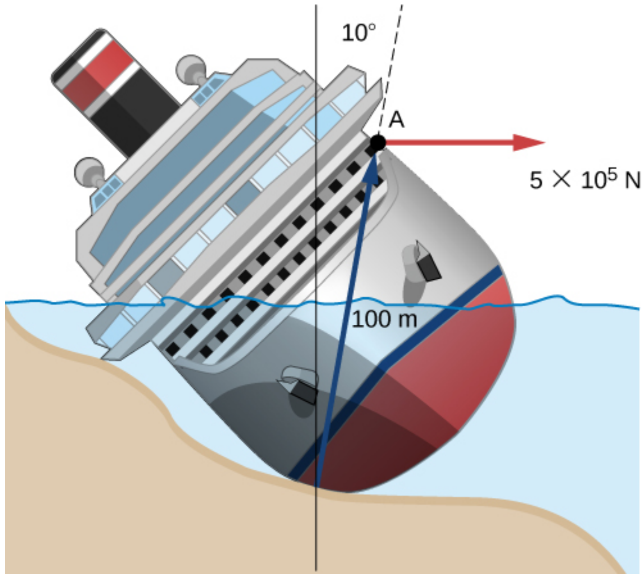


Figure 9.12 A ship runs aground and tilts, requiring torque to be applied to return the vessel to an upright position.

9.3 | Newton's Second Law for Rotations

Learning Objectives

By the end of this section, you will be able to:

- Calculate the torques on rotating systems about a fixed axis to find the angular acceleration
- Explain how changes in the moment of inertia of a rotating system affect angular acceleration with a fixed applied torque

In this section, we put together all the pieces learned so far in this chapter to analyze the dynamics of rotating rigid bodies. We introduce the rotational equivalent to Newton's second law of motion and apply it to rigid bodies with fixed-axis rotation.

Newton's Second Law for Rotation

We have thus far found many counterparts to the translational terms used throughout this text, most recently, torque, the rotational analog to force. This raises the question: Is there an analogous equation to Newton's second law, $\Sigma \vec{\mathbf{F}} = m\vec{\mathbf{a}}$, which involves torque and rotational motion? To investigate this, we start with Newton's second law for a single particle rotating around an axis and executing circular motion. Let's exert a force $\vec{\mathbf{F}}$ on a point mass *m* that is at a distance r from a pivot point (**Figure 9.13**). The particle is constrained to move in a circular path with fixed radius and the force is tangent to the circle. We apply Newton's second law to determine the magnitude of the acceleration $a = F/m$ in the direction of $\vec{\mathbf{F}}$. Recall that the magnitude of the tangential acceleration is proportional to the magnitude of the angular acceleration by $a = r\alpha$. Substituting this expression into Newton's second law, we obtain

$$F = mr\alpha.$$

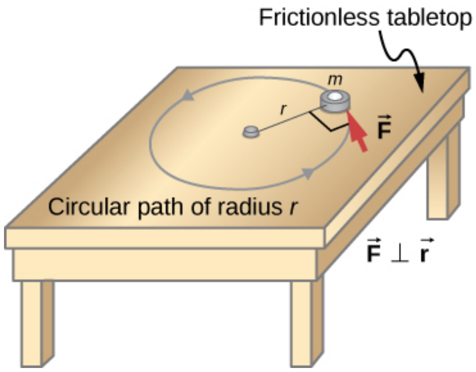


Figure 9.13 An object is supported by a horizontal frictionless table and is attached to a pivot point by a cord that supplies centripetal force. A force $\vec{\mathbf{F}}$ is applied to the object perpendicular to the radius r, causing it to accelerate about the pivot point. The force is perpendicular to r.

Multiply both sides of this equation by r,

$$rF = mr^2\alpha.$$

Note that the left side of this equation is the torque about the axis of rotation, where r is the lever arm and F is the force, perpendicular to r. Recall that the moment of inertia for a point particle is $I = mr^2$. The torque applied perpendicularly to the point mass in **Figure 9.13** is therefore

$$\tau = I\alpha.$$

The torque on the particle is equal to the moment of inertia about the rotation axis times the angular acceleration. We can generalize this equation to a rigid body rotating about a fixed axis.

Newton's Second Law for Rotation

If more than one torque acts on a rigid body about a fixed axis, then the sum of the torques equals the moment of inertia times the angular acceleration:

$$\sum_i \tau_i = I\alpha. \quad (9.7)$$

The term $I\alpha$ is a scalar quantity and can be positive or negative (counterclockwise or clockwise) depending upon the sign of the net torque. Remember the convention that counterclockwise angular acceleration is positive. Thus, if a rigid body is rotating clockwise and experiences a positive torque (counterclockwise), the angular acceleration is positive.

Equation 9.7 is **Newton's second law for rotation** and tells us how to relate torque, moment of inertia, and rotational kinematics. This is called the equation for **rotational dynamics**. With this equation, we can solve a whole class of problems involving force and rotation. It makes sense that the relationship for how much force it takes to rotate a body would include the moment of inertia, since that is the quantity that tells us how easy or hard it is to change the rotational motion of an object.

Applying the Rotational Dynamics Equation

Before we apply the rotational dynamics equation to some everyday situations, let's review a general problem-solving strategy for use with this category of problems.

Problem-Solving Strategy: Rotational Dynamics

1. Examine the situation to determine that torque and mass are involved in the rotation. Draw a careful sketch of the situation.
2. Determine the system of interest.
3. Draw a free-body diagram. That is, draw and label all external forces acting on the system of interest.
4. Identify the pivot point. If the object is in equilibrium, it must be in equilibrium for all possible pivot points—chose the one that simplifies your work the most.
5. Apply $\sum_i \tau_i = I\alpha$, the rotational equivalent of Newton's second law, to solve the problem. Care must be taken to use the correct moment of inertia and to consider the torque about the point of rotation.
6. As always, check the solution to see if it is reasonable.

Example 9.6

Calculating the Effect of Mass Distribution on a Merry-Go-Round

Consider the father pushing a playground merry-go-round in Figure 9.14. He exerts a force of 250 N at the edge of the 50.0-kg merry-go-round, which has a 1.50-m radius. Calculate the angular acceleration produced (a) when no one is on the merry-go-round and (b) when an 18.0-kg child sits 1.25 m away from the center. Consider the merry-go-round itself to be a uniform disk with negligible friction.

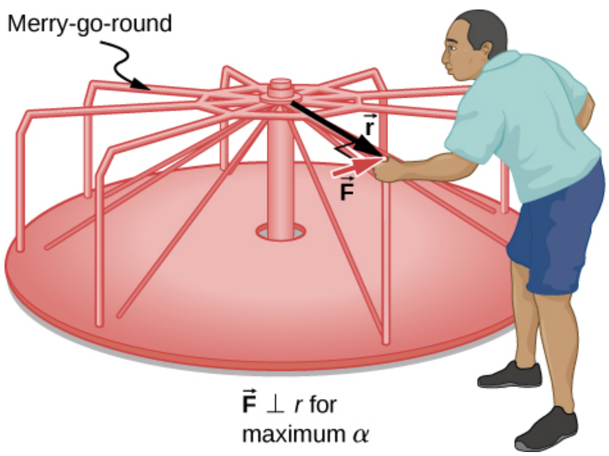


Figure 9.14 A father pushes a playground merry-go-round at its edge and perpendicular to its radius to achieve maximum torque.

Strategy

The net torque is given directly by the expression $\sum_i \tau_i = I\alpha$, To solve for α , we must first calculate the net torque τ (which is the same in both cases) and moment of inertia I (which is greater in the second case).

Solution

a. The moment of inertia of a solid disk about this axis is given in Figure 9.4 to be

$$\frac{1}{2}MR^2.$$

We have $M = 50.0\text{ kg}$ and $R = 1.50\text{ m}$, so

$$I = (0.500)(50.0\ \text{kg})(1.50\ \text{m})^2 = 56.25\ \text{kg-m}^2.$$

To find the net torque, we note that the applied force is perpendicular to the radius and friction is negligible, so that

$$\tau = rF\sin\theta = (1.50\ \text{m})(250.0\ \text{N}) = 375.0\ \text{N-m}.$$

Now, after we substitute the known values, we find the angular acceleration to be

$$\alpha = \frac{\tau}{I} = \frac{375.0\ \text{N-m}}{56.25\ \text{kg-m}^2} = 6.67\frac{\text{rad}}{\text{s}^2}.$$

b. We expect the angular acceleration for the system to be less in this part because the moment of inertia is greater when the child is on the merry-go-round. To find the total moment of inertia I, we first find the child's moment of inertia I_c by approximating the child as a point mass at a distance of 1.25 m from the axis. Then

$$I_c = mR^2 = (18.0\ \text{kg})(1.25\ \text{m})^2 = 28.13\ \text{kg-m}^2.$$

The total moment of inertia is the sum of the moments of inertia of the merry-go-round and the child (about the same axis):

$$I = 28.13\ \text{kg-m}^2 + 56.25\ \text{kg-m}^2 = 84.38\ \text{kg-m}^2.$$

Substituting known values into the equation for α gives

$$\alpha = \frac{\tau}{I} = \frac{375.0\ \text{N-m}}{84.38\ \text{kg-m}^2} = 4.44\frac{\text{rad}}{\text{s}^2}.$$

Significance

The angular acceleration is less when the child is on the merry-go-round than when the merry-go-round is empty, as expected. The angular accelerations found are quite large, partly due to the fact that friction was considered to be negligible. If, for example, the father kept pushing perpendicularly for 2.00 s, he would give the merry-go-round an angular velocity of 13.3 rad/s when it is empty but only 8.89 rad/s when the child is on it. In terms of revolutions per second, these angular velocities are 2.12 rev/s and 1.41 rev/s, respectively. The father would end up running at about 50 km/h in the first case.

9.3 Check Your Understanding The fan blades on a jet engine have a moment of inertia $30.0\ \text{kg-m}^2$. In 10 s, they rotate counterclockwise from rest up to a rotation rate of 20 rev/s. (a) What torque must be applied to the blades to achieve this angular acceleration? (b) What is the torque required to bring the fan blades rotating at 20 rev/s to a rest in 20 s?

CHAPTER 9 REVIEW

KEY TERMS

lever arm perpendicular distance from the line that the force vector lies on to a given axis

moment of inertia rotational mass of rigid bodies that relates to how easy or hard it will be to change the angular velocity of the rotating rigid body

Newton's second law for rotation sum of the torques on a rotating system equals its moment of inertia times its angular acceleration

parallel axis axis of rotation that is parallel to an axis about which the moment of inertia of an object is known

parallel-axis theorem if the moment of inertia is known for a given axis, it can be found for any axis parallel to it

rotational dynamics analysis of rotational motion using the net torque and moment of inertia to find the angular acceleration

torque cross product of a force and a lever arm to a given axis

KEY EQUATIONS

Moment of inertia of a group of point masses	$I = \sum_j m_j r_j^2.$
Parallel axis theorem	$I_{\text{parallel-axis}} = I_{\text{center of mass}} + md^2.$
Moment of inertia of a compound object	$I_{\text{total}} = \sum_i I_i.$
Torque exerted by a single force	$\lvert \vec{\tau} \rvert = rF\sin\theta = r_\perp F$
Net torque due to multiple forces	$\vec{\tau}_{\text{net}} = \sum_i \vec{\tau}_i.$
Newton's Second Law for Rotations	$\sum_i \tau_i = I\alpha.$

SUMMARY

9.1 Moment of Inertia

- The moment of inertia for a system of point particles rotating about a fixed axis is $I = \sum_j m_j r_j^2$, where m_j is the mass of the point particle and r_j is the distance of the point particle to the rotation axis. Because of the r^2 term, the moment of inertia increases as the square of the distance to the fixed rotational axis. The moment of inertia is the rotational counterpart to the mass in linear motion.
- For objects of uniform density with regular, simple shapes, integral calculus can be used to calculate their moments of inertia. Examples of these results are given in **Figure 9.4**.
- Moment of inertia is larger when an object's mass is farther from the axis of rotation.
- It is possible to find the moment of inertia of an object about a new axis of rotation once it is known for a parallel axis. This is called the parallel axis theorem given by $I_{\text{parallel-axis}} = I_{\text{center of mass}} + md^2$, where d is the distance from the initial axis to the parallel axis.
- Moment of inertia for a compound object is simply the sum of the moments of inertia for each individual object that

makes up the compound object.

9.2 Torque

- The magnitude of a torque about a fixed axis is calculated by finding the lever arm to the point where the force is applied and using the relation $|\vec{\tau}| = r_\perp F$, where $r_\perp$ is the perpendicular distance from the axis to the line upon which the force vector lies.
- The sign of the torque is positive if it acts in the counterclockwise direction, and negative if it acts in the clockwise direction.
- The net torque can be found from summing the individual torques about a given axis.

9.3 Newton's Second Law for Rotations

- Newton’s second law for rotation, $\sum_i \tau_i = I\alpha$, says that the sum of the torques on a rotating system about a fixed axis equals the product of the moment of inertia and the angular acceleration. This is the rotational analog to Newton’s second law of linear motion.

CONCEPTUAL QUESTIONS

9.1 Moment of Inertia

1. What if another planet the same size as Earth were put into orbit around the Sun along with Earth. Would the moment of inertia of the system increase, decrease, or stay the same?

2. A vast nebula of matter, out of which a solar system will eventually be born, is shown in the following figure. The matter is rotating around the center of the nebula. We know that, over time, because of the gravitational force, the matter in this nebula will collapse to occupy a smaller portion of space. What will happen to the moment of inertia of the system as it collapses?

Figure 9.15 Image courtesy of NASA-JPL/Caltech

3. A solid sphere is rotating about an axis through its center. Another hollow sphere of the same mass and radius is rotating about its axis through the center. Which sphere has a greater moment of inteita?

4. If a child walks toward the center of a merry-go-round, does the moment of inertia increase or decrease?

5. A discus thrower rotates with a discus in his hand before letting it go. (a) How does his moment of inertia change after releasing the discus? (b) What would be a good approximation to use in calculating the moment of inertia of the discus thrower and discus?

6. Does increasing the number of blades on a propeller increase or decrease its moment of inertia, and why?

7. The moment of inertia of a long rod spun around an axis through one end perpendicular to its length is $mL^2/3$. Why is this moment of inertia greater than it would be if you spun a point mass m at the location of the center of mass of the rod (at $L/2$) (that would be $mL^2/4$)?

8. Why is the moment of inertia of a hoop that has a mass M and a radius R greater than the moment of inertia of a disk that has the same mass and radius?

9.2 Torque

9. What three factors affect the torque created by a force relative to a specific pivot point?

10. Give an example in which a small force exerts a large torque. Give another example in which a large force exerts a small torque.

11. When reducing the mass of a racing bike, the greatest benefit is realized from reducing the mass of the tires and wheel rims. Why does this allow a racer to achieve greater accelerations than would an identical reduction in the mass of the bicycle’s frame?

12. Can a single force produce a zero torque?

13. Can a set of forces have a net torque that is zero and a net force that is not zero?

14. Can a set of forces have a net force that is zero and a net torque that is not zero?

15. In the expression $\vec{\mathbf{r}} \times \vec{\mathbf{F}}$ can $|\vec{\mathbf{r}}|$ ever be less than the lever arm? Can it be equal to the lever arm?

9.3 Newton's Second Law for Rotations

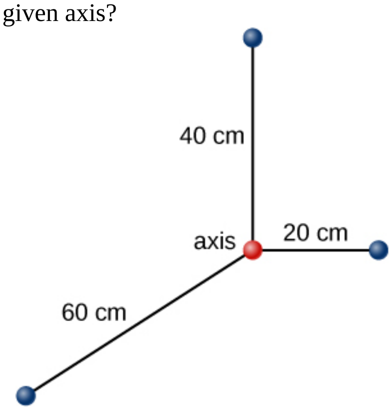
16. If you were to stop a spinning wheel with a constant force, where on the wheel would you apply the force to produce the maximum negative acceleration?

17. A rod is pivoted about one end. Two forces $\vec{\mathbf{F}}$ and $-\vec{\mathbf{F}}$ are applied to it. Under what circumstances will the rod not rotate?

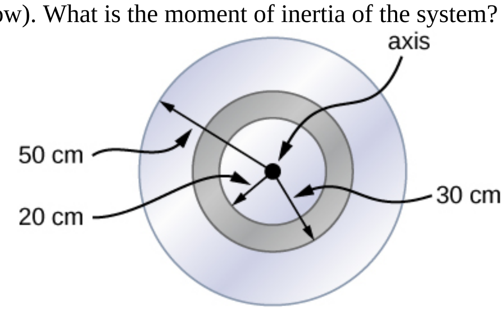
PROBLEMS

9.1 Moment of Inertia

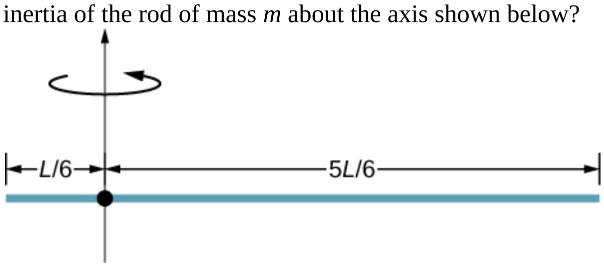
18. A system of point particles is shown in the following figure. Each particle has mass 0.3 kg and they all lie in the same plane. What is the moment of inertia of the system about the given axis?



19. A system consists of a disk of mass 2.0 kg and radius 50 cm upon which is mounted an annular cylinder of mass 1.0 kg with inner radius 20 cm and outer radius 30 cm (see below). What is the moment of inertia of the system?

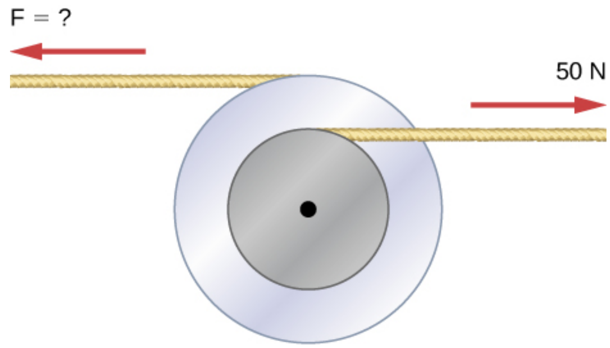


20. Using the parallel axis theorem, what is the moment of inertia of the rod of mass *m* about the axis shown below?



9.2 Torque

21. Two flywheels of negligible mass and different radii are bonded together and rotate about a common axis (see below). The smaller flywheel of radius 30 cm has a cord that has a pulling force of 50 N on it. What pulling force needs to be applied to the cord connecting the larger flywheel of radius 50 cm such that the combination does not rotate?

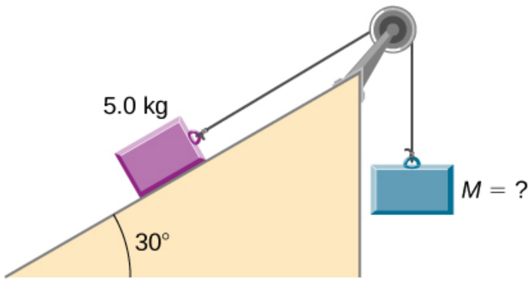


22. The cylindrical head bolts on a car are to be tightened with a torque of 62.0 $N \cdot m$. If a mechanic uses a wrench of length 20 cm, what perpendicular force must he exert on the end of the wrench to tighten a bolt correctly?

23. (a) When opening a door, you push on it perpendicularly with a force of 55.0 N at a distance of 0.850 m from the hinges. What torque are you exerting relative to the hinges? (b) Does it matter if you push at the same height as the hinges? There is only one pair of hinges.

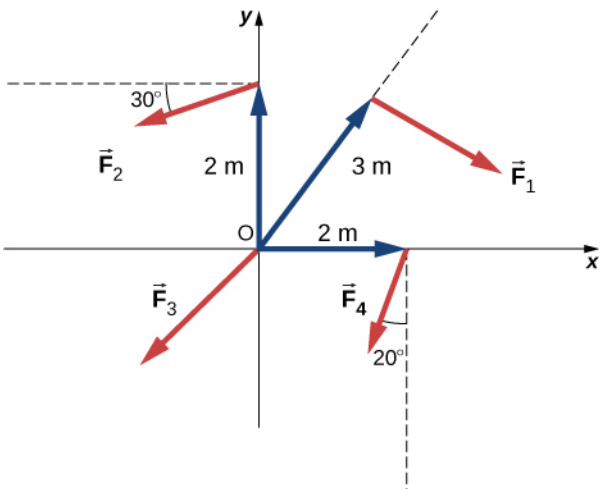
24. When tightening a bolt, you push perpendicularly on a wrench with a force of 165 N at a distance of 0.140 m from the center of the bolt. How much torque are you exerting in newton-meters (relative to the center of the bolt)?

25. What hanging mass must be placed on the cord to keep the pulley from rotating (see the following figure)? The mass on the frictionless plane is 5.0 kg. The inner radius of the pulley is 20 cm and the outer radius is 30 cm.

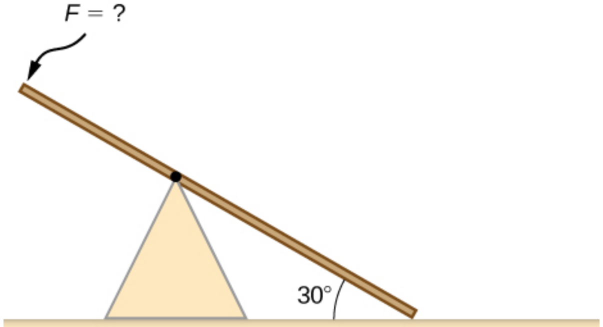


26. A simple pendulum consists of a massless tether 50 cm in length connected to a pivot and a small mass of 1.0 kg attached at the other end. What is the torque about the pivot when the pendulum makes an angle of $40°$ with respect to the vertical?

27. Calculate the torque about the *z*-axis that is out of the page at the origin in the following figure, given that $F_1 = 3\,\text{N}$, $F_2 = 2\,\text{N}$, $F_3 = 3\,\text{N}$, $F_4 = 1.8\,\text{N}$.

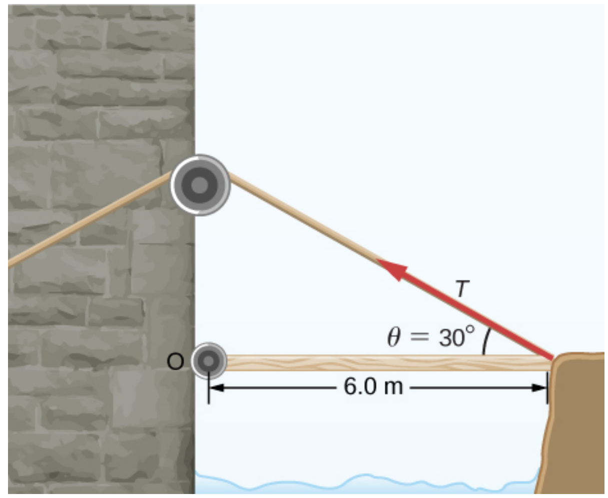


28. A seesaw has length 10.0 m and uniform mass 10.0 kg and is resting at an angle of $30°$ with respect to the ground (see the following figure). The pivot is located at 6.0 m. What magnitude of force needs to be applied perpendicular to the seesaw at the raised end so as to allow the seesaw to barely start to rotate?



29. A pendulum consists of a rod of mass 1 kg and length 1 m connected to a pivot with a solid sphere attached at the other end with mass 0.5 kg and radius 30 cm. What is the torque about the pivot when the pendulum makes an angle of $30°$ with respect to the vertical?

30. A torque of $5.00 \times 10^3\,\text{N} \cdot \text{m}$ is required to raise a drawbridge (see the following figure). What is the tension necessary to produce this torque? Would it be easier to raise the drawbridge if the angle θ were larger or smaller?



9.3 Newton's Second Law for Rotations

31. You have a grindstone (a disk) that is 90.0 kg, has a 0.340-m radius, and is turning at 90.0 rpm, and you press a steel axe against it with a radial force of 20.0 N. (a) Assuming the kinetic coefficient of friction between steel and stone is 0.20, calculate the angular acceleration of the grindstone. (b) How many turns will the stone make before coming to rest?

32. Suppose you exert a force of 180 N tangential to a 0.280-m-radius, 75.0-kg grindstone (a solid disk). (a)What torque is exerted? (b) What is the angular acceleration assuming negligible opposing friction? (c) What is the angular acceleration if there is an opposing frictional force of 20.0 N exerted 1.50 cm from the axis?

33. A flywheel ($I = 50\,\text{kg-m}^2$) starting from rest acquires an angular velocity of 200.0 rad/s while subject to a constant torque from a motor for 5 s. (a) What is the angular acceleration of the flywheel? (b) What is the magnitude of the torque?

34. A constant torque is applied to a rigid body whose moment of inertia is $4.0\,\text{kg-m}^2$ around the axis of rotation. If the wheel starts from rest and attains an angular velocity of 20.0 rad/s in 10.0 s, what is the applied torque?

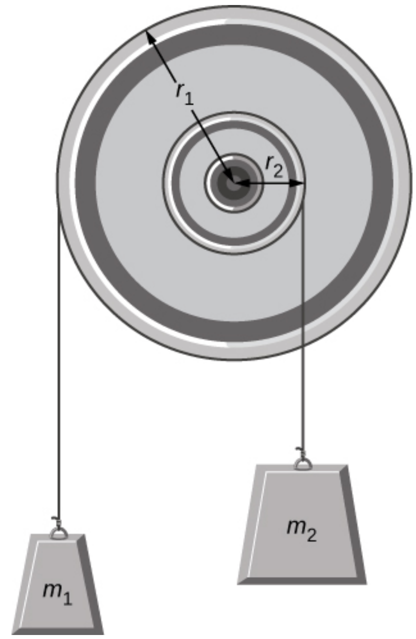
35. A torque of 50.0 N-m is applied to a grinding wheel ($I = 20.0\,\text{kg-m}^2$) for 20 s. (a) If it starts from rest, what is the angular velocity of the grinding wheel after the torque is removed? (b) Through what angle does the wheel move while the torque is applied?

36. A flywheel ($I = 100.0\,\text{kg-m}^2$) rotating at 500.0 rev/min is brought to rest by friction in 2.0 min. What is the frictional torque on the flywheel?

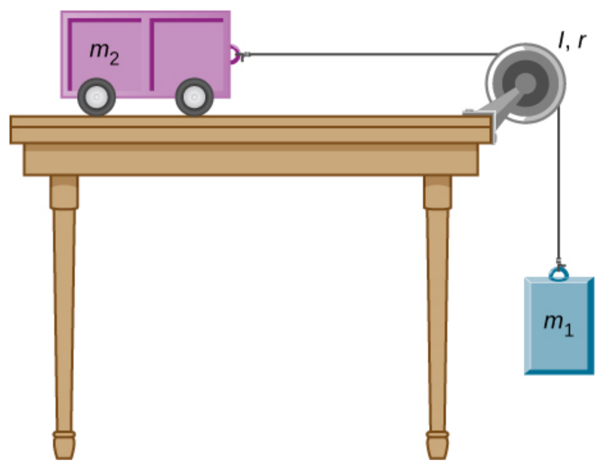
37. A uniform cylindrical grinding wheel of mass 50.0 kg and diameter 1.0 m is turned on by an electric motor. The friction in the bearings is negligible. (a) What torque must be applied to the wheel to bring it from rest to 120 rev/min in 20 revolutions? (b) A tool whose coefficient of kinetic friction with the wheel is 0.60 is pressed perpendicularly against the wheel with a force of 40.0 N. What torque must be supplied by the motor to keep the wheel rotating at a constant angular velocity?

38. Suppose when Earth was created, it was not rotating. However, after the application of a uniform torque after 6 days, it was rotating at 1 rev/day. (a) What was the angular acceleration during the 6 days? (b) What torque was applied to Earth during this period? (c) What force tangent to Earth at its equator would produce this torque?

39. A pulley of moment of inertia $2.0\,\text{kg-m}^2$ is mounted on a wall as shown in the following figure. Light strings are wrapped around two circumferences of the pulley and weights are attached. What are (a) the angular acceleration of the pulley and (b) the linear acceleration of the weights? Assume the following data: $r_1 = 50\,\text{cm}, \quad r_2 = 20\,\text{cm}, \quad m_1 = 1.0\,\text{kg}, \quad m_2 = 2.0\,\text{kg}$.



40. The cart shown below moves across the table top as the block falls. What is the acceleration of the cart? Neglect friction and assume the following data: $m_1 = 2.0\,\text{kg},\ m_2 = 4.0\,\text{kg},\ I = 0.4\,\text{kg-m}^2,\ r = 20\,\text{cm}$



10 | WORK AND ENERGY

Figure 10.1 A sprinter exerts her maximum power with the greatest force in the short time her foot is in contact with the ground. This adds to her kinetic energy, preventing her from slowing down during the race. Pushing back hard on the track generates a reaction force that propels the sprinter forward to win at the finish. (credit: modification of work by Marie-Lan Nguyen)

Chapter Outline

Introduction

In this chapter, we discuss some basic physical concepts involved in every physical motion in the universe, going beyond the concepts of force and change in motion, which we discussed in **Motion in Two and Three Dimensions** and **Newton's Synthesis**. These concepts are work, kinetic energy, and power. We explain how these quantities are related to one another, which will lead us to a fundamental relationship called the work-energy theorem. In the next chapter, we generalize this idea to the broader principle of conservation of energy.

The application of Newton's laws usually requires solving differential equations that relate the forces acting on an object to the accelerations they produce. Often, an analytic solution is intractable or impossible, requiring lengthy numerical solutions or simulations to get approximate results. In such situations, more general relations, like the work-energy theorem (or the conservation of energy), can still provide useful answers to many questions and require a more modest amount of mathematical calculation. In particular, you will see how the work-energy theorem is useful in relating the speeds of a particle, at different points along its trajectory, to the forces acting on it, even when the trajectory is otherwise too complicated to deal with. Thus, some aspects of motion can be addressed with fewer equations and without vector decompositions.

10.1 | Work

Learning Objectives

By the end of this section, you will be able to:

- Represent the work done by any force
- Evaluate the work done for various forces

In physics, **work** represents a type of energy. Work is done when a **force** acts on something that undergoes a **displacement** from one position to another. Forces can vary as a function of position, and displacements can be along various paths between two points. The mathematical expression for work involves the magnitude and direction of the force vector, the magnitude and direction of the displacement vector. The work done by a force is the scalar product (or "dot product") between the force vector and a vector representing the displacement of the object. In moving an object from location $\mathbf{r_1}$ to $\mathbf{r_2}$, the work done by the force **F** is:

$$W = \vec{\mathbf{F}} \cdot \vec{\mathbf{\Delta r}} = \vec{\mathbf{F}} \cdot \left(\vec{\mathbf{r}}_2 - \vec{\mathbf{r}}_1\right) = F\Delta \mathrm{r} \cos(\theta) \qquad (10.1)$$

where F is the magnitude of the force vector, $\Delta \mathrm{r}$ is the magnitude of the displacement vector, and θ is the angle between the two vectors.

The units of work are units of force multiplied by units of length, which in the SI system is newtons times meters, $\mathrm{N \cdot m}$. This combination is called a joule, for historical reasons that we will mention later, and is abbreviated as J. In the English system, still used in the United States, the unit of force is the pound (lb) and the unit of distance is the foot (ft), so the unit of work is the foot-pound $(\mathrm{ft \cdot lb})$.

Example

Figure 10.2(a) shows a person exerting a constant force $\vec{\mathbf{F}}$ along the handle of a lawn mower, which makes an angle θ with the horizontal. The horizontal displacement of the lawn mower, over which the force acts, is $\vec{\mathbf{d}}$. The work done on the lawn mower is $W = \vec{\mathbf{F}} \cdot \vec{\mathbf{d}} = Fd \cos\theta$, which the figure also illustrates as the horizontal component of the force times the magnitude of the displacement.

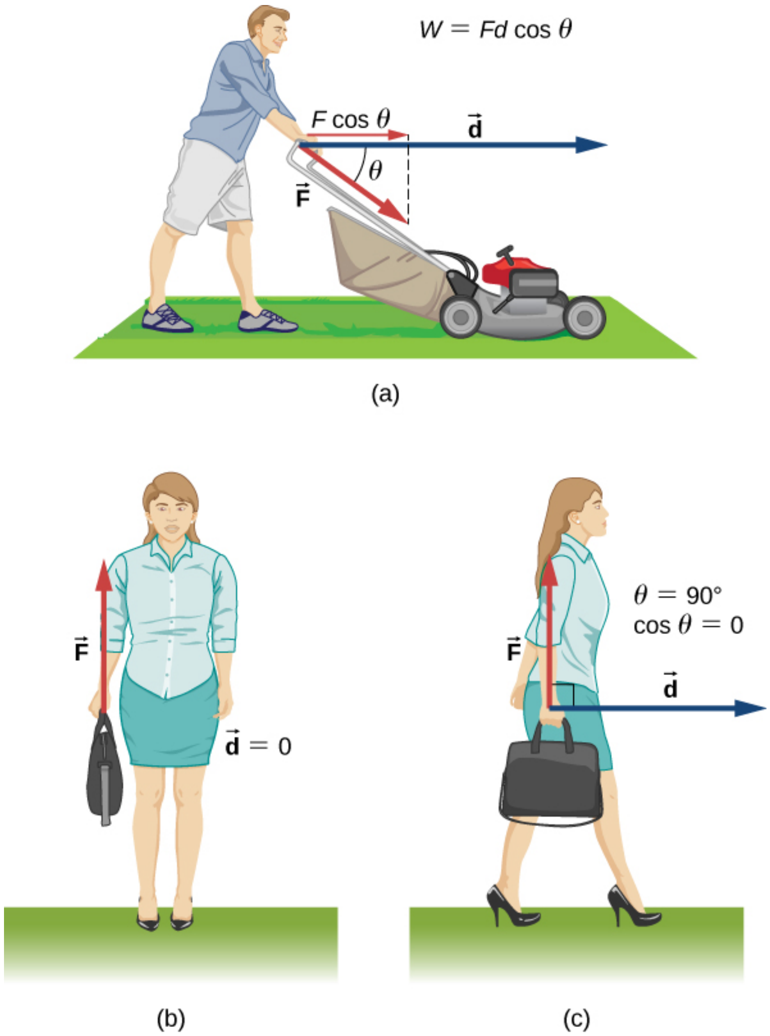


Figure 10.2 Work done by a constant force. (a) A person pushes a lawn mower with a constant force. The component of the force parallel to the displacement is the work done, as shown in the equation in the figure. (b) A person holds a briefcase. No work is done because the displacement is zero. (c) The person in (b) walks horizontally while holding the briefcase. No work is done because $\cos\theta$ is zero.

Figure 10.2(b) shows a person holding a briefcase. The person must exert an upward force, equal in magnitude to the weight of the briefcase, but this force does no work, because the displacement over which it acts is zero. So why do you eventually feel tired just holding the briefcase, if you're not doing any work on it? The answer is that muscle fibers in your arm are contracting and doing work inside your arm, even though the force your muscles exert externally on the briefcase doesn't do any work on it. (Part of the force you exert could also be tension in the bones and ligaments of your arm, but other muscles in your body would be doing work to maintain the position of your arm.)

In **Figure 10.2**(c), where the person in (b) is walking horizontally with constant speed, the work done by the person on the briefcase is still zero, but now because the angle between the force exerted and the displacement is $90°$ ($\vec{\mathbf{F}}$ perpendicular to $\vec{\mathbf{d}}$) and $\cos 90° = 0$.

Example 10.1

Calculating the Work You Do to Push a Lawn Mower

How much work is done on the lawn mower by the person in **Figure 10.2**(a) if he exerts a constant force of 75.0

N at an angle $35°$ below the horizontal and pushes the mower 25.0 m on level ground?

Strategy

We can solve this problem by substituting the given values into the definition of work done on an object by a constant force, stated in the equation $W = Fd\cos\theta$. The force, angle, and displacement are given, so that only the work W is unknown.

Solution

The equation for the work is

$$W = Fd\cos\theta.$$

Substituting the known values gives

$$W = (75.0\,\text{N})(25.0\,\text{m})\cos(35.0°) = 1.54 \times 10^3\ \text{J}.$$

Significance

Even though one and a half kilojoules may seem like a lot of work, it's only about as much work as you could do by burning one sixth of a gram of fat.

When you mow the grass, other forces act on the lawn mower besides the force you exert—namely, the contact force of the ground and the gravitational force of Earth. Let's consider the work done by these forces in general. For an object moving on a surface, the displacement $\Delta\vec{\mathbf{r}}$ is tangent to the surface. The part of the contact force on the object that is perpendicular to the surface is the normal force $\vec{\mathbf{N}}$. Since the cosine of the angle between the normal and the tangent to a surface is zero, we have

$$W_\text{N} = \vec{\mathbf{N}} \cdot \Delta\vec{\mathbf{r}} = 0$$

The normal force never does work under these circumstances. (Note that if the displacement $\Delta\vec{\mathbf{r}}$ did have a relative component perpendicular to the surface, the object would either leave the surface or break through it, and there would no longer be any normal contact force. However, if the object is more than a particle, and has an internal structure, the normal contact force can do work on it, for example, by displacing it or deforming its shape.)

The part of the contact force on the object that is parallel to the surface is friction, $\vec{\mathbf{f}}$. For this object sliding along the surface, kinetic friction $\vec{\mathbf{f}}_\text{k}$ is opposite to $\Delta\vec{\mathbf{r}}$, relative to the surface, so the work done by kinetic friction is negative. If the magnitude of $\vec{\mathbf{f}}_\text{k}$ is constant (as it would be if all the other forces on the object were constant), then the work done by friction is

$$W_\text{fr} = -f_k|l_{AB}|, \qquad (10.2)$$

where $|l_{AB}|$ is the path length on the surface. The minus sign comes from the fact that the force of friction and the displacement vectors are in opposite directions, hence $\cos(\theta) = -1$.(Note that, especially if the work done *by* a force is negative, people may refer to the work done *against* this force, where $W_\text{against} = -W_\text{by}$. The work done against a force may also be viewed as the work required to overcome this force, as in "How much work is required to overcome…?") The force of static friction, however, can do positive or negative work. When you walk, the force of static friction exerted by the ground on your back foot accelerates you for part of each step. If you're slowing down, the force of the ground on your front foot decelerates you. If you're driving your car at the speed limit on a straight, level stretch of highway, the negative work done by kinetic friction of air resistance is balanced by the positive work done by the static friction of the road on the drive wheels. You can pull the rug out from under an object in such a way that it slides backward relative to the rug, but forward relative to the floor. In this case, kinetic friction exerted by the rug on the object could be in the same direction as the displacement of the object, relative to the floor, and do positive work. The bottom line is that you need to analyze each particular case to determine the work done by the forces, whether positive, negative or zero.

Example 10.2

Moving a Couch

You decide to move your couch to a new position on your horizontal living room floor. The normal force on the couch is 1 kN and the coefficient of friction is 0.6. (a) You first push the couch 3 m parallel to a wall and then 1 m perpendicular to the wall (*A* to *B* in **Figure 10.3**). How much work is done by the frictional force? (b) You don't like the new position, so you move the couch straight back to its original position (*B* to *A* in **Figure 10.3**). What was the total work done against friction moving the couch away from its original position and back again?

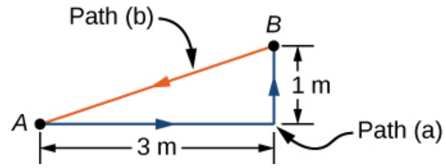


Figure 10.3 Top view of paths for moving a couch.

Strategy

The magnitude of the force of kinetic friction on the couch is constant, equal to the coefficient of friction times the normal force, $f_K = \mu_K N$. Therefore, the work done by it is $W_{\text{fr}} = -f_K d$, where *d* is the path length traversed. The segments of the paths are the sides of a right triangle, so the path lengths are easily calculated. In part (b), you can use the fact that the work done against a force is the negative of the work done by the force.

Solution

a. The work done by friction is

$$W = -(0.6)(1\text{ kN})(3\text{ m} + 1\text{ m}) = -2.4\text{ kJ}.$$

b. The length of the path along the hypotenuse is $\sqrt{10}\text{ m}$, so the total work done against friction is

$$W = (0.6)(1\text{ kN})(3\text{ m} + 1\text{ m} + \sqrt{10}\text{ m}) = 4.3\text{ kJ}.$$

Significance

The total path over which the work of friction was evaluated began and ended at the same point (it was a closed path), so that the total displacement of the couch was zero. However, the total work was not zero. The reason is that forces like friction are classified as nonconservative forces, or dissipative forces, as we discuss in the next chapter.

10.1 Check Your Understanding Can kinetic friction ever be a constant force for all paths?

The other force on the lawn mower mentioned above was Earth's gravitational force, or the weight of the mower. Near the surface of Earth, the gravitational force on an object of mass *m* has a constant magnitude, *mg*, and constant direction, vertically down. Therefore, the work done by gravity on an object is the dot product of its weight and its displacement. In many cases, it is convenient to express the dot product for gravitational work in terms of the *x*-, *y*-, and *z*-components of the vectors. A typical coordinate system has the *x*-axis horizontal and the *y*-axis vertically up. Then the gravitational force is $-mg\,\hat{\mathbf{j}}$, so the work done by gravity, over any path from *A* to *B*, is

$$W_{\text{grav},\,AB} = -mg\,\hat{\mathbf{j}} \cdot (\vec{\mathbf{r}}_B - \vec{\mathbf{r}}_A) = -mg(y_B - y_A). \tag{10.3}$$

The work done by a constant force of gravity on an object depends only on the object's weight and the difference in height through which the object is displaced. Gravity does negative work on an object that moves upward ($y_B > y_A$), or, in other words, you must do positive work against gravity to lift an object upward. Alternately, gravity does positive work on an object that moves downward ($y_B < y_A$), or you do negative work against gravity to "lift" an object downward, controlling its descent so it doesn't drop to the ground. ("Lift" is used as opposed to "drop".)

Example 10.3

Shelving a Book

You lift an oversized library book, weighing 20 N, 1 m vertically down from a shelf, and carry it 3 m horizontally to a table (**Figure 10.4**). How much work does gravity do on the book? (b) When you're finished, you move the book in a straight line back to its original place on the shelf. What was the total work done against gravity, moving the book away from its original position on the shelf and back again?

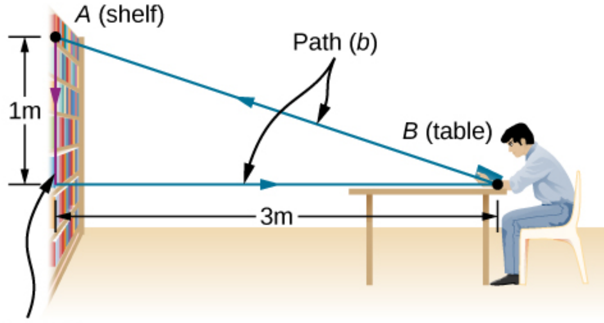


Figure 10.4 Side view of the paths for moving a book to and from a shelf.

Strategy

We have just seen that the work done by a constant force of gravity depends only on the weight of the object moved and the difference in height for the path taken, $W_{AB} = -mg(y_B - y_A)$. We can evaluate the difference in height to answer (a) and (b).

Solution

a. Since the book starts on the shelf and is lifted down $y_B - y_A = -1\text{ m}$, we have

$$W = -(20\,\text{N})(-1\text{ m}) = 20\,\text{J}.$$

b. There is zero difference in height for any path that begins and ends at the same place on the shelf, so $W = 0$.

Significance

Gravity does positive work (20 J) when the book moves down from the shelf. The gravitational force between two objects is an attractive force, which does positive work when the objects get closer together. Gravity does zero work (0 J) when the book moves horizontally from the shelf to the table and negative work (−20 J) when the book moves from the table back to the shelf. The total work done by gravity is zero $[20\,\text{J} + 0\,\text{J} + (-20\,\text{J}) = 0]$.

Unlike friction or other dissipative forces, described in **Example 10.2**, the total work done against gravity, over any closed path, is zero. Positive work is done against gravity on the upward parts of a closed path, but an equal amount of negative work is done against gravity on the downward parts. In other words, work done *against* gravity, lifting an object *up*, is "given back" when the object comes back down. Forces like gravity (those that do zero work over any closed path) are classified as conservative forces and play an important role in physics.

10.2 Check Your Understanding Can Earth's gravity ever be a constant force for all paths?

10.2 | Kinetic Energy

Learning Objectives

By the end of this section, you will be able to:

- Calculate the kinetic energy of a particle given its mass and its velocity or momentum
- Evaluate the kinetic energy of a body, relative to different frames of reference
- Describe the differences between rotational and translational kinetic energy

It's plausible to suppose that the greater the velocity of a body, the greater effect it could have on other bodies. This does not depend on the direction of the velocity, only its magnitude. At the end of the seventeenth century, a quantity was introduced into mechanics to explain collisions between two perfectly elastic bodies, in which one body makes a head-on collision with an identical body at rest. The first body stops, and the second body moves off with the initial velocity of the first body. (If you have ever played billiards or croquet, or seen a model of Newton's Cradle, you have observed this type of collision.) The idea behind this quantity was related to the forces acting on a body and was referred to as "the energy of motion." Later on, during the eighteenth century, the name **kinetic energy** was given to energy of motion.

With this history in mind, we can now state the classical definition of kinetic energy. Note that when we say "classical," we mean non-relativistic, that is, at speeds much less that the speed of light. At speeds comparable to the speed of light, the special theory of relativity requires a different expression for the kinetic energy of a particle.

Since objects (or systems) of interest vary in complexity, we first define the kinetic energy of a particle with mass *m*.

Kinetic Energy

The kinetic energy of a particle is one-half the product of the particle's mass *m* and the square of its speed *v*:

$$K = \frac{1}{2}mv^2. \tag{10.4}$$

We then extend this definition to any system of particles by adding up the kinetic energies of all the constituent particles:

$$K = \sum \frac{1}{2}mv^2. \tag{10.5}$$

The units of kinetic energy are mass times the square of speed, or $\text{kg} \cdot \text{m}^2/\text{s}^2$. But the units of force are mass times acceleration, $\text{kg} \cdot \text{m/s}^2$, so the units of kinetic energy are also the units of force times distance, which are the units of work, or joules. You will see in the next section that work and kinetic energy have the same units, because they are different forms of the same, more general, physical property.

Example 10.4

Kinetic Energy of an Object

(a) What is the kinetic energy of an 80-kg athlete, running at 10 m/s? (b) The Chicxulub crater in Yucatan, one of the largest existing impact craters on Earth, is thought to have been created by an asteroid, traveling at 22 km/s and releasing 4.2×10^{23} J of kinetic energy upon impact. What was its mass? (c) In nuclear reactors, thermal neutrons, traveling at about 2.2 km/s, play an important role. What is the kinetic energy of such a particle?

Strategy

To answer these questions, you can use the definition of kinetic energy in Equation 10.4. You also have to look up the mass of a neutron.

Solution

Don't forget to convert km into m to do these calculations, although, to save space, we omitted showing these conversions.

a. $K = \frac{1}{2}(80\,\text{kg})(10\,\text{m/s})^2 = 4.0\,\text{kJ}.$

b. $m = 2K/v^2 = 2(4.2 \times 10^{23}\,\text{J})/(22\,\text{km/s})^2 = 1.7 \times 10^{15}\,\text{kg}.$

c. $K = \frac{1}{2}(1.68 \times 10^{-27}\,\text{kg})(2.2\,\text{km/s})^2 = 4.1 \times 10^{-21}\,\text{J}.$

Significance

In this example, we used the way mass and speed are related to kinetic energy, and we encountered a very wide range of values for the kinetic energies. Different units are commonly used for such very large and very small values. The energy of the impactor in part (b) can be compared to the explosive yield of TNT and nuclear explosions, $1\,\text{megaton} = 4.18 \times 10^{15}\,\text{J}$. The Chicxulub asteroid's kinetic energy was about a hundred million megatons. At the other extreme, the energy of subatomic particle is expressed in electron-volts, $1\,\text{eV} = 1.6 \times 10^{-19}\,\text{J}$. The thermal neutron in part (c) has a kinetic energy of about one fortieth of an electron-volt.

10.3 Check Your Understanding (a) A car and a truck are each moving with the same kinetic energy. Assume that the truck has more mass than the car. Which has the greater speed? (b) A car and a truck are each moving with the same speed. Which has the greater kinetic energy?

The kinetic energy of a particle is a single quantity, but the kinetic energy of a system of particles can sometimes be divided into various types, depending on the system and its motion. For example, if all the particles in a system have the same velocity, the system is undergoing translational motion and has translational kinetic energy. If an object is rotating, it could have rotational kinetic energy, or if it's vibrating, it could have vibrational kinetic energy. The kinetic energy of a system, relative to an internal frame of reference, may be called internal kinetic energy. The kinetic energy associated with random molecular motion may be called thermal energy. These names will be used in later chapters of the book, when appropriate. Regardless of the name, every kind of kinetic energy is the same physical quantity, representing energy associated with motion.

Example 10.5

Special Names for Kinetic Energy

(a) A player lobs a mid-court pass with a 624-g basketball, which covers 15 m in 2 s. What is the basketball's horizontal translational kinetic energy while in flight? (b) An average molecule of air, in the basketball in part (a), has a mass of 29 u, and an average speed of 500 m/s, relative to the basketball. There are about 3×10^{23} molecules inside it, moving in random directions, when the ball is properly inflated. What is the average translational kinetic energy of the random motion of all the molecules inside, relative to the basketball? (c) How fast would the basketball have to travel relative to the court, as in part (a), so as to have a kinetic energy equal to the amount in part (b)?

Strategy

In part (a), first find the horizontal speed of the basketball and then use the definition of kinetic energy in terms of mass and speed, $K = \frac{1}{2}mv^2$. Then in part (b), convert unified units to kilograms and then use $K = \frac{1}{2}mv^2$ to get the average translational kinetic energy of one molecule, relative to the basketball. Then multiply by the number of molecules to get the total result. Finally, in part (c), we can substitute the amount of kinetic energy in part (b), and the mass of the basketball in part (a), into the definition $K = \frac{1}{2}mv^2$, and solve for v.

Solution

a. The horizontal speed is (15 m)/(2 s), so the horizontal kinetic energy of the basketball is

$$\frac{1}{2}(0.624\,\text{kg})(7.5\,\text{m/s})^2 = 17.6\,\text{J}.$$

b. The average translational kinetic energy of a molecule is

$$\frac{1}{2}(29\text{ u})(1.66 \times 10^{-27}\text{ kg/u})(500\text{ m/s})^2 = 6.02 \times 10^{-21}\text{ J},$$

and the total kinetic energy of all the molecules is

$$(3 \times 10^{23})(6.02 \times 10^{-21}\text{ J}) = 1.80\text{ kJ}.$$

c. $v = \sqrt{2(1.8\text{ kJ})/(0.624\text{ kg})} = 76.0\text{ m/s}.$

Significance

In part (a), this kind of kinetic energy can be called the horizontal kinetic energy of an object (the basketball), relative to its surroundings (the court). If the basketball were spinning, all parts of it would have not just the average speed, but it would also have rotational kinetic energy. Part (b) reminds us that this kind of kinetic energy can be called internal or thermal kinetic energy. Notice that this energy is about a hundred times the energy in part (a). How to make use of thermal energy will be the subject of the chapters on thermodynamics. In part (c), since the energy in part (b) is about 100 times that in part (a), the speed should be about 10 times as big, which it is (76 compared to 7.5 m/s).

Rotational Kinetic Energy

Any moving object has kinetic energy. We know how to calculate this for a body undergoing translational motion, but how about for a rigid body undergoing rotation? This might seem complicated because each point on the rigid body has a different velocity. However, we can make use of angular velocity—which is the same for the entire rigid body—to express the kinetic energy for a rotating object. **Figure 10.5** shows an example of a very energetic rotating body: an electric grindstone propelled by a motor. Sparks are flying, and noise and vibration are generated as the grindstone does its work. This system has considerable energy, some of it in the form of heat, light, sound, and vibration. However, most of this energy is in the form of **rotational kinetic energy**.

Figure 10.5 The rotational kinetic energy of the grindstone is converted to heat, light, sound, and vibration. (credit: Zachary David Bell, US Navy)

Energy in rotational motion is not a new form of energy; rather, it is the energy associated with rotational motion, the same as kinetic energy in translational motion. However, because kinetic energy is given by $K = \frac{1}{2}mv^2$, and velocity is a quantity that is different for every point on a rotating body about an axis, it makes sense to find a way to write kinetic energy in terms of the variable ω, which is the same for all points on a rigid rotating body. For a single particle rotating around a fixed axis, this is straightforward to calculate. We can relate the angular velocity to the magnitude of the translational velocity using the relation $v_{\text{t}} = \omega r$, where *r* is the distance of the particle from the axis of rotation and v_{t} is its tangential speed. Substituting into the equation for kinetic energy, we find

$$K = \frac{1}{2}mv_{\text{t}}^2 = \frac{1}{2}m(\omega r)^2 = \frac{1}{2}(mr^2)\omega^2.$$

In the case of a rigid rotating body, we can divide up any body into a large number of smaller masses, each with a mass m_j and distance to the axis of rotation r_j, such that the total mass of the body is equal to the sum of the individual masses: $M = \sum_j m_j$. Each smaller mass has tangential speed v_j, where we have dropped the subscript *t* for the moment. The total kinetic energy of the rigid rotating body is

$$K = \sum_j \frac{1}{2}m_j v_j^2 = \sum_j \frac{1}{2}m_j (r_j \omega_j)^2$$

and since $\omega_j = \omega$ for all masses,

$$K = \frac{1}{2}\left(\sum_j m_j r_j^2\right)\omega^2. \quad (10.6)$$

where we recognize the expression for the moment of inertia in parenthesis. The units of **Equation 10.6** are joules (J). Written in terms of the moment of inertia:

$$K = \frac{1}{2}I\omega^2. \quad (10.7)$$

We see from this equation that the kinetic energy of a rotating rigid body is directly proportional to the moment of inertia and the square of the angular velocity. This is exploited in flywheel energy-storage devices, which are designed to store large amounts of rotational kinetic energy. Many carmakers are now testing flywheel energy storage devices in their automobiles, such as the flywheel, or kinetic energy recovery system, shown in **Figure 10.6**.

Figure 10.6 A KERS (kinetic energy recovery system) flywheel used in cars. (credit: “cmonville”/Flickr)

The rotational and translational quantities for kinetic energy and inertia are summarized in **Table 10.1**.

Rotational	Translational
$I = \sum_j m_j r_j^2$	m
$K = \frac{1}{2}I\omega^2$	$K = \frac{1}{2}mv^2$

Table 10.1 Rotational and Translational Kinetic Energies and Inertia

10.3 | Work-Energy Theorem

Learning Objectives

By the end of this section, you will be able to:

- Apply the work-energy theorem to find information about the motion of a particle, given the forces acting on it
- Use the work-energy theorem to find information about the forces acting on a particle, given information about its motion

We have discussed how to find the work done on a particle by the forces that act on it, but how is that work manifested in the motion of the particle? According to Newton’s second law of motion, the sum of all the forces acting on a particle, or the net force, determines the acceleration of the particle, or the change in its motion. Therefore, we should consider the work done by all the forces acting on a particle, or the **net work**, to see what effect it has on the particle’s motion.

We already know that work and energy are measured in the same units (Joules). Work is done when a force acts on an object and the object experiences a displacement along the direction of the force. It seems obvious that, in such a case, the object's speed along the direction of this motion will also change. Therefore, its kinetic energy will also change. We quantify this connection between the work done on an object and the change in its kinetic energy as follows:

Work-Energy Theorem

The net work done on a particle equals the change in the particle's kinetic energy:

$$W_{\text{net}} = K_B - K_A. \tag{10.8}$$

Figure 10.7 Horse pulls are common events at state fairs. The work done by the horses pulling on the load results in a change in kinetic energy of the load, ultimately going faster. (credit: modification of work by "Jassen"/ Flickr)

According to this theorem, when an object slows down, its final kinetic energy is less than its initial kinetic energy, the change in its kinetic energy is negative, and so is the net work done on it. If an object speeds up, the net work done on it is positive. When calculating the net work, you must include all the forces that act on an object. If you leave out any forces that act on an object, or if you include any forces that don't act on it, you will get a wrong result.

One importance of the work-energy theorem, and the further generalizations to which it leads, is that it makes some types of calculations much simpler to accomplish than they would be by trying to solve Newton's second law. Also, in situations where we either do not know the explicit values of the individual forces involved, or where the values of those forces are not constant in time, the work-energy approach can allow us to find information about changes in velocity without ever calculating the value of an acceleration.

Problem-Solving Strategy: Work-Energy Theorem

1. Draw a free-body diagram for each force on the object.
2. Determine whether or not each force does work over the displacement in the diagram. Be sure to keep any positive or negative signs in the work done.
3. Add up the total amount of work done by each force.
4. Set this total work equal to the change in kinetic energy and solve for any unknown parameter.
5. Check your answers. If the object is traveling at a constant speed or zero acceleration, the total work done should be zero and match the change in kinetic energy. If the total work is positive, the object must have sped up or increased kinetic energy. If the total work is negative, the object must have slowed down or decreased kinetic energy.

Example 10.6

Loop-the-Loop

The frictionless track for a toy car includes a loop-the-loop of radius *R*. How high, measured from the bottom of the loop, must the car be placed to start from rest on the approaching section of track and go all the way around the loop?

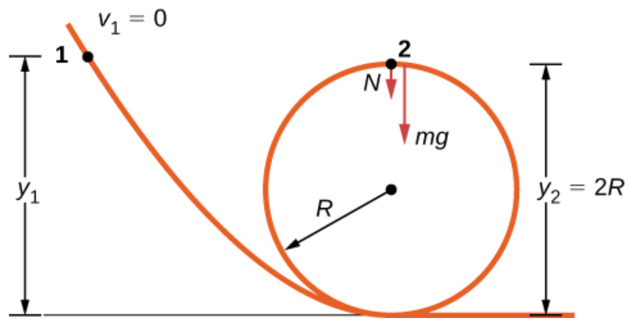


Figure 10.8 A frictionless track for a toy car has a loop-the-loop in it. How high must the car start so that it can go around the loop without falling off?

Strategy

The free-body diagram at the final position of the object is drawn in **Figure 10.8**. The gravitational work is the only work done over the displacement that is not zero. Since the weight points in the same direction as the net vertical displacement, the total work done by the gravitational force is positive. From the work-energy theorem, the starting height determines the speed of the car at the top of the loop,

$$-mg(y_2 - y_1) = \frac{1}{2}mv_2{}^2,$$

where the notation is shown in the accompanying figure. At the top of the loop, the normal force and gravity are both down and the acceleration is centripetal, so

$$a_{\text{top}} = \frac{F}{m} = \frac{N + mg}{m} = \frac{v_2^2}{R}.$$

The condition for maintaining contact with the track is that there must be some normal force, however slight; that is, $N > 0$. Substituting for v_2^2 and *N*, we can find the condition for y_1.

Solution

Implement the steps in the strategy to arrive at the desired result:

$$N = \frac{-mg + mv_2^2}{R} = \frac{-mg + 2mg(y_1 - 2R)}{R} > 0 \quad \text{or} \quad y_1 > \frac{5R}{2}.$$

Significance

On the surface of the loop, the normal component of gravity and the normal contact force must provide the centripetal acceleration of the car going around the loop. The tangential component of gravity slows down or speeds up the car. A child would find out how high to start the car by trial and error, but now that you know the work-energy theorem, you can predict the minimum height (as well as other more useful results) from physical principles. By using the work-energy theorem, you did not have to solve a differential equation to determine the height.

10.4 Check Your Understanding Suppose the radius of the loop-the-loop in **Example 10.6** is 15 cm and the toy car starts from rest at a height of 45 cm above the bottom. What is its speed at the top of the loop?

Visit Carleton College's site to see a **video (https://openstax.org/l/21carcollvidrol)** of a looping rollercoaster.

In situations where the motion of an object is known, but the values of one or more of the forces acting on it are not known, you may be able to use the work-energy theorem to get some information about the forces. Work depends on the force and the distance over which it acts, so the information is provided via their product.

Example 10.7

Determining a Stopping Force

A bullet from a 0.22LR-caliber cartridge has a mass of 40 grains (2.60 g) and a muzzle velocity of 1100 ft./s (335 m/s). It can penetrate eight 1-inch pine boards, each with thickness 0.75 inches. What is the average stopping force exerted by the wood, as shown in **Figure 10.9**?

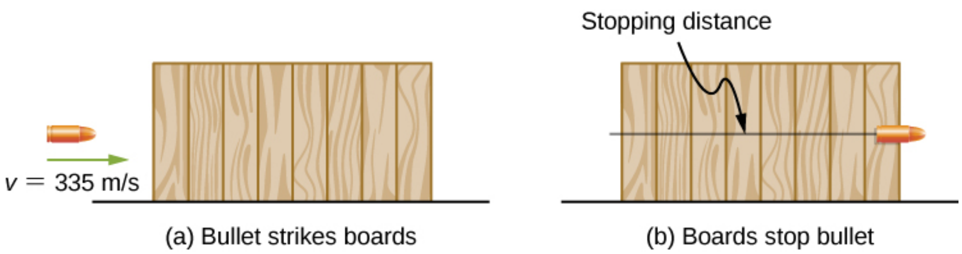


Figure 10.9 The boards exert a force to stop the bullet. As a result, the boards do work and the bullet loses kinetic energy.

Strategy

We can assume that under the general conditions stated, the bullet loses all its kinetic energy penetrating the boards, so the work-energy theorem says its initial kinetic energy is equal to the average stopping force times the distance penetrated. The change in the bullet's kinetic energy and the net work done stopping it are both negative, so when you write out the work-energy theorem, with the net work equal to the average force times the stopping distance, that's what you get. The total thickness of eight 1-inch pine boards that the bullet penetrates is $8 \times \frac{3}{4}\,\text{in.} = 6\,\text{in.} = 15.2\,\text{cm}.$

Solution

Applying the work-energy theorem, we get

$$W_{\text{net}} = -F_{\text{ave}}\,\Delta s_{\text{stop}} = -K_{\text{initial}},$$

so

$$F_{\text{ave}} = \frac{\frac{1}{2}mv^2}{\Delta s_{\text{stop}}} = \frac{\frac{1}{2}(2.6 \times 10^{-3}\,\text{kg})(335\,\text{m/s})^2}{0.152\,\text{m}} = 960\,\text{N}.$$

Significance

We could have used Newton's second law and kinematics in this example, but the work-energy theorem also supplies an answer to less simple situations. The penetration of a bullet, fired vertically upward into a block of wood, is discussed in one section of Asif Shakur's recent article ["Bullet-Block Science Video Puzzle." *The Physics Teacher* (January 2015) 53(1): 15-16]. If the bullet is fired dead center into the block, it loses all its kinetic energy and penetrates slightly farther than if fired off-center. The reason is that if the bullet hits off-center, it has a little kinetic energy after it stops penetrating, because the block rotates. The work-energy theorem implies that a smaller change in kinetic energy results in a smaller penetration. You will understand more of the physics in this interesting article after you finish reading **Angular Momentum**.

Learn more about work and energy in this **PhET simulation (https://openstax.org/l/21PhETSimRamp)** called "the ramp." Try changing the force pushing the box and the frictional force along the incline. The work and energy plots can be examined to note the total work done and change in kinetic energy of the box.

10.4 | Power

Learning Objectives

By the end of this section, you will be able to:

- Relate the work done during a time interval to the power delivered
- Find the power expended by a force acting on a moving body

The concept of work involves force and displacement; the work-energy theorem relates the net work done on a body to the difference in its kinetic energy, calculated between two points on its trajectory. None of these quantities or relations involves time explicitly, yet we know that the time available to accomplish a particular amount of work is frequently just as important to us as the amount itself. In the chapter-opening figure, several sprinters may have achieved the same velocity at the finish, and therefore did the same amount of work, but the winner of the race did it in the least amount of time.

We express the relation between work done and the time interval involved in doing it, by introducing the concept of power. Since work can vary as a function of time, we first define **average power** as the work done during a time interval, divided by the interval,

$$P_{\text{ave}} = \frac{\Delta W}{\Delta t}. \tag{10.9}$$

Then, we can define the **instantaneous power** (frequently referred to as just plain **power**).

Power

Power is defined as the rate of doing work, or the limit of the average power for time intervals approaching zero,. Using the concept of the derivative from calculus:

$$P = \frac{dW}{dt}. \tag{10.10}$$

If the power is constant over a time interval, the average power for that interval equals the instantaneous power, and the work done by the agent supplying the power is:

$$W = P\Delta t$$

The work-energy theorem relates how work can be transformed into kinetic energy. Since there are other forms of energy as well, as we discuss in the next chapter, we can also define power as the rate of transfer of energy. Work and energy are measured in units of joules, so power is measured in units of joules per second, which has been given the SI name watts, abbreviation W: $1 \text{ J/s} = 1 \text{ W}$. Another common unit for expressing the power capability of everyday devices is horsepower: $1 \text{ hp} = 746 \text{ W}$.

Example 10.8

Pull-Up Power

An 80-kg army trainee does 10 pull-ups in 10 s (**Figure 10.10**). How much average power do the trainee's muscles supply moving his body? (*Hint:* Make reasonable estimates for any quantities needed.)

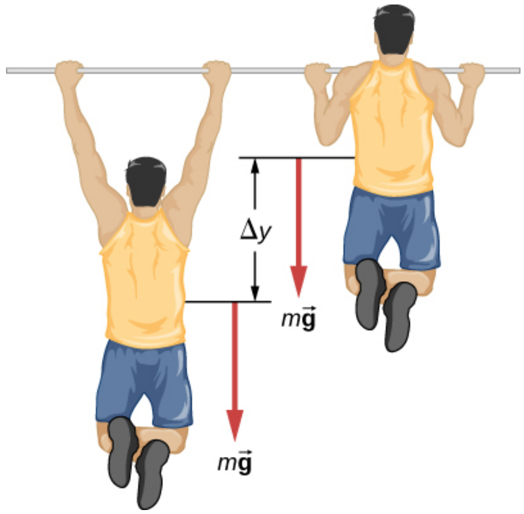


Figure 10.10 What is the power expended in doing ten pull-ups in ten seconds?

Strategy

The work done against gravity, going up or down a distance Δy, is $mg\Delta y.$ (If you lift and lower yourself at constant speed, the force you exert cancels gravity over the whole pull-up cycle.) Thus, the work done by the trainee's muscles (moving, but not accelerating, his body) for a complete repetition (up and down) is $2mg\Delta y.$ Let's assume that $\Delta y = 2\text{ft} \approx 60\,\text{cm}.$ Also, assume that the arms comprise 10% of the body mass and are not included in the moving mass. With these assumptions, we can calculate the work done for 10 pull-ups and divide by 10 s to get the average power.

Solution

The result we get, applying our assumptions, is

$$P_{\text{ave}} = \frac{10 \times 2(0.9 \times 80\,\text{kg})(9.8\,\text{m/s}^2)(0.6\,\text{m})}{10\,\text{s}} = 850\,\text{W}.$$

Significance

This is typical for power expenditure in strenuous exercise; in everyday units, it's somewhat more than one horsepower $(1\,\text{hp} = 746\,\text{W})$.

10.5 Check Your Understanding Estimate the power expended by a weightlifter raising a 150-kg barbell 2 m in 3 s.

The power involved in moving a body can also be expressed in terms of the forces acting on it. If a force $\mathbf{F}$ acts on a body that is displaced $\Delta\mathbf{r}$ in a time Δt, the power expended by the force is

$$P = \frac{\Delta W}{\Delta t} = \frac{\mathbf{F}\Delta\mathbf{r}\cos\theta}{\Delta t} = Fv\cos\theta \tag{10.11}$$

where v is the velocity of the body and θ is the angle between the force and the velocity.

Example 10.9

Automotive Power Driving Uphill

How much power must an automobile engine expend to move a 1200-kg car up a 15% grade at 90 km/h (**Figure 10.11**)? Assume that 25% of this power is dissipated overcoming air resistance and friction.

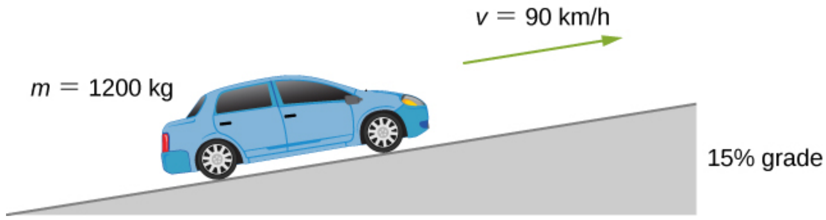


Figure 10.11 We want to calculate the power needed to move a car up a hill at constant speed.

Strategy

At constant velocity, there is no change in kinetic energy, so the net work done to move the car is zero. Therefore the power supplied by the engine to move the car equals the power expended against gravity and air resistance. By assumption, 75% of the power is supplied against gravity, which equals $mgv\cos(90^{\circ} - \theta) = mgv\sin\theta,$ where θ is the angle of the incline. A 15% grade means $\tan\theta = 0.15.$ This reasoning allows us to solve for the power required.

Solution

Carrying out the suggested steps, we find

$$0.75\,P = mgv\sin(\tan^{-1} 0.15),$$

or

$$P = \frac{(1200 \times 9.8\text{ N})(90\text{ m}/3.6\text{ s})\sin(8.53^{\circ})}{0.75} = 58\text{ kW},$$

or about 78 hp. (You should supply the steps used to convert units.)

Significance

This is a reasonable amount of power for the engine of a small to mid-size car to supply $(1\text{ hp} = 0.746\text{ kW}).$ Note that this is only the power expended to move the car. Much of the engine's power goes elsewhere, for example, into waste heat. That's why cars need radiators. Any remaining power could be used for acceleration, or to operate the car's accessories.

10.5 | Potential Energy

Learning Objectives

By the end of this section, you will be able to:

- Relate the difference of potential energy to work done on a particle for a system without friction or air drag
- Explain the meaning of the zero of the potential energy function for a system
- Calculate and apply the gravitational potential energy for an object near Earth's surface and the elastic potential energy of a mass-spring system
- Determine changes in gravitational potential energy over great distances

In **Work**, we saw that the work done on an object by the constant gravitational force, near the surface of Earth, over any displacement is a function only of the difference in the positions of the end-points of the displacement. This property allows us to define a different kind of energy for the system than its kinetic energy, which is called **potential energy**. We consider various properties and types of potential energy in the following subsections.

Potential Energy Basics

In **Projectile Motion**, we analyzed the motion of a projectile, like kicking a football in **Figure 10.12**. For this example, let's ignore friction and air resistance. As the football rises, the work done by the gravitational force on the football is negative, because the ball's displacement is positive vertically and the force due to gravity is negative vertically. We also noted that the ball slowed down until it reached its highest point in the motion, thereby decreasing the ball's kinetic energy. This loss in kinetic energy translates to a gain in gravitational potential energy of the football-Earth system.

As the football falls toward Earth, the work done on the football is now positive, because the displacement and the gravitational force both point vertically downward. The ball also speeds up, which indicates an increase in kinetic energy. Therefore, energy is converted from gravitational potential energy back into kinetic energy.

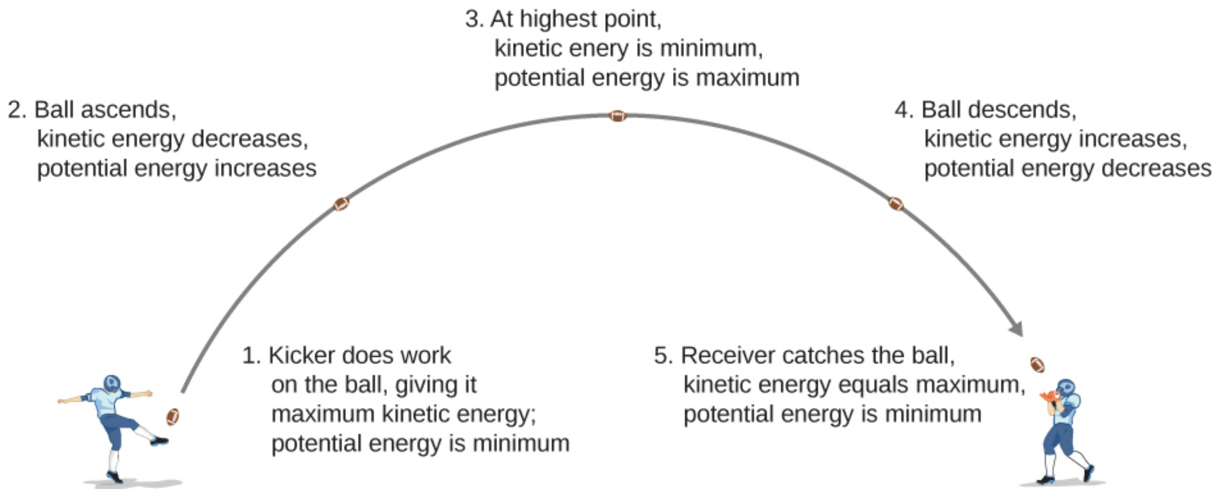


Figure 10.12 As a football starts its descent toward the wide receiver, gravitational potential energy is converted back into kinetic energy.

Based on this scenario, we can define the difference of potential energy from point A to point B as the negative of the work done:

$$\Delta U_{AB} = U_B - U_A = -W_{AB}. \tag{10.12}$$

This formula explicitly states a **potential energy difference**, not just an absolute potential energy. Therefore, we need to define potential energy at a given position in such a way as to state standard values of potential energy on their own, rather than potential energy differences. We do this by rewriting the potential energy function in terms of an arbitrary constant,

$$\Delta U = U(\vec{\mathbf{r}}) - U(\vec{\mathbf{r}}_0). \tag{10.13}$$

The choice of the potential energy at a starting location of $\vec{\mathbf{r}}_0$ is made out of convenience in the given problem. Most importantly, whatever choice is made should be stated and kept consistent throughout the given problem. There are some well-accepted choices of initial potential energy. For example, the lowest height in a problem is usually defined as zero potential energy, or if an object is in space, the farthest point away from the system is often defined as zero potential energy. Then, the potential energy, with respect to zero at $\vec{\mathbf{r}}_0$, is just $U(\vec{r})$.

As long as there is no friction or air resistance, the amount of change in kinetic energy of the football equals the amount of change in gravitational potential energy of the football. This can be generalized to any potential energy:

$$\Delta K_{AB} = -\Delta U_{AB}. \tag{10.14}$$

The minus sign is important! If the change in kinetic energy is positive (i.e. the system gains kinetic energy), then the change in potential energy must be negative (it loses potential energy). And, vice-versa.

Systems of Several Particles

In general, a system of interest could consist of several particles. The difference in the potential energy of the system is the negative of the work done by gravitational or elastic forces, which, as we will see in the next section, are conservative forces. The potential energy difference depends only on the initial and final positions of the particles, and on some parameters that characterize the interaction (like mass for gravity or the spring constant for a Hooke's law force).

It is important to remember that potential energy is a property of the interactions between objects in a chosen system, and not just a property of each object. This is especially true for electric forces, although in the examples of potential energy we consider below, parts of the system are either so big (like Earth, compared to an object on its surface) or so small (like a massless spring), that the changes those parts undergo are negligible if included in the system.

Types of Potential Energy

For each type of interaction present in a system, you can label a corresponding type of potential energy. The total potential energy of the system is the sum of the potential energies of all the types. (This follows from the additive property of the dot product in the expression for the work done.) Let's look at some specific examples of types of potential energy discussed in **Work**. First, we consider each of these forces when acting separately, and then when both act together.

Gravitational potential energy near Earth's surface

The system of interest consists of our planet, Earth, and one or more particles near its surface (or bodies small enough to be considered as particles, compared to Earth). The gravitational force on each particle (or body) is just its weight *mg* near the surface of Earth, acting vertically down. According to Newton's third law, each particle exerts a force on Earth of equal magnitude but in the opposite direction. Newton's second law tells us that the magnitude of the acceleration produced by each of these forces on Earth is *mg* divided by Earth's mass. Since the ratio of the mass of any ordinary object to the mass of Earth is vanishingly small, the motion of Earth can be completely neglected. Therefore, we consider this system to be a group of single-particle systems, subject to the uniform gravitational force of Earth.

In **Work**, the work done on a body by Earth's uniform gravitational force, near its surface, depended on the mass of the body, the acceleration due to gravity, and the difference in height the body traversed, as given by **Equation 10.3**. By definition, this work is the negative of the difference in the gravitational potential energy, so that difference is

$$\Delta U_{\text{grav}} = -W_{\text{grav},\ AB} = mg(y_B - y_A). \tag{10.15}$$

You can see from this that the gravitational potential energy function, near Earth's surface, is

$$U(y) = mgy + \text{const.} \tag{10.16}$$

You can choose the value of the constant, as described in the discussion of **Equation 10.13**; however, for solving most problems, the most convenient constant to choose is zero for when $y = 0$, which is the lowest vertical position in the problem.

Example 10.10

Gravitational Potential Energy of a Hiker

The summit of Great Blue Hill in Milton, MA, is 147 m above its base and has an elevation above sea level of 195 m (**Figure 10.13**). (Its Native American name, *Massachusett*, was adopted by settlers for naming the Bay Colony and state near its location.) A 75-kg hiker ascends from the base to the summit. What is the gravitational potential energy of the hiker-Earth system with respect to zero gravitational potential energy at base height, when the hiker is (a) at the base of the hill, (b) at the summit, and (c) at sea level, afterward?

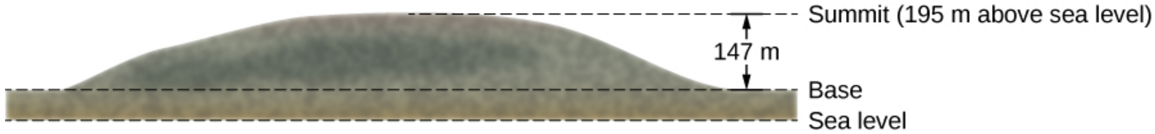


Figure 10.13 Sketch of the profile of Great Blue Hill, Milton, MA. The altitudes of the three levels are indicated.

Strategy

First, we need to pick an origin for the *y*-axis and then determine the value of the constant that makes the potential energy zero at the height of the base. Then, we can determine the potential energies from **Equation 10.16**, based on the relationship between the zero potential energy height and the height at which the hiker is located.

Solution

a. Let's choose the origin for the *y*-axis at base height, where we also want the zero of potential energy to be. This choice makes the constant equal to zero and

$$U(\text{base}) = U(0) = 0.$$

b. At the summit, $y = 147\text{ m}$, so

$$U(\text{summit}) = U(147\text{ m}) = mgh = (75 \times 9.8\text{ N})(147\text{ m}) = 108\text{ kJ}.$$

c. At sea level, $y = (147 - 195)\text{m} = -48\text{ m}$, so

$$U(\text{sea-level}) = (75 \times 9.8\text{ N})(-48\text{ m}) = -35.3\text{ kJ}.$$

Significance

Besides illustrating the use of **Equation 10.15** and **Equation 10.16**, the values of gravitational potential energy we found are reasonable. The gravitational potential energy is higher at the summit than at the base, and lower at sea level than at the base. Gravity does work on you on your way up, too! It does negative work and not quite as much (in magnitude), as your muscles do. But it certainly does work. Similarly, your muscles do work on your way down, as negative work. The numerical values of the potential energies depend on the choice of zero of potential energy, but the physically meaningful differences of potential energy do not. [Note that since **Equation 10.13** is a difference, the numerical values do not depend on the origin of coordinates.]

10.6 Check Your Understanding What are the values of the gravitational potential energy of the hiker at the base, summit, and sea level, with respect to a sea-level zero of potential energy?

View this **simulation (https://openstax.org/l/21conenerskat)** to learn about conservation of energy with a skater! Build tracks, ramps and jumps for the skater and view the kinetic energy, potential energy and friction as he moves. You can also take the skater to different planets or even space!

A sample chart of a variety of energies is shown in **Table 10.2** to give you an idea about typical energy values associated with certain events. Some of these are calculated using kinetic energy, whereas others are calculated by using quantities found in a form of potential energy that may not have been discussed at this point.

Object/phenomenon	Energy in joules
Big Bang	10^{68}
Annual world energy use	4.0×10^{20}

Table 10.2 Energy of Various Objects and Phenomena

Object/phenomenon	Energy in joules
Large fusion bomb (9 megaton)	3.8×10^{16}
Hiroshima-size fission bomb (10 kiloton)	4.2×10^{13}
1 barrel crude oil	5.9×10^{9}
1 ton TNT	4.2×10^{9}
1 gallon of gasoline	1.2×10^{8}
Daily adult food intake (recommended)	1.2×10^{7}
1000-kg car at 90 km/h	3.1×10^{5}
Tennis ball at 100 km/h	22
Mosquito $\left(10^{-2}\text{ g at }0.5\text{ m/s}\right)$	1.3×10^{-6}
Single electron in a TV tube beam	4.0×10^{-15}
Energy to break one DNA strand	10^{-19}

Table 10.2 Energy of Various Objects and Phenomena

Gravitational Potential Energy Beyond Earth

We have defined gravitational potential energy near the surface of the Earth, where the force of gravity is essentially constant (mg). However, an expression that is correct over larger distances must take into account the fact that Newton's universal force of gravitation varies inversely as the square of the distance from Earth's center ($\frac{-GM_E m}{r^2}$).

The change in potential energy when an object of mass m moves from a distance r_1 to a distance r_2 away from Earth's center is given by:

$$\Delta U = GM_\text{E} m\left(\frac{1}{r_1} - \frac{1}{r_2}\right).$$

Since $\Delta U = U_2 - U_1$, we can adopt a simple expression for U :

$$U = -\frac{GM_\text{E} m}{r}. \tag{10.17}$$

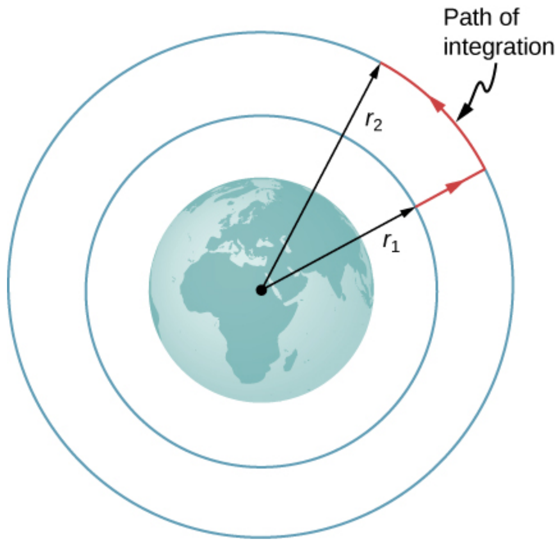


Figure 10.14 The change in gravitational potential energy can be evaluated for a displacement from a distance r_1 to a distance r_2.

Note two important items with this definition. First, $U \to 0$ as $r \to \infty$. The potential energy is zero when the two masses are infinitely far apart. Only the difference in U is important, so the choice of $U = 0$ for $r = \infty$ is merely one of convenience. (Recall that in earlier gravity problems, you were free to take $U = 0$ at the top or bottom of a building, or anywhere.) Second, note that U becomes increasingly more negative as the masses get closer. That is consistent with what you learned about potential energy. As the two masses are separated, positive work must be done against the force of gravity, and hence, U increases (becomes less negative). All masses naturally fall together under the influence of gravity, falling from a higher to a lower potential energy.

Example 10.11

Lifting a Payload

How much energy is required to lift the 9000-kg *Soyuz* vehicle from Earth's surface to the height of the ISS, 400 km above the surface?

Strategy

Use **Equation 10.16** to find the change in potential energy of the payload. That amount of work or energy must be supplied to lift the payload.

Solution

Paying attention to the fact that we start at Earth's surface and end at 400 km above the surface, the change in U is

$$\Delta U = U_{\text{orbit}} - U_{\text{Earth}} = -\frac{GM_{\text{E}}m}{R_{\text{E}} + 400\,\text{km}} - \left(-\frac{GM_{\text{E}}m}{R_{\text{E}}}\right).$$

We insert the values

$$m = 9000\,\text{kg}, \quad M_{\text{E}} = 5.96 \times 10^{24}\,\text{kg}, \quad R_{\text{E}} = 6.37 \times 10^{6}\ \text{m}$$

and convert 400 km into 4.00×10^5 m. We find $\Delta U = 3.32 \times 10^{10}$ J. It is positive, indicating an increase in potential energy, as we would expect.

Significance

For perspective, consider that the average US household energy use in 2013 was 909 kWh per month. That is energy of

$$909\ \text{kWh} \times 1000\ \text{W/kW} \times 3600\ \text{s/h} = 3.27 \times 10^9\ \text{J per month.}$$

So our result is an energy expenditure equivalent to 10 months. But this is just the energy needed to raise the payload 400 km. If we want the *Soyuz* to be in orbit so it can rendezvous with the ISS and not just fall back to Earth, it needs a lot of kinetic energy. As we see in the next section, that kinetic energy is about five times that of ΔU. In addition, far more energy is expended lifting the propulsion system itself. Space travel is not cheap.

10.7 Check Your Understanding Why not use the simpler expression $\Delta U = mg(y_2 - y_1)$? How significant would the error be? (Recall the previous result, in **Example 8.21**, that the value of g at 400 km above the Earth is $8.67\ \text{m/s}^2$.)

10.6 | Conservation of Energy

Learning Objectives

By the end of this section, you will be able to:

- Formulate the principle of conservation of mechanical energy, with or without the presence of non-conservative forces
- Use the conservation of mechanical energy to calculate various properties of simple systems
- Determine changes in gravitational potential energy over great distances
- Apply conservation of energy to determine escape velocity
- Determine whether astronomical bodies are gravitationally bound

In this section, we elaborate and extend the result we derived in the section on **Potential Energy**, where we re-wrote the work-energy theorem in terms of the change in the kinetic and potential energies of a particle. This will lead us to a discussion of the important principle of the conservation of mechanical energy. As you continue to examine other topics in physics, in later chapters of this book, you will see how this conservation law is generalized to encompass other types of energy and energy transfers. The last section of this chapter provides a preview.

The terms ‘conserved quantity’ and ‘conservation law’ have specific, scientific meanings in physics, which are different from the everyday meanings associated with the use of these words. (The same comment is also true about the scientific and everyday uses of the word ‘work.’) In everyday usage, you could conserve water by not using it, or by using less of it, or by re-using it. Water is composed of molecules consisting of two atoms of hydrogen and one of oxygen. Bring these atoms together to form a molecule and you create water; dissociate the atoms in such a molecule and you destroy water. However, in scientific usage, a **conserved quantity** for a system stays constant, changes by a definite amount that is transferred to other systems, and/or is converted into other forms of that quantity. A conserved quantity, in the scientific sense, can be transformed, but not strictly created or destroyed. Thus, there is no physical law of conservation of water.

Systems with a Single Particle or Object

We first consider a system with a single particle or object. Returning to our development of **Equation 10.13**, recall that we first separated all the forces acting on a particle into conservative and non-conservative types, and wrote the work done by each type of force as a separate term in the work-energy theorem. We then replaced the work done by the conservative forces by the change in the potential energy of the particle, combining it with the change in the particle’s kinetic energy to get **Equation 10.13**. Now, we write this equation without the middle step and define the sum of the kinetic and potential energies, $K + U = E;$ to be the **mechanical energy** of the particle.

Conservation of Energy

The mechanical energy E of a particle stays constant unless forces outside the system or non-conservative forces do work on it, in which case, the change in the mechanical energy is equal to the work done by the non-conservative forces:

$$W_{\text{nc, }AB} = \Delta(K + U)_{AB} = \Delta E_{AB}. \quad (10.18)$$

This statement expresses the concept of **energy conservation** for a classical particle as long as there is no non-conservative work. Recall that a classical particle is just a point mass that obeys Newton's laws of motion.

It is sometimes convenient to separate the case where the work done by non-conservative forces is zero, either because no such forces are assumed present, or, like the normal force, they do zero work when the motion is parallel to the surface. Then

$$0 = W_{\text{nc, }AB} = \Delta(K + U)_{AB} = \Delta E_{AB}. \quad (10.19)$$

In this case, the conservation of mechanical energy can be expressed as follows: The mechanical energy of a particle does not change if all the non-conservative forces that may act on it do no work. Understanding the concept of energy conservation is the important thing, not the particular equation you use to express it.

Problem-Solving Strategy: Conservation of Energy

1. Identify the body or bodies to be studied (the system). Often, in applications of the principle of mechanical energy conservation, we study more than one body at the same time.
2. Identify all forces acting on the body or bodies.
3. Determine whether each force that does work is conservative. If a non-conservative force (e.g., friction) is doing work, then mechanical energy is not conserved. The system must then be analyzed with non-conservative work, Equation 10.19.
4. For every force that does work, choose a reference point and determine the potential energy function for the force. The reference points for the various potential energies do not have to be at the same location.
5. Apply the principle of mechanical energy conservation by setting the sum of the kinetic energies and potential energies equal at every point of interest.

Example 10.12

Simple Pendulum

A particle of mass m is hung from the ceiling by a massless string of length 1.0 m, as shown in Figure 10.15. The particle is released from rest, when the angle between the string and the downward vertical direction is $30°$. What is its speed when it reaches the lowest point of its arc?

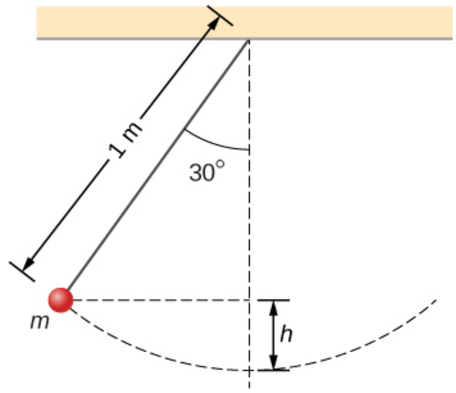


Figure 10.15 A particle hung from a string constitutes a simple pendulum. It is shown when released from rest, along with some distances used in analyzing the motion.

Strategy

Using our problem-solving strategy, the first step is to define that we are interested in the particle-Earth system. Second, only the gravitational force is acting on the particle, which is conservative (step 3). We neglect air resistance in the problem, and no work is done by the string tension, which is perpendicular to the arc of the motion. Therefore, the mechanical energy of the system is conserved, as represented by **Equation 10.19**, $0 = \Delta(K + U)$. Because the particle starts from rest, the increase in the kinetic energy is just the kinetic energy at the lowest point. This increase in kinetic energy equals the decrease in the gravitational potential energy, which we can calculate from the geometry. In step 4, we choose a reference point for zero gravitational potential energy to be at the lowest vertical point the particle achieves, which is mid-swing. Lastly, in step 5, we set the sum of energies at the highest point (initial) of the swing to the lowest point (final) of the swing to ultimately solve for the final speed.

Solution

We are neglecting non-conservative forces, so we write the energy conservation formula relating the particle at the highest point (initial) and the lowest point in the swing (final) as

$$K_\text{i} + U_\text{i} = K_\text{f} + U_\text{f}.$$

Since the particle is released from rest, the initial kinetic energy is zero. At the lowest point, we define the gravitational potential energy to be zero. Therefore our conservation of energy formula reduces to

$$\begin{aligned} 0 + mgh &= \frac{1}{2}mv^2 + 0 \\ v &= \sqrt{2gh}. \end{aligned}$$

The vertical height of the particle is not given directly in the problem. This can be solved for by using trigonometry and two givens: the length of the pendulum and the angle through which the particle is vertically pulled up. Looking at the diagram, the vertical dashed line is the length of the pendulum string. The vertical height is labeled *h*. The other partial length of the vertical string can be calculated with trigonometry. That piece is solved for by

$$\cos\theta = x/L,\ x = L\cos\theta.$$

Therefore, by looking at the two parts of the string, we can solve for the height *h*,

$$\begin{aligned} x + h &= L \\ L\cos\theta + h &= L \\ h &= L - L\cos\theta = L(1 - \cos\theta). \end{aligned}$$

We substitute this height into the previous expression solved for speed to calculate our result:

$$v = \sqrt{2gL(1 - \cos\theta)} = \sqrt{2\left(9.8\ \text{m/s}^2\right)(1\ \text{m})(1 - \cos 30°)} = 1.62\ \text{m/s}.$$

Significance

We found the speed directly from the conservation of mechanical energy, without having to solve the (complicated) differential equation for the motion of a pendulum. We can approach this problem in terms of bar graphs of total energy. Initially, the particle has all potential energy, being at the highest point, and no kinetic energy. When the particle crosses the lowest point at the bottom of the swing, the energy moves from the potential energy column to the kinetic energy column. Therefore, we can imagine a progression of this transfer as the particle moves between its highest point, lowest point of the swing, and back to the highest point (**Figure 10.16**). As the particle travels from the lowest point in the swing to the highest point on the far right hand side of the diagram, the energy bars go in reverse order from (c) to (b) to (a).

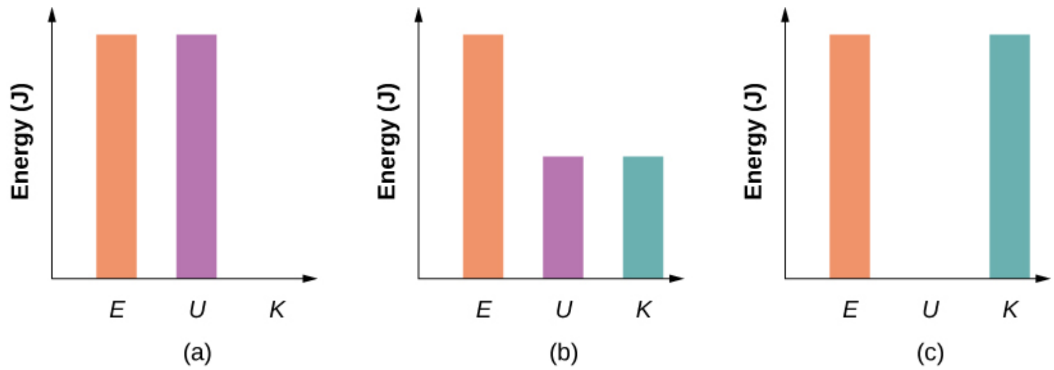


Figure 10.16 Bar graphs representing the total energy (E), potential energy (U), and kinetic energy (K) of the particle in different positions. (a) The total energy of the system equals the potential energy and the kinetic energy is zero, which is found at the highest point the particle reaches. (b) The particle is midway between the highest and lowest point, so the kinetic energy plus potential energy bar graphs equal the total energy. (c) The particle is at the lowest point of the swing, so the kinetic energy bar graph is the highest and equal to the total energy of the system.

10.8 Check Your Understanding How high above the bottom of its arc is the particle in the simple pendulum above, when its speed is $0.81\ \text{m/s}$?

Example 10.13

Air Resistance on a Falling Object

A helicopter is hovering at an altitude of $1\ \text{km}$ when a panel from its underside breaks loose and plummets to the ground (**Figure 10.17**). The mass of the panel is $15\ \text{kg},$ and it hits the ground with a speed of $45\ \text{m/s}$. How much mechanical energy was dissipated by air resistance during the panel's descent?

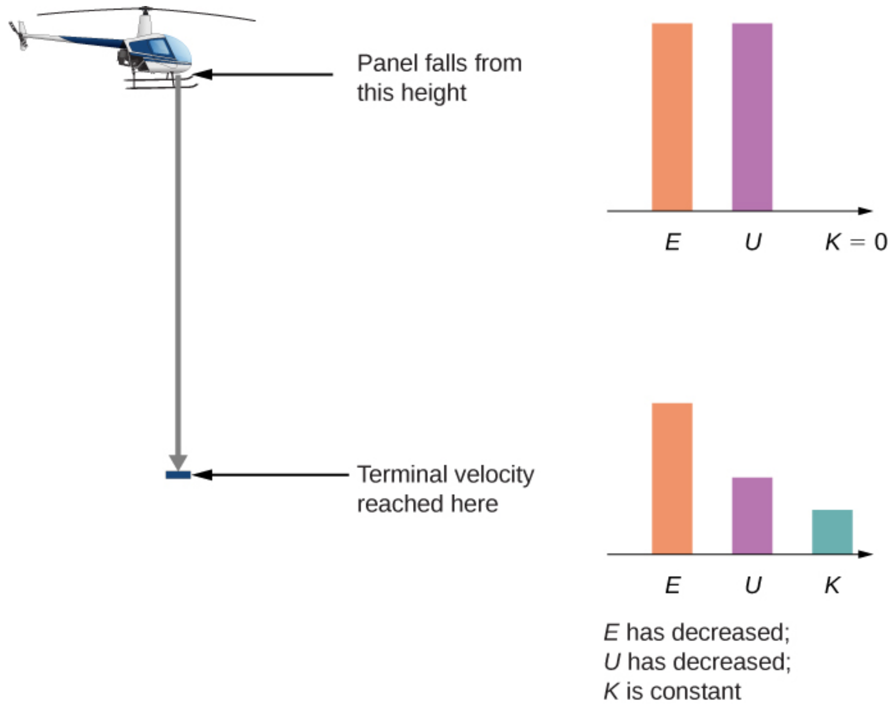


Figure 10.17 A helicopter loses a panel that falls until it reaches terminal velocity of 45 m/s. How much did air resistance contribute to the dissipation of energy in this problem?

Strategy

Step 1: Here only one body is being investigated.

Step 2: Gravitational force is acting on the panel, as well as air resistance, which is stated in the problem.

Step 3: Gravitational force is conservative; however, the non-conservative force of air resistance does negative work on the falling panel, so we can use the conservation of mechanical energy, in the form expressed by **Equation 10.18**, to find the energy dissipated. This energy is the magnitude of the work:

$$\Delta E_{\text{diss}} = \left|W_{\text{nc,if}}\right| = \left|\Delta(K+U)_{\text{if}}\right|.$$

Step 4: The initial kinetic energy, at $y_{\text{i}} = 1\text{ km},$ is zero. We set the gravitational potential energy to zero at ground level out of convenience.

Step 5: The non-conservative work is set equal to the energies to solve for the work dissipated by air resistance.

Solution

The mechanical energy dissipated by air resistance is the algebraic sum of the gain in the kinetic energy and loss in potential energy. Therefore the calculation of this energy is

$$\begin{aligned}\Delta E_{\text{diss}} &= \left|K_{\text{f}} - K_{\text{i}} + U_{\text{f}} - U_{\text{i}}\right| \\ &= \left|\tfrac{1}{2}(15\text{ kg})(45\text{ m/s})^2 - 0 + 0 - (15\text{ kg})\left(9.8\text{ m/s}^2\right)(1000\text{ m})\right| = 130\text{ kJ}.\end{aligned}$$

Significance

Most of the initial mechanical energy of the panel $\left(U_{\text{i}}\right)$, 147 kJ, was lost to air resistance. Notice that we were able to calculate the energy dissipated without knowing what the force of air resistance was, only that it was dissipative.

Example 10.14

Lifting a Payload

How much energy is required to lift the 9000-kg *Soyuz* vehicle from Earth’s surface to the height of the ISS, 400 km above the surface?

Strategy

Use **Note** to find the change in potential energy of the payload. That amount of work or energy must be supplied to lift the payload.

Solution

Paying attention to the fact that we start at Earth’s surface and end at 400 km above the surface, the change in U is

$$\Delta U = U_{\text{orbit}} - U_{\text{Earth}} = -\frac{GM_E m}{R_E + 400\ \text{km}} - \left(-\frac{GM_E m}{R_E}\right).$$

We insert the values

$$m = 9000\ \text{kg}, \quad M_E = 5.96 \times 10^{24}\ \text{kg}, \quad R_E = 6.37 \times 10^6\ \text{m}$$

and convert 400 km into $4.00 \times 10^5\ \text{m}$. We find $\Delta U = 3.32 \times 10^{10}\ \text{J}$. It is positive, indicating an increase in potential energy, as we would expect.

Significance

For perspective, consider that the average US household energy use in 2013 was 909 kWh per month. That is energy of

$$909\ \text{kWh} \times 1000\ \text{W/kW} \times 3600\ \text{s/h} = 3.27 \times 10^9\ \text{J per month.}$$

So our result is an energy expenditure equivalent to 10 months. But this is just the energy needed to raise the payload 400 km. If we want the *Soyuz* to be in orbit so it can rendezvous with the ISS and not just fall back to Earth, it needs a lot of kinetic energy. As we see in the next section, that kinetic energy is about five times that of ΔU. In addition, far more energy is expended lifting the propulsion system itself. Space travel is not cheap.

Exercise 10.1

Check Your Understanding Why not use the simpler expression $\Delta U = mg(y_2 - y_1)$? How significant would the error be? (Recall the previous result, in **Example 8.21**, that the value g at 400 km above the Earth is $8.67\ \text{m/s}^2$.)

Solution

The value of g drops by about 10% over this change in height. So $\Delta U = mg(y_2 - y_1)$ will give too large a value. If we use $g = 9.80\ \text{m/s}$, then we get

$$\Delta U = mg(y_2 - y_1) = 3.53 \times 10^{10}\ \text{J}$$

which is about 6% greater than that found with the correct method.

10.9 Check Your Understanding You probably recall that, neglecting air resistance, if you throw a projectile straight up, the time it takes to reach its maximum height equals the time it takes to fall from the maximum height back to the starting height. Suppose you cannot neglect air resistance, as in **Example 10.13**. Is the time the projectile takes to go up (a) greater than, (b) less than, or (c) equal to the time it takes to come back down? Explain.

10.10 Check Your Understanding What potential energy $U(x)$ can you substitute in **Equation 10.19** that will result in motion with constant velocity of 2 m/s for a particle of mass 1 kg and mechanical energy 1 J?

Energy Conservation and Universal Gravitation

The principles and problem-solving strategies we have discussed here apply equally well to problems in which the potential energy arises from the force of universal gravitation. The only change is to place the new expression for potential energy into the conservation of energy equation, $E = K_1 + U_1 = K_2 + U_2$.

$$\frac{1}{2}mv_1^2 - \frac{GMm}{r_1} = \frac{1}{2}mv_2^2 - \frac{GMm}{r_2} \tag{10.20}$$

Note that we use *M*, rather than M_E, as a reminder that we are not restricted to problems involving Earth. However, we still assume that $m << M$. (For problems in which this is not true, we need to include the kinetic energy of both masses and use conservation of momentum to relate the velocities to each other. But the principle remains the same.)

Escape Velocity

Escape velocity is often defined to be the *minimum* initial velocity of an object that is required to escape the surface of a planet (or any large body like a moon) and never return. As usual, we assume no energy lost to an atmosphere, should there be any.

Consider the case where an object is launched from the surface of a planet with an initial velocity directed away from the planet. With the *minimum* velocity needed to escape, the object would *just* come to rest infinitely far away, that is, the object gives up the last of its kinetic energy just as it reaches infinity, where the force of gravity becomes zero. Since $U \to 0$ as $r \to \infty$, this means the total energy is zero. Thus, we find the escape velocity from the surface of an astronomical body of mass *M* and radius *R* by setting the total energy equal to zero. At the surface of the body, the object is located at $r_1 = R$ and it has escape velocity $v_1 = v_{esc}$. It reaches $r_2 = \infty$ with velocity $v_2 = 0$. Substituting into **Equation 10.20**, we have

$$\frac{1}{2}mv_{esc}^2 - \frac{GMm}{R} = \frac{1}{2}m0^2 - \frac{GMm}{\infty} = 0.$$

Solving for v_{esc}, we obtain

$$v_{esc} = \sqrt{\frac{2GM}{R}}. \tag{10.21}$$

Notice that *m* has canceled out of the equation. The escape velocity is the same for all objects, regardless of mass. Also, we are not restricted to the surface of the planet; *R* can be any starting point beyond the surface of the planet.

Example 10.15

Escape from Earth

What is the escape speed from the surface of Earth? Assume there is no energy loss from air resistance. Compare this to the escape speed from the Sun, starting from Earth's orbit.

Strategy

We use **Equation 10.21**, clearly defining the values of *R* and *M*. To escape Earth, we need the mass and radius of Earth. For escaping the Sun, we need the mass of the Sun, and the orbital distance between Earth and the Sun.

Solution

Substituting the values for Earth's mass and radius directly into **Equation 10.21**, we obtain

$$v_{esc} = \sqrt{\frac{2GM}{R}} = \sqrt{\frac{2(6.67 \times 10^{-11}\ \text{N} \cdot \text{m}^2/\text{kg}^2)(5.96 \times 10^{24}\ \text{kg})}{6.37 \times 10^6\ \text{m}}} = 1.12 \times 10^4\ \text{m/s}.$$

That is about 11 km/s or 25,000 mph. To escape the Sun, starting from Earth's orbit, we use $R = R_{ES} = 1.50 \times 10^{11}\ \text{m}$ and $M_{Sun} = 1.99 \times 10^{30}\ \text{kg}$. The result is $v_{esc} = 4.21 \times 10^4\ \text{m/s}$ or about 42 km/s.

Significance

The speed needed to escape the Sun (leave the solar system) is nearly four times the escape speed from Earth's surface. But there is help in both cases. Earth is rotating, at a speed of nearly 1.7 km/s at the equator, and we can use that velocity to help escape, or to achieve orbit. For this reason, many commercial space companies maintain launch facilities near the equator. To escape the Sun, there is even more help. Earth revolves about the Sun at a speed of approximately 30 km/s. By launching in the direction that Earth is moving, we need only an additional 12 km/s. The use of gravitational assist from other planets, essentially a gravity slingshot technique, allows space probes to reach even greater speeds. In this slingshot technique, the vehicle approaches the planet and is accelerated by the planet's gravitational attraction. It has its greatest speed at the closest point of approach, although it decelerates in equal measure as it moves away. But relative to the planet, the vehicle's speed far before the approach, and long after, are the same. If the directions are chosen correctly, that can result in a significant increase (or decrease if needed) in the vehicle's speed relative to the rest of the solar system.

Visit this **website (https://openstax.org/l/21escapevelocit)** to learn more about escape velocity.

10.11 Check Your Understanding If we send a probe out of the solar system starting from Earth's surface, do we only have to escape the Sun?

Energy and gravitationally bound objects

As stated previously, escape velocity can be defined as the initial velocity of an object that can escape the surface of a moon or planet. More generally, it is the speed at *any* position such that the *total* energy is zero. If the total energy is zero or greater, the object escapes. If the total energy is negative, the object cannot escape. Let's see why that is the case.

As noted earlier, we see that $U \to 0$ as $r \to \infty$. If the total energy is zero, then as m reaches a value of r that approaches infinity, U becomes zero and so must the kinetic energy. Hence, m comes to rest infinitely far away from M. It has "just escaped" M. If the total energy is positive, then kinetic energy remains at $r = \infty$ and certainly m does not return. When the total energy is zero or greater, then we say that m is not gravitationally bound to M.

On the other hand, if the total energy is negative, then the kinetic energy must reach zero at some finite value of r, where U is negative and equal to the total energy. The object can never exceed this finite distance from M, since to do so would require the kinetic energy to become negative, which is not possible. We say m is **gravitationally bound** to M.

We have simplified this discussion by assuming that the object was headed directly away from the planet. What is remarkable is that the result applies for any velocity. Energy is a scalar quantity and hence **Equation 10.20** is a scalar equation—the direction of the velocity plays no role in conservation of energy. It is possible to have a gravitationally bound system where the masses do not "fall together," but maintain an orbital motion about each other.

We have one important final observation. Earlier we stated that if the total energy is zero or greater, the object escapes. Strictly speaking, **Equation 10.20** and **Equation 10.21** apply for point objects. They apply to finite-sized, spherically symmetric objects as well, provided that the value for r in **Equation 10.20** is always greater than the sum of the radii of the two objects. If r becomes less than this sum, then the objects collide. (Even for greater values of r, but near the sum of the radii, gravitational tidal forces could create significant effects if both objects are planet sized. Neither positive nor negative total energy precludes finite-sized masses from colliding. For real objects, direction is important.

Example 10.16

How Far Can an Object Escape?

Let's consider the preceding example again, where we calculated the escape speed from Earth and the Sun, starting from Earth's orbit. We noted that Earth already has an orbital speed of 30 km/s. As we see in the next section, that is the tangential speed needed to stay in circular orbit. If an object had this speed at the distance of Earth's orbit, but was headed directly away from the Sun, how far would it travel before coming to rest? Ignore the gravitational effects of any other bodies.

Strategy

The object has initial kinetic and potential energies that we can calculate. When its speed reaches zero, it is at its maximum distance from the Sun. We use **Equation 10.20**, conservation of energy, to find the distance at which kinetic energy is zero.

Solution

The initial position of the object is Earth's radius of orbit and the intial speed is given as 30 km/s. The final velocity is zero, so we can solve for the distance at that point from the conservation of energy equation. Using $R_{ES} = 1.50 \times 10^{11}\ \text{m}$ and $M_{Sun} = 1.99 \times 10^{30}\ \text{kg}$, we have

$$\frac{1}{2}mv_1^2 - \frac{GMm}{r_1} = \frac{1}{2}mv_2^2 - \frac{GMm}{r_2}$$

$$\frac{1}{2}\cancel{m}(3.0\times10^3\ \text{m/s})^2 - \frac{(6.67\times10^{-11}\ \text{N}\cdot\text{m/kg}^2)(1.99\times10^{30}\ \text{kg})\cancel{m}}{1.50\times10^{11}\ \text{m}}$$

$$= \frac{1}{2}\cancel{m}0^2 - \frac{(6.67\times10^{-11}\ \text{N}\cdot\text{m/kg}^2)(1.99\times10^{30}\ \text{kg})\cancel{m}}{r_2}$$

where the mass *m* cancels. Solving for r_2 we get $r_2 = 3.0 \times 10^{11}\ \text{m}$. Note that this is twice the initial distance from the Sun and takes us past Mars's orbit, but not quite to the asteroid belt.

Systems with Several Particles or Objects

Systems generally consist of more than one particle or object. However, the conservation of mechanical energy, in one of the forms in **Equation 10.18** or **Equation 10.19**, is a fundamental law of physics and applies to any system. You just have to include the kinetic and potential energies of all the particles, and the work done by all the non-conservative forces acting on them.

10.7 | Sources of Energy

Learning Objectives

By the end of this section, you will be able to:

- Describe energy transformations and conversions in general terms
- Explain what it means for an energy source to be renewable or nonrenewable

In this chapter, we have studied energy. We learned that energy can take different forms and can be transferred from one form to another. You will find that energy is discussed in many everyday, as well as scientific, contexts, because it is involved in all physical processes. It will also become apparent that many situations are best understood, or most easily conceptualized, by considering energy. So far, no experimental results have contradicted the conservation of energy. In fact, whenever measurements have appeared to conflict with energy conservation, new forms of energy have been discovered or recognized in accordance with this principle.

What are some other forms of energy? Many of these are covered in later chapters (also see **Figure 10.18**), but let's detail a few here:

- Atoms and molecules inside all objects are in random motion. The internal kinetic energy from these random motions is called *thermal energy*, because it is related to the temperature of the object. Note that thermal energy can also be transferred from one place to another, not transformed or converted, by the familiar processes of conduction, convection, and radiation. In this case, the energy is known as *heat energy*.
- *Electrical energy* is a common form that is converted to many other forms and does work in a wide range of practical situations.
- Fuels, such as gasoline and food, have *chemical energy*, which is potential energy arising from their molecular structure. Chemical energy can be converted into thermal energy by reactions like oxidation. Chemical reactions can also produce electrical energy, such as in batteries. Electrical energy can, in turn, produce thermal energy and light, such as in an electric heater or a light bulb.
- Light is just one kind of electromagnetic radiation, or *radiant energy*, which also includes radio, infrared, ultraviolet, X-rays, and gamma rays. All bodies with thermal energy can radiate energy in electromagnetic waves.
- *Nuclear energy* comes from reactions and processes that convert measurable amounts of mass into energy. Nuclear energy is transformed into radiant energy in the Sun, into thermal energy in the boilers of nuclear power plants, and then into electrical energy in the generators of power plants. These and all other forms of energy can be transformed into one another and, to a certain degree, can be converted into mechanical work.

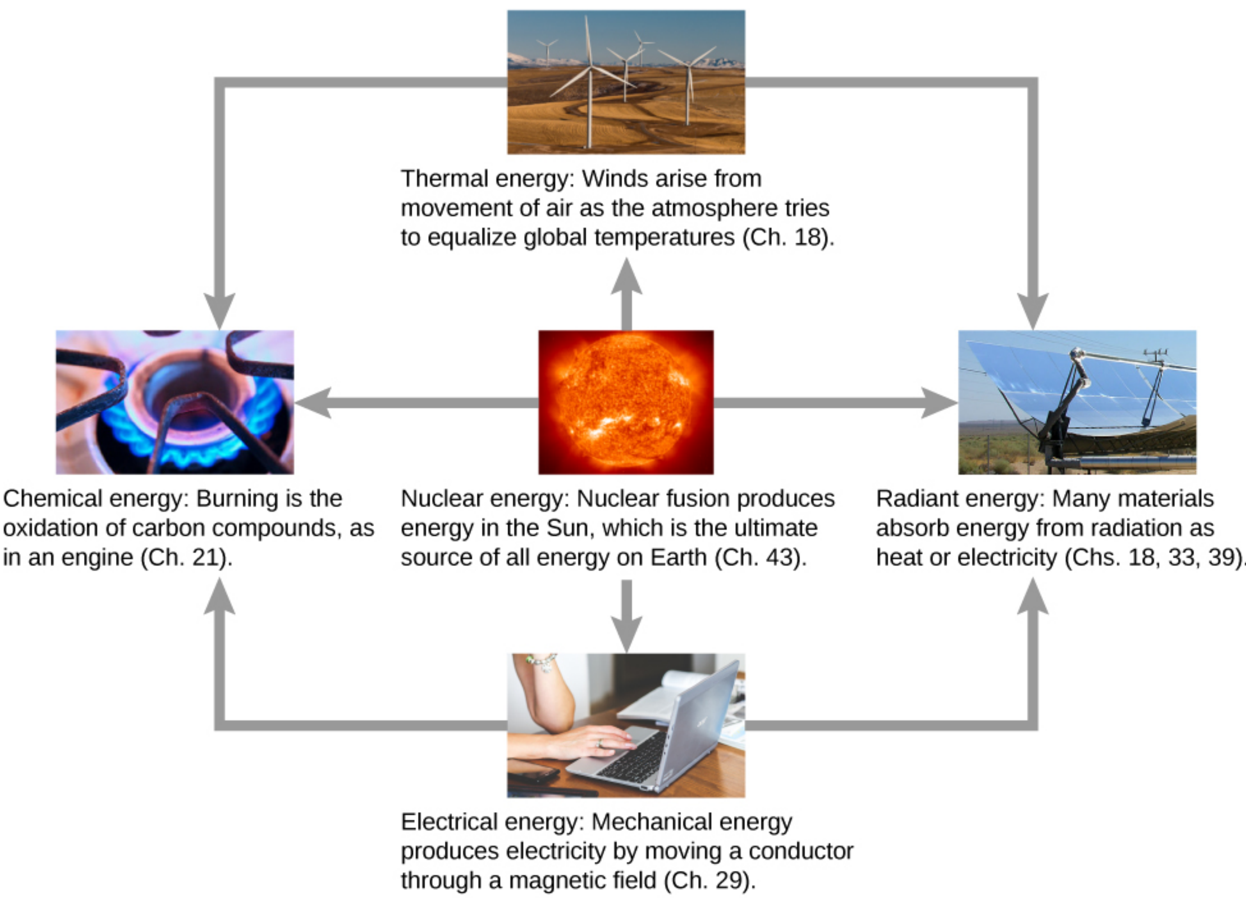


Figure 10.18 Energy that we use in society takes many forms, which be converted from one into another depending on the process involved. We will study many of these forms of energy in later chapters in this text. (credit "sun": modification of work by EIT - SOHO Consortium, ESA, NASA credit "solar panels": "modification of work by "kjkolb"/Wikimedia Commons; credit "gas burner": modification of work by Steven Depolo)

The transformation of energy from one form into another happens all the time. The chemical energy in food is converted into thermal energy through metabolism; light energy is converted into chemical energy through photosynthesis. Another example of energy conversion occurs in a solar cell. Sunlight impinging on a solar cell produces electricity, which can be used to run electric motors or heat water. In an example encompassing many steps, the chemical energy contained in coal is converted into thermal energy as it burns in a furnace, to transform water into steam, in a boiler. Some of the thermal energy in the steam is then converted into mechanical energy as it expands and spins a turbine, which is connected to a generator to produce electrical energy. In these examples, not all of the initial energy is converted into the forms mentioned, because some energy is always transferred to the environment.

Energy is an important element at all levels of society. We live in a very interdependent world, and access to adequate and reliable energy resources is crucial for economic growth and for maintaining the quality of our lives. The principal energy resources used in the world are shown in **Figure 10.19**. The figure distinguishes between two major types of energy sources: **renewable** and **non-renewable**, and further divides each type into a few more specific kinds. Renewable sources are energy sources that are replenished through naturally occurring, ongoing processes, on a time scale that is much shorter than the anticipated lifetime of the civilization using the source. Non-renewable sources are depleted once some of the energy they contain is extracted and converted into other kinds of energy. The natural processes by which non-renewable sources are formed typically take place over geological time scales.

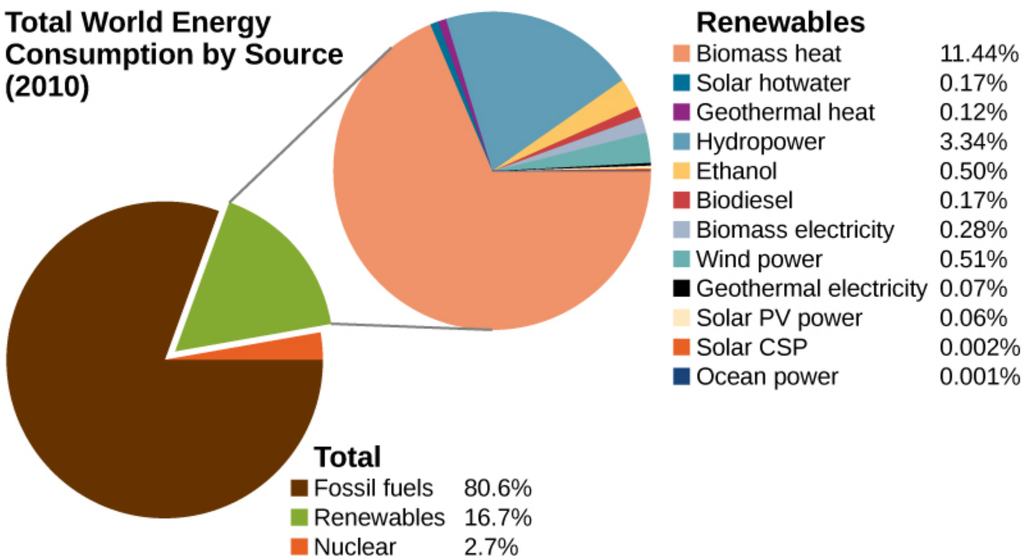


Figure 10.19 World energy consumption by source; the percentage of renewables is increasing, accounting for 19% in 2012.

Our most important non-renewable energy sources are fossil fuels, such as coal, petroleum, and natural gas. These account for about 81% of the world's energy consumption, as shown in the figure. Burning fossil fuels creates chemical reactions that transform potential energy, in the molecular structures of the reactants, into thermal energy and products. This thermal energy can be used to heat buildings or to operate steam-driven machinery. Internal combustion and jet engines convert some of the energy of rapidly expanding gases, released from burning gasoline, into mechanical work. Electrical power generation is mostly derived from transferring energy in expanding steam, via turbines, into mechanical work, which rotates coils of wire in magnetic fields to generate electricity. Nuclear energy is the other non-renewable source shown in **Figure 10.19** and supplies about 3% of the world's consumption. Nuclear reactions release energy by transforming potential energy, in the structure of nuclei, into thermal energy, analogous to energy release in chemical reactions. The thermal energy obtained from nuclear reactions can be transferred and converted into other forms in the same ways that energy from fossil fuels are used.

An unfortunate byproduct of relying on energy produced from the combustion of fossil fuels is the release of carbon dioxide into the atmosphere and its contribution to global warming. Nuclear energy poses environmental problems as well, including the safety and disposal of nuclear waste. Besides these important consequences, reserves of non-renewable sources of energy are limited and, given the rapidly growing rate of world energy consumption, may not last for more than a few hundred years. Considerable effort is going on to develop and expand the use of renewable sources of energy, involving a significant percentage of the world's physicists and engineers.

Four of the renewable energy sources listed in **Figure 10.19**—those using material from plants as fuel (biomass heat, ethanol, biodiesel, and biomass electricity)—involve the same types of energy transformations and conversions as just discussed for fossil and nuclear fuels. The other major types of renewable energy sources are hydropower, wind power, geothermal power, and solar power.

Hydropower is produced by converting the gravitational potential energy of falling or flowing water into kinetic energy and then into work to run electric generators or machinery. Converting the mechanical energy in ocean surface waves and tides is in development. Wind power also converts kinetic energy into work, which can be used directly to generate electricity, operate mills, and propel sailboats.

The interior of Earth has a great deal of thermal energy, part of which is left over from its original formation (gravitational potential energy converted into thermal energy) and part of which is released from radioactive minerals (a form of natural nuclear energy). It will take a very long time for this geothermal energy to escape into space, so people generally regard it as a renewable source, when actually, it’s just inexhaustible on human time scales.

The source of solar power is energy carried by the electromagnetic waves radiated by the Sun. Most of this energy is carried by visible light and infrared (heat) radiation. When suitable materials absorb electromagnetic waves, radiant energy is converted into thermal energy, which can be used to heat water, or when concentrated, to make steam and generate electricity (**Figure 10.20**). However, in another important physical process, known as the photoelectric effect, energetic radiation impinging on certain materials is directly converted into electricity. Materials that do this are called photovoltaics (PV in **Figure 10.19**). Some solar power systems use lenses or mirrors to concentrate the Sun’s rays, before converting their energy through photovoltaics, and these are qualified as CSP in **Figure 10.19**.

Figure 10.20 Solar cell arrays found in a sunny area converting the solar energy into stored electrical energy. (credit: modification of work by Sarah Swenty, U.S. Fish and Wildlife Service)

As we finish this chapter on energy and work, it is relevant to draw some distinctions between two sometimes misunderstood terms in the area of energy use. As we mentioned earlier, the “law of conservation of energy” is a very useful principle in analyzing physical processes. It cannot be proven from basic principles but is a very good bookkeeping device, and no exceptions have ever been found. It states that the total amount of energy in an isolated system always remains constant. Related to this principle, but remarkably different from it, is the important philosophy of energy conservation. This concept has to do with seeking to decrease the amount of energy used by an individual or group through reducing activities (e.g., turning down thermostats, diving fewer kilometers) and/or increasing conversion efficiencies in the performance of a particular task, such as developing and using more efficient room heaters, cars that have greater miles-per-gallon ratings, energy-efficient compact fluorescent lights, etc.

Since energy in an isolated system is not destroyed, created, or generated, you might wonder why we need to be concerned about our energy resources, since energy is a conserved quantity. The problem is that the final result of most energy transformations is waste heat, that is, work that has been “degraded” in the energy transformation. We will discuss this idea in more detail in the chapters on thermodynamics.

CHAPTER 10 REVIEW

KEY TERMS

average power work done in a time interval divided by the time interval

conserved quantity one that cannot be created or destroyed, but may be transformed between different forms of itself

energy conservation total energy of an isolated system is constant

escape velocity initial velocity an object needs to escape the gravitational pull of another; it is more accurately defined as the velocity of an object with zero total mechanical energy

gravitationally bound two object are gravitationally bound if their orbits are closed; gravitationally bound systems have a negative total mechanical energy

kinetic energy energy of motion, one-half an object's mass times the square of its speed

mechanical energy sum of the kinetic and potential energies

net work work done by all the forces acting on an object

non-renewable energy source that is not renewable, but is depleted by human consumption

potential energy function of position, energy possessed by an object relative to the system considered

potential energy difference negative of the work done acting between two points in space

power (or instantaneous power) rate of doing work

renewable energy source that is replenished by natural processes, over human time scales

rotational kinetic energy kinetic energy due to the rotation of an object; this is part of its total kinetic energy

work done when a force acts on something that undergoes a displacement from one position to another

work done by a force the dot product of the force vector and the displacement vector (from the initial position to the final position) along the path over which the force acts

work-energy theorem net work done on a particle is equal to the change in its kinetic energy

KEY EQUATIONS

Work done by a constant force	$W = \vec{\mathbf{F}} \cdot \vec{\Delta\mathbf{r}} = \vec{\mathbf{F}} \cdot (\vec{\mathbf{r}}_2 - \vec{\mathbf{r}}_1) = F\Delta r\cos(\theta)$
Work done by a constant force of kinetic friction	$W_{\text{fr}} = -f_k \vert l_{AB} \vert$
Work done going from *A* to *B* by Earth's gravity, near its surface	$W_{\text{grav},AB} = -mg(y_B - y_A)$
Kinetic energy of a non-relativistic particle	$K = \frac{1}{2}mv^2$
Work-energy theorem	$W_{\text{net}} = K_B - K_A$
Power as rate of doing work	$P = \frac{\Delta W}{\Delta t}$
Power found from the force and velocity	$P = Fv\cos\theta$
Difference of potential energy	$\Delta U_{AB} = U_B - U_A = -W_{AB}$
Potential energy with respect to zero of potential energy at $\vec{\mathbf{r}}_0$	$\Delta U = U(\vec{\mathbf{r}}) - U(\vec{\mathbf{r}}_0)$
Gravitational potential energy near Earth's surface	$U(y) = mgy + \text{const.}$

Gravitational potential energy beyond Earth's surface	$U = -\frac{GM_E m}{r}$
Conservation of energy with no non-conservative forces	$0 = W_{nc,\ AB} = \Delta(K + U)_{AB} = \Delta E_{AB}.$
Escape velocity from the surface of a planet	$v_{esc} = \sqrt{\frac{2GM}{R}}.$

SUMMARY

10.1 Work

- The work done by a force, acting over some displacement, is the dot product of the force and the displacement.
- The work done *against* a force is the negative of the work done *by* the force.
- The work done by a normal or frictional contact force must be determined in each particular case.
- The work done by the force of gravity, on an object near the surface of Earth, depends only on the weight of the object and the difference in height through which it moved.

10.2 Kinetic Energy

- The kinetic energy of a particle is the product of one-half its mass and the square of its speed, for non-relativistic speeds.
- The kinetic energy of a system is the sum of the kinetic energies of all the particles in the system.
- The rotational kinetic energy is the kinetic energy of rotation of a rotating rigid body or system of particles, and is given by $K = \frac{1}{2}I\omega^2$, where I is the moment of inertia, or "rotational mass" of the rigid body or system of particles.

10.3 Work-Energy Theorem

- The net work done on a particle is equal to the change in the particle's kinetic energy. This is the work-energy theorem.
- You can use the work-energy theorem to find certain properties of a system, without having to solve the Newton's second law.

10.4 Power

- Power is the rate of doing work; that is, the amount of work done per unit time.
- The power delivered by a force, acting on a moving particle, is the product of the work done by that force times the particle's displacement. This power can also be found as the product of the magnitude of the force times the magnitude of the particle's velocity times the cosine of the angle between the force and the velocity.

10.5 Potential Energy

- For a single-particle system, the difference of potential energy is the opposite of the work done by the forces acting on the particle as it moves from one position to another.
- Since only differences of potential energy are physically meaningful, the zero of the potential energy function can be chosen at a convenient location.
- The potential energies for Earth's constant gravity, near its surface, and for a Hooke's law force are linear and quadratic functions of position, respectively.
- The force of gravity changes as we move away from Earth, and the expression for gravitational potential energy must reflect this change.

10.6 Conservation of Energy

- A conserved quantity is a physical property that stays constant regardless of the path taken.

- A form of the work-energy theorem says that the change in the mechanical energy of a particle equals the work done on it by non-conservative forces.
- If non-conservative forces do no work and there are no external forces, the mechanical energy of a particle stays constant. This is a statement of the conservation of mechanical energy and there is no change in the total mechanical energy.
- The total energy of a system is the sum of kinetic and gravitational potential energy, and this total energy is conserved in orbital motion.
- Objects must have a minimum velocity, the escape velocity, to leave a planet and not return.
- Objects with total energy less than zero are bound; those with zero or greater are unbounded.

10.7 Sources of Energy

- Energy can be transferred from one system to another and transformed or converted from one type into another. Some of the basic types of energy are kinetic, potential, thermal, and electromagnetic.
- Renewable energy sources are those that are replenished by ongoing natural processes, over human time scales. Examples are wind, water, geothermal, and solar power.
- Non-renewable energy sources are those that are depleted by consumption, over human time scales. Examples are fossil fuel and nuclear power.

CONCEPTUAL QUESTIONS

10.1 Work

2. Give an example of something we think of as work in everyday circumstances that is not work in the scientific sense. Is energy transferred or changed in form in your example? If so, explain how this is accomplished without doing work.

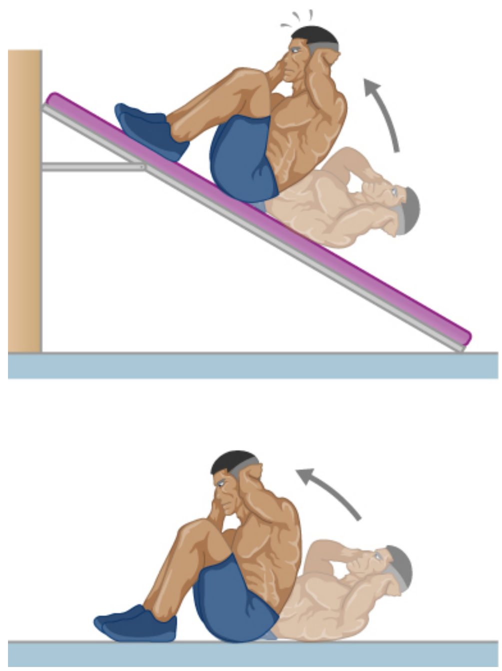
3. Give an example of a situation in which there is a force and a displacement, but the force does no work. Explain why it does no work.

4. Describe a situation in which a force is exerted for a long time but does no work. Explain.

5. A body moves in a circle at constant speed. Does the centripetal force that accelerates the body do any work? Explain.

6. Suppose you throw a ball upward and catch it when it returns at the same height. How much work does the gravitational force do on the ball over its entire trip?

7. Why is it more difficult to do sit-ups while on a slant board than on a horizontal surface? (See below.)



8. As a young man, Tarzan climbed up a vine to reach his tree house. As he got older, he decided to build and use a staircase instead. Since the work of the gravitational force *mg* is path independent, what did the King of the Apes gain in using stairs?

10.2 Kinetic Energy

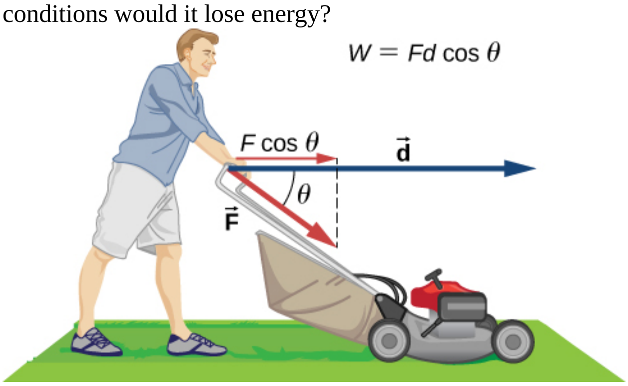
9. One particle has mass *m* and a second particle has mass 2*m*. The second particle is moving with speed *v* and the first with speed 2*v*. How do their kinetic energies compare?

10. A person drops a pebble of mass m_1 from a height h, and it hits the floor with kinetic energy K. The person drops another pebble of mass m_2 from a height of $2h$, and it hits the floor with the same kinetic energy K. How do the masses of the pebbles compare?

11. A solid sphere is rotating about an axis through its center at a constant rotation rate. Another hollow sphere of the same mass and radius is rotating about its axis through the center at the same rotation rate. Which sphere has a greater rotational kinetic energy?

10.3 Work-Energy Theorem

12. The person shown below does work on the lawn mower. Under what conditions would the mower gain energy from the person pushing the mower? Under what conditions would it lose energy?



13. Work done on a system puts energy into it. Work done by a system removes energy from it. Give an example for each statement.

14. Two marbles of masses m and $2m$ are dropped from a height h. Compare their kinetic energies when they reach the ground.

15. Compare the work required to accelerate a car of mass 2000 kg from 30.0 to 40.0 km/h with that required for an acceleration from 50.0 to 60.0 km/h.

16. Suppose you are jogging at constant velocity. Are you doing any work on the environment and vice versa?

17. Two forces act to double the speed of a particle, initially moving with kinetic energy of 1 J. One of the forces does 4 J of work. How much work does the other force do?

10.4 Power

18. Most electrical appliances are rated in watts. Does this rating depend on how long the appliance is on? (When off, it is a zero-watt device.) Explain in terms of the definition of power.

19. Explain, in terms of the definition of power, why energy consumption is sometimes listed in kilowatt-hours rather than joules. What is the relationship between these two energy units?

20. A spark of static electricity, such as that you might receive from a doorknob on a cold dry day, may carry a few hundred watts of power. Explain why you are not injured by such a spark.

21. Does the work done in lifting an object depend on how fast it is lifted? Does the power expended depend on how fast it is lifted?

22. Can the power expended by a force be negative?

23. How can a 50-W light bulb use more energy than a 1000-W oven?

10.5 Potential Energy

24. The kinetic energy of a system must always be positive or zero. Explain whether this is true for the potential energy of a system.

25. The force exerted by a diving board is conservative, provided the internal friction is negligible. Assuming friction is negligible, describe changes in the potential energy of a diving board as a swimmer drives from it, starting just before the swimmer steps on the board until just after his feet leave it.

26. Describe the gravitational potential energy transfers and transformations for a javelin, starting from the point at which an athlete picks up the javelin and ending when the javelin is stuck into the ground after being thrown.

27. A couple of soccer balls of equal mass are kicked off the ground at the same speed but at different angles. Soccer ball A is kicked off at an angle slightly above the horizontal, whereas ball B is kicked slightly below the vertical. How do each of the following compare for ball A and ball B? (a) The initial kinetic energy and (b) the change in gravitational potential energy from the ground to the highest point? If the energy in part (a) differs from part (b), explain why there is a difference between the two energies.

28. What is the dominant factor that affects the speed of an object that started from rest down a frictionless incline if the only work done on the object is from gravitational forces?

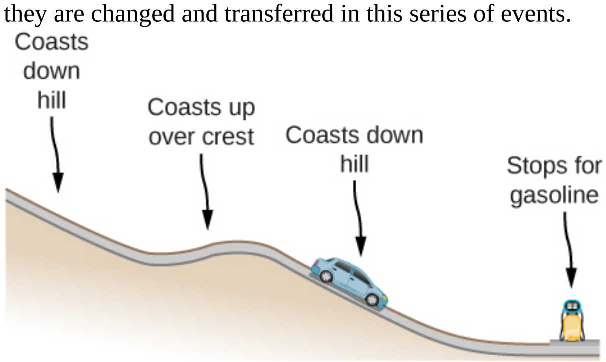
29. Two people observe a leaf falling from a tree. One person is standing on a ladder and the other is on the ground. If each person were to compare the energy of the leaf observed, would each person find the following to be the same or different for the leaf, from the point where it falls off the tree to when it hits the ground: (a) the kinetic energy of the leaf; (b) the change in gravitational potential energy; (c) the final gravitational potential energy?

30. It was shown that the energy required to lift a satellite into a *low* Earth orbit (the change in potential energy) is only a small fraction of the kinetic energy needed to keep it in orbit. Is this true for larger orbits? Is there a trend to the ratio of kinetic energy to change in potential energy as the size of the orbit increases?

10.6 Conservation of Energy

31. When a body slides down an inclined plane, does the work of friction depend on the body's initial speed? Answer the same question for a body sliding down a curved surface.

32. Consider the following scenario. A car for which friction is *not* negligible accelerates from rest down a hill, running out of gasoline after a short distance (see below). The driver lets the car coast farther down the hill, then up and over a small crest. He then coasts down that hill into a gas station, where he brakes to a stop and fills the tank with gasoline. Identify the forms of energy the car has, and how they are changed and transferred in this series of events.

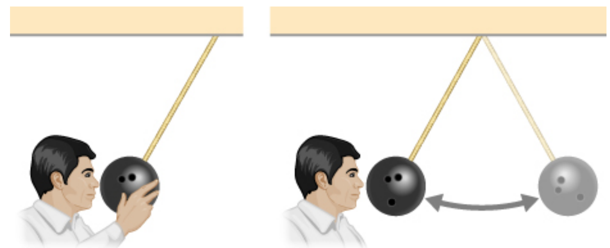


33. A dropped ball bounces to one-half its original height. Discuss the energy transformations that take place.

34. “$E = K + U =$ constant is a special case of the work-energy theorem.” Discuss this statement.

35. In a common physics demonstration, a bowling ball is suspended from the ceiling by a rope.

The professor pulls the ball away from its equilibrium position and holds it adjacent to his nose, as shown below. He releases the ball so that it swings directly away from him. Does he get struck by the ball on its return swing? What is he trying to show in this demonstration?



36. A child jumps up and down on a bed, reaching a higher height after each bounce. Explain how the child can increase his maximum gravitational potential energy with each bounce.

37. Can a non-conservative force increase the mechanical energy of the system?

38. Neglecting air resistance, how much would I have to raise the vertical height if I wanted to double the impact speed of a falling object?

39. It was stated that a satellite with negative total energy is in a bound orbit, whereas one with zero or positive total energy is in an unbounded orbit. Why is this true? What choice for gravitational potential energy was made such that this is true?

40. It was shown that the energy required to lift a satellite into a *low* Earth orbit (the change in potential energy) is only a small fraction of the kinetic energy needed to keep it in orbit. Is this true for larger orbits? Is there a trend to the ratio of kinetic energy to change in potential energy as the size of the orbit increases?

PROBLEMS

10.1 Work

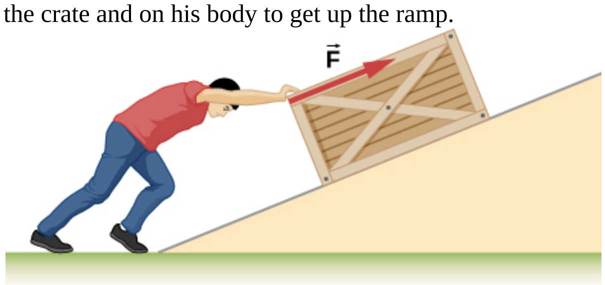
41. How much work does a supermarket checkout attendant do on a can of soup he pushes 0.600 m horizontally with a force of 5.00 N?

42. A 75.0-kg person climbs stairs, gaining 2.50 m in height. Find the work done to accomplish this task.

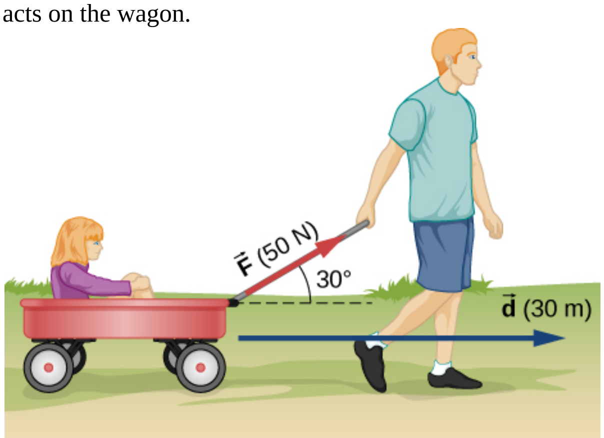
43. (a) Calculate the work done on a 1500-kg elevator car by its cable to lift it 40.0 m at constant speed, assuming friction averages 100 N. (b) What is the work done on the lift by the gravitational force in this process? (c) What is the total work done on the lift?

44. Suppose a car travels 108 km at a speed of 30.0 m/s, and uses 2.0 gal of gasoline. Only 30% of the gasoline goes into useful work by the force that keeps the car moving at constant speed despite friction. (The energy content of gasoline is about 140 MJ/gal.) (a) What is the magnitude of the force exerted to keep the car moving at constant speed? (b) If the required force is directly proportional to speed, how many gallons will be used to drive 108 km at a speed of 28.0 m/s?

45. Calculate the work done by an 85.0-kg man who pushes a crate 4.00 m up along a ramp that makes an angle of $20.0°$ with the horizontal (see below). He exerts a force of 500 N on the crate parallel to the ramp and moves at a constant speed. Be certain to include the work he does on the crate and on his body to get up the ramp.



46. How much work is done by the boy pulling his sister 30.0 m in a wagon as shown below? Assume no friction acts on the wagon.



47. A shopper pushes a grocery cart 20.0 m at constant speed on level ground, against a 35.0 N frictional force. He pushes in a direction $25.0°$ below the horizontal. (a) What is the work done on the cart by friction? (b) What is the work done on the cart by the gravitational force? (c) What is the work done on the cart by the shopper? (d) Find the force the shopper exerts, using energy considerations. (e) What is the total work done on the cart?

48. A constant 20-N force pushes a small ball in the direction of the force over a distance of 5.0 m. What is the work done by the force?

49. A toy cart is pulled a distance of 6.0 m in a straight line across the floor. The force pulling the cart has a magnitude of 20 N and is directed at $37°$ above the horizontal. What is the work done by this force?

50. A 5.0-kg box rests on a horizontal surface. The coefficient of kinetic friction between the box and surface is $\mu_K = 0.50$. A horizontal force pulls the box at constant velocity for 10 cm. Find the work done by (a) the applied horizontal force, (b) the frictional force, and (c) the net force.

51. A sled plus passenger with total mass 50 kg is pulled 20 m across the snow $(\mu_k = 0.20)$ at constant velocity by a force directed $25°$ above the horizontal. Calculate (a) the work of the applied force, (b) the work of friction, and (c) the total work.

52. Suppose that the sled plus passenger of the preceding problem is pushed 20 m across the snow at constant velocity by a force directed $30°$ below the horizontal. Calculate (a) the work of the applied force, (b) the work of friction, and (c) the total work.

53. How much work is done against the gravitational force on a 5.0-kg briefcase when it is carried from the ground floor to the roof of the Empire State Building, a vertical climb of 380 m?

10.2 Kinetic Energy

54. Compare the kinetic energy of a 20,000-kg truck moving at 110 km/h with that of an 80.0-kg astronaut in orbit moving at 27,500 km/h.

55. (a) How fast must a 3000-kg elephant move to have the same kinetic energy as a 65.0-kg sprinter running at 10.0 m/s? (b) Discuss how the larger energies needed for the movement of larger animals would relate to metabolic rates.

56. Estimate the kinetic energy of a 90,000-ton aircraft carrier moving at a speed of at 30 knots. You will need to look up the definition of a nautical mile to use in converting the unit for speed, where 1 knot equals 1 nautical mile per hour.

57. Calculate the kinetic energies of (a) a 2000.0-kg automobile moving at 100.0 km/h; (b) an 80.-kg runner sprinting at 10. m/s; and (c) a 9.1×10^{-31} -kg electron moving at 2.0×10^7 m/s.

58. A 5.0-kg body has three times the kinetic energy of an 8.0-kg body. Calculate the ratio of the speeds of these bodies.

59. An 8.0-g bullet has a speed of 800 m/s. (a) What is its kinetic energy? (b) What is its kinetic energy if the speed is halved?

60. (a) Calculate the rotational kinetic energy of Earth on its axis. (b) What is the rotational kinetic energy of Earth in its orbit around the Sun?

61. Calculate the rotational kinetic energy of a 12-kg motorcycle wheel if its angular velocity is 120 rad/s and its inner radius is 0.280 m and outer radius 0.330 m.

62. A baseball pitcher throws the ball in a motion where there is rotation of the forearm about the elbow joint as well as other movements. If the linear velocity of the ball relative to the elbow joint is 20.0 m/s at a distance of 0.480 m from the joint and the moment of inertia of the forearm is $0.500\ \text{kg-m}^2$, what is the rotational kinetic energy of the forearm?

63. A diver goes into a somersault during a dive by tucking her limbs. If her rotational kinetic energy is 100 J and her moment of inertia in the tuck is $9.0\ \text{kg}\cdot\text{m}^2$, what is her rotational rate during the somersault?

64. A neutron star of mass $2\times10^{30}\ \text{kg}$ and radius 10 km rotates with a period of 0.02 seconds. What is its rotational kinetic energy?

65. An electric sander consisting of a rotating disk of mass 0.7 kg and radius 10 cm rotates at 15 rev/sec. When applied to a rough wooden wall the rotation rate decreases by 20%. (a) What is the final rotational kinetic energy of the rotating disk? (b) How much has its rotational kinetic energy decreased?

10.3 Work-Energy Theorem

66. (a) Calculate the force needed to bring a 950-kg car to rest from a speed of 90.0 km/h in a distance of 120 m (a fairly typical distance for a non-panic stop). (b) Suppose instead the car hits a concrete abutment at full speed and is brought to a stop in 2.00 m. Calculate the force exerted on the car and compare it with the force found in part (a).

67. A car's bumper is designed to withstand a 4.0-km/h (1.1-m/s) collision with an immovable object without damage to the body of the car. The bumper cushions the shock by absorbing the force over a distance. Calculate the magnitude of the average force on a bumper that collapses 0.200 m while bringing a 900-kg car to rest from an initial speed of 1.1 m/s.

68. Boxing gloves are padded to lessen the force of a blow. (a) Calculate the force exerted by a boxing glove on an opponent's face, if the glove and face compress 7.50 cm during a blow in which the 7.00-kg arm and glove are brought to rest from an initial speed of 10.0 m/s. (b) Calculate the force exerted by an identical blow in the gory old days when no gloves were used, and the knuckles and face would compress only 2.00 cm. Assume the change in mass by removing the glove is negligible. (c) Discuss the magnitude of the force with glove on. Does it seem high enough to cause damage even though it is lower than the force with no glove?

69. Using energy considerations, calculate the average force a 60.0-kg sprinter exerts backward on the track to accelerate from 2.00 to 8.00 m/s in a distance of 25.0 m, if he encounters a headwind that exerts an average force of 30.0 N against him.

70. A 5.0-kg box has an acceleration of $2.0\ \text{m/s}^2$ when it is pulled by a horizontal force across a surface with $\mu_K = 0.50.$ Find the work done over a distance of 10 cm by (a) the horizontal force, (b) the frictional force, and (c) the net force. (d) What is the change in kinetic energy of the box?

71. A constant 10-N horizontal force is applied to a 20-kg cart at rest on a level floor. If friction is negligible, what is the speed of the cart when it has been pushed 8.0 m?

72. In the preceding problem, the 10-N force is applied at an angle of $45°$ below the horizontal. What is the speed of the cart when it has been pushed 8.0 m?

73. Compare the work required to stop a 100-kg crate sliding at 1.0 m/s and an 8.0-g bullet traveling at 500 m/s.

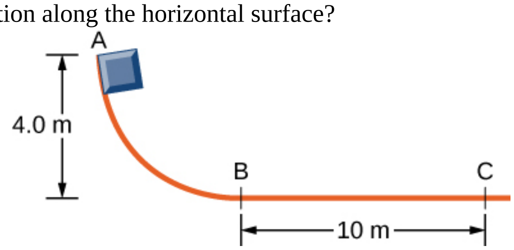
74. A wagon with its passenger sits at the top of a hill. The wagon is given a slight push and rolls 100 m down a $10°$ incline to the bottom of the hill. What is the wagon's speed when it reaches the end of the incline. Assume that the retarding force of friction is negligible.

75. An 8.0-g bullet with a speed of 800 m/s is shot into a wooden block and penetrates 20 cm before stopping. What is the average force of the wood on the bullet? Assume the block does not move.

76. A 2.0-kg block starts with a speed of 10 m/s at the bottom of a plane inclined at $37°$ to the horizontal. The coefficient of sliding friction between the block and plane is $\mu_k = 0.30.$ (a) Use the work-energy principle to determine how far the block slides along the plane before momentarily coming to rest. (b) After stopping, the block slides back down the plane. What is its speed when it reaches the bottom? (*Hint:* For the round trip, only the force of friction does work on the block.)

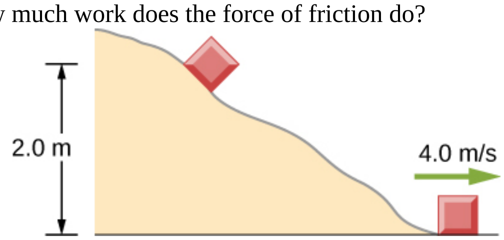
77. When a 3.0-kg block is pushed against a massless spring of force constant constant 4.5×10^3 N/m, the spring is compressed 8.0 cm. The block is released, and it slides 2.0 m (from the point at which it is released) across a horizontal surface before friction stops it. What is the coefficient of kinetic friction between the block and the surface?

78. A small block of mass 200 g starts at rest at A, slides to B where its speed is $v_B = 8.0$ m/s, then slides along the horizontal surface a distance 10 m before coming to rest at C. (See below.) (a) What is the work of friction along the curved surface? (b) What is the coefficient of kinetic friction along the horizontal surface?



79. A small object is placed at the top of an incline that is essentially frictionless. The object slides down the incline onto a rough horizontal surface, where it stops in 5.0 s after traveling 60 m. (a) What is the speed of the object at the bottom of the incline and its acceleration along the horizontal surface? (b) What is the height of the incline?

80. When released, a 100-g block slides down the path shown below, reaching the bottom with a speed of 4.0 m/s. How much work does the force of friction do?



81. A 0.22LR-caliber bullet like that mentioned in **Example 10.7** is fired into a door made of a single thickness of 1-inch pine boards. How fast would the bullet be traveling after it penetrated through the door?

82. A sled starts from rest at the top of a snow-covered incline that makes a $22°$ angle with the horizontal. After sliding 75 m down the slope, its speed is 14 m/s. Use the work-energy theorem to calculate the coefficient of kinetic friction between the runners of the sled and the snowy surface.

10.4 Power

83. A person in good physical condition can put out 100 W of useful power for several hours at a stretch, perhaps by pedaling a mechanism that drives an electric generator. Neglecting any problems of generator efficiency and practical considerations such as resting time: (a) How many people would it take to run a 4.00-kW electric clothes dryer? (b) How many people would it take to replace a large electric power plant that generates 800 MW?

84. What is the cost of operating a 3.00-W electric clock for a year if the cost of electricity is \$0.0900 per $\text{kW} \cdot \text{h}$?

85. A large household air conditioner may consume 15.0 kW of power. What is the cost of operating this air conditioner 3.00 h per day for 30.0 d if the cost of electricity is \$0.110 per $\text{kW} \cdot \text{h}$?

86. (a) What is the average power consumption in watts of an appliance that uses 5.00 $\text{kW} \cdot \text{h}$ of energy per day? (b) How many joules of energy does this appliance consume in a year?

87. (a) What is the average useful power output of a person who does 6.00×10^6 J of useful work in 8.00 h? (b) Working at this rate, how long will it take this person to lift 2000 kg of bricks 1.50 m to a platform? (Work done to lift his body can be omitted because it is not considered useful output here.)

88. A 500-kg dragster accelerates from rest to a final speed of 110 m/s in 400 m (about a quarter of a mile) and encounters an average frictional force of 1200 N. What is its average power output in watts and horsepower if this takes 7.30 s?

89. (a) How long will it take an 850-kg car with a useful power output of 40.0 hp (1 hp equals 746 W) to reach a speed of 15.0 m/s, neglecting friction? (b) How long will this acceleration take if the car also climbs a 3.00-m high hill in the process?

90. (a) Find the useful power output of an elevator motor that lifts a 2500-kg load a height of 35.0 m in 12.0 s, if it also increases the speed from rest to 4.00 m/s. Note that the total mass of the counterbalanced system is 10,000 kg—so that only 2500 kg is raised in height, but the full 10,000 kg is accelerated. (b) What does it cost, if electricity is \$0.0900 per $\text{kW} \cdot \text{h}$?

91. (a) How long would it take a 1.50×10^5-kg airplane with engines that produce 100 MW of power to reach a speed of 250 m/s and an altitude of 12.0 km if air resistance were negligible? (b) If it actually takes 900 s, what is the power? (c) Given this power, what is the average force of air resistance if the airplane takes 1200 s? (*Hint:* You must find the distance the plane travels in 1200 s assuming constant acceleration.)

92. Calculate the power output needed for a 950-kg car to climb a $2.00°$ slope at a constant 30.0 m/s while encountering wind resistance and friction totaling 600 N.

93. A man of mass 80 kg runs up a flight of stairs 20 m high in 10 s. (a) how much power is used to lift the man? (b) If the man's body is 25% efficient, how much power does he expend?

94. The man of the preceding problem consumes approximately 1.05×10^7 J (2500 food calories) of energy per day in maintaining a constant weight. What is the average power he produces over a day? Compare this with his power production when he runs up the stairs.

95. An electron in a television tube is accelerated uniformly from rest to a speed of 8.4×10^7 m/s over a distance of 2.5 cm. What is the power delivered to the electron at the instant that its displacement is 1.0 cm?

96. Coal is lifted out of a mine a vertical distance of 50 m by an engine that supplies 500 W to a conveyer belt. How much coal per minute can be brought to the surface? Ignore the effects of friction.

97. A girl pulls her 15-kg wagon along a flat sidewalk by applying a 10-N force at $37°$ to the horizontal. Assume that friction is negligible and that the wagon starts from rest. (a) How much work does the girl do on the wagon in the first 2.0 s. (b) How much instantaneous power does she exert at $t = 2.0\text{ s}$?

98. A typical automobile engine has an efficiency of 25%. Suppose that the engine of a 1000-kg automobile has a maximum power output of 140 hp. What is the maximum grade that the automobile can climb at 50 km/h if the frictional retarding force on it is 300 N?

99. When jogging at 13 km/h on a level surface, a 70-kg man uses energy at a rate of approximately 850 W. Using the facts that the "human engine" is approximately 25% efficient, determine the rate at which this man uses energy when jogging up a $5.0°$ slope at this same speed. Assume that the frictional retarding force is the same in both cases.

10.5 Potential Energy

100. Using values from **Table 10.2**, how many DNA molecules could be broken by the energy carried by a single electron in the beam of an old-fashioned TV tube? (These electrons were not dangerous in themselves, but they did create dangerous X-rays. Later-model tube TVs had shielding that absorbed X-rays before they escaped and exposed viewers.)

101. If the energy in fusion bombs were used to supply the energy needs of the world, how many of the 9-megaton variety would be needed for a year's supply of energy (using data from **???**)?

102. A camera weighing 10 N falls from a small drone hovering 20 m overhead and enters free fall. What is the gravitational potential energy change of the camera from the drone to the ground if you take a reference point of (a) the ground being zero gravitational potential energy? (b) The drone being zero gravitational potential energy? What is the gravitational potential energy of the camera (c) before it falls from the drone and (d) after the camera lands on the ground if the reference point of zero gravitational potential energy is taken to be a second person looking out of a building 30 m from the ground?

103. Someone drops a $50-\text{g}$ pebble off of a docked cruise ship, 70.0 m from the water line. A person on a dock 3.0 m from the water line holds out a net to catch the pebble. (a) How much work is done on the pebble by gravity during the drop? (b) What is the change in the gravitational potential energy during the drop? If the gravitational potential energy is zero at the water line, what is the gravitational potential energy (c) when the pebble is dropped? (d) When it reaches the net? What if the gravitational potential energy was 30.0 Joules at water level? (e) Find the answers to the same questions in (c) and (d).

104. A cat's crinkle ball toy of mass 15 g is thrown straight up with an initial speed of 3 m/s. Assume in this problem that air drag is negligible. (a) What is the kinetic energy of the ball as it leaves the hand? (b) How much work is done by the gravitational force during the ball's rise to its peak? (c) What is the change in the gravitational potential energy of the ball during the rise to its peak? (d) If the gravitational potential energy is taken to be zero at the point where it leaves your hand, what is the gravitational potential energy when it reaches the maximum height? (e) What if the gravitational potential energy is taken to be zero at the maximum height the ball reaches, what would the gravitational potential energy be when it leaves the hand? (f) What is the maximum height the ball reaches?

10.6 Conservation of Energy

105. A boy throws a ball of mass 0.25 kg straight upward with an initial speed of 20 m/s When the ball returns to the boy, its speed is 17 m/s How much much work does air resistance do on the ball during its flight?

106. A mouse of mass 200 g falls 100 m down a vertical mine shaft and lands at the bottom with a speed of 8.0 m/s. During its fall, how much work is done on the mouse by air resistance?

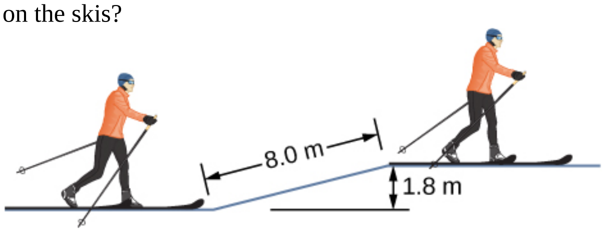
107. Using energy considerations and assuming negligible air resistance, show that a rock thrown from a bridge 20.0 m above water with an initial speed of 15.0 m/s strikes the water with a speed of 24.8 m/s independent of the direction thrown. (*Hint:* show that $K_i + U_i = K_f + U_f$)

108. A 1.0-kg ball at the end of a 2.0-m string swings in a vertical plane. At its lowest point the ball is moving with a speed of 10 m/s. (a) What is its speed at the top of its path? (b) What is the tension in the string when the ball is at the bottom and at the top of its path?

109. Ignoring details associated with friction, extra forces exerted by arm and leg muscles, and other factors, we can consider a pole vault as the conversion of an athlete's running kinetic energy to gravitational potential energy. If an athlete is to lift his body 4.8 m during a vault, what speed must he have when he plants his pole?

110. Tarzan grabs a vine hanging vertically from a tall tree when he is running at $9.0\ \text{m/s}$. (a) How high can he swing upward? (b) Does the length of the vine affect this height?

111. A $100-\text{kg}$ man is skiing across level ground at a speed of $8.0\ \text{m/s}$ when he comes to the small slope 1.8 m higher than ground level shown in the following figure. (a) If the skier coasts up the hill, what is his speed when he reaches the top plateau? Assume friction between the snow and skis is negligible. (b) What is his speed when he reaches the upper level if an $80-\text{N}$ frictional force acts on the skis?

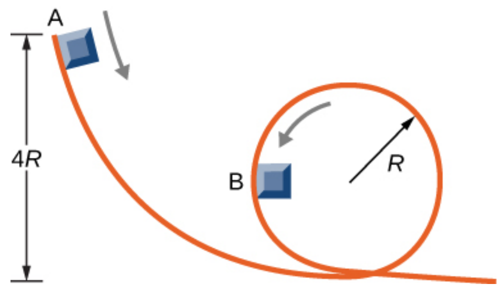


112. A sled of mass 70 kg starts from rest and slides down a $10°$ incline $80\ \text{m}$ long. It then travels for 20 m horizontally before starting back up an $8°$ incline. It travels 80 m along this incline before coming to rest. What is the magnitude of the net work done on the sled by friction?

113. A girl on a skateboard (total mass of 40 kg) is moving at a speed of 10 m/s at the bottom of a long ramp. The ramp is inclined at $20°$ with respect to the horizontal. If she travels 14.2 m upward along the ramp before stopping, what is the net frictional force on her?

114. A baseball of mass 0.25 kg is hit at home plate with a speed of 40 m/s. When it lands in a seat in the left-field bleachers a horizontal distance 120 m from home plate, it is moving at 30 m/s. If the ball lands 20 m above the spot where it was hit, how much work is done on it by air resistance?

115. A small block of mass *m* slides without friction around the loop-the-loop apparatus shown below. (a) If the block starts from rest at *A*, what is its speed at *B*? (b) What is the force of the track on the block at *B*?



116. A small ball is tied to a string and set rotating with negligible friction in a vertical circle. Prove that the tension in the string at the bottom of the circle exceeds that at the top of the circle by eight times the weight of the ball. Assume the ball's speed is zero as it sails over the top of the circle and there is no additional energy added to the ball during rotation.

117. Find the escape speed of a projectile from the surface of Mars.

118. Find the escape speed of a projectile from the surface of Jupiter.

119. What is the escape speed of a satellite located at the Moon's orbit about Earth? Assume the Moon is not nearby.

120. (a) Evaluate the gravitational potential energy between two 5.00-kg spherical steel balls separated by a center-to-center distance of 15.0 cm. (b) Assuming that they are both initially at rest relative to each other in deep space, use conservation of energy to find how fast will they be traveling upon impact. Each sphere has a radius of 5.10 cm.

121. An average-sized asteroid located $5.0 \times 10^7\ \text{km}$ from Earth with mass $2.0 \times 10^{13}\ \text{kg}$ is detected headed directly toward Earth with speed of 2.0 km/s. What will its speed be just before it hits our atmosphere? (You may ignore the size of the asteroid.)

122. (a) What will be the kinetic energy of the asteroid in the previous problem just before it hits Earth? b) Compare this energy to the output of the largest fission bomb, 2100 TJ. What impact would this have on Earth?

123. (a) What is the change in energy of a 1000-kg payload taken from rest at the surface of Earth and placed at rest on the surface of the Moon? (b) What would be the answer if the payload were taken from the Moon's surface to Earth? Is this a reasonable calculation of the energy needed to move a payload back and forth?

10.7 Sources of Energy

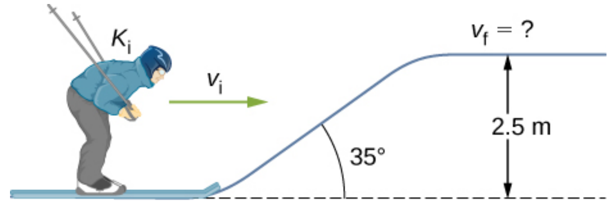
124. In the cartoon movie ***Pocahontas* (https://openstax.org/l/21pocahontclip)**, Pocahontas runs to the edge of a cliff and jumps off, showcasing the fun side of her personality. (a) If she is running at 3.0 m/s before jumping off the cliff and she hits the water at the bottom of the cliff at 20.0 m/s, how high is the cliff? Assume negligible air drag in this cartoon. (b) If she jumped off the same cliff from a standstill, how fast would she be falling right before she hit the water?

125. In the **Back to the Future movies (https://openstax.org/l/21bactofutclip)**, a DeLorean car of mass 1230 kg travels at 88 miles per hour to venture back to the future. (a) What is the kinetic energy of the DeLorean? (b) What spring constant would be needed to stop this DeLorean in a distance of 0.1m?

126. In a **"Top Fail" video (https://openstax.org/l/21topfailvideo)**, two women run at each other and collide by hitting exercise balls together. If each woman has a mass of 50 kg, which includes the exercise ball, and one woman runs to the right at 2.0 m/s and the other is running toward her at 1.0 m/s, (a) how much total kinetic energy is there in the system? (b) If energy is conserved after the collision and each exercise ball has a mass of 2.0 kg, how fast would the balls fly off toward the camera?

127. In an iconic movie scene, **Forrest Gump (https://openstax.org/l/21ForrGumpvid)** runs around the country. If he is running at a constant speed of 3 m/s, would it take him more or less energy to run uphill or downhill and why?

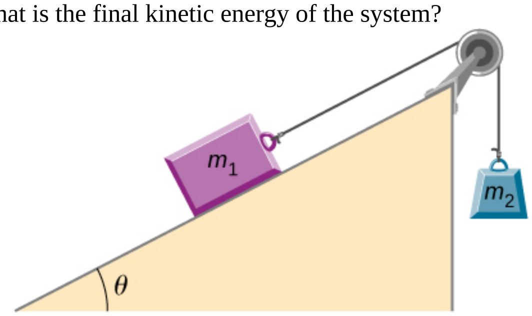
128. A 60.0-kg skier with an initial speed of 12.0 m/s coasts up a 2.50-m high rise as shown. Find her final speed at the top, given that the coefficient of friction between her skis and the snow is 0.80.



129. (a) How high a hill can a car coast up (engines disengaged) if work done by friction is negligible and its initial speed is 110 km/h? (b) If, in actuality, a 750-kg car with an initial speed of 110 km/h is observed to coast up a hill to a height 22.0 m above its starting point, how much thermal energy was generated by friction? (c) What is the average force of friction if the hill has a slope of $2.5°$ above the horizontal?

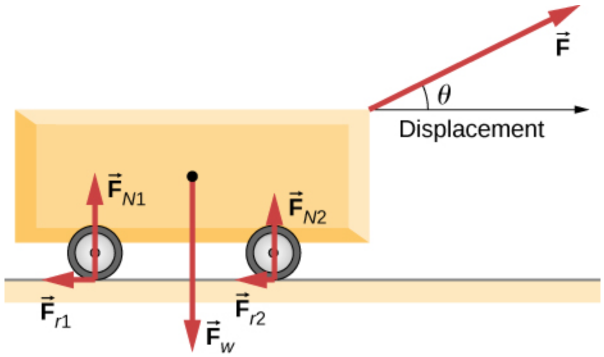
130. A T-shirt cannon launches a shirt at 5.00 m/s from a platform height of 3.00 m from ground level. How fast will the shirt be traveling if it is caught by someone whose hands are (a) 1.00 m from ground level? (b) 4.00 m from ground level? Neglect air drag.

131. Shown below is a box of mass m_1 that sits on a frictionless incline at an angle above the horizontal $\theta = 30°$. This box is connected by a relatively massless string, over a frictionless pulley, and finally connected to a box at rest over the ledge, labeled m_2. If m_1 and m_2 are a height h above the ground and $m_2 >> m_1$: (a) What is the initial gravitational potential energy of the system? (b) What is the final kinetic energy of the system?



ADDITIONAL PROBLEMS

132. A cart is pulled a distance D on a flat, horizontal surface by a constant force F that acts at an angle θ with the horizontal direction. The other forces on the object during this time are gravity (F_w), normal forces (F_{N1}) and (F_{N2}), and rolling frictions F_{r1} and F_{r2}, as shown below. What is the work done by each force?



133. Consider a particle on which several forces act, one of which is known to be constant in time: $\vec{\mathbf{F}}_1 = (3\text{ N})\hat{\mathbf{i}} + (4\text{ N})\hat{\mathbf{j}}$. As a result, the particle moves along the *x*-axis from $x = 0$ to $x = 5\text{ m}$ in some time interval. What is the work done by $\vec{\mathbf{F}}_1$?

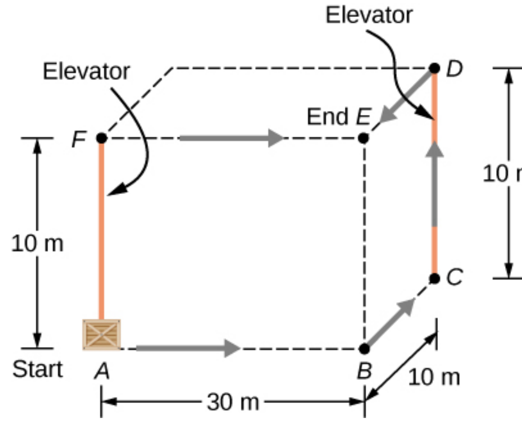
134. Consider a particle on which several forces act, one of which is known to be constant in time: $\vec{\mathbf{F}}_1 = (3\text{ N})\hat{\mathbf{i}} + (4\text{ N})\hat{\mathbf{j}}$. As a result, the particle moves first along the *x*-axis from $x = 0$ to $x = 5\text{ m}$ and then parallel to the *y*-axis from $y = 0$ to $y = 6\text{ m}$. What is the work done by $\vec{\mathbf{F}}_1$?

135. Consider a particle on which several forces act, one of which is known to be constant in time: $\vec{\mathbf{F}}_1 = (3\text{ N})\hat{\mathbf{i}} + (4\text{ N})\hat{\mathbf{j}}$. As a result, the particle moves along a straight path from a Cartesian coordinate of (0 m, 0 m) to (5 m, 6 m). What is the work done by $\vec{\mathbf{F}}_1$?

136. Consider a particle on which a force acts that depends on the position of the particle. This force is given by $\vec{\mathbf{F}}_1 = (2y)\hat{\mathbf{i}} + (3x)\hat{\mathbf{j}}$. Find the work done by this force when the particle moves from the origin to a point 5 meters to the right on the *x*-axis.

137. A boy pulls a 5-kg cart with a 20-N force at an angle of $30°$ above the horizontal for a length of time. Over this time frame, the cart moves a distance of 12 m on the horizontal floor. (a) Find the work done on the cart by the boy. (b) What will be the work done by the boy if he pulled with the same force horizontally instead of at an angle of $30°$ above the horizontal over the same distance?

138. A crate of mass 200 kg is to be brought from a site on the ground floor to a third floor apartment. The workers know that they can either use the elevator first, then slide it along the third floor to the apartment, or first slide the crate to another location marked C below, and then take the elevator to the third floor and slide it on the third floor a shorter distance. The trouble is that the third floor is very rough compared to the ground floor. Given that the coefficient of kinetic friction between the crate and the ground floor is 0.100 and between the crate and the third floor surface is 0.300, find the work needed by the workers for each path shown from *A* to *E*. Assume that the force the workers need to do is just enough to slide the crate at constant velocity (zero acceleration). *Note:* The work by the elevator against the force of gravity is not done by the workers.



139. A hockey puck of mass 0.17 kg is shot across a rough floor with the roughness different at different places, which can be described by a position-dependent coefficient of kinetic friction. For a puck moving along the *x*-axis, the coefficient of kinetic friction is the following function of *x*, where *x* is in m: $\mu(x) = 0.1 + 0.05x$. Find the work done by the kinetic frictional force on the hockey puck when it has moved (a) from $x = 0$ to $x = 2\text{ m}$, and (b) from $x = 2\text{ m}$ to $x = 4\text{ m}$.

140. A horizontal force of 20 N is required to keep a 5.0 kg box traveling at a constant speed up a frictionless incline for a vertical height change of 3.0 m. (a) What is the work done by gravity during this change in height? (b) What is the work done by the normal force? (c) What is the work done by the horizontal force?

141. A 7.0-kg box slides along a horizontal frictionless floor at 1.7 m/s and collides with a relatively massless spring that compresses 23 cm before the box comes to a stop. (a) How much kinetic energy does the box have before it collides with the spring? (b) Calculate the work done by the spring. (c) Determine the spring constant of the spring.

142. You are driving your car on a straight road with a coefficient of friction between the tires and the road of 0.55. A large piece of debris falls in front of your view and you immediate slam on the brakes, leaving a skid mark of 30.5 m (100-feet) long before coming to a stop. A policeman sees your car stopped on the road, looks at the skid mark, and gives you a ticket for traveling over the 13.4 m/s (30 mph) speed limit. Should you fight the speeding ticket in court?

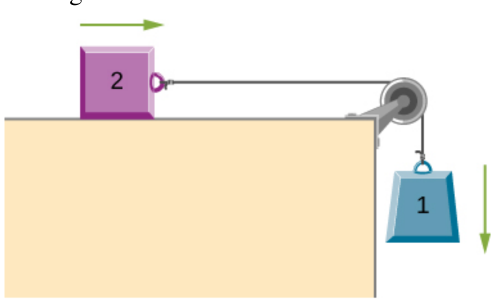
143. A crate is being pushed across a rough floor surface. If no force is applied on the crate, the crate will slow down and come to a stop. If the crate of mass 50 kg moving at speed 8 m/s comes to rest in 10 seconds, what is the rate at which the frictional force on the crate takes energy away from the crate?

144. Suppose a horizontal force of 20 N is required to maintain a speed of 8 m/s of a 50 kg crate. (a) What is the power of this force? (b) Note that the acceleration of the crate is zero despite the fact that 20 N force acts on the crate horizontally. What happens to the energy given to the crate as a result of the work done by this 20 N force?

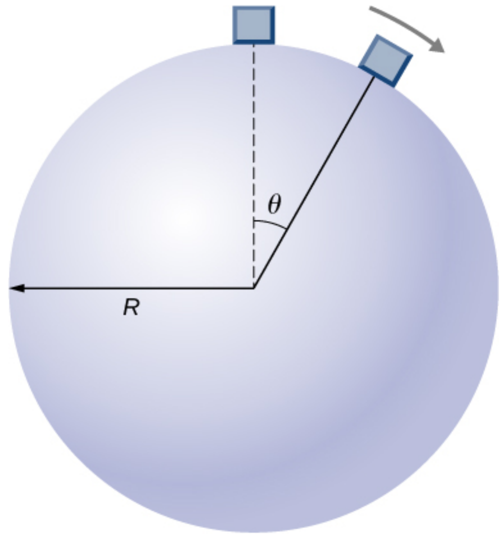
145. Grains from a hopper falls at a rate of 10 kg/s vertically onto a conveyor belt that is moving horizontally at a constant speed of 2 m/s. (a) What force is needed to keep the conveyor belt moving at the constant velocity? (b) What is the minimum power of the motor driving the conveyor belt?

146. A cyclist in a race must climb a $5°$ hill at a speed of 8 m/s. If the mass of the bike and the biker together is 80 kg, what must be the power output of the biker to achieve the goal?

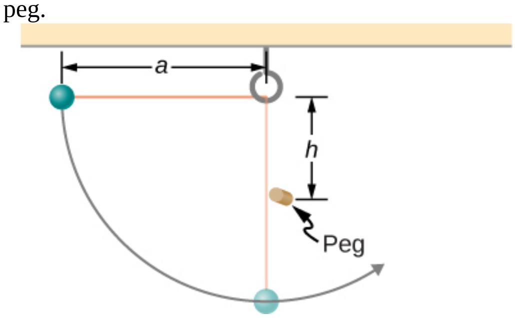
147. Block 2 shown below slides along a frictionless table as block 1 falls. Both blocks are attached by a frictionless pulley. Find the speed of the blocks after they have each moved 2.0 m. Assume that they start at rest and that the pulley has negligible mass. Use $m_1 = 2.0\,\text{kg}$ and $m_2 = 4.0\,\text{kg}$.



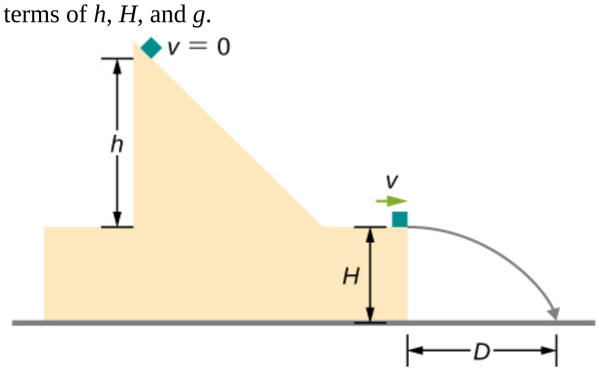
148. A body of mass *m* and negligible size starts from rest and slides down the surface of a frictionless solid sphere of radius *R*. (See below.) Prove that the body leaves the sphere when $\theta = \cos^{-1}(2/3)$.



149. Shown below is a small ball of mass *m* attached to a string of length *a*. A small peg is located a distance *h* below the point where the string is supported. If the ball is released when the string is horizontal, show that *h* must be greater than 3*a*/5 if the ball is to swing completely around the peg.



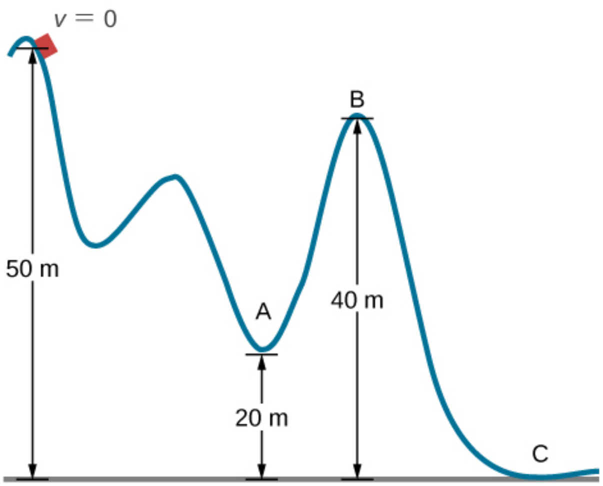
150. A block leaves a frictionless inclined surface horizontally after dropping off by a height *h*. Find the horizontal distance *D* where it will land on the floor, in terms of *h*, *H*, and *g*.



151. A skier starts from rest and slides downhill. What will be the speed of the skier if he drops by 20 meters in vertical height? Ignore any air resistance (which will, in reality, be quite a lot), and any friction between the skis and the snow.

152. Repeat the preceding problem, but this time, suppose that the work done by air resistance cannot be ignored. Let the work done by the air resistance when the skier goes from *A* to *B* along the given hilly path be −2000 J. The work done by air resistance is negative since the air resistance acts in the opposite direction to the displacement. Supposing the mass of the skier is 50 kg, what is the speed of the skier at point *B*?

153. In an amusement park, a car rolls in a track as shown below. Find the speed of the car at *A*, *B*, and *C*. Note that the work done by the rolling friction is zero since the displacement of the point at which the rolling friction acts on the tires is momentarily at rest and therefore has a zero displacement.



154. A 200-g steel ball is tied to a 2.00-m "massless" string and hung from the ceiling to make a pendulum, and then, the ball is brought to a position making a $30°$ angle with the vertical direction and released from rest. Ignoring the effects of the air resistance, find the speed of the ball when the string (a) is vertically down, (b) makes an angle of $20°$ with the vertical and (c) makes an angle of $10°$ with the vertical.

155. A 300 g hockey puck is shot across an ice-covered pond. Before the hockey puck was hit, the puck was at rest. After the hit, the puck has a speed of 40 m/s. The puck comes to rest after going a distance of 30 m. (a) Describe how the energy of the puck changes over time, giving the numerical values of any work or energy involved. (b) Find the magnitude of the net friction force.

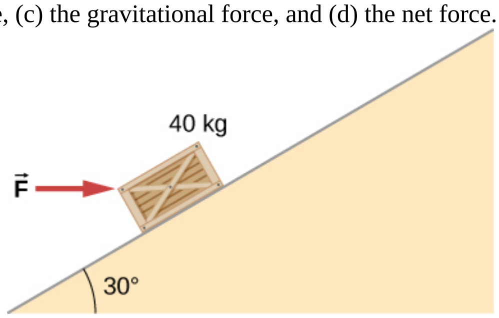
156. A projectile of mass 2 kg is fired with a speed of 20 m/s at an angle of $30°$ with respect to the horizontal. (a) Calculate the initial total energy of the projectile given that the reference point of zero gravitational potential energy at the launch position. (b) Calculate the kinetic energy at the highest vertical position of the projectile. (c) Calculate the gravitational potential energy at the highest vertical position. (d) Calculate the maximum height that the projectile reaches. Compare this result by solving the same problem using your knowledge of projectile motion.

157. An artillery shell is fired at a target 200 m above the ground. When the shell is 100 m in the air, it has a speed of 100 m/s. What is its speed when it hits its target? Neglect air friction.

158. How much energy is lost to a dissipative drag force if a 60-kg person falls at a constant speed for 15 meters?

CHALLENGE PROBLEMS

159. Shown below is a 40-kg crate that is pushed at constant velocity a distance 8.0 m along a $30°$ incline by the horizontal force $\vec{\mathbf{F}}$. The coefficient of kinetic friction between the crate and the incline is $\mu_k = 0.40$. Calculate the work done by (a) the applied force, (b) the frictional force, (c) the gravitational force, and (d) the net force.



160. The surface of the preceding problem is modified so that the coefficient of kinetic friction is decreased. The same horizontal force is applied to the crate, and after being pushed 8.0 m, its speed is 5.0 m/s. How much work is now done by the force of friction? Assume that the crate starts at rest.

161. Constant power *P* is delivered to a car of mass *m* by its engine. Show that if air resistance can be ignored, the distance covered in a time *t* by the car, starting from rest, is given by $s = (8P/9m)^{1/2} t^{3/2}$.

162. Suppose that the air resistance a car encounters is independent of its speed. When the car travels at 15 m/s, its engine delivers 20 hp to its wheels. (a) What is the power delivered to the wheels when the car travels at 30 m/s? (b) How much energy does the car use in covering 10 km at 15 m/s? At 30 m/s? Assume that the engine is 25% efficient. (c) Answer the same questions if the force of air resistance is proportional to the speed of the automobile. (d) What do these results, plus your experience with gasoline consumption, tell you about air resistance?

11 | LINEAR MOMENTUM

Figure 11.1 The concepts of impulse, momentum, and center of mass are crucial for a major-league baseball player to successfully get a hit. If he misjudges these quantities, he might break his bat instead. (credit: modification of work by “Cathy T”/Flickr)

Chapter Outline

Introduction

The concepts of work, energy, and the work-energy theorem are valuable for two primary reasons: First, they are powerful computational tools, making it much easier to analyze complex physical systems than is possible using Newton’s laws directly (for example, systems with nonconstant forces); and second, the observation that the total energy of a closed system is conserved means that the system can only evolve in ways that are consistent with energy conservation. In other words, a system cannot evolve randomly; it can only change in ways that conserve energy.

In this chapter, we develop and define another conserved quantity, called *linear momentum*, and another relationship (the *impulse-momentum theorem*), which will put an additional constraint on how a system evolves in time. Conservation of momentum is useful for understanding collisions, such as that shown in the above image. It is just as powerful, just as important, and just as useful as conservation of energy and the work-energy theorem.

11.1 | Linear Momentum

Learning Objectives

By the end of this section, you will be able to:

- Explain what momentum is, physically
- Calculate the momentum of a moving object

Our study of kinetic energy showed that a complete understanding of an object's motion must include both its mass and its velocity ($K = (1/2)mv^2$). However, as powerful as this concept is, it does not include any information about the direction of the moving object's velocity vector. We'll now define a physical quantity that includes direction.

Like kinetic energy, this quantity includes both mass and velocity; like kinetic energy, it is a way of characterizing the "quantity of motion" of an object. It is given the name **momentum** (from the Latin word *movimentum*, meaning "movement"), and it is represented by the symbol *p*.

Momentum

The momentum *p* of an object is the product of its mass and its velocity:

$$\vec{\mathbf{p}} = m\vec{\mathbf{v}}. \quad (11.1)$$

Figure 11.2 The velocity and momentum vectors for the ball are in the same direction. The mass of the ball is about 0.5 kg, so the momentum vector is about half the length of the velocity vector because momentum is velocity time mass. (credit: modification of work by Ben Sutherland)

As shown in **Figure 11.2**, momentum is a vector quantity (since velocity is). The direction of the momentum vector is the same as the direction of the associated velocity vector. In one dimensional motion, this vector nature may simply be denoted by a sign (postive or negative) indicating the direction of the object's velocity. This is one of the things that makes momentum useful and not a duplication of kinetic energy. It is perhaps most useful when determining whether an object’s motion is difficult to change (**Figure 11.3**) or easy to change (**Figure 11.4**).

Figure 11.3 This supertanker transports a huge mass of oil; as a consequence, it takes a long time for a force to change its (comparatively small) velocity. (credit: modification of work by “the_tahoe_guy”/Flickr)

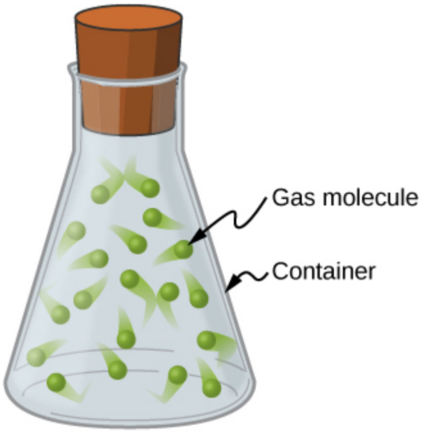


Figure 11.4 Gas molecules can have very large velocities, but these velocities change nearly instantaneously when they collide with the container walls or with each other. This is primarily because their masses are so tiny.

Unlike kinetic energy, momentum depends equally on an object’s mass and velocity. For example, as you will learn when you study thermodynamics, the average speed of an air molecule at room temperature is approximately 500 m/s, with an average molecular mass of 6×10^{-25} kg ; its momentum is thus

$$p_{\text{molecule}} = \left(6 \times 10^{-25}\ \text{kg}\right)\left(500\ \frac{\text{m}}{\text{s}}\right) = 3 \times 10^{-22}\ \frac{\text{kg} \cdot \text{m}}{\text{s}}.$$

For comparison, a typical automobile might have a speed of only 15 m/s, but a mass of 1400 kg, giving it a momentum of

$$p_{\text{car}} = (1400\ \text{kg})\left(15\ \frac{\text{m}}{\text{s}}\right) = 21{,}000\ \frac{\text{kg} \cdot \text{m}}{\text{s}}.$$

These momenta are different by 27 orders of magnitude, or a factor of a billion billion billion!

11.2 | Impulse and Collisions

Learning Objectives

By the end of this section, you will be able to:

- Explain what an impulse is, physically
- Describe what an impulse does
- Relate impulses to collisions
- Apply the impulse-momentum theorem to solve problems

We have defined momentum to be the product of mass and velocity. Therefore, if an object's velocity should change (due to the application of a force on the object), then necessarily, its momentum changes as well. This indicates a connection between momentum and force. The purpose of this section is to explore and describe that connection.

Suppose you apply a force on a free object for some amount of time. Clearly, the larger the force, the larger the object's change of momentum will be. Alternatively, the more time you spend applying this force, again the larger the change of momentum will be, as depicted in **Figure 11.5**. The amount by which the object's motion changes is therefore proportional to the magnitude of the force, and also to the time interval over which the force is applied.

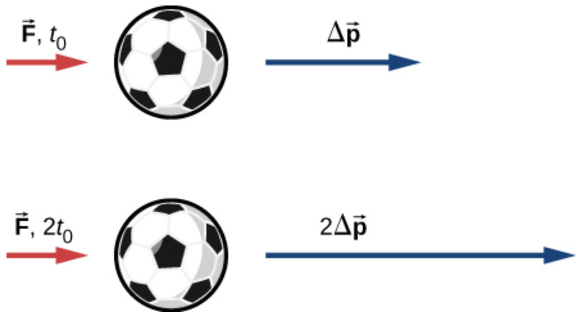


Figure 11.5 The change in momentum of an object is proportional to the length of time during which the force is applied. If a force is exerted on the lower ball for twice as long as on the upper ball, then the change in the momentum of the lower ball is twice that of the upper ball.

Mathematically, if a quantity is proportional to two (or more) things, then it is proportional to the product of those things. The product of a force and a time interval (over which that force acts) is called **impulse**, and is given the symbol $\vec{J}$.

Impulse

Let $\vec{F}(t)$ be the force applied to an object over some differential time interval dt (**Figure 11.6**). The resulting impulse on the object is defined as

$$d\vec{J} \equiv \vec{F}(t)dt. \tag{11.2}$$

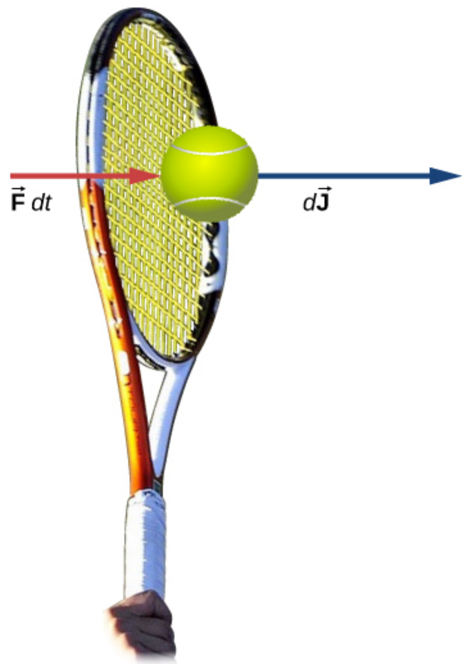


Figure 11.6 A force applied by a tennis racquet to a tennis ball over a time interval generates an impulse acting on the ball.

Now, if the applied force F remains constant throughout a finite time interval Δt, then the total impulse

$$\vec{\mathbf{J}} = \vec{F}\Delta t \tag{11.3}$$

Even if the impulsive force isn't completely constant in time, such forces usually act for very short time intervals, so it is reasonable to use the relation

$$\vec{\mathbf{J}} = \vec{\mathbf{F}}_{\text{ave}}\Delta t. \tag{11.4}$$

The idea here is that you can calculate the impulse on the object even if you don't know the details of the force as a function of time; you only need the average force. In fact, though, the process is usually reversed: You determine the impulse (by measurement or calculation) and then calculate the average force that caused that impulse.

To calculate the impulse, a useful result follows from writing the force in **Equation 11.3** as $\vec{\mathbf{F}}(t) = m\vec{\mathbf{a}}(t)$:

$$\vec{\mathbf{J}} = \vec{F}\Delta t = m\vec{\mathbf{a}}\Delta t = m\left[\vec{\mathbf{v}}(t_f) - \vec{\mathbf{v}}_i\right].$$

For a constant force $\vec{\mathbf{F}}_{\text{ave}} = \vec{\mathbf{F}} = m\vec{\mathbf{a}}$, this simplifies to

$$\vec{\mathbf{J}} = m\vec{\mathbf{a}}\Delta t = m\vec{\mathbf{v}}_f - m\vec{\mathbf{v}}_i = m(\vec{\mathbf{v}}_f - \vec{\mathbf{v}}_i).$$

That is,

$$\vec{\mathbf{J}} = m\Delta\vec{\mathbf{v}}. \tag{11.5}$$

Note that the integral form, **Equation 11.3**, applies to constant forces as well; in that case, since the force is independent of time, it comes out of the integral, which can then be trivially evaluated.

Example 11.1

The Arizona Meteor Crater

Approximately 50,000 years ago, a large (radius of 25 m) iron-nickel meteorite collided with Earth at an estimated speed of 1.28×10^4 m/s in what is now the northern Arizona desert, in the United States. The impact produced a crater that is still visible today (Figure 11.7); it is approximately 1200 m (three-quarters of a mile) in diameter, 170 m deep, and has a rim that rises 45 m above the surrounding desert plain. Iron-nickel meteorites typically have a density of $\rho = 7970\,\text{kg/m}^3$. Use impulse considerations to estimate the average force that the meteor applied to Earth during the impact.

Figure 11.7 The Arizona Meteor Crater in Flagstaff, Arizona (often referred to as the Barringer Crater after the person who first suggested its origin and whose family owns the land). (credit: modification of work by "Shane.torgerson"/Wikimedia Commons)

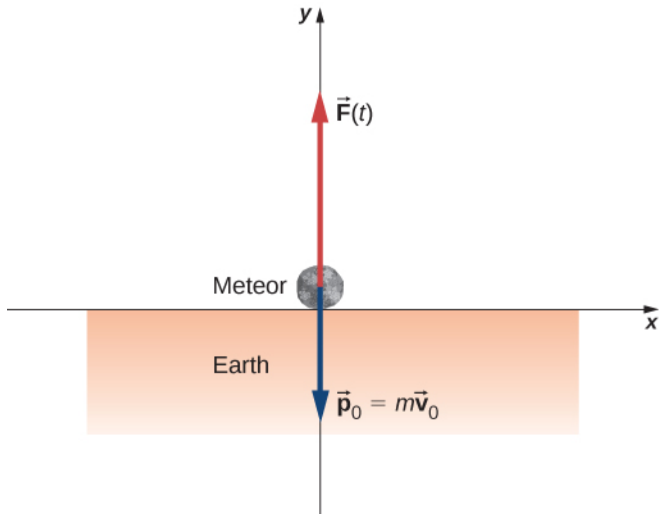
Strategy

It is conceptually easier to reverse the question and calculate the force that Earth applied on the meteor in order to stop it. Therefore, we'll calculate the force on the meteor and then use Newton's third law to argue that the force from the meteor on Earth was equal in magnitude and opposite in direction.

Using the given data about the meteor, and making reasonable guesses about the shape of the meteor and impact time, we first calculate the impulse using Equation 11.5. We then use the relationship between force and impulse Equation 11.4 to estimate the average force during impact. Next, we choose a reasonable force function for the impact event, calculate the average value of that function ???, and set the resulting expression equal to the calculated average force. This enables us to solve for the maximum force.

Solution

Define upward to be the +*y*-direction. For simplicity, assume the meteor is traveling vertically downward prior to impact. In that case, its initial velocity is $\vec{\mathbf{v}}_i = -v_i \hat{\mathbf{j}}$, and the force Earth exerts on the meteor points upward, $\vec{\mathbf{F}}(t) = +F(t)\hat{\mathbf{j}}$. The situation at $t = 0$ is depicted below.



The average force during the impact is related to the impulse by

$$\vec{\mathbf{F}}_{\text{ave}} = \frac{\vec{\mathbf{J}}}{\Delta t}.$$

From **Equation 11.5**, $\vec{\mathbf{J}} = m\Delta\vec{\mathbf{v}}$, so we have

$$\vec{\mathbf{F}}_{\text{ave}} = \frac{m\Delta\vec{\mathbf{v}}}{\Delta t}.$$

The mass is equal to the product of the meteor's density and its volume:

$$m = \rho V.$$

If we assume (guess) that the meteor was roughly spherical, we have

$$V = \frac{4}{3}\pi R^3.$$

Thus we obtain

$$\vec{\mathbf{F}}_{\text{ave}} = \frac{\rho V\Delta\vec{\mathbf{v}}}{\Delta t} = \frac{\rho\left(\frac{4}{3}\pi R^3\right)\left(\vec{\mathbf{v}}_{\text{f}} - \vec{\mathbf{v}}_{\text{i}}\right)}{\Delta t}.$$

The problem says the velocity at impact was $-1.28 \times 10^4\ \text{m/s}\,\hat{\mathbf{j}}$ (the final velocity is zero); also, we guess that the primary impact lasted about $t_{\text{max}} = 2\ \text{s}$. Substituting these values gives

$$\begin{aligned}\vec{\mathbf{F}}_{\text{ave}} &= \frac{\left(7970\ \frac{\text{kg}}{\text{m}^3}\right)\left[\frac{4}{3}\ \pi\ (25\ \text{m})^3\right]\left[0\ \frac{\text{m}}{\text{s}} - \left(-1.28 \times 10^4\ \frac{\text{m}}{\text{s}}\hat{\mathbf{j}}\right)\right]}{2\ \text{s}}.\\ &= +\left(3.33 \times 10^{12}\ \text{N}\right)\hat{\mathbf{j}}\end{aligned}$$

This is the average force applied during the collision. Notice that this force vector points in the same direction as the change of velocity vector $\Delta\vec{\mathbf{v}}$.

Example 11.2

The Benefits of Impulse

A car traveling at 27 m/s collides with a building. The collision with the building causes the car to come to a stop in approximately 1 second. The driver, who weighs 860 N, is protected by a combination of a variable-tension seatbelt and an airbag (Figure 11.8). (In effect, the driver collides with the seatbelt and airbag and *not* with the building.) The airbag and seatbelt slow his velocity, such that he comes to a stop in approximately 2.5 s.

a. What average force does the driver experience during the collision?

b. Without the seatbelt and airbag, his collision time (with the steering wheel) would have been approximately 0.20 s. What force would he experience in this case?

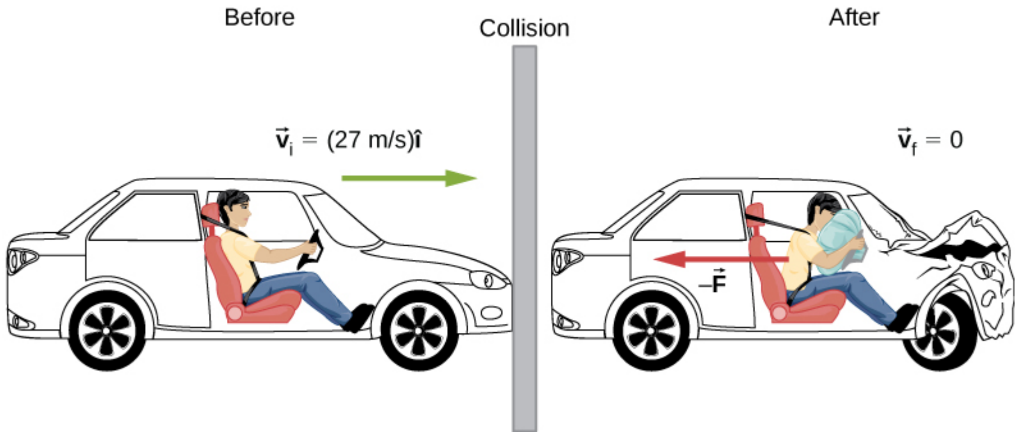


Figure 11.8 The motion of a car and its driver at the instant before and the instant after colliding with the wall. The restrained driver experiences a large backward force from the seatbelt and airbag, which causes his velocity to decrease to zero. (The forward force from the seatback is much smaller than the backward force, so we neglect it in the solution.)

Strategy

We are given the driver's weight, his initial and final velocities, and the time of collision; we are asked to calculate a force. Impulse seems the right way to tackle this; we can combine Equation 11.4 and Equation 11.5.

Solution

a. Define the +*x*-direction to be the direction the car is initially moving. We know

$$\vec{J} = \vec{F}\,\Delta t$$

and

$$\vec{J} = m\Delta\vec{v}.$$

Since *J* is equal to both those things, they must be equal to each other:

$$\vec{F}\,\Delta t = m\Delta\vec{v}.$$

We need to convert this weight to the equivalent mass, expressed in SI units:

$$\frac{860\text{ N}}{9.8\text{ m/s}^2} = 87.8\text{ kg}.$$

Remembering that $\Delta\vec{v} = \vec{v}_f - \vec{v}_i$, and noting that the final velocity is zero, we solve for the force:

$$\vec{F} = m\frac{0 - v_i\,\hat{i}}{\Delta t} = (87.8\text{ kg})\left(\frac{-(27\text{ m/s})\,\hat{i}}{2.5\text{ s}}\right) = -(948\text{ N})\,\hat{i}.$$

The negative sign implies that the force slows him down. For perspective, this is about 1.1 times his own weight.

b. Same calculation, just the different time interval:

$$\vec{\mathbf{F}} = (87.8\ \text{kg})\left(\frac{-(27\ \text{m/s})\ \hat{\mathbf{i}}}{0.20\ \text{s}}\right) = -(11{,}853\ \text{N})\ \hat{\mathbf{i}}$$

which is about 14 times his own weight. Big difference!

Significance

You see that the value of an airbag is how greatly it reduces the force on the vehicle occupants. For this reason, they have been required on all passenger vehicles in the United States since 1991, and have been commonplace throughout Europe and Asia since the mid-1990s. The change of momentum in a crash is the same, with or without an airbag; the force, however, is vastly different.

Effect of Impulse

Since an impulse is a force acting for some amount of time, it causes an object's motion to change. Recall **Equation 11.5**:

$$\vec{\mathbf{J}} = m\Delta \vec{\mathbf{v}}.$$

Because $m\vec{\mathbf{v}}$ is the momentum of a system, $m\Delta\vec{\mathbf{v}}$ is the *change* of momentum $\Delta\vec{\mathbf{p}}$. This gives us the following relation, called the **impulse-momentum theorem** (or relation).

Impulse-Momentum Theorem

An impulse applied to a system changes the system's momentum, and that change of momentum is exactly equal to the impulse that was applied:

$$\vec{\mathbf{J}} = \Delta\vec{\mathbf{p}}. \quad (11.6)$$

The impulse-momentum theorem is depicted graphically in **Figure 11.9**.

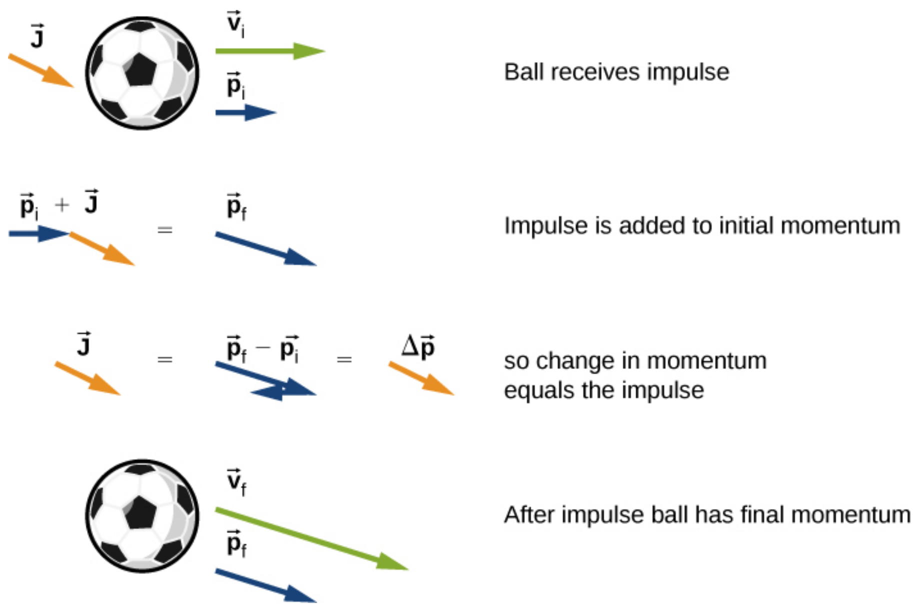


Figure 11.9 Illustration of impulse-momentum theorem. (a) A ball with initial velocity $\vec{\mathbf{v}}_0$ and momentum $\vec{\mathbf{p}}_0$ receives an impulse $\vec{\mathbf{J}}$. (b) This impulse is added vectorially to the initial momentum. (c) Thus, the impulse equals the change in momentum, $\vec{\mathbf{J}} = \Delta\vec{\mathbf{p}}$. (d) After the impulse, the ball moves off with its new momentum $\vec{\mathbf{p}}_f$.

There are two crucial concepts in the impulse-momentum theorem:

1. Impulse is a vector quantity; an impulse of, say, $-(10\,\text{N}\cdot\text{s})\,\hat{\mathbf{i}}$ is very different from an impulse of $+(10\,\text{N}\cdot\text{s})\,\hat{\mathbf{i}}$; they cause completely opposite changes of momentum.
2. An impulse does not cause momentum; rather, it causes a *change* in the momentum of an object. Thus, you must subtract the final momentum from the initial momentum, and—since momentum is also a vector quantity—you must take careful account of the signs of the momentum vectors.

The most common questions asked in relation to impulse are to calculate the applied force, or the change of velocity that occurs as a result of applying an impulse. The general approach is the same.

Problem-Solving Strategy: Impulse-Momentum Theorem

1. Express the impulse as force times the relevant time interval.
2. Express the impulse as the change of momentum, usually $m\Delta v$.
3. Equate these and solve for the desired quantity.

Example 11.3

Moving the *Enterprise*

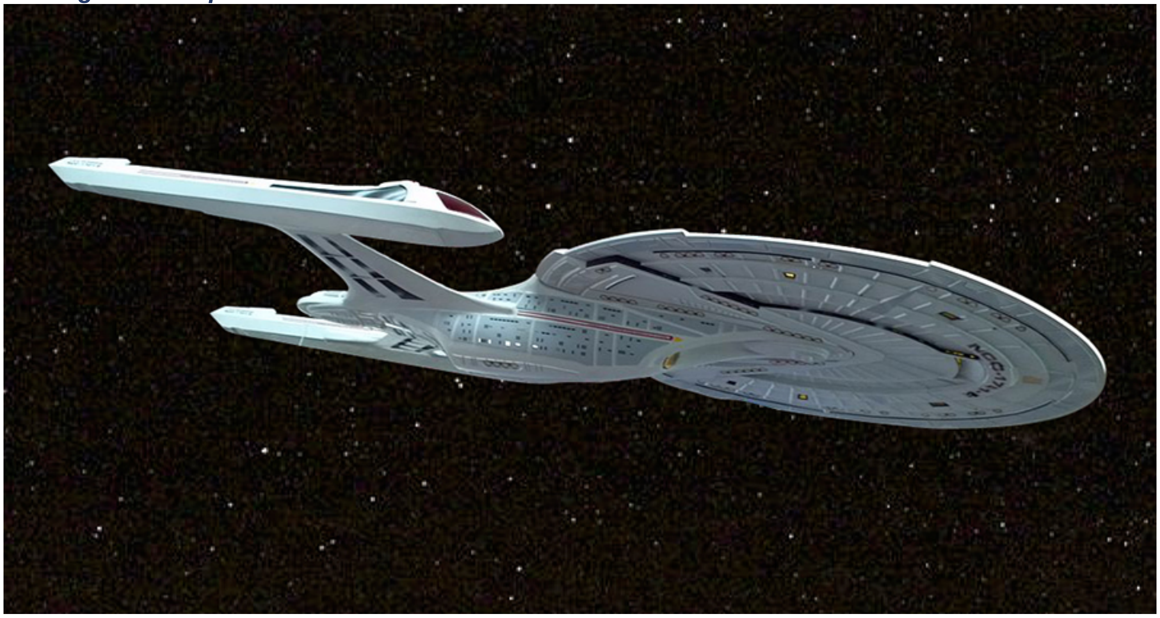


Figure 11.10 The fictional starship *Enterprise* from the Star Trek adventures operated on so-called "impulse engines" that combined matter with antimatter to produce energy.

"Mister Sulu, take us out; ahead one-quarter impulse." With this command, Captain Kirk of the starship *Enterprise* (**Figure 11.10**) has his ship start from rest to a final speed of $v_{\text{f}} = 1/4\left(3.0 \times 10^8 \text{ m/s}\right)$. Assuming this maneuver is completed in 60 s, what average force did the impulse engines apply to the ship?

Strategy

We are asked for a force; we know the initial and final speeds (and hence the change in speed), and we know the time interval over which this all happened. In particular, we know the amount of time that the force acted. This suggests using the impulse-momentum relation. To use that, though, we need the mass of the *Enterprise*. An internet search gives a best estimate of the mass of the *Enterprise* (in the 2009 movie) as $2 \times 10^9 \text{ kg}$.

Solution

Because this problem involves only one direction (i.e., the direction of the force applied by the engines), we only need the scalar form of the impulse-momentum theorem **Equation 11.6**, which is

$$\Delta p = J$$

with

$$\Delta p = m\Delta v$$

and

$$J = F\Delta t.$$

Equating these expressions gives

$$F\Delta t = m\Delta v.$$

Solving for the magnitude of the force and inserting the given values leads to

$$F = \frac{m\Delta v}{\Delta t} = \frac{\left(2 \times 10^9 \text{ kg}\right)\left(7.5 \times 10^7 \text{ m/s}\right)}{60 \text{ s}} = 2.5 \times 10^{15} \text{ N}.$$

Significance

This is an unimaginably huge force. It goes almost without saying that such a force would kill everyone on board instantly, as well as destroying every piece of equipment. Fortunately, the *Enterprise* has "inertial dampeners." It is left as an exercise for the reader's imagination to determine how these work.

11.1 Check Your Understanding The U.S. Air Force uses "10*gs*" (an acceleration equal to $10 \times 9.8\ \text{m/s}^2$) as the maximum acceleration a human can withstand (but only for several seconds) and survive. How much time must the *Enterprise* spend accelerating if the humans on board are to experience an average of at most 10*gs* of acceleration? (Assume the inertial dampeners are offline.)

Example 11.4

The iPhone Drop

Apple released its iPhone 6 Plus in November 2014. According to many reports, it was originally supposed to have a screen made from sapphire, but that was changed at the last minute for a hardened glass screen. Reportedly, this was because the sapphire screen cracked when the phone was dropped. What force did the iPhone 6 Plus experience as a result of being dropped?

Strategy

The force the phone experiences is due to the impulse applied to it by the floor when the phone collides with the floor. Our strategy then is to use the impulse-momentum relationship. We calculate the impulse, estimate the impact time, and use this to calculate the force.

We need to make a couple of reasonable estimates, as well as find technical data on the phone itself. First, let's suppose that the phone is most often dropped from about chest height on an average-height person. Second, assume that it is dropped from rest, that is, with an initial vertical velocity of zero. Finally, we assume that the phone bounces very little—the height of its bounce is assumed to be negligible.

Solution

Define upward to be the +*y*-direction. A typical height is approximately $h = 1.5\ \text{m}$ and, as stated, $\vec{\mathbf{v}}_\text{i} = (0\ \text{m/s})\hat{\mathbf{i}}$. The average force on the phone is related to the impulse the floor applies on it during the collision:

$$\vec{\mathbf{F}}_\text{ave} = \frac{\vec{\mathbf{J}}}{\Delta t}.$$

The impulse $\vec{\mathbf{J}}$ equals the change in momentum,

$$\vec{\mathbf{J}} = \Delta \vec{\mathbf{p}}$$

so

$$\vec{\mathbf{F}}_\text{ave} = \frac{\Delta \vec{\mathbf{p}}}{\Delta t}.$$

Next, the change of momentum is

$$\Delta \vec{\mathbf{p}} = m\Delta \vec{\mathbf{v}}.$$

We need to be careful with the velocities here; this is the change of velocity due to the collision with the floor. But the phone also has an initial drop velocity [$\vec{\mathbf{v}}_\text{i} = (0\ \text{m/s})\hat{\mathbf{j}}$], so we label our velocities. Let:

- $\vec{\mathbf{v}}_\text{i} =$ the initial velocity with which the phone was dropped (zero, in this example)

- $\vec{v}_1 =$ the velocity the phone had the instant just before it hit the floor
- $\vec{v}_2 =$ the final velocity of the phone as a result of hitting the floor

Figure 11.11 shows the velocities at each of these points in the phone's trajectory.

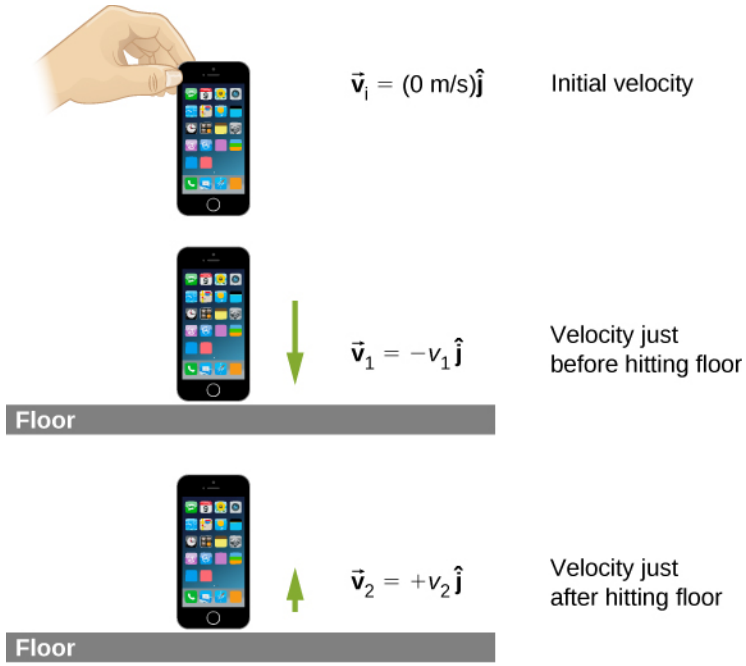


Figure 11.11 (a) The initial velocity of the phone is zero, just after the person drops it. (b) Just before the phone hits the floor, its velocity is $\vec{v}_1$, which is unknown at the moment, except for its direction, which is downward $(-\hat{j})$. (c) After bouncing off the floor, the phone has a velocity $\vec{v}_2$, which is also unknown, except for its direction, which is upward $(+\hat{j})$.

With these definitions, the change of momentum of the phone during the collision with the floor is

$$m\Delta\vec{v} = m(\vec{v}_2 - \vec{v}_1).$$

Since we assume the phone doesn't bounce at all when it hits the floor (or at least, the bounce height is negligible), then $\vec{v}_2$ is zero, so

$$\begin{aligned} m\Delta\vec{v} &= m\left[0 - \left(-v_1\hat{j}\right)\right] \\ m\Delta\vec{v} &= +mv_1\hat{j}. \end{aligned}$$

We can get the speed of the phone just before it hits the floor using either kinematics or conservation of energy. We'll use conservation of energy here; you should re-do this part of the problem using kinematics and prove that you get the same answer.

First, define the zero of potential energy to be located at the floor. Conservation of energy then gives us:

$$\begin{aligned} E_\text{i} &= E_1 \\ K_\text{i} + U_\text{i} &= K_1 + U_1 \\ \frac{1}{2}mv_\text{i}^2 + mgh_\text{drop} &= \frac{1}{2}mv_1^2 + mgh_\text{floor}. \end{aligned}$$

Defining $h_\text{floor} = 0$ and using $\vec{\mathbf{v}}_\text{i} = (0 \text{ m/s})\hat{\mathbf{j}}$ gives

$$\begin{aligned} \frac{1}{2}mv_1^2 &= mgh_\text{drop} \\ v_1 &= \pm\sqrt{2gh_\text{drop}}. \end{aligned}$$

Because v_1 is a vector magnitude, it must be positive. Thus, $m\Delta v = mv_1 = m\sqrt{2gh_\text{drop}}$. Inserting this result into the expression for force gives

$$\begin{aligned} \vec{\mathbf{F}} &= \frac{\Delta \vec{\mathbf{p}}}{\Delta t} \\ &= \frac{m\Delta \vec{\mathbf{v}}}{\Delta t} \\ &= \frac{+mv_1 \hat{\mathbf{j}}}{\Delta t} \\ &= \frac{m\sqrt{2gh}}{\Delta t}\hat{\mathbf{j}}. \end{aligned}$$

Finally, we need to estimate the collision time. One common way to estimate a collision time is to calculate how long the object would take to travel its own length. The phone is moving at 5.4 m/s just before it hits the floor, and it is 0.14 m long, giving an estimated collision time of 0.026 s. Inserting the given numbers, we obtain

$$\vec{\mathbf{F}} = \frac{(0.172 \text{ kg})\sqrt{2(9.8 \text{ m/s}^2)(1.5 \text{ m})}}{0.026 \text{ s}}\hat{\mathbf{j}} = (36 \text{ N})\hat{\mathbf{j}}.$$

Significance

The iPhone itself weighs just $(0.172 \text{ kg})(9.81 \text{ m/s}^2) = 1.68 \text{ N}$; the force the floor applies to it is therefore over 20 times its weight.

11.2 Check Your Understanding What if we had assumed the phone *did* bounce on impact? Would this have increased the force on the iPhone, decreased it, or made no difference?

Momentum and Force

In **Example 11.3**, we obtained an important relationship:

$$\vec{\mathbf{F}}_\text{ave} = \frac{\Delta \vec{\mathbf{p}}}{\Delta t}. \tag{11.7}$$

In words, the average force applied to an object is equal to the change of the momentum that the force causes, divided by the time interval over which this change of momentum occurs. This relationship is very useful in situations where the collision time Δt is small, but measureable; typical values would be 1/10th of a second, or even one thousandth of a second. Car crashes, punting a football, or collisions of subatomic particles would meet this criterion.

This says that the rate of change of the system's momentum (implying that momentum is a function of time) is exactly equal to the net applied force (also, in general, a function of time). This is, in fact, Newton's second law, written in terms of momentum rather than acceleration. This is the relationship Newton himself presented in his *Principia Mathematica* (although he called it "quantity of motion" rather than "momentum").

If the mass of the system remains constant, Equation 11.3 reduces to the more familiar form of Newton's second law. We can see this by substituting the definition of momentum:

$$\vec{\mathbf{F}} = \frac{\Delta(m\vec{\mathbf{v}})}{\Delta t} = m\frac{\Delta\vec{\mathbf{v}}}{\Delta t} = m\vec{\mathbf{a}}.$$

The assumption of constant mass allowed us to pull m out of the derivative. If the mass is not constant, we cannot use this form of the second law, but instead must start from Equation 11.3. Thus, one advantage to expressing force in terms of changing momentum is that it allows for the mass of the system to change, as well as the velocity; this is a concept we'll explore when we study the motion of rockets.

Newton's Second Law of Motion in Terms of Momentum

The net external force on a system is equal to the rate of change of the momentum of that system caused by the force. The correct, calculus-based equation, reads:

$$\vec{\mathbf{F}} = \frac{d\vec{\mathbf{p}}}{dt}.$$

Example 11.5

Calculating Force: Venus Williams' Tennis Serve

During the 2007 French Open, Venus Williams hit the fastest recorded serve in a premier women's match, reaching a speed of 58 m/s (209 km/h). What is the average force exerted on the 0.057-kg tennis ball by Venus Williams' racquet? Assume that the ball's speed just after impact is 58 m/s, as shown in Figure 11.12, that the initial horizontal component of the velocity before impact is negligible, and that the ball remained in contact with the racquet for 5.0 ms.

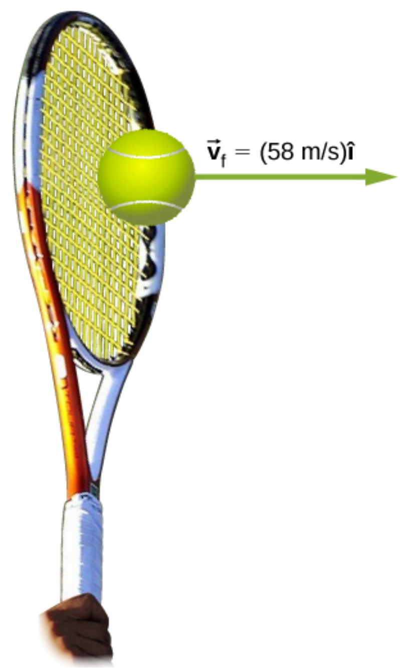


Figure 11.12 The final velocity of the tennis ball is $\vec{\mathbf{v}}_f = (58 \text{ m/s})\hat{\mathbf{i}}$.

Strategy

This problem involves only one dimension because the ball starts from having no horizontal velocity component before impact. Newton's second law stated in terms of momentum is then written as

$$\vec{\mathbf{F}} = \frac{\Delta \vec{\mathbf{p}}}{\Delta t}.$$

As noted above, when mass is constant, the change in momentum is given by

$$\Delta p = m\Delta v = m(v_f - v_i)$$

where we have used scalars because this problem involves only one dimension. In this example, the velocity just after impact and the time interval are given; thus, once Δp is calculated, we can use $F = \frac{\Delta p}{\Delta t}$ to find the force.

Solution

To determine the change in momentum, insert the values for the initial and final velocities into the equation above:

$$\begin{aligned}\Delta p &= m(v_f - v_i) \\ &= (0.057 \text{ kg})(58 \text{ m/s} - 0 \text{ m/s}) \\ &= 3.3 \frac{\text{kg} \cdot \text{m}}{\text{s}}.\end{aligned}$$

Now the magnitude of the net external force can be determined by using

$$F = \frac{\Delta p}{\Delta t} = \frac{3.3 \frac{\text{kg} \cdot \text{m}}{\text{s}}}{5.0 \times 10^{-3} \text{ s}} = 6.6 \times 10^2 \text{ N}.$$

where we have retained only two significant figures in the final step.

Significance

This quantity was the average force exerted by Venus Williams' racquet on the tennis ball during its brief impact (note that the ball also experienced the 0.57-N force of gravity, but that force was not due to the racquet). This problem could also be solved by first finding the acceleration and then using $F = ma$, but one additional step would be required compared with the strategy used in this example.

11.3 | Conservation of Linear Momentum

Learning Objectives

By the end of this section, you will be able to:

- Explain the meaning of "conservation of momentum"
- Correctly identify if a system is, or is not, closed
- Define a system whose momentum is conserved
- Mathematically express conservation of momentum for a given system
- Calculate an unknown quantity using conservation of momentum

Recall Newton's third law: When two objects of masses m_1 and m_2 interact (meaning that they apply forces on each other), the force that object 2 applies to object 1 is equal in magnitude and opposite in direction to the force that object 1 applies on object 2. Let:

- $\vec{\mathbf{F}}_{21} =$ the force on m_1 from m_2
- $\vec{\mathbf{F}}_{12} =$ the force on m_2 from m_1

Then, in symbols, Newton's third law says

$$\begin{aligned} \vec{\mathbf{F}}_{21} &= -\vec{\mathbf{F}}_{12} \\ m_1 \vec{\mathbf{a}}_1 &= -m_2 \vec{\mathbf{a}}_2. \end{aligned} \tag{11.8}$$

(Recall that these two forces do not cancel because they are applied to different objects. F_{21} causes m_1 to accelerate, and F_{12} causes m_2 to accelerate.)

Although the magnitudes of the forces on the objects are the same, the accelerations are not, simply because the masses (in general) are different. Therefore, the changes in velocity of each object are different:

$$\frac{d\vec{\mathbf{v}}_1}{dt} \neq \frac{d\vec{\mathbf{v}}_2}{dt}.$$

However, the products of the mass and the change of velocity *are* equal (in magnitude):

$$m_1 \frac{d\vec{\mathbf{v}}_1}{dt} = -m_2 \frac{d\vec{\mathbf{v}}_2}{dt}. \tag{11.9}$$

It's a good idea, at this point, to make sure you're clear on the physical meaning of the derivatives in **Equation 11.3**. Because of the interaction, each object ends up getting its velocity changed, by an amount *dv*. Furthermore, the interaction occurs over a time interval *dt*, which means that the change of velocities also occurs over *dt*. This time interval is the same for each object.

Let's assume, for the moment, that the masses of the objects do not change during the interaction. (We'll relax this restriction later.) In that case, we can pull the masses inside the derivatives:

$$\frac{d}{dt}\left(m_1 \vec{\mathbf{v}}_1\right) = -\frac{d}{dt}\left(m_2 \vec{\mathbf{v}}_2\right) \tag{11.10}$$

and thus

$$\frac{d\vec{\mathbf{p}}_1}{dt} = -\frac{d\vec{\mathbf{p}}_2}{dt}. \tag{11.11}$$

This says that *the rate at which momentum changes is the same for both objects.* The masses are different, and the changes of velocity are different, but the rate of change of the product of *m* and $\vec{\mathbf{v}}$ are the same.

Physically, this means that during the interaction of the two objects (m_1 and m_2), both objects have their momentum changed; but those changes are identical in magnitude, though opposite in sign. For example, the momentum of object 1 might increase, which means that the momentum of object 2 decreases by exactly the same amount.

In light of this, let's re-write **Equation 11.10** in a more suggestive form:

$$\frac{d\vec{\mathbf{p}}_1}{dt} + \frac{d\vec{\mathbf{p}}_2}{dt} = 0. \tag{11.12}$$

This says that during the interaction, although object 1's momentum changes, and object 2's momentum also changes, these two changes cancel each other out, so that the total change of momentum of the two objects together is zero.

Since the total combined momentum of the two objects together never changes, then we could write

$$\frac{d}{dt}\left(\vec{\mathbf{p}}_1 + \vec{\mathbf{p}}_2\right) = 0 \tag{11.13}$$

from which it follows that

$$\vec{\mathbf{p}}_1 + \vec{\mathbf{p}}_2 = \text{constant}. \tag{11.14}$$

As shown in **Figure 11.13**, the total momentum of the system before and after the collision remains the same.

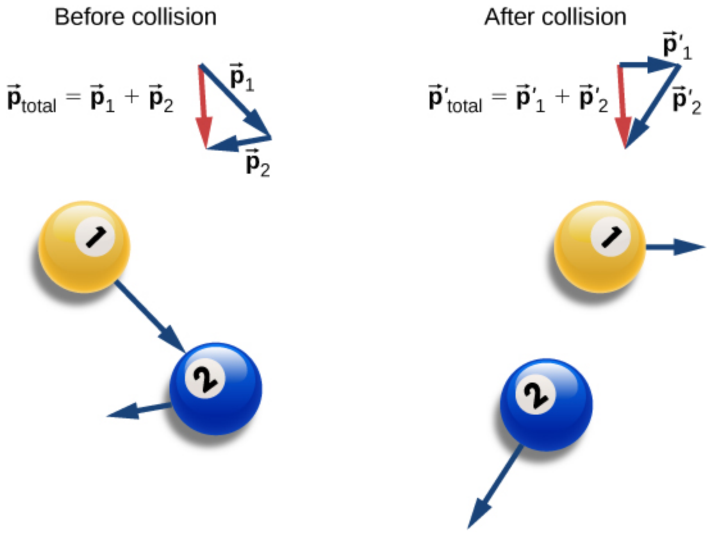


Figure 11.13 Before the collision, the two billiard balls travel with momenta $\vec{\mathbf{p}}_1$ and $\vec{\mathbf{p}}_3$. The total momentum of the system is the sum of these, as shown by the red vector labeled $\vec{\mathbf{p}}_{total}$ on the left. After the collision, the two billiard balls travel with different momenta $\vec{\mathbf{p}}'_1$ and $\vec{\mathbf{p}}'_3$. The total momentum, however, has not changed, as shown by the red vector arrow $\vec{\mathbf{p}}'_{total}$ on the right.

Generalizing this result to *N* objects, we obtain

$$\begin{aligned} \vec{\mathbf{p}}_1 + \vec{\mathbf{p}}_2 + \vec{\mathbf{p}}_3 + \cdots + \vec{\mathbf{p}}_N &= \text{constant} \\ \sum_{j=1}^{N} \vec{\mathbf{p}}_j &= \text{constant}. \end{aligned} \tag{11.15}$$

Equation 11.15 is the definition of the total (or net) momentum of a system of *N* interacting objects, along with the statement that the total momentum of a system of objects is constant in time—or better, is conserved.

Conservation Laws

If the value of a physical quantity is constant in time, we say that the quantity is conserved.

Important Simplification

The foregoing derivation of the constancy (conservation) of momentum was obtained in the most general case, i.e. taking the momentum to be a general vector quantity in (up to) three dimensions. For the remainder of this chapter, we will limit our discussions to examples in which the motion is one dimensional, usually either purely horizontal or purely vertical. In such cases, the vector nature of the momentum is simply accounted for by using positive or negative signs to indicate the direction of each momentum vector. All that is needed is for you to choose (arbitrarily) one direction to be positive and the opposite direction to be negative.

Requirements for Momentum Conservation

There is a complication, however. A system must meet two requirements for its momentum to be conserved:

1. *The mass of the system must remain constant during the interaction.*
   As the objects interact (apply forces on each other), they may *transfer* mass from one to another; but any mass one object gains is balanced by the loss of that mass from another. The total mass of the system of objects, therefore, remains unchanged as time passes:
   $$\left[\frac{dm}{dt}\right]_{\text{system}} = 0.$$
2. *The net external force on the system must be zero.*
   As the objects collide, or explode, and move around, they exert forces on each other. However, all of these forces are internal to the system, and thus each of these internal forces is balanced by another internal force that is equal in magnitude and opposite in sign. As a result, the change in momentum caused by each internal force is cancelled by another momentum change that is equal in magnitude and opposite in direction. Therefore, internal forces cannot change the total momentum of a system because the changes sum to zero. However, if there is some external force that acts on all of the objects (gravity, for example, or friction), then this force changes the momentum of the system as a whole; that is to say, the momentum of the system is changed by the external force. Thus, for the momentum of the system to be conserved, we must have
   $$\vec{\mathbf{F}}_{\mathbf{ext}} = \vec{\mathbf{0}}.$$

A system of objects that meets these two requirements is said to be a **closed system** (also called an isolated system). Thus, the more compact way to express this is shown below.

Law of Conservation of Momentum

The total momentum of a closed system is conserved:

$$\sum_{j=1}^{N} \vec{\mathbf{p}}_{\mathbf{j}} = \text{constant}.$$

This statement is called the **Law of Conservation of Momentum**. Along with the conservation of energy, it is one of the foundations upon which all of physics stands. All our experimental evidence supports this statement: from the motions of galactic clusters to the quarks that make up the proton and the neutron, and at every scale in between. *In a closed system, the total momentum never changes.*

Note that there absolutely *can* be external forces acting on the system; but for the system's momentum to remain constant, these external forces have to cancel, so that the *net* external force is zero. Billiard balls on a table all have a weight force acting on them, but the weights are balanced (canceled) by the normal forces, so there is no *net* force.

The Meaning of 'System'

A **system** (mechanical) is the collection of objects in whose motion (kinematics and dynamics) you are interested. If you are analyzing the bounce of a ball on the ground, you are probably only interested in the motion of the ball, and not of Earth; thus, the ball is your system. If you are analyzing a car crash, the two cars together compose your system (**Figure 11.14**).

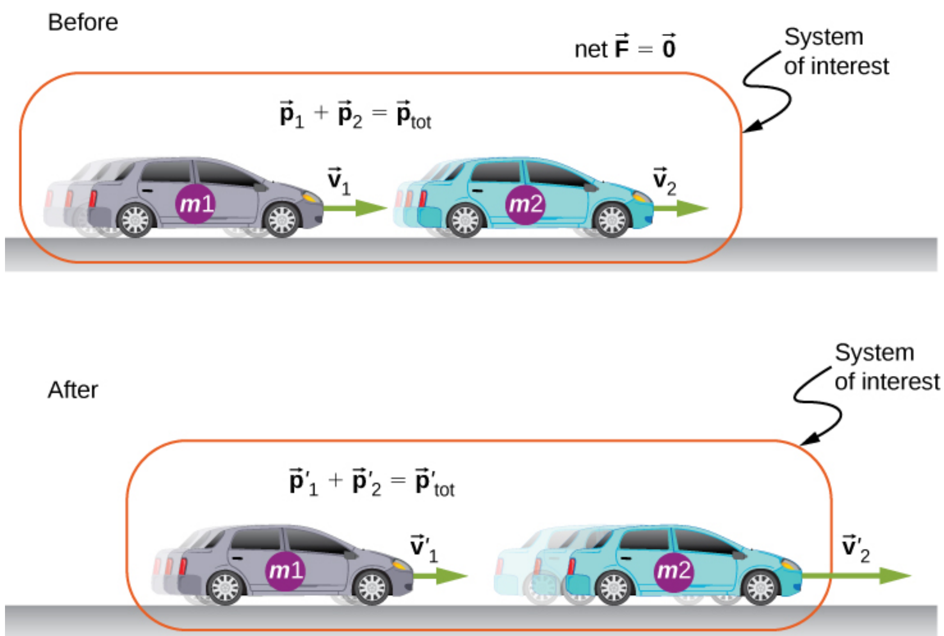


Figure 11.14 The two cars together form the system that is to be analyzed. It is important to remember that the contents (the mass) of the system do not change before, during, or after the objects in the system interact.

Problem-Solving Strategy: Conservation of Momentum

Using conservation of momentum requires four basic steps. The first step is crucial:

1. Identify a closed system (total mass is constant, no net external force acts on the system).
2. Write down an expression representing the total momentum of the system before the “event” (explosion or collision).
3. Write down an expression representing the total momentum of the system after the “event.”
4. Set these two expressions equal to each other, and solve this equation for the desired quantity.

Example 11.6

Colliding Carts

Two carts in a physics lab roll on a level track, with negligible friction. These carts have small magnets at their ends, so that when they collide, they stick together (**Figure 11.15**). The first cart has a mass of 675 grams and is rolling at 0.75 m/s to the right; the second has a mass of 500 grams and is rolling at 1.33 m/s, also to the right. After the collision, what is the velocity of the two joined carts?

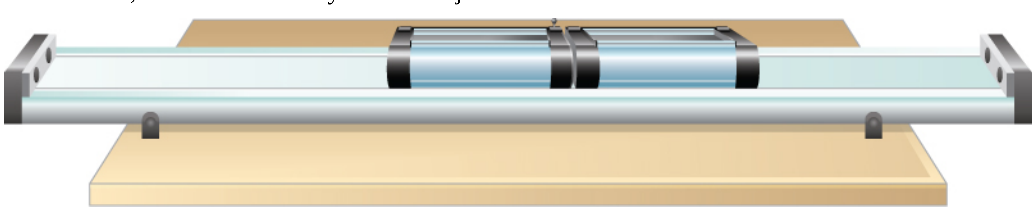


Figure 11.15 Two lab carts collide and stick together after the collision.

Strategy

We have a collision. We're given masses and initial velocities; we're asked for the final velocity. This all suggests using conservation of momentum as a method of solution. However, we can only use it if we have a closed system. So we need to be sure that the system we choose has no net external force on it, and that its mass is not changed by the collision.

Defining the system to be the two carts meets the requirements for a closed system: The combined mass of the two carts certainly doesn't change, and while the carts definitely exert forces on each other, those forces are internal to the system, so they do not change the momentum of the system as a whole. In the vertical direction, the weights of the carts are canceled by the normal forces on the carts from the track.

Solution

Conservation of momentum is

$$\vec{\mathbf{p}}_{\mathbf{f}} = \vec{\mathbf{p}}_{\mathbf{i}}.$$

Define the direction of their initial velocity vectors to be the +*x*-direction. The initial momentum is then

$$\vec{\mathbf{p}}_{\mathbf{i}} = m_1 v_1 \hat{\mathbf{i}} + m_2 v_2 \hat{\mathbf{i}}.$$

The final momentum of the now-linked carts is

$$\vec{\mathbf{p}}_{\mathbf{f}} = (m_1 + m_2)\vec{\mathbf{v}}_{\mathbf{f}}.$$

Equating:

$$\begin{aligned}(m_1 + m_2)\vec{\mathbf{v}}_{\mathbf{f}} &= m_1 v_1 \hat{\mathbf{i}} + m_2 v_2 \hat{\mathbf{i}} \\ \vec{\mathbf{v}}_{\mathbf{f}} &= \left(\frac{m_1 v_1 + m_2 v_2}{m_1 + m_2}\right)\hat{\mathbf{i}}.\end{aligned}$$

Substituting the given numbers:

$$\begin{aligned}\vec{\mathbf{v}}_{\mathbf{f}} &= \left[\frac{(0.675\ \text{kg})(0.75\ \text{m/s}) + (0.5\ \text{kg})(1.33\ \text{m/s})}{1.175\ \text{kg}}\right]\hat{\mathbf{i}} \\ &= (0.997\ \text{m/s})\hat{\mathbf{i}}.\end{aligned}$$

Significance

The principles that apply here to two laboratory carts apply identically to all objects of whatever type or size. Even for photons, the concepts of momentum and conservation of momentum are still crucially important even at that scale. (Since they are massless, the momentum of a photon is defined very differently from the momentum of ordinary objects. You will learn about this when you study quantum physics.)

11.3 Check Your Understanding Suppose the second, smaller cart had been initially moving to the left. What would the sign of the final velocity have been in this case?

Example 11.7

A Bouncing Superball

A superball of mass 0.25 kg is dropped from rest from a height of $h = 1.50\ \text{m}$ above the floor. It bounces with no loss of energy and returns to its initial height (**Figure 11.16**).

a. What is the superball's change of momentum during its bounce on the floor?

b. What was Earth's change of momentum due to the ball colliding with the floor?

c. What was Earth's change of velocity as a result of this collision?

(This example shows that you have to be careful about defining your system.)

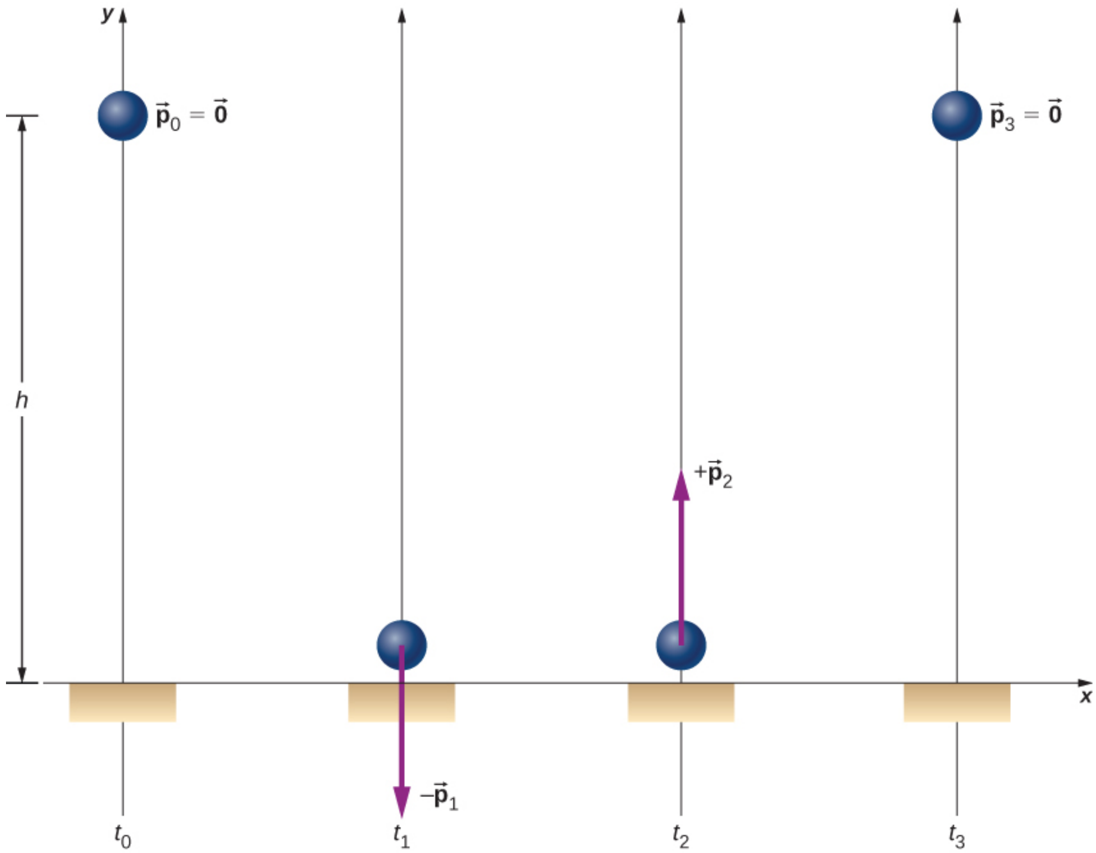


Figure 11.16 A superball is dropped to the floor (t_0), hits the floor (t_1), bounces (t_2), and returns to its initial height (t_3).

Strategy

Since we are asked only about the ball's change of momentum, we define our system to be the ball. But this is clearly not a closed system; gravity applies a downward force on the ball while it is falling, and the normal force from the floor applies a force during the bounce. Thus, we cannot use conservation of momentum as a strategy. Instead, we simply determine the ball's momentum just before it collides with the floor and just after, and calculate the difference. We have the ball's mass, so we need its velocities.

Solution

a. Since this is a one-dimensional problem, we use the scalar form of the equations. Let:
   - $p_0 =$ the magnitude of the ball's momentum at time t_0, the moment it was released; since it was dropped from rest, this is zero.
   - $p_1 =$ the magnitude of the ball's momentum at time t_1, the instant just before it hits the floor.
   - $p_2 =$ the magnitude of the ball's momentum at time t_2, just after it loses contact with the floor after the bounce.

   The ball's change of momentum is

$$\begin{aligned}\Delta \vec{\mathbf{p}} &= \vec{\mathbf{p}}_2 - \vec{\mathbf{p}}_1\\ &= p_2 \hat{\mathbf{j}} - \left(-p_1 \hat{\mathbf{j}}\right)\\ &= (p_2 + p_1)\hat{\mathbf{j}}.\end{aligned}$$

Its velocity just before it hits the floor can be determined from either conservation of energy or kinematics. We use kinematics here; you should re-solve it using conservation of energy and confirm you get the same result.

We want the velocity just before it hits the ground (at time t_1). We know its initial velocity $v_0 = 0$ (at time t_0), the height it falls, and its acceleration; we don't know the fall time. We could calculate that, but instead we use

$$\vec{\mathbf{v}}_1 = -\hat{\mathbf{j}}\sqrt{2gy} = -5.4\text{ m/s}\,\hat{\mathbf{j}}.$$

Thus the ball has a momentum of

$$\begin{aligned}\vec{\mathbf{p}}_1 &= -(0.25\text{ kg})\left(-5.4\text{ m/s}\,\hat{\mathbf{j}}\right)\\ &= -(1.4\text{ kg}\cdot\text{m/s})\,\hat{\mathbf{j}}.\end{aligned}$$

We don't have an easy way to calculate the momentum after the bounce. Instead, we reason from the symmetry of the situation.

Before the bounce, the ball starts with zero velocity and falls 1.50 m under the influence of gravity, achieving some amount of momentum just before it hits the ground. On the return trip (after the bounce), it starts with some amount of momentum, rises the same 1.50 m it fell, and ends with zero velocity. Thus, the motion after the bounce was the mirror image of the motion before the bounce. From this symmetry, it must be true that the ball's momentum after the bounce must be equal and opposite to its momentum before the bounce. (This is a subtle but crucial argument; make sure you understand it before you go on.) Therefore,

$$\vec{\mathbf{p}}_2 = -\vec{\mathbf{p}}_1 = +(1.4\text{ kg}\cdot\text{m/s})\,\hat{\mathbf{j}}.$$

Thus, the ball's change of velocity during the bounce is

$$\begin{aligned}\Delta \vec{\mathbf{p}} &= \vec{\mathbf{p}}_2 - \vec{\mathbf{p}}_1\\ &= (1.4\text{ kg}\cdot\text{m/s})\,\hat{\mathbf{j}} - (-1.4\text{ kg}\cdot\text{m/s})\,\hat{\mathbf{j}}\\ &= +(2.8\text{ kg}\cdot\text{m/s})\,\hat{\mathbf{j}}.\end{aligned}$$

b. What was Earth's change of momentum due to the ball colliding with the floor?

Your instinctive response may well have been either "zero; the Earth is just too massive for that tiny ball to have affected it" or possibly, "more than zero, but utterly negligible." But no—if we re-define our system to be the Superball + Earth, then this system is closed (neglecting the gravitational pulls of the Sun, the Moon, and the other planets in the solar system), and therefore the total change of momentum of this new system must be zero. Therefore, Earth's change of momentum is exactly the same magnitude:

$$\Delta \vec{\mathbf{p}}_{\text{Earth}} = -2.8\text{ kg}\cdot\text{m/s}\,\hat{\mathbf{j}}.$$

c. What was Earth's change of velocity as a result of this collision?

This is where your instinctive feeling is probably correct:

$$\Delta \vec{v}_{\text{Earth}} = \frac{\Delta \vec{p}_{\text{Earth}}}{M_{\text{Earth}}} = -\frac{2.8 \text{ kg} \cdot \text{m/s}}{5.97 \times 10^{24} \text{ kg}} \hat{j} = -\left(4.7 \times 10^{-25} \text{ m/s}\right)\hat{j}.$$

This change of Earth's velocity *is* utterly negligible.

Significance

It is important to realize that the answer to part (c) is not a velocity; it is a change of velocity, which is a very different thing. Nevertheless, to give you a feel for just how small that change of velocity is, suppose you were moving with a velocity of 4.7×10^{-25} m/s. At this speed, it would take you about 7 million years to travel a distance equal to the diameter of a hydrogen atom.

11.4 Check Your Understanding Would the ball's change of momentum have been larger, smaller, or the same, if it had collided with the floor and stopped (without bouncing)?

Would the ball's change of momentum have been larger, smaller, or the same, if it had collided with the floor and stopped (without bouncing)?

Example 11.8

Ice Hockey 1

Two hockey pucks of identical mass are on a flat, horizontal ice hockey rink. The red puck is motionless; the blue puck is moving at 2.5 m/s to the left (**Figure 11.17**). It collides with the motionless red puck. The pucks have a mass of 15 g. After the collision, the red puck is moving at 2.5 m/s, to the left. What is the final velocity of the blue puck?

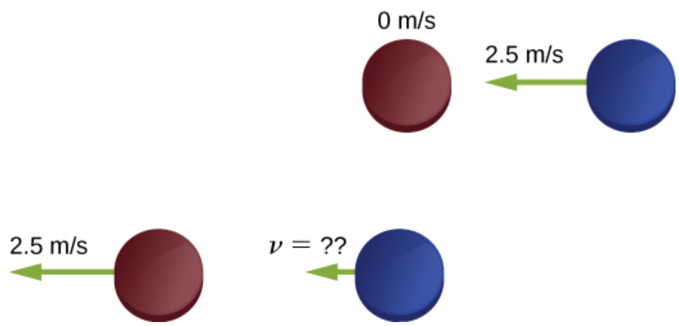


Figure 11.17 Two identical hockey pucks colliding. The top diagram shows the pucks the instant before the collision, and the bottom diagram show the pucks the instant after the collision. The net external force is zero.

Strategy

We're told that we have two colliding objects, we're told the masses and initial velocities, and one final velocity; we're asked for both final velocities. Conservation of momentum seems like a good strategy. Define the system to be the two pucks; there's no friction, so we have a closed system.

Before you look at the solution, what do you think the answer will be?

The blue puck final velocity will be:

- zero

- 2.5 m/s to the left
- 2.5 m/s to the right
- 1.25 m/s to the left
- 1.25 m/s to the right
- something else

Solution

Define the +*x*-direction to point to the right. Conservation of momentum then reads

$$\vec{\mathbf{p}}_{\mathbf{f}} = \vec{\mathbf{p}}_{\mathbf{i}}$$
$$mv_{r_f}\,\hat{\mathbf{i}} + mv_{b_f}\,\hat{\mathbf{i}} = mv_{r_i}\,\hat{\mathbf{i}} - mv_{b_i}\,\hat{\mathbf{i}}.$$

Before the collision, the momentum of the system is entirely and only in the blue puck. Thus,

$$mv_{r_f}\,\hat{\mathbf{i}} + mv_{b_f}\,\hat{\mathbf{i}} = -mv_{b_i}\,\hat{\mathbf{i}}$$

$$v_{r_f}\,\hat{\mathbf{i}} + v_{b_f}\,\hat{\mathbf{i}} = -v_{b_i}\,\hat{\mathbf{i}}.$$

(Remember that the masses of the pucks are equal.) Substituting numbers:

$$-(2.5\text{ m/s})\,\hat{\mathbf{i}} + \vec{\mathbf{v}}_{b_f} = -(2.5\text{ m/s})\,\hat{\mathbf{i}}$$

$$\vec{\mathbf{v}}_{b_f} = 0.$$

Significance

Evidently, the two pucks simply exchanged momentum. The blue puck transferred all of its momentum to the red puck. In fact, this is what happens in similar collision where $m_1 = m_2$.

11.5 Check Your Understanding Even if there were some friction on the ice, it is still possible to use conservation of momentum to solve this problem, but you would need to impose an additional condition on the problem. What is that additional condition?

Example 11.9

Landing of *Philae*

On November 12, 2014, the European Space Agency successfully landed a probe named *Philae* on Comet 67P/Churyumov/Gerasimenko (**Figure 11.18**). During the landing, however, the probe actually landed three times, because it bounced twice. Let’s calculate how much the comet’s speed changed as a result of the first bounce.

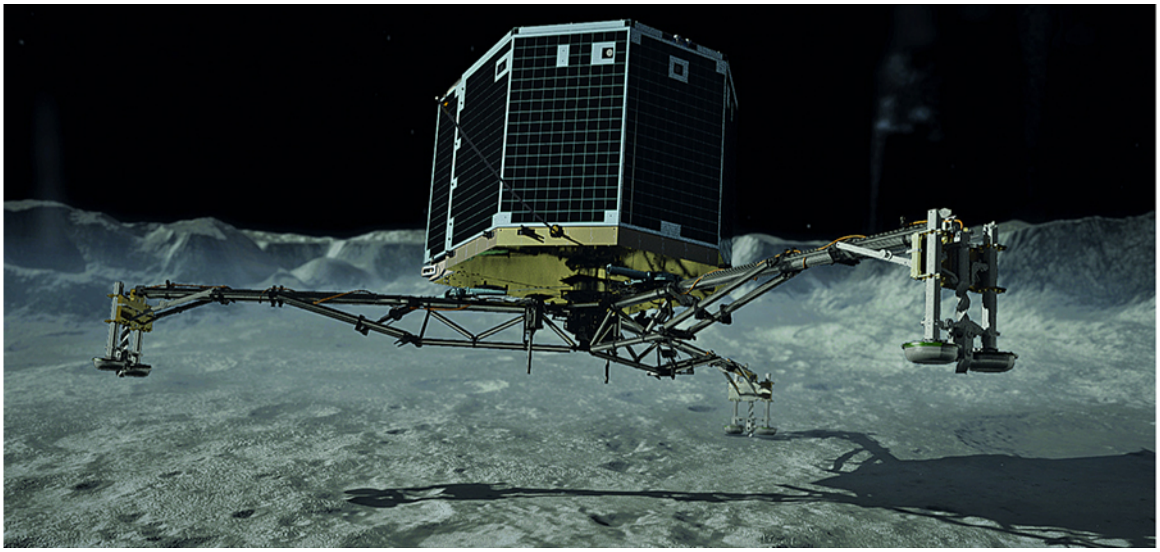


Figure 11.18 An artist's rendering of *Philae* landing on a comet. (credit: modification of work by "DLR German Aerospace Center"/Flickr)

Let's define upward to be the +*y*-direction, perpendicular to the surface of the comet, and $y = 0$ to be at the surface of the comet. Here's what we know:

- The mass of Comet 67P: $M_c = 1.0 \times 10^{13}$ kg
- The acceleration due to the comet's gravity: $\vec{\mathbf{a}} = -\left(5.0 \times 10^{-3} \text{ m/s}^2\right)\hat{\mathbf{j}}$
- *Philae's* mass: $M_p = 96$ kg
- Initial touchdown speed: $\vec{\mathbf{v}}_1 = -(1.0 \text{ m/s})\,\hat{\mathbf{j}}$
- Initial upward speed due to first bounce: $\vec{\mathbf{v}}_2 = (0.38 \text{ m/s})\,\hat{\mathbf{j}}$
- Landing impact time: $\Delta t = 1.3$ s

Strategy

We're asked for how much the comet's speed changed, but we don't know much about the comet, beyond its mass and the acceleration its gravity causes. However, we *are* told that the *Philae* lander collides with (lands on) the comet, and bounces off of it. A collision suggests momentum as a strategy for solving this problem.

If we define a system that consists of both *Philae* and Comet 67/P, then there is no net external force on this system, and thus the momentum of this system is conserved. (We'll neglect the gravitational force of the sun.) Thus, if we calculate the change of momentum of the lander, we automatically have the change of momentum of the comet. Also, the comet's change of velocity is directly related to its change of momentum as a result of the lander "colliding" with it.

Solution

Let $\vec{\mathbf{p}}_1$ be *Philae's* momentum at the moment just before touchdown, and $\vec{\mathbf{p}}_2$ be its momentum just after the first bounce. Then its momentum just before landing was

$$\vec{\mathbf{p}}_1 = M_p \vec{\mathbf{v}}_1 = (96 \text{ kg})\left(-1.0 \text{ m/s}\,\hat{\mathbf{j}}\right) = -(96 \text{ kg} \cdot \text{m/s})\,\hat{\mathbf{j}}$$

and just after was

$$\vec{\mathbf{p}}_2 = M_p \vec{\mathbf{v}}_2 = (96\text{ kg})\left(+0.38\text{ m/s}\,\hat{\mathbf{j}}\right) = (36.5\text{ kg}\cdot\text{m/s})\,\hat{\mathbf{j}}.$$

Therefore, the lander's change of momentum during the first bounce is

$$\begin{aligned}\Delta\vec{\mathbf{p}} &= \vec{\mathbf{p}}_2 - \vec{\mathbf{p}}_1 \\ &= (36.5\text{ kg}\cdot\text{m/s})\,\hat{\mathbf{j}} - \left(-96.0\text{ kg}\cdot\text{m/s}\,\hat{\mathbf{j}}\right) = (133\text{ kg}\cdot\text{m/s})\,\hat{\mathbf{j}}\end{aligned}$$

Notice how important it is to include the negative sign of the initial momentum.

Now for the comet. Since momentum of the system must be conserved, the *comet's* momentum changed by exactly the negative of this:

$$\Delta\vec{\mathbf{p}}_c = -\Delta\vec{\mathbf{p}} = -(133\text{ kg}\cdot\text{m/s})\,\hat{\mathbf{j}}.$$

Therefore, its change of velocity is

$$\Delta\vec{\mathbf{v}}_c = \frac{\Delta\vec{\mathbf{p}}_c}{M_c} = \frac{-(133\text{ kg}\cdot\text{m/s})\,\hat{\mathbf{j}}}{1.0\times10^{13}\text{ kg}} = -\left(1.33\times10^{-11}\text{ m/s}\right)\hat{\mathbf{j}}.$$

Significance

This is a very small change in velocity, about a thousandth of a billionth of a meter per second. Crucially, however, it is *not* zero.

11.6 Check Your Understanding The changes of momentum for *Philae* and for Comet 67/P were equal (in magnitude). Were the impulses experienced by *Philae* and the comet equal? How about the forces? How about the changes of kinetic energies?

11.4 | Types of Collisions

Learning Objectives

By the end of this section, you will be able to:

- Identify the type of collision
- Correctly label a collision as elastic or inelastic
- Use kinetic energy along with momentum and impulse to analyze a collision

Although momentum is conserved in all interactions, not all interactions (collisions or explosions) are the same. The possibilities include:

- A single object can explode into multiple objects (one-to-many).
- Multiple objects can collide and stick together, forming a single object (many-to-one).
- Multiple objects can collide and bounce off of each other, remaining as multiple objects (many-to-many). If they do bounce off each other, then they may recoil at the same speeds with which they approached each other before the collision, or they may move off more slowly.

It's useful, therefore, to categorize different types of interactions, according to how the interacting objects move before and after the interaction.

One-to-Many

The first possibility is that a single object may break apart into two or more pieces. An example of this is a firecracker, or a bow and arrow, or a rocket rising through the air toward space. These can be difficult to analyze if the number of fragments after the collision is more than about three or four; but nevertheless, the total momentum of the system before and after the explosion is identical.

Note that if the object is initially motionless, then the system (which is just the object) has no momentum and no kinetic energy. After the explosion, the net momentum of all the pieces of the object must sum to zero (since the momentum of this closed system cannot change). However, the system *will* have a great deal of kinetic energy after the explosion, although it had none before. Thus, we see that, although the momentum of the system is conserved in an explosion, the kinetic energy of the system most definitely is not; it increases. This interaction—one object becoming many, with an increase of kinetic energy of the system—is called an **explosion**.

Where does the energy come from? Does conservation of energy still hold? Yes; some form of potential energy is converted to kinetic energy. In the case of gunpowder burning and pushing out a bullet, chemical potential energy is converted to kinetic energy of the bullet, and of the recoiling gun. For a bow and arrow, it is elastic potential energy in the bowstring.

Many-to-One

The second possibility is the reverse: that two or more objects collide with each other and stick together, thus (after the collision) forming one single composite object. The total mass of this composite object is the sum of the masses of the original objects, and the new single object moves with a velocity dictated by the conservation of momentum. However, it turns out again that, although the total momentum of the system of objects remains constant, the kinetic energy doesn't; but this time, the kinetic energy decreases. This type of collision is called **inelastic**.

In the extreme case, multiple objects collide, stick together, and remain motionless after the collision. Since the objects are all motionless after the collision, the final kinetic energy is also zero; the loss of kinetic energy is a maximum. Such a collision is said to be **perfectly inelastic**.

Many-to-Many

The extreme case on the other end is if two or more objects approach each other, collide, and bounce off each other, moving away from each other at the same relative speed at which they approached each other. In this case, the total kinetic energy of the system is conserved. Such an interaction is called **elastic**.

In any interaction of a closed system of objects, the total momentum of the system is conserved ($\vec{\mathbf{p}}_f = \vec{\mathbf{p}}_\text{i}$) but the kinetic energy may not be:

- If $0 < K_\text{f} < K_\text{i}$, the collision is inelastic.
- If $K_\text{f} = 0$, the collision is perfectly inelastic.
- If $K_\text{f} = K_\text{i}$, the collision is elastic.
- If $K_\text{f} > K_\text{i}$, the interaction is an explosion.

The point of all this is that, in analyzing a collision or explosion, you can use both momentum and kinetic energy.

Problem-Solving Strategy: Collisions

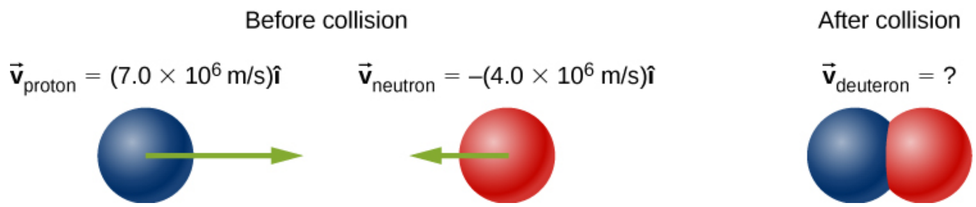
A closed system always conserves momentum; it might also conserve kinetic energy, but very often it doesn't. Energy-momentum problems confined to a plane (as ours are) usually have two unknowns. Generally, this approach works well:

1. Define a closed system.
2. Write down the expression for conservation of momentum.
3. If kinetic energy is conserved, write down the expression for conservation of kinetic energy; if not, write down the expression for the change of kinetic energy.
4. You now have two equations in two unknowns, which you solve by standard methods.

Example 11.10

Formation of a Deuteron

A proton (mass 1.67×10^{-27} kg) collides with a neutron (with essentially the same mass as the proton) to form a particle called a *deuteron*. What is the velocity of the deuteron if it is formed from a proton moving with velocity 7.0×10^{6} m/s to the left and a neutron moving with velocity 4.0×10^{6} m/s to the right?



Strategy

Define the system to be the two particles. This is a collision, so we should first identify what kind. Since we are told the two particles form a single particle after the collision, this means that the collision is perfectly inelastic. Thus, kinetic energy is not conserved, but momentum is. Thus, we use conservation of energy to determine the final velocity of the system.

Solution

Treat the two particles as having identical masses M. Use the subscripts p, n, and d for proton, neutron, and deuteron, respectively. This is a one-dimensional problem, so we have

$$Mv_p - Mv_n = 2Mv_d.$$

The masses divide out:

$$\begin{aligned} v_p - v_n &= 2v_d \\ 7.0 \times 10^6 \text{ m/s} - 4.0 \times 10^6 \text{ m/s} &= 2v_d \\ v_d &= 1.5 \times 10^6 \text{ m/s}. \end{aligned}$$

The velocity is thus $\vec{v}_d = (1.5 \times 10^6 \text{ m/s})\hat{i}$.

Significance

This is essentially how particle colliders like the Large Hadron Collider work: They accelerate particles up to very high speeds (large momenta), but in opposite directions. This maximizes the creation of so-called "daughter particles."

Example 11.11

Ice Hockey 2

(This is a variation of an earlier example.)

Two ice hockey pucks of different masses are on a flat, horizontal hockey rink. The red puck has a mass of 15 grams, and is motionless; the blue puck has a mass of 12 grams, and is moving at 2.5 m/s to the left. It collides with the motionless red puck (Figure 11.19). If the collision is perfectly elastic, what are the final velocities of the two pucks?

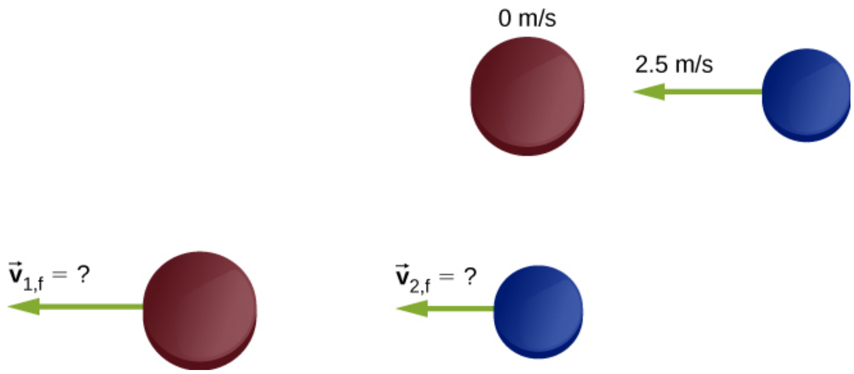


Figure 11.19 Two different hockey pucks colliding. The top diagram shows the pucks the instant before the collision, and the bottom diagram show the pucks the instant after the collision. The net external force is zero.

Strategy

We're told that we have two colliding objects, and we're told their masses and initial velocities; we're asked for both final velocities. Conservation of momentum seems like a good strategy; define the system to be the two pucks. There is no friction, so we have a closed system. We have two unknowns (the two final velocities), but only one equation. The comment about the collision being perfectly elastic is the clue; it suggests that kinetic energy is also conserved in this collision. That gives us our second equation.

The initial momentum and initial kinetic energy of the system resides entirely and only in the second puck (the blue one); the collision transfers some of this momentum and energy to the first puck.

Solution

Conservation of momentum, in this case, reads

$$\begin{aligned} p_\text{i} &= p_\text{f} \\ m_2 v_{2,\text{i}} &= m_1 v_{1,\text{f}} + m_2 v_{2,\text{f}}. \end{aligned}$$

Conservation of kinetic energy reads

$$\begin{aligned} K_\text{i} &= K_\text{f} \\ \frac{1}{2} m_2 v_{2,\text{i}}^2 &= \frac{1}{2} m_1 v_{1,\text{f}}^2 + \frac{1}{2} m_2 v_{2,\text{f}}^2. \end{aligned}$$

There are our two equations in two unknowns. The algebra is tedious but not terribly difficult; you definitely should work it through. The solution is

$$\begin{aligned} v_{1,\text{f}} &= \frac{(m_1 - m_2)v_{1,\text{i}} + 2m_2 v_{2,\text{i}}}{m_1 + m_2} \\ v_{2\text{f}} &= \frac{(m_2 - m_1)v_{2,\text{i}} + 2m_1 v_{1,\text{i}}}{m_1 + m_2}. \end{aligned}$$

Substituting the given numbers, we obtain

$$\begin{aligned} v_{1,\text{f}} &= 2.22 \, \frac{\text{m}}{\text{s}} \\ v_{2,\text{f}} &= -0.28 \, \frac{\text{m}}{\text{s}}. \end{aligned}$$

Significance

Notice that after the collision, the blue puck is moving to the right; its direction of motion was reversed. The red puck is now moving to the left.

11.7 Check Your Understanding There is a second solution to the system of equations solved in this example (because the energy equation is quadratic): $v_{1,\text{f}} = -2.5\text{ m/s}$, $v_{2,\text{f}} = 0$. This solution is unacceptable on physical grounds; what's wrong with it?

Example 11.12

Thor vs. Iron Man

The 2012 movie "The Avengers" has a scene where Iron Man and Thor fight. At the beginning of the fight, Thor throws his hammer at Iron Man, hitting him and throwing him slightly up into the air and against a small tree, which breaks. From the video, Iron Man is standing still when the hammer hits him. The distance between Thor and Iron Man is approximately 10 m, and the hammer takes about 1 s to reach Iron Man after Thor releases it. The tree is about 2 m behind Iron Man, which he hits in about 0.75 s. Also from the video, Iron Man's trajectory to the tree is very close to horizontal. Assuming Iron Man's total mass is 200 kg:

a. Estimate the mass of Thor's hammer
b. Estimate how much kinetic energy was lost in this collision

Strategy

After the collision, Thor's hammer is in contact with Iron Man for the entire time, so this is a perfectly inelastic collision. Thus, with the correct choice of a closed system, we expect momentum is conserved, but not kinetic energy. We use the given numbers to estimate the initial momentum, the initial kinetic energy, and the final kinetic energy. Because this is a one-dimensional problem, we can go directly to the scalar form of the equations.

Solution

a. First, we posit conservation of momentum. For that, we need a closed system. The choice here is the system (hammer + Iron Man), from the time of collision to the moment just before Iron Man and the hammer hit the tree. Let:

   - $M_\text{H} =$ mass of the hammer
   - $M_\text{I} =$ mass of Iron Man
   - $v_\text{H} =$ velocity of the hammer before hitting Iron Man
   - $v =$ combined velocity of Iron Man + hammer after the collision

   Again, Iron Man's initial velocity was zero. Conservation of momentum here reads:

$$M_\text{H} v_\text{H} = (M_\text{H} + M_\text{I})v.$$

   We are asked to find the mass of the hammer, so we have

$$\begin{aligned} M_\text{H} v_\text{H} &= M_\text{H} v + M_\text{I} v \\ M_\text{H}(v_\text{H} - v) &= M_\text{I} v \\ M_\text{H} &= \frac{M_\text{I} v}{v_\text{H} - v} \\ &= \frac{(200\text{ kg})\left(\frac{2\text{ m}}{0.75\text{ s}}\right)}{10\,\frac{\text{m}}{\text{s}} - \left(\frac{2\text{ m}}{0.75\text{ s}}\right)} \\ &= 73\text{ kg}. \end{aligned}$$

   Considering the uncertainties in our estimates, this should be expressed with just one significant figure; thus, $M_\text{H} = 7 \times 10^1\text{ kg}$.

b. The initial kinetic energy of the system, like the initial momentum, is all in the hammer:

$$\begin{aligned} K_{\text{i}} &= \frac{1}{2}M_{\text{H}}v_{\text{H}}^2 \\ &= \frac{1}{2}(70\ \text{kg})(10\ \text{m/s})^2 \\ &= 3500\ \text{J}. \end{aligned}$$

After the collision,

$$\begin{aligned} K_{\text{f}} &= \frac{1}{2}(M_{\text{H}} + M_{\text{I}})v^2 \\ &= \frac{1}{2}(70\ \text{kg} + 200\ \text{kg})(2.67\ \text{m/s})^2 \\ &= 960\ \text{J}. \end{aligned}$$

Thus, there was a loss of $3500\ \text{J} - 960\ \text{J} = 2540\ \text{J}$.

Significance

From other scenes in the movie, Thor apparently can control the hammer's velocity with his mind. It is possible, therefore, that he mentally causes the hammer to maintain its initial velocity of 10 m/s while Iron Man is being driven backward toward the tree. If so, this would represent an external force on our system, so it would not be closed. Thor's mental control of his hammer is beyond the scope of this book, however.

Example 11.13

Analyzing a Car Crash

At a stoplight, a large truck (3000 kg) collides with a motionless small car (1200 kg). The truck comes to an instantaneous stop; the car slides straight ahead, coming to a stop after sliding 10 meters. The measured coefficient of friction between the car's tires and the road was 0.62. How fast was the truck moving at the moment of impact?

Strategy

At first it may seem we don't have enough information to solve this problem. Although we know the initial speed of the car, we don't know the speed of the truck (indeed, that's what we're asked to find), so we don't know the initial momentum of the system. Similarly, we know the final speed of the truck, but not the speed of the car immediately after impact. The fact that the car eventually slid to a speed of zero doesn't help with the final momentum, since an external friction force caused that. Nor can we calculate an impulse, since we don't know the collision time, or the amount of time the car slid before stopping. A useful strategy is to impose a restriction on the analysis.

Suppose we define a system consisting of just the truck and the car. The momentum of this system isn't conserved, because of the friction between the car and the road. But if we *could* find the speed of the car the instant after impact—before friction had any measurable effect on the car—then we could consider the momentum of the system to be conserved, with that restriction.

Can we find the final speed of the car? Yes; we invoke the work-kinetic energy theorem.

Solution

First, define some variables. Let:

- M_{c} and M_{T} be the masses of the car and truck, respectively
- $v_{\text{T,i}}$ and$v_{\text{T,f}}$ be the velocities of the truck before and after the collision, respectively
- $v_{\text{c,i}}$ and$v_{\text{c,f}}$ Z be the velocities of the car before and after the collision, respectively
- K_{i} and K_{f} be the kinetic energies of the car immediately after the collision, and after the car has stopped sliding (so $K_{\text{f}} = 0$).

- d be the distance the car slides after the collision before eventually coming to a stop.

Since we actually want the initial speed of the truck, and since the truck is not part of the work-energy calculation, let's start with conservation of momentum. For the car + truck system, conservation of momentum reads

$$\begin{aligned} p_i &= p_f \\ M_c v_{c,i} + M_T v_{T,i} &= M_c v_{c,f} + M_T v_{T,f}. \end{aligned}$$

Since the car's initial velocity was zero, as was the truck's final velocity, this simplifies to

$$v_{T,i} = \frac{M_c}{M_T} v_{c,f}.$$

So now we need the car's speed immediately after impact. Recall that

$$W = \Delta K$$

where

$$\begin{aligned} \Delta K &= K_f - K_i \\ &= 0 - \frac{1}{2} M_c v_{c,f}^2. \end{aligned}$$

Also,

$$W = \vec{\mathbf{F}} \cdot \vec{\mathbf{d}} = Fd\cos\theta.$$

The work is done over the distance the car slides, which we've called d. Equating:

$$Fd\cos\theta = -\frac{1}{2} M_c v_{c,f}^2.$$

Friction is the force on the car that does the work to stop the sliding. With a level road, the friction force is

$$F = \mu_k M_c g.$$

Since the angle between the directions of the friction force vector and the displacement d is $180°$, and $\cos(180°) = -1,$ we have

$$-(\mu_k M_c g)d = -\frac{1}{2} M_c v_{c,f}^2$$

(Notice that the car's mass divides out; evidently the mass of the car doesn't matter.)

Solving for the car's speed immediately after the collision gives

$$v_{c,f} = \sqrt{2\mu_k g d}.$$

Substituting the given numbers:

$$\begin{aligned} v_{c,f} &= \sqrt{2(0.62)\left(9.81\,\frac{\text{m}}{\text{s}^2}\right)(10\,\text{m})}. \\ &= 11.0\ \text{m/s}. \end{aligned}$$

Now we can calculate the initial speed of the truck:

$$v_{T,i} = \left(\frac{1200\ \text{kg}}{3000\ \text{kg}}\right)\left(11.0\,\frac{\text{m}}{\text{s}}\right) = 4.4\ \text{m/s}.$$

Significance

This is an example of the type of analysis done by investigators of major car accidents. A great deal of legal and financial consequences depend on an accurate analysis and calculation of momentum and energy.

11.8 Check Your Understanding Suppose there had been no friction (the collision happened on ice); that would make μ_k zero, and thus $v_{c,f} = \sqrt{2\mu_k gd} = 0$, which is obviously wrong. What is the mistake in this conclusion?

11.5 | Center of Mass

Learning Objectives

By the end of this section, you will be able to:

- Explain the meaning and usefulness of the concept of center of mass
- Calculate the center of mass of a simple system
- Calculate the velocity and acceleration of the center of mass

We have been avoiding an important issue up to now: When we say that an object moves (more correctly, accelerates) in a way that obeys Newton's second law, we have been ignoring the fact that all objects are actually made of many constituent particles. A car has an engine, steering wheel, seats, passengers; a football is leather and rubber surrounding air; a brick is made of atoms. There are many different types of particles, and they are generally not distributed uniformly in the object. How do we include these facts into our calculations?

Then too, an extended object might change shape as it moves, such as a water balloon or a cat falling (Figure 11.20). This implies that the constituent particles are applying internal forces on each other, in addition to the external force that is acting on the object as a whole. We want to be able to handle this, as well.

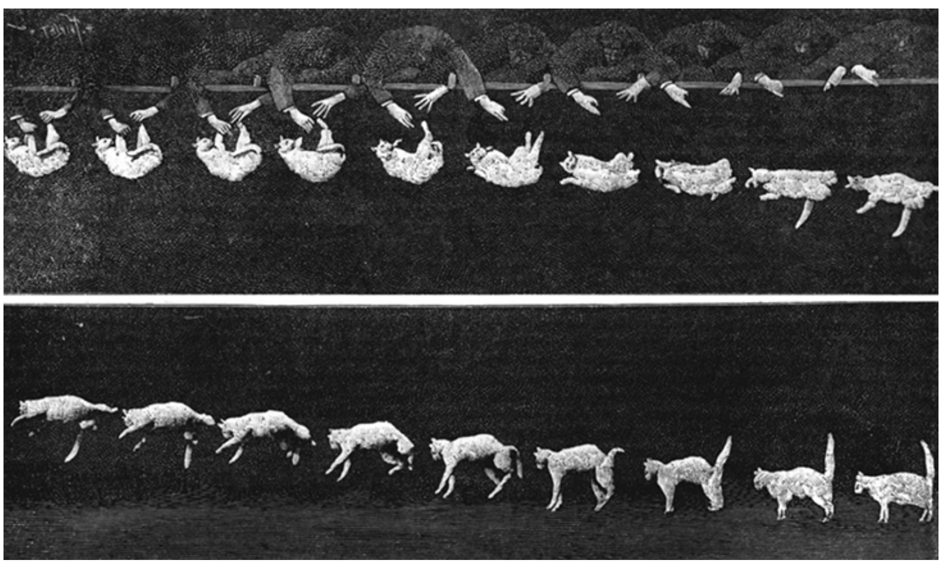


Figure 11.20 As the cat falls, its body performs complicated motions so it can land on its feet, but one point in the system moves with the simple uniform acceleration of gravity.

The problem before us, then, is to determine what part of an extended object is obeying Newton's second law when an external force is applied and to determine how the motion of the object as a whole is affected by both the internal and external forces.

Be warned: To treat this new situation correctly, we must be rigorous and completely general. We won't make any assumptions about the nature of the object, or of its constituent particles, or either the internal or external forces. Thus, the arguments will be complex.

Internal and External Forces

Suppose we have an extended object of mass *M*, made of *N* interacting particles. Let's label their masses as m_j, where

$j = 1, 2, 3, \ldots, N$. Note that

$$M = \sum_{j=1}^{N} m_j. \tag{11.16}$$

If we apply some net **external force** $\vec{\mathbf{F}}_{\text{ext}}$ on the object, every particle experiences some "share" or some fraction of that external force. Let:

$$\vec{\mathbf{f}}_j^{\text{ext}} = \text{the fraction of the external force that the } j\text{th particle experiences.}$$

Notice that these fractions of the total force are not necessarily equal; indeed, they virtually never are. (They *can* be, but they usually aren't.) In general, therefore,

$$\vec{\mathbf{f}}_1^{\text{ext}} \neq \vec{\mathbf{f}}_2^{\text{ext}} \neq \cdots \neq \vec{\mathbf{f}}_N^{\text{ext}}.$$

Next, we assume that each of the particles making up our object can interact (apply forces on) every other particle of the object. We won't try to guess what kind of forces they are; but since these forces are the result of particles of the object acting on other particles of the same object, we refer to them as **internal forces** $\vec{\mathbf{f}}_j^{\text{int}}$; thus:

$$\vec{\mathbf{f}}_j^{\text{int}} = \text{the net internal force that the } j\text{th particle experiences from all the other particles that make up the object.}$$

Now, the *net* force, internal plus external, on the *j*th particle is the vector sum of these:

$$\vec{\mathbf{f}}_j = \vec{\mathbf{f}}_j^{\text{int}} + \vec{\mathbf{f}}_j^{\text{ext}}. \tag{11.17}$$

where again, this is for all *N* particles; $j = 1, 2, 3, \ldots, N$.

As a result of this fractional force, the momentum of each particle gets changed:

$$\begin{aligned} \vec{\mathbf{f}}_j &= \frac{d\vec{\mathbf{p}}_j}{dt} \\ \vec{\mathbf{f}}_j^{\text{int}} + \vec{\mathbf{f}}_j^{\text{ext}} &= \frac{d\vec{\mathbf{p}}_j}{dt}. \end{aligned} \tag{11.18}$$

The net force $\vec{\mathbf{F}}$ on the *object* is the vector sum of these forces:

$$\begin{aligned} \vec{\mathbf{F}}_{\text{net}} &= \sum_{j=1}^{N} \left(\vec{\mathbf{f}}_{\mathbf{j}}^{\mathbf{int}} + \vec{\mathbf{f}}_{\mathbf{j}}^{\mathbf{ext}} \right) \\ &= \sum_{j=1}^{N} \vec{\mathbf{f}}_{\mathbf{j}}^{\mathbf{int}} + \sum_{j=1}^{N} \vec{\mathbf{f}}_{\mathbf{j}}^{\mathbf{ext}}. \end{aligned} \tag{11.19}$$

This net force changes the momentum of the object as a whole, and the net change of momentum of the object must be the vector sum of all the individual changes of momentum of all of the particles:

$$\vec{\mathbf{F}}_{\text{net}} = \sum_{j=1}^{N} \frac{d\vec{\mathbf{p}}_j}{dt}. \tag{11.20}$$

Combining **Equation 11.19** and **Equation 11.20** gives

$$\sum_{j=1}^{N} \vec{\mathbf{f}}_j^{\text{int}} + \sum_{j=1}^{N} \vec{\mathbf{f}}_j^{\text{ext}} = \sum_{j=1}^{N} \frac{d\vec{\mathbf{p}}_j}{dt}. \tag{11.21}$$

Let's now think about these summations. First consider the internal forces term; remember that each $\vec{\mathbf{f}}_j^{\text{int}}$ is the force on the *j*th particle from the other particles in the object. But by Newton's third law, for every one of these forces, there must be another force that has the same magnitude, but the opposite sign (points in the opposite direction). These forces do not cancel; however, that's not what we're doing in the summation. Rather, we're simply *mathematically adding up* all the internal force vectors. That is, in general, the internal forces for any individual part of the object won't cancel, but when all the internal forces are added up, the internal forces must cancel in pairs. It follows, therefore, that the sum of all the internal forces must be zero:

$$\sum_{j=1}^{N} \vec{\mathbf{f}}_j^{\,\text{int}} = 0.$$

(This argument is subtle, but crucial; take plenty of time to completely understand it.)

For the external forces, this summation is simply the total external force that was applied to the whole object:

$$\sum_{j=1}^{N} \vec{\mathbf{f}}_j^{\,\text{ext}} = \vec{\mathbf{F}}_{\text{ext}}.$$

As a result,

$$\vec{\mathbf{F}}_{\text{ext}} = \sum_{j=1}^{N} \frac{d\vec{\mathbf{p}}_j}{dt}. \tag{11.22}$$

This is an important result. **Equation 11.22** tells us that the total change of momentum of the entire object (all *N* particles) is due only to the external forces; the internal forces do not change the momentum of the object as a whole. This is why you can't lift yourself in the air by standing in a basket and pulling up on the handles: For the system of you + basket, your upward pulling force is an internal force.

Force and Momentum

Remember that our actual goal is to determine the equation of motion for the entire object (the entire system of particles). To that end, let's define:

$\vec{\mathbf{p}}_{\text{CM}} =$ the total momentum of the system of *N* particles (the reason for the subscript will become clear shortly)

Then we have

$$\vec{\mathbf{p}}_{\text{CM}} \equiv \sum_{j=1}^{N} \vec{\mathbf{p}}_j,$$

and therefore **Equation 11.22** can be written simply as

$$\vec{\mathbf{F}} = \frac{d\vec{\mathbf{p}}_{\text{CM}}}{dt}. \tag{11.23}$$

Since this change of momentum is caused by only the net external force, we have dropped the "ext" subscript.

This is Newton's second law, but now for the entire extended object. If this feels a bit anticlimactic, remember what is hiding inside it: $\vec{\mathbf{p}}_{\text{CM}}$ is the vector sum of the momentum of (in principle) hundreds of thousands of billions of billions of particles (6.02×10^{23}), all caused by one simple net external force—a force that you can calculate.

Center of Mass

Our next task is to determine what part of the extended object, if any, is obeying **Equation 11.23**.

It's tempting to take the next step; does the following equation mean anything?

$$\vec{\mathbf{F}} = M\vec{\mathbf{a}} \tag{11.24}$$

If it *does* mean something (acceleration of what, exactly?), then we could write

$$M\vec{\mathbf{a}} = \frac{d\vec{\mathbf{p}}_{\text{CM}}}{dt}$$

and thus

$$M\vec{\mathbf{a}} = \sum_{j=1}^{N} \frac{d\vec{\mathbf{p}}_j}{dt} = \frac{d}{dt}\sum_{j=1}^{N}\vec{\mathbf{p}}_j.$$

which follows because the derivative of a sum is equal to the sum of the derivatives.

Now, $\vec{\mathbf{p}}_j$ is the momentum of the *j*th particle. Defining the positions of the constituent particles (relative to some coordinate system) as $\vec{\mathbf{r}}_{\mathbf{j}} = (x_j, y_j, z_j)$, we thus have

$$\vec{\mathbf{p}}_j = m_j \vec{\mathbf{v}}_j = m_j \frac{d\vec{\mathbf{r}}_j}{dt}.$$

Substituting back, we obtain

$$\begin{aligned} M\vec{\mathbf{a}} &= \frac{d}{dt}\sum_{j=1}^{N} m_j \frac{d\vec{\mathbf{r}}_j}{dt} \\ &= \frac{d^2}{dt^2}\sum_{j=1}^{N} m_j \vec{\mathbf{r}}_j. \end{aligned}$$

Dividing both sides by M (the total mass of the extended object) gives us

$$\vec{\mathbf{a}} = \frac{d^2}{dt^2}\left(\frac{1}{M}\sum_{j=1}^{N} m_j \vec{\mathbf{r}}_j\right). \tag{11.25}$$

Thus, the point in the object that traces out the trajectory dictated by the applied force in **Equation 11.24** is inside the parentheses in **Equation 11.25**.

Looking at this calculation, notice that (inside the parentheses) we are calculating the product of each particle's mass with its position, adding all N of these up, and dividing this sum by the total mass of particles we summed. This is reminiscent of an average; inspired by this, we'll (loosely) interpret it to be the weighted average position of the mass of the extended object. It's actually called the **center of mass** of the object. Notice that the position of the center of mass has units of meters; that suggests a definition:

$$\vec{\mathbf{r}}_{\text{CM}} \equiv \frac{1}{M}\sum_{j=1}^{N} m_j \vec{\mathbf{r}}_j. \tag{11.26}$$

So, the point that obeys **Equation 11.23** (and therefore **Equation 11.24** as well) is the center of mass of the object, which is located at the position vector $\vec{\mathbf{r}}_{\text{CM}}$.

It may surprise you to learn that there does not have to be any actual mass at the center of mass of an object. For example, a hollow steel sphere with a vacuum inside it is spherically symmetrical (meaning its mass is uniformly distributed about the center of the sphere); all of the sphere's mass is out on its surface, with no mass inside. But it can be shown that the center of mass of the sphere is at its geometric center, which seems reasonable. Thus, there is no mass at the position of the center of mass of the sphere. (Another example is a doughnut.) The procedure to find the center of mass is illustrated in **Figure 11.21**.

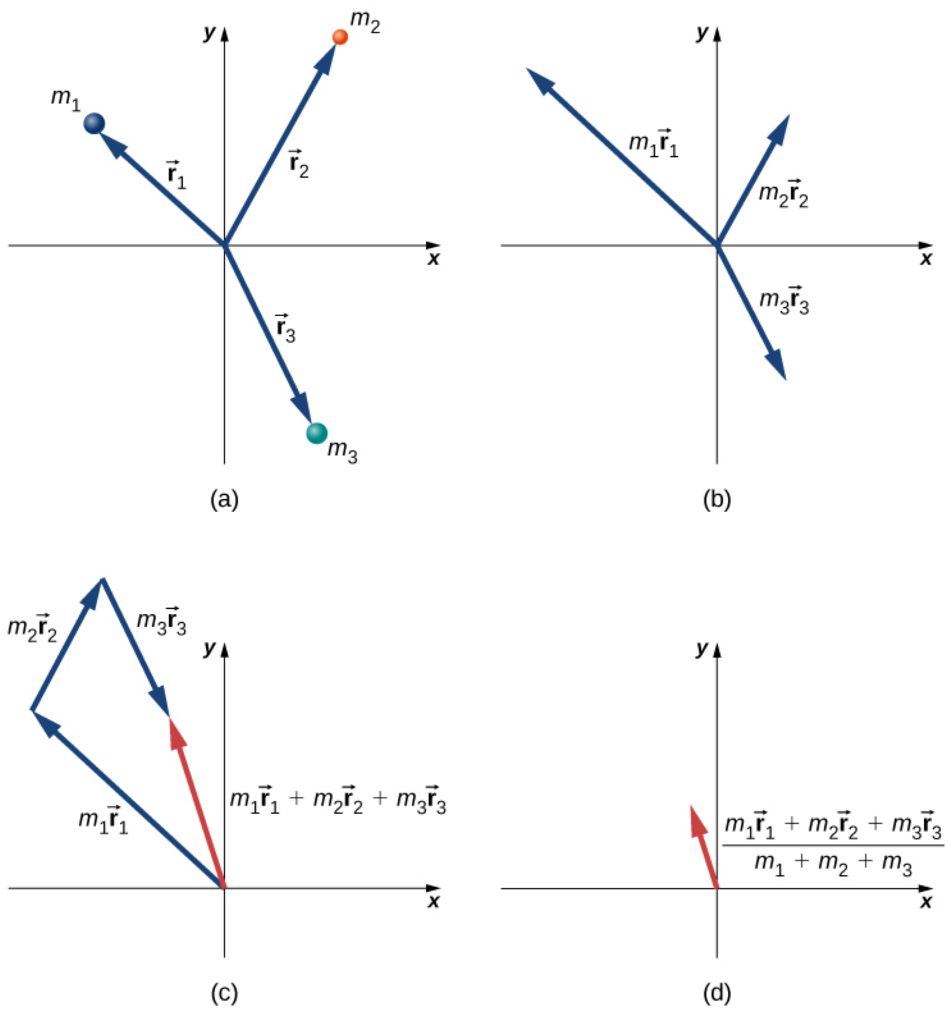


Figure 11.21 Finding the center of mass of a system of three different particles. (a) Position vectors are created for each object. (b) The position vectors are multiplied by the mass of the corresponding object. (c) The scaled vectors from part (b) are added together. (d) The final vector is divided by the total mass. This vector points to the center of mass of the system. Note that no mass is actually present at the center of mass of this system.

Since $\vec{\mathbf{r}}_j = x_j \hat{\mathbf{i}} + y_j \hat{\mathbf{j}} + z_j \hat{\mathbf{k}}$, it follows that:

$$r_{\text{CM},x} = \frac{1}{M} \sum_{j=1}^{N} m_j x_j \tag{11.27}$$

$$r_{\text{CM},y} = \frac{1}{M} \sum_{j=1}^{N} m_j y_j \tag{11.28}$$

$$r_{\text{CM},z} = \frac{1}{M} \sum_{j=1}^{N} m_j z_j \tag{11.29}$$

and thus

$$\vec{\mathbf{r}}_{\text{CM}} = r_{\text{CM},x} \hat{\mathbf{i}} + r_{\text{CM},y} \hat{\mathbf{j}} + r_{\text{CM},z} \hat{\mathbf{k}}$$
$$r_{\text{CM}} = \left| \vec{\mathbf{r}}_{\text{CM}} \right| = \left(r_{\text{CM},x}^2 + r_{\text{CM},y}^2 + r_{\text{CM},z}^2 \right)^{1/2}.$$

Therefore, you can calculate the components of the center of mass vector individually.

Finally, to complete the kinematics, the instantaneous velocity of the center of mass is calculated exactly as you might suspect:

$$\vec{\mathbf{v}}_{CM} = \frac{d}{dt}\left(\frac{1}{M}\sum_{j=1}^{N} m_j \vec{\mathbf{r}}_j\right) = \frac{1}{M}\sum_{j=1}^{N} m_j \vec{\mathbf{v}}_j \quad (11.30)$$

and this, like the position, has *x*-, *y*-, and *z*-components.

To calculate the center of mass in actual situations, we recommend the following procedure:

Problem-Solving Strategy: Calculating the Center of Mass

The center of mass of an object is a position vector. Thus, to calculate it, do these steps:

1. Define your coordinate system. Typically, the origin is placed at the location of one of the particles. This is not required, however.
2. Determine the *x*, *y*, *z*-coordinates of each particle that makes up the object.
3. Determine the mass of each particle, and sum them to obtain the total mass of the object. Note that the mass of the object at the origin *must* be included in the total mass.
4. Calculate the *x*-, *y*-, and *z*-components of the center of mass vector, using **Equation 11.27**, **Equation 11.28**, and **Equation 11.29**.
5. If required, use the Pythagorean theorem to determine its magnitude.

Here is an example that will give you a feel for what the center of mass is.

Example 11.14

Center of Mass of the Earth-Moon System

Using data from text appendix, determine how far the center of mass of the Earth-moon system is from the center of Earth. Compare this distance to the radius of Earth, and comment on the result. Ignore the other objects in the solar system.

Strategy

We get the masses and separation distance of the Earth and moon, impose a coordinate system, and use **Equation 11.26** with just $N = 2$ objects. We use a subscript "e" to refer to Earth, and subscript "m" to refer to the moon.

Solution

Define the origin of the coordinate system as the center of Earth. Then, with just two objects, **Equation 11.26** becomes

$$R = \frac{m_e r_e + m_m r_m}{m_e + m_m}.$$

From **Appendix D**,

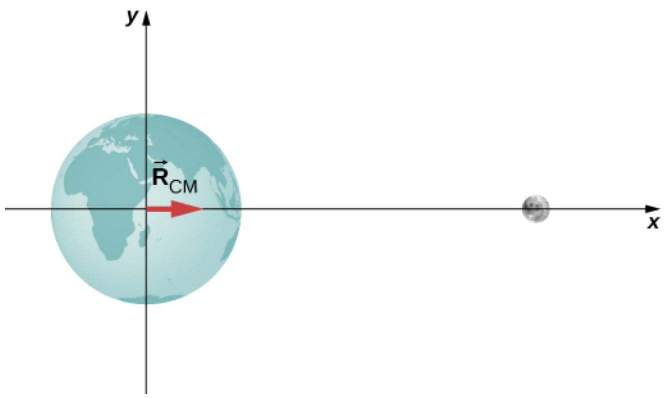
$$m_e = 5.97 \times 10^{24} \text{ kg}$$
$$m_m = 7.36 \times 10^{22} \text{ kg}$$
$$r_m = 3.82 \times 10^{8} \text{ m}.$$

We defined the center of Earth as the origin, so $r_e = 0 \text{ m}$. Inserting these into the equation for *R* gives

$$\begin{aligned} R &= \frac{\left(5.97 \times 10^{24}\ \text{kg}\right)(0\ \text{m}) + \left(7.36 \times 10^{22}\ \text{kg}\right)\left(3.82 \times 10^{8}\ \text{m}\right)}{5.97 \times 10^{24}\ \text{kg} + 7.36 \times 10^{22}\ \text{kg}} \\ &= 4.64 \times 10^{6}\ \text{m}. \end{aligned}$$

Significance

The radius of Earth is 6.37×10^6 m, so the center of mass of the Earth-moon system is (6.37 − 4.64) $\times 10^6\ \text{m} = 1.73 \times 10^6\ \text{m} = 1730\,\text{km}$ (roughly 1080 miles) *below* the surface of Earth. The location of the center of mass is shown (not to scale).



11.9 Check Your Understanding Suppose we included the sun in the system. Approximately where would the center of mass of the Earth-moon-sun system be located? (Feel free to actually calculate it.)

Two crucial concepts come out of this example:

1. As with all problems, you must define your coordinate system and origin. For center-of-mass calculations, it often makes sense to choose your origin to be located at one of the masses of your system. That choice automatically defines its distance in **Equation 11.26** to be zero. However, you must still include the mass of the object at your origin in your calculation of *M*, the total mass **Equation 11.16**. In the Earth-moon system example, this means including the mass of Earth. If you hadn't, you'd have ended up with the center of mass of the system being at the center of the moon, which is clearly wrong.
2. Had the problem been to find the location of the center of mass of the Earth-Moon system, where the masses of the Moon and Earth are more similar, note that there would be no mass at all at the location of the center of mass.

Center of Mass of Continuous Objects

If the object in question has its mass distributed uniformly in space, rather than as a collection of discrete particles, then calculus must be used, and the summation of **Equation 11.26** becomes an integral. For our purposes, suffice it to say that a regular, symmetrically shaped object will have its center of mass at the intersection of its axes of symmetry.

Center of Mass and Conservation of Momentum

How does all this connect to conservation of momentum?

Suppose you have *N* objects with masses $m_1, m_2, m_3, \ldots m_N$ and initial velocities $\vec{\mathbf{v}}_1, \vec{\mathbf{v}}_2, \vec{\mathbf{v}}_3, \ldots, \vec{\mathbf{v}}_N$. The center of mass of the objects is

$$\vec{\mathbf{r}}_{\text{CM}} = \frac{1}{M} \sum_{j=1}^{N} m_j \vec{\mathbf{r}}_j.$$

Its velocity is

$$\vec{\mathbf{v}}_{\text{CM}} = \frac{d\vec{\mathbf{r}}_{\text{CM}}}{dt} = \frac{1}{M}\sum_{j=1}^{N} m_j \frac{d\vec{\mathbf{r}}_j}{dt} \tag{11.31}$$

and thus the initial momentum of the center of mass is

$$\begin{aligned}\left[M\frac{d\vec{\mathbf{r}}_{\text{CM}}}{dt}\right]_{\text{i}} &= \sum_{j=1}^{N} m_j \frac{d\vec{\mathbf{r}}_{j,\text{i}}}{dt} \\ M\vec{\mathbf{v}}_{\text{CM,i}} &= \sum_{j=1}^{N} m_j \vec{\mathbf{v}}_{j,\text{i}}.\end{aligned}$$

After these masses move and interact with each other, the momentum of the center of mass is

$$M\vec{\mathbf{v}}_{\text{CM,f}} = \sum_{j=1}^{N} m_j \vec{\mathbf{v}}_{j,\text{f}}.$$

But conservation of momentum tells us that the right-hand side of both equations must be equal, which says

$$M\vec{\mathbf{v}}_{\text{CM,f}} = M\vec{\mathbf{v}}_{\text{CM,i}}. \tag{11.32}$$

This result implies that conservation of momentum is expressed in terms of the center of mass of the system. Notice that as an object moves through space with no net external force acting on it, an individual particle of the object may accelerate in various directions, with various magnitudes, depending on the net internal force acting on that object at any time. (Remember, it is only the vector sum of all the internal forces that vanishes, not the internal force on a single particle.) Thus, such a particle's momentum will not be constant—but the momentum of the entire extended object will be, in accord with **Equation 11.32**.

Equation 11.32 implies another important result: Since M represents the mass of the entire system of particles, it is necessarily constant. (If it isn't, we don't have a closed system, so we can't expect the system's momentum to be conserved.) As a result, **Equation 11.32** implies that, for a closed system,

$$\vec{\mathbf{v}}_{\text{CM,f}} = \vec{\mathbf{v}}_{\text{CM,i}}. \tag{11.33}$$

That is to say, *in the absence of an external force, the velocity of the center of mass never changes*.

You might be tempted to shrug and say, "Well yes, that's just Newton's first law," but remember that Newton's first law discusses the constant velocity of a particle, whereas **Equation 11.33** applies to the center of mass of a (possibly vast) collection of interacting particles, and that there may not be any particle at the center of mass at all! So, this really is a remarkable result.

Example 11.15

Fireworks Display

When a fireworks rocket explodes, thousands of glowing fragments fly outward in all directions, and fall to Earth in an elegant and beautiful display (**Figure 11.22**). Describe what happens, in terms of conservation of momentum and center of mass.

Figure 11.22 These exploding fireworks are a vivid example of conservation of momentum and the motion of the center of mass.

The picture shows radial symmetry about the central points of the explosions; this suggests the idea of center of mass. We can also see the parabolic motion of the glowing particles; this brings to mind projectile motion ideas.

Solution

Initially, the fireworks rocket is launched and flies more or less straight upward; this is the cause of the more-or-less-straight, white trail going high into the sky below the explosion in the upper-right of the picture (the yellow explosion). This trail is not parabolic because the explosive shell, during its launch phase, is actually a rocket; the impulse applied to it by the ejection of the burning fuel applies a force on the shell during the rise-time interval. (This is a phenomenon we will study in the next section.) The shell has multiple forces on it; thus, it is not in free-fall prior to the explosion.

At the instant of the explosion, the thousands of glowing fragments fly outward in a radially symmetrical pattern. The symmetry of the explosion is the result of all the internal forces summing to zero $\left(\sum_j \vec{\mathbf{f}}_j^{\text{int}} = 0\right)$; for every internal force, there is another that is equal in magnitude and opposite in direction.

However, as we learned above, these internal forces cannot change the momentum of the center of mass of the (now exploded) shell. Since the rocket force has now vanished, the center of mass of the shell is now a projectile (the only force on it is gravity), so its trajectory does become parabolic. The two red explosions on the left show the path of their centers of mass at a slightly longer time after explosion compared to the yellow explosion on the upper right.

In fact, if you look carefully at all three explosions, you can see that the glowing trails are not truly radially symmetric; rather, they are somewhat denser on one side than the other. Specifically, the yellow explosion and the lower middle explosion are slightly denser on their right sides, and the upper-left explosion is denser on its left side. This is because of the momentum of their centers of mass; the differing trail densities are due to the momentum each piece of the shell had at the moment of its explosion. The fragment for the explosion on the upper left of the picture had a momentum that pointed upward and to the left; the middle fragment's momentum pointed upward and slightly to the right; and the right-side explosion clearly upward and to the right (as evidenced by the white rocket exhaust trail visible below the yellow explosion).

Finally, each fragment is a projectile on its own, thus tracing out thousands of glowing parabolas.

Significance

In the discussion above, we said, "…the center of mass of the shell is now a projectile (the only force on it is gravity)…." This is not quite accurate, for there may not be any mass at all at the center of mass; in which case, there could not be a force acting on it. This is actually just verbal shorthand for describing the fact that the gravitational forces on all the particles act so that the center of mass changes position exactly as if all the mass of the shell were always located at the position of the center of mass.

11.10 Check Your Understanding How would the firework display change in deep space, far away from any source of gravity?

You may sometimes hear someone describe an explosion by saying something like, "the fragments of the exploded object always move in a way that makes sure that the center of mass continues to move on its original trajectory." This makes it sound as if the process is somewhat magical: how can it be that, in *every* explosion, it *always* works out that the fragments move in just the right way so that the center of mass' motion is unchanged? Phrased this way, it would be hard to believe no explosion ever does anything differently.

The explanation of this apparently astonishing coincidence is: We defined the center of mass precisely so this is exactly what we would get. Recall that first we defined the momentum of the system:

$$\vec{\mathbf{p}}_{\text{CM}} = \sum_{j=1}^{N} \frac{d\vec{\mathbf{p}}_j}{dt}.$$

We then concluded that the net external force on the system (if any) changed this momentum:

$$\vec{\mathbf{F}} = \frac{d\vec{\mathbf{p}}_{\text{CM}}}{dt}$$

and then—and here's the point—we defined an acceleration that would obey Newton's second law. That is, we demanded that we should be able to write

$$\vec{\mathbf{a}} = \frac{\vec{\mathbf{F}}}{M}$$

which requires that

$$\vec{\mathbf{a}} = \frac{d^2}{dt^2}\left(\frac{1}{M}\sum_{j=1}^{N} m_j \vec{\mathbf{r}}_j\right).$$

where the quantity inside the parentheses is the center of mass of our system. So, it's not astonishing that the center of mass obeys Newton's second law; we defined it so that it would.

CHAPTER 11 REVIEW

KEY TERMS

center of mass weighted average position of the mass

closed system system for which the mass is constant and the net external force on the system is zero

elastic collision that conserves kinetic energy

explosion single object breaks up into multiple objects; kinetic energy is not conserved in explosions

external force force applied to an extended object that changes the momentum of the extended object as a whole

impulse effect of applying a force on a system for a time interval; this time interval is usually small, but does not have to be

impulse-momentum theorem change of momentum of a system is equal to the impulse applied to the system

inelastic collision that does not conserve kinetic energy

internal force force that the simple particles that make up an extended object exert on each other. Internal forces can be attractive or repulsive

Law of Conservation of Momentum total momentum of a closed system cannot change

linear mass density λ, expressed as the number of kilograms of material per meter

momentum measure of the quantity of motion that an object has; it takes into account both how fast the object is moving, and its mass; specifically, it is the product of mass and velocity; it is a vector quantity

perfectly inelastic collision after which all objects are motionless, the final kinetic energy is zero, and the loss of kinetic energy is a maximum

system object or collection of objects whose motion is currently under investigation; however, your system is defined at the start of the problem, you must keep that definition for the entire problem

KEY EQUATIONS

Linear momentum	$\vec{\mathbf{p}} = m\vec{\mathbf{v}}.$
Definition of impulse	$d\vec{\mathbf{J}} \equiv \vec{\mathbf{F}}(t)dt.$
Impulse-Momentum theorem	$\vec{\mathbf{J}} = \Delta\vec{\mathbf{p}}.$
Momentum and force	$\vec{\mathbf{F}}_{\text{ave}} = \frac{\Delta\vec{\mathbf{p}}}{\Delta t}.$
Conservation of momentum	$\sum_{j=1}^{N} \vec{\mathbf{p}}_{\mathbf{j}} = \text{constant}.$
Center of mass	$\vec{\mathbf{r}}_{\text{CM}} \equiv \frac{1}{M}\sum_{j=1}^{N} m_j \vec{\mathbf{r}}_j.$

SUMMARY

11.1 Linear Momentum

- The motion of an object depends on its mass as well as its velocity. Momentum is a concept that describes this. It is a useful and powerful concept, both computationally and theoretically. The SI unit for momentum is kg · m/s.

11.2 Impulse and Collisions

- When a force is applied on an object for some amount of time, the object experiences an impulse.
- This impulse is equal to the object's change of momentum.
- Newton's second law in terms of momentum states that the net force applied to a system equals the rate of change of the momentum that the force causes.

11.3 Conservation of Linear Momentum

- The law of conservation of momentum says that the momentum of a closed system is constant in time (conserved).
- A closed (or isolated) system is defined to be one for which the mass remains constant, and the net external force is zero.
- The total momentum of a system is conserved *only* when the system is closed.

11.4 Types of Collisions

- An elastic collision is one that conserves kinetic energy.
- An inelastic collision does not conserve kinetic energy.
- Momentum is conserved regardless of whether or not kinetic energy is conserved.
- Analysis of kinetic energy changes and conservation of momentum together allow the final velocities to be calculated in terms of initial velocities and masses in one-dimensional, two-body collisions.

11.5 Center of Mass

- An extended object (made up of many objects) has a defined position vector called the center of mass.
- The center of mass can be thought of, loosely, as the average location of the total mass of the object.
- The center of mass of an object traces out the trajectory dictated by Newton's second law, due to the net external force.
- The internal forces within an extended object cannot alter the momentum of the extended object as a whole.

CONCEPTUAL QUESTIONS

11.1 Linear Momentum

1. An object that has a small mass and an object that has a large mass have the same momentum. Which object has the largest kinetic energy?

2. An object that has a small mass and an object that has a large mass have the same kinetic energy. Which mass has the largest momentum?

11.2 Impulse and Collisions

3. Is it possible for a small force to produce a larger impulse on a given object than a large force? Explain.

4. Why is a 10-m fall onto concrete far more dangerous than a 10-m fall onto water?

5. What external force is responsible for changing the momentum of a car moving along a horizontal road?

6. A piece of putty and a tennis ball with the same mass are thrown against a wall with the same velocity. Which object experience a greater impulse from the wall or are the impulses equal? Explain.

11.3 Conservation of Linear Momentum

7. Under what circumstances is momentum conserved?

8. Can momentum be conserved for a system if there are external forces acting on the system? If so, under what conditions? If not, why not?

9. Explain in terms of momentum and Newton's laws how a car's air resistance is due in part to the fact that it pushes air in its direction of motion.

10. Can objects in a system have momentum while the momentum of the system is zero? Explain your answer.

11. A sprinter accelerates out of the starting blocks. Can you consider him as a closed system? Explain.

12. A rocket in deep space (zero gravity) accelerates by firing hot gas out of its thrusters. Does the rocket constitute a closed system? Explain.

11.4 Types of Collisions

13. Two objects of equal mass are moving with equal and opposite velocities when they collide. Can all the kinetic energy be lost in the collision?

14. Describe a system for which momentum is conserved but mechanical energy is not. Now the reverse: Describe a system for which kinetic energy is conserved but momentum is not.

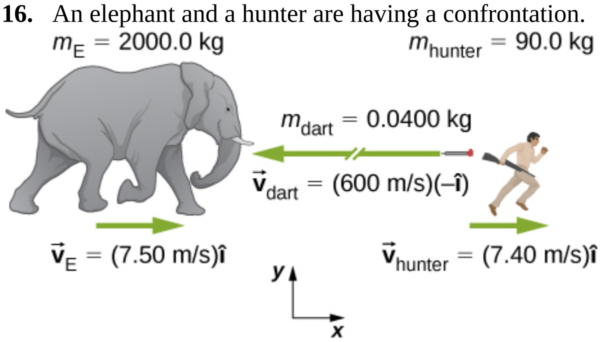
11.5 Center of Mass

15. Suppose a fireworks shell explodes, breaking into three large pieces for which air resistance is negligible. How does the explosion affect the motion of the center of mass? How would it be affected if the pieces experienced significantly more air resistance than the intact shell?

PROBLEMS

11.1 Linear Momentum

16. An elephant and a hunter are having a confrontation.

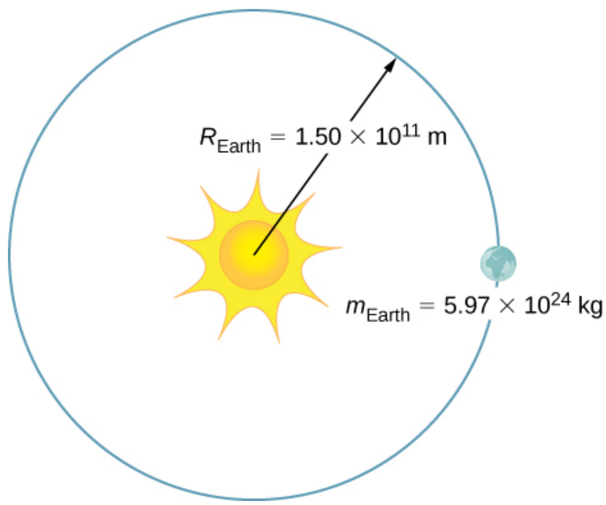


a. Calculate the momentum of the 2000.0-kg elephant charging the hunter at a speed of 7.50 m/s.
b. Calculate the ratio of the elephant's momentum to the momentum of a 0.0400-kg tranquilizer dart fired at a speed of 600 m/s.
c. What is the momentum of the 90.0-kg hunter running at 7.40 m/s after missing the elephant?

17. A skater of mass 40 kg is carrying a box of mass 5 kg. The skater has a speed of 5 m/s with respect to the floor and is gliding without any friction on a smooth surface.
a. Find the momentum of the box with respect to the floor.
b. Find the momentum of the box with respect to the floor after she puts the box down on the frictionless skating surface.

18. A car of mass 2000 kg is moving with a constant velocity of 10 m/s due east. What is the momentum of the car?

19. The mass of Earth is 5.97×10^{24} kg and its orbital radius is an average of 1.50×10^{11} m. Calculate the magnitude of its linear momentum at the location in the diagram.



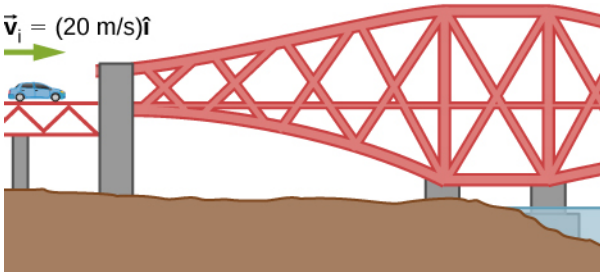
20. If a rainstorm drops 1 cm of rain over an area of 10 km^2 in the period of 1 hour, what is the momentum of the rain that falls in one second? Assume the terminal velocity of a raindrop is 10 m/s.

21. What is the average momentum of an avalanche that moves a 40-cm-thick layer of snow over an area of 100 m by 500 m over a distance of 1 km down a hill in 5.5 s? Assume a density of 350 kg/m^3 for the snow.

22. What is the average momentum of a 70.0-kg sprinter who runs the 100-m dash in 9.65 s?

11.2 Impulse and Collisions

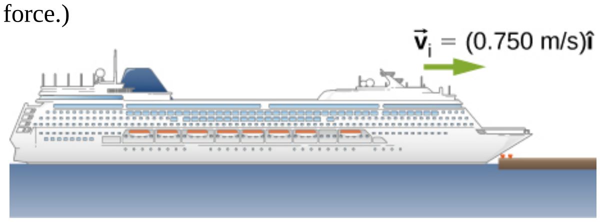
23. A 75.0-kg person is riding in a car moving at 20.0 m/s when the car runs into a bridge abutment (see the following figure).



a. Calculate the average force on the person if he is stopped by a padded dashboard that compresses an average of 1.00 cm.
b. Calculate the average force on the person if he is stopped by an air bag that compresses an average of 15.0 cm.

24. One hazard of space travel is debris left by previous missions. There are several thousand objects orbiting Earth that are large enough to be detected by radar, but there are far greater numbers of very small objects, such as flakes of paint. Calculate the force exerted by a 0.100-mg chip of paint that strikes a spacecraft window at a relative speed of 4.00×10^3 m/s, given the collision lasts 6.00×10^{-8} s.

25. A cruise ship with a mass of 1.00×10^7 kg strikes a pier at a speed of 0.750 m/s. It comes to rest after traveling 6.00 m, damaging the ship, the pier, and the tugboat captain's finances. Calculate the average force exerted on the pier using the concept of impulse. (*Hint*: First calculate the time it took to bring the ship to rest, assuming a constant force.)



26. Calculate the final speed of a 110-kg rugby player who is initially running at 8.00 m/s but collides head-on with a padded goalpost and experiences a backward force of 1.76×10^4 N for 5.50×10^{-2} s.

27. Water from a fire hose is directed horizontally against a wall at a rate of 50.0 kg/s and a speed of 42.0 m/s. Calculate the force exerted on the wall, assuming the water's horizontal momentum is reduced to zero.

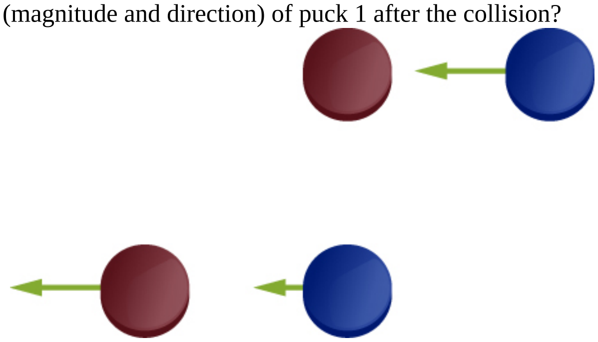
28. A 0.450-kg hammer is moving horizontally at 7.00 m/s when it strikes a nail and comes to rest after driving the nail 1.00 cm into a board. Assume constant acceleration of the hammer-nail pair.
a. Calculate the duration of the impact.
b. What was the average force exerted on the nail?

29. What is the momentum (as a function of time) of a 5.0-kg particle moving with a velocity $\vec{v}(t) = \left(2.0\,\hat{i} + 4.0t\,\hat{j}\right)$ m/s? What is the net force acting on this particle?

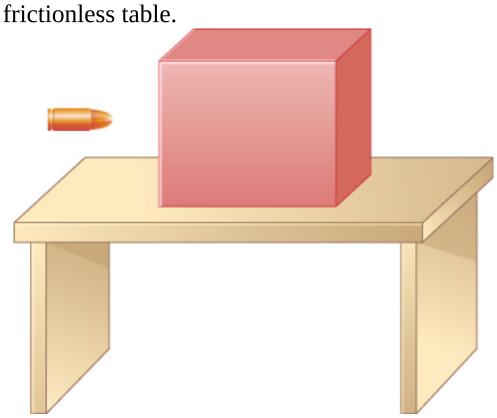
11.3 Conservation of Linear Momentum

30. Train cars are coupled together by being bumped into one another. Suppose two loaded train cars are moving toward one another, the first having a mass of 1.50×10^5 kg and a velocity of $(0.30\text{ m/s})\,\hat{i}$, and the second having a mass of 1.10×10^5 kg and a velocity of $-(0.12\text{ m/s})\,\hat{i}$. What is their final velocity?

31. Two identical pucks collide elastically on an air hockey table. Puck 1 was originally at rest; puck 2 has an incoming speed of 6.00 m/s and scatters at an angle of $30°$ with respect to its incoming direction. What is the velocity (magnitude and direction) of puck 1 after the collision?



32. The figure below shows a bullet of mass 200 g traveling horizontally towards the east with speed 400 m/s, which strikes a block of mass 1.5 kg that is initially at rest on a frictionless table.



After striking the block, the bullet is embedded in the block and the block and the bullet move together as one unit.

a. What is the magnitude and direction of the velocity of the block/bullet combination immediately after the impact?
b. What is the magnitude and direction of the impulse by the block on the bullet?
c. What is the magnitude and direction of the impulse from the bullet on the block?
d. If it took 3 ms for the bullet to change the speed from 400 m/s to the final speed after impact, what is the average force between the block and the bullet during this time?

33. A 20-kg child is coasting at 3.3 m/s over flat ground in a 4.0-kg wagon. The child drops a 1.0-kg ball out the back of the wagon. What is the final speed of the child and wagon?

34. A 4.5 kg puffer fish expands to 40% of its mass by taking in water. When the puffer fish is threatened, it releases the water toward the threat to move quickly forward. What is the ratio of the speed of the puffer fish forward to the speed of the expelled water backwards?

35. Explain why a cannon recoils when it fires a shell.

36. Two figure skaters are coasting in the same direction, with the leading skater moving at 5.5 m/s and the trailing skating moving at 6.2 m/s. When the trailing skater catches up with the leading skater, he picks her up without applying any horizontal forces on his skates. If the trailing skater is 50% heavier than the 50-kg leading skater, what is their speed after he picks her up?

37. A 2000-kg railway freight car coasts at 4.4 m/s underneath a grain terminal, which dumps grain directly down into the freight car. If the speed of the loaded freight car must not go below 3.0 m/s, what is the maximum mass of grain that it can accept?

11.4 Types of Collisions

38. A 90.0-kg ice hockey player hits a 0.150-kg puck, giving the puck a velocity of 45.0 m/s. If both are initially at rest and if the ice is frictionless, how far does the player recoil in the time it takes the puck to reach the goal 15.0 m away?

39. A 100-g firecracker is launched vertically into the air and explodes into two pieces at the peak of its trajectory. If a 72-g piece is projected horizontally to the left at 20 m/s, what is the speed and direction of the other piece?

40. In an elastic collision, a 400-kg bumper car collides directly from behind with a second, identical bumper car that is traveling in the same direction. The initial speed of the leading bumper car is 5.60 m/s and that of the trailing car is 6.00 m/s. Assuming that the mass of the drivers is much, much less than that of the bumper cars, what are their final speeds?

41. Repeat the preceding problem if the mass of the leading bumper car is 30.0% greater than that of the trailing bumper car.

42. An alpha particle (^{4}He) undergoes an elastic collision with a stationary uranium nucleus (^{235}U). What percent of the kinetic energy of the alpha particle is transferred to the uranium nucleus? Assume the collision is one-dimensional.

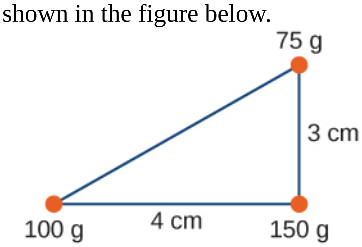
43. You are standing on a very slippery icy surface and throw a 1-kg football horizontally at a speed of 6.7 m/s. What is your velocity when you release the football? Assume your mass is 65 kg.

44. A 35-kg child rides a relatively massless sled down a hill and then coasts along the flat section at the bottom, where a second 35-kg child jumps on the sled as it passes by her. If the speed of the sled is 3.5 m/s before the second child jumps on, what is its speed after she jumps on?

45. A boy sleds down a hill and onto a frictionless ice-covered lake at 10.0 m/s. In the middle of the lake is a 1000-kg boulder. When the sled crashes into the boulder, he is propelled backwards from the boulder. The collision is an elastic collision. If the boy's mass is 40.0 kg and the sled's mass is 2.50 kg, what is the speed of the sled and the boulder after the collision?

11.5 Center of Mass

46. Three point masses are placed at the corners of a triangle as shown in the figure below.



Find the center of mass of the three-mass system.

47. Two particles of masses m_1 and m_2 move uniformly in different circles of radii R_1 and R_2 about the origin in the x, y-plane. The coordinates of the two particles in meters are given as follows ($z = 0$ for both). Here t is in seconds:

$$x_1(t) = 4\cos(2t)$$
$$y_1(t) = 4\sin(2t)$$
$$x_2(t) = 2\cos\left(3t - \frac{\pi}{2}\right)$$
$$y_2(t) = 2\sin\left(3t - \frac{\pi}{2}\right)$$

a. Find the radii of the circles of motion of both particles.
b. Find the *x*- and *y*-coordinates of the center of mass.
c. Decide if the center of mass moves in a circle by plotting its trajectory.

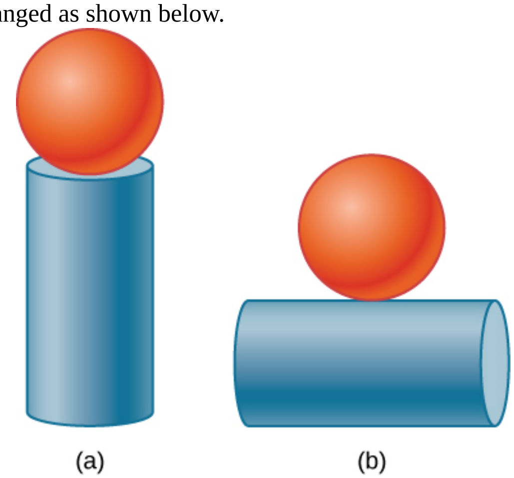
48. Find the center of mass of a one-meter long rod, made of 50 cm of iron (density $8\,\frac{\text{g}}{\text{cm}^3}$) and 50 cm of aluminum (density $2.7\,\frac{\text{g}}{\text{cm}^3}$).

49. Find the center of mass of a cone of uniform density that has a radius *R* at the base, height *h*, and mass *M*. Let the origin be at the center of the base of the cone and have +*z* going through the cone vertex.

50. Find the center of mass of a thin wire of mass *m* and length *L* bent in a semicircular shape. Let the origin be at the center of the semicircle and have the wire arc from the +*x* axis, cross the +*y* axis, and terminate at the −*x* axis.

51. Find the center of mass of a uniform thin semicircular plate of radius *R*. Let the origin be at the center of the semicircle, the plate arc from the +*x* axis to the −x axis, and the *z* axis be perpendicular to the plate.

52. Find the center of mass of a sphere of mass *M* and radius *R* and a cylinder of mass *m*, radius *r*, and height *h* arranged as shown below.



Express your answers in a coordinate system that has the origin at the center of the cylinder.

12 | ANGULAR MOMENTUM

Figure 12.1 A helicopter has its main lift blades rotating to keep the aircraft airborne. Due to conservation of angular momentum, the body of the helicopter would want to rotate in the opposite sense to the blades, if it were not for the small rotor on the tail of the aircraft, which provides thrust to stabilize it.

Chapter Outline

Introduction

Angular momentum is the rotational counterpart of linear momentum. Any massive object that rotates about an axis carries angular momentum, including rotating flywheels, planets, stars, hurricanes, tornadoes, whirlpools, and so on. The helicopter shown in the chapter-opening picture can be used to illustrate the concept of angular momentum. The lift blades spin about a vertical axis through the main body and carry angular momentum. The body of the helicopter tends to rotate in the opposite sense in order to conserve angular momentum. The small rotors at the tail of the aircraft provide a counter thrust against the body to prevent this from happening, and the helicopter stabilizes itself. The concept of conservation of angular momentum is discussed later in this chapter. In the main part of this chapter, we explore the intricacies of angular momentum of rigid bodies such as a top, and also of point particles and systems of particles.

12.1 | Angular Momentum

Learning Objectives

By the end of this section, you will be able to:

- Determine the vector sign of angular momentum of an object in orbit or rotation
- Find the total angular momentum about a designated origin of a system of particles
- Calculate the angular momentum of a rigid body rotating about a fixed axis
- Use conservation of angular momentum in the analysis of objects that change their rotation rate

Why does Earth keep on spinning? What started it spinning to begin with? Why doesn't Earth's gravitational attraction not bring the Moon crashing in toward Earth? And how does an ice skater manage to spin faster and faster simply by pulling her arms in? Why does she not have to exert a torque to spin faster? Questions like these have answers based in **angular momentum**, the rotational analog to linear momentum.

A complete determination of the angular momentum of a rigid body involves:

- Dividing the extended body into infinitesimal pieces of mass;
- Calculating the velocity vector of each piece of mass relative to the origin defined by the rotation axis;
- Taking the vector cross product of each of the resulting momentum vectors with its corresponding displacement vector relative to the origin; and
- integrating these results over the entire volume of the rigid body.

We present here, without the need for the full derivation, two simple definitions, one for the angular momentum of a single particle, the other for the angular momentum of a rotating rigid body.

Angular Momentum of a Particle in Circular Motion

For a particle of mass m orbiting about a center at radius r at tangential velocity v, its angular momentum is

$$L = mvr \tag{12.1}$$

The angular momentum is a vector quantity, and its direction is considered to be positive for counterclockwise motion and negative for clockwise motion.

For a collection of particles, orbiting together, the total angular momentum of the system is just the sum of the individual angular momenta.

$$\vec{\mathbf{L}} = \sum_i \vec{\mathbf{L}}_i \tag{12.2}$$

Angular Momentum of a Rigid Body

The net angular momentum of the rigid body which has moment of inertia I and angular velocity ω about the axis of rotation is

$$L = I\omega. \tag{12.3}$$

This equation is analogous to the magnitude of the linear momentum $p = mv$. The direction of the angular momentum vector is directed along the axis of rotation given by the right-hand rule. Counterclockwise rotations have positive angular momenta, whereas clockwise rotations have negative angular momenta.

Example 12.1

Angular Momentum of a Robot Arm

A robot arm on a Mars rover like *Curiosity* shown in **Figure 7.1** is 1.0 m long and has forceps at the free end to pick up rocks. The mass of the arm is 2.0 kg and the mass of the forceps is 1.0 kg. See **Figure 12.2**. The robot arm and forceps move from rest to $\omega = 0.1\pi$ rad/s in 0.1 s. It rotates down and picks up a Mars rock that has mass 1.5 kg. The axis of rotation is the point where the robot arm connects to the rover. (a) What is the angular momentum of the robot arm by itself about the axis of rotation after 0.1 s when the arm has stopped accelerating? (b) What is the angular momentum of the robot arm when it has the Mars rock in its forceps and is rotating upwards

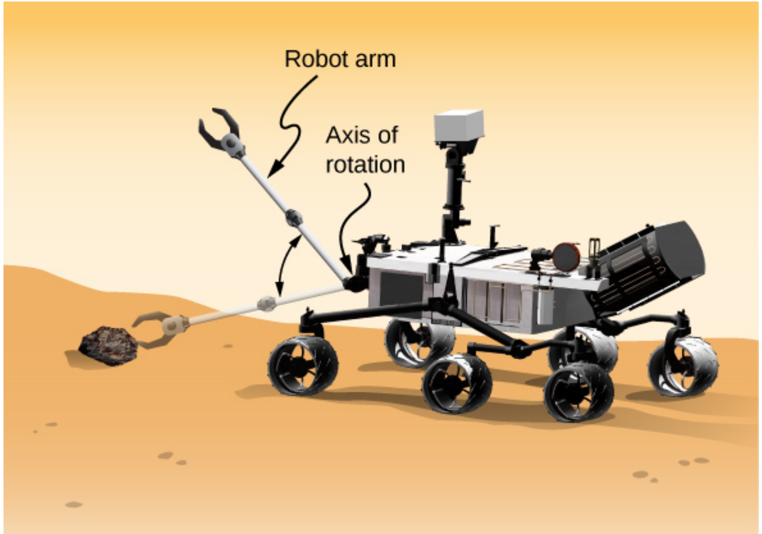


Figure 12.2 A robot arm on a Mars rover swings down and picks up a Mars rock. (credit: modification of work by NASA/JPL-Caltech)

Strategy

We use **Equation 12.3** to find angular momentum in the various configurations. When the arm is rotating downward, the right-hand rule gives the angular momentum vector directed out of the page, which we will call the positive *z*-direction. When the arm is rotating upward, the right-hand rule gives the direction of the angular momentum vector into the page or in the negative *z*-direction. The moment of inertia is the sum of the individual moments of inertia. The arm can be approximated with a solid rod, and the forceps and Mars rock can be approximated as point masses located at a distance of 1 m from the origin.

Solution

a. Writing down the individual moments of inertia, we have

Robot arm: $I_R = \frac{1}{3} m_R r^2 = \frac{1}{3}(2.00\,\text{kg})(1.00\,\text{m})^2 = \frac{2}{3}\text{kg}\cdot\text{m}^2$.

Forceps: $I_F = m_F r^2 = (1.0\,\text{kg})(1.0\,\text{m})^2 = 1.0\,\text{kg}\cdot\text{m}^2$.

Mars rock: $I_{MR} = m_{MR} r^2 = (1.5\,\text{kg})(1.0\,\text{m})^2 = 1.5\,\text{kg}\cdot\text{m}^2$.

Therefore, without the Mars rock, the total moment of inertia is

$$I_{\text{Total}} = I_R + I_F = 1.67\,\text{kg}\cdot\text{m}^2$$

and the magnitude of the angular momentum is

$$L = I\omega = 1.67\,\text{kg}\cdot\text{m}^2(0.1\pi\,\text{rad/s}) = 0.17\pi\,\text{kg}\cdot\text{m}^2/\text{s}.$$

The angular momentum vector is directed out of the page in the $\hat{\mathbf{k}}$ direction since the robot arm is rotating counterclockwise.

b. We must include the Mars rock in the calculation of the moment of inertia, so we have

$$I_{\text{Total}} = I_R + I_F + I_{MR} = 3.17\,\text{kg}\cdot\text{m}^2$$

and

$$L = I\omega = 3.17\,\text{kg}\cdot\text{m}^2(0.1\pi\,\text{rad/s}) = 0.32\pi\,\text{kg}\cdot\text{m}^2/\text{s}.$$

And, the angular momentum vector is negative, since the robot arm is now rotating clockwise.

Significance

The angular momentum in (a) is less than that of (b) due to the fact that the moment of inertia in (b) is greater than (a), while the angular velocity is the same.

12.1 Check Your Understanding Which has greater angular momentum: a solid sphere of mass *m* rotating at a constant angular frequency ω_0 about the *z*-axis, or a solid cylinder of same mass and rotation rate about the *z*-axis?

Angular Momentum and Rotational Kinetic Energy

In **Section 10.2** we saw that the kinetic energy of a rotating rigid body is given by

$$K = \frac{1}{2}I\omega^2 \tag{12.4}$$

Having now defined the angular momentum of the rotating rigid body in **Equation 12.3**, we can write an alternate expression for the kinetic energy as

$$K = \frac{L^2}{2I} \tag{12.5}$$

Visit the **University of Colorado's Interactive Simulation of Angular Momentum (https://openstax.org/l/21angmomintsim)** to learn more about angular momentum.

12.2 | Conservation of Angular Momentum

Learning Objectives

By the end of this section, you will be able to:

- Apply conservation of angular momentum to determine the angular velocity of a rotating system in which the moment of inertia is changing
- Explain how the rotational kinetic energy changes when a system undergoes changes in both moment of inertia and angular velocity

So far, we have looked at the angular momentum of systems consisting of point particles and rigid bodies. If the body or system of particles we are examining is completely isolated from any external forces, or even if there are forces present but there is no net external torque on it, then there is **conservation of angular momentum**.

Law of Conservation of Angular Momentum

The angular momentum of a system of particles around a point in a fixed inertial reference frame is conserved if there is no net external torque around that point:

$$\frac{d\vec{\mathbf{L}}}{dt} = 0 \tag{12.6}$$

or

$$\vec{\mathbf{L}} = \vec{\mathbf{L}}_1 + \vec{\mathbf{L}}_2 + \cdots + \vec{\mathbf{L}}_N = \text{constant.} \tag{12.7}$$

Note that the *total* angular momentum $\vec{\mathbf{L}}$ is conserved. Any of the individual angular momenta can change as long as their sum remains constant. This law is analogous to linear momentum being conserved when the external force on a system is zero.

As an example of conservation of angular momentum, **Figure 12.3** shows an ice skater executing a spin. The net torque on her is very close to zero because there is relatively little friction between her skates and the ice. Also, the friction is exerted very close to the pivot point, giving it almost no lever arm.Consequently, she can spin for quite some time. She can also increase her rate of spin by pulling her arms and legs in. Why does pulling her arms and legs in increase her rate of spin? The answer is that her angular momentum is constant, so that

$$L' = L$$

or

$$I'\omega' = I\omega,$$

where the primed quantities refer to conditions after she has pulled in her arms and reduced her moment of inertia. Because I' is smaller, the angular velocity ω' must increase to keep the angular momentum constant.

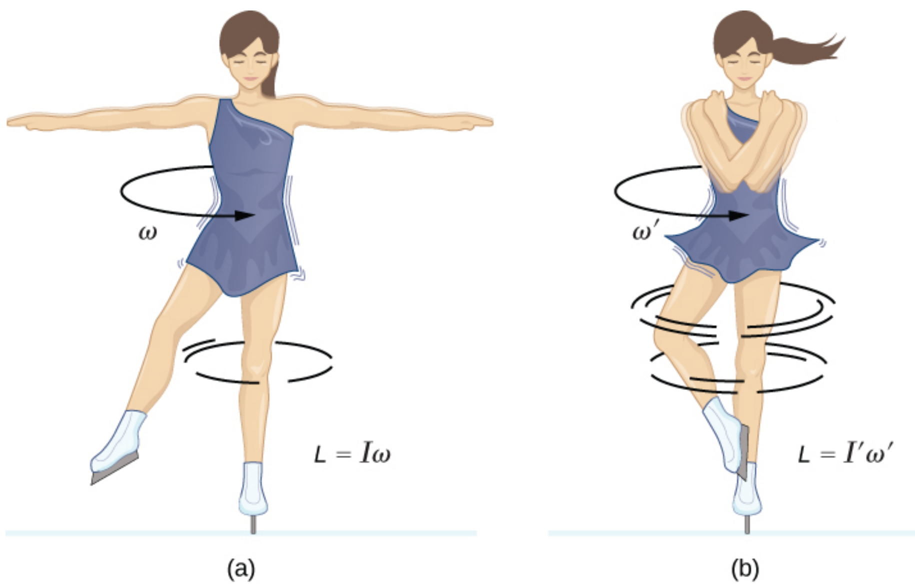


Figure 12.3 (a) An ice skater is spinning on the tip of her skate with her arms extended. Her angular momentum is conserved because the net torque on her is negligibly small. (b) Her rate of spin increases greatly when she pulls in her arms, decreasing her moment of inertia. The work she does to pull in her arms results in an increase in rotational kinetic energy.

It is interesting to see how the rotational kinetic energy of the skater changes when she pulls her arms in. Her initial rotational energy is

$$K_{\text{Rot}} = \frac{1}{2}I\omega^2,$$

whereas her final rotational energy is

$$K'_{\text{Rot}} = \frac{1}{2}I'(\omega')^2.$$

Since $I'\omega' = I\omega,$ we can substitute for ω' and find

$$K'_{\text{Rot}} = \frac{1}{2}I'(\omega')^2 = \frac{1}{2}I'\left(\frac{I}{I'}\omega\right)^2 = \frac{1}{2}I\omega^2\left(\frac{I}{I'}\right) = K_{\text{Rot}}\left(\frac{I}{I'}\right).$$

Because her moment of inertia has decreased, $I' < I,$ her final rotational kinetic energy has increased. The source of this additional rotational kinetic energy is the work required to pull her arms inward. Note that the skater's arms do not move in a perfect circle—they spiral inward. This work causes an increase in the rotational kinetic energy, while her angular momentum remains constant. Since she is in a frictionless environment, no energy escapes the system. Thus, if she were to extend her arms to their original positions, she would rotate at her original angular velocity and her kinetic energy would return to its original value.

The solar system is another example of how conservation of angular momentum works in our universe. Our solar system was born from a huge cloud of gas and dust that initially had rotational energy. Gravitational forces caused the cloud to contract, and the rotation rate increased as a result of conservation of angular momentum (**Figure 12.4**).

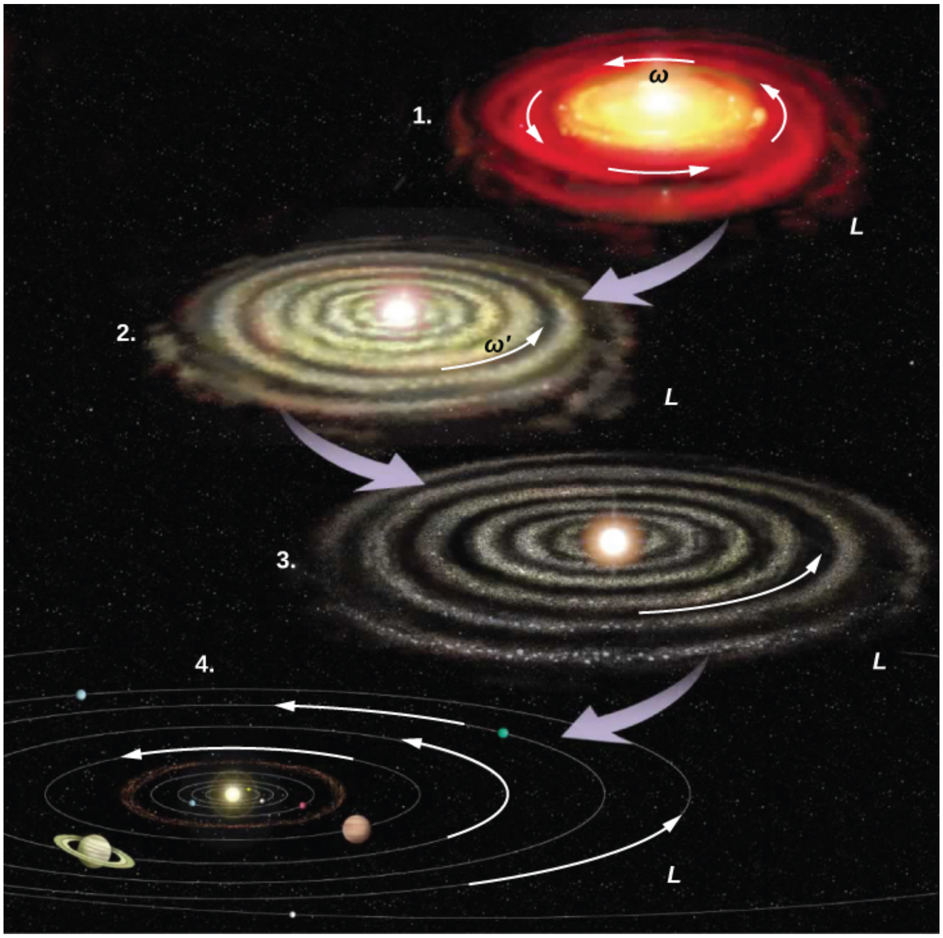


Figure 12.4 The solar system coalesced from a cloud of gas and dust that was originally rotating. The orbital motions and spins of the planets are in the same direction as the original spin and conserve the angular momentum of the parent cloud. (credit: modification of work by NASA)

We continue our discussion with an example that has applications to engineering.

Example 12.2

Coupled Flywheels

A flywheel rotates without friction at an angular velocity $\omega_0 = 600\,\text{rev/min}$ on a frictionless, vertical shaft of negligible rotational inertia. A second flywheel, which is at rest and has a moment of inertia three times that of the rotating flywheel, is dropped onto it (**Figure 12.5**). Because friction exists between the surfaces, the flywheels very quickly reach the same rotational velocity, after which they spin together. (a) Use the law of conservation of angular momentum to determine the angular velocity ω of the combination. (b) What fraction of the initial kinetic energy is lost in the coupling of the flywheels?

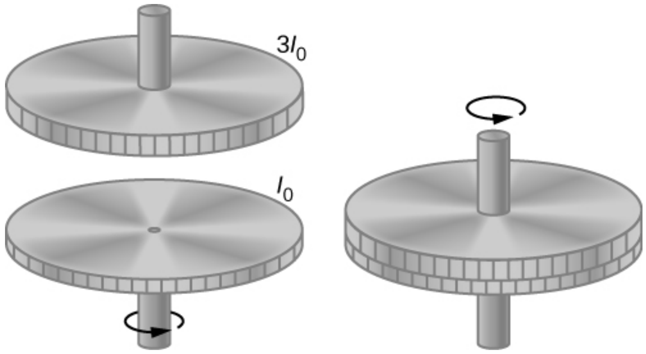


Figure 12.5 Two flywheels are coupled and rotate together.

Strategy

Part (a) is straightforward to solve for the angular velocity of the coupled system. We use the result of (a) to compare the initial and final kinetic energies of the system in part (b).

Solution

a. No external torques act on the system. The force due to friction produces an internal torque, which does not affect the angular momentum of the system. Therefore conservation of angular momentum gives

$$I_0\omega_0 = (I_0 + 3I_0)\omega,$$
$$\omega = \frac{1}{4}\omega_0 = 150 \text{ rev/min} = 15.7 \text{ rad/s}.$$

b. Before contact, only one flywheel is rotating. The rotational kinetic energy of this flywheel is the initial rotational kinetic energy of the system, $\frac{1}{2}I_0\omega_0^2$. The final kinetic energy is $\frac{1}{2}(4I_0)\omega^2 = \frac{1}{2}(4I_0)\left(\frac{\omega_0}{4}\right)^2 = \frac{1}{8}I_0\omega_0^2$.

Therefore, the ratio of the final kinetic energy to the initial kinetic energy is

$$\frac{\frac{1}{8}I_0\omega_0^2}{\frac{1}{2}I_0\omega_0^2} = \frac{1}{4}.$$

Thus, 3/4 of the initial kinetic energy is lost to the coupling of the two flywheels.

Significance

Since the rotational inertia of the system increased, the angular velocity decreased, as expected from the law of conservation of angular momentum. In this example, we see that the final kinetic energy of the system has decreased, as energy is lost to the coupling of the flywheels. Compare this to the example of the skater in Figure 12.3 doing work to bring her arms inward and adding rotational kinetic energy.

12.2 Check Your Understanding A merry-go-round at a playground is rotating at 4.0 rev/min. Three children jump on and increase the moment of inertia of the merry-go-round/children rotating system by 25%. What is the new rotation rate?

Example 12.3

Dismount from a High Bar

An 80.0-kg gymnast dismounts from a high bar. He starts the dismount at full extension, then tucks to complete a number of revolutions before landing. His moment of inertia when fully extended can be approximated as a rod of length 1.8 m and when in the tuck a rod of half that length. If his rotation rate at full extension is 1.0 rev/s and he enters the tuck when his center of mass is at 3.0 m height moving horizontally to the floor, how many revolutions can he execute if he comes out of the tuck at 1.8 m height? See **Figure 12.6**.

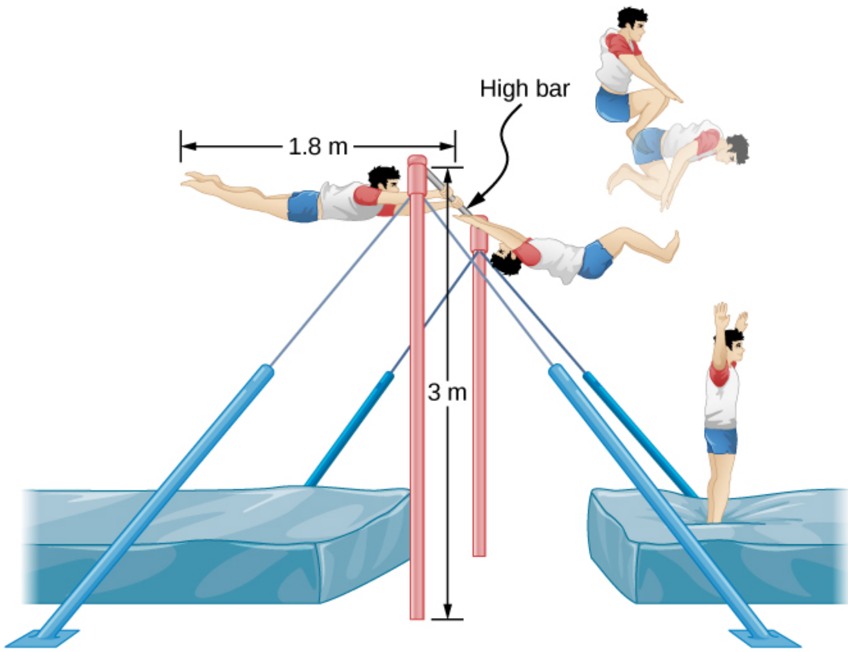


Figure 12.6 A gymnast dismounts from a high bar and executes a number of revolutions in the tucked position before landing upright.

Strategy

Using conservation of angular momentum, we can find his rotation rate when in the tuck. Using the equations of kinematics, we can find the time interval from a height of 3.0 m to 1.8 m. Since he is moving horizontally with respect to the ground, the equations of free fall simplify. This will allow the number of revolutions that can be executed to be calculated. Since we are using a ratio, we can keep the units as rev/s and don't need to convert to radians/s.

Solution

The moment of inertia at full extension is $I_0 = \frac{1}{12}mL^2 = \frac{1}{12}80.0\,\text{kg}(1.8\,\text{m})^2 = 21.6\,\text{kg} \cdot \text{m}^2$.

The moment of inertia in the tuck is $I_\text{f} = \frac{1}{12}mL_\text{f}^2 = \frac{1}{12}80.0\,\text{kg}(0.9\,\text{m})^2 = 5.4\,\text{kg} \cdot \text{m}^2$.

Conservation of angular momentum: $I_\text{f}\omega_\text{f} = I_0\omega_0 \Rightarrow \omega_\text{f} = \frac{I_0\omega_0}{I_\text{f}} = \frac{21.6\,\text{kg} \cdot \text{m}^2(1.0\,\text{rev/s})}{5.4\,\text{kg} \cdot \text{m}^2} = 4.0\,\text{rev/s}$.

Time interval in the tuck: $t = \sqrt{\frac{2h}{g}} = \sqrt{\frac{2(3.0 - 1.8)\text{m}}{9.8\,\text{m/s}}} = 0.5\,\text{s}$.

In 0.5 s, he will be able to execute two revolutions at 4.0 rev/s.

Significance

Note that the number of revolutions he can complete will depend on how long he is in the air. In the problem, he is exiting the high bar horizontally to the ground. He could also exit at an angle with respect to the ground, giving him more or less time in the air depending on the angle, positive or negative, with respect to the ground. Gymnasts must take this into account when they are executing their dismounts.

Example 12.4

Conservation of Angular Momentum of a Collision

A bullet of mass $m = 2.0\text{ g}$ is moving horizontally with a speed of 500.0 m/s. The bullet strikes and becomes embedded in the edge of a solid disk of mass $M = 3.2\text{ kg}$ and radius $R = 0.5\text{ m}$. The cylinder is free to rotate around its axis and is initially at rest (**Figure 12.7**). What is the angular velocity of the disk immediately after the bullet is embedded?

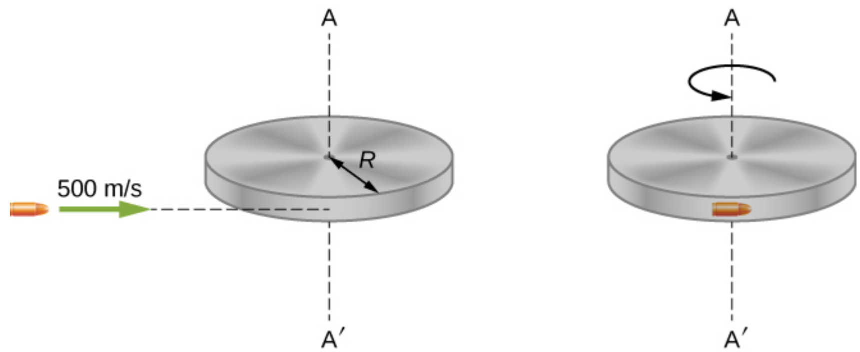


Figure 12.7 A bullet is fired horizontally and becomes embedded in the edge of a disk that is free to rotate about its vertical axis.

Strategy

For the system of the bullet and the cylinder, no external torque acts along the vertical axis through the center of the disk. Thus, the angular momentum along this axis is conserved. The initial angular momentum of the bullet is mvR, which is taken about the rotational axis of the disk the moment before the collision. The initial angular momentum of the cylinder is zero. Thus, the net angular momentum of the system is mvR. Since angular momentum is conserved, the initial angular momentum of the system is equal to the angular momentum of the bullet embedded in the disk immediately after impact.

Solution

The initial angular momentum of the system is

$$L_i = mvR.$$

The moment of inertia of the system with the bullet embedded in the disk is

$$I = mR^2 + \frac{1}{2}MR^2 = \left(m + \frac{M}{2}\right)R^2.$$

The final angular momentum of the system is

$$L_f = I\omega_f.$$

Thus, by conservation of angular momentum, $L_i = L_f$ and

$$mvR = \left(m + \frac{M}{2}\right)R^2\omega_f.$$

Solving for ω_f,

$$\omega_f = \frac{mvR}{(m + M/2)R^2} = \frac{(2.0 \times 10^{-3}\text{ kg})(500.0\text{ m/s})}{(2.0 \times 10^{-3}\text{ kg} + 1.6\text{ kg})(0.50\text{ m})} = 1.2\text{ rad/s}.$$

Significance

The system is composed of both a point particle and a rigid body. Care must be taken when formulating the angular momentum before and after the collision. Just before impact the angular momentum of the bullet is taken about the rotational axis of the disk.

Kepler's Second Law Revisited

Recall that Kepler's second law states that a planet sweeps out equal areas in equal times, that is, the area divided by time, called the areal velocity, is constant. Consider **Figure 12.8**. The time it takes a planet to move from position *A* to *B*, sweeping out area A_1, is exactly the time taken to move from position *C* to *D*, sweeping area A_2, and to move from *E* to *F*, sweeping out area A_3. These areas are the same: $A_1 = A_2 = A_3$.

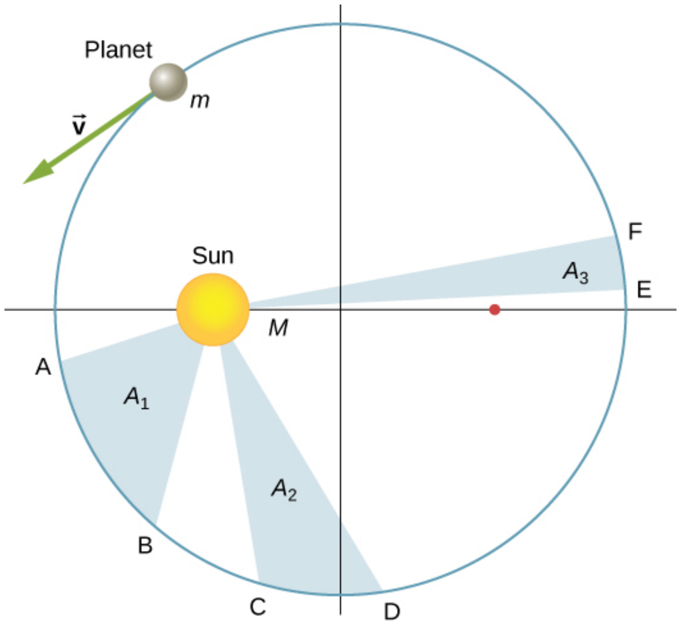


Figure 12.8 The shaded regions shown have equal areas and represent the same time interval.

We saw in **Kepler's Laws of Planetary Motion** that this implies that an orbiting planet must speed up as it gets closer to the Sun and slow down as it moves farther away. We now see that Kepler's second law is just a consequence of the conservation of angular momentum, which holds for any system with only radial forces.

Recall from the definition of angular momentum that, for the orbiting planet, $L = mvr$. If that product of $v \times r$ is to remain constant, then when a planet moves to a larger r, it must be moving at a slower tangential speed v.

Now consider **Figure 12.9**. A small triangular area ΔA is swept out in time Δt. The velocity is along the path and it makes an angle θ with the radial direction. Hence, the tangential (or "perpendicular") velocity is given by $v_{\text{perp}} = v\sin\theta$. The planet moves a distance $\Delta s = v\Delta t\sin\theta$ projected along the direction perpendicular to *r*. Since the area of a triangle is one-half the base (*r*) times the height (Δs), for a small displacement, the area is given by $\Delta A = \frac{1}{2}r\Delta s$. Substituting for Δs, multiplying by *m* in the numerator and denominator, and rearranging, we obtain

$$\Delta A = \frac{1}{2}r\Delta s = \frac{1}{2}r(v\Delta t\sin\theta) = \frac{1}{2m}r(mv\sin\theta\Delta t) = \frac{1}{2m}r(mv_{\text{perp}}\Delta t) = \frac{L}{2m}\Delta t.$$

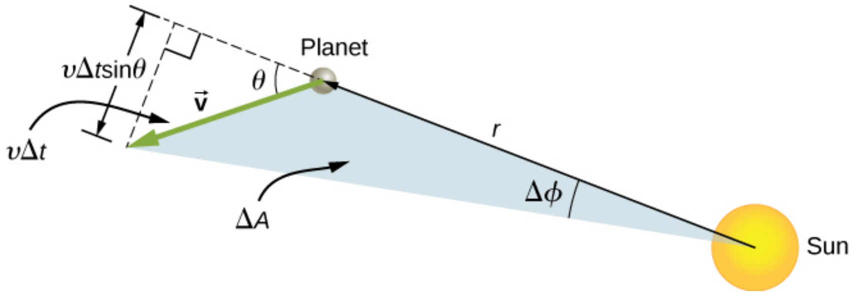


Figure 12.9 The element of area ΔA swept out in time Δt as the planet moves through angle $\Delta\phi$. The angle between the radial direction and $\vec{v}$ is θ.

The areal velocity is simply the rate of change of area with time, so we have

$$\text{areal velocity} = \frac{\Delta A}{\Delta t} = \frac{L}{2m}.$$

Since the angular momentum is constant, the areal velocity must also be constant. This is Kepler's second law in its original form.

You can view an **animated version (https://openstax.org/l/21animationgrav)** of **Figure 12.8**, and many other interesting animations as well, at the School of Physics (University of New South Wales) site.

CHAPTER 12 REVIEW

KEY TERMS

angular momentum rotational analog of linear momentum, found by taking the product of moment of inertia and angular velocity

law of conservation of angular momentum angular momentum is conserved, that is, the initial angular momentum is equal to the final angular momentum when no external torque is applied to the system

KEY EQUATIONS

Angular momentum of a particle	$L = mvr$
Angular momentum of a rigid body	$L = I\omega.$
Rotational kinetic energy	$K = \frac{L^2}{2I}$
Conservation of angular momentum	$\frac{d\vec{\mathbf{L}}}{dt} = 0$

SUMMARY

12.1 Angular Momentum

- The angular momentum $L = mvr$ for a single particle orbiting in a circular path.
- The angular momentum $\vec{\mathbf{L}} = \sum_i \vec{\mathbf{L}}_i$ of a system of particles about a designated origin is the vector sum of the individual momenta of the particles that make up the system.
- A rigid rotating body has angular momentum $L = I\omega$ directed along the axis of rotation. The time derivative of the angular momentum $\frac{dL}{dt} = \sum \tau$ gives the net torque on a rigid body and is directed along the axis of rotation.
- The rotational kinetic energy of a rotating rigid body is given by $K = \frac{L^2}{2I}$

12.2 Conservation of Angular Momentum

- In the absence of external torques, a system's total angular momentum is conserved. This is the rotational counterpart to linear momentum being conserved when the external force on a system is zero.
- For a rigid body that changes its angular momentum in the absence of a net external torque, conservation of angular momentum gives $I_f \omega_f = I_i \omega_i$. This equation says that the angular velocity is inversely proportional to the moment of inertia. Thus, if the moment of inertia decreases, the angular velocity must increase to conserve angular momentum.
- Systems containing both point particles and rigid bodies can be analyzed using conservation of angular momentum. The angular momentum of all bodies in the system must be taken about a common axis.

CONCEPTUAL QUESTIONS

12.1 Angular Momentum

1. Can you assign an angular momentum to a particle without first defining a reference point?

2. For a particle traveling in a straight line, are there any points about which the angular momentum is zero? Assume the line intersects the origin.

3. Under what conditions does a rigid body have angular momentum but not linear momentum?

4. If a particle is moving with respect to a chosen origin it has linear momentum. What conditions must exist for this particle's angular momentum to be zero about the chosen origin?

5. If you know the velocity of a particle, can you say anything about the particle's angular momentum?

12.2 Conservation of Angular Momentum

6. What is the purpose of the small propeller at the back of a helicopter that rotates in the plane perpendicular to the large propeller?

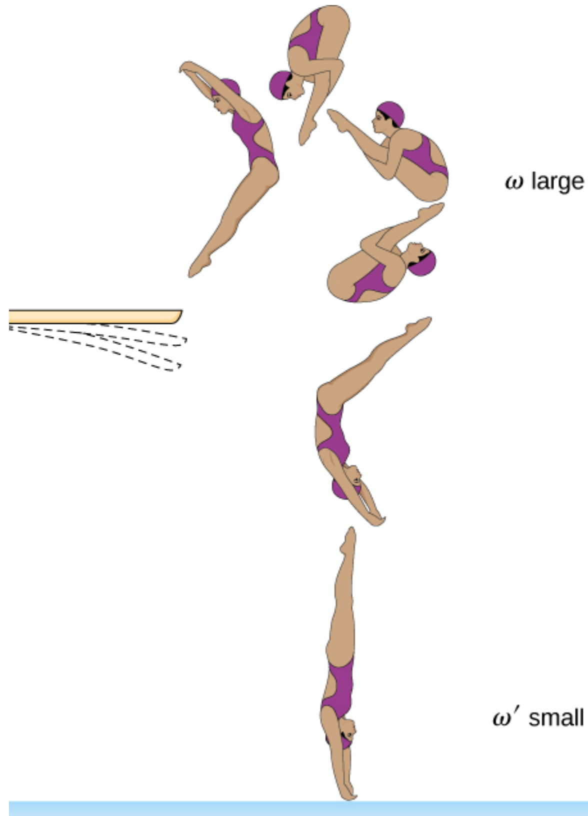
7. Suppose a child walks from the outer edge of a rotating merry-go-round to the inside. Does the angular velocity of the merry-go-round increase, decrease, or remain the same? Explain your answer. Assume the merry-go-round is spinning without friction.

8. As the rope of a tethered ball winds around a pole, what happens to the angular velocity of the ball?

9. Suppose the polar ice sheets broke free and floated toward Earth's equator without melting. What would happen to Earth's angular velocity?

10. Explain why stars spin faster when they collapse.

11. Competitive divers pull their limbs in and curl up their bodies when they do flips. Just before entering the water, they fully extend their limbs to enter straight down (see below). Explain the effect of both actions on their angular velocities. Also explain the effect on their angular momentum.

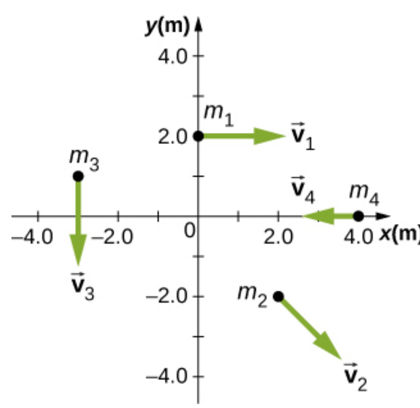


PROBLEMS

12.1 Angular Momentum

12. A Formula One race car with mass 750.0 kg is speeding through a course in Monaco and enters a circular turn at 220.0 km/h in the counterclockwise direction about the origin of the circle. At another part of the course, the car enters a second circular turn at 180 km/h also in the counterclockwise direction. If the radius of curvature of the first turn is 130.0 m and that of the second is 100.0 m, compare the angular momenta of the race car in each turn taken about the origin of the circular turn.

13. Determine the signs (positive or negative) of the angular momenta about the origin of the particles as shown below.



14. (a) Calculate the angular momentum of Earth in its orbit around the Sun. (b) Compare this angular momentum with the angular momentum of Earth about its axis.

15. A satellite is spinning at 6.0 rev/s. The satellite consists of a main body in the shape of a sphere of radius 2.0 m and mass 10,000 kg, and two antennas projecting out from the center of mass of the main body that can be approximated with rods of length 3.0 m each and mass 10 kg. The antenna's lie in the plane of rotation. What is the angular momentum of the satellite?

16. A propeller consists of two blades each 3.0 m in length and mass 120 kg each. The propeller can be approximated by a single rod rotating about its center of mass. The propeller starts from rest and rotates up to 1200 rpm in 30 seconds at a constant rate. What is the angular momentum of the propeller at $t = 10\,\text{s}; t = 20\,\text{s}$?

17. A pulsar is a rapidly rotating neutron star. The Crab nebula pulsar in the constellation Taurus has a period of 33.5×10^{-3} s, radius 10.0 km, and mass 2.8×10^{30} kg. The pulsar's rotational period will increase over time due to the release of electromagnetic radiation, which doesn't change its radius but reduces its rotational energy. (a) What is the angular momentum of the pulsar?

18. The blades of a wind turbine are 30 m in length and rotate at a maximum rotation rate of 20 rev/min. (a) If the blades are 6000 kg each and the rotor assembly has three blades, calculate the angular momentum of the turbine at this rotation rate.

19. A roller coaster has mass 3000.0 kg and needs to make it safely through a vertical circular loop of radius 50.0 m. What is the minimum angular momentum of the coaster at the bottom of the loop to make it safely through? Neglect friction on the track. Take the coaster to be a point particle.

20. A mountain biker takes a jump in a race and goes airborne. The mountain bike is travelling at 10.0 m/s before it goes airborne. If the mass of the front wheel on the bike is 750 g and has radius 35 cm, what is the angular momentum of the spinning wheel in the air the moment the bike leaves the ground?

12.2 Conservation of Angular Momentum

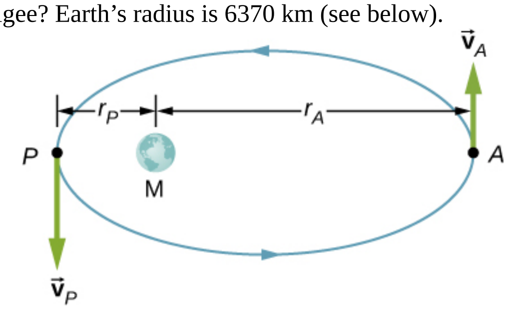
21. A disk of mass 2.0 kg and radius 60 cm with a small mass of 0.05 kg attached at the edge is rotating at 2.0 rev/s. The small mass suddenly separates from the disk. What is the disk's final rotation rate?

22. The Sun's mass is $2.0 \times 10^{30}\,\text{kg}$, its radius is 7.0×10^{5} km, and it has a rotational period of approximately 28 days. If the Sun should collapse into a white dwarf of radius 3.5×10^{3} km, what would its period be if no mass were ejected and a sphere of uniform density can model the Sun both before and after?

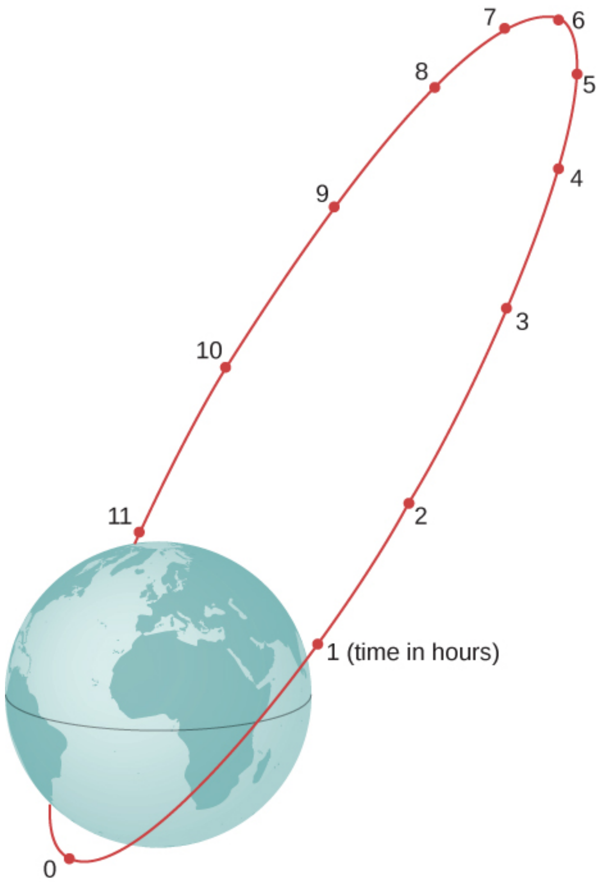
23. A cylinder with rotational inertia $I_1 = 2.0\,\text{kg} \cdot \text{m}^2$ rotates clockwise about a vertical axis through its center with angular speed $\omega_1 = 5.0\,\text{rad/s}$. A second cylinder with rotational inertia $I_2 = 1.0\,\text{kg} \cdot \text{m}^2$ rotates counterclockwise about the same axis with angular speed $\omega_2 = 8.0\,\text{rad/s}$. If the cylinders couple so they have the same rotational axis what is the angular speed of the combination? What percentage of the original kinetic energy is lost to friction?

24. A diver off the high board imparts an initial rotation with his body fully extended before going into a tuck and executing three back somersaults before hitting the water. If his moment of inertia before the tuck is $16.9\,\text{kg} \cdot \text{m}^2$ and after the tuck during the somersaults is $4.2\,\text{kg} \cdot \text{m}^2$, what rotation rate must he impart to his body directly off the board and before the tuck if he takes 1.4 s to execute the somersaults before hitting the water?

25. An Earth satellite has its apogee at 2500 km above the surface of Earth and perigee at 500 km above the surface of Earth. At apogee its speed is 730 m/s. What is its speed at perigee? Earth's radius is 6370 km (see below).



26. A Molniya orbit is a highly eccentric orbit of a communication satellite so as to provide continuous communications coverage for Scandinavian countries and adjacent Russia. The orbit is positioned so that these countries have the satellite in view for extended periods in time (see below). If a satellite in such an orbit has an apogee at 40,000.0 km as measured from the center of Earth and a velocity of 3.0 km/s, what would be its velocity at perigee measured at 200.0 km altitude?



27. A bug of mass 0.020 kg is at rest on the edge of a solid cylindrical disk $(M = 0.10\,\text{kg}, R = 0.10\,\text{m})$ rotating in a horizontal plane around the vertical axis through its center. The disk is rotating at 10.0 rad/s. The bug crawls to the center of the disk. (a) What is the new angular velocity of the disk? (b) What is the change in the kinetic energy of the system? (c) If the bug crawls back to the outer edge of the disk, what is the angular velocity of the disk then? (d) What is the new kinetic energy of the system? (e) What is the cause of the increase and decrease of kinetic energy?

28. A uniform rod of mass 200 g and length 100 cm is free to rotate in a horizontal plane around a fixed vertical axis through its center, perpendicular to its length. Two small beads, each of mass 20 g, are mounted in grooves along the rod. Initially, the two beads are held by catches on opposite sides of the rod's center, 10 cm from the axis of rotation. With the beads in this position, the rod is rotating with an angular velocity of 10.0 rad/s. When the catches are released, the beads slide outward along the rod. (a) What is the rod's angular velocity when the beads reach the ends of the rod? (b) What is the rod's angular velocity if the beads fly off the rod?

29. A playground merry-go-round has a mass of 120 kg and a radius of 1.80 m and it is rotating with an angular velocity of 0.500 rev/s. What is its angular velocity after a 22.0-kg child gets onto it by grabbing its outer edge? The child is initially at rest.

30. Three children are riding on the edge of a merry-go-round that is 100 kg, has a 1.60-m radius, and is spinning at 20.0 rpm. The children have masses of 22.0, 28.0, and 33.0 kg. If the child who has a mass of 28.0 kg moves to the center of the merry-go-round, what is the new angular velocity in rpm?

31. In 2015, in Warsaw, Poland, Olivia Oliver of Nova Scotia broke the world record for being the fastest spinner on ice skates. She achieved a record 342 rev/min, beating the existing Guinness World Record by 34 rotations. If an ice skater extends her arms at that rotation rate, what would be her new rotation rate? Assume she can be approximated by a 45-kg rod that is 1.7 m tall with a radius of 15 cm in the record spin. With her arms stretched take the approximation of a rod of length 130 cm with 10% of her body mass aligned perpendicular to the spin axis. Neglect frictional forces.

32. A satellite in a geosynchronous circular orbit is 42,164.0 km from the center of Earth. A small asteroid collides with the satellite sending it into an elliptical orbit of apogee 45,000.0 km. What is the speed of the satellite at apogee? Assume its angular momentum is conserved.

33. A gymnast does cartwheels along the floor and then launches herself into the air and executes several flips in a tuck while she is airborne. If her moment of inertia when executing the cartwheels is $13.5\,\text{kg}\cdot\text{m}^2$ and her spin rate is 0.5 rev/s, how many revolutions does she do in the air if her moment of inertia in the tuck is $3.4\,\text{kg}\cdot\text{m}^2$ and she has 2.0 s to do the flips in the air?

34. The centrifuge at NASA Ames Research Center has a radius of 8.8 m and can produce forces on its payload of 20 *g*s or 20 times the force of gravity on Earth. (a) What is the angular momentum of a 20-kg payload that experiences 10 *g*s in the centrifuge? (b) If the driver motor was turned off in (a) and the payload lost 10 kg, what would be its new spin rate, taking into account there are no frictional forces present?

35. A ride at a carnival has four spokes to which pods are attached that can hold two people. The spokes are each 15 m long and are attached to a central axis. Each spoke has mass 200.0 kg, and the pods each have mass 100.0 kg. If the ride spins at 0.2 rev/s with each pod containing two 50.0-kg children, what is the new spin rate if all the children jump off the ride?

36. An ice skater is preparing for a jump with turns and has his arms extended. His moment of inertia is $1.8\,\text{kg}\cdot\text{m}^2$ while his arms are extended, and he is spinning at 0.5 rev/s. If he launches himself into the air at 9.0 m/s at an angle of $45°$ with respect to the ice, how many revolutions can he execute while airborne if his moment of inertia in the air is $0.5\,\text{kg}\cdot\text{m}^2$?

37. A space station consists of a giant rotating hollow cylinder of mass 10^6 kg including people on the station and a radius of 100.00 m. It is rotating in space at 3.30 rev/min in order to produce artificial gravity. If 100 people of an average mass of 65.00 kg spacewalk to an awaiting spaceship, what is the new rotation rate when all the people are off the station?

38. Neptune has a mass of 1.0×10^{26} kg and is 4.5×10^9 km from the Sun with an orbital period of 165 years. Planetesimals in the outer primordial solar system 4.5 billion years ago coalesced into Neptune over hundreds of millions of years. If the primordial disk that evolved into our present day solar system had a radius of 10^{11} km and if the matter that made up these planetesimals that later became Neptune was spread out evenly on the edges of it, what was the orbital period of the outer edges of the primordial disk?

13 | THE NATURE OF LIGHT

Chapter Outline

Introduction

Figure 13.1 Waves in the ocean behave similarly to all other types of waves. (credit: Steve Jurveston, Flickr)

What do we mean when we say something is a wave? The most intuitive and easiest wave to imagine is the familiar water wave. More precisely, a **wave** is a disturbance that propagates, or moves from the place it was created. For water waves, the disturbance is in the surface of the water, perhaps created by a rock thrown into a pond or by a swimmer splashing the surface repeatedly. For sound waves, the disturbance is a change in air pressure, perhaps created by the oscillating cone inside a speaker. For earthquakes, there are several types of disturbances, including disturbance of Earth's surface and pressure disturbances under the surface.

In the nineteenth century, the culmination of the study of electrodynamics by James Clerk Maxwell showed that light, too, is a wave, an **electromagnetic wave**. And, we will learn that the term "light" may be used more generally, and may refer not only to the light visible to the human eye, but to electromagnetic waves whose frequency and wavelength make them invisible. Examples of these include radio waves, infrared, microwaves, ultraviolet, and x-radiation.

All waves exhibit common characteristics such as amplitude, period, frequency and energy. All wave characteristics can be described by a small set of underlying principles.

13.1 | The Wave Behavior of Light

Learning Objectives

By the end of this section, you will be able to:

- Explain the evidence for Maxwell's electromagnetic model of light.
- Describe the relationship between wavelength, frequency, and speed of light.
- Discuss the particle model of light and the definition of photon.

Coded into the light and other kinds of radiation that reach us from objects in the universe is a wide range of information about what those objects are like and how they work. If we can decipher this code and read the messages it contains, we can learn an enormous amount about the cosmos without ever having to leave Earth or its immediate environment.

The visible light and other radiation we receive from the stars and planets is generated by processes at the atomic level—by changes in the way the parts of an atom interact and move. Thus, to appreciate how light is generated, we must explore how atoms work. There is a bit of irony in the fact that in order to understand some of the largest structures in the universe, we must become acquainted with some of the smallest.

Notice that we have twice used the phrase "light and other radiation." One of the key ideas explored in this chapter is that visible light is not unique; it is merely the most familiar example of a much larger family of radiation that can carry information to us.

The word " radiation" will be used frequently in this book, so it is important to understand what it means. In everyday language, "radiation" is often used to describe certain kinds of energetic subatomic particles released by radioactive materials in our environment. (An example is the kind of radiation used to treat some cancers.) But this is not what we mean when we use the word "radiation" in an astronomy text. *Radiation*, as used in this book, is a general term for waves (including light waves) that *radiate* outward from a source.

As we saw in **Newton's Law of Universal Gravitation**, Newton's theory of gravity accounts for the motions of planets as well as objects on Earth. Application of this theory to a variety of problems dominated the work of scientists for nearly two centuries. In the nineteenth century, many physicists turned to the study of electricity and magnetism, which are intimately connected with the production of light.

The scientist who played a role in this field comparable to Newton's role in the study of gravity was physicist James Clerk Maxwell, born and educated in Scotland (**Figure 13.2**). Inspired by a number of ingenious experiments that showed an intimate relationship between electricity and magnetism, Maxwell developed a theory that describes both electricity and magnetism with only a small number of elegant equations. It is this theory that gives us important insights into the nature and behavior of light.

Figure 13.2 Maxwell unified the rules governing electricity and magnetism into a coherent theory.

Maxwell's Theory of Electromagnetism

We will look at the structure of the atom in more detail later, but we begin by noting that the typical atom consists of several types of particles, a number of which have not only mass but an additional property called electric charge. In the nucleus (central part) of every atom are *protons*, which are positively charged; outside the nucleus are electrons, which have a negative charge.

Maxwell's theory deals with these electric charges and their effects, especially when they are moving. In the vicinity of an electron charge, another charge feels a force of attraction or repulsion: opposite charges attract; like charges repel. When charges are not in motion, we observe only this electric attraction or repulsion. If charges are in motion, however (as they are inside every atom and in a wire carrying a current), then we measure another force called *magnetism*.

Magnetism was well known for much of recorded human history, but its cause was not understood until the nineteenth century. Experiments with electric charges demonstrated that magnetism was the result of moving charged particles. Sometimes, the motion is clear, as in the coils of heavy wire that make an industrial electromagnet. Other times, it is more subtle, as in the kind of magnet you buy in a hardware store, in which many of the electrons inside the atoms are spinning in roughly the same direction; it is the alignment of their motion that causes the material to become magnetic.

Physicists use the word *field* to describe the action of forces that one object exerts on other distant objects. For example, we say the Sun produces a *gravitational field* that controls Earth's orbit, even though the Sun and Earth do not come directly into contact. Using this terminology, we can say that stationary electric charges produce *electric fields*, and moving electric charges also produce *magnetic fields*.

Actually, the relationship between electric and magnetic phenomena is even more profound. Experiments showed that changing magnetic fields could produce electric currents (and thus changing electric fields), and changing electric currents could in turn produce changing magnetic fields. So once begun, electric and magnetic field changes could continue to trigger each other.

Maxwell analyzed what would happen if electric charges were oscillating (moving constantly back and forth) and found that the resulting pattern of electric and magnetic fields would spread out and travel rapidly through space. Something similar happens when a raindrop strikes the surface of water or a frog jumps into a pond. The disturbance moves outward and creates a pattern we call a *wave* in the water (**Figure 13.3**). You might, at first, think that there must be very few situations in nature where electric charges oscillate, but this is not at all the case. As we shall see, atoms and molecules (which consist of charged particles) oscillate back and forth all the time. The resulting electromagnetic disturbances are among the most common phenomena in the universe.

Figure 13.3 An oscillation in a pool of water creates an expanding disturbance called a wave. (credit: modification of work by "vastateparksstaff"/Flickr)

Maxwell was able to calculate the speed at which an electromagnetic disturbance moves through space; he found that it is equal to the speed of light, which had been measured experimentally. On that basis, he speculated that light was one form of a family of possible electromagnetic disturbances called **electromagnetic radiation**, a conclusion that was again confirmed in laboratory experiments. When light (reflected from the pages of an astronomy textbook, for example) enters a human eye, its changing electric and magnetic fields stimulate nerve endings, which then transmit the information contained in these changing fields to the brain. The science of astronomy is primarily about analyzing radiation from distant objects to understand what they are and how they work.

The Wave-Like Characteristics of Light

The changing electric and magnetic fields in light are similar to the waves that can be set up in a quiet pool of water. In both cases, the disturbance travels rapidly outward from the point of origin and can use its energy to disturb other things farther away. (For example, in water, the expanding ripples moving away from our frog could disturb the peace of a dragonfly resting on a leaf in the same pool.) In the case of electromagnetic waves, the radiation generated by a transmitting antenna full of charged particles and moving electrons at your local radio station can, sometime later, disturb a group of electrons in your car radio antenna and bring you the news and weather while you are driving to class or work in the morning.

The waves generated by charged particles differ from water waves in some profound ways, however. Water waves require water to travel in. The sound waves we hear, to give another example, are pressure disturbances that require air to travel though. But electromagnetic waves do not require water or air: the fields generate each other and so can move through a vacuum (such as outer space). This was such a disturbing idea to nineteenth-century scientists that they actually made up a substance to fill all of space—one for which there was not a single shred of evidence—just so light waves could have something to travel through: they called it the *aether*. Today, we know that there is no aether and that electromagnetic waves have no trouble at all moving through empty space (as all the starlight visible on a clear night must surely be doing).

The other difference is that *all* electromagnetic waves move at the same speed in empty space (the speed of light—approximately 300,000 kilometers per second, or 300,000,000 meters per second, which can also be written as 3×10^8 m/s), which turns out to be the fastest possible speed in the universe. No matter where electromagnetic waves are generated from and no matter what other properties they have, when they are moving (and not interacting with matter), they move at the speed of light. Yet you know from everyday experience that there are different kinds of light. For example, we perceive that light waves differ from one another in a property we call color. Let's see how we can denote the differences among the whole broad family of electromagnetic waves.

The nice thing about a wave is that it is a repeating phenomenon. Whether it is the up-and-down motion of a water wave or the changing electric and magnetic fields in a wave of light, the pattern of disturbance repeats in a cyclical way. Thus, any wave motion can be characterized by a series of crests and troughs (**Figure 13.4**). Moving from one crest through a trough to the next crest completes one cycle. The horizontal length covered by one cycle is called the **wavelength**. Ocean waves provide an analogy: the wavelength is the distance that separates successive wave crests.

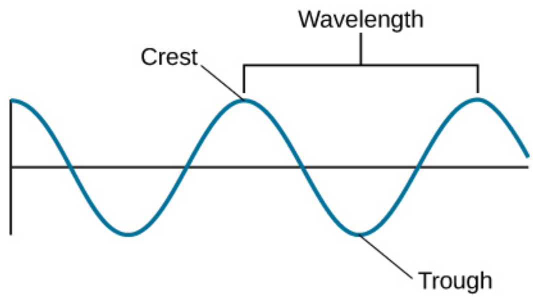


Figure 13.4 Electromagnetic radiation has wave-like characteristics. The wavelength (λ) is the distance between crests, the frequency (*f*) is the number of cycles per second, and the speed (*c*) is the distance the wave covers during a specified period of time (e.g., kilometers per second).

For visible light, our eyes perceive different wavelengths as different colors: red, for example, is the longest visible wavelength, and violet is the shortest. The main colors of visible light from longest to shortest wavelength can be remembered using the mnemonic ROY G BIV—for Red, Orange, Yellow, Green, Blue, Indigo, and Violet. Other invisible forms of electromagnetic radiation have different wavelengths, as we will see in the next section.

We can also characterize different waves by their **frequency**, the number of wave cycles that pass by per second. If you count 10 crests moving by each second, for example, then the frequency is 10 cycles per second (cps). In honor of Heinrich Hertz, the physicist who—inspired by Maxwell's work—discovered radio waves, a cps is also called a *hertz* (Hz). Take a look at your radio, for example, and you will see the channel assigned to each radio station is characterized by its frequency, usually in units of KHz (kilohertz, or thousands of hertz) or MHz (megahertz, or millions of hertz).

Wavelength (λ) and frequency (*f*) are related because all electromagnetic waves travel at the same speed. To see how this works, imagine a parade in which everyone is forced by prevailing traffic conditions to move at exactly the same speed. You stand on a corner and watch the waves of marchers come by. First you see row after row of miniature ponies. Because they are not very large and, therefore, have a shorter wavelength, a good number of the ponies can move past you each minute; we can say they have a high frequency. Next, however, come several rows of circus elephants. The elephants are large and marching at the same speed as the ponies, so far fewer of them can march past you per minute: Because they have a wider spacing (longer wavelength), they represent a lower frequency.

The formula for this relationship can be expressed as follows: for any wave motion, the speed at which a wave moves equals the frequency times the wavelength. Waves with longer wavelengths have lower frequencies. Mathematically, we can express this as

$$c = \lambda f$$

where the Greek letter for "l"—lambda, λ—is used to denote wavelength and *c* is the scientific symbol for the speed of light. Solving for the wavelength, this is expressed as:

$$\lambda = \frac{c}{f}.$$

Example 13.1

Deriving and Using the Wave Equation

The equation for the relationship between the speed and other characteristics of a wave can be derived from our basic understanding of motion. The average speed of anything that is moving is:

$$\text{average speed} = \frac{\text{distance}}{\text{time}}$$

(So, for example, a car on the highway traveling at a speed of 100 km/h covers 100 km during the time of 1 h.) For an electromagnetic wave to travel the distance of one of its wavelengths, λ, at the speed of light, c, we have $c = \lambda/t$. The frequency of a wave is the number of cycles per second. If a wave has a frequency of a million cycles per second, then the time for each cycle to go by is a millionth of a second. So, in general, $t = 1/f$. Substituting into our wave equation, we get $c = \lambda \times f$. Now let's use this to calculate an example. What is the wavelength of visible light that has a frequency of 5.66×10^{14} Hz?

Solution

Solving the wave equation for wavelength, we find:

$$\lambda = \frac{c}{f}$$

Substituting our values gives:

$$\lambda = \frac{3.00 \times 10^8 \text{ m/s}}{5.66 \times 10^{14} \text{ Hz}} = 5.30 \times 10^{-7} \text{ m}$$

This answer can also be written as 530 nm, which is in the yellow-green part of the visible spectrum (nm stands for nanometers, where the term "nano" means "billionths").

Check Your Learning

"Tidal waves," or tsunamis, are waves caused by earthquakes that travel rapidly through the ocean. If a tsunami travels at the speed of 600 km/h and approaches a shore at a rate of one wave crest every 15 min (4 waves/h), what would be the distance between those wave crests at sea?

Answer:

$$\lambda = \frac{600 \text{ km/h}}{4 \text{ waves/h}} = 150 \text{ km}$$

Light as a Photon

The electromagnetic wave model of light (as formulated by Maxwell) was one of the great triumphs of nineteenth-century science. In 1887, when Heinrich Hertz actually made invisible electromagnetic waves (what today are called radio waves) on one side of a room and detected them on the other side, it ushered in a new era that led to the modern age of telecommunications. His experiment ultimately led to the technologies of television, cell phones, and today's wireless networks around the globe.

However, by the beginning of the twentieth century, more sophisticated experiments had revealed that light behaves in certain ways that cannot be explained by the wave model. Reluctantly, physicists had to accept that sometimes light behaves more like a "particle"—or at least a self-contained packet of energy—than a wave. We call such a packet of electromagnetic energy a **photon**.

The fact that light behaves like a wave in certain experiments and like a particle in others was a very surprising and unlikely idea. After all, our common sense says that waves and particles are opposite concepts. On one hand, a wave is a repeating disturbance that, by its very nature, is not in only one place, but spreads out. A particle, on the other hand, is something that can be in only one place at any given time. Strange as it sounds, though, countless experiments now confirm that electromagnetic radiation can sometimes behave like a wave and at other times like a particle.

Then, again, perhaps we shouldn't be surprised that something that always travels at the "speed limit" of the universe and doesn't need a medium to travel through might not obey our everyday common sense ideas. The confusion that this wave-particle duality of light caused in physics was eventually resolved by the introduction of a more complicated theory of waves and particles, now called quantum mechanics. (This is one of the most interesting fields of modern science, but it is mostly beyond the scope of our book. If you are interested in it, see some of the suggested resources at the end of this chapter.)

In any case, you should now be prepared when scientists (or the authors of this book) sometimes discuss electromagnetic radiation as if it consisted of waves and at other times refer to it as a stream of photons. A photon (being a packet of energy) carries a specific amount of energy. We can use the idea of energy to connect the photon and wave models. How much energy a photon has depends on its frequency when you think about it as a wave. A low-energy radio wave has a low frequency as a wave, while a high-energy X-ray at your dentist's office is a high-frequency wave. Among the colors of visible light, violet-light photons have the highest energy and red-light photons have the lowest.

Test whether the connection between photons and waves is clear to you. In the above example, which photon would have the longer wavelength as a wave: the radio wave or the X-ray? If you answered the radio wave, you are correct. Radio waves have a lower frequency, so the wave cycles are longer (they are elephants, not miniature ponies).

13.2 | Huygens' Principle

Learning Objectives

By the end of this section, you will be able to:

- Describe Huygens's principle
- Use Huygens's principle to explain the law of reflection
- Use Huygens's principle to explain the law of refraction
- Use Huygens's principle to explain diffraction

In the preceding chapters, we have been discussing optical phenomena using the ray model of light. However, some phenomena require analysis and explanations based on the wave characteristics of light. This is particularly true when the wavelength is not negligible compared to the dimensions of an optical device, such as a slit in the case of *diffraction*. Huygens's principle is an indispensable tool for this analysis.

Figure 13.5 shows how a transverse wave looks as viewed from above and from the side. A light wave can be imagined to propagate like this, although we do not actually see it wiggling through space. From above, we view the wave fronts (or wave crests) as if we were looking down on ocean waves. The side view would be a graph of the electric or magnetic field. The view from above is perhaps more useful in developing concepts about **wave optics**.

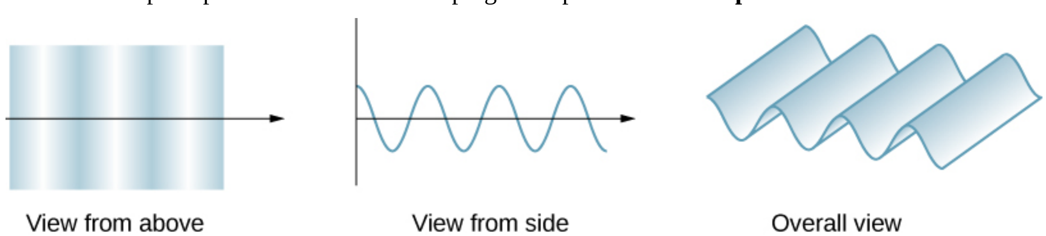


Figure 13.5 A transverse wave, such as an electromagnetic light wave, as viewed from above and from the side. The direction of propagation is perpendicular to the wave fronts (or wave crests) and is represented by a ray.

The Dutch scientist Christiaan Huygens (1629–1695) developed a useful technique for determining in detail how and where waves propagate. Starting from some known position, **Huygens's principle** states that every point on a wave front is a source of wavelets that spread out in the forward direction at the same speed as the wave itself. The new wave front is tangent to all of the wavelets.

Figure 13.6 shows how Huygens's principle is applied. A wave front is the long edge that moves, for example, with the crest or the trough. Each point on the wave front emits a semicircular wave that moves at the propagation speed v. We can draw these wavelets at a time t later, so that they have moved a distance $s = vt$. The new wave front is a plane tangent to the wavelets and is where we would expect the wave to be a time t later. Huygens's principle works for all types of waves, including water waves, sound waves, and light waves. It is useful not only in describing how light waves propagate but also in explaining the laws of reflection and refraction. In addition, we will see that Huygens's principle tells us how and where light rays interfere.

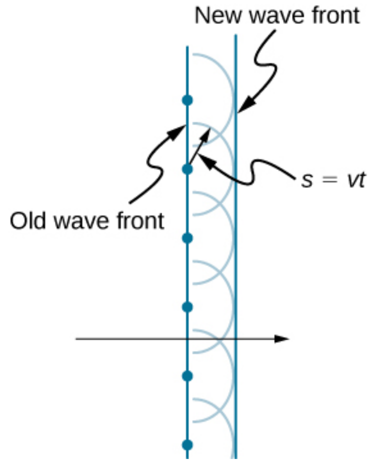


Figure 13.6 Huygens's principle applied to a straight wave front. Each point on the wave front emits a semicircular wavelet that moves a distance $s = vt$. The new wave front is a line tangent to the wavelets.

Reflection

Figure 13.7 shows how a mirror reflects an incoming wave at an angle equal to the incident angle, verifying the law of reflection. As the wave front strikes the mirror, wavelets are first emitted from the left part of the mirror and then from the right. The wavelets closer to the left have had time to travel farther, producing a wave front traveling in the direction shown.

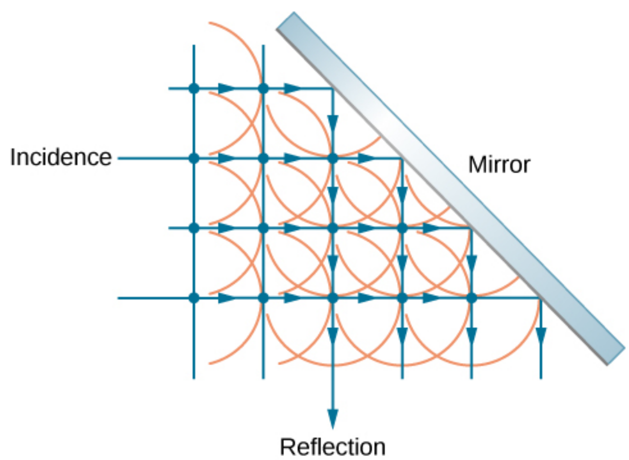


Figure 13.7 Huygens's principle applied to a plane wave front striking a mirror. The wavelets shown were emitted as each point on the wave front struck the mirror. The tangent to these wavelets shows that the new wave front has been reflected at an angle equal to the incident angle. The direction of propagation is perpendicular to the wave front, as shown by the downward-pointing arrows.

Refraction

The law of refraction can be explained by applying Huygens's principle to a wave front passing from one medium to another (**Figure 13.8**). Each wavelet in the figure was emitted when the wave front crossed the interface between the media. Since the speed of light is smaller in the second medium, the waves do not travel as far in a given time, and the new wave front changes direction as shown. This explains why a ray changes direction to become closer to the perpendicular when light slows down. Snell's law can be derived from the geometry in **Figure 13.8** (**Example 13.2**).

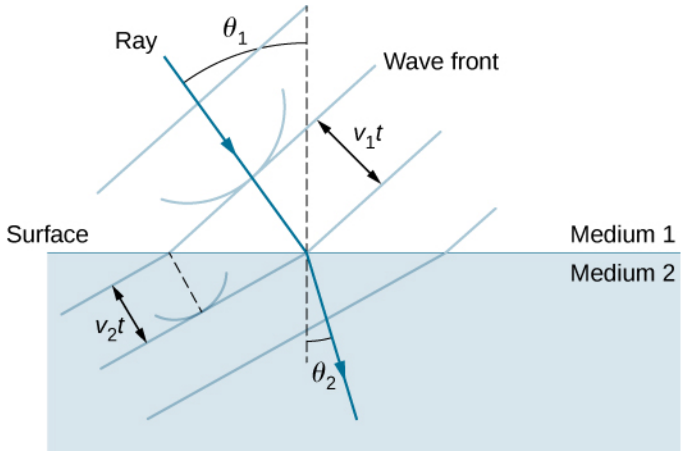


Figure 13.8 Huygens's principle applied to a plane wave front traveling from one medium to another, where its speed is less. The ray bends toward the perpendicular, since the wavelets have a lower speed in the second medium.

Example 13.2

Deriving the Law of Refraction

By examining the geometry of the wave fronts, derive the law of refraction.

Strategy

Consider **Figure 13.9**, which expands upon **Figure 13.8**. It shows the incident wave front just reaching the surface at point *A*, while point *B* is still well within medium 1. In the time Δt it takes for a wavelet from *B* to reach B' on the surface at speed $v_1 = c/n_1$, a wavelet from *A* travels into medium 2 a distance of $AA' = v_2 \Delta t$, where $v_2 = c/n_2$. Note that in this example, v_2 is slower than v_1 because $n_1 < n_2$.

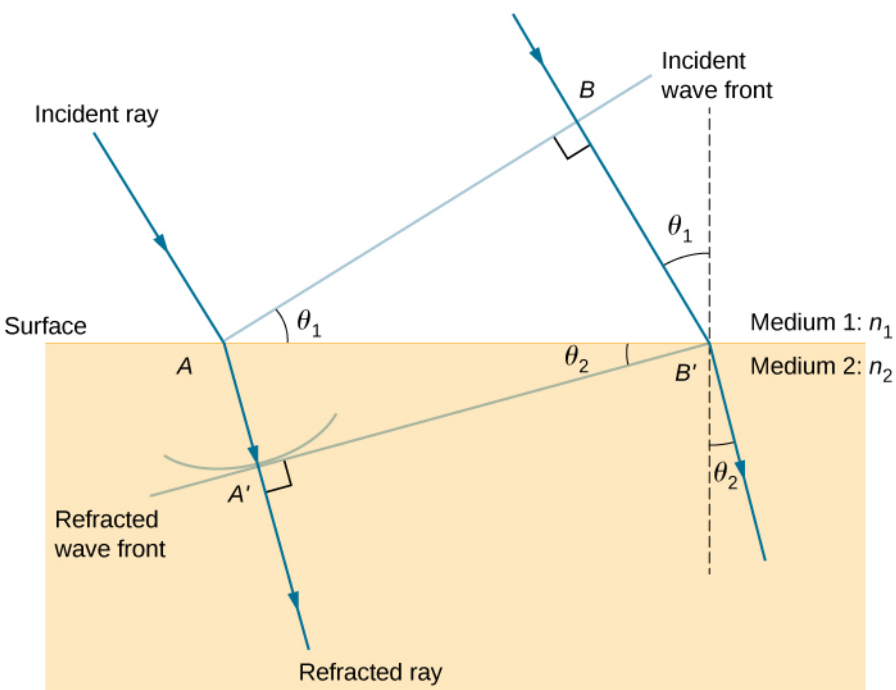


Figure 13.9 Geometry of the law of refraction from medium 1 to medium 2.

Solution

The segment on the surface AB' is shared by both the triangle ABB' inside medium 1 and the triangle $AA'B'$ inside medium 2. Note that from the geometry, the angle $\angle BAB'$ is equal to the angle of incidence, θ_1. Similarly, $\angle AB'A'$ is θ_2.

The length of AB' is given in two ways as

$$AB' = \frac{BB'}{\sin\theta_1} = \frac{AA'}{\sin\theta_2}.$$

Inverting the equation and substituting $AA' = c\Delta t/n_2$ from above and similarly $BB' = c\Delta t/n_1$, we obtain

$$\frac{\sin\theta_1}{c\Delta t/n_1} = \frac{\sin\theta_2}{c\Delta t/n_2}.$$

Cancellation of $c\Delta t$ allows us to simplify this equation into the familiar form

$$n_1 \sin\theta_1 = n_2 \sin\theta_2.$$

Significance

Although the law of refraction was established experimentally by Snell and stated in **Refraction**, its derivation here requires Huygens's principle and the understanding that the speed of light is different in different media.

13.1 Check Your Understanding In **Example 13.2**, we had $n_1 < n_2$. If n_2 were decreased such that $n_1 > n_2$ and the speed of light in medium 2 is faster than in medium 1, what would happen to the length of AA'? What would happen to the wave front $A'B'$ and the direction of the refracted ray?

This **applet (https://openstax.org/l/21walfedaniref)** by Walter Fendt shows an animation of reflection and refraction using Huygens's wavelets while you control the parameters. Be sure to click on "Next step" to display the wavelets. You can see the reflected and refracted wave fronts forming.

Diffraction

What happens when a wave passes through an opening, such as light shining through an open door into a dark room? For light, we observe a sharp shadow of the doorway on the floor of the room, and no visible light bends around corners into other parts of the room. When sound passes through a door, we hear it everywhere in the room and thus observe that sound spreads out when passing through such an opening (**Figure 13.10**). What is the difference between the behavior of sound waves and light waves in this case? The answer is that light has very short wavelengths and acts like a ray. Sound has wavelengths on the order of the size of the door and bends around corners (for frequency of 1000 Hz,

$$\lambda = \frac{c}{f} = \frac{330\text{ m/s}}{1000\text{ s}^{-1}} = 0.33\text{ m},$$

about three times smaller than the width of the doorway).

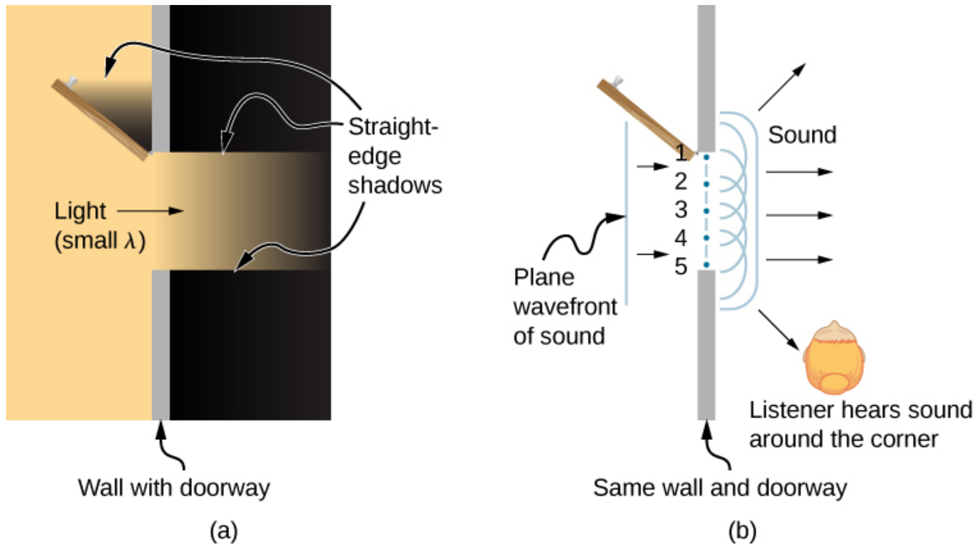


Figure 13.10 (a) Light passing through a doorway makes a sharp outline on the floor. Since light's wavelength is very small compared with the size of the door, it acts like a ray. (b) Sound waves bend into all parts of the room, a wave effect, because their wavelength is similar to the size of the door.

If we pass light through smaller openings such as slits, we can use Huygens's principle to see that light bends as sound does (**Figure 13.11**). The bending of a wave around the edges of an opening or an obstacle is called diffraction. Diffraction is a wave characteristic and occurs for all types of waves. If diffraction is observed for some phenomenon, it is evidence that the phenomenon is a wave. Thus, the horizontal diffraction of the laser beam after it passes through the slits in **Figure 13.11** is evidence that light is a wave. You will learn about diffraction in much more detail in the chapter on **Diffraction**.

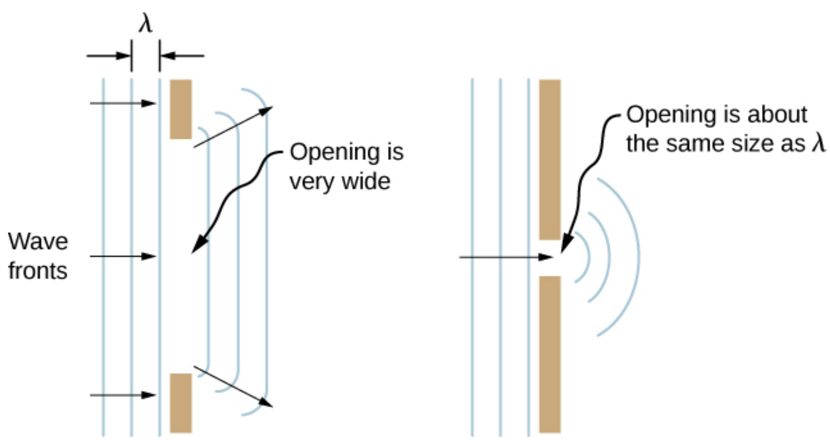


Figure 13.11 Huygens's principle applied to a plane wave front striking an opening. The edges of the wave front bend after passing through the opening, a process called diffraction. The amount of bending is more extreme for a small opening, consistent with the fact that wave characteristics are most noticeable for interactions with objects about the same size as the wavelength.

13.3 | The Electromagnetic Spectrum

Learning Objectives

By the end of this section, you will be able to:

- Understand the bands of the electromagnetic spectrum and how they differ from one another
- Understand how each part of the spectrum interacts with Earth's atmosphere
- Explain how and why the light emitted by an object depends on its temperature

Objects in the universe send out an enormous range of electromagnetic radiation. Scientists call this range the **electromagnetic spectrum**, which they have divided into a number of categories. The spectrum is shown in Figure 13.12, with some information about the waves in each part or band.

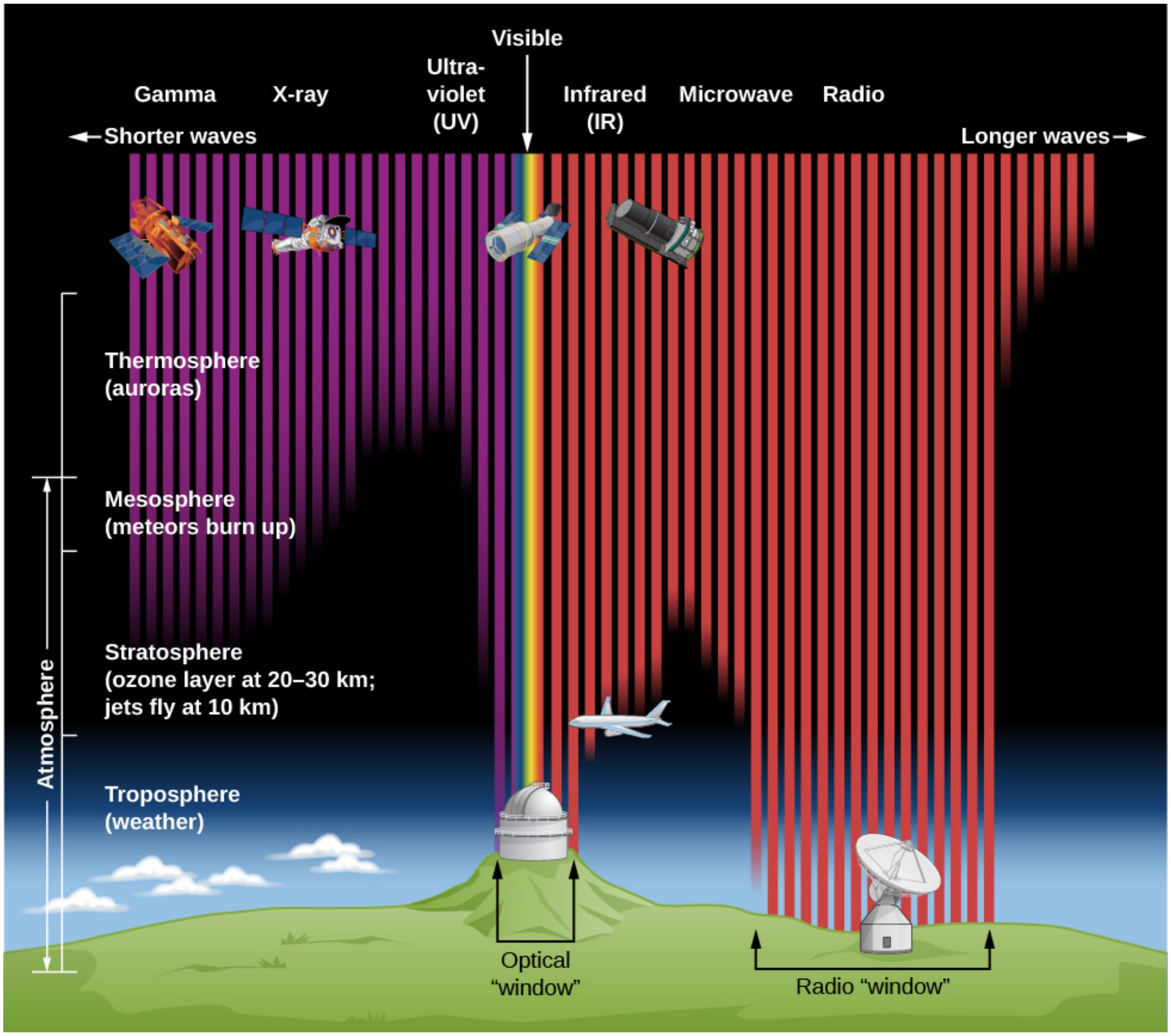


Figure 13.12 This figure shows the bands of the electromagnetic spectrum and how well Earth's atmosphere transmits them. Note that high-frequency waves from space do not make it to the surface and must therefore be observed from space. Some infrared and microwaves are absorbed by water and thus are best observed from high altitudes. Low-frequency radio waves are blocked by Earth's ionosphere. (credit: modification of work by STScI/JHU/NASA)

Types of Electromagnetic Radiation

Electromagnetic radiation with the shortest wavelengths, no longer than 0.01 nanometer, is categorized as **gamma rays** (1 nanometer = 10^{-9} meters; see **Appendix B**). The name *gamma* comes from the third letter of the Greek alphabet: gamma rays were the third kind of radiation discovered coming from radioactive atoms when physicists first investigated their behavior. Because gamma rays carry a lot of energy, they can be dangerous for living tissues. Gamma radiation is generated deep in the interior of stars, as well as by some of the most violent phenomena in the universe, such as the deaths of stars and the merging of stellar corpses. Gamma rays coming to Earth are absorbed by our atmosphere before they reach the ground (which is a good thing for our health); thus, they can only be studied using instruments in space.

Electromagnetic radiation with wavelengths between 0.01 nanometer and 20 nanometers is referred to as **X-rays**. Being more energetic than visible light, X-rays are able to penetrate soft tissues but not bones, and so allow us to make images of the shadows of the bones inside us. While X-rays can penetrate a short length of human flesh, they are stopped by the large numbers of atoms in Earth's atmosphere with which they interact. Thus, X-ray astronomy (like gamma-ray astronomy) could not develop until we invented ways of sending instruments above our atmosphere (**Figure 13.13**).

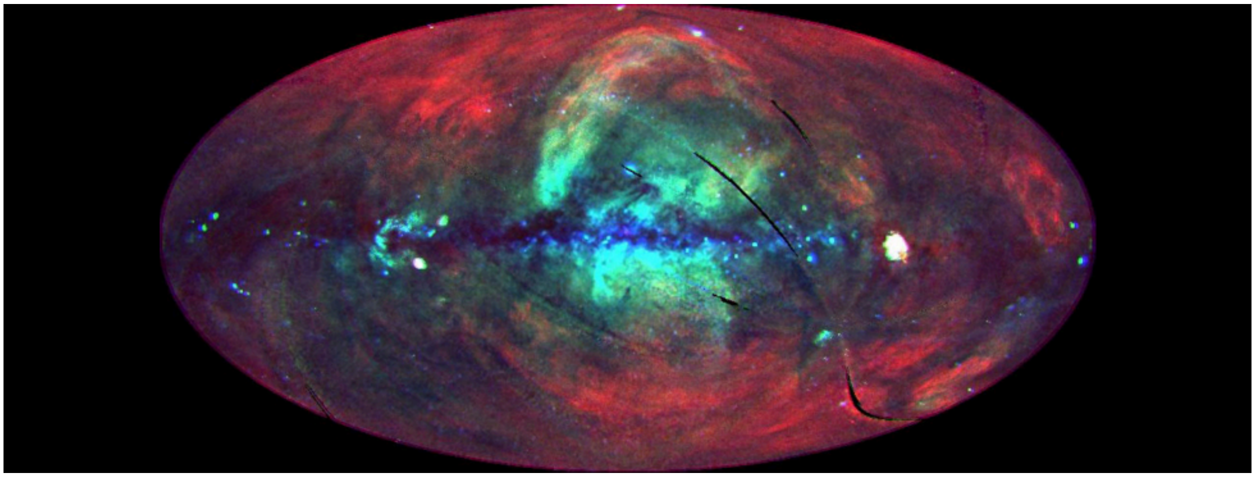


Figure 13.13 This is a map of the sky tuned to certain types of X-rays (seen from above Earth's atmosphere). The map tilts the sky so that the disk of our Milky Way Galaxy runs across its center. It was constructed and artificially colored from data gathered by the European ROSAT satellite. Each color (red, yellow, and blue) shows X-rays of different frequencies or energies. For example, red outlines the glow from a hot local bubble of gas all around us, blown by one or more exploding stars in our cosmic vicinity. Yellow and blue show more distant sources of X-rays, such as remnants of other exploded stars or the active center of our Galaxy (in the middle of the picture). (credit: modification of work by NASA)

Radiation intermediate between X-rays and visible light is **ultraviolet** (meaning higher energy than violet). Outside the world of science, ultraviolet light is sometimes called "black light" because our eyes cannot see it. Ultraviolet radiation is mostly blocked by the ozone layer of Earth's atmosphere, but a small fraction of ultraviolet rays from our Sun do penetrate to cause sunburn or, in extreme cases of overexposure, skin cancer in human beings. Ultraviolet astronomy is also best done from space.

Electromagnetic radiation with wavelengths between roughly 400 and 700 nm is called **visible light** because these are the waves that human vision can perceive. This is also the band of the electromagnetic spectrum that most readily reaches Earth's surface. These two observations are not coincidental: human eyes evolved to see the kinds of waves that arrive from the Sun most effectively. Visible light penetrates Earth's atmosphere effectively, except when it is temporarily blocked by clouds.

Between visible light and radio waves are the wavelengths of **infrared** or heat radiation. Astronomer William Herschel first discovered infrared in 1800 while trying to measure the temperatures of different colors of sunlight spread out into a spectrum. He noticed that when he accidently positioned his thermometer beyond the reddest color, it still registered heating due to some invisible energy coming from the Sun. This was the first hint about the existence of the other (invisible) bands of the electromagnetic spectrum, although it would take many decades for our full understanding to develop.

A heat lamp radiates mostly infrared radiation, and the nerve endings in our skin are sensitive to this band of the electromagnetic spectrum. Infrared waves are absorbed by water and carbon dioxide molecules, which are more concentrated low in Earth's atmosphere. For this reason, infrared astronomy is best done from high mountaintops, high-flying airplanes, and spacecraft.

After infrared comes the familiar **microwave**, used in short-wave communication and microwave ovens. (Wavelengths vary from 1 millimeter to 1 meter and are absorbed by water vapor, which makes them effective in heating foods.) The "micro-" prefix refers to the fact that microwaves are small in comparison to radio waves, the next on the spectrum. You may remember that tea—which is full of water—heats up quickly in your microwave oven, while a ceramic cup—from which water has been removed by baking—stays cool in comparison.

All electromagnetic waves longer than microwaves are called **radio waves**, but this is so broad a category that we generally divide it into several subsections. Among the most familiar of these are radar waves, which are used in radar guns by traffic officers to determine vehicle speeds, and AM radio waves, which were the first to be developed for broadcasting. The wavelengths of these different categories range from over a meter to hundreds of meters, and other radio radiation can have wavelengths as long as several kilometers.

With such a wide range of wavelengths, not all radio waves interact with Earth's atmosphere in the same way. FM and TV waves are not absorbed and can travel easily through our atmosphere. AM radio waves are absorbed or reflected by a layer in Earth's atmosphere called the ionosphere (the ionosphere is a layer of charged particles at the top of our atmosphere, produced by interactions with sunlight and charged particles that are ejected from the Sun).

We hope this brief survey has left you with one strong impression: although visible light is what most people associate with astronomy, the light that our eyes can see is only a tiny fraction of the broad range of waves generated in the universe. Today, we understand that judging some astronomical phenomenon by using only the light we can see is like hiding under the table at a big dinner party and judging all the guests by nothing but their shoes. There's a lot more to each person than meets our eye under the table. It is very important for those who study astronomy today to avoid being "visible light chauvinists"—to respect only the information seen by their eyes while ignoring the information gathered by instruments sensitive to other bands of the electromagnetic spectrum.

Table 13.1 summarizes the bands of the electromagnetic spectrum and indicates the temperatures and typical astronomical objects that emit each kind of electromagnetic radiation. While at first, some of the types of radiation listed in the table may seem unfamiliar, you will get to know them better as your astronomy course continues. You can return to this table as you learn more about the types of objects astronomers study.

Types of Electromagnetic Radiation

Type of Radiation	Wavelength Range (nm)	Radiated by Objects at This Temperature	Typical Sources
Gamma rays	Less than 0.01	More than 10^8 K	Produced in nuclear reactions; require very high-energy processes
X-rays	0.01–20	10^6–10^8 K	Gas in clusters of galaxies, supernova remnants, solar corona
Ultraviolet	20–400	10^4–10^6 K	Supernova remnants, very hot stars
Visible	400–700	10^3–10^4 K	Stars
Infrared	10^3–10^6	10–10^3 K	Cool clouds of dust and gas, planets, moons
Microwave	10^6–10^9	Less than 10 K	Active galaxies, pulsars, cosmic background radiation
Radio	More than 10^9	Less than 10 K	Supernova remnants, pulsars, cold gas

Table 13.1

Radiation and Temperature

Some astronomical objects emit mostly infrared radiation, others mostly visible light, and still others mostly ultraviolet radiation. What determines the type of electromagnetic radiation emitted by the Sun, stars, and other dense astronomical objects? The answer often turns out to be their *temperature*.

At the microscopic level, everything in nature is in motion. A solid is composed of molecules and atoms in continuous vibration: they move back and forth in place, but their motion is much too small for our eyes to make out. A gas consists of atoms and/or molecules that are flying about freely at high speed, continually bumping into one another and bombarding the surrounding matter. The hotter the solid or gas, the more rapid the motion of its molecules or atoms. The temperature of something is thus a measure of the average motion energy of the particles that make it up.

This motion at the microscopic level is responsible for much of the electromagnetic radiation on Earth and in the universe. As atoms and molecules move about and collide, or vibrate in place, their electrons give off electromagnetic radiation. The characteristics of this radiation are determined by the temperature of those atoms and molecules. In a hot material, for example, the individual particles vibrate in place or move rapidly from collisions, so the emitted waves are, on average, more energetic. And recall that higher energy waves have a higher frequency. In very cool material, the particles have low-energy atomic and molecular motions and thus generate lower-energy waves.

Check out the **NASA briefing (https://science.nasa.gov/ems/01_intro)** or **NASA's 5-minute introductory video (https://openstax.org/l/30elmagsp2)** to learn more about the electromagnetic spectrum.

Radiation Laws

To understand, in more quantitative detail, the relationship between temperature and electromagnetic radiation, we imagine an idealized object called a **blackbody**. Such an object (unlike your sweater or your astronomy instructor's head) does not reflect or scatter any radiation, but absorbs all the electromagnetic energy that falls onto it. The energy that is absorbed causes the atoms and molecules in it to vibrate or move around at increasing speeds. As it gets hotter, this object will radiate electromagnetic waves until absorption and radiation are in balance. We want to discuss such an idealized object because, as you will see, stars behave in very nearly the same way.

The radiation from a blackbody has several characteristics, as illustrated in **Figure 13.14**. The graph shows the power emitted at each wavelength by objects of different temperatures. In science, the word *power* means the energy coming off per second (and it is typically measured in *watts*, which you are probably familiar with from buying lightbulbs).

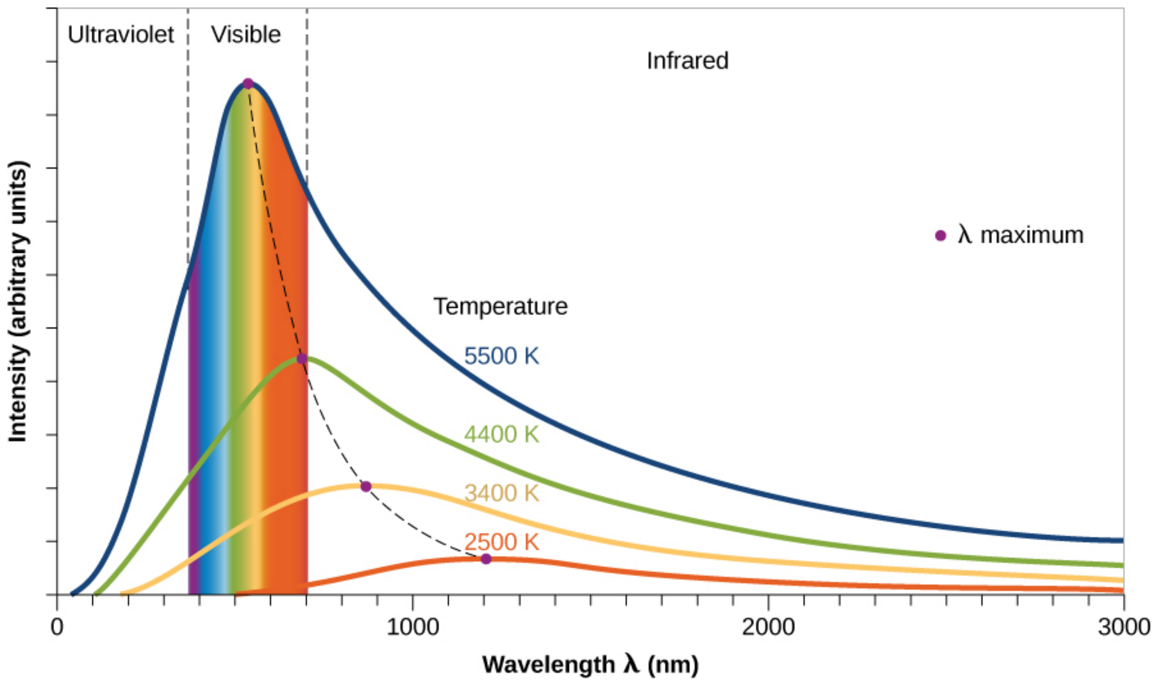


Figure 13.14 This graph shows in arbitrary units how many photons are given off at each wavelength for objects at four different temperatures. The wavelengths corresponding to visible light are shown by the colored bands. Note that at hotter temperatures, more energy (in the form of photons) is emitted at all wavelengths. The higher the temperature, the shorter the wavelength at which the peak amount of energy is radiated (this is known as Wien's law).

First of all, notice that the curves show that, at each temperature, our blackbody object emits radiation (photons) at all wavelengths (all colors). This is because in any solid or denser gas, some molecules or atoms vibrate or move between collisions slower than average and some move faster than average. So when we look at the electromagnetic waves emitted, we find a broad range, or spectrum, of energies and wavelengths. More energy is emitted at the average vibration or motion rate (the highest part of each curve), but if we have a large number of atoms or molecules, some energy will be detected at each wavelength.

Second, note that an object at a higher temperature emits more power at all wavelengths than does a cooler one. In a hot gas (the taller curves in **Figure 13.14**), for example, the atoms have more collisions and give off more energy. In the real world of stars, this means that hotter stars give off more energy at every wavelength than do cooler stars.

Third, the graph shows us that the higher the temperature, the shorter the wavelength at which the maximum power is emitted. Remember that a shorter wavelength means a higher frequency and energy. It makes sense, then, that hot objects give off a larger fraction of their energy at shorter wavelengths (higher energies) than do cool objects. You may have observed examples of this rule in everyday life. When a burner on an electric stove is turned on low, it emits only heat, which is infrared radiation, but does not glow with visible light. If the burner is set to a higher temperature, it starts to glow a dull red. At a still-higher setting, it glows a brighter orange-red (shorter wavelength). At even higher temperatures, which cannot be reached with ordinary stoves, metal can appear brilliant yellow or even blue-white.

We can use these ideas to come up with a rough sort of "thermometer" for measuring the temperatures of stars. Because many stars give off most of their energy in visible light, the color of light that dominates a star's appearance is a rough indicator of its temperature. If one star looks red and another looks blue, which one has the higher temperature? Because blue is the shorter-wavelength color, it is the sign of a hotter star. (Note that the temperatures we associate with different colors in science are not the same as the ones artists use. In art, red is often called a "hot" color and blue a "cool" color. Likewise, we commonly see red on faucet or air conditioning controls to indicate hot temperatures and blue to indicate cold temperatures. Although these are common uses to us in daily life, in nature, it's the other way around.)

We can develop a more precise star thermometer by measuring how much energy a star gives off at each wavelength and by constructing diagrams like **Figure 13.14**. The location of the peak (or maximum) in the power curve of each star can tell us its temperature. The average temperature at the surface of the Sun, which is where the radiation that we see is emitted, turns out to be 5800 K. (Throughout this section, we use the kelvin or absolute temperature scale. On this scale, water freezes at 273 K and boils at 373 K. All molecular motion ceases at 0 K.) There are stars cooler than the Sun and stars hotter than the Sun.

The wavelength at which maximum power is emitted can be calculated according to the equation

Wien's Law

$$\lambda_{max} = \frac{2.9 \times 10^6}{T} \tag{13.1}$$

where the wavelength is in nanometers (one billionth of a meter) and the temperature is in K (the constant 2.9×10^6 has units of nm × K). This relationship is called **Wien's law**. For the Sun, the wavelength at which the maximum energy is emitted is 520 nanometers, which is near the middle of that portion of the electromagnetic spectrum called visible light. Characteristic temperatures of other astronomical objects, and the wavelengths at which they emit most of their power, are listed in **Table 13.1**.

Example 13.3

Calculating the Temperature of a Blackbody

We can use Wien's law to calculate the temperature of a star provided we know the wavelength of peak intensity for its spectrum. If the emitted radiation from a red dwarf star has a wavelength of maximum power at 1200 nm, what is the temperature of this star, assuming it is a blackbody?

Solution

Solving Wien's law for temperature gives:

$$T = \frac{2.9 \times 10^6 \text{ nm K}}{\lambda_{max}} = \frac{2.9 \times 10^6 \text{ nm K}}{1200 \text{ nm}} = 2400 \text{ K}$$

Since this star has a peak wavelength that is at a shorter wavelength (in the ultraviolet part of the spectrum) than that of our Sun (in the visible part of the spectrum), it should come as no surprise that its surface temperature is much hotter than our Sun's.

13.2 What is the temperature of a star whose maximum light is emitted at a much shorter wavelength of 290 nm?

We can also describe our observation that hotter objects radiate more power at all wavelengths in a mathematical form. If we sum up the contributions from all parts of the electromagnetic spectrum, we obtain the total energy emitted by a blackbody.

What we usually measure from a large object like a star is the **energy flux**, the power emitted per square meter. The word *flux* means "flow" here: we are interested in the flow of power into an area (like the area of a telescope mirror). It turns out that the energy flux from a blackbody at temperature T is proportional to the fourth power of its absolute temperature. This relationship is known as the **Stefan-Boltzmann law** and can be written in the form of an equation as

Stefan-Boltzmann Law

$$F = \sigma T^4 \quad (13.2)$$

where F stands for the energy flux (in units of watts per square meter), T is given in Kelvins, and σ (Greek letter sigma) is a constant number (5.67×10^{-8} W/m^2-K^4).

Notice how impressive this result is. Increasing the temperature of a star would have a tremendous effect on the power it radiates. If the Sun, for example, were twice as hot—that is, if it had a temperature of 11,600 K—it would radiate 2^4, or 16 times more power than it does now. Tripling the temperature would raise the power output 81 times. Hot stars really shine away a tremendous amount of energy.

Example 13.4

Calculating the Power of a Star

While energy flux tells us how much power a star emits per square meter, we would often like to know how much total power is emitted by the star. We can determine that by multiplying the energy flux by the number of square meters on the surface of the star. Stars are mostly spherical, so we can use the formula $4\pi R^2$ for the surface area, where R is the radius of the star. The total power emitted by the star (which we call the star's "absolute luminosity") can be found by multiplying the formula for energy flux and the formula for the surface area:

Luminosity of a Star

$$L = 4\pi R^2 \sigma T^4 \quad (13.3)$$

Note the use of the symbol L, which comes from the fact that, in astronomy, the power of a star is called the **luminosity** (as we have seen in the section on **The Brightness of Stars**).

Two stars have the same size and are the same distance from us. Star A has a surface temperature of 6000 K, and star B has a surface temperature twice as high, 12,000 K. How much more luminous is star B compared to star A?

Solution

$$L_{\text{A}} = 4\pi {R_{\text{A}}}^2 \sigma {T_{\text{A}}}^4 \text{ and } L_{\text{B}} = 4\pi {R_{\text{B}}}^2 \sigma {T_{\text{B}}}^4$$

Take the ratio of the luminosity of Star A to Star B:

$$\frac{L_{\text{B}}}{L_{\text{A}}} = \frac{4\pi {R_{\text{B}}}^2 \sigma {T_{\text{B}}}^4}{4\pi {R_{\text{A}}}^2 \sigma {T_{\text{A}}}^4} = \frac{{R_{\text{B}}}^2 {T_{\text{B}}}^4}{{R_{\text{A}}}^2 {T_{\text{A}}}^4}$$

Because the two stars are the same size, $R_{\text{A}} = R_{\text{B}}$, leaving

$$\frac{{T_{\text{B}}}^4}{{T_{\text{A}}}^4} = \frac{(12{,}000\text{ K})^4}{(6000\text{ K})^4} = 2^4 = 16$$

13.3 Two stars with identical diameters are the same distance away. One has a temperature of 8700 K and the other has a temperature of 2900 K. Which is brighter? How much brighter is it?

13.4 | The Brightness of Stars

Learning Objectives

- Explain how and why the amount of light we see from an object depends upon its distance
- Explain the difference between luminosity and apparent brightness
- Understand how astronomers specify brightness with magnitudes

Luminosity

Perhaps the most important characteristic of a star is its **luminosity**—the total amount of energy at all wavelengths that it emits per second. Earlier, we saw that the Sun puts out a tremendous amount of energy every second. (And there are stars far more luminous than the Sun out there.) To make the comparison among stars easy, astronomers express the luminosity of other stars in terms of the Sun's luminosity. For example, the luminosity of Sirius is about 25 times that of the Sun. We use the symbol L_{Sun} to denote the Sun's luminosity; hence, that of Sirius can be written as 25 L_{Sun}. In a later chapter, we will see that if we can measure how much energy a star emits and we also know its mass, then we can calculate how long it can continue to shine before it exhausts its nuclear energy and begins to die.

Propagation of Light

Let's think for a moment about how light from a lightbulb moves through space. As waves expand, they travel away from the bulb, not just toward your eyes but in all directions. They must therefore cover an ever-widening space. Yet the total amount of light available can't change once the light has left the bulb. This means that, as the same expanding shell of light covers a larger and larger area, there must be less and less of it in any given place. Light (and all other electromagnetic radiation) gets weaker and weaker as it gets farther from its source.

The increase in the area that the light must cover is proportional to the square of the distance that the light has traveled (Figure 13.15). If we stand twice as far from the source, our eyes will intercept two-squared (2 × 2), or four times less light. If we stand 10 times farther from the source, we get 10-squared, or 100 times less light. You can see how this weakening means trouble for sources of light at astronomical distances. One of the nearest stars, Alpha Centauri A, emits about the same total energy as the Sun. But it is about 270,000 times farther away, and so it appears about 73 billion times fainter. No wonder the stars, which close-up would look more or less like the Sun, look like faint pinpoints of light from far away.

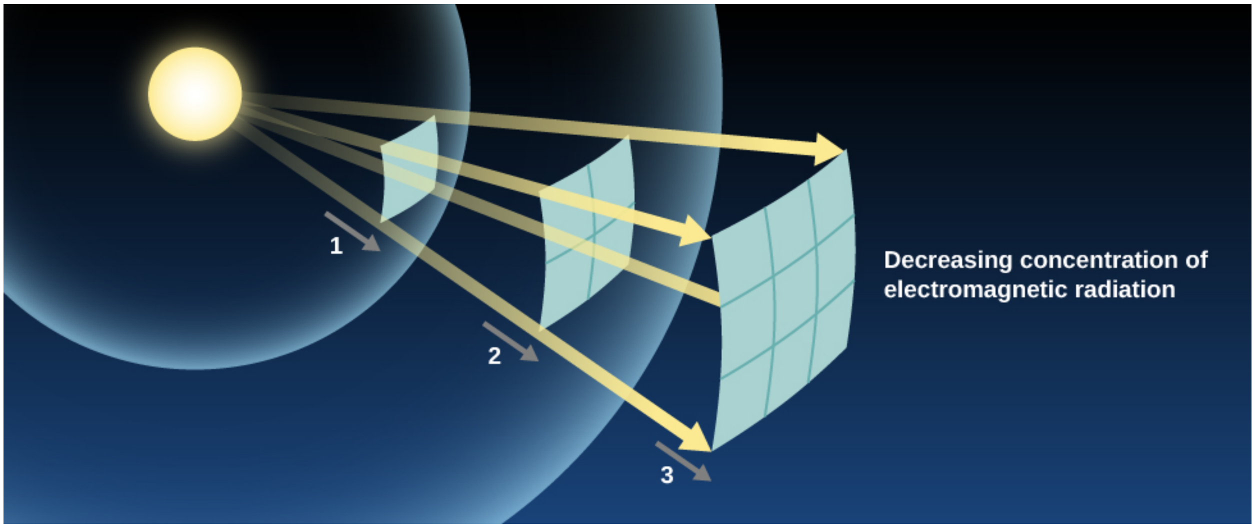


Figure 13.15 As light radiates away from its source, it spreads out in such a way that the energy per unit area (the amount of energy passing through one of the small squares) decreases as the square of the distance from its source.

This idea—that the apparent brightness of a source (how bright it looks to us) gets weaker with distance in the way we have described—is known as the **inverse square law** for light propagation. In this respect, the propagation of light is similar to the effects of gravity. Remember that the force of gravity between two attracting masses is also inversely proportional to the square of their separation.

Inverse Square Law for Light

As an equation:

$$b = \frac{L}{4\pi d^2} \tag{13.4}$$

Example 13.5

The Inverse Square Law for Light

The intensity of a 120-W lightbulb observed from a distance 2 m away is 2.4 W/m^2. What would be the intensity if this distance was doubled?

Solution

If we move twice as far away, then the answer will change according to the inverse square of the distance, so the new intensity will be $(1/2)^2 = 1/4$ of the original intensity, or 0.6 W/m^2.

Check Your Learning

How many times brighter or fainter would a star appear if it were moved to:

a. twice its present distance?
b. ten times its present distance?
c. half its present distance?

Answer:

a. $\left(\frac{1}{2}\right)^2 = \frac{1}{4}$; b. $\left(\frac{1}{10}\right)^2 = \frac{1}{100}$; c. $\left(\frac{1}{1/2}\right)^2 = 4$

Apparent Brightness

Astronomers are careful to distinguish between the luminosity of the star (the total energy output) and the amount of energy that happens to reach our eyes or a telescope on Earth. Stars are democratic in how they produce radiation; they emit the same amount of energy in every direction in space. Consequently, only a minuscule fraction of the energy given off by a star actually reaches an observer on Earth. We call the amount of a star's energy that reaches a given area (say, one square meter) each second here on Earth its **apparent brightness**. If you look at the night sky, you see a wide range of apparent brightnesses among the stars. Most stars, in fact, are so dim that you need a telescope to detect them.

If all stars were the same luminosity—if they were like standard bulbs with the same light output—we could use the difference in their apparent brightnesses to tell us something we very much want to know: how far away they are. Imagine you are in a big concert hall or ballroom that is dark except for a few dozen 25-watt bulbs placed in fixtures around the walls. Since they are all 25-watt bulbs, their luminosity (energy output) is the same. But from where you are standing in one corner, they do *not* have the same apparent brightness. Those close to you appear brighter (more of their light reaches your eye), whereas those far away appear dimmer (their light has spread out more before reaching you). In this way, you can tell which bulbs are closest to you. In the same way, if all the stars had the same luminosity, we could immediately infer that the brightest-appearing stars were close by and the dimmest-appearing ones were far away.

To pin down this idea more precisely, we know exactly how light fades with increasing distance. The energy we receive is inversely proportional to the square of the distance. If, for example, we have two stars of the same luminosity and one is twice as far away as the other, it will look four times dimmer than the closer one. If it is three times farther away, it will look nine (three squared) times dimmer, and so forth.

Alas, the stars do not all have the same luminosity. (Actually, we are pretty glad about that because having many different types of stars makes the universe a much more interesting place.) But this means that if a star looks dim in the sky, we cannot tell whether it appears dim because it has a low luminosity but is relatively nearby, or because it has a high luminosity but is very far away. To measure the luminosities of stars, we must first compensate for the dimming effects of distance on light, and to do that, we must know how far away they are. Distance is among the most difficult of all astronomical measurements. We will return to how it is determined after we have learned more about the stars. For now, we will describe how astronomers specify the apparent brightness of stars.

The Magnitude Scale

The process of measuring the apparent brightness of stars is called *photometry* (from the Greek *photo* meaning "light" and *–metry* meaning "to measure"). Astronomical photometry began with Hipparchus. Around 150 B.C.E., he erected an observatory on the island of Rhodes in the Mediterranean. There he prepared a catalog of nearly 1000 stars that included not only their positions but also estimates of their apparent brightnesses.

Hipparchus did not have a telescope or any instrument that could measure apparent brightness accurately, so he simply made estimates with his eyes. He sorted the stars into six brightness categories, each of which he called a **magnitude**. He referred to the brightest stars in his catalog as first-magnitudes stars, whereas those so faint he could barely see them were sixth-magnitude stars. During the nineteenth century, astronomers attempted to make the scale more precise by establishing exactly how much the apparent brightness of a sixth-magnitude star differs from that of a first-magnitude star. Measurements showed that we receive about 100 times more light from a first-magnitude star than from a sixth-magnitude star. Based on this measurement, astronomers then defined an accurate magnitude system in which a difference of five magnitudes corresponds exactly to a brightness ratio of 100:1. In addition, the magnitudes of stars are decimalized; for example, a star isn't just a "second-magnitude star," it has a magnitude of 2.0 (or 2.1, 2.3, and so forth). So what number is it that, when multiplied together five times, gives you this factor of 100? Play on your calculator and see if you can get it. The answer turns out to be about 2.5, which is the fifth root of 100. This means that a magnitude 1.0 star and a magnitude 2.0 star differ in brightness by a factor of about 2.5. Likewise, we receive about 2.5 times as much light from a magnitude 2.0 star as from a magnitude 3.0 star. What about the difference between a magnitude 1.0 star and a magnitude 3.0 star? Since the difference is 2.5 times for each "step" of magnitude, the total difference in brightness is $2.5 \times 2.5 = 6.25$ times.

Here are a few rules of thumb that might help those new to this system. If two stars differ by 0.75 magnitudes, they differ by a factor of about 2 in brightness. If they are 2.5 magnitudes apart, they differ in brightness by a factor of 10, and a 4-magnitude difference corresponds to a difference in brightness of a factor of 40.You might be saying to yourself at this point, "Why do astronomers continue to use this complicated system from more than 2000 years ago?" That's an excellent question and, as we shall discuss, astronomers today can use other ways of expressing how bright a star looks. But because this system is still used in many books, star charts, and computer apps, we felt we had to introduce students to it (even though we were very tempted to leave it out.)

Apparent Magnitude

In fact, the correct term for this quantity is the **apparent magnitude**, because we are talking about the magnitude as it appears to (or is measured by) an observer on Earth. The symbol for apparent magnitude us usually a lower case m.

For reasons that will become clear shortly, to avoid ambiguity, it would be best if astronomers always used the terminology of **apparent magnitude**. Unfortunately, this quantity is sometimes refereed to simply as the **magnitude** of a star or other astronomical object. We shall treat the two terms as synonymous.

The brightest stars, those that were traditionally referred to as first-magnitude stars, actually turned out (when measured accurately) not to be identical in brightness. For example, the brightest star in the sky, Sirius, sends us about 10 times as much light as the average first-magnitude star. On the modern magnitude scale, Sirius, the star with the brightest apparent magnitude, has been assigned a magnitude of −1.5. Other objects in the sky can appear even brighter. Venus at its brightest is of magnitude −4.4, while the Sun has a magnitude of −26.8. **Figure 13.16** shows the range of observed magnitudes from the brightest to the faintest, along with the actual magnitudes of several well-known objects. The important fact to remember when using magnitude is that the system goes backward: the *larger* the magnitude, the *fainter* the object you are observing.

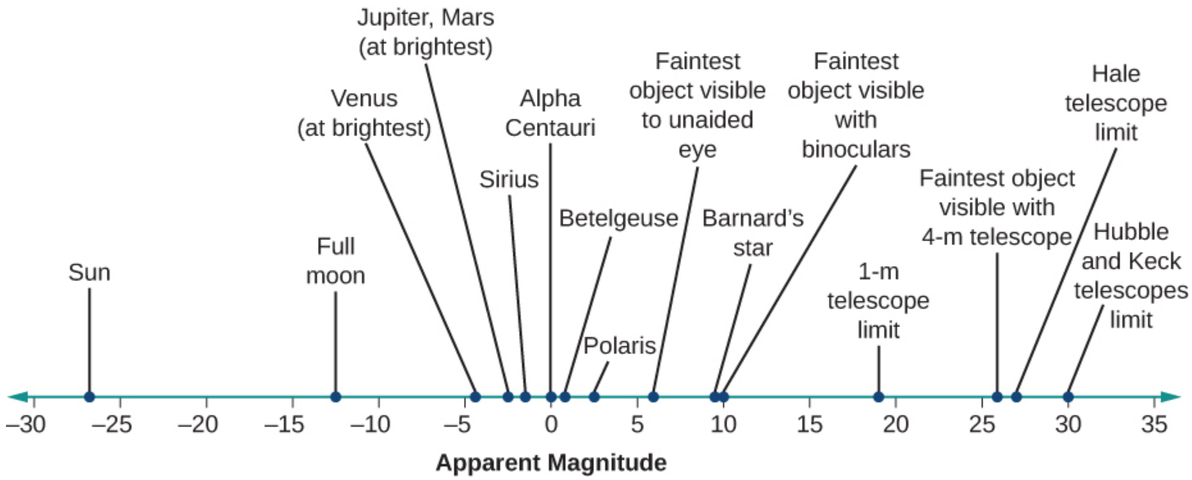


Figure 13.16 The faintest magnitudes that can be detected by the unaided eye, binoculars, and large telescopes are also shown.

The Modern Magnitude Equation

Even scientists can't calculate fifth roots in their heads, so astronomers have summarized the above discussion in an equation to help calculate the difference in brightness for stars with different magnitudes. If m_1 and m_2 are the magnitudes of two stars, then we can calculate the ratio of their brightness $\left(\frac{b_2}{b_1}\right)$ using this equation:

$$m_1 - m_2 = 2.5 \log\left(\frac{b_2}{b_1}\right) \quad \text{or} \quad \frac{b_2}{b_1} = 2.5^{m_1 - m_2} \tag{13.5}$$

Here is another way to write this equation:

$$\frac{b_2}{b_1} = \left(100^{0.2}\right)^{m_1 - m_2}$$

Let's do a real example, just to show how this works. Imagine that an astronomer has discovered something special about a dim star (magnitude 8.5), and she wants to tell her students how much dimmer the star is than Sirius. Star 1 in the equation will be our dim star and star 2 will be Sirius.

Solution

Remember, Sirius has a magnitude of −1.5. In that case:

$$\begin{aligned}\frac{b_2}{b_1} &= \left(100^{0.2}\right)^{8.5 - (-1.5)} = \left(100^{0.2}\right)^{10} \\ &= (100)^2 = 100 \times 100 = 10{,}000\end{aligned}$$

13.4 It is a common misconception that Polaris (magnitude 2.0) is the brightest star in the sky, but, as we saw, that distinction actually belongs to Sirius (magnitude −1.5). How does Sirius' apparent brightness compare to that of Polaris?

(Hint: If you only have a basic calculator, you may wonder how to take 100 to the 0.7th power. But this is something you can ask Google to do. Google now accepts mathematical questions and will answer them. So try it for yourself. Ask Google, "What is 100 to the 0.7th power?")

Our calculation shows that Sirius' apparent brightness is 25 times greater than Polaris' apparent brightness.

Absolute Magnitude

We already understand that objects with a large apparent brightness are not necessarily the most luminous objects (and vice versa), because of the fact that the objects may lie at different distances from Earth. The core ideas here may also be expressed using the concept of an **absolute magnitude**.

The **absolute magnitude** of an astronomical object is defined as the magnitude that would be measured by an observer located exactly 10 parsecs (about 32.6 light-years) away from that object.

By defining absolute magnitude in this way, we see that it places all objects the same distance from a hypothetical observer. Thus, a comparison of the absolute magnitudes of any two object is a direct comparison of their luminosities. As explained before, this makes for a logarithmic (or power law) comparison when expressed mathematically. The algebraic symbol used for the absolute magnitude of an object is a capital M.

By analogy with our discussion of brightness above, we can calculate the difference in luminosity for two stars with different absolute magnitudes. If M_1 and M_2 are the absolute magnitudes of two stars, then, by analogy with **Equation 13.5** we can calculate the ratio of their luminosities $\left(\frac{L_2}{L_1}\right)$ as:

$$\frac{L_2}{L_1} = \left(100^{0.2}\right)^{M_1 - M_2} \quad (13.6)$$

Now, we should probably point out the obvious here, namely that **absolute magnitude** is more of a theoretical quantity than an experimental one. After all, to actually measure it we would have to ride a spaceship to a location that was precisely 10 parsecs away for the object we want to study, and only then perform a measurement of its apparent magnitude. Still, as we shall see in subsequent chapters, there are various theories that allow us to make good estimates of the absolute magnitude of a star or galaxy. So, it has been a useful quantity for astronomers, in popular use for more than a century.

Let's return to the stars Sirius and Polaris for a numerical example. Sirius has an absolute magnitude of 1.4, while the absolute magnitude of Polaris is -3.6. So, the ratio of their luminosities is:

$$\frac{L_{\text{Polaris}}}{L_{\text{Sirius}}} = \left(100^{0.2}\right)^{M_{\text{Sirius}} - M_{\text{Polaris}}} = \left(100^{0.2}\right)^{1.4 - (-3.6)} = \left(100^{0.2}\right)^{5} = 100^1 = 100$$

So, Polaris is 100 times more luminous than Sirius. But, how can that be? (Sirius is much brighter in our sky than Polaris!) The answer is the obvious one: Polaris is much farther away from Earth (around 400 light years) than Sirius (which is less than 10 light years away).

Other Units of Brightness

Although the magnitude scale is still used for visual astronomy, it is not used at all in newer branches of the field. In radio astronomy, for example, no equivalent of the magnitude system has been defined. Rather, radio astronomers measure the amount of energy being collected each second by each square meter of a radio telescope and express the brightness of each source in terms of, for example, watts per square meter.

Similarly, most researchers in the fields of infrared, X-ray, and gamma-ray astronomy use energy per area per second rather than magnitudes to express the results of their measurements. Nevertheless, astronomers in all fields are careful to distinguish between the *luminosity* of the source (even when that luminosity is all in X-rays) and the amount of energy that happens to reach us on Earth. After all, the luminosity is a really important characteristic that tells us a lot about the object in question, whereas the energy that reaches Earth is an accident of cosmic geography.

To make the comparison among stars easy, in this text, we avoid the use of magnitudes as much as possible and will express the luminosity of other stars in terms of the Sun's luminosity. For example, the luminosity of Sirius is 25 times that of the Sun. We use the symbol L_{Sun} to denote the Sun's luminosity; hence, that of Sirius can be written as 25 L_{Sun}.

Color Filters, Magnitudes, and Measurements of Stellar Brightness

As we have said, there have been many measurements of the brightness of stars expressed in terms of an **apparent magnitude**. The large number of such measurements is in part due to the fact that **photometry**, a simple, direct measurement of apparent brightness, is significantly easier to carry out than **spectroscopy**, the complete determination of intensity vs. wavelength, like the temperature curves in **Figure 13.14**. (You will learn more about the significance of such measurements in the chapter on **Spectroscopy**.)

As we will learn in the section on **Colors of Stars**, that property is very important. (We have already seen that Wien's Law provides a connection between a star's color and its temperature.) In order to make a crude determination of the color of a star, it is possible to make photometric measurements of its apparent magnitude by passing the star's light through various colored filters, each one designed to transmit light over a limited range of wavelengths. The standard set of filters are labeled U,B,V, R and I. The initials stand for ultraviolet, blue, visible, infrared and far-infrared respectively.

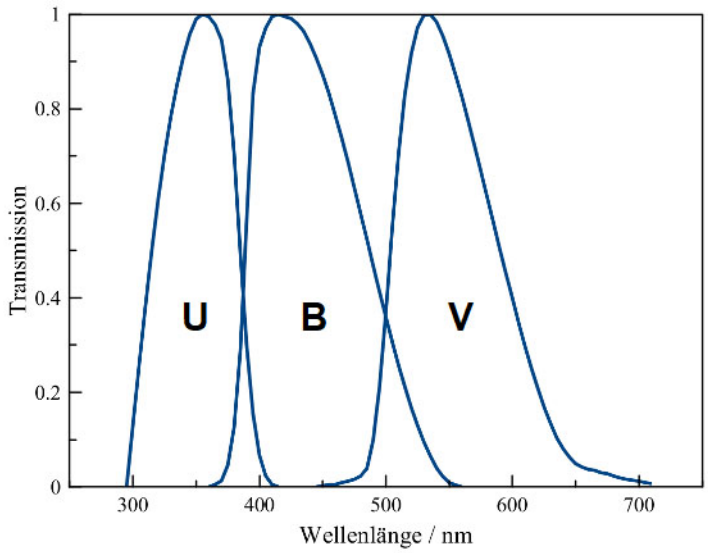


Figure 13.17 These curves show the wavelength distributions for light transmission through U, B and V filters. (Note that the horizontal axis label "Wellenlänge" is German for "Wavelength".) credit: Michael Oestreicher, Wikimedia Commons

Figure 13.17 shows the wavelength ranges for transmission by U, B and V filters. As can be seen by comparison with **Figure 13.14**, the U filter passes ultraviolet light, the B filter passes primarily blue light, and the V filter passes most visible wavelengths. Often, the apparent magnitude of an astronomical object measured using one of these filters is referred to by the filter letter, e.g. a "B magnitude".

Suppose that we make two measurements of the same star, one through a V filter, and the other through a B filter. If we call the two measured magnitudes V and B respectively, then let's consider the significance of the quantity $B - V$.

Recalling that apparent magnitudes are a "backwards quantity", i.e. the smaller the number the brighter the star, we see that B-V is a measure of the "redness" of a star. Why? Suppose that we have two stars with equal visible brightness. Then, their V magnitudes will be the same. But, if one is a very blue star, and the other a very red star, the blue star will have a smaller B magnitude, and therefore a smaller value of B-V. The red star will have a larger B magnitude, and therefore a larger value of B-V. For this reason, the quantity B-V is called a **color index**.

CHAPTER 13 REVIEW

KEY TERMS

absolute magnitude the apparent magnitude that would be measured by an observer located at a distance of 10 parsecs from a luminous object

apparent brightness a measure of the amount of light received by Earth from a star or other object—that is, how bright an object appears in the sky, as contrasted with its luminosity

apparent magnitude synonym for magnitude

blackbody an idealized object that absorbs all electromagnetic energy that falls onto it

color index The mathematical difference *B*-*V* between the B and V magnitudes of a star, which is an indication of its color. A smaller value corresponds to a bluer star, a larger value to a redder star.

electromagnetic radiation radiation consisting of waves propagated through regularly varying electric and magnetic fields and traveling at the speed of light

electromagnetic spectrum the whole array or family of electromagnetic waves, from radio to gamma rays

energy flux the amount of energy passing through a unit area (for example, 1 square meter) per second; the units of flux are watts per square meter

frequency the number of waves that cross a given point per unit time (in radiation)

gamma rays photons (of electromagnetic radiation) of energy with wavelengths no longer than 0.01 nanometer; the most energetic form of electromagnetic radiation

Huygens's principle every point on a wave front is a source of wavelets that spread out in the forward direction at the same speed as the wave itself; the new wave front is a plane tangent to all of the wavelets

infrared electromagnetic radiation of wavelength 10^3–10^6 nanometers; longer than the longest (red) wavelengths that can be perceived by the eye, but shorter than radio wavelengths

inverse square law (for light) the amount of energy (light) flowing through a given area in a given time decreases in proportion to the square of the distance from the source of energy or light

luminosity the rate at which a star or other object emits electromagnetic energy into space; the total power output of an object

magnitude an older system of measuring the amount of light we receive from a star or other luminous object; the larger the magnitude, the less radiation we receive from the object

microwave electromagnetic radiation of wavelengths from 1 millimeter to 1 meter; longer than infrared but shorter than radio waves

photon a discrete unit (or "packet") of electromagnetic energy

radio waves all electromagnetic waves longer than microwaves, including radar waves and AM radio waves

Stefan-Boltzmann law a formula from which the rate at which a blackbody radiates energy can be computed; the total rate of energy emission from a unit area of a blackbody is proportional to the fourth power of its absolute temperature: $F = \sigma T^4$

ultraviolet electromagnetic radiation of wavelengths 10 to 400 nanometers; shorter than the shortest visible wavelengths

visible light electromagnetic radiation with wavelengths of roughly 400–700 nanometers; visible to the human eye

wave optics part of optics dealing with the wave aspect of light

wavelength the distance from crest to crest or trough to trough in a wave

Wien's law formula that relates the temperature of a blackbody to the wavelength at which it emits the greatest intensity of radiation

X-rays electromagnetic radiation with wavelengths between 0.01 nanometer and 20 nanometers; intermediate between those of ultraviolet radiation and gamma rays

KEY EQUATIONS

Wien's Law	$\lambda_{\max} = \frac{2.9 \times 10^6}{T}$
Stefan-Boltzmann law	$F = \sigma T^4$
Luminosity of a Star	$L = 4\pi R^2 \sigma T^4$
Apparent brightness	$b = \frac{L}{4\pi d^2}$
Brightness ratio	$\frac{b_2}{b_1} = \left(100^{0.2}\right)^{m_1 - m_2}$
Luminosity ratio	$\frac{L_2}{L_1} = \left(100^{0.2}\right)^{M_1 - M_2}$

SUMMARY

13.1 The Wave Behavior of Light

- James Clerk Maxwell showed that whenever charged particles change their motion, they give off waves of energy.
- Light is one form of this electromagnetic radiation.
- The wavelength of light determines the color of visible radiation.
- Wavelength (λ) is related to frequency (f) and the speed of light (c) by the equation $c = \lambda f$.
- Electromagnetic radiation sometimes behaves like waves, but at other times, it behaves as if it were a particle—a little packet of energy, called a photon.

13.2 Huygens' Principle

- According to Huygens's principle, every point on a wave front is a source of wavelets that spread out in the forward direction at the same speed as the wave itself. The new wave front is tangent to all of the wavelets.
- A mirror reflects an incoming wave at an angle equal to the incident angle, verifying the law of reflection.
- The law of refraction can be explained by applying Huygens's principle to a wave front passing from one medium to another.
- The bending of a wave around the edges of an opening or an obstacle is called diffraction.

13.3 The Electromagnetic Spectrum

- The electromagnetic spectrum consists of gamma rays, X-rays, ultraviolet radiation, visible light, infrared, and radio radiation.
- Many of these wavelengths cannot penetrate the layers of Earth's atmosphere and must be observed from space, whereas others—such as visible light, FM radio and TV—can penetrate to Earth's surface.
- The emission of electromagnetic radiation is intimately connected to the temperature of the source. The higher the temperature of an idealized emitter of electromagnetic radiation, the shorter is the wavelength at which the maximum amount of radiation is emitted. The mathematical equation describing this relationship is known as Wien's law: $\lambda_{max} = (3 \times 10^6)/T$.
- The total power emitted per square meter increases with increasing temperature. The relationship between emitted energy flux and temperature is known as the Stefan-Boltzmann law: $F = \sigma T^4$.
- A star's apparent magnitude may be measured through colored filters that pass light primarily in the ultraviolet (U), blue (B) or visible (V) region of the spectrum.
- The quantity B-V is a color index, with a smaller value corresponding to a bluer star, and a larger value

corresponding to a redder star.

13.4 The Brightness of Stars

- The total energy emitted per second by a star is called its luminosity.
- How bright a star looks from the perspective of Earth is its apparent brightness.
- The apparent brightness of a star depends on both its luminosity and its distance from Earth. Thus, the determination of apparent brightness and measurement of the distance to a star provide enough information to calculate its luminosity.
- The apparent brightnesses of stars are often expressed in terms of magnitudes, which is an old system based on how human vision interprets relative light intensity.

CONCEPTUAL QUESTIONS

13.1 The Wave Behavior of Light

1. What distinguishes one type of electromagnetic radiation from another? What are the main categories (or bands) of the electromagnetic spectrum?

2. What is a wave? Use the terms *wavelength* and *frequency* in your definition.

13.2 Huygens' Principle

3. How do wave effects depend on the size of the object with which the wave interacts? For example, why does sound bend around the corner of a building while light does not?

4. Does Huygens's principle apply to all types of waves?

5. If diffraction is observed for some phenomenon, it is evidence that the phenomenon is a wave. Does the reverse hold true? That is, if diffraction is not observed, does that mean the phenomenon is not a wave?

13.3 The Electromagnetic Spectrum

6. Is your textbook the kind of idealized object (described in section on radiation laws) that absorbs all the radiation falling on it? Explain. How about the black sweater worn by one of your classmates?

7. Explain how we can deduce the temperature of a star by determining its color.

8. Go outside on a clear night, wait 15 minutes for your eyes to adjust to the dark, and look carefully at the brightest stars. Some should look slightly red and others slightly blue. The primary factor that determines the color of a star is its temperature. Which is hotter: a blue star or a red one? Explain

9. Water faucets are often labeled with a red dot for hot water and a blue dot for cold. Given Wien's law, does this labeling make sense?

10. Two stars that reside in a particular star cluster have different values of the color index *B*-*V*. For Star 1, the value of *B*-*V* = -0.5. For Star 2, the value of *B*-*V* = +0.5.

What can you conclude about the colors (and temperatures) of these two stars?

13.4 The Brightness of Stars

11. What two factors determine how bright a star appears to be in the sky?

12. If the star Sirius emits 23 times more energy than the Sun, why does the Sun appear brighter in the sky?

13. What star appears the brightest in the sky (other than the Sun)? The second brightest? Use **Appendix D** to find the answers.

14.

Star	Absolute Magnitude (M)	Apparent Magnitude (m)
Aldebaran	-0.2	+0.9
Alpha Centauri A	+4.4	0.0
Antares	-4.5	+0.9
Canopus	-3.1	-0.7
Regulus	-0.6	+1.4
Spica	-3.6	+0.9

Table 13.2 Bright Stars Apparent and Absolute Magnitudes of Several Bright Stars

Consider the table above. Which of the listed stars appears brightest in the sky? Which is the dimmest? Which is the most luminous? Which is the least luminous?

PROBLEMS

13.1 The Wave Behavior of Light

15. What is the wavelength of the carrier wave of a campus radio station, broadcasting at a frequency of 97.2 MHz (million cycles per second or million hertz)?

16. What is the frequency of a red laser beam, with a wavelength of 670 nm, which your astronomy instructor might use to point to slides during a lecture on galaxies?

17. You go to a dance club to forget how hard your astronomy midterm was. What is the frequency of a wave of ultraviolet light coming from a blacklight in the club, if its wavelength is 150 nm?

13.3 The Electromagnetic Spectrum

18. If the emitted infrared radiation from Pluto, has a wavelength of maximum intensity at 75,000 nm, what is the temperature of Pluto assuming it follows Wien's law?

19. What is the temperature of a star whose maximum light is emitted at a wavelength of 290 nm?

13.4 The Brightness of Stars

20. In **Appendix D**, how much more luminous is the most luminous of the stars than the least luminous?

For **Exercise 13.21** through **Exercise 13.25**, use the equations relating magnitude and apparent brightness.

21. Verify that if two stars have a difference of five magnitudes, this corresponds to a factor of 100 in the ratio $\left(\frac{b_2}{b_1}\right)$; that 2.5 magnitudes corresponds to a factor of 10; and that 0.75 magnitudes corresponds to a factor of 2.

22. As seen from Earth, the Sun has an apparent magnitude of about −26.7. What is the apparent magnitude of the Sun as seen from Saturn, about 10 AU away? (Remember that one AU is the distance from Earth to the Sun and that the brightness decreases as the inverse square of the distance.) Would the Sun still be the brightest star in the sky?

23. An astronomer is investigating a faint star that has recently been discovered in very sensitive surveys of the sky. The star has a magnitude of 16. How much less bright is it than Antares, a star with magnitude roughly equal to 1?

24. The center of a faint but active galaxy has magnitude 26. How much less bright does it look than the very faintest star that our eyes can see, roughly magnitude 6?

25. Do the previous problem again, this time using the information that the Sun is 150,000,000 km away. You will get a very large number of km as your answer. To get a better feeling for how the distances compare, try calculating the time it takes light at a speed of 299,338 km/s to travel from the Sun to Earth and from Alpha Centauri to Earth. For Alpha Centauri, figure out how long the trip will take in years as well as in seconds.

26. Star A and Star B have different apparent brightnesses but identical luminosities. If Star A is 20 light-years away from Earth and Star B is 40 light-years away from Earth, which star appears brighter and by what factor?

27. Star A and Star B have different apparent brightnesses but identical luminosities. Star A is 10 light-years away from Earth and appears 36 times brighter than Star B. How far away is Star B?

28. The star Sirius A has an apparent magnitude of −1.5. Sirius A has a dim companion, Sirius B, which is 10,000 times less bright than Sirius A. What is the apparent magnitude of Sirius B? Can Sirius B be seen with the naked eye?

29. Referring to **Table 13.2**, what is the ratio of the luminosities of the stars Antares and Aldebaran?

14 | SPECTROSCOPY

Figure 14.1 This long time exposure shows the colors of the stars. The circular motion of the stars across the image is provided by Earth’s rotation. The various colors of the stars are caused by their different temperatures. (credit: modification of work by ESO/A.Santerne)

Chapter Outline

Introduction

Here's another view of the night sky. It is obvious that stars do not all appear equally bright, but we understand that this has to do with each star's intrinsic luminosity and its distance away from Earth. However, as we have noted before, the stars we observe are also not all the same color. To understand why, we must delve deeper into the nature and properties of light as an electromagnetic wave. That study will lead us to even more information about stars, including their temperatures and their velocities toward or away from us.

14.1 | Spectroscopy

Learning Objectives

By the end of this section you will be able to:

- Describe the properties of light
- Explain how astronomers learn the composition of a gas by examining its spectral lines
- Discuss the various types of spectra

Electromagnetic radiation carries a lot of information about the nature of stars and other astronomical objects. To extract this information, however, astronomers must be able to study the amounts of energy we receive at different wavelengths of light in fine detail. Let’s examine how we can do this and what we can learn.

Properties of Light

Light exhibits certain behaviors that are important to the design of telescopes and other instruments. For example, light can be *reflected* from a surface. If the surface is smooth and shiny, as with a mirror, the direction of the reflected light beam can be calculated accurately from knowledge of the shape of the reflecting surface. Light is also bent, or *refracted*, when it passes from one kind of transparent material into another—say, from the air into a glass lens.

Reflection and refraction of light are the basic properties that make possible all *optical* instruments (devices that help us to see things better)—from eyeglasses to giant astronomical telescopes. Such instruments are generally combinations of glass lenses, which bend light according to the principles of refraction, and curved mirrors, which depend on the properties of reflection. Small optical devices, such as eyeglasses or binoculars, generally use lenses, whereas large telescopes depend almost entirely on mirrors for their main optical elements. We discussed some astronomical instruments in **Section 31.7**. For now, we turn to another behavior of light, one that is essential for the decoding of light.

In 1672, in the first paper that he submitted to the Royal Society, Sir Isaac Newton described an experiment in which he permitted sunlight to pass through a small hole and then through a prism. Newton found that sunlight, which looks white to us, is actually made up of a mixture of all the colors of the rainbow (**Figure 14.2**).

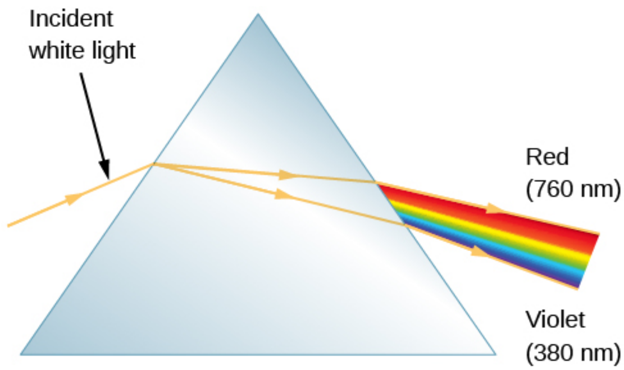


Figure 14.2 When we pass a beam of white sunlight through a prism, we see a rainbow-colored band of light that we call a continuous spectrum.

Figure 14.2 shows how light is separated into different colors with a prism—a piece of glass in the shape of a triangle with refracting surfaces. Upon entering one face of the prism, the path of the light is refracted (bent), but not all of the colors are bent by the same amount. The bending of the beam depends on the wavelength of the light as well as the properties of the material, and as a result, different wavelengths (or colors of light) are bent by different amounts and therefore follow slightly different paths through the prism. The violet light is bent more than the red. As we saw in **Section 30.5**, this phenomenon explains Newton's rainbow experiment.

Upon leaving the opposite face of the prism, the light is bent again and further dispersed. If the light leaving the prism is focused on a screen, the different wavelengths or colors that make up white light are lined up side by side just like a rainbow (**Figure 14.3**). (In fact, a rainbow is formed by the dispersion of light though raindrops; see **The Rainbow** feature box.) Because this array of colors is a spectrum of light, the instrument used to disperse the light and form the spectrum is called a **spectrometer**.

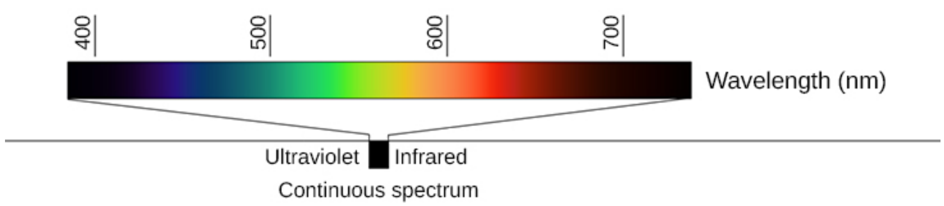


Figure 14.3 When white light passes through a prism, it is dispersed and forms a continuous spectrum of all the colors. Although it is hard to see in this printed version, in a well-dispersed spectrum, many subtle gradations in color are visible as your eye scans from one end (violet) to the other (red).

The Value of Stellar Spectra

When Newton described the laws of refraction and dispersion in optics, and observed the solar spectrum, all he could see was a continuous band of colors. If the spectrum of the white light from the Sun and stars were simply a continuous rainbow of colors, astronomers would have little interest in the detailed study of a star's spectrum once they had learned its average surface temperature. In 1802, however, William Wollaston built an improved spectrometer that included a lens to focus the Sun's spectrum on a screen. With this device, Wollaston saw that the colors were not spread out uniformly, but instead, some ranges of color were missing, appearing as dark bands in the solar spectrum. He mistakenly attributed these lines to natural boundaries between the colors. In 1815, German physicist Joseph Fraunhofer, upon a more careful examination of the solar spectrum, found about 600 such dark lines (missing colors), which led scientists to rule out the boundary hypothesis (**Figure 14.4**).

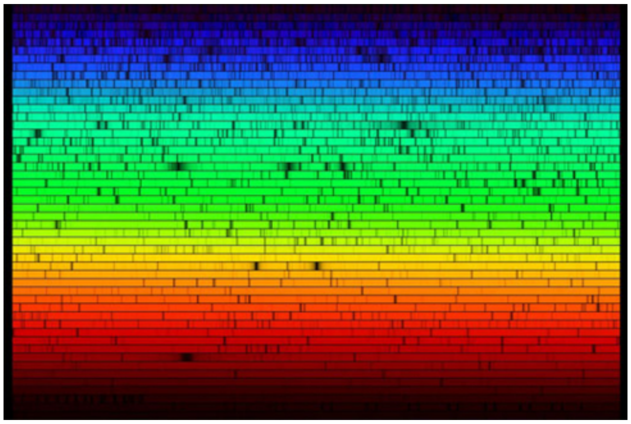


Figure 14.4 Our star's spectrum is crossed by dark lines produced by atoms in the solar atmosphere that absorb light at certain wavelengths. (credit: modification of work by Nigel Sharp, NOAO/National Solar Observatory at Kitt Peak/AURA, and the National Science Foundation)

Later, researchers found that similar dark lines could be produced in the spectra ("spectra" is the plural of "spectrum") of artificial light sources. They did this by passing their light through various apparently transparent substances—usually containers with just a bit of thin gas in them.

These gases turned out not to be transparent at *all* colors: they were quite opaque at a few sharply defined wavelengths. Something in each gas had to be absorbing just a few colors of light and no others. All gases did this, but each different element absorbed a different set of colors and thus showed different dark lines. If the gas in a container consisted of two elements, then light passing through it was missing the colors (showing dark lines) for both of the elements. So it became clear that certain lines in the spectrum "go with" certain elements. This discovery was one of the most important steps forward in the history of astronomy.

What would happen if there were no continuous spectrum for our gases to remove light from? What if, instead, we heated the same thin gases until they were hot enough to glow with their own light? When the gases were heated, a spectrometer revealed no continuous spectrum, but several separate bright lines. That is, these hot gases emitted light only at certain specific wavelengths or colors.

When the gas was pure hydrogen, it would emit one pattern of colors; when it was pure sodium, it would emit a different pattern. A mixture of hydrogen and sodium emitted both sets of spectral lines. The colors the gases emitted when they were heated were the very same colors as those they had absorbed when a continuous source of light was behind them. From such experiments, scientists began to see that different substances showed distinctive *spectral signatures* by which their presence could be detected (**Figure 14.5**). Just as your signature allows the bank to identify you, the unique pattern of colors for each type of atom (its spectrum) can help us identify which element or elements are in a gas.

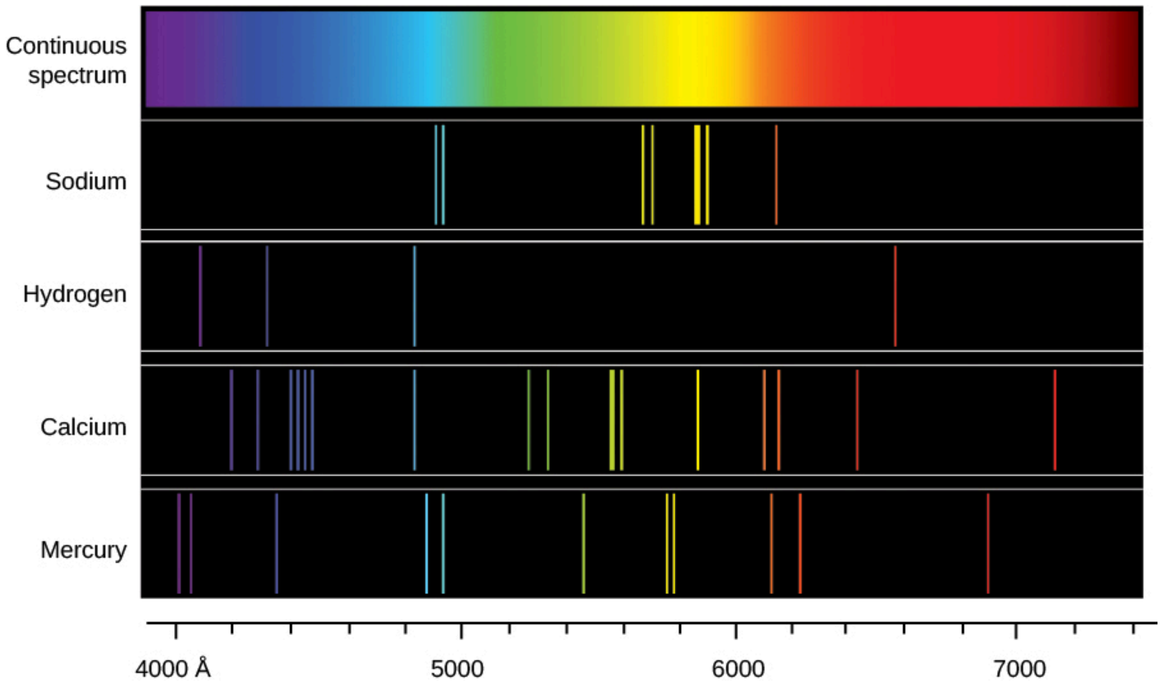


Figure 14.5 Each type of glowing gas (each element) produces its own unique pattern of lines, so the composition of a gas can be identified by its spectrum. The spectra of sodium, hydrogen, calcium, and mercury gases are shown here.

Types of Spectra

In these experiments, then, there were three different types of spectra. A **continuous spectrum** (formed when a solid or very dense gas gives off radiation) is an array of all wavelengths or colors of the rainbow. A continuous spectrum can serve as a backdrop from which the atoms of much less dense gas can absorb light. A dark line, or **absorption spectrum**, consists of a series or pattern of dark lines—missing colors—superimposed upon the continuous spectrum of a source. A bright line, or **emission spectrum**, appears as a pattern or series of bright lines; it consists of light in which only certain discrete wavelengths are present. (**Figure 14.4** shows an absorption spectrum, whereas **Figure 14.5** shows the emission spectrum of a number of common elements along with an example of a continuous spectrum.)

When we have a hot, thin gas, each particular chemical element or compound produces its own characteristic pattern of spectral lines—its spectral signature. No two types of atoms or molecules give the same patterns. In other words, each particular gas can absorb or emit only certain wavelengths of the light peculiar to that gas. In contrast, absorption spectra occur when passing white light through a cool, thin gas. The temperature and other conditions determine whether the lines are bright or dark (whether light is absorbed or emitted), but the wavelengths of the lines for any element are the same in either case. It is the precise pattern of wavelengths that makes the signature of each element unique. Liquids and solids can also generate spectral lines or bands, but they are broader and less well defined—and hence, more difficult to interpret. Spectral analysis, however, can be quite useful. It can, for example, be applied to light reflected off the surface of a nearby asteroid as well as to light from a distant galaxy.

The dark lines in the solar spectrum thus give evidence of certain chemical elements between us and the Sun absorbing those wavelengths of sunlight. Because the space between us and the Sun is pretty empty, astronomers realized that the atoms doing the absorbing must be in a thin atmosphere of cooler gas around the Sun. This outer atmosphere is not all that different from the rest of the Sun, just thinner and cooler. Thus, we can use what we learn about its composition as an indicator of what the whole Sun is made of. Similarly, we can use the presence of absorption and emission lines to analyze the composition of other stars and clouds of gas in space.

Such analysis of spectra is the key to modern astronomy. Only in this way can we “sample” the stars, which are too far away for us to visit. Encoded in the electromagnetic radiation from celestial objects is clear information about the chemical makeup of these objects. Only by understanding what the stars were made of could astronomers begin to form theories about what made them shine and how they evolved.

In 1860, German physicist Gustav Kirchhoff became the first person to use spectroscopy to identify an element in the Sun when he found the spectral signature of sodium gas. In the years that followed, astronomers found many other chemical elements in the Sun and stars. In fact, the element helium was found first in the Sun from its spectrum and only later identified on Earth. (The word “helium” comes from *helios*, the Greek name for the Sun.)

Why are there specific lines for each element? The answer to that question was not found until the twentieth century; it required the development of a model for the atom. We therefore turn next to a closer examination of the atoms that make up all matter.

The Rainbow

Rainbows are an excellent illustration of the dispersion of sunlight. You have a good chance of seeing a rainbow any time you are between the Sun and a rain shower, as illustrated in **Figure 14.6**. The raindrops act like little prisms and break white light into the spectrum of colors. Suppose a ray of sunlight encounters a raindrop and passes into it. The light changes direction—is refracted—when it passes from air to water; the blue and violet light are refracted more than the red. Some of the light is then reflected at the backside of the drop and reemerges from the front, where it is again refracted. As a result, the white light is spread out into a rainbow of colors.

Figure 14.6 (a) This diagram shows how light from the Sun, which is located behind the observer, can be refracted by raindrops to produce (b) a rainbow. (c) Refraction separates white light into its component colors.

Note that violet light lies above the red light after it emerges from the raindrop. When you look at a rainbow, however, the red light is higher in the sky. Why? Look again at **Figure 14.6**. If the observer looks at a raindrop that is high in the sky, the violet light passes over her head and the red light enters her eye. Similarly, if the observer looks at a raindrop that is low in the sky, the violet light reaches her eye and the drop appears violet, whereas the red light from that same drop strikes the ground and is not seen. Colors of intermediate wavelengths are refracted to the eye by drops that are intermediate in altitude between the drops that appear violet and the ones that appear red. Thus, a single rainbow always has red on the outside and violet on the inside.

14.2 | The Doppler Effect

Learning Objectives

By the end of this section you will be able to:

- Explain why the spectral lines of photons we observe from an object will change as a result of the object's motion toward or away from us
- Describe how we can use the Doppler effect to deduce how fast astronomical objects are moving through space
- Calculate the Doppler shift of a particular wavelength given the radial velocity of the source.

The preceding section introduced you to many new concepts, and we hope that through those, you have seen one major idea emerge. Astronomers can learn about the elements in stars and galaxies by decoding the information in their spectral lines. There is a complicating factor in learning how to decode the message of starlight, however. If a star is moving toward or away from us, its lines will be in a slightly different place in the spectrum from where they would be in a star at rest. And most objects in the universe do have some motion relative to the Sun.

Motion Affects Waves

In 1842, Christian Doppler first measured the effect of motion on waves by hiring a group of musicians to play on an open railroad car as it was moving along the track. He then applied what he learned to all waves, including light, and pointed out that if a light source is approaching or receding from the observer, the light waves will be, respectively, crowded more closely together or spread out. The general principle, now known as the **Doppler effect**, is illustrated in Figure 14.7.

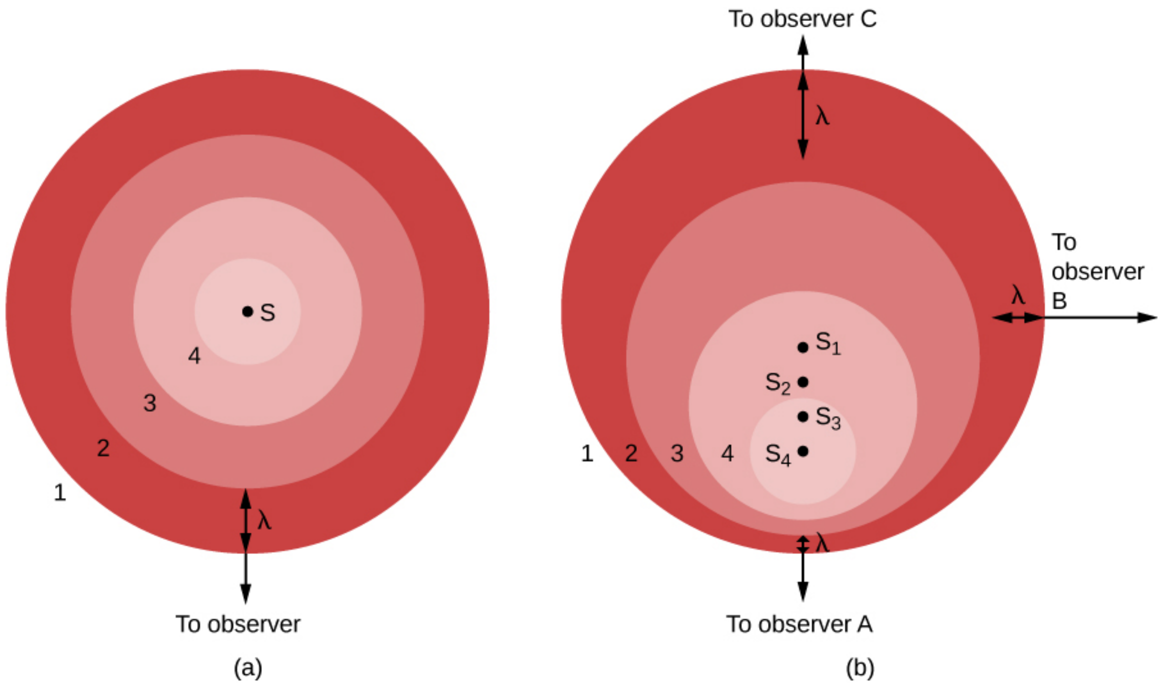


Figure 14.7 (a) A source, S, makes waves whose numbered crests (1, 2, 3, and 4) wash over a stationary observer. (b) The source S now moves toward observer *A* and away from observer *C*. Wave crest 1 was emitted when the source was at position S1, crest 2 at position S2, and so forth. Observer *A* sees waves compressed by this motion and sees a blueshift (if the waves are light). Observer *C* sees the waves stretched out by the motion and sees a redshift. Observer *B*, whose line of sight is perpendicular to the source's motion, sees no change in the waves (and feels left out).

In part (a) of the figure, the light source (S) is at rest with respect to the observer. The source gives off a series of waves, whose crests we have labeled 1, 2, 3, and 4. The light waves spread out evenly in all directions, like the ripples from a splash in a pond. The crests are separated by a distance, λ, where λ is the wavelength. The observer, who happens to be located in the direction of the bottom of the image, sees the light waves coming nice and evenly, one wavelength apart. Observers located anywhere else would see the same thing.

On the other hand, if the source of light is moving with respect to the observer, as seen in part (b), the situation is more complicated. Between the time one crest is emitted and the next one is ready to come out, the source has moved a bit, toward the bottom of the page. From the point of view of observer *A*, this motion of the source has decreased the distance between crests—it's squeezing the crests together, this observer might say.

In part (b), we show the situation from the perspective of three observers. The source is seen in four positions, S_1, S_2, S_3, and S_4, each corresponding to the emission of one wave crest. To observer *A*, the waves seem to follow one another more closely, at a decreased wavelength and thus increased frequency. (Remember, all light waves travel at the speed of light through empty space, no matter what. This means that motion cannot affect the speed, but only the wavelength and the frequency. As the wavelength decreases, the frequency must increase. If the waves are shorter, more will be able to move by during each second.)

The situation is not the same for other observers. Let's look at the situation from the point of view of observer *C*, located opposite observer *A* in the figure. For her, the source is moving away from her location. As a result, the waves are not squeezed together but instead are spread out by the motion of the source. The crests arrive with an increased wavelength and decreased frequency. To observer *B*, in a direction at right angles to the motion of the source, no effect is observed. The wavelength and frequency remain the same as they were in part (a) of the figure.

We can see from this illustration that the Doppler effect is produced only by a motion toward or away from the observer, a motion called **radial velocity**. Sideways motion does not produce such an effect. Observers between *A* and *B* would observe some shortening of the light waves for that part of the motion of the source that is along their line of sight. Observers between *B* and *C* would observe lengthening of the light waves that are along their line of sight.

You may have heard the Doppler effect with sound waves. When a train whistle or police siren approaches you and then moves away, you will notice a decrease in the pitch (which is how human senses interpret sound wave frequency) of the sound waves. Compared to the waves at rest, they have changed from slightly more frequent when coming toward you, to slightly less frequent when moving away from you.

A nice example of this change in the sound of a train whistle can be heard at the end of the classic Beach Boys song "Caroline, No" on their album *Pet Sounds*. To hear this sound, go to this **YouTube (https://openstax.org/l/30BBtrain)** version of the song. The sound of the train begins at approximately 2:20.

Color Shifts

When the source of waves moves toward you, the wavelength decreases a bit. If the waves involved are visible light, then the colors of the light change slightly. As wavelength decreases, they shift toward the blue end of the spectrum: astronomers call this a *blueshift* (since the end of the spectrum is really violet, the term should probably be *violetshift*, but blue is a more common color). When the source moves away from you and the wavelength gets longer, we call the change in colors a *redshift*. Because the Doppler effect was first used with visible light in astronomy, the terms " blueshift" and " redshift" became well established. Today, astronomers use these words to describe changes in the wavelengths of radio waves or X-rays as comfortably as they use them to describe changes in visible light.

The greater the motion toward or away from us, the greater the Doppler shift. If the relative motion is entirely along the line of sight, the formula for the Doppler shift of light is

$$\frac{\Delta\lambda}{\lambda} = \frac{v}{c}$$

where λ is the wavelength emitted by the source, Δλ is the difference between λ and the wavelength measured by the observer, c is the speed of light, and v is the relative speed of the observer and the source in the line of sight. The variable v is counted as positive if the velocity is one of recession, and negative if it is one of approach. Solving this equation for the velocity, we find $v = c \times \Delta\lambda/\lambda$.

If a star approaches or recedes from us, the wavelengths of light in its continuous spectrum appear shortened or lengthened, respectively, as do those of the dark lines. However, unless its speed is tens of thousands of kilometers per second, the star does not appear noticeably bluer or redder than normal. The Doppler shift is thus not easily detected in a continuous spectrum and cannot be measured accurately in such a spectrum. The wavelengths of the absorption lines can be measured accurately, however, and their Doppler shift is relatively simple to detect.

Example 14.1

The Doppler Effect

We can use the Doppler effect equation to calculate the radial velocity of an object if we know three things: the speed of light, the original (unshifted) wavelength of the light emitted, and the difference between the wavelength of the emitted light and the wavelength we observe. For particular absorption or emission lines, we usually know exactly what wavelength the line has in our laboratories on Earth, where the source of light is not moving. We can measure the new wavelength with our instruments at the telescope, and so we know the difference in wavelength due to Doppler shifting. Since the speed of light is a universal constant, we can then calculate the radial velocity of the star.

A particular emission line of hydrogen is originally emitted with a wavelength of 656.3 nm from a gas cloud. At our telescope, we observe the wavelength of the emission line to be 656.6 nm. How fast is this gas cloud moving toward or away from Earth?

Solution

Because the light is shifted to a longer wavelength (redshifted), we know this gas cloud is moving away from us. The speed can be calculated using the Doppler shift formula:

$$\begin{aligned} v &= c \times \frac{\Delta\lambda}{\lambda} = \left(3.0 \times 10^8 \text{ m/s}\right)\left(\frac{0.3\text{ nm}}{656.3\text{ nm}}\right) = \left(3.0 \times 10^8 \text{ m/s}\right)\left(\frac{0.3 \times 10^{-9}\text{ m}}{656.3 \times 10^{-9}\text{ m}}\right) \\ &= 140{,}000\text{ m/s} = 140\text{ km/s} \end{aligned}$$

14.1 Suppose a spectral line of hydrogen, normally at 500 nm, is observed in the spectrum of a star to be at 500.1 nm. How fast is the star moving toward or away from Earth?

You may now be asking: if all the stars are moving and motion changes the wavelength of each spectral line, won’t this be a disaster for astronomers trying to figure out what elements are present in the stars? After all, it is the precise wavelength (or color) that tells astronomers which lines belong to which element. And we first measure these wavelengths in containers of gas in our laboratories, which are not moving. If every line in a star’s spectrum is now shifted by its motion to a different wavelength (color), how can we be sure which lines and which elements we are looking at in a star whose speed we do not know?

Take heart. This situation sounds worse than it really is. Astronomers rarely judge the presence of an element in an astronomical object by a single line. It is the *pattern* of lines unique to hydrogen or calcium that enables us to determine that those elements are part of the star or galaxy we are observing. The Doppler effect does not change the pattern of lines from a given element—it only shifts the whole pattern slightly toward redder or bluer wavelengths. The shifted pattern is still quite easy to recognize. Best of all, when we do recognize a familiar element’s pattern, we get a bonus: the amount the pattern is shifted can enable us to determine the speed of the objects in our line of sight.

The training of astronomers includes much work on learning to decode light (and other electromagnetic radiation). A skillful “decoder” can learn the temperature of a star, what elements are in it, and even its speed in a direction toward us or away from us. That’s really an impressive amount of information for stars that are light-years away.

CHAPTER 14 REVIEW

KEY TERMS

absorption spectrum a series or pattern of dark lines superimposed on a continuous spectrum

continuous spectrum a spectrum of light composed of radiation of a continuous range of wavelengths or colors, rather than only certain discrete wavelengths

dispersion separation of different wavelengths of white light through refraction of different amounts

Doppler effect the apparent change in wavelength or frequency of the radiation from a source due to its relative motion away from or toward the observer

emission spectrum a series or pattern of bright lines superimposed on a continuous spectrum

radial velocity motion toward or away from the observer; the component of relative velocity that lies in the line of sight

spectrometer an instrument for obtaining a spectrum; in astronomy, usually attached to a telescope to record the spectrum of a star, galaxy, or other astronomical object

KEY EQUATIONS

Doppler Wavelength Shift $\frac{\Delta\lambda}{\lambda} = \frac{v}{c}$

SUMMARY

14.1 Spectroscopy

- A spectrometer is a device that forms a spectrum, often utilizing the phenomenon of dispersion.
- The light from an astronomical source can consist of a continuous spectrum, an emission (bright line) spectrum, or an absorption (dark line) spectrum.
- Because each element leaves its spectral signature in the pattern of lines we observe, spectral analyses reveal the composition of the Sun and stars.

14.2 The Doppler Effect

- If a source of light is moving toward us and produces a spectral line, we see that line shifted slightly toward the blue of its normal wavelength in a spectrum.
- If the source is moving away, we see the line shifted toward the red.
- This shift is known as the Doppler effect and can be used to measure the radial velocities of distant objects.

CONCEPTUAL QUESTIONS

14.1 Spectroscopy

1. Explain what dispersion is and how astronomers use this phenomenon to study a star's light.

2. Explain why glass prisms disperse light.

3. Explain what Joseph Fraunhofer discovered about stellar spectra.

4. Explain how we use spectral absorption and emission lines to determine the composition of a gas.

14.2 The Doppler Effect

5. Where in an atom would you expect to find electrons? Protons? Neutrons?

6. Explain how emission lines and absorption lines are formed. In what sorts of cosmic objects would you expect to see each?

7. Explain how the Doppler effect works for sound waves and give some familiar examples.

8. What kind of motion for a star does not produce a Doppler effect? Explain.

9. Describe how Bohr's model used the work of Maxwell.

10. Explain why light is referred to as electromagnetic radiation.

11. Explain the difference between radiation as it is used in most everyday language and radiation as it is used in an astronomical context.

12. What are the differences between light waves and sound waves?

13. Which type of wave has a longer wavelength: AM radio waves (with frequencies in the kilohertz range) or FM radio waves (with frequencies in the megahertz range)? Explain.

14. Explain why astronomers long ago believed that space must be filled with some kind of substance (the "aether") instead of the vacuum we know it is today.

15. Explain what the ionosphere is and how it interacts with some radio waves.

16. Which is more dangerous to living things, gamma rays or X-rays? Explain.

17. Explain why we have to observe stars and other astronomical objects from above Earth's atmosphere in order to fully learn about their properties.

18. Explain why hotter objects tend to radiate more energetic photons compared to cooler objects.

19. Explain the results of Rutherford's gold foil experiment and how they changed our model of the atom.

20. Is it possible for two different atoms of carbon to have different numbers of neutrons in their nuclei? Explain.

21. What are the three isotopes of hydrogen, and how do they differ?

22. Explain how electrons use light energy to move among energy levels within an atom.

23. Explain why astronomers use the term "blueshifted" for objects moving toward us and "redshifted" for objects moving away from us.

24. If spectral line wavelengths are changing for objects based on the radial velocities of those objects, how can we deduce which type of atom is responsible for a particular absorption or emission line?

25. Suppose you are standing at the exact center of a park surrounded by a circular road. An ambulance drives completely around this road, with siren blaring. How does the pitch of the siren change as it circles around you?

26. How could you measure Earth's orbital speed by photographing the spectrum of a star at various times throughout the year? (Hint: Suppose the star lies in the plane of Earth's orbit.)

PROBLEMS

14.2 The Doppler Effect

27. Suppose you have a spectrometer that can measure wavelengths to a precision of 0.3%. What is the minimum speed at which a light source must travel toward you for you to be able to determine that its wavelength is Doppler shifted? That is, what speed produces a shift of 0.300% ?

28. The spectral line in hydrogen that, in the laboratory, has a wavelength of 656.3 nm, is detected in the spectrum of a distant object at a wavelength of 657.0 nm. How fast is this object receding from Earth?

29. Light from distant nebulae contains radio-frequency radiation that originates in hydrogen atoms and has a wavelength of 21 cm. Suppose such a nebula is approaching Earth at a speed of 15,000 km/s. At what wavelength would our radio telescopes measure this radiation?

15 | THE ORIGINS OF LIGHT

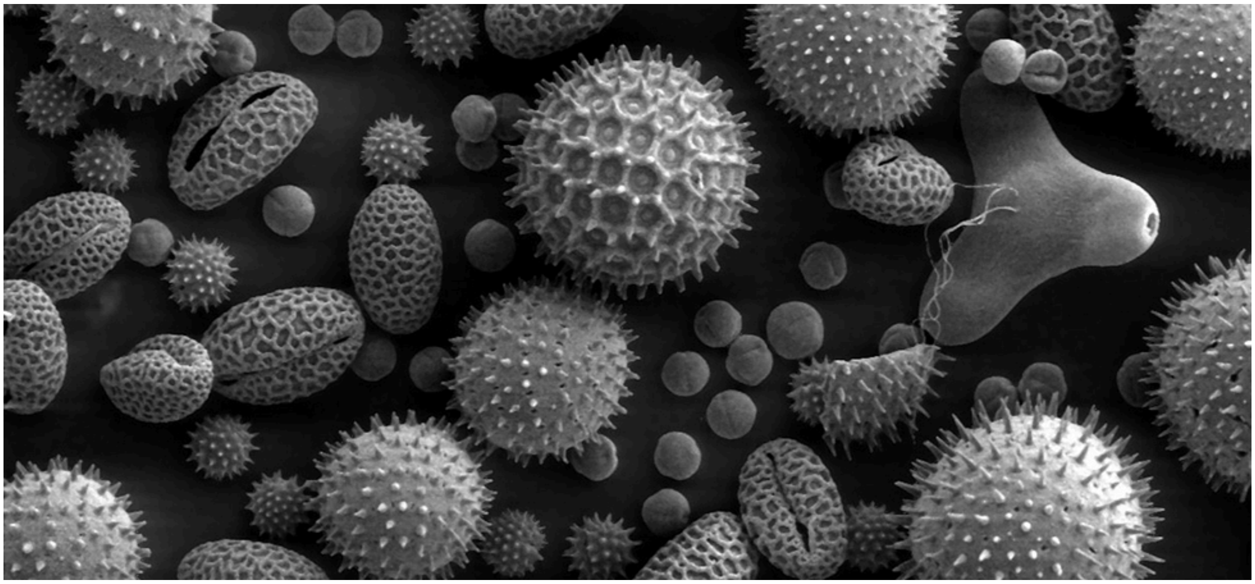


Figure 15.1 In this image of pollen taken with an electron microscope, the bean-shaped grains are about $50\mu m$ long. Electron microscopes can have a much higher resolving power than a conventional light microscope because electron wavelengths can be 100,000 times shorter than the wavelengths of visible-light photons. (credit: modification of work by Dartmouth College Electron Microscope Facility)

Chapter Outline

Introduction

Two of the most revolutionary concepts of the twentieth century were the description of light as a collection of particles, and the treatment of particles as waves. These wave properties of matter have led to the discovery of technologies such as electron microscopy, which allows us to examine submicroscopic objects such as grains of pollen, as shown above.

In this chapter, you will learn about the energy quantum, a concept that was introduced in 1900 by the German physicist Max Planck to explain blackbody radiation. But our ultimate goal is to answer a very basic question: Where does light come from? .

15.1 | Blackbody Radiation

Learning Objectives

By the end of this section you will be able to:

- Apply Wien's and Stefan's laws to analyze radiation emitted by a blackbody
- Explain Planck's hypothesis of energy quanta

All bodies emit electromagnetic radiation over a range of wavelengths. In **Section 14.1**, we learned that there are three kinds of spectra: continuous spectra, line-emission spectra, and line-absorption spectra. Now, we know that a cooler body radiates less energy than a warmer body. We also know by observation that when a body is heated and its temperature rises, the perceived wavelength of its emitted radiation changes from infrared to red, and then from red to orange, and so forth. As its temperature rises, the body glows with the colors corresponding to ever-smaller wavelengths of the electromagnetic spectrum. This is the underlying principle of the incandescent light bulb: A hot metal filament glows red, and when heating continues, its glow eventually covers the entire visible portion of the electromagnetic spectrum. The temperature (T) of the object that emits radiation, or the **emitter**, determines the wavelength at which the radiated energy is at its maximum. For example, the Sun, whose surface temperature is in the range between 5000 K and 6000 K, radiates most strongly in a range of wavelengths about 560 nm in the visible part of the electromagnetic spectrum. Your body, when at its normal temperature of about 300 K, radiates most strongly in the infrared part of the spectrum.

Our goal in this section is to quantify what we know about the radiation of continuous spectra. Despite the fact that such spectra had been studied extensively in the 19th century, a consistent theory that explained them was not found until, in 1900, Max Planck put forth his quantum hypothesis.

Radiation that is incident on an object is partially absorbed and partially reflected. At thermodynamic equilibrium, the rate at which an object absorbs radiation is the same as the rate at which it emits it. Therefore, a good **absorber** of radiation (any object that absorbs radiation) is also a good emitter. A perfect absorber absorbs all electromagnetic radiation incident on it; such an object is called a **blackbody**.

Although the blackbody is an idealization, because no physical object absorbs 100% of incident radiation, we can construct a close realization of a blackbody in the form of a small hole in the wall of a sealed enclosure known as a cavity radiator, as shown in **Figure 15.2**. The inside walls of a cavity radiator are rough and blackened so that any radiation that enters through a tiny hole in the cavity wall becomes trapped inside the cavity. At thermodynamic equilibrium (at temperature T), the cavity walls absorb exactly as much radiation as they emit. Furthermore, inside the cavity, the radiation entering the hole is balanced by the radiation leaving it. The emission spectrum of a blackbody can be obtained by analyzing the light radiating from the hole. Electromagnetic waves emitted by a blackbody are called **blackbody radiation**.

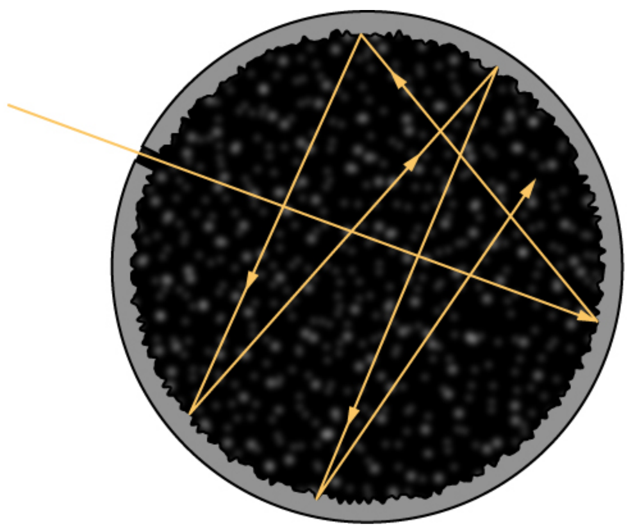


Figure 15.2 A blackbody is physically realized by a small hole in the wall of a cavity radiator.

The intensity $I(\lambda, T)$ of blackbody radiation depends on the wavelength λ of the emitted radiation and on the temperature T of the blackbody (**Figure 15.3**). The function $I(\lambda, T)$ is the **power intensity** that is radiated per unit wavelength; in other words, it is the power radiated per unit area of the hole in a cavity radiator per unit wavelength. According to this definition, $I(\lambda, T)d\lambda$ is the power per unit area that is emitted in the wavelength interval from λ to $\lambda + d\lambda$. The intensity distribution among wavelengths of radiation emitted by cavities was studied experimentally at the end of the nineteenth century. Generally, radiation emitted by materials only approximately follows the blackbody radiation curve (**Figure 15.4**); however, spectra of common stars do follow the blackbody radiation curve very closely.

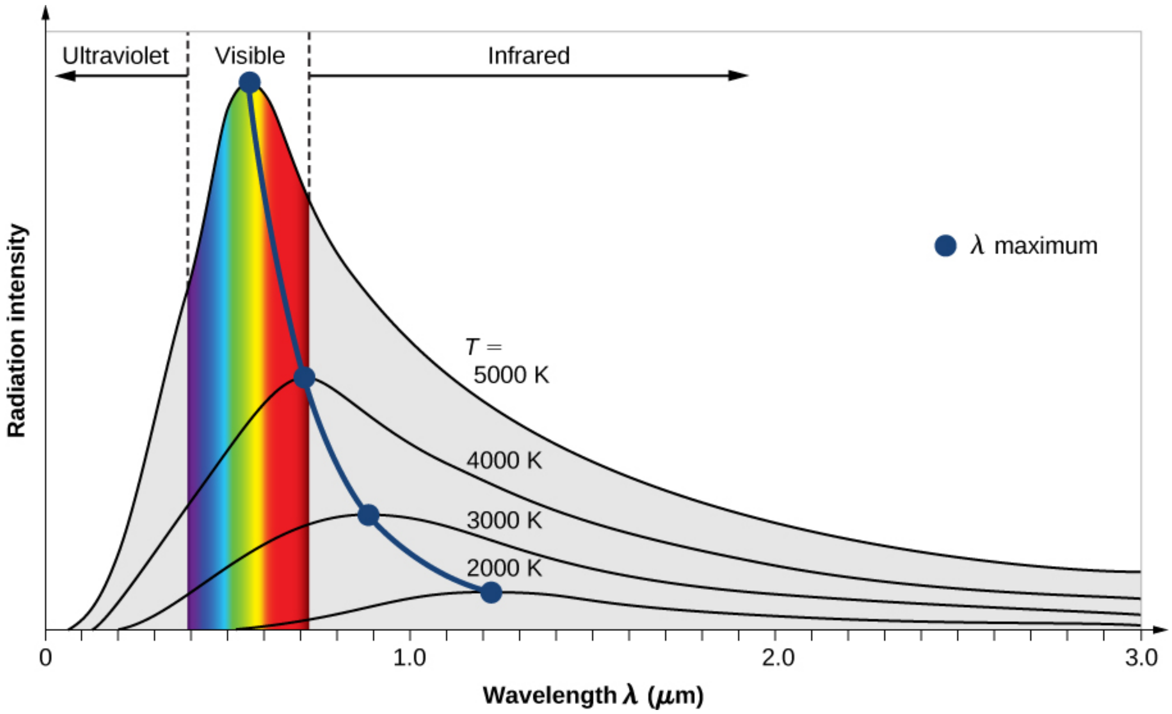


Figure 15.3 The intensity of blackbody radiation versus the wavelength of the emitted radiation. Each curve corresponds to a different blackbody temperature, starting with a low temperature (the lowest curve) to a high temperature (the highest curve).

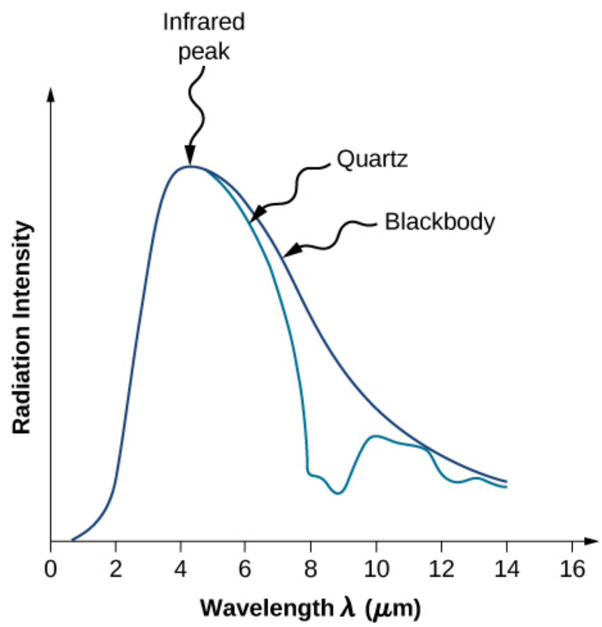


Figure 15.4 The spectrum of radiation emitted from a quartz surface (blue curve) and the blackbody radiation curve (black curve) at 600 K.

Two important laws, which we have already encountered in our discussion of **The Electromagnetic Spectrum**, summarize the experimental findings of blackbody radiation: *Wien's displacement law* and *Stefan's law*. Wien's displacement law is illustrated in **Figure 15.3** by the curve connecting the maxima on the intensity curves. In these curves, we see that the hotter the body, the shorter the wavelength corresponding to the emission peak in the radiation curve. Quantitatively, to four significant figures, Wien's law reads

$$\lambda_{\text{max}} T = 2.898 \times 10^{-3}\,\text{m} \cdot \text{K} \tag{15.1}$$

where λ_{max} is the position of the maximum in the radiation curve. In other words, λ_{max} is the wavelength at which a blackbody radiates most strongly at a given temperature T. Note that in **Equation 15.1**, the temperature is in kelvins. Wien's displacement law allows us to estimate the temperatures of distant stars by measuring the wavelength of radiation they emit. (Note that this differs from **Equation 13.1** only in the units used for the wavelength.)

Of course, Wien's Law is easy to derive using calculus, Without detailing the calculation here, it it simply a relationship that identifies the maximum point in the intensity distribution $I(\lambda, T)$. All that is needed is to take the derivative of $I(\lambda, T)$ with respect to T and to then set that derivative equal to zero.

Example 15.1

Temperatures of Distant Stars

On a clear evening during the winter months, if you happen to be in the Northern Hemisphere and look up at the sky, you can see the constellation Orion (The Hunter). One star in this constellation, Rigel, flickers in a blue color and another star, Betelgeuse, has a reddish color, as shown in **Figure 15.5**. Which of these two stars is cooler, Betelgeuse or Rigel?

Strategy

We treat each star as a blackbody. Then according to Wien's law, its temperature is inversely proportional to the wavelength of its peak intensity. The wavelength $\lambda_{\text{max}}^{(\text{blue})}$ of blue light is shorter than the wavelength $\lambda_{\text{max}}^{(\text{red})}$ of red light. Even if we do not know the precise wavelengths, we can still set up a proportion.

Solution

Writing Wien's law for the blue star and for the red star, we have

$$\lambda_{\text{max}}^{(\text{red})} T_{(\text{red})} = 2.898 \times 10^{-3}\,\text{m} \cdot \text{K} = \lambda_{\text{max}}^{(\text{blue})} T_{(\text{blue})} \tag{15.2}$$

When simplified, **Equation 15.2** gives

$$T_{(\text{red})} = \frac{\lambda_{\text{max}}^{(\text{blue})}}{\lambda_{\text{max}}^{(\text{red})}} T_{(\text{blue})} < T_{(\text{blue})} \tag{15.3}$$

Therefore, Betelgeuse is cooler than Rigel.

Significance

Note that Wien's displacement law tells us that the higher the temperature of an emitting body, the shorter the wavelength of the radiation it emits. The qualitative analysis presented in this example is generally valid for any emitting body, whether it is a big object such as a star or a small object such as the glowing filament in an incandescent lightbulb.

15.1 Check Your Understanding The flame of a peach-scented candle has a yellowish color and the flame of a Bunsen's burner in a chemistry lab has a bluish color. Which flame has a higher temperature?

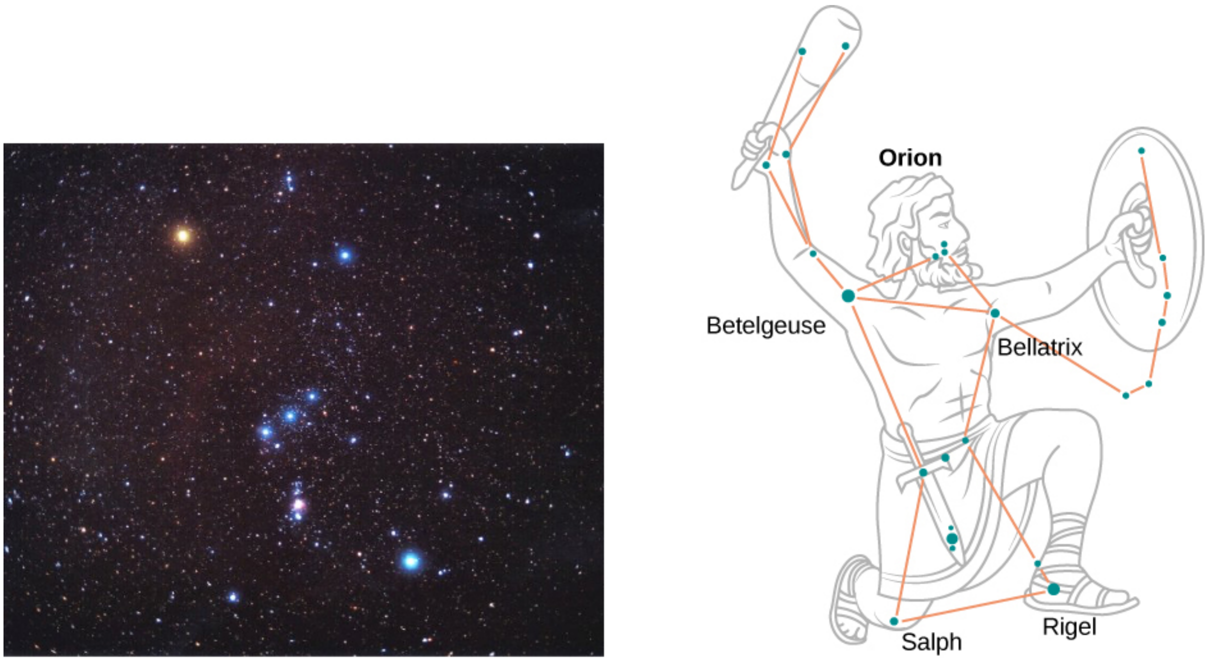


Figure 15.5 In the Orion constellation, the red star Betelgeuse, which usually takes on a yellowish tint, appears as the figure's right shoulder (in the upper left). The giant blue star on the bottom right is Rigel, which appears as the hunter's left foot. (credit left: modification of work by Matthew Spinelli, NASA APOD)

The second experimental relation is Stefan's law**,** which concerns the total power of blackbody radiation emitted across the entire spectrum of wavelengths at a given temperature. In **Figure 15.3**, this total power is represented by the area under the blackbody radiation curve for a given *T*. As the temperature of a blackbody increases, the total emitted power also increases. Quantitatively, Stefan's law expresses this relation as

$$P(T) = \sigma A T^4 \tag{15.4}$$

where A is the surface area of a blackbody, T is its temperature (in kelvins), and σ is the **Stefan–Boltzmann constant**, and to four significant figures $\sigma = 5.670 \times 10^{-8}\,\text{W/(m}^2 \cdot \text{K}^4)$. Stefan's law enables us to estimate how much energy a star is radiating by remotely measuring its temperature. (Note that this equation is equivalent to **Equation 13.2**.)

Example 15.2

Power Radiated by Stars

A star such as our Sun will eventually evolve to a "red giant" star and then to a "white dwarf" star. A typical white dwarf is approximately the size of Earth, and its surface temperature is about $2.5 \times 10^4\,\text{K}$. A typical red giant has a surface temperature of $3.0 \times 10^3\,\text{K}$ and a radius ~100,000 times larger than that of a white dwarf. What is the average radiated power per unit area and the total power radiated by each of these types of stars? How do they compare?

Strategy

If we treat the star as a blackbody, then according to Stefan's law, the total power that the star radiates is proportional to the fourth power of its temperature. To find the power radiated per unit area of the surface, we do not need to make any assumptions about the shape of the star because *P*/*A* depends only on temperature. However, to compute the total power, we need to make an assumption that the energy radiates through a spherical surface enclosing the star, so that the surface area is $A = 4\pi R^2$, where *R* is its radius.

Solution

A simple proportion based on Stefan's law gives

$$\frac{P_{\text{dwarf}}/A_{\text{dwarf}}}{P_{\text{giant}}/A_{\text{giant}}} = \frac{\sigma T^4_{\text{dwarf}}}{\sigma T^4_{\text{giant}}} = \left(\frac{T_{\text{dwarf}}}{T_{\text{giant}}}\right)^4 = \left(\frac{2.5\times 10^4}{3.0\times 10^3}\right)^4 = 4820 \quad (15.5)$$

The power emitted per unit area by a white dwarf is about 5000 times that the power emitted by a red giant. Denoting this ratio by $a = 4.8\times 10^3$, **Equation 15.5** gives

$$\frac{P_{\text{dwarf}}}{P_{\text{giant}}} = a\frac{A_{\text{dwarf}}}{A_{\text{giant}}} = a\frac{4\pi R^2_{\text{dwarf}}}{4\pi R^2_{\text{giant}}} = a\left(\frac{R_{\text{dwarf}}}{R_{\text{giant}}}\right)^2 = 4.8\times 10^3\left(\frac{R_{\text{dwarf}}}{10^5 R_{\text{dwarf}}}\right)^2 = 4.8\times 10^{-7} \quad (15.6)$$

We see that the total power emitted by a white dwarf is a tiny fraction of the total power emitted by a red giant. Despite its relatively lower temperature, the overall power radiated by a red giant far exceeds that of the white dwarf because the red giant has a much larger surface area. To estimate the absolute value of the emitted power per unit area, we again use Stefan's law. For the white dwarf, we obtain

$$\frac{P_{\text{dwarf}}}{A_{\text{dwarf}}} = \sigma T^4_{\text{dwarf}} = 5.670\times 10^{-8}\frac{\text{W}}{\text{m}^2\cdot\text{K}^4}\left(2.5\times 10^4\,\text{K}\right)^4 = 2.2\times 10^{10}\,\text{W/m}^2 \quad (15.7)$$

The analogous result for the red giant is obtained by scaling the result for a white dwarf:

$$\frac{P_{\text{giant}}}{A_{\text{giant}}} = \frac{2.2\times 10^{10}}{4.82\times 10^3}\frac{\text{W}}{\text{m}^2} = 4.56\times 10^6\frac{\text{W}}{\text{m}^2} \cong 4.6\times 10^6\frac{\text{W}}{\text{m}^2} \quad (15.8)$$

Significance

To estimate the total power emitted by a white dwarf, in principle, we could use **Equation 15.7**. However, to find its surface area, we need to know the average radius, which is not given in this example. Therefore, the solution stops here. The same is also true for the red giant star.

15.2 Check Your Understanding An iron poker is being heated. As its temperature rises, the poker begins to glow—first dull red, then bright red, then orange, and then yellow. Use either the blackbody radiation curve or Wien's law to explain these changes in the color of the glow.

15.3 Check Your Understanding Suppose that two stars, α and β, radiate exactly the same total power. If the radius of star α is three times that of star β, what is the ratio of the surface temperatures of these stars? Which one is hotter?

The term "blackbody" was coined by Gustav R. Kirchhoff in 1862. The blackbody radiation curve was known experimentally, but its shape eluded physical explanation until the year 1900. The physical model of a blackbody at temperature T is that of the electromagnetic waves enclosed in a cavity (see **Figure 15.2**) and at thermodynamic equilibrium with the cavity walls. The waves can exchange energy with the walls. The objective here is to find the energy density distribution among various modes of vibration at various wavelengths (or frequencies). In other words, we want to know how much energy is carried by a single wavelength or a band of wavelengths. Once we know the energy distribution, we can use standard statistical methods (similar to those studied in a previous chapter) to obtain the blackbody radiation curve, Stefan's law, and Wien's displacement law. When the physical model is correct, the theoretical predictions should be the same as the experimental curves.

In a classical approach to the blackbody radiation problem, in which radiation is treated as waves (as you have studied in previous chapters), the modes of electromagnetic waves trapped in the cavity are in equilibrium and continually exchange their energies with the cavity walls. There is no physical reason why a wave should do otherwise: Any amount of energy can be exchanged, either by being transferred from the wave to the material in the wall or by being received by the wave from the material in the wall. This classical picture is the basis of the model developed by Lord Rayleigh and, independently, by Sir James Jeans. The result of this classical model for blackbody radiation curves is known as the *Rayleigh–Jeans law*.

However, as shown in **Figure 15.6**, the Rayleigh–Jeans law fails to correctly reproduce experimental results. In the limit of short wavelengths, the Rayleigh–Jeans law predicts infinite radiation intensity, which is inconsistent with the experimental results in which radiation intensity has finite values in the ultraviolet region of the spectrum. This divergence between the results of classical theory and experiments, which came to be called the *ultraviolet catastrophe*, shows how classical physics fails to explain the mechanism of blackbody radiation.

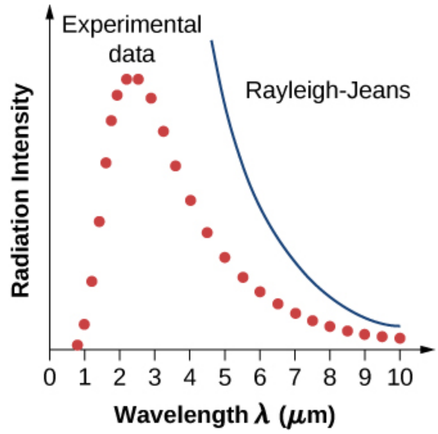


Figure 15.6 The ultraviolet catastrophe: The Rayleigh–Jeans law does not explain the observed blackbody emission spectrum.

The blackbody radiation problem was solved in 1900 by Max Planck. Planck used the same idea as the Rayleigh–Jeans model in the sense that he treated the electromagnetic waves between the walls inside the cavity classically, and assumed that the radiation is in equilibrium with the cavity walls. The innovative idea that Planck introduced in his model is the assumption that the cavity radiation originates from atomic oscillations inside the cavity walls, and that these oscillations can have only *discrete* values of energy. Therefore, the radiation trapped inside the cavity walls can exchange energy with the walls only in discrete amounts. Planck's hypothesis of discrete energy values, which he called *quanta*, assumes that the oscillators inside the cavity walls have **quantized energies**. This was a brand new idea that went beyond the classical physics of the nineteenth century because, as you learned in a previous chapter, in the classical picture, the energy of an oscillator can take on any continuous value. Planck assumed that the energy of an oscillator (E_n) can have only discrete, or quantized, values:

$$E_n = nhf, \quad \text{where } n = 1, 2, 3, ... \tag{15.9}$$

In **Equation 15.9**, *f* is the frequency of Planck's oscillator. The natural number *n* that enumerates these discrete energies is called a **quantum number**. The physical constant *h* is called *Planck's constant*:

$$h = 6.626 \times 10^{-34}\,\text{J} \cdot \text{s} = 4.136 \times 10^{-15}\,\text{eV} \cdot \text{s} \tag{15.10}$$

Each discrete energy value corresponds to a **quantum state of a Planck oscillator**. Quantum states are enumerated by quantum numbers. For example, when Planck's oscillator is in its first $n = 1$ quantum state, its energy is $E_1 = hf$; when it is in the $n = 2$ quantum state, its energy is $E_2 = 2hf$; when it is in the $n = 3$ quantum state, $E_3 = 3hf$; and so on.

Note that **Equation 15.9** shows that there are infinitely many quantum states, which can be represented as a sequence {*hf*, 2*hf*, 3*hf*,…, (*n* – 1)*hf*, *nhf*, (*n* + 1)*hf*,…}. Each two consecutive quantum states in this sequence are separated by an energy jump, $\Delta E = hf$. An oscillator in the wall can receive energy from the radiation in the cavity (absorption), or it can give away energy to the radiation in the cavity (emission). The absorption process sends the oscillator to a higher quantum state, and the emission process sends the oscillator to a lower quantum state. Whichever way this exchange of energy goes, the smallest amount of energy that can be exchanged is *hf*. There is no upper limit to how much energy can be exchanged, but whatever is exchanged must be an integer multiple of *hf*. If the energy packet does not have this exact amount, it is neither absorbed nor emitted at the wall of the blackbody.

Planck's Quantum Hypothesis

Planck's hypothesis of energy quanta states that the amount of energy emitted by the oscillator is carried by the quantum of radiation, ΔE :

$$\Delta E = hf$$

Recall that the frequency of electromagnetic radiation is related to its wavelength and to the speed of light by the fundamental relation $f\lambda = c$. This means that we can express **Equation 15.10** equivalently in terms of wavelength λ. When included in the computation of the energy density of a blackbody, Planck's hypothesis gives the following theoretical expression for the power intensity of emitted radiation per unit wavelength:

$$I(\lambda, T) = \frac{2\pi hc^2}{\lambda^5} \frac{1}{e^{hc/\lambda k_B T} - 1} \tag{15.11}$$

where c is the speed of light in vacuum and k_B is Boltzmann's constant, $k_B = 1.380 \times 10^{-23}$ J/K. The theoretical formula expressed in **Equation 15.11** is called *Planck's blackbody radiation law*. This law is in agreement with the experimental blackbody radiation curve (see **Figure 15.7**). In addition, Wien's displacement law and Stefan's law can both be derived from **Equation 15.11**. To derive Wien's displacement law, we use differential calculus to find the maximum of the radiation intensity curve $I(\lambda, T)$. To derive Stefan's law and find the value of the Stefan–Boltzmann constant, we use integral calculus and integrate $I(\lambda, T)$ to find the total power radiated by a blackbody at one temperature in the entire spectrum of wavelengths from $\lambda = 0$ to $\lambda = \infty$. This derivation is left as an exception later in this chapter.

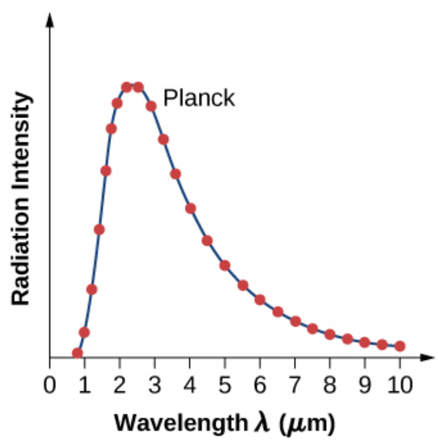


Figure 15.7 Planck's theoretical result (continuous curve) and the experimental blackbody radiation curve (dots).

Example 15.3

Planck's Quantum Oscillator

A quantum oscillator in the cavity wall in **Figure 15.2** is vibrating at a frequency of 5.0×10^{14} Hz. Calculate the spacing between its energy levels.

Strategy

Energy states of a quantum oscillator are given by **Equation 15.9**. The energy spacing ΔE is obtained by finding the energy difference between two adjacent quantum states for quantum numbers $n + 1$ and n.

Solution

We can substitute the given frequency and Planck's constant directly into the equation:

$$\Delta E = E_{n+1} - E_n = (n+1)hf - nhf = hf = (6.626 \times 10^{-34}\,\text{J} \cdot \text{s})(5.0 \times 10^{14}\,\text{Hz}) = 3.3 \times 10^{-19}\,\text{J}$$

Significance

Note that we do not specify what kind of material was used to build the cavity. Here, a quantum oscillator is a theoretical model of an atom or molecule of material in the wall.

15.4 Check Your Understanding A molecule is vibrating at a frequency of $5.0 \times 10^{14}\,\text{Hz}$. What is the smallest spacing between its vibrational energy levels?

Example 15.4

Quantum Theory Applied to a Classical Oscillator

A 1.0-kg mass oscillates at the end of a spring with a spring constant of 1000 N/m. The amplitude of these oscillations is 0.10 m. Use the concept of quantization to find the energy spacing for this classical oscillator. Is the energy quantization significant for macroscopic systems, such as this oscillator?

Strategy

We use **Equation 15.10** as though the system were a quantum oscillator, but with the frequency *f* of the mass vibrating on a spring. To evaluate whether or not quantization has a significant effect, we compare the quantum energy spacing with the macroscopic total energy of this classical oscillator.

Solution

For the spring constant, $k = 1.0 \times 10^3\,\text{N/m},$ the frequency *f* of the mass, $m = 1.0\text{kg},$ is

$$f = \frac{1}{2\pi}\sqrt{\frac{k}{m}} = \frac{1}{2\pi}\sqrt{\frac{1.0 \times 10^3\,\text{N/m}}{1.0\text{kg}}} \simeq 5.0\,\text{Hz}$$

The energy quantum that corresponds to this frequency is

$$\Delta E = hf = (6.626 \times 10^{-34}\,\text{J} \cdot \text{s})(5.0\text{Hz}) = 3.3 \times 10^{-33}\,\text{J}$$

When vibrations have amplitude $A = 0.10\text{m},$ the energy of oscillations is

$$E = \frac{1}{2}kA^2 = \frac{1}{2}(1000\text{N/m})(0.1\text{m})^2 = 5.0\text{J}$$

Significance

Thus, for a classical oscillator, we have $\Delta E/E \approx 10^{-34}$. We see that the separation of the energy levels is immeasurably small. Therefore, for all practical purposes, the energy of a classical oscillator takes on continuous values. This is why classical principles may be applied to macroscopic systems encountered in everyday life without loss of accuracy.

15.5 Check Your Understanding Would the result in **Example 15.4** be different if the mass were not 1.0 kg but a tiny mass of 1.0 μg, and the amplitude of vibrations were 0.10 μm?

When Planck first published his result, the hypothesis of energy quanta was not taken seriously by the physics community because it did not follow from any established physics theory at that time. It was perceived, even by Planck himself, as a useful mathematical trick that led to a good theoretical "fit" to the experimental curve. This perception was changed in 1905 when Einstein published his explanation of the photoelectric effect, in which he gave Planck's energy quantum a new meaning: that of a particle of light.

15.2 | Bohr's Model of the Hydrogen Atom

Learning Objectives

By the end of this section, you will be able to:

- Explain the difference between the absorption spectrum and the emission spectrum of radiation emitted by atoms
- Describe the Rutherford gold foil experiment and the discovery of the atomic nucleus
- Explain the atomic structure of hydrogen
- Describe the postulates of the early quantum theory for the hydrogen atom
- Summarize how Bohr's quantum model of the hydrogen atom explains the radiation spectrum of atomic hydrogen

In **Section 15.1** we examined the theory behind the emission of continuous spectra by heated objects (like stars). In this section, we examine the theory behind line spectra (either line-emission or line-absorption spectra). The successful theory, originally put forth by Niels Bohr, also relies on Planck's quantum hypothesis.

Historically, Bohr's model of the hydrogen atom is the very first model of atomic structure that correctly explained the radiation spectra of atomic hydrogen. The model has a special place in the history of physics because it introduced an early quantum theory, which brought about new developments in scientific thought and later culminated in the development of quantum mechanics. To understand the specifics of Bohr's model, we must first review the nineteenth-century discoveries that prompted its formulation.

When we use a prism to analyze white light coming from the sun, several dark lines in the solar spectrum are observed (**Figure 15.8**). Solar absorption lines are called **Fraunhofer lines** after Joseph von Fraunhofer, who accurately measured their wavelengths. During 1854–1861, Gustav Kirchhoff and Robert Bunsen discovered that for the various chemical elements, the line **emission spectrum** of an element exactly matches its line **absorption spectrum**. The difference between the absorption spectrum and the emission spectrum is explained in **Figure 15.9**. An absorption spectrum is observed when light passes through a gas. This spectrum appears as black lines that occur only at certain wavelengths on the background of the continuous spectrum of white light (**Figure 15.8**). The missing wavelengths tell us which wavelengths of the radiation are absorbed by the gas. The emission spectrum is observed when light is emitted by a gas. This spectrum is seen as colorful lines on the black background (see **Figure 15.10** and **Figure 15.11**). Positions of the emission lines tell us which wavelengths of the radiation are emitted by the gas. Each chemical element has its own characteristic emission spectrum. For each element, the positions of its emission lines are exactly the same as the positions of its absorption lines. This means that atoms of a specific element absorb radiation only at specific wavelengths and radiation that does not have these wavelengths is not absorbed by the element at all. This also means that the radiation emitted by atoms of each element has exactly the same wavelengths as the radiation they absorb.

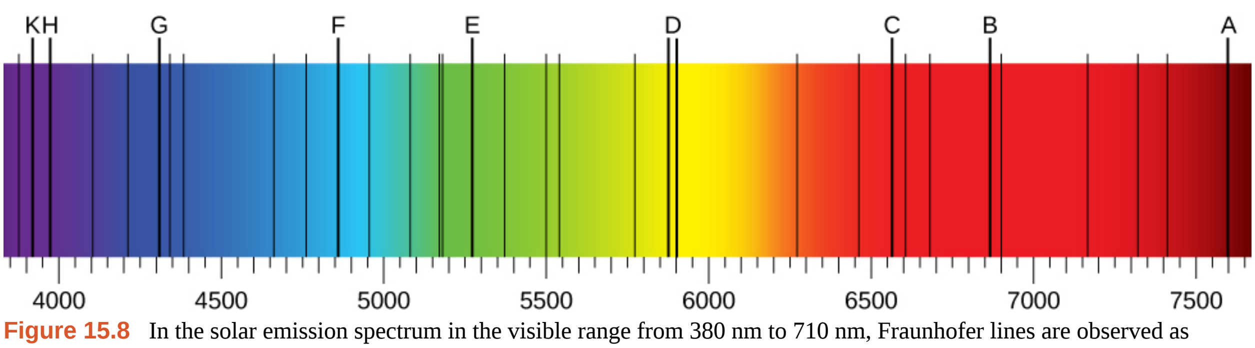


Figure 15.8 In the solar emission spectrum in the visible range from 380 nm to 710 nm, Fraunhofer lines are observed as vertical black lines at specific spectral positions in the continuous spectrum. Highly sensitive modern instruments observe thousands of such lines.

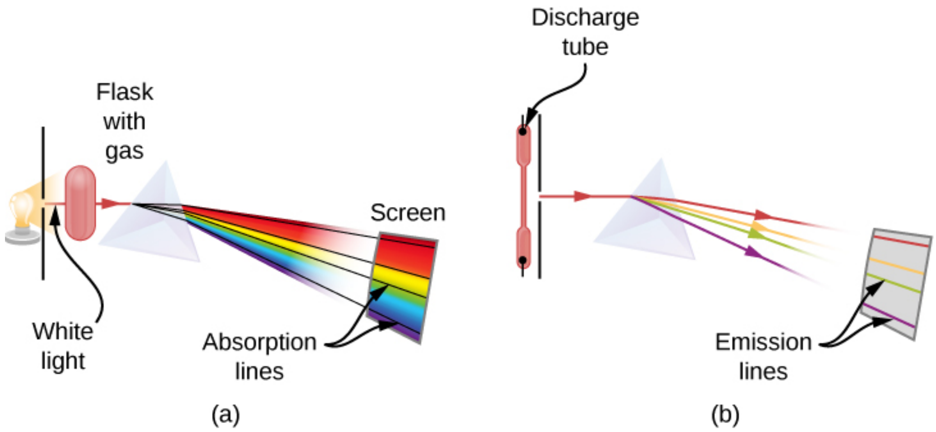


Figure 15.9 Observation of line spectra: (a) setup to observe absorption lines; (b) setup to observe emission lines. (a) White light passes through a cold gas that is contained in a glass flask. A prism is used to separate wavelengths of the passed light. In the spectrum of the passed light, some wavelengths are missing, which are seen as black absorption lines in the continuous spectrum on the viewing screen. (b) A gas is contained in a glass discharge tube that has electrodes at its ends. At a high potential difference between the electrodes, the gas glows and the light emitted from the gas passes through the prism that separates its wavelengths. In the spectrum of the emitted light, only specific wavelengths are present, which are seen as colorful emission lines on the screen.

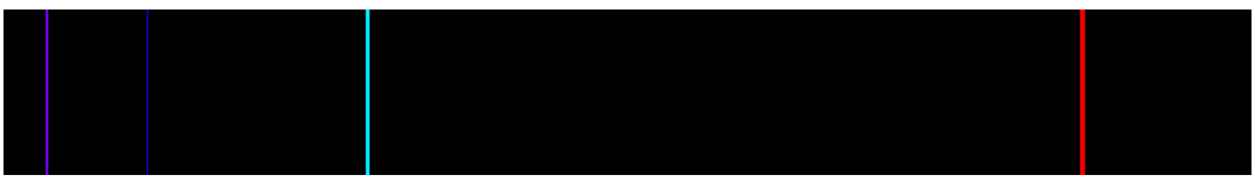


Figure 15.10 The emission spectrum of atomic hydrogen: The spectral positions of emission lines are characteristic for hydrogen atoms. (credit: "Merikanto"/Wikimedia Commons)

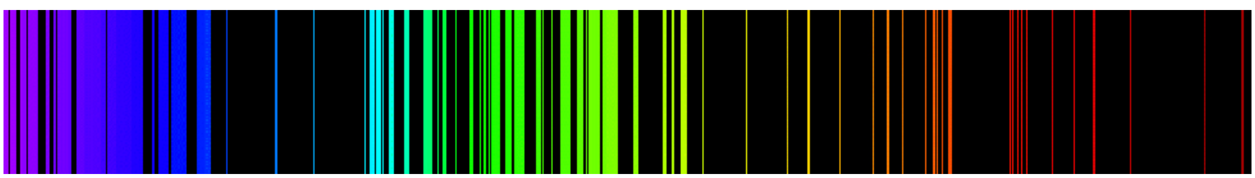


Figure 15.11 The emission spectrum of atomic iron: The spectral positions of emission lines are characteristic for iron atoms.

Emission spectra of the elements have complex structures; they become even more complex for elements with higher atomic numbers. The simplest spectrum, shown in **Figure 15.10**, belongs to the hydrogen atom. Only four lines are visible to the human eye. As you read from right to left in **Figure 15.10**, these lines are: red (656 nm), called the H-α line; aqua (486 nm), blue (434 nm), and violet (410 nm). The lines with wavelengths shorter than 400 nm appear in the ultraviolet part of the spectrum (**Figure 15.10**, far left) and are invisible to the human eye. There are infinitely many invisible spectral lines in the series for hydrogen.

An empirical formula to describe the positions (wavelengths) λ of the hydrogen emission lines in this series was discovered in 1885 by Johann Balmer. It is known as the **Balmer formula**:

$$\frac{1}{\lambda} = R_H\left(\frac{1}{2^2} - \frac{1}{n^2}\right). \tag{15.12}$$

The constant $R_H = 1.09737 \times 10^7 \text{ m}^{-1}$ is called the **Rydberg constant for hydrogen**. In **Equation 15.12**, the positive integer n takes on values $n = 3, 4, 5, 6$ for the four visible lines in this series. The series of emission lines given by the Balmer formula is called the **Balmer series** for hydrogen. Other emission lines of hydrogen that were discovered in the twentieth century are described by the **Rydberg formula**, which summarizes all of the experimental data:

$$\frac{1}{\lambda} = R_{\text{H}}\left(\frac{1}{n_f^2} - \frac{1}{n_i^2}\right), \text{ where } n_i = n_f + 1,\, n_f + 2,\, n_f + 3,\, \ldots \tag{15.13}$$

When $n_f = 1,$ the series of spectral lines is called the **Lyman series**. When $n_f = 2,$ the series is called the Balmer series, and in this case, the Rydberg formula coincides with the Balmer formula. When $n_f = 3,$ the series is called the **Paschen series**. When $n_f = 4,$ the series is called the **Brackett series**. When $n_f = 5,$ the series is called the **Pfund series**. When $n_f = 6,$ we have the **Humphreys series**. As you may guess, there are infinitely many such spectral bands in the spectrum of hydrogen because n_f can be any positive integer number.

The Rydberg formula for hydrogen gives the exact positions of the spectral lines as they are observed in a laboratory; however, at the beginning of the twentieth century, nobody could explain why it worked so well. The Rydberg formula remained unexplained until the first successful model of the hydrogen atom was proposed in 1913.

Example 15.5

Limits of the Balmer Series

Calculate the longest and the shortest wavelengths in the Balmer series.

Strategy

We can use either the Balmer formula or the Rydberg formula. The longest wavelength is obtained when $1/n_i$ is largest, which is when $n_i = n_f + 1 = 3,$ because $n_f = 2$ for the Balmer series. The smallest wavelength is obtained when $1/n_i$ is smallest, which is $1/n_i \to 0$ when $n_i \to \infty$.

Solution

The long-wave limit:

$$\frac{1}{\lambda} = R_{\text{H}}\left(\frac{1}{2^2} - \frac{1}{3^2}\right) = (1.09737 \times 10^7)\frac{1}{\text{m}}\left(\frac{1}{4} - \frac{1}{9}\right) \Rightarrow \lambda = 656.3 \text{ nm}$$

The short-wave limit:

$$\frac{1}{\lambda} = R_{\text{H}}\left(\frac{1}{2^2} - 0\right) = (1.09737 \times 10^7)\frac{1}{\text{m}}\left(\frac{1}{4}\right) \Rightarrow \lambda = 364.6 \text{ nm}$$

Significance

Note that there are infinitely many spectral lines lying between these two limits.

15.6 Check Your Understanding What are the limits of the Lyman series? Can you see these spectral lines?

The key to unlocking the mystery of atomic spectra is in understanding atomic structure. Scientists have long known that matter is made of atoms. According to nineteenth-century science, atoms are the smallest indivisible quantities of matter. This scientific belief was shattered by a series of groundbreaking experiments that proved the existence of subatomic particles, such as electrons, protons, and neutrons.

The electron was discovered and identified as the smallest quantity of electric charge by J.J. Thomson in 1897 in his cathode ray experiments, also known as β-ray experiments: A ***β*-ray** is a beam of electrons. In 1904, Thomson proposed the first model of atomic structure, known as the "plum pudding" model, in which an atom consisted of an unknown positively charged matter with negative electrons embedded in it like plums in a pudding. Around 1900, E. Rutherford, and independently, Paul Ulrich Villard, classified all radiation known at that time as **α-rays**, β-rays, and **γ-rays** (a γ-ray is a beam of highly energetic photons). In 1907, Rutherford and Thomas Royds used spectroscopy methods to show that positively charged particles of α-radiation (called **α-particles**) are in fact doubly ionized atoms of helium. In 1909, Rutherford, Ernest Marsden, and Hans Geiger used α-particles in their famous scattering experiment that disproved Thomson's model.

In the **Rutherford gold foil experiment** (also known as the Geiger–Marsden experiment), α-particles were incident on a thin gold foil and were scattered by gold atoms inside the foil (see **Figure 15.12**). The outgoing particles were detected by a $360°$ scintillation screen surrounding the gold target. When a scattered particle struck the screen, a tiny flash of light (scintillation) was observed at that location. By counting the scintillations seen at various angles with respect to the direction of the incident beam, the scientists could determine what fraction of the incident particles were scattered and what fraction were not deflected at all. If the plum pudding model were correct, there would be no back-scattered α-particles. However, the results of the Rutherford experiment showed that, although a sizable fraction of α-particles emerged from the foil not scattered at all as though the foil were not in their way, a significant fraction of α-particles were back-scattered toward the source. This kind of result was possible only when most of the mass and the entire positive charge of the gold atom were concentrated in a tiny space inside the atom.

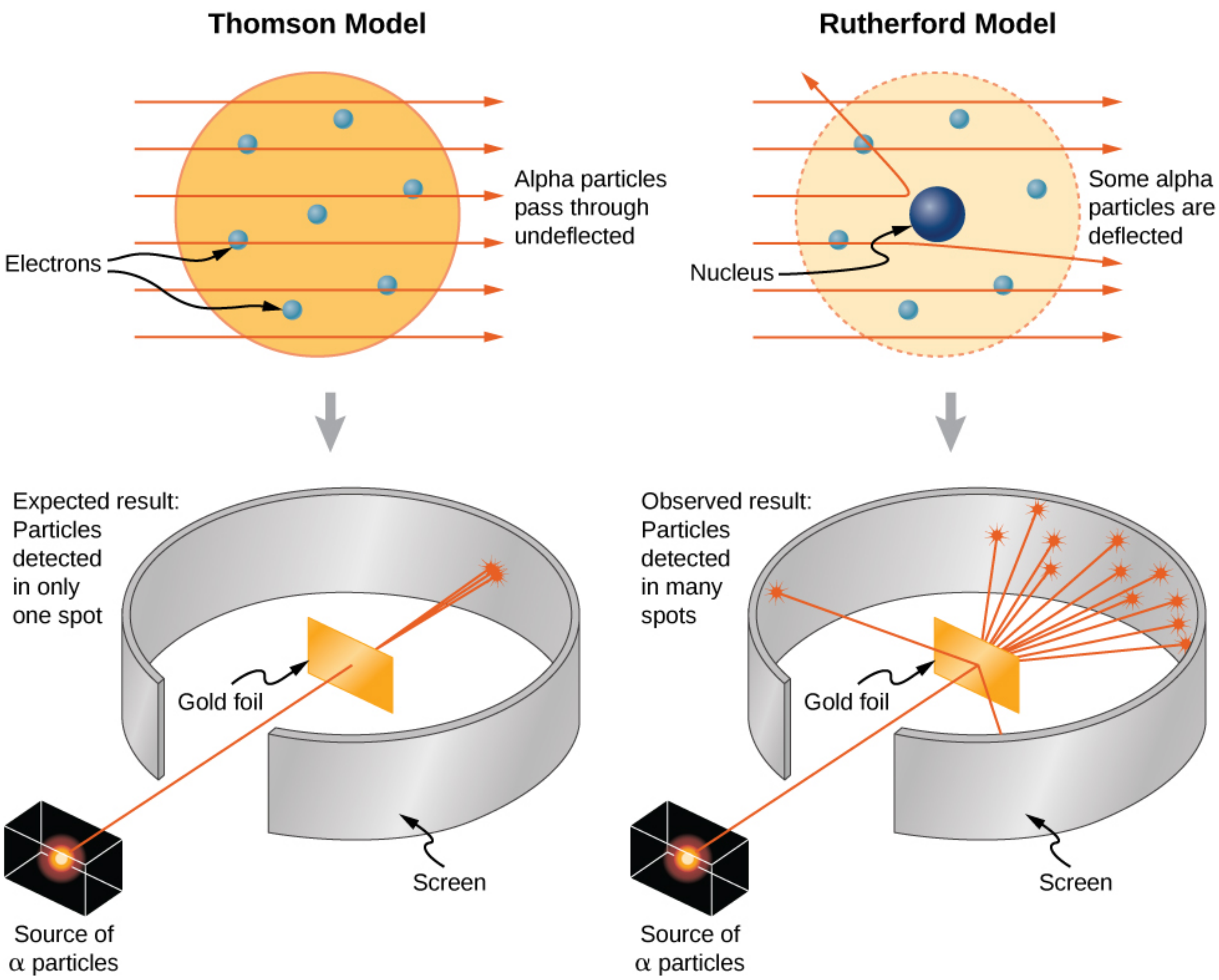


Figure 15.12 The Thomson and Rutherford models of the atom. The Thomson model predicted that nearly all of the incident alpha-particles would be scattered and at small angles. Rutherford and Geiger found that nearly none of the alpha particles were scattered, but those few that were deflected did so through very large angles. The results of Rutherford's experiments were inconsistent with the Thomson model. Rutherford used conservation of momentum and energy to develop a new, and better model of the atom—the nuclear model.

In 1911, Rutherford proposed a **nuclear model of the atom**. In Rutherford's model, an atom contained a positively charged nucleus of negligible size, almost like a point, but included almost the entire mass of the atom. The atom also contained negative electrons that were located within the atom but relatively far away from the nucleus. Ten years later, Rutherford coined the name *proton* for the nucleus of hydrogen and the name *neutron* for a hypothetical electrically neutral particle that would mediate the binding of positive protons in the nucleus (the neutron was discovered in 1932 by James Chadwick). Rutherford is credited with the discovery of the atomic nucleus; however, the Rutherford model of atomic structure does not explain the Rydberg formula for the hydrogen emission lines.

Bohr's model of the hydrogen atom, proposed by Niels Bohr in 1913, was the first quantum model that correctly explained the hydrogen emission spectrum. Bohr's model combines the classical mechanics of planetary motion with the quantum concept of photons. Once Rutherford had established the existence of the atomic nucleus, Bohr's intuition that the negative electron in the hydrogen atom must revolve around the positive nucleus became a logical consequence of the inverse-square-distance law of electrostatic attraction. Recall that Coulomb's law describing the attraction between two opposite charges has a similar form to Newton's universal law of gravitation in the sense that the gravitational force and the electrostatic force are both decreasing as $1/r^2$, where *r* is the separation distance between the bodies. In the same way as Earth revolves around the sun, the negative electron in the hydrogen atom can revolve around the positive nucleus. However, an accelerating charge radiates its energy. Classically, if the electron moved around the nucleus in a planetary fashion, it would be undergoing centripetal acceleration, and thus would be radiating energy that would cause it to spiral down into the nucleus. Such a planetary hydrogen atom would not be stable, which is contrary to what we know about ordinary hydrogen atoms that do not disintegrate. Moreover, the classical motion of the electron is not able to explain the discrete emission spectrum of hydrogen.

To circumvent these two difficulties, Bohr proposed the following three **postulates of Bohr's model**:

1. The negative electron moves around the positive nucleus (proton) in a circular orbit. All electron orbits are centered at the nucleus. Not all classically possible orbits are available to an electron bound to the nucleus.
2. The allowed electron orbits satisfy the *first quantization condition*: In the *n*th orbit, the angular momentum L_n of the electron can take only discrete values:

$$L_n = n\hbar, \text{ where } n = 1, 2, 3, \ldots \tag{15.14}$$

   This postulate says that the electron's angular momentum is quantized. Denoted by r_n and v_n, respectively, the radius of the *n*th orbit and the electron's speed in it, the first quantization condition can be expressed explicitly as

$$m_e v_n r_n = n\hbar. \tag{15.15}$$

3. An electron is allowed to make transitions from one orbit where its energy is E_n to another orbit where its energy is E_m. When an atom absorbs a photon, the electron makes a transition to a higher-energy orbit. When an atom emits a photon, the electron transits to a lower-energy orbit. Electron transitions with the simultaneous photon absorption or photon emission take place *instantaneously*. The allowed electron transitions satisfy the *second quantization condition*:

$$hf = |E_n - E_m| \tag{15.16}$$

   where hf is the energy of either an emitted or an absorbed photon with frequency *f*. The second quantization condition states that an electron's change in energy in the hydrogen atom is quantized.

These three postulates of the early quantum theory of the hydrogen atom allow us to derive not only the Rydberg formula, but also the value of the Rydberg constant and other important properties of the hydrogen atom such as its energy levels, its ionization energy, and the sizes of electron orbits. Note that in Bohr's model, along with two nonclassical quantization postulates, we also have the classical description of the electron as a particle that is subjected to the Coulomb force, and its motion must obey Newton's laws of motion. The hydrogen atom, as an isolated system, must obey the laws of conservation of energy and momentum in the way we know from classical physics. Having this theoretical framework in mind, we are ready to proceed with our analysis.

Electron Orbits

To obtain the size r_n of the electron's *n*th orbit and the electron's speed v_n in it, we turn to Newtonian mechanics. As a charged particle, the electron experiences an electrostatic pull toward the positively charged nucleus in the center of its circular orbit. This electrostatic pull is the centripetal force that causes the electron to move in a circle around the nucleus. Therefore, the magnitude of centripetal force is identified with the magnitude of the electrostatic force:

$$\frac{m_e v_n^2}{r_n} = \frac{1}{4\pi\varepsilon_0}\frac{e^2}{r_n^2}. \qquad (15.17)$$

Here, e denotes the value of the elementary charge. The negative electron and positive proton have the same value of charge, $|q| = e.$ When **Equation 15.17** is combined with the first quantization condition given by **Equation 15.15**, we can solve for the speed, $v_n,$ and for the radius, r_n :

$$v_n = \frac{1}{4\pi\varepsilon_0}\frac{e^2}{\hbar}\frac{1}{n} \qquad (15.18)$$

$$r_n = 4\pi\varepsilon_0\frac{\hbar^2}{m_e e^2}n^2. \qquad (15.19)$$

Note that these results tell us that the electron's speed as well as the radius of its orbit depend only on the index *n* that enumerates the orbit because all other quantities in the preceding equations are fundamental constants. We see from **Equation 15.19** that the size of the orbit grows as the square of *n*. This means that the second orbit is four times as large as the first orbit, and the third orbit is nine times as large as the first orbit, and so on. We also see from **Equation 15.18** that the electron's speed in the orbit decreases as the orbit size increases. The electron's speed is largest in the first Bohr orbit, for $n = 1,$ which is the orbit closest to the nucleus. The radius of the first Bohr orbit is called the **Bohr radius of hydrogen**, denoted as $a_0.$ Its value is obtained by setting $n = 1$ in **Equation 15.19**:

$$a_0 = 4\pi\varepsilon_0\frac{\hbar^2}{m_e e^2} = 5.29 \times 10^{-11}\,\text{m} = 0.529\,\text{Å}. \qquad (15.20)$$

We can substitute a_0 in **Equation 15.19** to express the radius of the *n*th orbit in terms of a_0 :

$$r_n = a_0 n^2. \qquad (15.21)$$

This result means that the electron orbits in hydrogen atom are *quantized* because the orbital radius takes on only specific values of $a_0, 4a_0, 9a_0, 16a_0, \ldots$ given by **Equation 15.21**, and no other values are allowed.

Electron Energies

The total energy E_n of an electron in the *n*th orbit is the sum of its kinetic energy K_n and its electrostatic potential energy $U_n.$ Utilizing **Equation 15.18**, we find that

$$K_n = \frac{1}{2}m_e v_n^2 = \frac{1}{32\pi^2\varepsilon_0^2}\frac{m_e e^4}{\hbar^2}\frac{1}{n^2}. \qquad (15.22)$$

Recall that the electrostatic potential energy of interaction between two charges q_1 and q_2 that are separated by a distance r_{12} is $(1/4\pi\varepsilon_0)q_1 q_2/r_{12}.$ Here, $q_1 = +e$ is the charge of the nucleus in the hydrogen atom (the charge of the proton), $q_2 = -e$ is the charge of the electron and $r_{12} = r_n$ is the radius of the *n*th orbit. Now we use **Equation 15.19** to find the potential energy of the electron:

$$U_n = -\frac{1}{4\pi\varepsilon_0}\frac{e^2}{r_n} = -\frac{1}{16\pi^2\varepsilon_0^2}\frac{m_e e^4}{\hbar^2}\frac{1}{n^2}. \tag{15.23}$$

The total energy of the electron is the sum of **Equation 15.22** and **Equation 15.23**:

$$E_n = K_n + U_n = -\frac{1}{32\pi^2\varepsilon_0^2}\frac{m_e e^4}{\hbar^2}\frac{1}{n^2}. \tag{15.24}$$

Note that the energy depends only on the index *n* because the remaining symbols in **Equation 15.24** are physical constants. The value of the constant factor in **Equation 15.24** is

$$E_0 = \frac{1}{32\pi^2\varepsilon_0^2}\frac{m_e e^4}{\hbar^2} = \frac{1}{8\varepsilon_0^2}\frac{m_e e^4}{h^2} = 2.17 \times 10^{-18}\,\text{J} = 13.6\,\text{eV}. \tag{15.25}$$

It is convenient to express the electron's energy in the *n*th orbit in terms of this energy, as

$$E_n = -E_0\frac{1}{n^2}. \tag{15.26}$$

Now we can see that the electron energies in the hydrogen atom are *quantized* because they can have only discrete values of $-E_0, -E_0/4, -E_0/9, -E_0/16, \ldots$ given by **Equation 15.26**, and no other energy values are allowed. This set of allowed electron energies is called the **energy spectrum of hydrogen** (**Figure 15.13**). The index *n* that enumerates energy levels in Bohr's model is called the energy **quantum number**. We identify the energy of the electron inside the hydrogen atom with the energy of the hydrogen atom. Note that the smallest value of energy is obtained for $n = 1,$ so the hydrogen atom cannot have energy smaller than that. This smallest value of the electron energy in the hydrogen atom is called the **ground state energy of the hydrogen atom** and its value is

$$E_1 = -E_0 = -13.6\,\text{eV}. \tag{15.27}$$

The hydrogen atom may have other energies that are higher than the ground state. These higher energy states are known as **excited energy states of a hydrogen atom**.

There is only one ground state, but there are infinitely many excited states because there are infinitely many values of *n* in **Equation 15.26**. We say that the electron is in the "first exited state" when its energy is E_2 (when $n = 2$), the second excited state when its energy is E_3 (when $n = 3$) and, in general, in the *n*th exited state when its energy is E_{n+1}. There is no highest-of-all excited state; however, there is a limit to the sequence of excited states. If we keep increasing *n* in **Equation 15.26**, we find that the limit is $-\lim_{n\to\infty} E_0/n^2 = 0.$ In this limit, the electron is no longer bound to the nucleus but becomes a free electron. An electron remains bound in the hydrogen atom as long as its energy is negative. An electron that orbits the nucleus in the first Bohr orbit, closest to the nucleus, is in the ground state, where its energy has the smallest value. In the ground state, the electron is most strongly bound to the nucleus and its energy is given by **Equation 15.27**. If we want to remove this electron from the atom, we must supply it with enough energy, $E_\infty,$ to at least balance out its ground state energy E_1:

$$E_\infty + E_1 = 0 \Rightarrow E_\infty = -E_1 = -(-E_0) = E_0 = 13.6\,\text{eV}. \tag{15.28}$$

The energy that is needed to remove the electron from the atom is called the **ionization energy**. The ionization energy E_∞ that is needed to remove the electron from the first Bohr orbit is called the **ionization limit of the hydrogen atom**. The ionization limit in **Equation 15.28** that we obtain in Bohr's model agrees with experimental value.

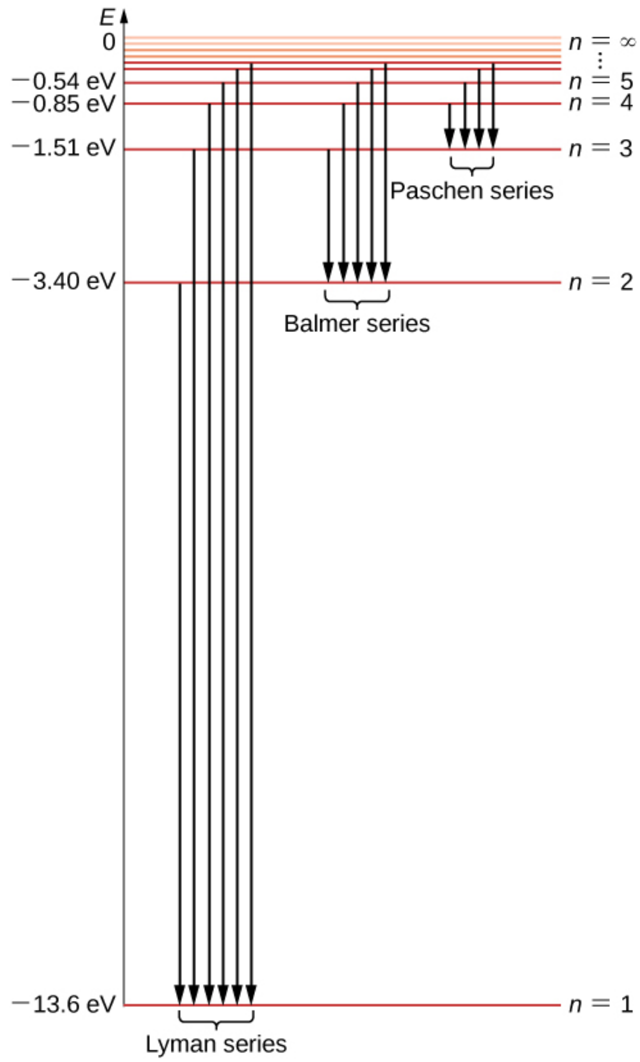


Figure 15.13 The energy spectrum of the hydrogen atom. Energy levels (horizontal lines) represent the bound states of an electron in the atom. There is only one ground state, $n = 1$, and infinite quantized excited states. The states are enumerated by the quantum number $n = 1, 2, 3, 4, \ldots.$ Vertical lines illustrate the allowed electron transitions between the states. Downward arrows illustrate transitions with an emission of a photon with a wavelength in the indicated spectral band.

Spectral Emission Lines of Hydrogen

To obtain the wavelengths of the emitted radiation when an electron makes a transition from the *n*th orbit to the *m*th orbit, we use the second of Bohr's quantization conditions and **Equation 15.26** for energies. The emission of energy from the atom can occur only when an electron makes a transition from an excited state to a lower-energy state. In the course of such a transition, the emitted photon carries away the difference of energies between the states involved in the transition. The transition cannot go in the other direction because the energy of a photon cannot be negative, which means that for emission we must have $E_n > E_m$ and $n > m$. Therefore, the third of Bohr's postulates gives

$$hf = \left|E_n - E_m\right| = E_n - E_m = -E_0\frac{1}{n^2} + E_0\frac{1}{m^2} = E_0\left(\frac{1}{m^2} - \frac{1}{n^2}\right). \tag{15.29}$$

Now we express the photon's energy in terms of its wavelength, $hf = hc/\lambda$, and divide both sides of **Equation 15.29** by hc. The result is

$$\frac{1}{\lambda}=\frac{E_0}{hc}\left(\frac{1}{m^2}-\frac{1}{n^2}\right). \quad (15.30)$$

The value of the constant in this equation is

$$\frac{E_0}{hc}=\frac{13.6\,\text{eV}}{(4.136\times10^{-15}\,\text{eV}\cdot\text{s})(2.997\times10^{8}\,\text{m/s})}=1.097\times10^{7}\,\frac{1}{\text{m}}. \quad (15.31)$$

This value is exactly the Rydberg constant R_H in the Rydberg heuristic formula **Equation 15.13**. In fact, **Equation 15.30** is identical to the Rydberg formula, because for a given *m*, we have $n=m+1,\ m+2,\ ....$ In this way, the Bohr quantum model of the hydrogen atom allows us to derive the experimental Rydberg constant from first principles and to express it in terms of fundamental constants. Transitions between the allowed electron orbits are illustrated in **Figure 15.13**.

We can repeat the same steps that led to **Equation 15.30** to obtain the wavelength of the absorbed radiation; this again gives **Equation 15.30** but this time for the positions of absorption lines in the absorption spectrum of hydrogen. The only difference is that for absorption, the quantum number *m* is the index of the orbit occupied by the electron before the transition (lower-energy orbit) and the quantum number *n* is the index of the orbit to which the electron makes the transition (higher-energy orbit). The difference between the electron energies in these two orbits is the energy of the absorbed photon.

Example 15.6

Size and Ionization Energy of the Hydrogen Atom in an Excited State

If a hydrogen atom in the ground state absorbs a 93.7-nm photon, corresponding to a transition line in the Lyman series, how does this affect the atom's energy and size? How much energy is needed to ionize the atom when it is in this excited state? Give your answers in absolute units, and relative to the ground state.

Strategy

Before the absorption, the atom is in its ground state. This means that the electron transition takes place from the orbit $m=1$ to some higher *n*th orbit. First, we must determine n for the absorbed wavelength $\lambda=93.7\text{ nm}$. Then, we can use **Equation 15.26** to find the energy E_n of the excited state and its ionization energy $E_{\infty,\,n}$, and use **Equation 15.21** to find the radius r_n of the atom in the excited state. To estimate *n*, we use **Equation 15.30**.

Solution

Substitute $m=1$ and $\lambda=93.7\text{ nm}$ in **Equation 15.30** and solve for n. You should not expect to obtain a perfect integer answer because of rounding errors, but your answer will be close to an integer, and you can estimate *n* by taking the integral part of your answer:

$$\frac{1}{\lambda}=R_H\left(\frac{1}{1^2}-\frac{1}{n^2}\right)\Rightarrow n=\frac{1}{\sqrt{1-\frac{1}{\lambda R_H}}}=\frac{1}{\sqrt{1-\frac{1}{(93.7\times10^{-9}\,\text{m})(1.097\times10^{7}\,\text{m}^{-1})}}}=6.07\Rightarrow n=6.$$

The radius of the $n=6$ orbit is

$$r_n=a_0n^2=a_0 6^2=36a_0=36(0.529\times10^{-10}\,\text{m})=19.04\times10^{-10}\,\text{m}\cong 19.0\,\text{Å}.$$

Thus, after absorbing the 93.7-nm photon, the size of the hydrogen atom in the excited $n=6$ state is 36 times larger than before the absorption, when the atom was in the ground state. The energy of the fifth excited state ($n=6$) is:

$$E_n=-\frac{E_0}{n^2}=-\frac{E_0}{6^2}=-\frac{E_0}{36}=-\frac{13.6\,\text{eV}}{36}\cong-0.378\,\text{eV}.$$

After absorbing the 93.7-nm photon, the energy of the hydrogen atom is larger than it was before the absorption. Ionization of the atom when it is in the fifth excited state ($n=6$) requites 36 times less energy than is needed when the atom is in the ground state:

$$E_{\infty,\,6} = -E_6 = -(-0.378\text{ eV}) = 0.378\text{ eV}.$$

Significance

We can analyze any spectral line in the spectrum of hydrogen in the same way. Thus, the experimental measurements of spectral lines provide us with information about the atomic structure of the hydrogen atom.

15.7 Check Your Understanding When an electron in a hydrogen atom is in the first excited state, what prediction does the Bohr model give about its orbital speed and kinetic energy? What is the magnitude of its orbital angular momentum?

Bohr's model of the hydrogen atom also correctly predicts the spectra of some hydrogen-like ions. **Hydrogen-like ions** are atoms of elements with an atomic number Z larger than one ($Z = 1$ for hydrogen) but with all electrons removed except one. For example, an electrically neutral helium atom has an atomic number $Z = 2.$ This means it has two electrons orbiting the nucleus with a charge of $q = +Ze.$ When one of the orbiting electrons is removed from the helium atom (we say, when the helium atom is singly ionized), what remains is a hydrogen-like atomic structure where the remaining electron orbits the nucleus with a charge of $q = +Ze.$ This type of situation is described by the Bohr model. Assuming that the charge of the nucleus is not $+e$ but $+Ze,$ we can repeat all steps, beginning with **Equation 15.17**, to obtain the results for a hydrogen-like ion:

$$r_n = \frac{a_0}{Z}n^2 \tag{15.32}$$

where a_0 is the Bohr orbit of hydrogen, and

$$E_n = -Z^2 E_0 \frac{1}{n^2} \tag{15.33}$$

where E_0 is the ionization limit of a hydrogen atom. These equations are good approximations as long as the atomic number Z is not too large.

The Bohr model is important because it was the first model to postulate the quantization of electron orbits in atoms. Thus, it represents an early quantum theory that gave a start to developing modern quantum theory. It introduced the concept of a quantum number to describe atomic states. The limitation of the early quantum theory is that it cannot describe atoms in which the number of electrons orbiting the nucleus is larger than one. The Bohr model of hydrogen is a semi-classical model because it combines the classical concept of electron orbits with the new concept of quantization. The remarkable success of this model prompted many physicists to seek an explanation for why such a model should work at all, and to seek an understanding of the physics behind the postulates of early quantum theory. This search brought about the onset of an entirely new concept of "matter waves."

15.3 | Atomic Spectra and X-rays

Learning Objectives

By the end of this section, you will be able to:

- Describe the absorption and emission of radiation in terms of atomic energy levels and energy differences
- Use quantum numbers to estimate the energy, frequency, and wavelength of photons produced by atomic transitions in multi-electron atoms
- Explain radiation concepts in the context of atomic fluorescence and X-rays

The study of atomic spectra provides most of our knowledge about atoms. In modern science, atomic spectra are used to identify species of atoms in a range of objects, from distant galaxies to blood samples at a crime scene.

The theoretical basis of atomic spectroscopy is the transition of electrons between energy levels in atoms. For example, if an electron in a hydrogen atom makes a transition from the $n = 3$ to the $n = 2$ shell, the atom emits a photon with a wavelength

$$\lambda = \frac{c}{f} = \frac{h \cdot c}{h \cdot f} = \frac{hc}{\Delta E} = \frac{hc}{E_3 - E_2}, \tag{15.34}$$

where $\Delta E = E_3 - E_2$ is energy carried away by the photon and $hc = 1940\,\text{eV} \cdot \text{nm}$. After this radiation passes through a spectrometer, it appears as a sharp spectral line on a screen. The Bohr model of this process is shown in **Figure 15.14**. If the electron later absorbs a photon with energy ΔE, the electron returns to the $n = 3$ shell. (We examined **The Bohr Model** earlier.)

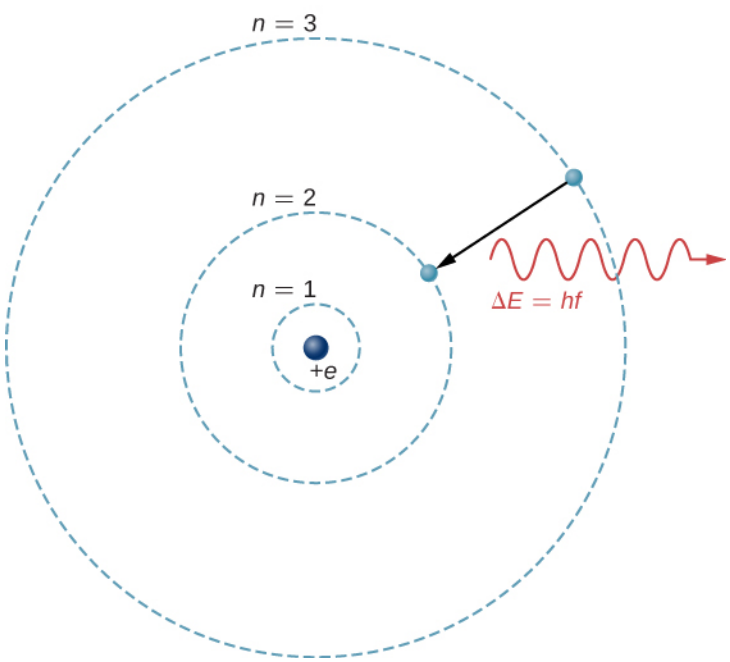


Figure 15.14 An electron transition from the $n = 3$ to the $n = 2$ shell of a hydrogen atom.

To understand atomic transitions in multi-electron atoms, it is necessary to consider many effects, including the Coulomb repulsion between electrons and internal magnetic interactions (spin-orbit and spin-spin couplings). Fortunately, many properties of these systems can be understood by neglecting interactions between electrons and representing each electron by its own single-particle wave function ψ_{nlm}.

Atomic transitions must obey **selection rules**. These rules follow from principles of quantum mechanics and symmetry.

Selection rules classify transitions as either allowed or forbidden. (Forbidden transitions do occur, but the probability of the typical forbidden transition is very small.) For a hydrogen-like atom, atomic transitions that involve electromagnetic interactions (the emission and absorption of photons) obey the following selection rule:

$$\Delta l = \pm 1, \tag{15.35}$$

where l is associated with the magnitude of orbital angular momentum,

$$L = \sqrt{l(l+1)}\hbar. \tag{15.36}$$

For multi-electron atoms, similar rules apply. To illustrate this rule, consider the observed atomic transitions in hydrogen (H), sodium (Na), and mercury (Hg) (**Figure 15.15**). The horizontal lines in this diagram correspond to atomic energy levels, and the transitions allowed by this selection rule are shown by lines drawn between these levels. The energies of these states are on the order of a few electron volts, and photons emitted in transitions are in the visible range. Technically, atomic transitions can violate the selection rule, but such transitions are uncommon.

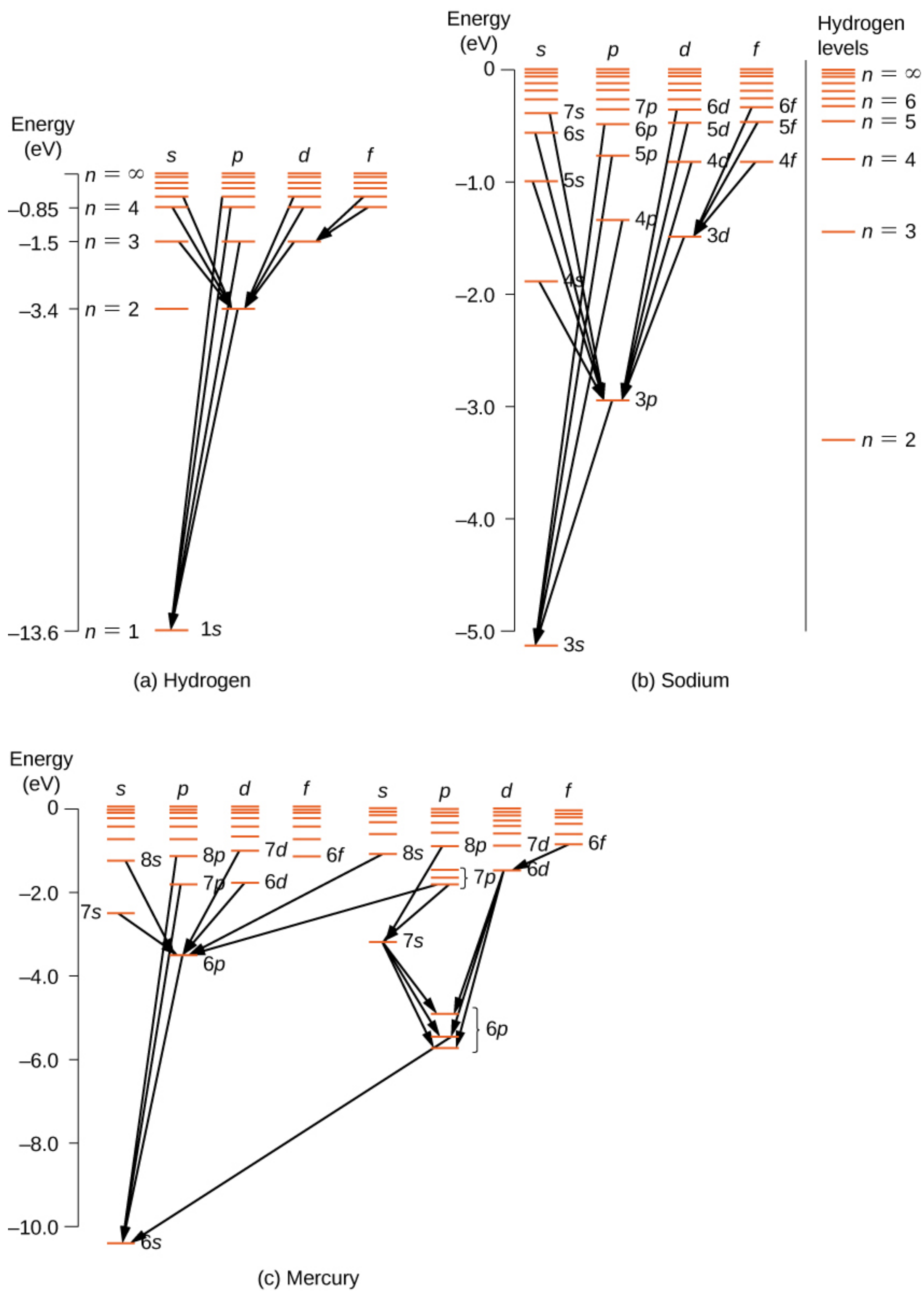


Figure 15.15 Energy-level diagrams for (a) hydrogen, (b) sodium, and (c) mercury. For comparison, hydrogen energy levels are shown in the sodium diagram.

The hydrogen atom has the simplest energy-level diagram. If we neglect electron spin, all states with the same value of n have the same total energy. However, spin-orbit coupling splits the $n = 2$ states into two angular momentum states (*s* and *p*) of slightly different energies. (These levels are not vertically displaced, because the energy splitting is too small to show up in this diagram.) Likewise, spin-orbit coupling splits the $n = 3$ states into three angular momentum states (*s*, *p*, and *d*).

The energy-level diagram for hydrogen is similar to sodium, because both atoms have one electron in the outer shell. The valence electron of sodium moves in the electric field of a nucleus shielded by electrons in the inner shells, so it does not experience a simple $1/r$ Coulomb potential and its total energy depends on both n and l. Interestingly, mercury has two separate energy-level diagrams; these diagrams correspond to two net spin states of its 6*s* (valence) electrons.

Example 15.7

The Sodium Doublet

The spectrum of sodium is analyzed with a spectrometer. Two closely spaced lines with wavelengths 589.00 nm and 589.59 nm are observed. (a) If the doublet corresponds to the excited (valence) electron that transitions from some excited state down to the 3*s* state, what was the original electron angular momentum? (b) What is the energy difference between these two excited states?

Strategy

Sodium and hydrogen belong to the same column or chemical group of the periodic table, so sodium is "hydrogen-like." The outermost electron in sodium is in the 3*s* ($l = 0$) subshell and can be excited to higher energy levels. As for hydrogen, subsequent transitions to lower energy levels must obey the selection rule:

$$\Delta l = \pm 1.$$

We must first determine the quantum number of the initial state that satisfies the selection rule. Then, we can use this number to determine the magnitude of orbital angular momentum of the initial state.

Solution

a. Allowed transitions must obey the selection rule. If the quantum number of the initial state is $l = 0$, the transition is forbidden because $\Delta l = 0$. If the quantum number of the initial state is $l = 2, 3, 4$,…the transition is forbidden because $\Delta l > 1$. Therefore, the quantum of the initial state must be $l = 1$. The orbital angular momentum of the initial state is

$$L = \sqrt{l(l+1)}\hbar = 1.41\hbar.$$

b. Because the final state for both transitions is the same (3*s*), the difference in energies of the photons is equal to the difference in energies of the two excited states. Using the equation

$$\Delta E = hf = h\left(\frac{c}{\lambda}\right),$$

we have

$$\begin{aligned}\Delta E &= hc\left(\frac{1}{\lambda_1} - \frac{1}{\lambda_2}\right) \\ &= \left(4.14 \times 10^{-15}\ \text{eVs}\right)\left(3.00 \times 10^{8}\ \text{m/s}\right) \times \left(\frac{1}{589.00 \times 10^{-9}\ \text{m}} - \frac{1}{589.59 \times 10^{-9}\ \text{m}}\right) \\ &= 2.11 \times 10^{-3}\ \text{eV}.\end{aligned}$$

Significance

To understand the difficulty of measuring this energy difference, we compare this difference with the average energy of the two photons emitted in the transition. Given an average wavelength of 589.30 nm, the average energy of the photons is

$$E = \frac{hc}{\lambda} = \frac{(4.14 \times 10^{-15}\ \text{eVs})(3.00 \times 10^{8}\ \text{m/s})}{589.30 \times 10^{-9}\ \text{m}} = 2.11\ \text{eV}.$$

The energy difference ΔE is about 0.1% (1 part in 1000) of this average energy. However, a sensitive spectrometer can measure the difference.

X-rays

The study of atomic energy transitions enables us to understand X-rays and X-ray technology. Like all electromagnetic radiation, X-rays are made of photons. X-ray photons are produced when electrons in the outermost shells of an atom drop to the inner shells. (Hydrogen atoms do not emit X-rays, because the electron energy levels are too closely spaced together to permit the emission of high-frequency radiation.) Transitions of this kind are normally forbidden because the lower states are already filled. However, if an inner shell has a vacancy (an inner electron is missing, perhaps from being knocked away by a high-speed electron), an electron from one of the outer shells can drop in energy to fill the vacancy. The energy gap for such a transition is relatively large, so wavelength of the radiated X-ray photon is relatively short.

X-rays can also be produced by bombarding a metal target with high-energy electrons, as shown in **Figure 15.16**. In the figure, electrons are boiled off a filament and accelerated by an electric field into a tungsten target. According to the classical theory of electromagnetism, *any* charged particle that accelerates emits radiation. Thus, when the electron strikes the tungsten target, and suddenly slows down, the electron emits **braking radiation**. (Braking radiation refers to radiation produced by any charged particle that is slowed by a medium.) In this case, braking radiation contains a continuous range of frequencies, because the electrons will collide with the target atoms in slightly different ways.

Braking radiation is not the only type of radiation produced in this interaction. In some cases, an electron collides with another inner-shell electron of a target atom, and knocks the electron out of the atom—billiard ball style. The empty state is filled when an electron in a higher shell drops into the state (drop in energy level) and emits an X-ray photon.

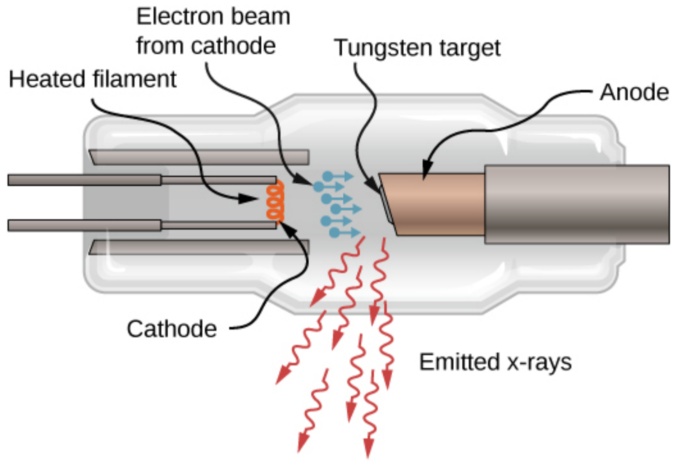


Figure 15.16 A sketch of an X-ray tube. X-rays are emitted from the tungsten target.

Historically, X-ray spectral lines were labeled with letters (K, L, M, N, …). These letters correspond to the atomic shells ($n = 1, 2, 3, 4, \ldots$). X-rays produced by a transition from any higher shell to the K ($n = 1$) shell are labeled as K X-rays. X-rays produced in a transition from the L ($n = 2$) shell are called K_α X-rays; X-rays produced in a transition from the M ($n = 3$) shell are called K_β X-rays; X-rays produced in a transition from the N ($n = 4$) shell are called K_γ X-rays; and so forth. Transitions from higher shells to L and M shells are labeled similarly. These transitions are represented by an energy-level diagram in **Figure 15.17**.

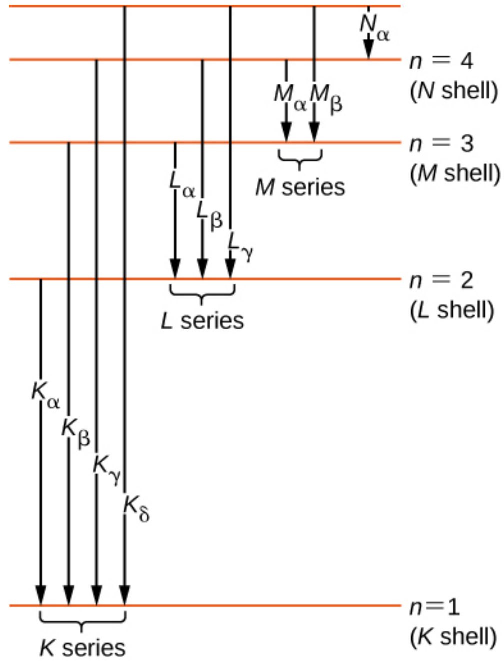


Figure 15.17 X-ray transitions in an atom.

The distribution of X-ray wavelengths produced by striking metal with a beam of electrons is given in **Figure 15.18**. X-ray transitions in the target metal appear as peaks on top of the braking radiation curve. Photon frequencies corresponding to the spikes in the X-ray distribution are called characteristic frequencies, because they can be used to identify the target metal. The sharp cutoff wavelength (just below the K_γ peak) corresponds to an electron that loses all of its energy to a single photon. Radiation of shorter wavelengths is forbidden by the conservation of energy.

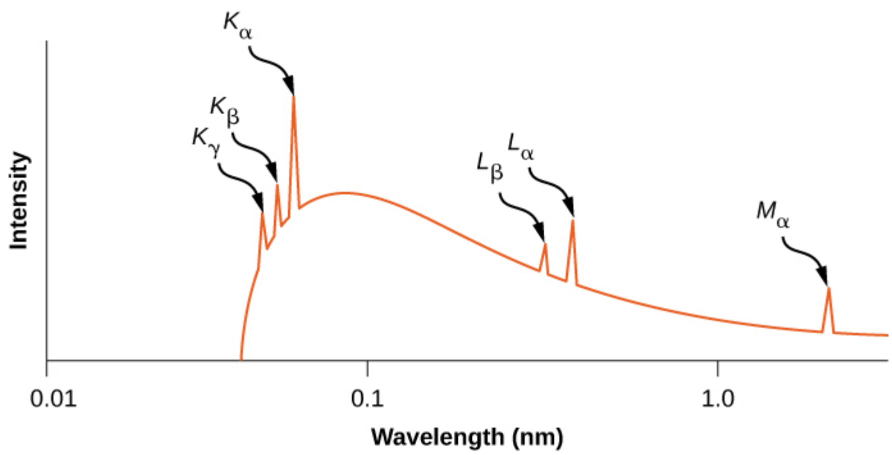


Figure 15.18 X-ray spectrum from a silver target. The peaks correspond to characteristic frequencies of X-rays emitted by silver when struck by an electron beam.

Example 15.8

X-Rays from Aluminum

Estimate the characteristic energy and frequency of the K_α X-ray for aluminum ($Z = 13$).

Strategy

A K_α X-ray is produced by the transition of an electron in the L ($n = 2$) shell to the K ($n = 1$) shell. An electron in the L shell "sees" a charge $Z = 13 - 1 = 12,$ because one electron in the K shell shields the nuclear charge. (Recall, two electrons are not in the K shell because the other electron state is vacant.) The frequency of the emitted photon can be estimated from the energy difference between the L and K shells.

Solution

The energy difference between the L and K shells in a hydrogen atom is 10.2 eV. Assuming that other electrons in the L shell or in higher-energy shells do not shield the nuclear charge, the energy difference between the L and K shells in an atom with $Z = 13$ is approximately

$$\Delta E_{L \to K} \approx (Z-1)^2(10.2\text{ eV}) = (13-1)^2(10.2\text{ eV}) = 1.47 \times 10^3\text{ eV}. \tag{15.37}$$

Based on the relationship $f = \left(\Delta E_{L \to K}\right)/h$, the frequency of the X-ray is

$$f = \frac{1.47 \times 10^3\text{ eV}}{4.14 \times 10^{-15}\text{ eV} \cdot \text{s}} = 3.55 \times 10^{17}\text{ Hz}.$$

Significance

The wavelength of the typical X-ray is 0.1–10 nm. In this case, the wavelength is:

$$\lambda = \frac{c}{f} = \frac{3.0 \times 10^8\text{ m/s}}{3.55 \times 10^{17}\text{ Hz}} = 8.5 \times 10^{-10} = 0.85\text{ nm}.$$

Hence, the transition $L \to K$ in aluminum produces X-ray radiation.

X-ray production provides an important test of quantum mechanics. According to the Bohr model, the energy of a K_α X-ray depends on the nuclear charge or atomic number, Z. If Z is large, Coulomb forces in the atom are large, energy differences (ΔE) are large, and, therefore, the energy of radiated photons is large. To illustrate, consider a single electron in a multi-electron atom. Neglecting interactions between the electrons, the allowed energy levels are

$$E_n = -\frac{Z^2(13.6\text{ eV})}{n^2}, \tag{15.38}$$

where n = 1, 2, …and Z is the atomic number of the nucleus. However, an electron in the L ($n = 2$) shell "sees" a charge $Z - 1$, because one electron in the K shell shields the nuclear charge. (Recall that there is only one electron in the K shell because the other electron was "knocked out.") Therefore, the approximate energies of the electron in the L and K shells are

$$\begin{aligned} E_L &\approx -\frac{(Z-1)^2(13.6\text{ eV})}{2^2} \\ E_K &\approx -\frac{(Z-1)^2(13.6\text{ eV})}{1^2}. \end{aligned}$$

The energy carried away by a photon in a transition from the L shell to the K shell is therefore

$$\begin{aligned} \Delta E_{L \to K} &= (Z-1)^2(13.6\text{ eV})\left(\frac{1}{1^2} - \frac{1}{2^2}\right) \\ &= (Z-1)^2(10.2\text{ eV}), \end{aligned}$$

where Z is the atomic number. In general, the X-ray photon energy for a transition from an outer shell to the K shell is

$$\Delta E_{L \to K} = hf = \text{constant} \times (Z-1)^2,$$

or

$$(Z-1) = \text{constant}\sqrt{f}, \tag{15.39}$$

where f is the frequency of a K_α X-ray. This equation is **Moseley's law**. For large values of Z, we have approximately

$$Z \approx \text{constant}\sqrt{f}.$$

This prediction can be checked by measuring f for a variety of metal targets. This model is supported if a plot of Z versus $\sqrt{f}$ data (called a **Moseley plot**) is linear. Comparison of model predictions and experimental results, for both the K and L series, is shown in **Figure 15.19**. The data support the model that X-rays are produced when an outer shell electron drops in energy to fill a vacancy in an inner shell.

15.8 Check Your Understanding X-rays are produced by bombarding a metal target with high-energy electrons. If the target is replaced by another with two times the atomic number, what happens to the frequency of X-rays?

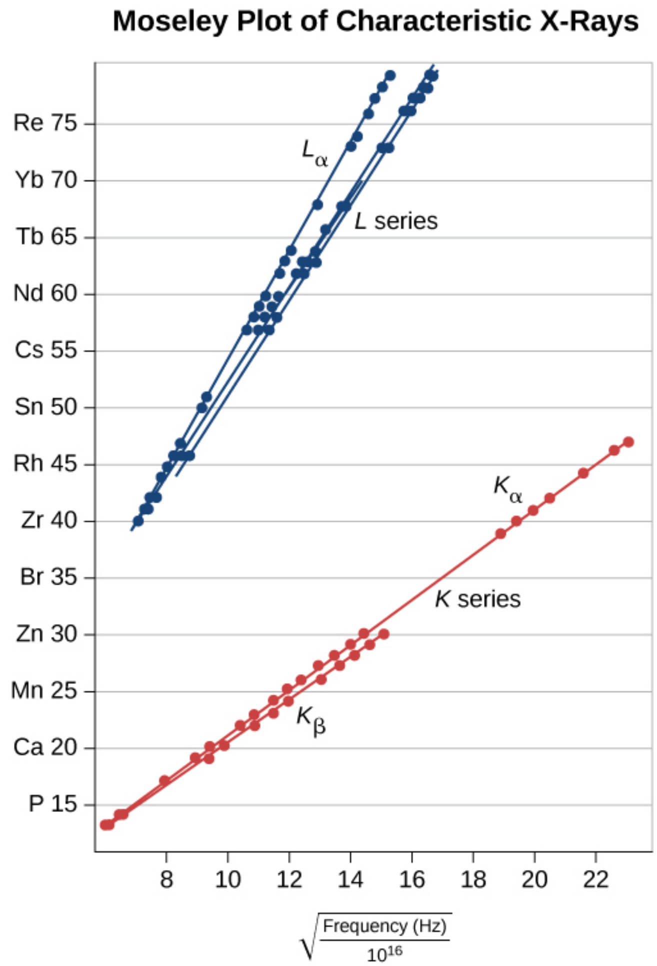


Figure 15.19 A Moseley plot. These data were adapted from Moseley's original data (H. G. J. Moseley, *Philos. Mag*. (6) 77:703, 1914).

Example 15.9

Characteristic X-Ray Energy

Calculate the approximate energy of a K_α X-ray from a tungsten anode in an X-ray tube.

Strategy

Two electrons occupy a filled K shell. A vacancy in this shell would leave one electron, so the effective charge for an electron in the L shell would be $Z - 1$ rather than Z. For tungsten, $Z = 74,$ so the effective charge is 73. This number can be used to calculate the energy-level difference between the L and K shells, and, therefore, the energy carried away by a photon in the transition $L \rightarrow K$.

Solution

The effective Z is 73, so the K_α X-ray energy is given by

$$E_{K_\alpha} = \Delta E = E_\text{i} - E_\text{f} = E_2 - E_1,$$

where

$$E_1 = -\frac{Z^2}{1^2}E_0 = -\frac{73^2}{1}(13.6\,\text{eV}) = -72.5\,\text{keV}$$

and

$$E_2 = -\frac{Z^2}{2^2}E_0 = -\frac{73^2}{4}(13.6\,\text{eV}) = -18.1\,\text{keV}.$$

Thus,

$$E_{K_\alpha} = -18.1\,\text{keV} - (-72.5\,\text{keV}) = 54.4\,\text{keV}.$$

Significance

This large photon energy is typical of X-rays. X-ray energies become progressively larger for heavier elements because their energy increases approximately as Z^2. An acceleration voltage of more than 50,000 volts is needed to "knock out" an inner electron from a tungsten atom.

15.4 | Molecular Spectra

Learning Objectives

By the end of this section, you will be able to:

- Use the concepts of vibrational and rotational energy to describe energy transitions in a diatomic molecule
- Explain key features of a vibrational-rotational energy spectrum of a diatomic molecule
- Estimate allowed energies of a rotating molecule
- Determine the equilibrium separation distance between atoms in a diatomic molecule from the vibrational-rotational absorption spectrum

Molecular energy levels are more complicated than atomic energy levels because molecules can also vibrate and rotate. The energies associated with such motions lie in different ranges and can therefore be studied separately. Electronic transitions are of order 1 eV, vibrational transitions are of order 10^{-2} eV, and rotational transitions are of order 10^{-3} eV. For complex molecules, these energy changes are difficult to characterize, so we begin with the simple case of a diatomic molecule.

As we saw in **Equation 12.5**, the energy of rotation of a diatomic molecule is given by

$$E_r = \frac{L^2}{2I}, \tag{15.40}$$

where I is the moment of inertia and L is the angular momentum. According to quantum mechanics, the rotational angular momentum is quantized:

$$L = \sqrt{l(l+1)}\hbar \; (l = 0, 1, 2, 3, ...), \tag{15.41}$$

where l is the orbital angular quantum number. The allowed **rotational energy level** of a diatomic molecule is therefore

$$E_r = l(l+1)\frac{\hbar^2}{2I} = l(l+1)E_{0r} \quad (l = 0,\ 1,\ 2,\ 3,...), \tag{15.42}$$

where the characteristic rotational energy of a molecule is defined as

$$E_{0r} = \frac{\hbar^2}{2I}. \tag{15.43}$$

For a diatomic molecule, the moment of inertia with reduced mass μ is

$$I = \mu r_0^2, \tag{15.44}$$

where r_0 is the total distance between the atoms. The energy difference between rotational levels is therefore

$$\Delta E_r = E_{l+1} - E_l = 2(l+1)\,E_{0r}. \tag{15.45}$$

A detailed study of transitions between rotational energy levels brought about by the absorption or emission of radiation (a so-called **electric dipole transition**) requires that

$$\Delta l = \pm 1. \tag{15.46}$$

This rule, known as a **selection rule**, limits the possible transitions from one quantum state to another. Equation 15.46 is the selection rule for rotational energy transitions. It applies only to diatomic molecules that have an electric dipole moment. For this reason, symmetric molecules such as H_2 and N_2 do not experience rotational energy transitions due to the absorption or emission of electromagnetic radiation.

Example 15.10

The Rotational Energy of HCl

Determine the lowest three rotational energy levels of a hydrogen chloride (HCl) molecule.

Strategy

Hydrogen chloride (HCl) is a diatomic molecule with an equilibrium separation distance of 0.127 nm. Rotational energy levels depend only on the momentum of inertia I and the orbital angular momentum quantum number l (in this case, $l = 0$, 1, and 2). The momentum of inertia depends, in turn, on the equilibrium separation distance (which is given) and the reduced mass, which depends on the masses of the H and Cl atoms.

Solution

First, we compute the reduced mass. If Particle 1 is hydrogen and Particle 2 is chloride, we have

$$\mu = \frac{m_1 m_2}{m_1 + m_2} = \frac{(1.0\text{ u})(35.4\text{ u})}{1.0\text{ u} + 35.4\text{ u}} = 0.97\text{ u} = 0.97\text{ u}\left(\frac{931.5\frac{\text{MeV}}{c^2}}{1\text{ u}}\right) = 906\frac{\text{MeV}}{c^2}.$$

The corresponding rest mass energy is therefore

$$\mu c^2 = 9.06 \times 10^8\text{ eV}.$$

This allows us to calculate the characteristic energy:

$$E_{0r} = \frac{\hbar^2}{2I} = \frac{\hbar^2}{2\left(\mu r_0^2\right)} = \frac{(\hbar c)^2}{2\left(\mu c^2\right)r_0^2} = \frac{(197.3\ \text{eV} \cdot \text{nm})^2}{2\left(9.06 \times 10^8\ \text{eV}\right)(0.127\ \text{nm})^2} = 1.33 \times 10^{-3}\ \text{eV}.$$

(Notice how this expression is written in terms of the rest mass energy. This technique is common in modern physics calculations.) The rotational energy levels are given by

$$E_r = l(l+1)\frac{\hbar^2}{2I} = l(l+1)E_{0r},$$

where l is the orbital quantum number. The three lowest rotational energy levels of an HCl molecule are therefore

$$l = 0;\ E_r = 0\ \text{eV}\ \text{(no rotation)},$$

$$l = 1;\ E_r = 2\,E_{0r} = 2.66 \times 10^{-3}\ \text{eV},$$

$$l = 2;\ E_r = 6\,E_{0r} = 7.99 \times 10^{-3}\ \text{eV}.$$

Significance

The rotational spectrum is associated with weak transitions (1/1000 to 1/100 of an eV). By comparison, the energy of an electron in the ground state of hydrogen is $-13.6\ \text{eV}$.

15.9 Check Your Understanding What does the energy separation between absorption lines in a rotational spectrum of a diatomic molecule tell you?

The **vibrational energy level**, which is the energy level associated with the vibrational energy of a molecule, is more difficult to estimate than the rotational energy level. However, we can estimate these levels by assuming that the two atoms in the diatomic molecule are connected by an ideal spring of spring constant k. The potential energy of this spring system is

$$U_{\text{osc}} = \frac{1}{2}k\,\Delta r^2, \tag{15.47}$$

Where Δr is a change in the "natural length" of the molecule along a line that connects the atoms. Solving Schrödinger's equation for this potential gives

$$E_n = \left(n + \frac{1}{2}\right)\hbar\omega\ (n = 0,\ 1,\ 2,\ \ldots), \tag{15.48}$$

Where ω is the natural angular frequency of vibration and n is the vibrational quantum number. The prediction that vibrational energy levels are evenly spaced $(\Delta E = \hbar\omega)$ turns out to be good at lower energies.

A detailed study of transitions between vibrational energy levels induced by the absorption or emission of radiation (and the specifically so-called electric dipole transition) requires that

$$\Delta n = \pm 1. \tag{15.49}$$

Equation 15.49 represents the selection rule for vibrational energy transitions. As mentioned before, this rule applies only to diatomic molecules that have an electric dipole moment. Symmetric molecules do not experience such transitions.

Due to the selection rules, the absorption or emission of radiation by a diatomic molecule involves a transition in vibrational and rotational states. Specifically, if the vibrational quantum number (n) changes by one unit, then the rotational quantum number (l) changes by one unit. An energy-level diagram of a possible transition is given in Figure 15.20. The absorption spectrum for such transitions in hydrogen chloride (HCl) is shown in Figure 15.21. The absorption peaks are due to transitions from the $n = 0$ to $n = 1$ vibrational states. Energy differences for the band of peaks at the left and right are, respectively, $\Delta E_{l \to l+1} = \hbar\omega + 2(l+1)E_{0r} = \hbar\omega + 2E_{0r},\ \hbar\omega + 4E_{0r},\ \hbar\omega + 6E_{0r},\ \ldots\ \text{(right band)}$ and $\Delta E_{l \to l-1} = \hbar\omega - 2lE_{0r} = \hbar\omega - 2E_{0r},\ \hbar\omega - 4E_{0r},\ \hbar\omega - 6E_{0r},\ \ldots\ \text{(left band)}$.

The moment of inertia can then be determined from the energy spacing between individual peaks $(2E_{0r})$ or from the gap between the left and right bands $(4E_{0r})$. The frequency at the center of this gap is the frequency of vibration.

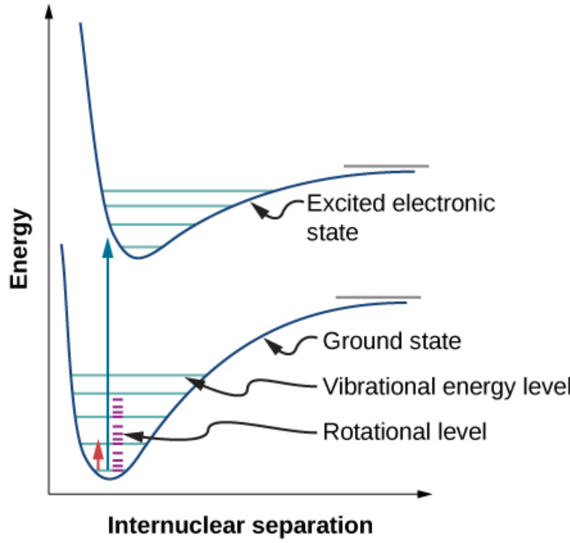


Figure 15.20 Three types of energy levels in a diatomic molecule: electronic, vibrational, and rotational. If the vibrational quantum number (*n*) changes by one unit, then the rotational quantum number (*l*) changes by one unit.

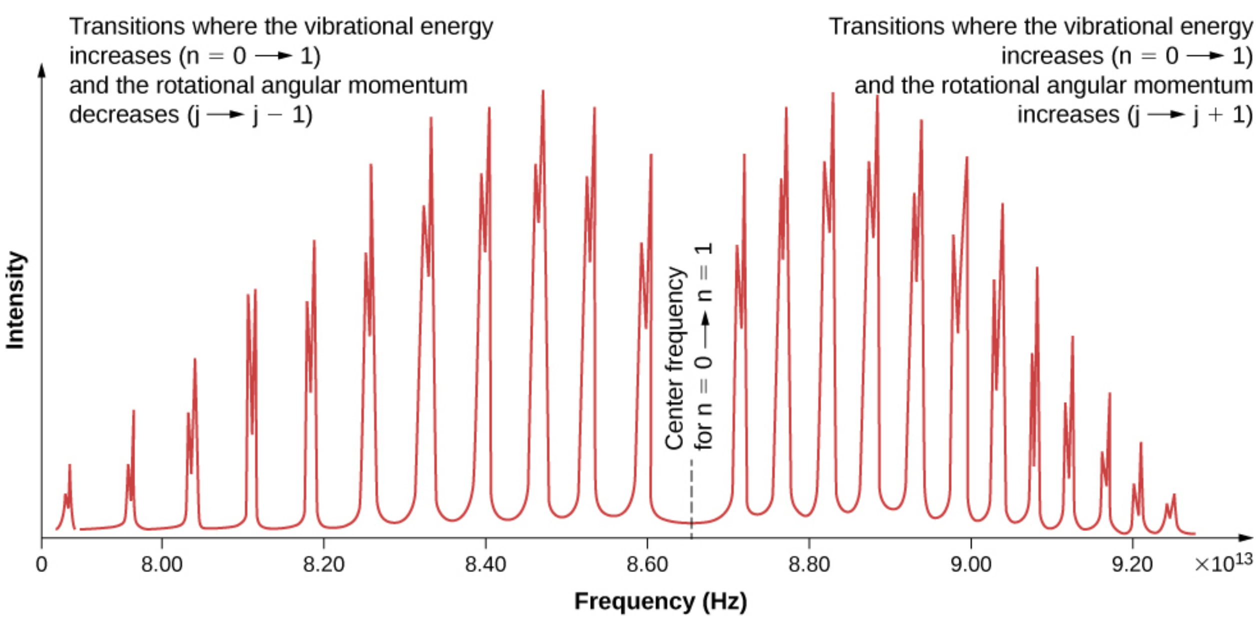


Figure 15.21 Absorption spectrum of hydrogen chloride (HCl) from the $n = 0$ to $n = 1$ vibrational levels. The discrete peaks indicate a quantization of the angular momentum of the molecule. The bands to the left indicate a decrease in angular momentum, whereas those to the right indicate an increase in angular momentum.

CHAPTER 15 REVIEW

KEY TERMS

absorber any object that absorbs radiation

absorption spectrum wavelengths of absorbed radiation by atoms and molecules

Balmer formula describes the emission spectrum of a hydrogen atom in the visible-light range

Balmer series spectral lines corresponding to electron transitions to/from the $n = 2$ state of the hydrogen atom, described by the Balmer formula

blackbody perfect absorber/emitter

blackbody radiation radiation emitted by a blackbody

Bohr radius of hydrogen radius of the first Bohr's orbit

Bohr's model of the hydrogen atom first quantum model to explain emission spectra of hydrogen

Brackett series spectral lines corresponding to electron transitions to/from the $n = 4$ state

braking radiation radiation produced by targeting metal with a high-energy electron beam (or radiation produced by the acceleration of any charged particle in a material)

electric dipole transition transition between energy levels brought by the absorption or emission of radiation

emission spectrum wavelengths of emitted radiation by atoms and molecules

emitter any object that emits radiation

energy spectrum of hydrogen set of allowed discrete energies of an electron in a hydrogen atom

excited energy states of the H atom energy state other than the ground state

Fraunhofer lines dark absorption lines in the continuum solar emission spectrum

ground state energy of the hydrogen atom energy of an electron in the first Bohr orbit of the hydrogen atom

Humphreys series spectral lines corresponding to electron transitions to/from the $n = 6$ state

hydrogen-like atom ionized atom with one electron remaining and nucleus with charge $+Ze$

ionization energy energy needed to remove an electron from an atom

ionization limit of the hydrogen atom ionization energy needed to remove an electron from the first Bohr orbit

Lyman series spectral lines corresponding to electron transitions to/from the ground state

Moseley plot plot of the atomic number versus the square root of X-ray frequency

Moseley's law relationship between the atomic number and X-ray photon frequency for X-ray production

nuclear model of the atom heavy positively charged nucleus at the center is surrounded by electrons, proposed by Rutherford

Paschen series spectral lines corresponding to electron transitions to/from the $n = 3$ state

Pfund series spectral lines corresponding to electron transitions to/from the $n = 5$ state

Planck's hypothesis of energy quanta energy exchanges between the radiation and the walls take place only in the form of discrete energy quanta

postulates of Bohr's model three assumptions that set a frame for Bohr's model

power intensity energy that passes through a unit surface per unit time

quantized energies discrete energies; not continuous

quantum number index that enumerates energy levels

quantum state of a Planck's oscillator any mode of vibration of Planck's oscillator, enumerated by quantum number

rotational energy level energy level associated with the rotational energy of a molecule

Rutherford's gold foil experiment first experiment to demonstrate the existence of the atomic nucleus

Rydberg constant for hydrogen physical constant in the Balmer formula

Rydberg formula experimentally found positions of spectral lines of hydrogen atom

selection rule rule that limits the possible transitions from one quantum state to another

selection rules rules that determine whether atomic transitions are allowed or forbidden (rare)

Stefan–Boltzmann constant physical constant in Stefan's law

vibrational energy level energy level associated with the vibrational energy of a molecule

α-particle doubly ionized helium atom

α-ray beam of α-particles (alpha-particles)

β-ray beam of electrons

γ-ray beam of highly energetic photons

KEY EQUATIONS

Wien's displacement law	$\lambda_{\max} T = 2.898 \times 10^{-3}\,\text{m} \cdot \text{K}$
Stefan's law	$P(T) = \sigma A T^4$
Planck's constant	$h = 6.626 \times 10^{-34}\,\text{J} \cdot \text{s} = 4.136 \times 10^{-15}\,\text{eV} \cdot \text{s}$
Energy quantum of radiation	$\Delta E = hf$
Planck's blackbody radiation law	$I(\lambda, T) = \frac{2\pi hc^2}{\lambda^5} \frac{1}{e^{hc/\lambda k_B T} - 1}$
Energy of a photon	$E_f = hf = \frac{hc}{\lambda}$
The Balmer formula	$\frac{1}{\lambda} = R_{\text{H}} \left(\frac{1}{2^2} - \frac{1}{n^2} \right)$
The Rydberg formula	$\frac{1}{\lambda} = R_{\text{H}} \left(\frac{1}{n_f^2} - \frac{1}{n_i^2} \right), n_i = n_f + 1, n_f + 2, \ \ldots$
Bohr's first quantization condition	$L_n = n\hbar, n = 1, 2, \ \ldots$
Bohr's second quantization condition	$hf = \lvert E_n - E_m \rvert$
Bohr's radius of hydrogen	$a_0 = 4\pi\varepsilon_0 \frac{\hbar^2}{m_e e^2} = 0.529\text{Å}$
Bohr's radius of the *n*th orbit	$r_n = a_0 n^2$
Ground-state energy value, ionization limit	$E_0 = \frac{1}{8\varepsilon_0^2} \frac{m_e e^4}{h^2} = 13.6\,\text{eV}$
Electron's energy in the *n*th orbit	$E_n = -E_0 \frac{1}{n^2}$

Ground state energy of hydrogen	$E_1 = -E_0 = -13.6\,\text{eV}$
The *n*th orbit of hydrogen-like ion	$r_n = \frac{a_0}{Z} n^2$
The *n*th energy of hydrogen-like ion	$E_n = -Z^2 E_0 \frac{1}{n^2}$

SUMMARY

15.1 Blackbody Radiation

- All bodies radiate energy. The amount of radiation a body emits depends on its temperature. The experimental Wien's displacement law states that the hotter the body, the shorter the wavelength corresponding to the emission peak in the radiation curve. The experimental Stefan's law states that the total power of radiation emitted across the entire spectrum of wavelengths at a given temperature is proportional to the fourth power of the Kelvin temperature of the radiating body.
- Absorption and emission of radiation are studied within the model of a blackbody. In the classical approach, the exchange of energy between radiation and cavity walls is continuous. The classical approach does not explain the blackbody radiation curve.
- To explain the blackbody radiation curve, Planck assumed that the exchange of energy between radiation and cavity walls takes place only in discrete quanta of energy. Planck's hypothesis of energy quanta led to the theoretical Planck's radiation law, which agrees with the experimental blackbody radiation curve; it also explains Wien's and Stefan's laws.

15.2 Bohr's Model of the Hydrogen Atom

- Positions of absorption and emission lines in the spectrum of atomic hydrogen are given by the experimental Rydberg formula. Classical physics cannot explain the spectrum of atomic hydrogen.
- The Bohr model of hydrogen was the first model of atomic structure to correctly explain the radiation spectra of atomic hydrogen. It was preceded by the Rutherford nuclear model of the atom. In Rutherford's model, an atom consists of a positively charged point-like nucleus that contains almost the entire mass of the atom and of negative electrons that are located far away from the nucleus.
- Bohr's model of the hydrogen atom is based on three postulates: (1) an electron moves around the nucleus in a circular orbit, (2) an electron's angular momentum in the orbit is quantized, and (3) the change in an electron's energy as it makes a quantum jump from one orbit to another is always accompanied by the emission or absorption of a photon. Bohr's model is semi-classical because it combines the classical concept of electron orbit (postulate 1) with the new concept of quantization (postulates 2 and 3).
- Bohr's model of the hydrogen atom explains the emission and absorption spectra of atomic hydrogen and hydrogen-like ions with low atomic numbers. It was the first model to introduce the concept of a quantum number to describe atomic states and to postulate quantization of electron orbits in the atom. Bohr's model is an important step in the development of quantum mechanics, which deals with many-electron atoms.

15.3 Atomic Spectra and X-rays

- Radiation is absorbed and emitted by atomic energy-level transitions.
- Quantum numbers can be used to estimate the energy, frequency, and wavelength of photons produced by atomic transitions.
- X-ray photons are produced when a vacancy in an inner shell of an atom is filled by an electron from the outer shell of the atom.
- The frequency of X-ray radiation is related to the atomic number Z of an atom.

15.4 Molecular Spectra

- Molecules possess vibrational and rotational energy.

- Energy differences between adjacent vibrational energy levels are larger than those between rotational energy levels.
- Separation between peaks in an absorption spectrum is inversely related to the moment of inertia.
- Transitions between vibrational and rotational energy levels follow selection rules.

CONCEPTUAL QUESTIONS

15.1 Blackbody Radiation

1. Which surface has a higher temperature – the surface of a yellow star or that of a red star?

2. Describe what you would see when looking at a body whose temperature is increased from 1000 K to 1,000,000 K.

3. Explain the color changes in a hot body as its temperature is increased.

4. Speculate as to why UV light causes sunburn, whereas visible light does not.

5. Two cavity radiators are constructed with walls made of different metals. At the same temperature, how would their radiation spectra differ?

6. Discuss why some bodies appear black, other bodies appear red, and still other bodies appear white.

7. If everything radiates electromagnetic energy, why can we not see objects at room temperature in a dark room?

8. How much does the power radiated by a blackbody increase when its temperature (in K) is tripled?

15.2 Bohr's Model of the Hydrogen Atom

9. Explain why the patterns of bright emission spectral lines have an identical spectral position to the pattern of dark absorption spectral lines for a given gaseous element.

10. Do the various spectral lines of the hydrogen atom overlap?

11. The Balmer series for hydrogen was discovered before either the Lyman or the Paschen series. Why?

12. When the absorption spectrum of hydrogen at room temperature is analyzed, absorption lines for the Lyman series are found, but none are found for the Balmer series. What does this tell us about the energy state of most hydrogen atoms at room temperature?

13. Hydrogen accounts for about 75% by mass of the matter at the surfaces of most stars. However, the absorption lines of hydrogen are strongest (of highest intensity) in the spectra of stars with a surface temperature of about 9000 K. They are weaker in the sun spectrum and are essentially nonexistent in very hot (temperatures above 25,000 K) or rather cool (temperatures below 3500 K) stars. Speculate as to why surface temperature affects the hydrogen absorption lines that we observe.

14. Discuss the similarities and differences between Thomson's model of the hydrogen atom and Bohr's model of the hydrogen atom.

15. Discuss the way in which Thomson's model is nonphysical. Support your argument with experimental evidence.

16. If, in a hydrogen atom, an electron moves to an orbit with a larger radius, does the energy of the hydrogen atom increase or decrease?

17. How is the energy conserved when an atom makes a transition from a higher to a lower energy state?

18. Suppose an electron in a hydrogen atom makes a transition from the $(n+1)$th orbit to the nth orbit. Is the wavelength of the emitted photon longer for larger values of n, or for smaller values of n?

19. Discuss why the allowed energies of the hydrogen atom are negative.

20. Can a hydrogen atom absorb a photon whose energy is greater than 13.6 eV?

21. Why can you see through glass but not through wood?

22. Do gravitational forces have a significant effect on atomic energy levels?

23. Show that Planck's constant has the dimensions of angular momentum.

15.3 Atomic Spectra and X-rays

24. Atomic and molecular spectra are discrete. What does discrete mean, and how are discrete spectra related to the quantization of energy and electron orbits in atoms and molecules?

25. Discuss the process of the absorption of light by matter in terms of the atomic structure of the absorbing medium.

26. NGC1763 is an emission nebula in the Large Magellanic Cloud just outside our Milky Way Galaxy. Ultraviolet light from hot stars ionize the hydrogen atoms in the nebula. As protons and electrons recombine, light in the visible range is emitted. Compare the energies of the photons involved in these two transitions.

27. Why are X-rays emitted only for electron transitions to inner shells? What type of photon is emitted for transitions between outer shells?

28. How do the allowed orbits for electrons in atoms differ from the allowed orbits for planets around the sun?

15.4 Molecular Spectra

29. Does the absorption spectrum of the diatomic molecule HCl depend on the isotope of chlorine contained in the molecule? Explain your reasoning.

30. Rank the energy spacing (ΔE) of the following transitions from least to greatest: an electron energy transition in an atom (atomic energy), the rotational energy of a molecule, or the vibrational energy of a molecule?

31. Explain key features of a vibrational-rotation energy spectrum of the diatomic molecule.

PROBLEMS

15.1 Blackbody Radiation

32. A 200-W heater emits a 1.5-µm radiation. (a) What value of the energy quantum does it emit? (b) Assuming that the specific heat of a 4.0-kg body is $0.83\text{kcal/kg}\cdot\text{K}$, how many of these photons must be absorbed by the body to increase its temperature by 2 K? (c) How long does the heating process in (b) take, assuming that all radiation emitted by the heater gets absorbed by the body?

33. A 900-W microwave generator in an oven generates energy quanta of frequency 2560 MHz. (a) How many energy quanta does it emit per second? (b) How many energy quanta must be absorbed by a pasta dish placed in the radiation cavity to increase its temperature by 45.0 K? Assume that the dish has a mass of 0.5 kg and that its specific heat is $0.9\ \text{kcal/kg}\cdot\text{K}$. (c) Assume that all energy quanta emitted by the generator are absorbed by the pasta dish. How long must we wait until the dish in (b) is ready?

34. (a) For what temperature is the peak of blackbody radiation spectrum at 400 nm? (b) If the temperature of a blackbody is 800 K, at what wavelength does it radiate the most energy?

35. The tungsten elements of incandescent light bulbs operate at 3200 K. At what frequency does the filament radiate maximum energy?

36. Interstellar space is filled with radiation of wavelength $970\mu\text{m}$. This radiation is considered to be a remnant of the "big bang." What is the corresponding blackbody temperature of this radiation?

37. The radiant energy from the sun reaches its maximum at a wavelength of about 500.0 nm. What is the approximate temperature of the sun's surface?

15.2 Bohr's Model of the Hydrogen Atom

38. Calculate the wavelength of the first line in the Lyman series and show that this line lies in the ultraviolet part of the spectrum.

39. Calculate the wavelength of the fifth line in the Lyman series and show that this line lies in the ultraviolet part of the spectrum.

40. Calculate the energy changes corresponding to the transitions of the hydrogen atom: (a) from $n = 3$ to $n = 4$; (b) from $n = 2$ to $n = 1$; and (c) from $n = 3$ to $n = \infty$.

41. Determine the wavelength of the third Balmer line (transition from $n = 5$ to $n = 2$).

42. What is the frequency of the photon absorbed when the hydrogen atom makes the transition from the ground state to the $n = 4$ state?

43. When a hydrogen atom is in its ground state, what are the shortest and longest wavelengths of the photons it can absorb without being ionized?

44. When a hydrogen atom is in its third excided state, what are the shortest and longest wavelengths of the photons it can emit?

45. What is the longest wavelength that light can have if it is to be capable of ionizing the hydrogen atom in its ground state?

46. For an electron in a hydrogen atom in the $n = 2$ state, compute: (a) the angular momentum; (b) the kinetic energy; (c) the potential energy; and (d) the total energy.

47. Find the ionization energy of a hydrogen atom in the fourth energy state.

48. It has been measured that it required 0.850 eV to remove an electron from the hydrogen atom. In what state was the atom before the ionization happened?

49. What is the radius of a hydrogen atom when the electron is in the first excited state?

50. Find the shortest wavelength in the Balmer series. In what part of the spectrum does this line lie?

51. Show that the entire Paschen series lies in the infrared part of the spectrum.

52. Do the Balmer series and the Lyman series overlap? Why? Why not? (Hint: calculate the shortest Balmer line and the longest Lyman line.)

53. (a) Which line in the Balmer series is the first one in the UV part of the spectrum? (b) How many Balmer lines lie in the visible part of the spectrum? (c) How many Balmer lines lie in the UV?

54. A $4.653\text{-}\mu\text{m}$ emission line of atomic hydrogen corresponds to transition between the states $n_f = 5$ and n_i. Find n_i.

15.3 Atomic Spectra and X-rays

55. What is the minimum frequency of a photon required to ionize: (a) a He^+ ion in its ground state? (b) A Li^{2+} ion in its first excited state?

56. The ion Li^{2+} makes an atomic transition from an $n = 4$ state to an $n = 2$ state. (a) What is the energy of the photon emitted during the transition? (b) What is the wavelength of the photon?

57. The red light emitted by a ruby laser has a wavelength of 694.3 nm. What is the difference in energy between the initial state and final state corresponding to the emission of the light?

58. The yellow light from a sodium-vapor street lamp is produced by a transition of sodium atoms from a 3*p* state to a *3s* state. If the difference in energies of those two states is 2.10 eV, what is the wavelength of the yellow light?

59. Estimate the wavelength of the K_α X-ray from calcium.

60. Estimate the frequency of the K_α X-ray from cesium.

61. X-rays are produced by striking a target with a beam of electrons. Prior to striking the target, the electrons are accelerated by an electric field through a potential energy difference:

$$\Delta U = -e\Delta V,$$

where *e* is the charge of an electron and ΔV is the voltage difference. If $\Delta V = 15{,}000$ volts, what is the minimum wavelength of the emitted radiation?

62. For the preceding problem, what happens to the minimum wavelength if the voltage across the X-ray tube is doubled?

63. Suppose the experiment in the preceding problem is conducted with muons. What happens to the minimum wavelength?

64. An X-ray tube accelerates an electron with an applied voltage of 50 kV toward a metal target. (a) What is the shortest-wavelength X-ray radiation generated at the target? (b) Calculate the photon energy in eV. (c) Explain the relationship of the photon energy to the applied voltage.

65. A color television tube generates some X-rays when its electron beam strikes the screen. What is the shortest wavelength of these X-rays, if a 30.0-kV potential is used to accelerate the electrons? (Note that TVs have shielding to prevent these X-rays from exposing viewers.)

66. An X-ray tube has an applied voltage of 100 kV. (a) What is the most energetic X-ray photon it can produce? Express your answer in electron volts and joules. (b) Find the wavelength of such an X-ray.

67. The maximum characteristic X-ray photon energy comes from the capture of a free electron into a *K* shell vacancy. What is this photon energy in keV for tungsten, assuming that the free electron has no initial kinetic energy?

68. What are the approximate energies of the K_α and K_β X-rays for copper?

69. Compare the X-ray photon wavelengths for copper and gold.

70. The approximate energies of the K_α and K_β X-rays for copper are $E_{K_\alpha} = 8.00\,\text{keV}$ and $E_{K_\beta} = 9.48\,\text{keV}$, respectively. Determine the ratio of X-ray frequencies of gold to copper, then use this value to estimate the corresponding energies of K_α and K_β X-rays for gold.

15.4 Molecular Spectra

71. In a physics lab, you measure the vibrational-rotational spectrum of HCl. The estimated separation between absorption peaks is $\Delta f \approx 5.5 \times 10^{11}\ \text{Hz}$. The central frequency of the band is $f_0 = 9.0 \times 10^{13}\ \text{Hz}$. (a) What is the moment of inertia (*I*)? (b) What is the energy of vibration for the molecule?

72. For the preceding problem, find the equilibrium separation of the H and Cl atoms. Compare this with the actual value.

73. The separation between oxygen atoms in an O_2 molecule is about 0.121 nm. Determine the characteristic energy of rotation in eV.

74. The characteristic energy of the N_2 molecule is $2.48 \times 10^{-4}\ \text{eV}$. Determine the separation distance between the nitrogen atoms

75. The characteristic energy for KCl is 1.4×10^{-5} eV. (a) Determine μ for the KCl molecule. (b) Find the separation distance between the K and Cl atoms.

76. A diatomic F_2 molecule is in the $l = 1$ state. (a) What is the energy of the molecule? (b) How much energy is radiated in a transition from a $l = 2$ to a $l = 1$ state?

77. In a physics lab, you measure the vibrational-rotational spectrum of potassium bromide (KBr). The estimated separation between absorption peaks is $\Delta f \approx 5.35 \times 10^{10}\ \text{Hz}$. The central frequency of the band is $f_0 = 8.75 \times 10^{12}\ \text{Hz}$. (a) What is the moment of inertia (*I*)? (b) What is the energy of vibration for the molecule?

16 | INTRODUCTORY THERMODYNAMICS

Figure 16.1 These snowshoers on Mount Hood in Oregon are enjoying the heat flow and light caused by high temperature. All three mechanisms of heat transfer are relevant to this picture. The heat flowing out of the fire also turns the solid snow to liquid water and vapor. (credit: modification of work by "Mt. Hood Territory"/Flickr)

Chapter Outline

Introduction

Heat and temperature are important concepts for each of us, every day. How we dress in the morning depends on whether the day is hot or cold, and most of what we do requires energy that ultimately comes from the Sun. The study of heat and temperature is part of an area of physics known as **thermodynamics**. The laws of thermodynamics govern the flow of energy throughout the universe. They are studied in all areas of science and engineering, and are essential to understanding our solar system, the stars and galaxies.

In this chapter, we explore heat and temperature. It is not always easy to distinguish these terms. Heat is the flow of energy from one object to another. This flow of energy is caused by a difference in temperature. The transfer of heat can change temperature, as can work, another kind of energy transfer that is central to thermodynamics. We return to these basic ideas several times throughout the next four chapters, and you will see that they affect everything from the behavior of atoms and molecules to cooking to our weather on Earth to the life cycles of stars.

16.1 | Temperature and Thermal Equilibrium

Learning Objectives

By the end of this section, you will be able to:

- Define temperature and describe it qualitatively
- Explain thermal equilibrium
- Explain the zeroth law of thermodynamics

Heat is familiar to all of us. We can feel heat entering our bodies from the summer Sun or from hot coffee or tea after a winter stroll. We can also feel heat leaving our bodies as we feel the chill of night or the cooling effect of sweat after exercise.

What is heat? How do we define it and how is it related to temperature? What are the effects of heat and how does it flow from place to place? We will find that, in spite of the richness of the phenomena, a small set of underlying physical principles unites these subjects and ties them to other fields. We start by examining temperature and how to define and measure it.

Temperature

The concept of temperature has evolved from the common concepts of hot and cold. The scientific definition of temperature explains more than our senses of hot and cold. As you may have already learned, many physical quantities are defined solely in terms of how they are observed or measured, that is, they are defined *operationally*. **Temperature** is operationally defined as the quantity of what we measure with a thermometer. As we will see in detail in a later chapter on the kinetic theory of gases, temperature is proportional to the average kinetic energy of translation, a fact that provides a more physical definition. Differences in temperature maintain the transfer of heat, or *heat transfe*r, throughout the universe. **Heat transfer** is the movement of energy from one place or material to another as a result of a difference in temperature. (You will learn more about heat transfer later in this chapter.)

Thermal Equilibrium

An important concept related to temperature is **thermal equilibrium**. Two objects are in thermal equilibrium if they are in close contact that allows either to gain energy from the other, but nevertheless, no net energy is transferred between them. Even when not in contact, they are in thermal equilibrium if, when they are placed in contact, no net energy is transferred between them. If two objects remain in contact for a long time, they typically come to equilibrium. In other words, two objects in thermal equilibrium do not exchange energy.

Experimentally, if object *A* is in equilibrium with object *B*, and object *B* is in equilibrium with object *C*, then (as you may have already guessed) object *A* is in equilibrium with object *C*. That statement of transitivity is called the **zeroth law of thermodynamics**. (The number "zeroth" was suggested by British physicist Ralph Fowler in the 1930s. The first, second, and third laws of thermodynamics were already named and numbered then. The zeroth law had seldom been stated, but it needs to be discussed before the others, so Fowler gave it a smaller number.) Consider the case where *A* is a thermometer. The zeroth law tells us that if *A* reads a certain temperature when in equilibrium with *B*, and it is then placed in contact with *C*, it will not exchange energy with *C*; therefore, its temperature reading will remain the same (**Figure 16.2**). In other words, *if two objects are in thermal equilibrium, they have the same temperature*.

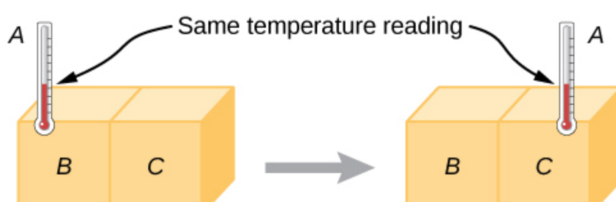


Figure 16.2 If thermometer *A* is in thermal equilibrium with object *B*, and *B* is in thermal equilibrium with *C*, then *A* is in thermal equilibrium with *C*. Therefore, the reading on *A* stays the same when *A* is moved over to make contact with *C*.

A thermometer measures its own temperature. It is through the concepts of thermal equilibrium and the zeroth law of thermodynamics that we can say that a thermometer measures the temperature of *something else,* and to make sense of the statement that two objects are at the same temperature.

In the rest of this chapter, we will often refer to "systems" instead of "objects." As in the chapter on linear momentum and collisions, a system consists of one or more objects—but in thermodynamics, we require a system to be macroscopic, that is, to consist of a huge number (such as 10^{23}) of molecules. Then we can say that a system is in thermal equilibrium with itself if all parts of it are at the same temperature. (We will return to the definition of a thermodynamic system in the chapter on the first law of thermodynamics.)

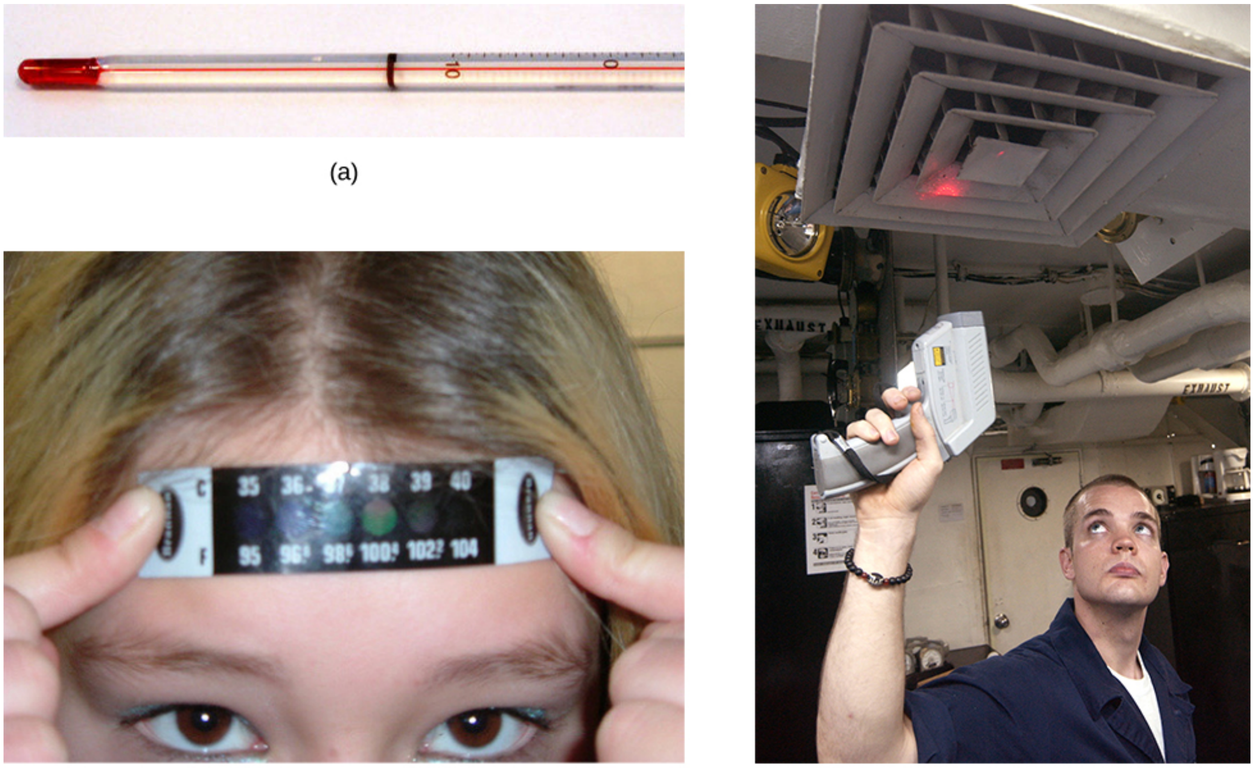
16.2 | Thermometers and Temperature Scales

Learning Objectives

By the end of this section, you will be able to:

- Describe several different types of thermometers
- Convert temperatures between the Celsius, Fahrenheit, and Kelvin scales

Any physical property that depends consistently and reproducibly on temperature can be used as the basis of a thermometer. For example, volume increases with temperature for most substances. This property is the basis for the common alcohol thermometer and the original mercury thermometers. Other properties used to measure temperature include electrical resistance, color, and the emission of infrared radiation (**Figure 16.3**).



(a) (b) (c)

Figure 16.3 Because many physical properties depend on temperature, the variety of thermometers is remarkable. (a) In this common type of thermometer, the alcohol, containing a red dye, expands more rapidly than the glass encasing it. When the thermometer's temperature increases, the liquid from the bulb is forced into the narrow tube, producing a large change in the length of the column for a small change in temperature. (b) Each of the six squares on this plastic (liquid crystal) thermometer contains a film of a different heat-sensitive liquid crystal material. Below $95\ °\text{F}$, all six squares are black. When the plastic thermometer is exposed to a temperature of $95\ °\text{F}$, the first liquid crystal square changes color. When the temperature reaches above $96.8\ °\text{F}$, the second liquid crystal square also changes color, and so forth. (c) A firefighter uses a pyrometer to check the temperature of an aircraft carrier's ventilation system. The pyrometer measures infrared radiation (whose emission varies with temperature) from the vent and quickly produces a temperature readout. Infrared thermometers are also frequently used to measure body temperature by gently placing them in the ear canal. Such thermometers are more accurate than the alcohol thermometers placed under the tongue or in the armpit. (credit b: modification of work by Tess Watson; credit c: modification of work by Lamel J. Hinton, U.S. Navy)

Thermometers measure temperature according to well-defined scales of measurement. The three most common temperature scales are Fahrenheit, Celsius, and Kelvin. Temperature scales are created by identifying two reproducible temperatures. The freezing and boiling temperatures of water at standard atmospheric pressure are commonly used.

On the **Celsius scale**, the freezing point of water is $0\ °\text{C}$ and the boiling point is $100\ °\text{C}.$ The unit of temperature on this scale is the **degree Celsius** $(°\text{C})$. The **Fahrenheit scale** (still the most frequently used for common purposes in the United States) has the freezing point of water at $32\ °\text{F}$ and the boiling point at $212\ °\text{F}.$ Its unit is the **degree Fahrenheit** $(°\text{F})$. You can see that 100 Celsius degrees span the same range as 180 Fahrenheit degrees. Thus, a temperature difference of one degree on the Celsius scale is 1.8 times as large as a difference of one degree on the Fahrenheit scale, or $\Delta T_\text{F} = \frac{9}{5} \Delta T_\text{C}.$

The definition of temperature in terms of molecular motion suggests that there should be a lowest possible temperature, where the average kinetic energy of molecules is zero (or the minimum allowed by quantum mechanics). Experiments confirm the existence of such a temperature, called **absolute zero**. An **absolute temperature scale** is one whose zero point is absolute zero. Such scales are convenient in science because several physical quantities, such as the volume of an ideal gas, are directly related to absolute temperature.

The **Kelvin scale** is the absolute temperature scale that is commonly used in science. The SI temperature unit is the *kelvin*, which is abbreviated K (not accompanied by a degree sign). Thus 0 K is absolute zero. The freezing and boiling points of water are 273.15 K and 373.15 K, respectively. Therefore, temperature differences are the same in units of kelvins and degrees Celsius, or $\Delta T_C = \Delta T_K$.

The relationships between the three common temperature scales are shown in **Figure 16.4**. Temperatures on these scales can be converted using the equations in **Table 16.1**.

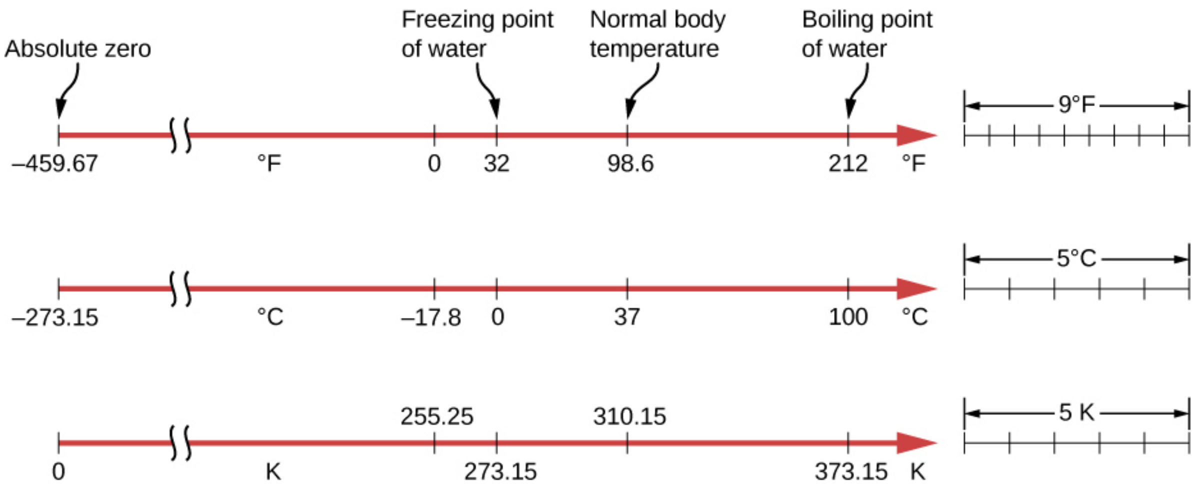


Figure 16.4 Relationships between the Fahrenheit, Celsius, and Kelvin temperature scales are shown. The relative sizes of the scales are also shown.

To convert from…	Use this equation…
Celsius to Fahrenheit	$T_F = \frac{9}{5}T_C + 32$
Fahrenheit to Celsius	$T_C = \frac{5}{9}(T_F - 32)$
Celsius to Kelvin	$T_K = T_C + 273.15$
Kelvin to Celsius	$T_C = T_K - 273.15$
Fahrenheit to Kelvin	$T_K = \frac{5}{9}(T_F - 32) + 273.15$
Kelvin to Fahrenheit	$T_F = \frac{9}{5}(T_K - 273.15) + 32$

Table 16.1 Temperature Conversions

To convert between Fahrenheit and Kelvin, convert to Celsius as an intermediate step.

Example 16.1

Converting between Temperature Scales: Room Temperature

“Room temperature” is generally defined in physics to be $25\,°C$. (a) What is room temperature in $°F$? (b) What is it in K?

Strategy

To answer these questions, all we need to do is choose the correct conversion equations and substitute the known values.

Solution

To convert from $°C$ to $°F$, use the equation

$$T_F = \frac{9}{5}T_C + 32.$$

Substitute the known value into the equation and solve:

$$T_F = \frac{9}{5}(25\ °C) + 32 = 77\ °F.$$

Similarly, we find that $T_K = T_C + 273.15 = 298\ K$.

The Kelvin scale is part of the SI system of units, so its actual definition is more complicated than the one given above. First, it is not defined in terms of the freezing and boiling points of water, but in terms of the **triple point**. The triple point is the unique combination of temperature and pressure at which ice, liquid water, and water vapor can coexist stably. As will be discussed in the section on phase changes, the coexistence is achieved by lowering the pressure and consequently the boiling point to reach the freezing point. The triple-point temperature is defined as 273.16 K. This definition has the advantage that although the freezing temperature and boiling temperature of water depend on pressure, there is only one triple-point temperature.

Second, even with two points on the scale defined, different thermometers give somewhat different results for other temperatures. Therefore, a standard thermometer is required. Metrologists (experts in the science of measurement) have chosen the *constant-volume gas thermometer* for this purpose. A vessel of constant volume filled with gas is subjected to temperature changes, and the measured temperature is proportional to the change in pressure. Using "TP" to represent the triple point,

$$T = \frac{p}{p_{TP}}T_{TP}.$$

The results depend somewhat on the choice of gas, but the less dense the gas in the bulb, the better the results for different gases agree. If the results are extrapolated to zero density, the results agree quite well, with zero pressure corresponding to a temperature of absolute zero.

Constant-volume gas thermometers are big and come to equilibrium slowly, so they are used mostly as standards to calibrate other thermometers.

Visit this **site (https://openstax.org/l/21consvolgasth)** to learn more about the constant-volume gas thermometer.

16.3 | Thermal Expansion

Learning Objectives

By the end of this section, you will be able to:

- Answer qualitative questions about the effects of thermal expansion
- Solve problems involving thermal expansion, including those involving thermal stress

The expansion of alcohol in a thermometer is one of many commonly encountered examples of **thermal expansion**, which is the change in size or volume of a given system as its temperature changes. The most visible example is the expansion of hot air. When air is heated, it expands and becomes less dense than the surrounding air, which then exerts an (upward) force on the hot air and makes steam and smoke rise, hot air balloons float, and so forth. The same behavior happens in all liquids and gases, driving natural heat transfer upward in homes, oceans, and weather systems, as we will discuss in an upcoming section. Solids also undergo thermal expansion. Railroad tracks and bridges, for example, have expansion joints to allow them to freely expand and contract with temperature changes, as shown in **Figure 16.5**.

(a) (b)

Figure 16.5 (a) Thermal expansion joints like these in the (b) Auckland Harbour Bridge in New Zealand allow bridges to change length without buckling. (credit: modification of works by “ŠJů”/Wikimedia Commons)

What is the underlying cause of thermal expansion? As previously mentioned, an increase in temperature means an increase in the kinetic energy of individual atoms. In a solid, unlike in a gas, the molecules are held in place by forces from neighboring molecules; the forces can be modeled as springs connecting the atoms together. The resulting potential energy increases more steeply when the molecules get closer to each other than when they get farther away. Thus, at a given kinetic energy, the distance moved is greater when neighbors move away from each other than when they move toward each other. The result is that increased kinetic energy (increased temperature) increases the average distance between molecules—the substance expands.

For most substances under ordinary conditions, it is an excellent approximation that there is no preferred direction (that is, the solid is “isotropic”), and an increase in temperature increases the solid’s size by a certain fraction in each dimension. Therefore, if the solid is free to expand or contract, its proportions stay the same; only its overall size changes.

Linear Thermal Expansion

According to experiments, the dependence of thermal expansion on temperature, substance, and original length is summarized in the equation

$$\frac{dL}{dT} = \alpha L \qquad (16.1)$$

where ΔL is the change in length L, ΔT is the change in temperature, and α is the **coefficient of linear expansion**, a material property that varies slightly with temperature. As α is nearly constant and also very small, for practical purposes, we use the linear approximation:

$$\Delta L = \alpha L \Delta T. \qquad (16.2)$$

Table 16.2 lists representative values of the coefficient of linear expansion. As noted earlier, ΔT is the same whether it is expressed in units of degrees Celsius or kelvins; thus, α may have units of $1/°\text{C}$ or 1/K with the same value in either case. Approximating α as a constant is quite accurate for small changes in temperature and sufficient for most practical purposes, even for large changes in temperature. We examine this approximation more closely in the next example.

Material	Coefficient of Linear Expansion $\alpha(1/°C)$	Coefficient of Volume Expansion $\beta(1/°C)$
Solids		
Aluminum	25×10^{-6}	75×10^{-6}
Brass	19×10^{-6}	56×10^{-6}
Copper	17×10^{-6}	51×10^{-6}
Gold	14×10^{-6}	42×10^{-6}
Iron or steel	12×10^{-6}	35×10^{-6}
Invar (nickel-iron alloy)	0.9×10^{-6}	2.7×10^{-6}
Lead	29×10^{-6}	87×10^{-6}
Silver	18×10^{-6}	54×10^{-6}
Glass (ordinary)	9×10^{-6}	27×10^{-6}
Glass (Pyrex®)	3×10^{-6}	9×10^{-6}
Quartz	0.4×10^{-6}	1×10^{-6}
Concrete, brick	$\sim 12 \times 10^{-6}$	$\sim 36 \times 10^{-6}$
Marble (average)	2.5×10^{-6}	7.5×10^{-6}
Liquids		
Ether		1650×10^{-6}
Ethyl alcohol		1100×10^{-6}
Gasoline		950×10^{-6}
Glycerin		500×10^{-6}
Mercury		180×10^{-6}
Water		210×10^{-6}
Gases		
Air and most other gases at atmospheric pressure		3400×10^{-6}

Table 16.2 Thermal Expansion Coefficients

Thermal expansion is exploited in the bimetallic strip (**Figure 16.6**). This device can be used as a thermometer if the curving strip is attached to a pointer on a scale. It can also be used to automatically close or open a switch at a certain temperature, as in older or analog thermostats.

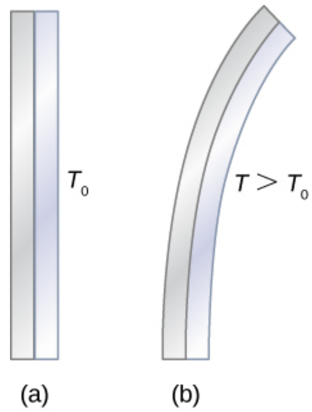


Figure 16.6 The curvature of a bimetallic strip depends on temperature. (a) The strip is straight at the starting temperature, where its two components have the same length. (b) At a higher temperature, this strip bends to the right, because the metal on the left has expanded more than the metal on the right. At a lower temperature, the strip would bend to the left.

Example 16.2

Calculating Linear Thermal Expansion

The main span of San Francisco's Golden Gate Bridge is 1275 m long at its coldest. The bridge is exposed to temperatures ranging from $-15\ °C$ to $40\ °C$. What is its change in length between these temperatures? Assume that the bridge is made entirely of steel.

Strategy

Use the equation for linear thermal expansion $\Delta L = \alpha L \Delta T$ to calculate the change in length, ΔL. Use the coefficient of linear expansion α for steel from **Table 16.2**, and note that the change in temperature ΔT is $55\ °C$.

Solution

Substitute all of the known values into the equation to solve for ΔL:

$$\Delta L = \alpha L \Delta T = \left(\frac{12 \times 10^{-6}}{°C}\right)(1275\ \text{m})(55\ °C) = 0.84\ \text{m}.$$

Significance

Although not large compared with the length of the bridge, this change in length is observable. It is generally spread over many expansion joints so that the expansion at each joint is small.

Thermal Expansion in Two and Three Dimensions

Unconstrained objects expand in all dimensions, as illustrated in **Figure 16.7**. That is, their areas and volumes, as well as their lengths, increase with temperature. Because the proportions stay the same, holes and container volumes also get larger with temperature. If you cut a hole in a metal plate, the remaining material will expand exactly as it would if the piece you removed were still in place. The piece would get bigger, so the hole must get bigger too.

Thermal Expansion in Two Dimensions

For small temperature changes, the change in area ΔA is given by

$$\Delta A = 2\alpha A \Delta T \qquad (16.3)$$

where ΔA is the change in area A, ΔT is the change in temperature, and α is the coefficient of linear expansion, which varies slightly with temperature. (The derivation of this equation is analogous to that of the more important equation for three dimensions, below.)

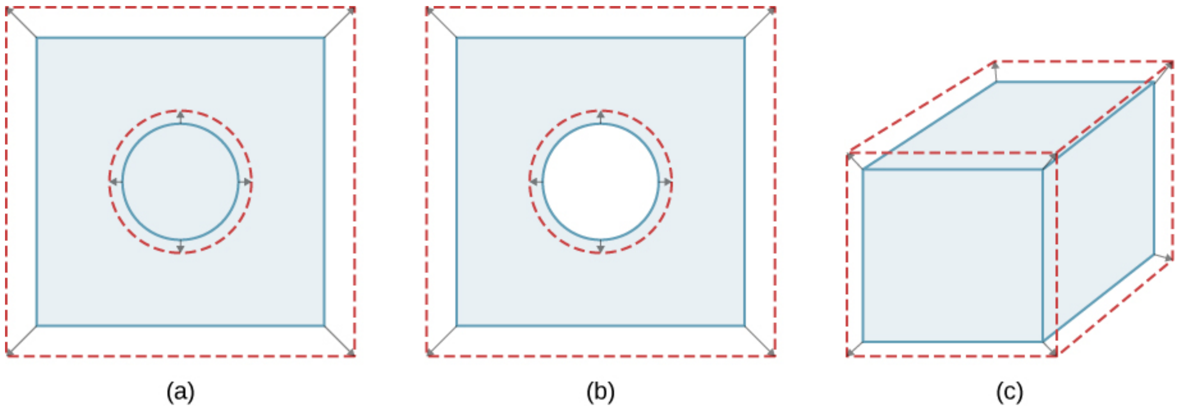


Figure 16.7 In general, objects expand in all directions as temperature increases. In these drawings, the original boundaries of the objects are shown with solid lines, and the expanded boundaries with dashed lines. (a) Area increases because both length and width increase. The area of a circular plug also increases. (b) If the plug is removed, the hole it leaves becomes larger with increasing temperature, just as if the expanding plug were still in place. (c) Volume also increases, because all three dimensions increase.

Thermal Expansion in Three Dimensions

The relationship between volume and temperature $\frac{dV}{dT}$ is given by $\frac{dV}{dT} = \beta V \Delta T$, where β is the **coefficient of volume expansion**. As you can show in **Exercise 16.57**, $\beta = 3\alpha$. This equation is usually written as

$$\Delta V = \beta V \Delta T. \qquad (16.4)$$

Note that the values of β in **Table 16.2** are equal to 3α except for rounding.

Volume expansion is defined for liquids, but linear and area expansion are not, as a liquid's changes in linear dimensions and area depend on the shape of its container. Thus, **Table 16.2** shows liquids' values of β but not α.

In general, objects expand with increasing temperature. Water is the most important exception to this rule. Water does expand with increasing temperature (its density *decreases*) at temperatures greater than $4\,°\text{C}\,(40\,°\text{F})$. However, it is densest at $+4\,°\text{C}$ and expands with *decreasing* temperature between $+4\,°\text{C}$ and $0\,°\text{C}$ ($40\,°\text{F}$ to $32\,°\text{F}$), as shown in **Figure 16.8**. A striking effect of this phenomenon is the freezing of water in a pond. When water near the surface cools down to $4\,°\text{C}$, it is denser than the remaining water and thus sinks to the bottom. This "turnover" leaves a layer of warmer water near the surface, which is then cooled. However, if the temperature in the surface layer drops below $4\,°\text{C}$, that water is less dense than the water below, and thus stays near the top. As a result, the pond surface can freeze over. The layer of ice insulates the liquid water below it from low air temperatures. Fish and other aquatic life can survive in $4\,°\text{C}$ water beneath ice, due to this unusual characteristic of water.

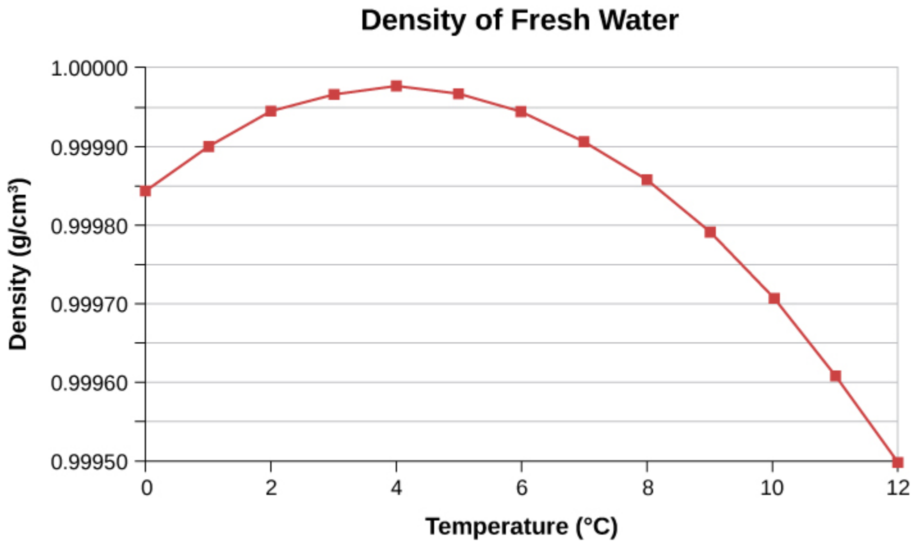


Figure 16.8 This curve shows the density of water as a function of temperature. Note that the thermal expansion at low temperatures is very small. The maximum density at $4\ °C$ is only 0.0075% greater than the density at $2\ °C$, and 0.012% greater than that at $0\ °C$. The decrease of density below $4\ °C$ occurs because the liquid water approachs the solid crystal form of ice, which contains more empty space than the liquid.

Example 16.3

Calculating Thermal Expansion

Suppose your 60.0-L (15.9 -gal -gal) steel gasoline tank is full of gas that is cool because it has just been pumped from an underground reservoir. Now, both the tank and the gasoline have a temperature of $15.0\ °C.$ How much gasoline has spilled by the time they warm to $35.0\ °C$?

Strategy

The tank and gasoline increase in volume, but the gasoline increases more, so the amount spilled is the difference in their volume changes. We can use the equation for volume expansion to calculate the change in volume of the gasoline and of the tank. (The gasoline tank can be treated as solid steel.)

Solution

1. Use the equation for volume expansion to calculate the increase in volume of the steel tank:
$$\Delta V_s = \beta_s V_s \Delta T.$$
2. The increase in volume of the gasoline is given by this equation:
$$\Delta V_{gas} = \beta_{gas} V_{gas} \Delta T.$$
3. Find the difference in volume to determine the amount spilled as
$$V_{spill} = \Delta V_{gas} - \Delta V_s.$$

Alternatively, we can combine these three equations into a single equation. (Note that the original volumes are equal.)

$$\begin{aligned} V_{spill} &= (\beta_{ga\ s} - \beta_s)V\Delta T \\ &= \left[(950 - 35) \times 10^{-6}/°C\right](60.0\ L)(20.0\ °C) \\ &= 1.10\ L. \end{aligned}$$

Significance

This amount is significant, particularly for a 60.0-L tank. The effect is so striking because the gasoline and steel expand quickly. The rate of change in thermal properties is discussed later in this chapter.

If you try to cap the tank tightly to prevent overflow, you will find that it leaks anyway, either around the cap or by bursting the tank. Tightly constricting the expanding gas is equivalent to compressing it, and both liquids and solids resist compression with extremely large forces. To avoid rupturing rigid containers, these containers have air gaps, which allow them to expand and contract without stressing them.

16.1 Check Your Understanding Does a given reading on a gasoline gauge indicate more gasoline in cold weather or in hot weather, or does the temperature not matter?

16.4 | Heat Transfer and Specific Heat

Learning Objectives

By the end of this section, you will be able to:

- Explain phenomena involving heat as a form of energy transfer
- Solve problems involving heat transfer

We have seen in previous chapters that energy is one of the fundamental concepts of physics. **Heat** is a type of energy transfer that is caused by a temperature difference, and it can change the temperature of an object. As we learned earlier in this chapter, heat transfer is the movement of energy from one place or material to another as a result of a difference in temperature. Heat transfer is fundamental to such everyday activities as home heating and cooking, as well as many industrial processes. It also forms a basis for the topics in the remainder of this chapter.

We also introduce the concept of internal energy, which can be increased or decreased by heat transfer. We discuss another way to change the internal energy of a system, namely doing work on it. Thus, we are beginning the study of the relationship of heat and work, which is the basis of engines and refrigerators and the central topic (and origin of the name) of thermodynamics.

Internal Energy and Heat

A thermal system has *internal energy* (also called thermal energy), which is the sum of the mechanical energies of its molecules. A system's internal energy is proportional to its temperature. As we saw earlier in this chapter, if two objects at different temperatures are brought into contact with each other, energy is transferred from the hotter to the colder object until the bodies reach thermal equilibrium (that is, they are at the same temperature). No work is done by either object because no force acts through a distance (as we discussed in the section on **Work**). These observations reveal that heat is energy transferred spontaneously due to a temperature difference. **Figure 16.9** shows an example of heat transfer.

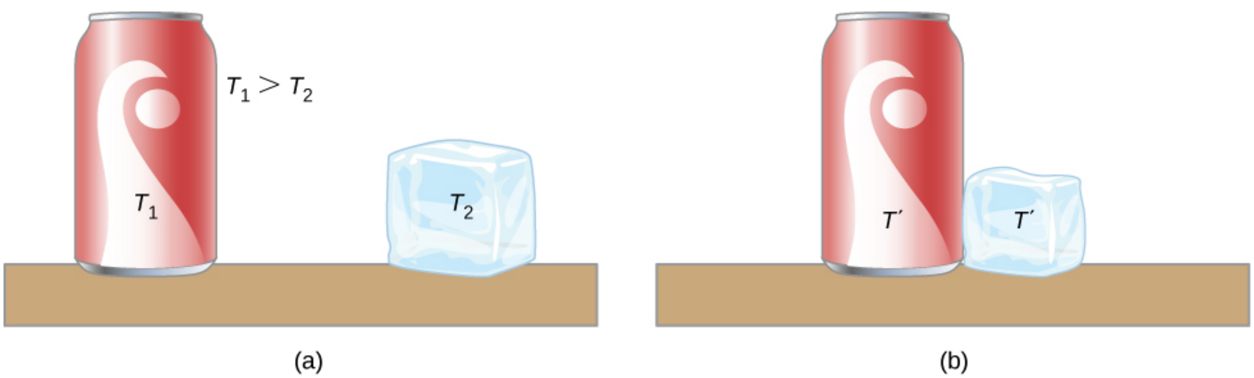


Figure 16.9 (a) Here, the soft drink has a higher temperature than the ice, so they are not in thermal equilibrium. (b) When the soft drink and ice are allowed to interact, heat is transferred from the drink to the ice due to the difference in temperatures until they reach the same temperature, T', achieving equilibrium. In fact, since the soft drink and ice are both in contact with the surrounding air and the bench, the ultimate equilibrium temperature will be the same as that of the surroundings.

The meaning of “heat” in physics is different from its ordinary meaning. For example, in conversation, we may say “the heat was unbearable,” but in physics, we would say that the temperature was high. Heat is a form of energy flow, whereas temperature is not. Incidentally, humans are sensitive to *heat flow* rather than to temperature.

Since heat is a form of energy, its SI unit is the joule (J). Another common unit of energy often used for heat is the **calorie** (cal), defined as the energy needed to change the temperature of 1.00 g of water by $1.00\,°\text{C}$ —specifically, between $14.5\,°\text{C}$ and $15.5\,°\text{C}$, since there is a slight temperature dependence. Also commonly used is the **kilocalorie** (kcal), which is the energy needed to change the temperature of 1.00 kg of water by $1.00\,°\text{C}$. Since mass is most often specified in kilograms, the kilocalorie is convenient. Confusingly, food calories (sometimes called “big calories,” abbreviated Cal) are actually kilocalories, a fact not easily determined from package labeling.

Mechanical Equivalent of Heat

It is also possible to change the temperature of a substance by doing work, which transfers energy into or out of a system. This realization helped establish that heat is a form of energy. James Prescott Joule (1818–1889) performed many experiments to establish the **mechanical equivalent of heat**—*the work needed to produce the same effects as heat transfer*. In the units used for these two quantities, the value for this equivalence is

$$1.000\,\text{kcal} = 4186\,\text{J}.$$

We consider this equation to represent the conversion between two units of energy. (Other numbers that you may see refer to calories defined for temperature ranges other than $14.5\,°\text{C}$ to $15.5\,°\text{C}$.)

Figure 16.10 shows one of Joule’s most famous experimental setups for demonstrating that work and heat can produce the same effects and measuring the mechanical equivalent of heat. It helped establish the principle of conservation of energy. Gravitational potential energy (U) was converted into kinetic energy (K), and then randomized by viscosity and turbulence into increased average kinetic energy of atoms and molecules in the system, producing a temperature increase. Joule’s contributions to thermodynamics were so significant that the SI unit of energy was named after him.

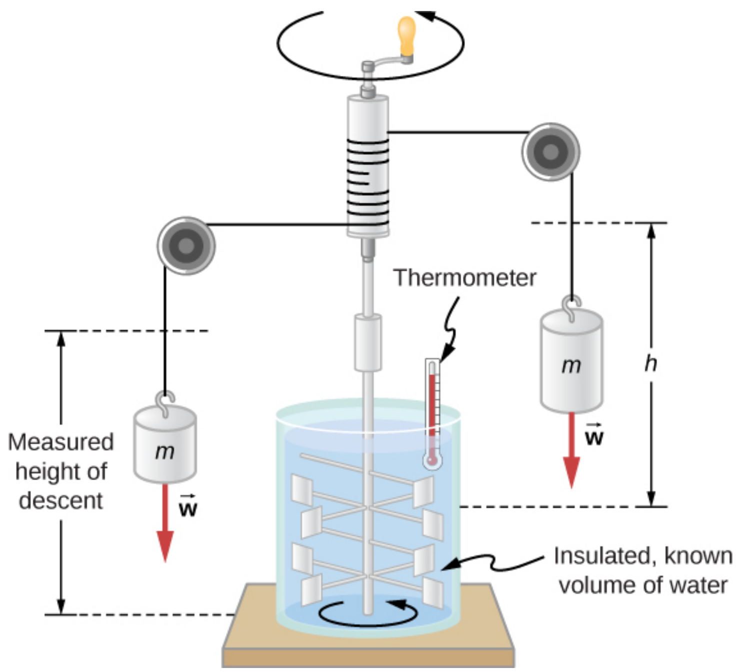


Figure 16.10 Joule’s experiment established the equivalence of heat and work. As the masses descended, they caused the paddles to do work, $W = mgh$, on the water. The result was a temperature increase, ΔT, measured by the thermometer. Joule found that ΔT was proportional to W and thus determined the mechanical equivalent of heat.

Increasing internal energy by heat transfer gives the same result as increasing it by doing work. Therefore, although a system has a well-defined internal energy, we cannot say that it has a certain “heat content” or “work content.” A well-defined quantity that depends only on the current state of the system, rather than on the history of that system, is known as a *state variable*. Temperature and internal energy are state variables. To sum up this paragraph, *heat and work are not state variables*.

Incidentally, increasing the internal energy of a system does not necessarily increase its temperature. As we'll see in the next section, the temperature does not change when a substance changes from one phase to another. An example is the melting of ice, which can be accomplished by adding heat or by doing frictional work, as when an ice cube is rubbed against a rough surface.

Temperature Change and Heat Capacity

We have noted that heat transfer often causes temperature change. Experiments show that with no phase change and no work done on or by the system, the transferred heat is typically directly proportional to the change in temperature and to the mass of the system, to a good approximation. (Below we show how to handle situations where the approximation is not valid.) The constant of proportionality depends on the substance and its phase, which may be gas, liquid, or solid. We omit discussion of the fourth phase, plasma, because although it is the most common phase in the universe, it is rare and short-lived on Earth.

We can understand the experimental facts by noting that the transferred heat is the change in the internal energy, which is the total energy of the molecules. Under typical conditions, the total kinetic energy of the molecules K_{total} is a constant fraction of the internal energy (for reasons and with exceptions that we'll see in the next chapter). The average kinetic energy of a molecule K_{ave} is proportional to the absolute temperature. Therefore, the change in internal energy of a system is typically proportional to the change in temperature and to the number of molecules, *N*. Mathematically, $\Delta U \propto \Delta K_{\text{total}} = NK_{\text{ave}} \propto N\Delta T$ The dependence on the substance results in large part from the different masses of atoms and molecules. We are considering its heat capacity in terms of its mass, but as we will see in the next chapter, in some cases, heat capacities *per molecule* are similar for different substances. The dependence on substance and phase also results from differences in the potential energy associated with interactions between atoms and molecules.

Heat Transfer and Temperature Change

A practical approximation for the relationship between heat transfer and temperature change is:

$$Q = mc\Delta T, \tag{16.5}$$

where Q is the symbol for heat transfer ("quantity of heat"), m is the mass of the substance, and ΔT is the change in temperature. The symbol c stands for the **specific heat** (also called "*specific heat capacity*") and depends on the material and phase. The specific heat is numerically equal to the amount of heat necessary to change the temperature of 1.00 kg of mass by $1.00\ °\text{C}$. The SI unit for specific heat is $\text{J/(kg} \times \text{K)}$ or $\text{J/(kg} \times °\text{C)}$. (Recall that the temperature change ΔT is the same in units of kelvin and degrees Celsius.)

Values of specific heat must generally be measured, because there is no simple way to calculate them precisely. **Table 16.3** lists representative values of specific heat for various substances. We see from this table that the specific heat of water is five times that of glass and 10 times that of iron, which means that it takes five times as much heat to raise the temperature of water a given amount as for glass, and 10 times as much as for iron. In fact, water has one of the largest specific heats of any material, which is important for sustaining life on Earth.

The specific heats of gases depend on what is maintained constant during the heating—typically either the volume or the pressure. In the table, the first specific heat value for each gas is measured at constant volume, and the second (in parentheses) is measured at constant pressure. We will return to this topic in the chapter on the kinetic theory of gases.

Substances	**Specific Heat (*c*)**	
Solids	$J/kg \cdot °C$	$kcal/kg \cdot °C$[2]
Aluminum	900	0.215
Asbestos	800	0.19
Concrete, granite (average)	840	0.20
Copper	387	0.0924
Glass	840	0.20
Gold	129	0.0308
Human body (average at $37\ °C$)	3500	0.83
Ice (average, $-50\ °C$ to $0\ °C$)	2090	0.50
Iron, steel	452	0.108
Lead	128	0.0305
Silver	235	0.0562
Wood	1700	0.40
Liquids		
Benzene	1740	0.415
Ethanol	2450	0.586
Glycerin	2410	0.576
Mercury	139	0.0333
Water $(15.0\ °C)$	4186	1.000
Gases[3]		
Air (dry)	721 (1015)	0.172 (0.242)
Ammonia	1670 (2190)	0.399 (0.523)
Carbon dioxide	638 (833)	0.152 (0.199)
Nitrogen	739 (1040)	0.177 (0.248)
Oxygen	651 (913)	0.156 (0.218)
Steam $(100\ °C)$	1520 (2020)	0.363 (0.482)

Table 16.3 Specific Heats of Various Substances[1] [1]The values for solids and liquids are at constant volume and $25\ °C$, except as noted. [2]These values are identical in units of $cal/g \cdot °C$. [3]Specific heats at constant volume and at $20.0\ °C$ except as noted, and at 1.00 atm pressure. Values in parentheses are specific heats at a constant pressure of 1.00 atm.

In general, specific heat also depends on temperature. Thus, a precise definition of c for a substance must be given in terms of an infinitesimal change in temperature. To do this, we note that $c = \frac{1}{m}\frac{\Delta Q}{\Delta T}$ and replace Δ with d:

$$c = \frac{1}{m}\frac{dQ}{dT}.$$

Except for gases, the temperature and volume dependence of the specific heat of most substances is weak at normal temperatures. Therefore, we will generally take specific heats to be constant at the values given in the table.

Example 16.4

Calculating the Required Heat

A 0.500-kg aluminum pan on a stove and 0.250 L of water in it are heated from $20.0\,°\text{C}$ to $80.0\,°\text{C}$. (a) How much heat is required? What percentage of the heat is used to raise the temperature of (b) the pan and (c) the water?

Strategy

We can assume that the pan and the water are always at the same temperature. When you put the pan on the stove, the temperature of the water and that of the pan are increased by the same amount. We use the equation for the heat transfer for the given temperature change and mass of water and aluminum. The specific heat values for water and aluminum are given in **Table 16.3**.

Solution

1. Calculate the temperature difference:

$$\Delta T = T_\text{f} - T_\text{i} = 60.0\,°\text{C}.$$

2. Calculate the mass of water. Because the density of water is $1000\,\text{kg/m}^3$, 1 L of water has a mass of 1 kg, and the mass of 0.250 L of water is $m_w = 0.250\,\text{kg}$.

3. Calculate the heat transferred to the water. Use the specific heat of water in **Table 16.3**:

$$Q_\text{w} = m_\text{w} c_\text{w} \Delta T = (0.250\,\text{kg})(4186\,\text{J/kg °C})(60.0\,°\text{C}) = 62.8\,\text{kJ}.$$

4. Calculate the heat transferred to the aluminum. Use the specific heat for aluminum in **Table 16.3**:

$$Q_\text{Al} = m_\text{Al} c_\text{Al} \Delta T = (0.500\,\text{kg})(900\,\text{J/kg °C})(60.0\,°\text{C}) = 27.0\,\text{kJ}.$$

5. Find the total transferred heat:

$$Q_\text{Total} = Q_\text{W} + Q_\text{Al} = 89.8\,\text{kJ}.$$

Significance

In this example, the heat transferred to the container is a significant fraction of the total transferred heat. Although the mass of the pan is twice that of the water, the specific heat of water is over four times that of aluminum. Therefore, it takes a bit more than twice as much heat to achieve the given temperature change for the water as for the aluminum pan.

Example 16.5 illustrates a temperature rise caused by doing work. (The result is the same as if the same amount of energy had been added with a blowtorch instead of mechanically.)

Example 16.5

Calculating the Temperature Increase from the Work Done on a Substance

Truck brakes used to control speed on a downhill run do work, converting gravitational potential energy into increased internal energy (higher temperature) of the brake material (**Figure 16.11**). This conversion prevents the gravitational potential energy from being converted into kinetic energy of the truck. Since the mass of the truck is much greater than that of the brake material absorbing the energy, the temperature increase may occur too fast for sufficient heat to transfer from the brakes to the environment; in other words, the brakes may overheat.

Figure 16.11 The smoking brakes on a braking truck are visible evidence of the mechanical equivalent of heat.

Calculate the temperature increase of 10 kg of brake material with an average specific heat of $800\ \text{J/kg} \cdot {}^\circ\text{C}$ if the material retains 10% of the energy from a 10,000-kg truck descending 75.0 m (in vertical displacement) at a constant speed.

Strategy

We calculate the gravitational potential energy (*Mgh*) that the entire truck loses in its descent, equate it to the increase in the brakes' internal energy, and then find the temperature increase produced in the brake material alone.

Solution

First we calculate the change in gravitational potential energy as the truck goes downhill:

$$Mgh = (10{,}000\ \text{kg})\left(9.80\ \text{m/s}^2\right)(75.0\ \text{m}) = 7.35 \times 10^6\ \text{J}.$$

Because the kinetic energy of the truck does not change, conservation of energy tells us the lost potential energy is dissipated, and we assume that 10% of it is transferred to internal energy of the brakes, so take $Q = Mgh/10$. Then we calculate the temperature change from the heat transferred, using

$$\Delta T = \frac{Q}{mc},$$

where *m* is the mass of the brake material. Insert the given values to find

$$\Delta T = \frac{7.35 \times 10^5\ \text{J}}{(10\ \text{kg})(800\ \text{J/kg}\ {}^\circ\text{C})} = 92\ {}^\circ\text{C}.$$

Significance

If the truck had been traveling for some time, then just before the descent, the brake temperature would probably be higher than the ambient temperature. The temperature increase in the descent would likely raise the temperature of the brake material very high, so this technique is not practical. Instead, the truck would use the technique of engine braking. A different idea underlies the recent technology of hybrid and electric cars, where mechanical energy (kinetic and gravitational potential energy) is converted by the brakes into electrical energy in the battery, a process called regenerative braking.

In a common kind of problem, objects at different temperatures are placed in contact with each other but isolated from everything else, and they are allowed to come into equilibrium. A container that prevents heat transfer in or out is called a **calorimeter**, and the use of a calorimeter to make measurements (typically of heat or specific heat capacity) is called **calorimetry**.

We will use the term "calorimetry problem" to refer to any problem in which the objects concerned are thermally isolated from their surroundings. An important idea in solving calorimetry problems is that during a heat transfer between objects isolated from their surroundings, the heat gained by the colder object must equal the heat lost by the hotter object, due to conservation of energy:

$$Q_{\text{cold}} + Q_{\text{hot}} = 0. \tag{16.6}$$

We express this idea by writing that the sum of the heats equals zero because the heat gained is usually considered positive; the heat lost, negative.

Example 16.6

Calculating the Final Temperature in Calorimetry

Suppose you pour 0.250 kg of 20.0-°C water (about a cup) into a 0.500-kg aluminum pan off the stove with a temperature of $150\,°\text{C}$. Assume no heat transfer takes place to anything else: The pan is placed on an insulated pad, and heat transfer to the air is neglected in the short time needed to reach equilibrium. Thus, this is a calorimetry problem, even though no isolating container is specified. Also assume that a negligible amount of water boils off. What is the temperature when the water and pan reach thermal equilibrium?

Strategy

Originally, the pan and water are not in thermal equilibrium: The pan is at a higher temperature than the water. Heat transfer restores thermal equilibrium once the water and pan are in contact; it stops once thermal equilibrium between the pan and the water is achieved. The heat lost by the pan is equal to the heat gained by the water—that is the basic principle of calorimetry.

Solution

1. Use the equation for heat transfer $Q = mc\Delta T$ to express the heat lost by the aluminum pan in terms of the mass of the pan, the specific heat of aluminum, the initial temperature of the pan, and the final temperature:
$$Q_{\text{hot}} = m_{\text{Al}} c_{\text{Al}} (T_{\text{f}} - 150\,°\text{C}).$$
2. Express the heat gained by the water in terms of the mass of the water, the specific heat of water, the initial temperature of the water, and the final temperature:
$$Q_{\text{cold}} = m_{\text{w}} c_{\text{w}} (T_{\text{f}} - 20.0\,°\text{C}).$$
3. Note that $Q_{\text{hot}} < 0$ and $Q_{\text{cold}} > 0$ and that as stated above, they must sum to zero:
$$\begin{aligned} Q_{\text{cold}} + Q_{\text{hot}} &= 0 \\ Q_{\text{cold}} &= -Q_{\text{hot}} \\ m_{\text{w}} c_{\text{w}} (T_{\text{f}} - 20.0\,°\text{C}) &= -m_{\text{Al}} c_{\text{Al}} (T_{\text{f}} - 150\,°\text{C}). \end{aligned}$$
4. Bring all terms involving T_{f} on the left hand side and all other terms on the right hand side. Solving for T_{f},
$$T_{\text{f}} = \frac{m_{\text{Al}} c_{\text{Al}} (150\,°\text{C}) + m_{\text{w}} c_{\text{w}} (20.0\,°\text{C})}{m_{\text{Al}} c_{\text{Al}} + m_{\text{w}} c_{\text{w}}},$$
and insert the numerical values:
$$T_{\text{f}} = \frac{(0.500\text{ kg})(900\text{ J/kg °C})(150\,°\text{C}) + (0.250\text{ kg})(4186\text{ J/kg °C})(20.0\,°\text{C})}{(0.500\text{ kg})(900\text{ J/kg °C}) + (0.250\text{ kg})(4186\text{ J/kg °C})} = 59.1\,°\text{C}.$$

Significance

Why is the final temperature so much closer to $20.0\,°\text{C}$ than to $150\,°\text{C}$? The reason is that water has a greater specific heat than most common substances and thus undergoes a smaller temperature change for a given heat transfer. A large body of water, such as a lake, requires a large amount of heat to increase its temperature appreciably. This explains why the temperature of a lake stays relatively constant during the day even when the temperature change of the air is large. However, the water temperature does change over longer times (e.g., summer to winter).

16.2 Check Your Understanding If 25 kJ is necessary to raise the temperature of a rock from $25\,°C$ to $30\,°C$, how much heat is necessary to heat the rock from $45\,°C$ to $50\,°C$?

16.5 | Phase Changes

Learning Objectives

By the end of this section, you will be able to:

- Describe phase transitions and equilibrium between phases
- Solve problems involving latent heat
- Solve calorimetry problems involving phase changes

Phase transitions play an important theoretical and practical role in the study of heat flow. In melting (or "fusion"), a solid turns into a liquid; the opposite process is freezing. In evaporation, a liquid turns into a gas; the opposite process is condensation.

A substance melts or freezes at a temperature called its melting point, and boils (evaporates rapidly) or condenses at its boiling point. These temperatures depend on pressure. High pressure favors the denser form, so typically, high pressure raises the melting point and boiling point, and low pressure lowers them. For example, the boiling point of water is $100\,°C$ at 1.00 atm. At higher pressure, the boiling point is higher, and at lower pressure, it is lower. The main exception is the melting and freezing of water, discussed in the next section.

Phase Diagrams

The phase of a given substance depends on the pressure and temperature. Thus, plots of pressure versus temperature showing the phase in each region provide considerable insight into thermal properties of substances. Such a *pT* graph is called a **phase diagram**.

Figure 16.12 shows the phase diagram for water. Using the graph, if you know the pressure and temperature, you can determine the phase of water. The solid curves—boundaries between phases—indicate phase transitions, that is, temperatures and pressures at which the phases coexist. For example, the boiling point of water is $100\,°C$ at 1.00 atm. As the pressure increases, the boiling temperature rises gradually to $374\,°C$ at a pressure of 218 atm. A pressure cooker (or even a covered pot) cooks food faster than an open pot, because the water can exist as a liquid at temperatures greater than $100\,°C$ without all boiling away. (As we'll see in the next section, liquid water conducts heat better than steam or hot air.) The boiling point curve ends at a certain point called the **critical point**—that is, a **critical temperature**, above which the liquid and gas phases cannot be distinguished; the substance is called a *supercritical fluid*. At sufficiently high pressure above the critical point, the gas has the density of a liquid but does not condense. Carbon dioxide, for example, is supercritical at all temperatures above $31.0\,°C$. **Critical pressure** is the pressure of the critical point.

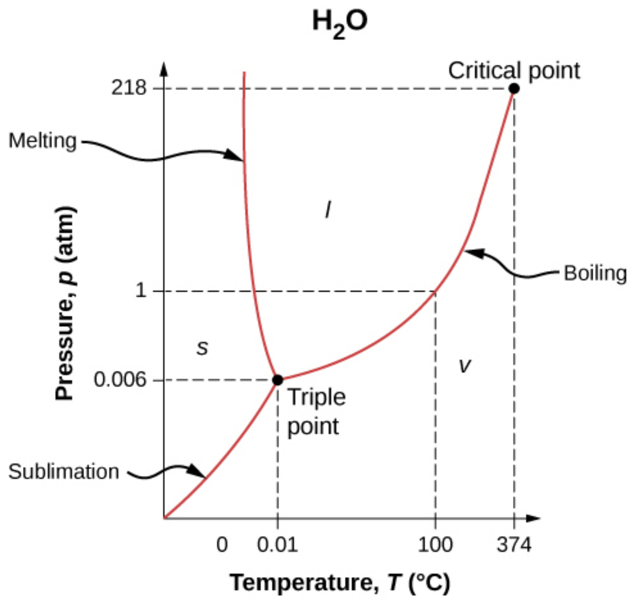


Figure 16.12 The phase diagram (*pT* graph) for water shows solid (s), liquid (l), and vapor (v) phases. At temperatures and pressure above those of the critical point, there is no distinction between liquid and vapor. Note that the axes are nonlinear and the graph is not to scale. This graph is simplified—it omits several exotic phases of ice at higher pressures. The phase diagram of water is unusual because the melting-point curve has a negative slope, showing that you can melt ice by *increasing* the pressure.

Similarly, the curve between the solid and liquid regions in **Figure 16.12** gives the melting temperature at various pressures. For example, the melting point is $0\,°\text{C}$ at 1.00 atm, as expected. Water has the unusual property that ice is less dense than liquid water at the melting point, so at a fixed temperature, you can change the phase from solid (ice) to liquid (water) by increasing the pressure. That is, the melting temperature of ice falls with increased pressure, as the phase diagram shows. For example, when a car is driven over snow, the increased pressure from the tires melts the snowflakes; afterwards, the water refreezes and forms an ice layer.

As you learned in the earlier section on thermometers and temperature scales, the triple point is the combination of temperature and pressure at which ice, liquid water, and water vapor can coexist stably—that is, all three phases exist in equilibrium. For water, the triple point occurs at $273.16\,\text{K}\,(0.01\,°\text{C})$ and 611.2 Pa; that is a more accurate calibration temperature than the melting point of water at 1.00 atm, or $273.15\,\text{K}\,(0.0\,°\text{C})$.

View this **video (https://openstax.org/l/21triplepoint)** to see a substance at its triple point.

At pressures below that of the triple point, there is no liquid phase; the substance can exist as either gas or solid. For water, there is no liquid phase at pressures below 0.00600 atm. The phase change from solid to gas is called **sublimation**. You may have noticed that snow can disappear into thin air without a trace of liquid water, or that ice cubes can disappear in a freezer. Both are examples of sublimation. The reverse also happens: Frost can form on very cold windows without going through the liquid stage. **Figure 16.13** shows the result, as well as showing a familiar example of sublimation. Carbon dioxide has no liquid phase at atmospheric pressure. Solid CO_2 is known as dry ice because instead of melting, it sublimes. Its sublimation temperature at atmospheric pressure is $-78\,°\text{C}$. Certain air fresheners use the sublimation of a solid to spread a perfume around a room. Some solids, such as osmium tetroxide, are so toxic that they must be kept in sealed containers to prevent human exposure to their sublimation-produced vapors.

(a) (b)

Figure 16.13 Direct transitions between solid and vapor are common, sometimes useful, and even beautiful. (a) Dry ice sublimes directly to carbon dioxide gas. The visible "smoke" consists of water droplets that condensed in the air cooled by the dry ice. (b) Frost forms patterns on a very cold window, an example of a solid formed directly from a vapor. (credit a: modification of work by Windell Oskay; credit b: modification of work by Liz West)

Equilibrium

At the melting temperature, the solid and liquid phases are in equilibrium. If heat is added, some of the solid will melt, and if heat is removed, some of the liquid will freeze. The situation is somewhat more complex for liquid-gas equilibrium. Generally, liquid and gas are in equilibrium at any temperature. We call the gas phase a **vapor** when it exists at a temperature below the boiling temperature, as it does for water at $20.0\,°\text{C}$. Liquid in a closed container at a fixed temperature evaporates until the pressure of the gas reaches a certain value, called the **vapor pressure**, which depends on the gas and the temperature. At this equilibrium, if heat is added, some of the liquid will evaporate, and if heat is removed, some of the gas will condense; molecules either join the liquid or form suspended droplets. If there is not enough liquid for the gas to reach the vapor pressure in the container, all the liquid eventually evaporates.

If the vapor pressure of the liquid is greater than the *total* ambient pressure, including that of any air (or other gas), the liquid evaporates rapidly; in other words, it boils. Thus, the boiling point of a liquid at a given pressure is the temperature at which its vapor pressure equals the ambient pressure. Liquid and gas phases are in equilibrium at the boiling temperature (**Figure 16.14**). If a substance is in a closed container at the boiling point, then the liquid is boiling and the gas is condensing at the same rate without net change in their amounts.

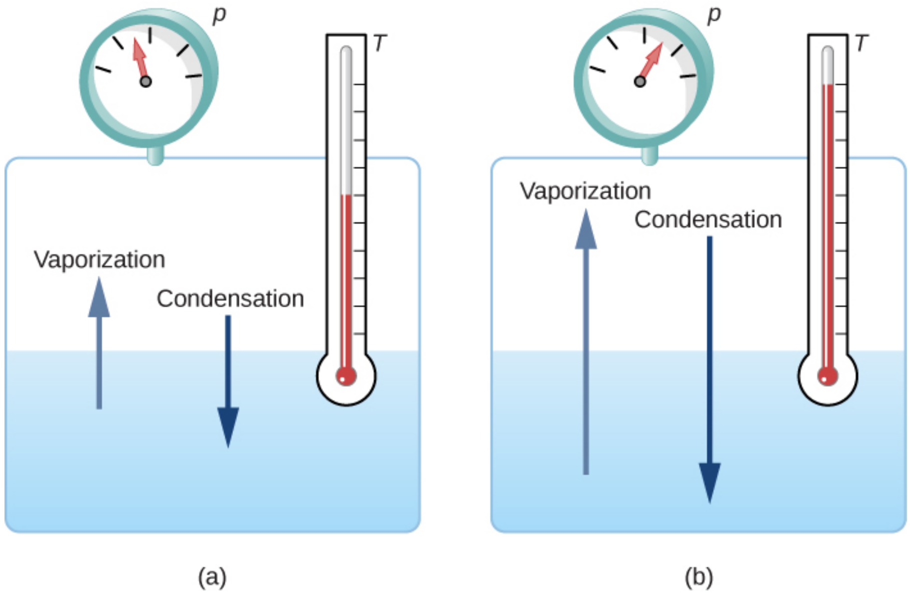


Figure 16.14 Equilibrium between liquid and gas at two different boiling points inside a closed container. (a) The rates of boiling and condensation are equal at this combination of temperature and pressure, so the liquid and gas phases are in equilibrium. (b) At a higher temperature, the boiling rate is faster, that is, the rate at which molecules leave the liquid and enter the gas is faster. This increases the number of molecules in the gas, which increases the gas pressure, which in turn increases the rate at which gas molecules condense and enter the liquid. The pressure stops increasing when it reaches the point where the boiling rate and the condensation rate are equal. The gas and liquid are in equilibrium again at this higher temperature and pressure.

For water, $100\,°C$ is the boiling point at 1.00 atm, so water and steam should exist in equilibrium under these conditions. Why does an open pot of water at $100\,°C$ boil completely away? The gas surrounding an open pot is not pure water: it is mixed with air. If pure water and steam are in a closed container at $100\,°C$ and 1.00 atm, they will coexist—but with air over the pot, there are fewer water molecules to condense, and water boils away. Another way to see this is that at the boiling point, the vapor pressure equals the ambient pressure. However, part of the ambient pressure is due to air, so the pressure of the steam is less than the vapor pressure at that temperature, and evaporation continues. Incidentally, the equilibrium vapor pressure of solids is not zero, a fact that accounts for sublimation.

16.3 Check Your Understanding Explain why a cup of water (or soda) with ice cubes stays at $0\,°C,$ even on a hot summer day.

Phase Change and Latent Heat

So far, we have discussed heat transfers that cause temperature change. However, in a phase transition, heat transfer does not cause any temperature change.

For an example of phase changes, consider the addition of heat to a sample of ice at $-20\,°C$ (**Figure 16.15**) and atmospheric pressure. The temperature of the ice rises linearly, absorbing heat at a constant rate of $2090\,J/kg \cdot °C$ until it reaches $0\,°C.$ Once at this temperature, the ice begins to melt and continues until it has all melted, absorbing 333 kJ/kg of heat. The temperature remains constant at $0\,°C$ during this phase change. Once all the ice has melted, the temperature of the liquid water rises, absorbing heat at a new constant rate of $4186\,J/kg \cdot °C.$ At $100\,°C,$ the water begins to boil. The temperature again remains constant during this phase change while the water absorbs 2256 kJ/kg of heat and turns into steam. When all the liquid has become steam, the temperature rises again, absorbing heat at a rate of $2020\,J/kg \cdot °C$. If we started with steam and cooled it to make it condense into liquid water and freeze into ice, the process would exactly reverse, with the temperature again constant during each phase transition.

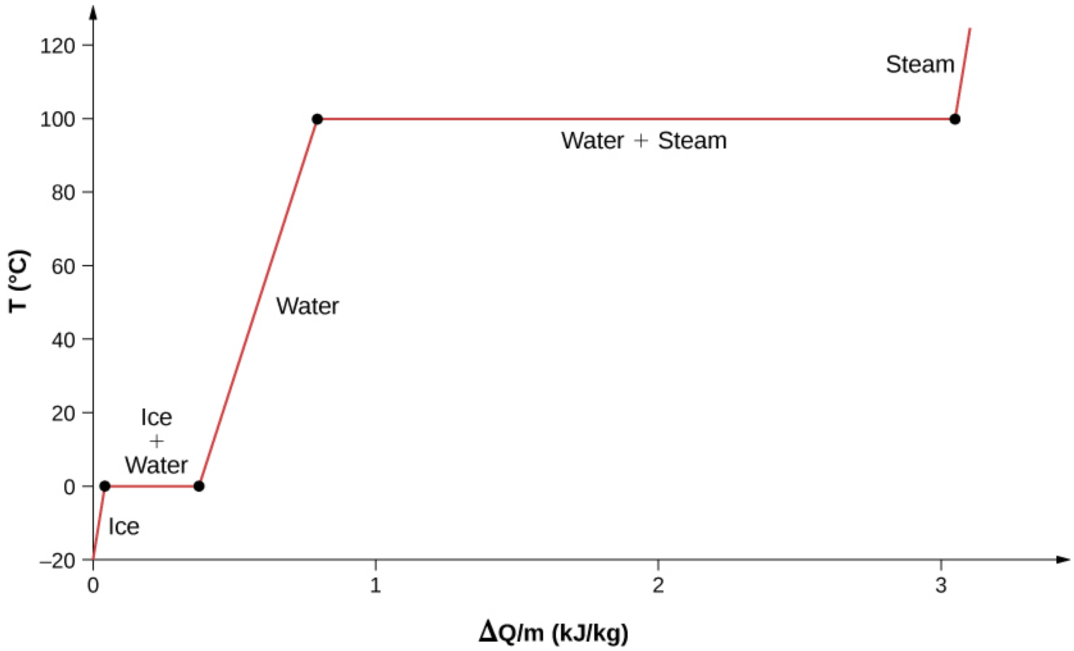


Figure 16.15 Temperature versus heat. The system is constructed so that no vapor evaporates while ice warms to become liquid water, and so that, when vaporization occurs, the vapor remains in the system. The long stretches of constant temperatures at $0\ °C$ and $100\ °C$ reflect the large amounts of heat needed to cause melting and vaporization, respectively.

Where does the heat added during melting or boiling go, considering that the temperature does not change until the transition is complete? Energy is required to melt a solid, because the attractive forces between the molecules in the solid must be broken apart, so that in the liquid, the molecules can move around at comparable kinetic energies; thus, there is no rise in temperature. Energy is needed to vaporize a liquid for similar reasons. Conversely, work is done by attractive forces when molecules are brought together during freezing and condensation. That energy must be transferred out of the system, usually in the form of heat, to allow the molecules to stay together (**Figure 16.18**). Thus, condensation occurs in association with cold objects—the glass in **Figure 16.16**, for example.

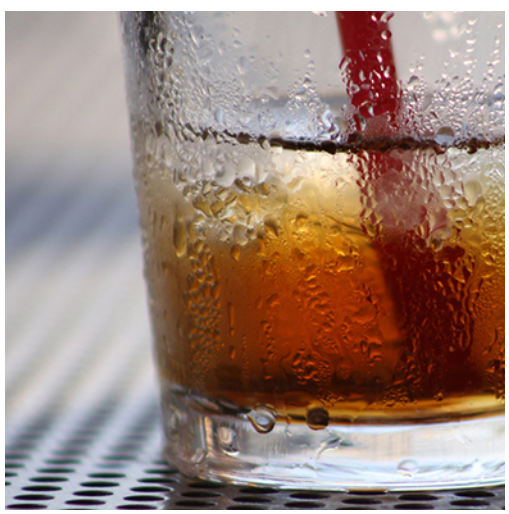


Figure 16.16 Condensation forms on this glass of iced tea because the temperature of the nearby air is reduced. The air cannot hold as much water as it did at room temperature, so water condenses. Energy is released when the water condenses, speeding the melting of the ice in the glass. (credit: Jenny Downing)

The energy released when a liquid freezes is used by orange growers when the temperature approaches $0\,°C$. Growers spray water on the trees so that the water freezes and heat is released to the growing oranges. This prevents the temperature inside the orange from dropping below freezing, which would damage the fruit (**Figure 16.17**).

Figure 16.17 The ice on these trees released large amounts of energy when it froze, helping to prevent the temperature of the trees from dropping below $0\,°C$. Water is intentionally sprayed on orchards to help prevent hard frosts. (credit: Hermann Hammer)

The energy involved in a phase change depends on the number of bonds or force pairs and their strength. The number of bonds is proportional to the number of molecules and thus to the mass of the sample. The energy per unit mass required to change a substance from the solid phase to the liquid phase, or released when the substance changes from liquid to solid, is known as the **heat of fusion**. The energy per unit mass required to change a substance from the liquid phase to the vapor phase is known as the **heat of vaporization**. The strength of the forces depends on the type of molecules. The heat *Q* absorbed or released in a phase change in a sample of mass *m* is given by

$$Q = mL_\text{f}(\text{melting/freezing}) \tag{16.7}$$

$$Q = mL_\text{v}(\text{vaporization/condensation}) \tag{16.8}$$

where the latent heat of fusion L_f and latent heat of vaporization L_v are material constants that are determined experimentally. (Latent heats are also called **latent heat coefficients** and heats of transformation.) These constants are "latent," or hidden, because in phase changes, energy enters or leaves a system without causing a temperature change in the system, so in effect, the energy is hidden.

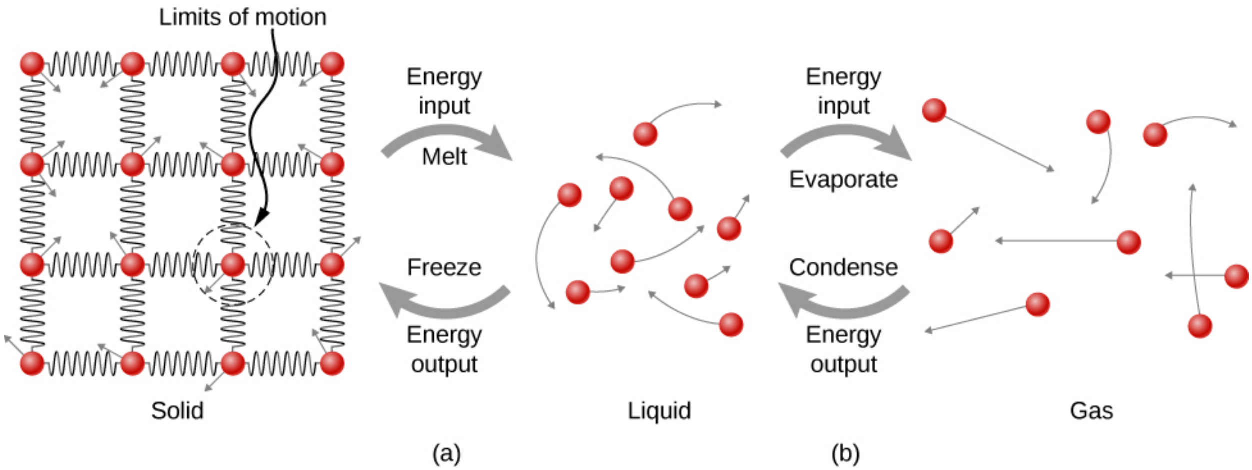


Figure 16.18 (a) Energy is required to partially overcome the attractive forces (modeled as springs) between molecules in a solid to form a liquid. That same energy must be removed from the liquid for freezing to take place. (b) Molecules become separated by large distances when going from liquid to vapor, requiring significant energy to completely overcome molecular attraction. The same energy must be removed from the vapor for condensation to take place.

Table 16.4 lists representative values of L_f and L_v in kJ/kg, together with melting and boiling points. Note that in general, $L_v > L_f$. The table shows that the amounts of energy involved in phase changes can easily be comparable to or greater than those involved in temperature changes, as **Figure 16.15** and the accompanying discussion also showed.

		L_f			L_v	
Substance	**Melting Point** (°C)	**kJ/kg**	**kcal/kg**	**Boiling Point** (°C)	**kJ/kg**	**kcal/kg**
Helium[2]	$-272.2\ (0.95\ \text{K})$	5.23	1.25	$-268.9(4.2\ \text{K})$	20.9	4.99
Hydrogen	$-259.3(13.9\ \text{K})$	58.6	14.0	$-252.9(20.2\ \text{K})$	452	108
Nitrogen	$-210.0(63.2\ \text{K})$	25.5	6.09	$-195.8(77.4\ \text{K})$	201	48.0
Oxygen	$-218.8(54.4\ \text{K})$	13.8	3.30	$-183.0(90.2\ \text{K})$	213	50.9
Ethanol	–114	104	24.9	78.3	854	204
Ammonia	–75	332	79.3	–33.4	1370	327
Mercury	–38.9	11.8	2.82	357	272	65.0
Water	0.00	334	79.8	100.0	2256[3]	539[4]
Sulfur	119	38.1	9.10	444.6	326	77.9
Lead	327	24.5	5.85	1750	871	208
Antimony	631	165	39.4	1440	561	134
Aluminum	660	380	90	2450	11400	2720
Silver	961	88.3	21.1	2193	2336	558

Table 16.4 Heats of Fusion and Vaporization[1] [1]Values quoted at the normal melting and boiling temperatures at standard atmospheric pressure ($1\ \text{atm}$). [2]Helium has no solid phase at atmospheric pressure. The melting point given is at a pressure of 2.5 MPa. [3]At $37.0\ °\text{C}$ (body temperature), the heat of vaporization L_v for water is 2430 kJ/kg or 580 kcal/kg. [4]At $37.0\ °\text{C}$ (body temperature), the heat of vaporization, L_v for water is 2430 kJ/kg or 580 kcal/kg.

		L_f			L_v	
Gold	1063	64.5	15.4	2660	1578	377
Copper	1083	134	32.0	2595	5069	1211
Uranium	1133	84	20	3900	1900	454
Tungsten	3410	184	44	5900	4810	1150

Table 16.4 Heats of Fusion and Vaporization[1] [1]Values quoted at the normal melting and boiling temperatures at standard atmospheric pressure (1 atm). [2]Helium has no solid phase at atmospheric pressure. The melting point given is at a pressure of 2.5 MPa. [3]At $37.0\,°\text{C}$ (body temperature), the heat of vaporization L_v for water is 2430 kJ/kg or 580 kcal/kg. [4]At $37.0\,°\text{C}$ (body temperature), the heat of vaporization, L_v for water is 2430 kJ/kg or 580 kcal/kg.

Phase changes can have a strong stabilizing effect on temperatures that are not near the melting and boiling points, since evaporation and condensation occur even at temperatures below the boiling point. For example, air temperatures in humid climates rarely go above approximately $38.0\,°\text{C}$ because most heat transfer goes into evaporating water into the air. Similarly, temperatures in humid weather rarely fall below the dew point—the temperature where condensation occurs given the concentration of water vapor in the air—because so much heat is released when water vapor condenses.

More energy is required to evaporate water below the boiling point than at the boiling point, because the kinetic energy of water molecules at temperatures below $100\,°\text{C}$ is less than that at $100\,°\text{C}$, so less energy is available from random thermal motions. For example, at body temperature, evaporation of sweat from the skin requires a heat input of 2428 kJ/kg, which is about 10% higher than the latent heat of vaporization at $100\,°\text{C}$. This heat comes from the skin, and this evaporative cooling effect of sweating helps reduce the body temperature in hot weather. However, high humidity inhibits evaporation, so that body temperature might rise, while unevaporated sweat might be left on your brow.

Example 16.7

Calculating Final Temperature from Phase Change

Three ice cubes are used to chill a soda at $20\,°\text{C}$ with mass $m_\text{soda} = 0.25\text{ kg}$. The ice is at $0\,°\text{C}$ and each ice cube has a mass of 6.0 g. Assume that the soda is kept in a foam container so that heat loss can be ignored and that the soda has the same specific heat as water. Find the final temperature when all ice has melted.

Strategy

The ice cubes are at the melting temperature of $0\,°\text{C}.$ Heat is transferred from the soda to the ice for melting. Melting yields water at $0\,°\text{C},$ so more heat is transferred from the soda to this water until the water plus soda system reaches thermal equilibrium.

The heat transferred to the ice is

$$Q_\text{ice} = m_\text{ice}L_\text{f} + m_\text{ice}c_\text{W}(T_\text{f} - 0\,°\text{C}).$$

The heat given off by the soda is

$$Q_\text{soda} = m_\text{soda}c_\text{W}(T_\text{f} - 20\,°\text{C}).$$

Since no heat is lost, $Q_\text{ice} = -Q_\text{soda},$ as in **Example 16.6**, so that

$$m_\text{ice}L_\text{f} + m_\text{ice}c_\text{W}(T_\text{f} - 0\,°\text{C}) = -m_\text{soda}c_\text{W}(T_\text{f} - 20\,°\text{C}).$$

Solve for the unknown quantity T_f:

$$T_\text{f} = \frac{m_\text{soda}c_\text{W}(20\,°\text{C}) - m_\text{ice}L_\text{f}}{(m_\text{soda} + m_\text{ice})c_\text{W}}.$$

Solution

First we identify the known quantities. The mass of ice is $m_{\text{ice}} = 3 \times 6.0\ \text{g} = 0.018\ \text{kg}$ and the mass of soda is $m_{\text{soda}} = 0.25\ \text{kg}$. Then we calculate the final temperature:

$$T_{\text{f}} = \frac{20{,}930\ \text{J} - 6012\ \text{J}}{1122\ \text{J/°C}} = 13\ \text{°C}.$$

Significance

This example illustrates the large energies involved during a phase change. The mass of ice is about 7% of the mass of the soda but leads to a noticeable change in the temperature of the soda. Although we assumed that the ice was at the freezing temperature, this is unrealistic for ice straight out of a freezer: The typical temperature is $-6\ \text{°C}$. However, this correction makes no significant change from the result we found. Can you explain why?

Like solid-liquid and and liquid-vapor transitions, direct solid-vapor transitions or sublimations involve heat. The energy transferred is given by the equation $Q = mL_{\text{s}}$, where L_{s} is the **heat of sublimation**, analogous to L_{f} and L_{v}. The heat of sublimation at a given temperature is equal to the heat of fusion plus the heat of vaporization at that temperature.

We can now calculate any number of effects related to temperature and phase change. In each case, it is necessary to identify which temperature and phase changes are taking place. Keep in mind that heat transfer and work can cause both temperature and phase changes.

Problem-Solving Strategy: The Effects of Heat Transfer

1. Examine the situation to determine that there is a change in the temperature or phase. Is there heat transfer into or out of the system? When it is not obvious whether a phase change occurs or not, you may wish to first solve the problem as if there were no phase changes, and examine the temperature change obtained. If it is sufficient to take you past a boiling or melting point, you should then go back and do the problem in steps—temperature change, phase change, subsequent temperature change, and so on.
2. Identify and list all objects that change temperature or phase.
3. Identify exactly what needs to be determined in the problem (identify the unknowns). A written list is useful.
4. Make a list of what is given or what can be inferred from the problem as stated (identify the knowns). If there is a temperature change, the transferred heat depends on the specific heat of the substance (**Heat Transfer and Specific Heat**), and if there is a phase change, the transferred heat depends on the latent heat of the substance (**Table 16.4**).
5. Solve the appropriate equation for the quantity to be determined (the unknown).
6. Substitute the knowns along with their units into the appropriate equation and obtain numerical solutions complete with units. You may need to do this in steps if there is more than one state to the process, such as a temperature change followed by a phase change. However, in a calorimetry problem, each step corresponds to a term in the single equation $Q_{\text{hot}} + Q_{\text{cold}} = 0$.
7. Check the answer to see if it is reasonable. Does it make sense? As an example, be certain that any temperature change does not also cause a phase change that you have not taken into account.

16.4 Check Your Understanding Why does snow often remain even when daytime temperatures are higher than the freezing temperature?

16.6 | Mechanisms of Heat Transfer

Learning Objectives

By the end of this section, you will be able to:

- Explain some phenomena that involve conductive, convective, and radiative heat transfer
- Solve problems on the relationships between heat transfer, time, and rate of heat transfer
- Solve problems using the formulas for conduction and radiation

Just as interesting as the effects of heat transfer on a system are the methods by which it occurs. Whenever there is a temperature difference, heat transfer occurs. It may occur rapidly, as through a cooking pan, or slowly, as through the walls of a picnic ice chest. So many processes involve heat transfer that it is hard to imagine a situation where no heat transfer occurs. Yet every heat transfer takes place by only three methods:

1. **Conduction** is heat transfer through stationary matter by physical contact. (The matter is stationary on a macroscopic scale—we know that thermal motion of the atoms and molecules occurs at any temperature above absolute zero.) Heat transferred from the burner of a stove through the bottom of a pan to food in the pan is transferred by conduction.
2. **Convection** is the heat transfer by the macroscopic movement of a fluid. This type of transfer takes place in a forced-air furnace and in weather systems, for example.
3. Heat transfer by **radiation** occurs when microwaves, infrared radiation, visible light, or another form of electromagnetic radiation is emitted or absorbed. An obvious example is the warming of Earth by the Sun. A less obvious example is thermal radiation from the human body.

In the illustration at the beginning of this chapter, the fire warms the snowshoers' faces largely by radiation. Convection carries some heat to them, but most of the air flow from the fire is upward (creating the familiar shape of flames), carrying heat to the food being cooked and into the sky. The snowshoers wear clothes designed with low conductivity to prevent heat flow out of their bodies.

In this section, we examine these methods in some detail. Each method has unique and interesting characteristics, but all three have two things in common: They transfer heat solely because of a temperature difference, and the greater the temperature difference, the faster the heat transfer (**Figure 16.19**).

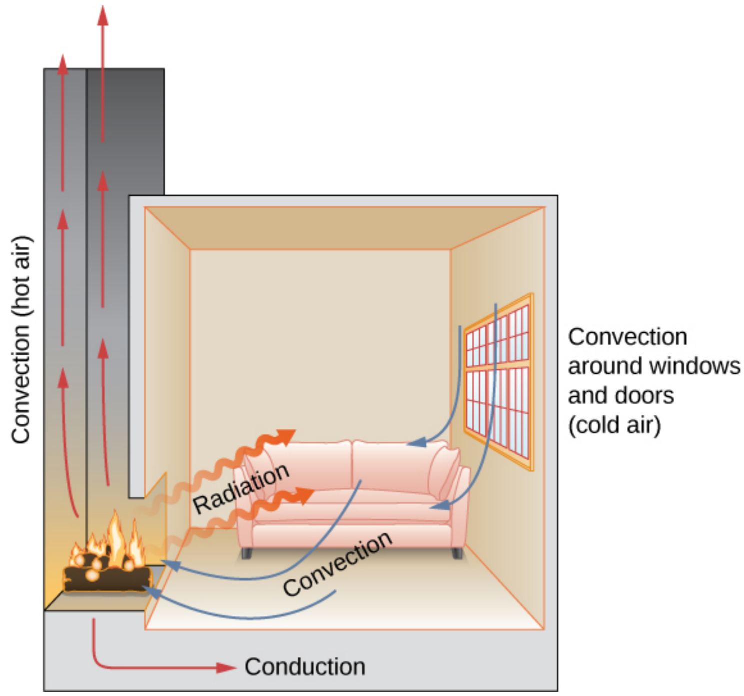


Figure 16.19 In a fireplace, heat transfer occurs by all three methods: conduction, convection, and radiation. Radiation is responsible for most of the heat transferred into the room. Heat transfer also occurs through conduction into the room, but much slower. Heat transfer by convection also occurs through cold air entering the room around windows and hot air leaving the room by rising up the chimney.

16.5 **Check Your Understanding** Name an example from daily life (different from the text) for each mechanism of heat transfer.

Conduction

As you walk barefoot across the living room carpet in a cold house and then step onto the kitchen tile floor, your feet feel colder on the tile. This result is intriguing, since the carpet and tile floor are both at the same temperature. The different sensation is explained by the different rates of heat transfer: The heat loss is faster for skin in contact with the tiles than with the carpet, so the sensation of cold is more intense.

Some materials conduct thermal energy faster than others. Figure 16.20 shows a material that conducts heat slowly—it is a good thermal insulator, or poor heat conductor—used to reduce heat flow into and out of a house.

Figure 16.20 Insulation is used to limit the conduction of heat from the inside to the outside (in winter) and from the outside to the inside (in summer). (credit: Giles Douglas)

A molecular picture of heat conduction will help justify the equation that describes it. **Figure 16.21** shows molecules in two bodies at different temperatures, T_{h} and $T_{\text{c}},$ for "hot" and "cold." The average kinetic energy of a molecule in the hot body is higher than in the colder body. If two molecules collide, energy transfers from the high-energy to the low-energy molecule. In a metal, the picture would also include free valence electrons colliding with each other and with atoms, likewise transferring energy. The cumulative effect of all collisions is a net flux of heat from the hotter body to the colder body. Thus, the rate of heat transfer increases with increasing temperature difference $\Delta T = T_{\text{h}} - T_{\text{c}}.$ If the temperatures are the same, the net heat transfer rate is zero. Because the number of collisions increases with increasing area, heat conduction is proportional to the cross-sectional area—a second factor in the equation.

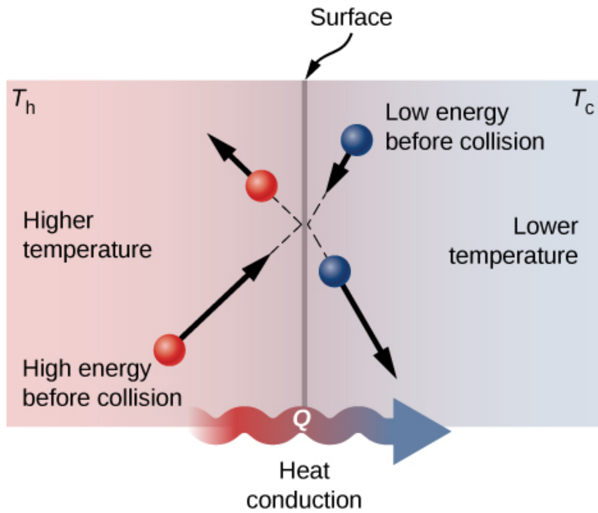


Figure 16.21 Molecules in two bodies at different temperatures have different average kinetic energies. Collisions occurring at the contact surface tend to transfer energy from high-temperature regions to low-temperature regions. In this illustration, a molecule in the lower-temperature region (right side) has low energy before collision, but its energy increases after colliding with a high-energy molecule at the contact surface. In contrast, a molecule in the higher-temperature region (left side) has high energy before collision, but its energy decreases after colliding with a low-energy molecule at the contact surface.

A third quantity that affects the conduction rate is the thickness of the material through which heat transfers. **Figure 16.22** shows a slab of material with a higher temperature on the left than on the right. Heat transfers from the left to the right by a series of molecular collisions. The greater the distance between hot and cold, the more time the material takes to transfer the same amount of heat.

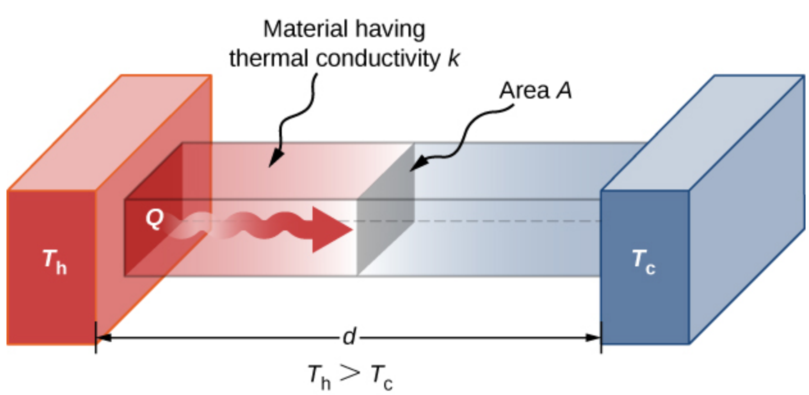


Figure 16.22 Heat conduction occurs through any material, represented here by a rectangular bar, whether window glass or walrus blubber.

All four of these quantities appear in a simple equation deduced from and confirmed by experiments. The **rate of conductive heat transfer** through a slab of material, such as the one in **Figure 16.22**, is given by

$$P = \frac{dQ}{dt} = \frac{kA(T_h - T_c)}{d} \quad (16.9)$$

where P is the power or rate of heat transfer in watts or in kilocalories per second, A and d are its surface area and thickness, as shown in **Figure 16.22**, $T_h - T_c$ is the temperature difference across the slab, and k is the **thermal conductivity** of the material. **Table 16.5** gives representative values of thermal conductivity.

More generally, we can write

$$P = -kA\frac{dT}{dx},$$

where x is the coordinate in the direction of heat flow. Since in **Figure 16.22**, the power and area are constant, dT/dx is constant, and the temperature decreases linearly from T_h to T_c.

Substance	Thermal Conductivity k $(\text{W/m}\cdot{}^\circ\text{C})$
Diamond	2000
Silver	420
Copper	390
Gold	318
Aluminum	220
Steel iron	80
Steel (stainless)	14
Ice	2.2
Glass (average)	0.84
Concrete brick	0.84
Water	0.6
Fatty tissue (without blood)	0.2
Asbestos	0.16
Plasterboard	0.16
Wood	0.08–0.16
Snow (dry)	0.10
Cork	0.042
Glass wool	0.042
Wool	0.04
Down feathers	0.025
Air	0.023
Polystyrene foam	0.010

Table 16.5 Thermal Conductivities of Common Substances Values are given for temperatures near $0\,^\circ\text{C}$.

Example 16.8

Calculating Heat Transfer through Conduction

A polystyrene foam icebox has a total area of $0.950\,\text{m}^2$ and walls with an average thickness of 2.50 cm. The box contains ice, water, and canned beverages at $0\,^\circ\text{C}$. The inside of the box is kept cold by melting ice. How much ice melts in one day if the icebox is kept in the trunk of a car at $35.0\,^\circ\text{C}$?

Strategy

This question involves both heat for a phase change (melting of ice) and the transfer of heat by conduction. To find the amount of ice melted, we must find the net heat transferred. This value can be obtained by calculating the rate of heat transfer by conduction and multiplying by time.

Solution

First we identify the knowns.

$k = 0.010\ \text{W/m}\cdot\text{°C}$ for polystyrene foam; $A = 0.950\ \text{m}^2$; $d = 2.50\ \text{cm} = 0.0250\ \text{m}$; ; $T_c = 0\ \text{°C}$; $T_h = 35.0\ \text{°C}$; $t = 1\ \text{day} = 24\ \text{hours - }86{,}400\ \text{s}$.

Then we identify the unknowns. We need to solve for the mass of the ice, *m*. We also need to solve for the net heat transferred to melt the ice, *Q*. The rate of heat transfer by conduction is given by

$$P = \frac{dQ}{dt} = \frac{kA(T_h - T_c)}{d}.$$

The heat used to melt the ice is $Q = mL_f$ .We insert the known values:

$$P = \frac{(0.010\ \text{W/m}\cdot\text{°C})\left(0.950\ \text{m}^2\right)(35.0\ \text{°C} - 0\ \text{°C})}{0.0250\ \text{m}} = 13.3\ \text{W}.$$

Multiplying the rate of heat transfer by the time we obtain

$$Q = Pt = (13.3\ \text{W})(86.400\ \text{s}) = 1.15 \times 10^6\ \text{J}.$$

We set this equal to the heat transferred to melt the ice, $Q = mL_f$, and solve for the mass *m*:

$$m = \frac{Q}{L_f} = \frac{1.15 \times 10^6\ \text{J}}{334 \times 10^3\ \text{J/kg}} = 3.44\ \text{kg}.$$

Significance

The result of 3.44 kg, or about 7.6 lb, seems about right, based on experience. You might expect to use about a 4 kg (7–10 lb) bag of ice per day. A little extra ice is required if you add any warm food or beverages.

Table 16.5 shows that polystyrene foam is a very poor conductor and thus a good insulator. Other good insulators include fiberglass, wool, and goosedown feathers. Like polystyrene foam, these all contain many small pockets of air, taking advantage of air's poor thermal conductivity.

In developing insulation, the smaller the conductivity *k* and the larger the thickness *d*, the better. Thus, the ratio *d/k*, called the *R factor*, is large for a good insulator. The rate of conductive heat transfer is inversely proportional to *R*. *R* factors are most commonly quoted for household insulation, refrigerators, and the like. Unfortunately, in the United States, *R* is still in non-metric units of $\text{ft}^2\cdot\text{°F}\cdot\text{h/Btu}$, although the unit usually goes unstated [1 British thermal unit (Btu) is the amount of energy needed to change the temperature of 1.0 lb of water by $1.0\ \text{°F}$, which is 1055.1 J]. A couple of representative values are an *R* factor of 11 for 3.5-inch-thick fiberglass batts (pieces) of insulation and an *R* factor of 19 for 6.5-inch-thick fiberglass batts (**Figure 16.23**). In the US, walls are usually insulated with 3.5-inch batts, whereas ceilings are usually insulated with 6.5-inch batts. In cold climates, thicker batts may be used.

Figure 16.23 The fiberglass batt is used for insulation of walls and ceilings to prevent heat transfer between the inside of the building and the outside environment. (credit: Tracey Nicholls)

Note that in **Table 16.5**, most of the best thermal conductors—silver, copper, gold, and aluminum—are also the best electrical conductors, because they contain many free electrons that can transport thermal energy. (Diamond, an electrical insulator, conducts heat by atomic vibrations.) Cooking utensils are typically made from good conductors, but the handles of those used on the stove are made from good insulators (bad conductors).

Example 16.9

Two Conductors End to End

A steel rod and an aluminum rod, each of diameter 1.00 cm and length 25.0 cm, are welded end to end. One end of the steel rod is placed in a large tank of boiling water at $100\,°\text{C}$, while the far end of the aluminum rod is placed in a large tank of water at $20\,°\text{C}$. The rods are insulated so that no heat escapes from their surfaces. What is the temperature at the joint, and what is the rate of heat conduction through this composite rod?

Strategy

The heat that enters the steel rod from the boiling water has no place to go but through the steel rod, then through the aluminum rod, to the cold water. Therefore, we can equate the rate of conduction through the steel to the rate of conduction through the aluminum.

We repeat the calculation with a second method, in which we use the thermal resistance *R* of the rod, since it simply adds when two rods are joined end to end. (We will use a similar method in the chapter on direct-current circuits.)

Solution

1. Identify the knowns and convert them to SI units.
   The length of each rod is $L_{\text{A1}} = L_{\text{steel}} = 0.25\text{ m}$, the cross-sectional area of each rod is $A_{\text{A1}} = A_{\text{steel}} = 7.85 \times 10^{-5}\text{ m}^2$, the thermal conductivity of aluminum is $k_{\text{A1}} = 220\text{ W/m}\cdot°\text{C}$, the thermal conductivity of steel is $k_{\text{steel}} = 80\text{ W/m}\cdot°\text{C}$, the temperature at the hot end is $T = 100\,°\text{C}$, and

the temperature at the cold end is $T = 20\ °\text{C}$.

2. Calculate the heat-conduction rate through the steel rod and the heat-conduction rate through the aluminum rod in terms of the unknown temperature T at the joint:

$$\begin{aligned} P_{\text{steel}} &= \frac{k_{\text{steel}} A_{\text{steel}} \Delta T_{\text{steel}}}{L_{\text{steel}}} \\ &= \frac{(80\ \text{W/m} \cdot °\text{C})\left(7.85 \times 10^{-5}\ \text{m}^2\right)(100\ °\text{C} - T)}{0.25\ \text{m}} \\ &= (0.0251\ \text{W/°C})(100\ °\text{C} - T); \end{aligned}$$

$$\begin{aligned} P_{\text{A1}} &= \frac{k_{\text{Al}} A_{\text{A1}} \Delta T_{\text{Al}}}{L_{\text{A1}}} \\ &= \frac{(220\ \text{W/m} \cdot °\text{C})\left(7.85 \times 10^{-5}\ \text{m}^2\right)(T - 20\ °\text{C})}{0.25\ \text{m}} \\ &= (0.0691\ \text{W/°C})(T - 20\ °\text{C}). \end{aligned}$$

3. Set the two rates equal and solve for the unknown temperature:

$$\begin{aligned} (0.0691\ \text{W/°C})(T - 20\ °\text{C}) &= (0.0251\ \text{W/°C})(100\ °\text{C} - T) \\ T &= 41.3\ °\text{C}. \end{aligned}$$

4. Calculate either rate:

$$P_{\text{steel}} = (0.0251\ \text{W/°C})(100\ °\text{C} - 41.3\ °\text{C}) = 1.47\ \text{W}.$$

5. If desired, check your answer by calculating the other rate.

Solution

1. Recall that $R = L/k$. Now $P = A\Delta T/R$, or $\Delta T = PR/A$.

2. We know that $\Delta T_{\text{steel}} + \Delta T_{\text{Al}} = 100\ °\text{C} - 20\ °\text{C} = 80\ °\text{C}$. We also know that $P_{\text{steel}} = P_{\text{Al}},$ and we denote that rate of heat flow by P. Combine the equations:

$$\frac{PR_{\text{steel}}}{A} + \frac{PR_{\text{Al}}}{A} = 80\ °\text{C}.$$

Thus, we can simply add R factors. Now, $P = \dfrac{80\ °\text{C}}{A\left(R_{\text{steel}} + R_{\text{Al}}\right)}$.

3. Find the R_s from the known quantities:

$$R_{\text{steel}} = 3.13 \times 10^{-3}\,\text{m}^2 \cdot °\text{C/W}$$

and

$$R_{\text{Al}} = 1.14 \times 10^{-3}\,\text{m}^2 \cdot °\text{C/W}.$$

4. Substitute these values in to find $P = 1.47\ \text{W}$ as before.

5. Determine ΔT for the aluminum rod (or for the steel rod) and use it to find T at the joint.

$$\Delta T_{\text{Al}} = \frac{PR_{\text{Al}}}{A} = \frac{(1.47\ \text{W})\left(1.14 \times 10^{-3}\ \text{m}^2 \cdot °\text{C/W}\right)}{7.85 \times 10^{-5}\ \text{m}^2} = 21.3\ °\text{C},$$

so $T = 20\ °\text{C} + 21.3\ °\text{C} = 41.3\ °\text{C}$, as in Solution 1.

6. If desired, check by determining ΔT for the other rod.

Significance

In practice, adding R values is common, as in calculating the R value of an insulated wall. In the analogous situation in electronics, the resistance corresponds to AR in this problem and is additive even when the areas are unequal, as is common in electronics. Our equation for heat conduction can be used only when the areas are equal; otherwise, we would have a problem in three-dimensional heat flow, which is beyond our scope.

16.6 Check Your Understanding How does the rate of heat transfer by conduction change when all spatial dimensions are doubled?

Conduction is caused by the random motion of atoms and molecules. As such, it is an ineffective mechanism for heat transport over macroscopic distances and short times. For example, the temperature on Earth would be unbearably cold during the night and extremely hot during the day if heat transport in the atmosphere were only through conduction. Also, car engines would overheat unless there was a more efficient way to remove excess heat from the pistons. The next module discusses the important heat-transfer mechanism in such situations.

Convection

In convection, thermal energy is carried by the large-scale flow of matter. It can be divided into two types. In *forced convection*, the flow is driven by fans, pumps, and the like. A simple example is a fan that blows air past you in hot surroundings and cools you by replacing the air heated by your body with cooler air. A more complicated example is the cooling system of a typical car, in which a pump moves coolant through the radiator and engine to cool the engine and a fan blows air to cool the radiator.

In *free* or *natural convection*, the flow is driven by buoyant forces: hot fluid rises and cold fluid sinks because density decreases as temperature increases. The house in **Figure 16.24** is kept warm by natural convection, as is the pot of water on the stove in **Figure 16.25**. Ocean currents and large-scale atmospheric circulation, which result from the buoyancy of warm air and water, transfer hot air from the tropics toward the poles and cold air from the poles toward the tropics. (Earth's rotation interacts with those flows, causing the observed eastward flow of air in the temperate zones.)

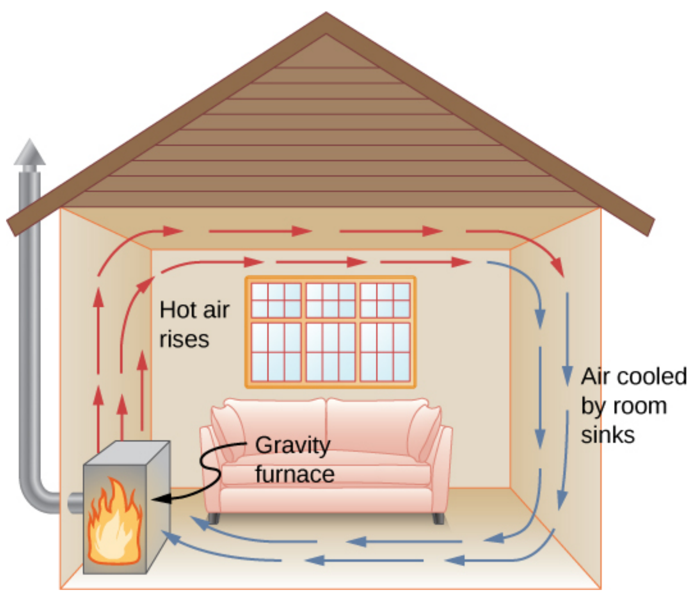


Figure 16.24 Air heated by a so-called gravity furnace expands and rises, forming a convective loop that transfers energy to other parts of the room. As the air is cooled at the ceiling and outside walls, it contracts, eventually becoming denser than room air and sinking to the floor. A properly designed heating system using natural convection, like this one, can heat a home quite efficiently.

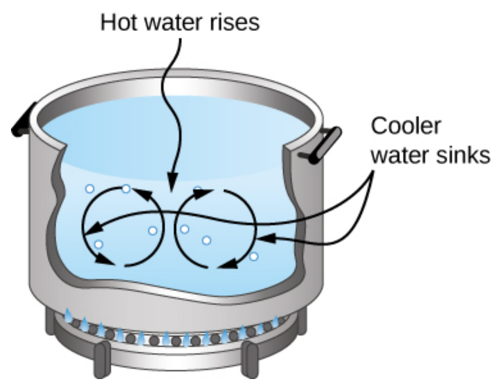


Figure 16.25 Natural convection plays an important role in heat transfer inside this pot of water. Once conducted to the inside, heat transfer to other parts of the pot is mostly by convection. The hotter water expands, decreases in density, and rises to transfer heat to other regions of the water, while colder water sinks to the bottom. This process keeps repeating.

Natural convection like that of Figure 16.24 and Figure 16.25, but acting on rock in Earth's mantle, drives plate tectonics (https://openstax.org/l/21platetecton) that are the motions that have shaped Earth's surface.

Convection is usually more complicated than conduction. Beyond noting that the convection rate is often approximately proportional to the temperature difference, we will not do any quantitative work comparable to the formula for conduction. However, we can describe convection qualitatively and relate convection rates to heat and time. Air is a poor conductor, so convection dominates heat transfer by air. Therefore, the amount of available space for airflow determines whether air transfers heat rapidly or slowly. There is little heat transfer in a space filled with air with a small amount of other material that prevents flow. The space between the inside and outside walls of a typical American house, for example, is about 9 cm (3.5 in.)—large enough for convection to work effectively. The addition of wall insulation prevents airflow, so heat loss (or gain) is decreased. On the other hand, the gap between the two panes of a double-paned window is about 1 cm, which largely prevents convection and takes advantage of air's low conductivity reduce heat loss. Fur, cloth, and fiberglass also take advantage of the low conductivity of air by trapping it in spaces too small to support convection (Figure 16.26).

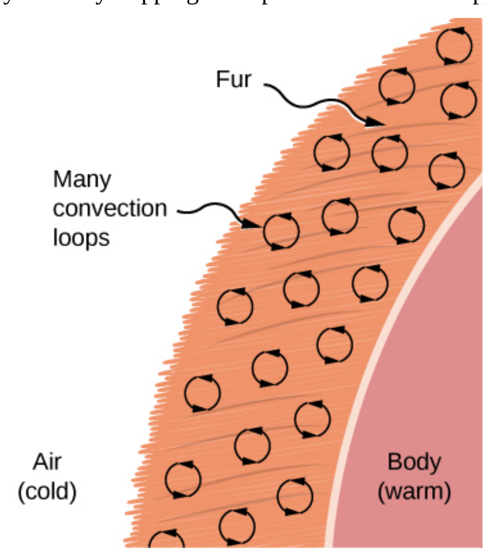


Figure 16.26 Fur is filled with air, breaking it up into many small pockets. Convection is very slow here, because the loops are so small. The low conductivity of air makes fur a very good lightweight insulator.

Some interesting phenomena happen when convection is accompanied by a phase change. The combination allows us to cool off by sweating even if the temperature of the surrounding air exceeds body temperature. Heat from the skin is required for sweat to evaporate from the skin, but without air flow, the air becomes saturated and evaporation stops. Air flow caused by convection replaces the saturated air by dry air and evaporation continues.

Example 16.10

Calculating the Flow of Mass during Convection

The average person produces heat at the rate of about 120 W when at rest. At what rate must water evaporate from the body to get rid of all this energy? (For simplicity, we assume this evaporation occurs when a person is sitting in the shade and surrounding temperatures are the same as skin temperature, eliminating heat transfer by other methods.)

Strategy

Energy is needed for this phase change ($Q = mL_v$). Thus, the energy loss per unit time is

$$\frac{Q}{t} = \frac{mL_\text{V}}{t} = 120\ \text{W} = 120\ \text{J/s}.$$

We divide both sides of the equation by L_v to find that the mass evaporated per unit time is

$$\frac{m}{t} = \frac{120\ \text{J/s}}{L_\text{v}}.$$

Solution

Insert the value of the latent heat from **Table 16.4**, $L_v = 2430\ \text{kJ/kg} = 2430\ \text{J/g}$. This yields

$$\frac{m}{t} = \frac{120\ \text{J/s}}{2430\ \text{J/g}} = 0.0494\ \text{g/s} = 2.96\ \text{g/min}.$$

Significance

Evaporating about 3 g/min seems reasonable. This would be about 180 g (about 7 oz.) per hour. If the air is very dry, the sweat may evaporate without even being noticed. A significant amount of evaporation also takes place in the lungs and breathing passages.

Another important example of the combination of phase change and convection occurs when water evaporates from the oceans. Heat is removed from the ocean when water evaporates. If the water vapor condenses in liquid droplets as clouds form, possibly far from the ocean, heat is released in the atmosphere. Thus, there is an overall transfer of heat from the ocean to the atmosphere. This process is the driving power behind thunderheads, those great cumulus clouds that rise as much as 20.0 km into the stratosphere (**Figure 16.27**). Water vapor carried in by convection condenses, releasing tremendous amounts of energy. This energy causes the air to expand and rise to colder altitudes. More condensation occurs in these regions, which in turn drives the cloud even higher. This mechanism is an example of positive feedback, since the process reinforces and accelerates itself. It sometimes produces violent storms, with lightning and hail. The same mechanism drives hurricanes.

This **time-lapse video (https://openstax.org/l/21convthuncurr)** shows convection currents in a thunderstorm, including “rolling” motion similar to that of boiling water.

Figure 16.27 Cumulus clouds are caused by water vapor that rises because of convection. The rise of clouds is driven by a positive feedback mechanism. (credit: “Amada44”/Wikimedia Commons)

16.7 **Check Your Understanding** Explain why using a fan in the summer feels refreshing.

Radiation

You can feel the heat transfer from the Sun. The space between Earth and the Sun is largely empty, so the Sun warms us without any possibility of heat transfer by convection or conduction. Similarly, you can sometimes tell that the oven is hot without touching its door or looking inside—it may just warm you as you walk by. In these examples, heat is transferred by radiation (**Figure 16.28**). That is, the hot body emits electromagnetic waves that are absorbed by the skin. No medium is required for electromagnetic waves to propagate. Different names are used for electromagnetic waves of different wavelengths: radio waves, microwaves, infrared radiation, visible light, ultraviolet radiation, X-rays, and gamma rays.

Figure 16.28 Most of the heat transfer from this fire to the observers occurs through infrared radiation. The visible light, although dramatic, transfers relatively little thermal energy. Convection transfers energy away from the observers as hot air rises, while conduction is negligibly slow here. Skin is very sensitive to infrared radiation, so you can sense the presence of a fire without looking at it directly. (credit: Daniel O’Neil)

The energy of electromagnetic radiation varies over a wide range, depending on the wavelength: A shorter wavelength (or higher frequency) corresponds to a higher energy. Because more heat is radiated at higher temperatures, higher temperatures produce more intensity at every wavelength but especially at shorter wavelengths. In visible light, wavelength determines color—red has the longest wavelength and violet the shortest—so a temperature change is accompanied by a color change. For example, an electric heating element on a stove glows from red to orange, while the higher-temperature steel in a blast furnace glows from yellow to white. Infrared radiation is the predominant form radiated by objects cooler than the electric element and the steel. The radiated energy as a function of wavelength depends on its intensity, which is represented in **Figure 16.29** by the height of the distribution. (See the section on the **Electromagnetic Spectrum**, and the section on **Blackbody Radiation**, which discusses why the decrease in wavelength corresponds to an increase in energy.)

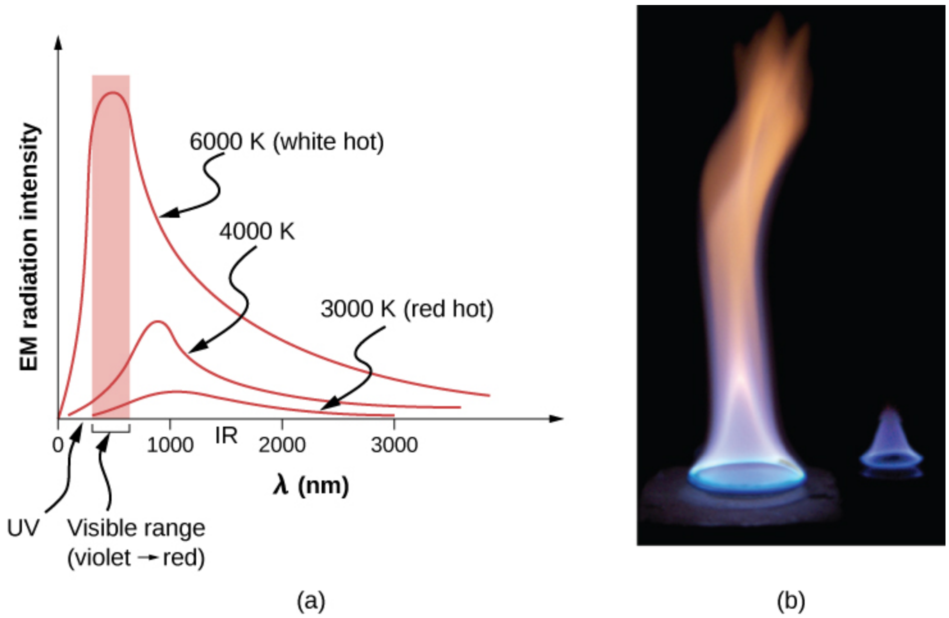


Figure 16.29 (a) A graph of the spectrum of electromagnetic waves emitted from an ideal radiator at three different temperatures. The intensity or rate of radiation emission increases dramatically with temperature, and the spectrum shifts down in wavelength toward the visible and ultraviolet parts of the spectrum. The shaded portion denotes the visible part of the spectrum. It is apparent that the shift toward the ultraviolet with temperature makes the visible appearance shift from red to white to blue as temperature increases. (b) Note the variations in color corresponding to variations in flame temperature.

The rate of heat transfer by radiation also depends on the object's color. Black is the most effective, and white is the least effective. On a clear summer day, black asphalt in a parking lot is hotter than adjacent gray sidewalk, because black absorbs better than gray (**Figure 16.30**). The reverse is also true—black radiates better than gray. Thus, on a clear summer night, the asphalt is colder than the gray sidewalk, because black radiates the energy more rapidly than gray. A perfectly black object would be an *ideal radiator* and an *ideal absorber*, as it would capture all the radiation that falls on it. In contrast, a perfectly white object or a perfect mirror would reflect all radiation, and a perfectly transparent object would transmit it all (**Figure 16.31**). Such objects would not emit any radiation. Mathematically, the color is represented by the **emissivity** *e*. A "blackbody" radiator would have an $e = 1$, whereas a perfect reflector or transmitter would have $e = 0$. For real examples, tungsten light bulb filaments have an *e* of about 0.5, and carbon black (a material used in printer toner) has an emissivity of about 0.95.

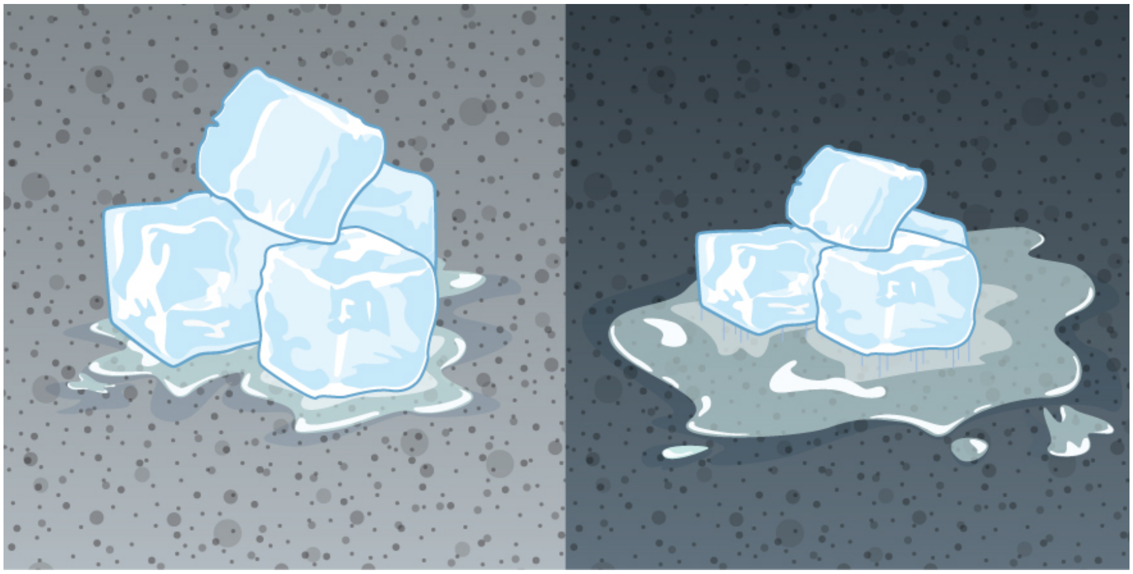


Figure 16.30 The darker pavement is hotter than the lighter pavement (much more of the ice on the right has melted), although both have been in the sunlight for the same time. The thermal conductivities of the pavements are the same.

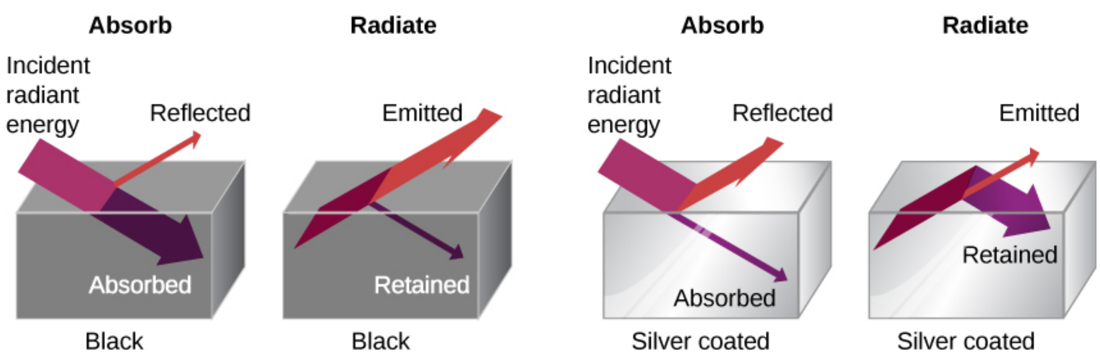


Figure 16.31 A black object is a good absorber and a good radiator, whereas a white, clear, or silver object is a poor absorber and a poor radiator.

To see that, consider a silver object and a black object that can exchange heat by radiation and are in thermal equilibrium. We know from experience that they will stay in equilibrium (the result of a principle called the Second Law of Thermodynamics). For the black object's temperature to stay constant, it must emit as much radiation as it absorbs, so it must be as good at radiating as absorbing. Similar considerations show that the silver object must radiate as little as it absorbs. Thus, one property, emissivity, controls both radiation and absorption.

Finally, the radiated heat is proportional to the object's surface area, since every part of the surface radiates. If you knock apart the coals of a fire, the radiation increases noticeably due to an increase in radiating surface area.

The rate of heat transfer by emitted radiation is described by the **Stefan-Boltzmann law of radiation**:

$$P = \sigma A e T^4,$$

where $\sigma = 5.67 \times 10^{-8} \text{ J/s} \cdot \text{m}^2 \cdot \text{K}^4$ is the Stefan-Boltzmann constant, a combination of fundamental constants of nature; A is the surface area of the object; and T is its temperature in kelvins.

The proportionality to the *fourth power* of the absolute temperature is a remarkably strong temperature dependence. It allows the detection of even small temperature variations. Images called *thermographs* can be used medically to detect regions of abnormally high temperature in the body, perhaps indicative of disease. Similar techniques can be used to detect heat leaks in homes (**Figure 16.32**), optimize performance of blast furnaces, improve comfort levels in work environments, and even remotely map Earth's temperature profile.

Figure 16.32 A thermograph of part of a building shows temperature variations, indicating where heat transfer to the outside is most severe. Windows are a major region of heat transfer to the outside of homes. (credit: US Army)

The Stefan-Boltzmann equation needs only slight refinement to deal with a simple case of an object's absorption of radiation from its surroundings. Assuming that an object with a temperature T_1 is surrounded by an environment with uniform temperature T_2, the **net rate of heat transfer by radiation** is

$$P_{\text{net}} = \sigma e A\left(T_2^{\ 4} - T_1^{\ 4}\right), \tag{16.10}$$

where *e* is the emissivity of the object alone. In other words, it does not matter whether the surroundings are white, gray, or black: The balance of radiation into and out of the object depends on how well it emits and absorbs radiation. When $T_2 > T_1,$ the quantity P_{net} is positive, that is, the net heat transfer is from hot to cold.

Before doing an example, we have a complication to discuss: different emissivities at different wavelengths. If the fraction of incident radiation an object reflects is the same at all visible wavelengths, the object is gray; if the fraction depends on the wavelength, the object has some other color. For instance, a red or reddish object reflects red light more strongly than other visible wavelengths. Because it absorbs less red, it radiates less red when hot. Differential reflection and absorption of wavelengths outside the visible range have no effect on what we see, but they may have physically important effects. Skin is a very good absorber and emitter of infrared radiation, having an emissivity of 0.97 in the infrared spectrum. Thus, in spite of the obvious variations in skin color, we are all nearly black in the infrared. This high infrared emissivity is why we can so easily feel radiation on our skin. It is also the basis for the effectiveness of night-vision scopes used by law enforcement and the military to detect human beings.

Example 16.11

Calculating the Net Heat Transfer of a Person

What is the rate of heat transfer by radiation of an unclothed person standing in a dark room whose ambient temperature is $22.0\ °\text{C}$? The person has a normal skin temperature of $33.0\ °\text{C}$ and a surface area of $1.50\ \text{m}^2$. The emissivity of skin is 0.97 in the infrared, the part of the spectrum where the radiation takes place.

Strategy

We can solve this by using the equation for the rate of radiative heat transfer.

Solution

Insert the temperature values $T_2 = 295\text{ K}$ and $T_1 = 306\text{ K}$, so that

$$\begin{aligned}\frac{Q}{t} &= \sigma e A\left(T_2{}^4 - T_1{}^4\right) \\ &= \left(5.67 \times 10^{-8}\text{ J/s}\cdot\text{m}^2\cdot\text{K}^4\right)(0.97)\left(1.50\text{ m}^2\right)\left[(295\text{ K})^4 - (306\text{ K})^4\right] \\ &= -99\text{ J/s} = -99\text{ W}.\end{aligned}$$

Significance

This value is a significant rate of heat transfer to the environment (note the minus sign), considering that a person at rest may produce energy at the rate of 125 W and that conduction and convection are also transferring energy to the environment. Indeed, we would probably expect this person to feel cold. Clothing significantly reduces heat transfer to the environment by all mechanisms, because clothing slows down both conduction and convection, and has a lower emissivity (especially if it is light-colored) than skin.

The average temperature of Earth is the subject of much current discussion. Earth is in radiative contact with both the Sun and dark space, so we cannot use the equation for an environment at a uniform temperature. Earth receives almost all its energy from radiation of the Sun and reflects some of it back into outer space. Conversely, dark space is very cold, about 3 K, so that Earth radiates energy into the dark sky. The rate of heat transfer from soil and grasses can be so rapid that frost may occur on clear summer evenings, even in warm latitudes.

The average temperature of Earth is determined by its energy balance. To a first approximation, it is the temperature at which Earth radiates heat to space as fast as it receives energy from the Sun.

An important parameter in calculating the temperature of Earth is its emissivity (e). On average, it is about 0.65, but calculation of this value is complicated by the great day-to-day variation in the highly reflective cloud coverage. Because clouds have lower emissivity than either oceans or land masses, they reflect some of the radiation back to the surface, greatly reducing heat transfer into dark space, just as they greatly reduce heat transfer into the atmosphere during the day. There is negative feedback (in which a change produces an effect that opposes that change) between clouds and heat transfer; higher temperatures evaporate more water to form more clouds, which reflect more radiation back into space, reducing the temperature.

The often-mentioned **greenhouse effect** is directly related to the variation of Earth's emissivity with wavelength (**Figure 16.33**). The greenhouse effect is a natural phenomenon responsible for providing temperatures suitable for life on Earth and for making Venus unsuitable for human life. Most of the infrared radiation emitted from Earth is absorbed by carbon dioxide (CO_2) and water (H_2O) in the atmosphere and then re-radiated into outer space or back to Earth. Re-radiation back to Earth maintains its surface temperature about $40\ °\text{C}$ higher than it would be if there were no atmosphere. (The glass walls and roof of a greenhouse increase the temperature inside by blocking convective heat losses, not radiative losses.)

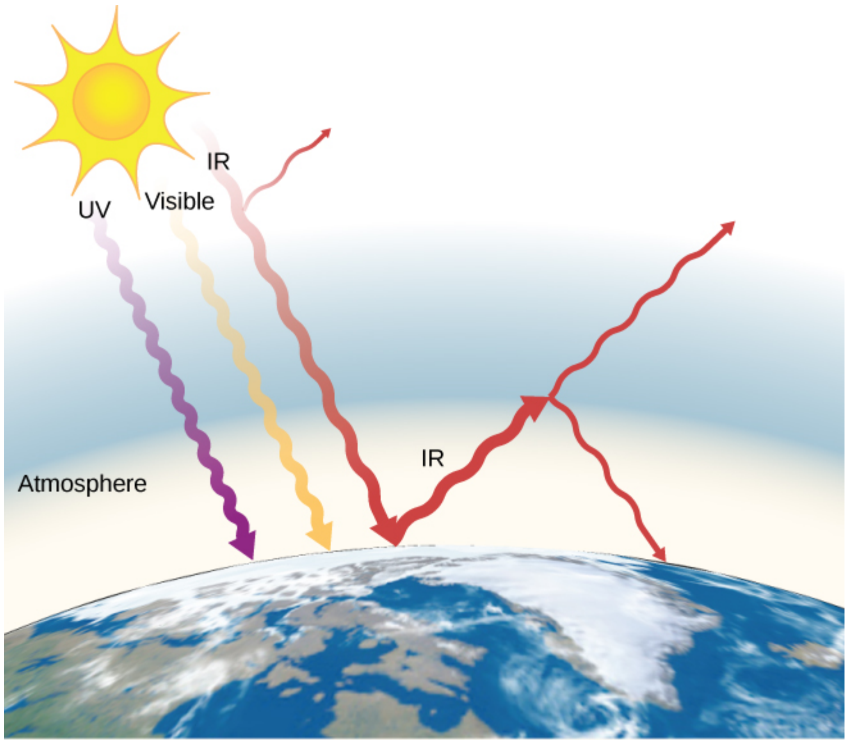


Figure 16.33 The greenhouse effect is the name given to the increase of Earth's temperature due to absorption of radiation in the atmosphere. The atmosphere is transparent to incoming visible radiation and most of the Sun's infrared. The Earth absorbs that energy and re-emits it. Since Earth's temperature is much lower than the Sun's, it re-emits the energy at much longer wavelengths, in the infrared. The atmosphere absorbs much of that infrared radiation and radiates about half of the energy back down, keeping Earth warmer than it would otherwise be. The amount of trapping depends on concentrations of trace gases such as carbon dioxide, and an increase in the concentration of these gases increases Earth's surface temperature.

The greenhouse effect is central to the discussion of global warming due to emission of carbon dioxide and methane (and other greenhouse gases) into Earth's atmosphere from industry, transportation, and farming. Changes in global climate could lead to more intense storms, precipitation changes (affecting agriculture), reduction in rain forest biodiversity, and rising sea levels.

You can explore **a simulation of the greenhouse effect (https://openstax.org/l/21simgreeneff)** that takes the point of view that the atmosphere scatters (redirects) infrared radiation rather than absorbing it and reradiating it. You may want to run the simulation first with no greenhouse gases in the atmosphere and then look at how adding greenhouse gases affects the infrared radiation from the Earth and the Earth's temperature.

Problem-Solving Strategy: Effects of Heat Transfer

1. Examine the situation to determine what type of heat transfer is involved.
2. Identify the type(s) of heat transfer—conduction, convection, or radiation.
3. Identify exactly what needs to be determined in the problem (identify the unknowns). A written list is useful.
4. Make a list of what is given or what can be inferred from the problem as stated (identify the knowns).
5. Solve the appropriate equation for the quantity to be determined (the unknown).
6. For conduction, use the equation $P = \frac{kA\Delta T}{d}$. **Table 16.5** lists thermal conductivities. For convection, determine the amount of matter moved and the equation $Q = mc\Delta T$, along with $Q = mL_f$ or $Q = mL_V$ if a

substance changes phase. For radiation, the equation $P_{\text{net}} = \sigma e A\left(T_2{}^4 - T_1{}^4\right)$ gives the net heat transfer rate.

7. Substitute the knowns along with their units into the appropriate equation and obtain numerical solutions complete with units.
8. Check the answer to see if it is reasonable. Does it make sense?

16.8 Check Your Understanding How much greater is the rate of heat radiation when a body is at the temperature $40\ °\text{C}$ than when it is at the temperature $20\ °\text{C}$?

CHAPTER 16 REVIEW

KEY TERMS

absolute temperature scale scale, such as Kelvin, with a zero point that is absolute zero

absolute zero temperature at which the average kinetic energy of molecules is zero

calorie (cal) energy needed to change the temperature of 1.00 g of water by $1.00\ °\text{C}$

calorimetry study of heat transfer inside a container impervious to heat

Celsius scale temperature scale in which the freezing point of water is $0\ °\text{C}$ and the boiling point of water is $100\ °\text{C}$

coefficient of linear expansion (α) material property that gives the change in length, per unit length, per $1\text{-}°\text{C}$ change in temperature; a constant used in the calculation of linear expansion; the coefficient of linear expansion depends to some degree on the temperature of the material

coefficient of volume expansion (β) similar to α but gives the change in volume, per unit volume, per $1\text{-}°\text{C}$ change in temperature

conduction heat transfer through stationary matter by physical contact

convection heat transfer by the macroscopic movement of fluid

critical point for a given substance, the combination of temperature and pressure above which the liquid and gas phases are indistinguishable

critical pressure pressure at the critical point

critical temperature temperature at the critical point

degree Celsius ($°\text{C}$) unit on the Celsius temperature scale

degree Fahrenheit ($°\text{F}$) unit on the Fahrenheit temperature scale

emissivity measure of how well an object radiates

Fahrenheit scale temperature scale in which the freezing point of water is $32\ °\text{F}$ and the boiling point of water is $212\ °\text{F}$

greenhouse effect warming of the earth that is due to gases such as carbon dioxide and methane that absorb infrared radiation from Earth's surface and reradiate it in all directions, thus sending some of it back toward Earth

heat energy transferred solely due to a temperature difference

heat of fusion energy per unit mass required to change a substance from the solid phase to the liquid phase, or released when the substance changes from liquid to solid

heat of sublimation energy per unit mass required to change a substance from the solid phase to the vapor phase

heat of vaporization energy per unit mass required to change a substance from the liquid phase to the vapor phase

heat transfer movement of energy from one place or material to another as a result of a difference in temperature

Kelvin scale (K) temperature scale in which 0 K is the lowest possible temperature, representing absolute zero

kilocalorie (kcal) energy needed to change the temperature of 1.00 kg of water between $14.5\ °\text{C}$ and $15.5\ °\text{C}$

latent heat coefficient general term for the heats of fusion, vaporization, and sublimation

mechanical equivalent of heat work needed to produce the same effects as heat transfer

net rate of heat transfer by radiation $P_{\text{net}} = \sigma e A\left(T_2{}^4 - T_1{}^4\right)$

phase diagram graph of pressure vs. temperature of a particular substance, showing at which pressures and temperatures the phases of the substance occur

radiation energy transferred by electromagnetic waves directly as a result of a temperature difference

rate of conductive heat transfer rate of heat transfer from one material to another

specific heat amount of heat necessary to change the temperature of 1.00 kg of a substance by $1.00\ °C$; also called "specific heat capacity"

Stefan-Boltzmann law of radiation $P = \sigma A e T^4$, where $\sigma = 5.67 \times 10^{-8}\ \text{J/s} \cdot \text{m}^2 \cdot \text{K}^4$ is the Stefan-Boltzmann constant, A is the surface area of the object, T is the absolute temperature, and e is the emissivity

sublimation phase change from solid to gas

temperature quantity measured by a thermometer, which reflects the mechanical energy of molecules in a system

thermal conductivity property of a material describing its ability to conduct heat

thermal equilibrium condition in which heat no longer flows between two objects that are in contact; the two objects have the same temperature

thermal expansion change in size or volume of an object with change in temperature

triple point pressure and temperature at which a substance exists in equilibrium as a solid, liquid, and gas

vapor gas at a temperature below the boiling temperature

vapor pressure pressure at which a gas coexists with its solid or liquid phase

zeroth law of thermodynamics law that states that if two objects are in thermal equilibrium, and a third object is in thermal equilibrium with one of those objects, it is also in thermal equilibrium with the other object

KEY EQUATIONS

Linear thermal expansion	$\Delta L = \alpha L \Delta T$
Thermal expansion in two dimensions	$\Delta A = 2\alpha A \Delta T$
Thermal expansion in three dimensions	$\Delta V = \beta V \Delta T$
Heat transfer	$Q = mc\Delta T$
Transfer of heat in a calorimeter	$Q_{\text{cold}} + Q_{\text{hot}} = 0$
Heat due to phase change (melting and freezing)	$Q = mL_{\text{f}}$
Heat due to phase change (evaporation and condensation)	$Q = mL_{\text{v}}$
Rate of conductive heat transfer	$P = \dfrac{kA(T_h - T_c)}{d}$
Net rate of heat transfer by radiation	$P_{\text{net}} = \sigma e A\left(T_2{}^4 - T_1{}^4\right)$

SUMMARY

16.1 Temperature and Thermal Equilibrium

- Temperature is operationally defined as the quantity measured by a thermometer. It is proportional to the average kinetic energy of atoms and molecules in a system.
- Thermal equilibrium occurs when two bodies are in contact with each other and can freely exchange energy. Systems are in thermal equilibrium when they have the same temperature.
- The zeroth law of thermodynamics states that when two systems, A and B, are in thermal equilibrium with each other, and B is in thermal equilibrium with a third system C, then A is also in thermal equilibrium with C.

16.2 Thermometers and Temperature Scales

- Three types of thermometers are alcohol, liquid crystal, and infrared radiation (pyrometer).
- The three main temperature scales are Celsius, Fahrenheit, and Kelvin. Temperatures can be converted from one scale to another using temperature conversion equations.
- The three phases of water (ice, liquid water, and water vapor) can coexist at a single pressure and temperature known as the triple point.

16.3 Thermal Expansion

- Thermal expansion is the increase of the size (length, area, or volume) of a body due to a change in temperature, usually a rise. Thermal contraction is the decrease in size due to a change in temperature, usually a fall in temperature.

16.4 Heat Transfer and Specific Heat

- Heat and work are the two distinct methods of energy transfer.
- Heat transfer to an object when its temperature changes is often approximated well by $Q = mc\Delta T,$ where *m* is the object's mass and *c* is the specific heat of the substance.

16.5 Phase Changes

- Most substances have three distinct phases (under ordinary conditions on Earth), and they depend on temperature and pressure.
- Two phases coexist (i.e., they are in thermal equilibrium) at a set of pressures and temperatures.
- Phase changes occur at fixed temperatures for a given substance at a given pressure, and these temperatures are called boiling, freezing (or melting), and sublimation points.

16.6 Mechanisms of Heat Transfer

- Heat is transferred by three different methods: conduction, convection, and radiation.
- Heat conduction is the transfer of heat between two objects in direct contact with each other.
- The rate of heat transfer *P* (energy per unit time) is proportional to the temperature difference $T_\text{h} - T_\text{c}$ and the contact area *A* and inversely proportional to the distance *d* between the objects.
- Convection is heat transfer by the macroscopic movement of mass. Convection can be natural or forced, and generally transfers thermal energy faster than conduction. Convection that occurs along with a phase change can transfer energy from cold regions to warm ones.
- Radiation is heat transfer through the emission or absorption of electromagnetic waves.
- The rate of radiative heat transfer is proportional to the emissivity *e*. For a perfect blackbody, $e = 1$, whereas a perfectly white, clear, or reflective body has $e = 0$, with real objects having values of *e* between 1 and 0.
- The rate of heat transfer depends on the surface area and the fourth power of the absolute temperature:

$$P = \sigma e A T^4,$$

where $\sigma = 5.67 \times 10^{-8}\ \text{J/s} \cdot \text{m}^2 \cdot \text{K}^4$ is the Stefan-Boltzmann constant and *e* is the emissivity of the body. The net rate of heat transfer from an object by radiation is

$$\frac{Q_\text{net}}{t} = \sigma e A \left(T_2^{\ 4} - T_1^{\ 4}\right),$$

where T_1 is the temperature of the object surrounded by an environment with uniform temperature T_2 and *e* is the emissivity of the object.

CONCEPTUAL QUESTIONS

16.1 Temperature and Thermal Equilibrium

1. What does it mean to say that two systems are in thermal equilibrium?

2. Give an example in which *A* has some kind of non-thermal equilibrium relationship with *B*, and *B* has the same relationship with *C*, but *A* does not have that relationship with *C*.

16.2 Thermometers and Temperature Scales

3. If a thermometer is allowed to come to equilibrium with the air, and a glass of water is not in equilibrium with the air, what will happen to the thermometer reading when it is placed in the water?

4. Give an example of a physical property that varies with temperature and describe how it is used to measure temperature.

16.3 Thermal Expansion

5. One method of getting a tight fit, say of a metal peg in a hole in a metal block, is to manufacture the peg slightly larger than the hole. The peg is then inserted when at a different temperature than the block. Should the block be hotter or colder than the peg during insertion? Explain your answer.

6. Does it really help to run hot water over a tight metal lid on a glass jar before trying to open it? Explain your answer.

7. When a cold alcohol thermometer is placed in a hot liquid, the column of alcohol goes *down* slightly before going up. Explain why.

8. Calculate the length of a 1-meter rod of a material with thermal expansion coefficient α when the temperature is raised from 300 K to 600 K. Taking your answer as the new initial length, find the length after the rod is cooled back down to 300 K. Is your answer 1 meter? Should it be? How can you account for the result you got?

16.4 Heat Transfer and Specific Heat

9. How is heat transfer related to temperature?

10. Describe a situation in which heat transfer occurs.

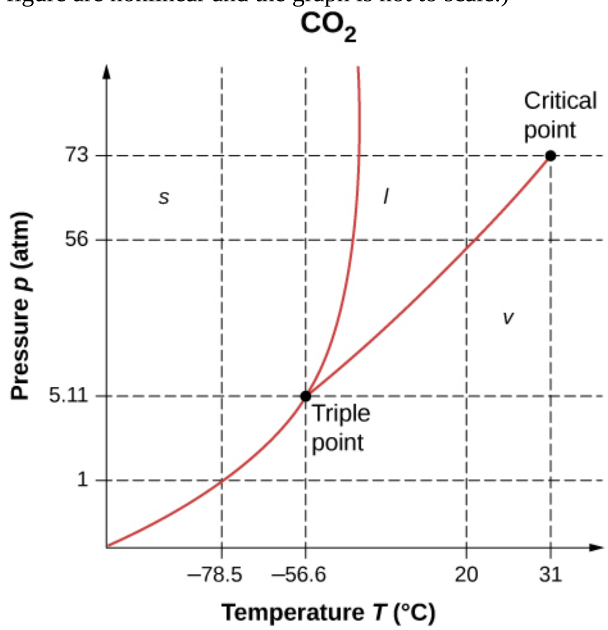
11. When heat transfers into a system, is the energy stored as heat? Explain briefly.

12. The brakes in a car increase in temperature by ΔT when bringing the car to rest from a speed *v*. How much greater would ΔT be if the car initially had twice the speed? You may assume the car stops fast enough that no heat transfers out of the brakes.

16.5 Phase Changes

13. A pressure cooker contains water and steam in equilibrium at a pressure greater than atmospheric pressure. How does this greater pressure increase cooking speed?

14. As shown below, which is the phase diagram for carbon dioxide, what is the vapor pressure of solid carbon dioxide (dry ice) at $-78.5\ °C$? (Note that the axes in the figure are nonlinear and the graph is not to scale.)



15. Can carbon dioxide be liquefied at room temperature ($20\ °C$)? If so, how? If not, why not? (See the phase diagram in the preceding problem.)

16. What is the distinction between gas and vapor?

17. Heat transfer can cause temperature and phase changes. What else can cause these changes?

18. How does the latent heat of fusion of water help slow the decrease of air temperatures, perhaps preventing temperatures from falling significantly below $0\ °C$, in the vicinity of large bodies of water?

19. What is the temperature of ice right after it is formed by freezing water?

20. If you place $0\,°\text{C}$ ice into $0\,°\text{C}$ water in an insulated container, what will the net result be? Will there be less ice and more liquid water, or more ice and less liquid water, or will the amounts stay the same?

21. What effect does condensation on a glass of ice water have on the rate at which the ice melts? Will the condensation speed up the melting process or slow it down?

22. In Miami, Florida, which has a very humid climate and numerous bodies of water nearby, it is unusual for temperatures to rise above about $38\,°\text{C}$ ($100\,°\text{F}$). In the desert climate of Phoenix, Arizona, however, temperatures rise above that almost every day in July and August. Explain how the evaporation of water helps limit high temperatures in humid climates.

23. In winter, it is often warmer in San Francisco than in Sacramento, 150 km inland. In summer, it is nearly always hotter in Sacramento. Explain how the bodies of water surrounding San Francisco moderate its extreme temperatures.

24. Freeze-dried foods have been dehydrated in a vacuum. During the process, the food freezes and must be heated to facilitate dehydration. Explain both how the vacuum speeds up dehydration and why the food freezes as a result.

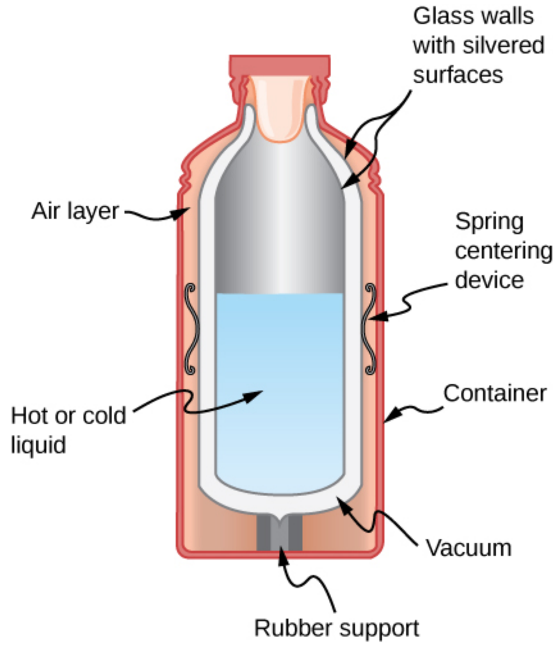
25. In a physics classroom demonstration, an instructor inflates a balloon by mouth and then cools it in liquid nitrogen. When cold, the shrunken balloon has a small amount of light blue liquid in it, as well as some snow-like crystals. As it warms up, the liquid boils, and part of the crystals sublime, with some crystals lingering for a while and then producing a liquid. Identify the blue liquid and the two solids in the cold balloon. Justify your identifications using data from **Table 16.4**.

16.6 Mechanisms of Heat Transfer

26. What are the main methods of heat transfer from the hot core of Earth to its surface? From Earth's surface to outer space?

27. When our bodies get too warm, they respond by sweating and increasing blood circulation to the surface to transfer thermal energy away from the core. What effect will those processes have on a person in a $40.0\text{-}°\text{C}$ hot tub?

28. Shown below is a cut-away drawing of a thermos bottle (also known as a Dewar flask), which is a device designed specifically to slow down all forms of heat transfer. Explain the functions of the various parts, such as the vacuum, the silvering of the walls, the thin-walled long glass neck, the rubber support, the air layer, and the stopper.



29. Some electric stoves have a flat ceramic surface with heating elements hidden beneath. A pot placed over a heating element will be heated, while the surface only a few centimeters away is safe to touch. Why is ceramic, with a conductivity less than that of a metal but greater than that of a good insulator, an ideal choice for the stove top?

30. Loose-fitting white clothing covering most of the body, shown below, is ideal for desert dwellers, both in the hot Sun and during cold evenings. Explain how such clothing is advantageous during both day and night.

31. One way to make a fireplace more energy-efficient is to have room air circulate around the outside of the fire box and back into the room. Detail the methods of heat transfer involved.

32. On cold, clear nights horses will sleep under the cover of large trees. How does this help them keep warm?

33. When watching a circus during the day in a large, dark-colored tent, you sense significant heat transfer from the tent. Explain why this occurs.

34. Satellites designed to observe the radiation from cold (3 K) dark space have sensors that are shaded from the Sun, Earth, and the Moon and are cooled to very low temperatures. Why must the sensors be at low temperature?

35. Why are thermometers that are used in weather stations shielded from the sunshine? What does a thermometer measure if it is shielded from the sunshine? What does it measure if it is not?

36. Putting a lid on a boiling pot greatly reduces the heat transfer necessary to keep it boiling. Explain why.

37. Your house will be empty for a while in cold weather, and you want to save energy and money. Should you turn the thermostat down to the lowest level that will protect the house from damage such as freezing pipes, or leave it at the normal temperature? (If you don't like coming back to a cold house, imagine that a timer controls the heating system so the house will be warm when you get back.) Explain your answer.

38. You pour coffee into an unlidded cup, intending to drink it 5 minutes later. You can add cream when you pour the cup or right before you drink it. (The cream is at the same temperature either way. Assume that the cream and coffee come into thermal equilibrium with each other very quickly.) Which way will give you hotter coffee? What feature of this question is different from the previous one?

39. Broiling is a method of cooking by radiation, which produces somewhat different results from cooking by conduction or convection. A gas flame or electric heating element produces a very high temperature close to the food and *above* it. Why is radiation the dominant heat-transfer method in this situation?

40. On a cold winter morning, why does the metal of a bike feel colder than the wood of a porch?

PROBLEMS

16.2 Thermometers and Temperature Scales

41. While traveling outside the United States, you feel sick. A companion gets you a thermometer, which says your temperature is 39. What scale is that on? What is your Fahrenheit temperature? Should you seek medical help?

42. What are the following temperatures on the Kelvin scale?

(a) $68.0\,°\text{F},$ an indoor temperature sometimes recommended for energy conservation in winter

(b) $134\,°\text{F},$ one of the highest atmospheric temperatures ever recorded on Earth (Death Valley, California, 1913)

(c) $9890\,°\text{F},$ the temperature of the surface of the Sun

43. (a) Suppose a cold front blows into your locale and drops the temperature by 40.0 Fahrenheit degrees. How many degrees Celsius does the temperature decrease when it decreases by $40.0\,°\text{F}$? (b) Show that any change in temperature in Fahrenheit degrees is nine-fifths the change in Celsius degrees

44. An Associated Press article on climate change said, "Some of the ice shelf's disappearance was probably during times when the planet was 36 degrees Fahrenheit (2 degrees Celsius) to 37 degrees Fahrenheit (3 degrees Celsius) warmer than it is today." What mistake did the reporter make?

45. (a) At what temperature do the Fahrenheit and Celsius scales have the same numerical value? (b) At what temperature do the Fahrenheit and Kelvin scales have the same numerical value?

46. A person taking a reading of the temperature in a freezer in Celsius makes two mistakes: first omitting the negative sign and then thinking the temperature is Fahrenheit. That is, the person reads $-x\,°\text{C}$ as $x\,°\text{F}$. Oddly enough, the result is the correct Fahrenheit temperature. What is the original Celsius reading? Round your answer to three significant figures.

16.3 Thermal Expansion

47. The height of the Washington Monument is measured to be 170.00 m on a day when the temperature is $35.0\,°\text{C}$. What will its height be on a day when the temperature falls to $-10.0\,°\text{C}$? Although the monument is made of limestone, assume that its coefficient of thermal expansion is the same as that of marble. Give your answer to five significant figures.

48. How much taller does the Eiffel Tower become at the end of a day when the temperature has increased by $15\ °C$? Its original height is 321 m and you can assume it is made of steel.

49. What is the change in length of a 3.00-cm-long column of mercury if its temperature changes from $37.0\ °C$ to $40.0\ °C$, assuming the mercury is constrained to a cylinder but unconstrained in length? Your answer will show why thermometers contain bulbs at the bottom instead of simple columns of liquid.

50. How large an expansion gap should be left between steel railroad rails if they may reach a maximum temperature $35.0\ °C$ greater than when they were laid? Their original length is 10.0 m.

51. You are looking to buy a small piece of land in Hong Kong. The price is "only" $60,000 per square meter. The land title says the dimensions are $20\ \text{m} \times 30\ \text{m}$. By how much would the total price change if you measured the parcel with a steel tape measure on a day when the temperature was $20\ °C$ above the temperature that the tape measure was designed for? The dimensions of the land do not change.

52. Global warming will produce rising sea levels partly due to melting ice caps and partly due to the expansion of water as average ocean temperatures rise. To get some idea of the size of this effect, calculate the change in length of a column of water 1.00 km high for a temperature increase of $1.00\ °C$. Assume the column is not free to expand sideways. As a model of the ocean, that is a reasonable approximation, as only parts of the ocean very close to the surface can expand sideways onto land, and only to a limited degree. As another approximation, neglect the fact that ocean warming is not uniform with depth.

53. (a) Suppose a meter stick made of steel and one made of aluminum are the same length at $0\ °C$. What is their difference in length at $22.0\ °C$? (b) Repeat the calculation for two 30.0-m-long surveyor's tapes.

54. (a) If a 500-mL glass beaker is filled to the brim with ethyl alcohol at a temperature of $5.00\ °C$, how much will overflow when the alcohol's temperature reaches the room temperature of $22.0\ °C$? (b) How much less water would overflow under the same conditions?

55. Most cars have a coolant reservoir to catch radiator fluid that may overflow when the engine is hot. A radiator is made of copper and is filled to its 16.0-L capacity when at $10.0\ °C$. What volume of radiator fluid will overflow when the radiator and fluid reach a temperature of $95.0\ °C$, given that the fluid's volume coefficient of expansion is $\beta = 400 \times 10^{-6}/°C$? (Your answer will be a conservative estimate, as most car radiators have operating temperatures greater than $95.0\ °C$).

56. A physicist makes a cup of instant coffee and notices that, as the coffee cools, its level drops 3.00 mm in the glass cup. Show that this decrease cannot be due to thermal contraction by calculating the decrease in level if the $350\ \text{cm}^3$ of coffee is in a 7.00-cm-diameter cup and decreases in temperature from $95.0\ °C$ to $45.0\ °C$. (Most of the drop in level is actually due to escaping bubbles of air.)

57. Show that $\beta = 3\alpha$, by calculating the infinitesimal change in volume dV of a cube with sides of length L when the temperature changes by dT.

16.4 Heat Transfer and Specific Heat

58. On a hot day, the temperature of an 80,000-L swimming pool increases by $1.50\ °C$. What is the net heat transfer during this heating? Ignore any complications, such as loss of water by evaporation.

59. To sterilize a 50.0-g glass baby bottle, we must raise its temperature from $22.0\ °C$ to $95.0\ °C$. How much heat transfer is required?

60. The same heat transfer into identical masses of different substances produces different temperature changes. Calculate the final temperature when 1.00 kcal of heat transfers into 1.00 kg of the following, originally at $20.0\ °C$: (a) water; (b) concrete; (c) steel; and (d) mercury.

61. Rubbing your hands together warms them by converting work into thermal energy. If a woman rubs her hands back and forth for a total of 20 rubs, at a distance of 7.50 cm per rub, and with an average frictional force of 40.0 N, what is the temperature increase? The mass of tissues warmed is only 0.100 kg, mostly in the palms and fingers.

62. A 0.250-kg block of a pure material is heated from $20.0\ °C$ to $65.0\ °C$ by the addition of 4.35 kJ of energy. Calculate its specific heat and identify the substance of which it is most likely composed.

63. Suppose identical amounts of heat transfer into different masses of copper and water, causing identical changes in temperature. What is the ratio of the mass of copper to water?

64. (a) The number of kilocalories in food is determined by calorimetry techniques in which the food is burned and the amount of heat transfer is measured. How many kilocalories per gram are there in a 5.00-g peanut if the energy from burning it is transferred to 0.500 kg of water held in a 0.100-kg aluminum cup, causing a $54.9\text{-}°\text{C}$ temperature increase? Assume the process takes place in an ideal calorimeter, in other words a perfectly insulated container. (b) Compare your answer to the following labeling information found on a package of dry roasted peanuts: a serving of 33 g contains 200 calories. Comment on whether the values are consistent.

65. Following vigorous exercise, the body temperature of an 80.0 kg person is $40.0\ °\text{C}$. At what rate in watts must the person transfer thermal energy to reduce the body temperature to $37.0\ °\text{C}$ in 30.0 min, assuming the body continues to produce energy at the rate of 150 W? $(1\text{ watt} = 1\text{ joule/second or } 1\text{ W} = 1\text{ J/s})$

66. In a study of healthy young men[1], doing 20 push-ups in 1 minute burned an amount of energy per kg that for a 70.0-kg man corresponds to 8.06 calories (kcal). How much would a 70.0-kg man's temperature rise if he did not lose any heat during that time?

67. A 1.28-kg sample of water at $10.0\ °\text{C}$ is in a calorimeter. You drop a piece of steel with a mass of 0.385 kg at $215\ °\text{C}$ into it. After the sizzling subsides, what is the final equilibrium temperature? (Make the reasonable assumptions that any steam produced condenses into liquid water during the process of equilibration and that the evaporation and condensation don't affect the outcome, as we'll see in the next section.)

68. Repeat the preceding problem, assuming the water is in a glass beaker with a mass of 0.200 kg, which turns it into a calorimeter. The beaker is initially at the same temperature as the water. Before doing the problem, should the answer be higher or lower than the preceding answer? Comparing the mass and specific heat of the beaker to those of the water, do you think the beaker will make much difference?

16.5 Phase Changes

69. How much heat transfer (in kilocalories) is required to thaw a 0.450-kg package of frozen vegetables originally at $0\ °\text{C}$ if their heat of fusion is the same as that of water?

70. A bag containing $0\ °\text{C}$ ice is much more effective in absorbing energy than one containing the same amount of $0\ °\text{C}$ water. (a) How much heat transfer is necessary to raise the temperature of 0.800 kg of water from $0\ °\text{C}$ to $30.0\ °\text{C}$? (b) How much heat transfer is required to first melt 0.800 kg of $0\ °\text{C}$ ice and then raise its temperature? (c) Explain how your answer supports the contention that the ice is more effective.

71. (a) How much heat transfer is required to raise the temperature of a 0.750-kg aluminum pot containing 2.50 kg of water from $30.0\ °\text{C}$ to the boiling point and then boil away 0.750 kg of water? (b) How long does this take if the rate of heat transfer is 500 W?

72. Condensation on a glass of ice water causes the ice to melt faster than it would otherwise. If 8.00 g of vapor condense on a glass containing both water and 200 g of ice, how many grams of the ice will melt as a result? Assume no other heat transfer occurs. Use L_v for water at $37\ °\text{C}$ as a better approximation than L_v for water at $100\ °\text{C}$.)

73. On a trip, you notice that a 3.50-kg bag of ice lasts an average of one day in your cooler. What is the average power in watts entering the ice if it starts at $0\ °\text{C}$ and completely melts to $0\ °\text{C}$ water in exactly one day?

74. On a certain dry sunny day, a swimming pool's temperature would rise by $1.50\ °\text{C}$ if not for evaporation. What fraction of the water must evaporate to carry away precisely enough energy to keep the temperature constant?

75. (a) How much heat transfer is necessary to raise the temperature of a 0.200-kg piece of ice from $-20.0\ °\text{C}$ to $130.0\ °\text{C}$, including the energy needed for phase changes? (b) How much time is required for each stage, assuming a constant 20.0 kJ/s rate of heat transfer? (c) Make a graph of temperature versus time for this process.

76. In 1986, an enormous iceberg broke away from the Ross Ice Shelf in Antarctica. It was an approximately rectangular prism 160 km long, 40.0 km wide, and 250 m thick. (a) What is the mass of this iceberg, given that the density of ice is 917 kg/m^3? (b) How much heat transfer (in joules) is needed to melt it? (c) How many years would it take sunlight alone to melt ice this thick, if the ice absorbs an average of 100 W/m^2, 12.00 h per day?

77. How many grams of coffee must evaporate from 350 g of coffee in a 100-g glass cup to cool the coffee and the cup from $95.0\ °\text{C}$ to $45.0\ °\text{C}$? Assume the coffee has the same thermal properties as water and that the average heat

1. JW Vezina, "An examination of the differences between two methods of estimating energy expenditure in resistance training activities," *Journal of Strength and Conditioning Research*, April 28, 2014, http://www.ncbi.nlm.nih.gov/pubmed/24402448

of vaporization is 2340 kJ/kg (560 kcal/g). Neglect heat losses through processes other than evaporation, as well as the change in mass of the coffee as it cools. Do the latter two assumptions cause your answer to be higher or lower than the true answer?

78. (a) It is difficult to extinguish a fire on a crude oil tanker, because each liter of crude oil releases 2.80×10^7 J of energy when burned. To illustrate this difficulty, calculate the number of liters of water that must be expended to absorb the energy released by burning 1.00 L of crude oil, if the water's temperature rises from $20.0\,^\circ\text{C}$ to $100\,^\circ\text{C}$, it boils, and the resulting steam's temperature rises to $300\,^\circ\text{C}$ at constant pressure. (b) Discuss additional complications caused by the fact that crude oil is less dense than water.

79. The energy released from condensation in thunderstorms can be very large. Calculate the energy released into the atmosphere for a small storm of radius 1 km, assuming that 1.0 cm of rain is precipitated uniformly over this area.

80. To help prevent frost damage, 4.00 kg of water at $0\,^\circ\text{C}$ is sprayed onto a fruit tree. (a) How much heat transfer occurs as the water freezes? (b) How much would the temperature of the 200-kg tree decrease if this amount of heat transferred from the tree? Take the specific heat to be $3.35\text{ kJ/kg}\cdot{}^\circ\text{C}$, and assume that no phase change occurs in the tree.

81. A 0.250-kg aluminum bowl holding 0.800 kg of soup at $25.0\,^\circ\text{C}$ is placed in a freezer. What is the final temperature if 388 kJ of energy is transferred from the bowl and soup, assuming the soup's thermal properties are the same as that of water?

82. A 0.0500-kg ice cube at $-30.0\,^\circ\text{C}$ is placed in 0.400 kg of $35.0\text{-}^\circ\text{C}$ water in a very well-insulated container. What is the final temperature?

83. If you pour 0.0100 kg of $20.0\,^\circ\text{C}$ water onto a 1.20-kg block of ice (which is initially at $-15.0\,^\circ\text{C}$), what is the final temperature? You may assume that the water cools so rapidly that effects of the surroundings are negligible.

84. Indigenous people sometimes cook in watertight baskets by placing hot rocks into water to bring it to a boil. What mass of $500\text{-}^\circ\text{C}$ granite must be placed in 4.00 kg of $15.0\text{-}^\circ\text{C}$ water to bring its temperature to $100\,^\circ\text{C}$, if 0.0250 kg of water escapes as vapor from the initial sizzle? You may neglect the effects of the surroundings.

85. What would the final temperature of the pan and water be in **Example 16.6** if 0.260 kg of water were placed in the pan and 0.0100 kg of the water evaporated immediately, leaving the remainder to come to a common temperature with the pan?

16.6 Mechanisms of Heat Transfer

86. (a) Calculate the rate of heat conduction through house walls that are 13.0 cm thick and have an average thermal conductivity twice that of glass wool. Assume there are no windows or doors. The walls' surface area is 120 m^2 and their inside surface is at $18.0\,^\circ\text{C}$, while their outside surface is at $5.00\,^\circ\text{C}$. (b) How many 1-kW room heaters would be needed to balance the heat transfer due to conduction?

87. The rate of heat conduction out of a window on a winter day is rapid enough to chill the air next to it. To see just how rapidly the windows transfer heat by conduction, calculate the rate of conduction in watts through a 3.00-m^2 window that is 0.634 cm thick (1/4 in.) if the temperatures of the inner and outer surfaces are $5.00\,^\circ\text{C}$ and $-10.0\,^\circ\text{C}$, respectively. (This rapid rate will not be maintained—the inner surface will cool, even to the point of frost formation.)

88. Calculate the rate of heat conduction out of the human body, assuming that the core internal temperature is $37.0\,^\circ\text{C}$, the skin temperature is $34.0\,^\circ\text{C}$, the thickness of the fatty tissues between the core and the skin averages 1.00 cm, and the surface area is 1.40 m^2.

89. Suppose you stand with one foot on ceramic flooring and one foot on a wool carpet, making contact over an area of 80.0 cm^2 with each foot. Both the ceramic and the carpet are 2.00 cm thick and are $10.0\,^\circ\text{C}$ on their bottom sides. At what rate must heat transfer occur from each foot to keep the top of the ceramic and carpet at $33.0\,^\circ\text{C}$?

90. A man consumes 3000 kcal of food in one day, converting most of it to thermal energy to maintain body temperature. If he loses half this energy by evaporating water (through breathing and sweating), how many kilograms of water evaporate?

91. A firewalker runs across a bed of hot coals without sustaining burns. Calculate the heat transferred by conduction into the sole of one foot of a firewalker given that the bottom of the foot is a 3.00-mm-thick callus with a conductivity at the low end of the range for wood and its density is 300 kg/m^3. The area of contact is 25.0 cm^2, the temperature of the coals is $700\,^\circ\text{C}$, and the time in contact is 1.00 s. Ignore the evaporative cooling of sweat.

92. (a) What is the rate of heat conduction through the 3.00-cm-thick fur of a large animal having a 1.40-m^2 surface area? Assume that the animal's skin temperature is $32.0\,°\text{C}$, that the air temperature is $-5.00\,°\text{C}$, and that fur has the same thermal conductivity as air. (b) What food intake will the animal need in one day to replace this heat transfer?

93. A walrus transfers energy by conduction through its blubber at the rate of 150 W when immersed in $-1.00\,°\text{C}$ water. The walrus's internal core temperature is $37.0\,°\text{C}$, and it has a surface area of $2.00\,\text{m}^2$. What is the average thickness of its blubber, which has the conductivity of fatty tissues without blood?

94. Compare the rate of heat conduction through a 13.0-cm-thick wall that has an area of $10.0\,\text{m}^2$ and a thermal conductivity twice that of glass wool with the rate of heat conduction through a 0.750-cm-thick window that has an area of $2.00\,\text{m}^2$, assuming the same temperature difference across each.

95. Suppose a person is covered head to foot by wool clothing with average thickness of 2.00 cm and is transferring energy by conduction through the clothing at the rate of 50.0 W. What is the temperature difference across the clothing, given the surface area is $1.40\,\text{m}^2$?

96. Some stove tops are smooth ceramic for easy cleaning. If the ceramic is 0.600 cm thick and heat conduction occurs through the same area and at the same rate as computed in **Example 16.9**, what is the temperature difference across it? Ceramic has the same thermal conductivity as glass and brick.

97. One easy way to reduce heating (and cooling) costs is to add extra insulation in the attic of a house. Suppose a single-story cubical house already had 15 cm of fiberglass insulation in the attic and in all the exterior surfaces. If you added an extra 8.0 cm of fiberglass to the attic, by what percentage would the heating cost of the house drop? Take the house to have dimensions 10 m by 15 m by 3.0 m. Ignore air infiltration and heat loss through windows and doors, and assume that the interior is uniformly at one temperature and the exterior is uniformly at another.

98. Many decisions are made on the basis of the payback period: the time it will take through savings to equal the capital cost of an investment. Acceptable payback times depend upon the business or philosophy one has. (For some industries, a payback period is as small as 2 years.) Suppose you wish to install the extra insulation in the preceding problem. If energy cost $\$1.00$ per million joules and the insulation was $4.00 per square meter, then calculate the simple payback time. Take the average ΔT for the 120-day heating season to be $15.0\,°\text{C}$.

ADDITIONAL PROBLEMS

99. In 1701, the Danish astronomer Ole Rømer proposed a temperature scale with two fixed points, freezing water at 7.5 degrees, and boiling water at 60.0 degrees. What is the boiling point of oxygen, 90.2 K, on the Rømer scale?

100. What is the percent error of thinking the melting point of tungsten is $3695\,°\text{C}$ instead of the correct value of 3695 K?

101. An engineer wants to design a structure in which the difference in length between a steel beam and an aluminum beam remains at 0.500 m regardless of temperature, for ordinary temperatures. What must the lengths of the beams be?

102. A mercury thermometer still in use for meteorology has a bulb with a volume of $0.780\,\text{cm}^3$ and a tube for the mercury to expand into of inside diameter 0.130 mm. (a) Neglecting the thermal expansion of the glass, what is the spacing between marks $1\,°\text{C}$ apart? (b) If the thermometer is made of ordinary glass (not a good idea), what is the spacing?

103. Even when shut down after a period of normal use, a large commercial nuclear reactor transfers thermal energy at the rate of 150 MW by the radioactive decay of fission products. This heat transfer causes a rapid increase in temperature if the cooling system fails $(1\text{ watt} = 1\text{ joule/second}$ or $1\text{ W} = 1\text{ J/s}$ and $1\text{ MW} = 1\text{ megawatt})$. (a) Calculate the rate of temperature increase in degrees Celsius per second $(°\text{C/s})$ if the mass of the reactor core is $1.60 \times 10^5\text{ kg}$ and it has an average specific heat of $0.3349\text{ kJ/kg}\cdot°\text{C}$. (b) How long would it take to obtain a temperature increase of $2000\,°\text{C}$, which could cause some metals holding the radioactive materials to melt? (The initial rate of temperature increase would be greater than that calculated here because the heat transfer is concentrated in a smaller mass. Later, however, the temperature increase would slow down because the 500,000-kg steel containment vessel would also begin to heat up.)

104. You leave a pastry in the refrigerator on a plate and ask your roommate to take it out before you get home so you can eat it at room temperature, the way you like it. Instead, your roommate plays video games for hours. When you return, you notice that the pastry is still cold, but the game console has become hot. Annoyed, and knowing that the pastry will not be good if it is microwaved, you warm up the pastry by unplugging the console and putting it in a clean trash bag (which acts as a perfect calorimeter) with the pastry on the plate. After a while, you find that the equilibrium temperature is a nice, warm $38.3\ °C$. You know that the game console has a mass of 2.1 kg. Approximate it as having a uniform initial temperature of $45\ °C$. The pastry has a mass of 0.16 kg and a specific heat of $3.0\ \text{k J/(kg} \cdot °\text{C)},$ and is at a uniform initial temperature of $4.0\ °C$. The plate is at the same temperature and has a mass of 0.24 kg and a specific heat of $0.90\ \text{J/(kg} \cdot °\text{C)}$. What is the specific heat of the console?

105. Two solid spheres, *A* and *B*, made of the same material, are at temperatures of $0\ °C$ and $100\ °C$, respectively. The spheres are placed in thermal contact in an ideal calorimeter, and they reach an equilibrium temperature of $20\ °C$. Which is the bigger sphere? What is the ratio of their diameters?

106. In some countries, liquid nitrogen is used on dairy trucks instead of mechanical refrigerators. A 3.00-hour delivery trip requires 200 L of liquid nitrogen, which has a density of $808\ \text{kg/m}^3$. (a) Calculate the heat transfer necessary to evaporate this amount of liquid nitrogen and raise its temperature to $3.00\ °C$. (Use c_P and assume it is constant over the temperature range.) This value is the amount of cooling the liquid nitrogen supplies. (b) What is this heat transfer rate in kilowatt-hours? (c) Compare the amount of cooling obtained from melting an identical mass of $0\text{-}°C$ ice with that from evaporating the liquid nitrogen.

107. Some gun fanciers make their own bullets, which involves melting lead and casting it into lead slugs. How much heat transfer is needed to raise the temperature and melt 0.500 kg of lead, starting from $25.0\ °C$?

108. A 0.800-kg iron cylinder at a temperature of $1.00 \times 10^3\ °C$ is dropped into an insulated chest of 1.00 kg of ice at its melting point. What is the final temperature, and how much ice has melted?

109. Repeat the preceding problem with 2.00 kg of ice instead of 1.00 kg.

110. Repeat the preceding problem with 0.500 kg of ice, assuming that the ice is initially in a copper container of mass 1.50 kg in equilibrium with the ice.

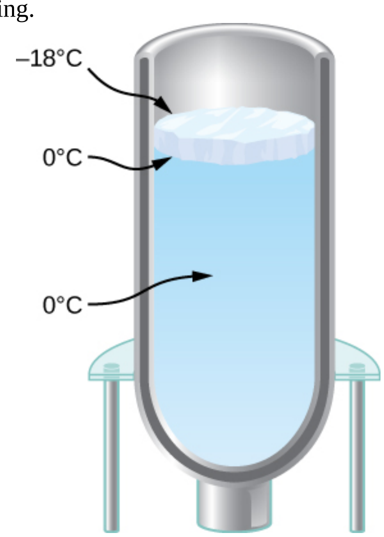
111. A 30.0-g ice cube at its melting point is dropped into an aluminum calorimeter of mass 100.0 g in equilibrium at $24.0\ °C$ with 300.0 g of an unknown liquid. The final temperature is $4.0\ °C$. What is the heat capacity of the liquid?

112. (a) Calculate the rate of heat conduction through a double-paned window that has a 1.50-m^2 area and is made of two panes of 0.800-cm-thick glass separated by a 1.00-cm air gap. The inside surface temperature is $15.0\ °C,$ while that on the outside is $-10.0\ °C.$ (*Hint:* There are identical temperature drops across the two glass panes. First find these and then the temperature drop across the air gap. This problem ignores the increased heat transfer in the air gap due to convection.) (b) Calculate the rate of heat conduction through a 1.60-cm-thick window of the same area and with the same temperatures. Compare your answer with that for part (a).

113. (a) An exterior wall of a house is 3 m tall and 10 m wide. It consists of a layer of drywall with an *R* factor of 0.56, a layer 3.5 inches thick filled with fiberglass batts, and a layer of insulated siding with an *R* factor of 2.6. The wall is built so well that there are no leaks of air through it. When the inside of the wall is at $22\ °C$ and the outside is at $-2\ °C$, what is the rate of heat flow through the wall? (b) More realistically, the 3.5-inch space also contains 2-by-4 studs—wooden boards 1.5 inches by 3.5 inches oriented so that 3.5-inch dimension extends from the drywall to the siding. They are “on 16-inch centers,” that is, the centers of the studs are 16 inches apart. What is the heat current in this situation? Don’t worry about one stud more or less.

114. For the human body, what is the rate of heat transfer by conduction through the body’s tissue with the following conditions: the tissue thickness is 3.00 cm, the difference in temperature is $2.00\ °C$, and the skin area is $1.50\ \text{m}^2$. How does this compare with the average heat transfer rate to the body resulting from an energy intake of about 2400 kcal per day? (No exercise is included.)

115. You have a Dewar flask (a laboratory vacuum flask) that has an open top and straight sides, as shown below. You fill it with water and put it into the freezer. It is effectively a perfect insulator, blocking all heat transfer, except on the top. After a time, ice forms on the surface of the water. The liquid water and the bottom surface of the ice, in contact with the liquid water, are at $0\ °C$. The top surface of the ice is at the same temperature as the air in the freezer, $-18\ °C.$ Set the rate of heat flow through the ice equal to the rate of loss of heat of fusion as the water freezes. When the ice layer is 0.700 cm thick, find the rate in m/s at which the ice is thickening.



116. An infrared heater for a sauna has a surface area of $0.050\ \text{m}^2$ and an emissivity of 0.84. What temperature must it run at if the required power is 360 W? Neglect the temperature of the environment.

117. (a) Determine the power of radiation from the Sun by noting that the intensity of the radiation at the distance of Earth is $1370\ \text{W/m}^2$. *Hint:* That intensity will be found everywhere on a spherical surface with radius equal to that of Earth's orbit. (b) Assuming that the Sun's temperature is 5780 K and that its emissivity is 1, find its radius.

CHALLENGE PROBLEMS

118. A pendulum is made of a rod of length *L* and negligible mass, but capable of thermal expansion, and a weight of negligible size. (a) Show that when the temperature increases by *dT*, the period of the pendulum increases by a fraction $\alpha L dT/2$. (b) A clock controlled by a brass pendulum keeps time correctly at $10\ °\text{C}$. If the room temperature is $30\ °\text{C}$, does the clock run faster or slower? What is its error in seconds per day?

119. In a calorimeter of negligible heat capacity, 200 g of steam at $150\ °\text{C}$ and 100 g of ice at $-40\ °\text{C}$ are mixed. The pressure is maintained at 1 atm. What is the final temperature, and how much steam, ice, and water are present?

120. An astronaut performing an extra-vehicular activity (space walk) shaded from the Sun is wearing a spacesuit that can be approximated as perfectly white $(e = 0)$ except for a $5\ \text{cm} \times 8\ \text{cm}$ patch in the form of the astronaut's national flag. The patch has emissivity 0.300. The spacesuit under the patch is 0.500 cm thick, with a thermal conductivity $k = 0.0600\ \text{W/m}\ °\text{C}$, and its inner surface is at a temperature of $20.0\ °\text{C}$. What is the temperature of the patch, and what is the rate of heat loss through it? Assume the patch is so thin that its outer surface is at the same temperature as the outer surface of the spacesuit under it. Also assume the temperature of outer space is 0 K. You will get an equation that is very hard to solve in closed form, so you can solve it numerically with a graphing calculator, with software, or even by trial and error with a calculator.

121. As the very first rudiment of climatology, estimate the temperature of Earth. Assume it is a perfect sphere and its temperature is uniform. Ignore the greenhouse effect. Thermal radiation from the Sun has an intensity (the "solar constant" *S*) of about $1370\ \text{W/m}^2$ at the radius of Earth's orbit. (a) Assuming the Sun's rays are parallel, what area must *S* be multiplied by to get the total radiation intercepted by Earth? It will be easiest to answer in terms of Earth's radius, *R*. (b) Assume that Earth reflects about 30% of the solar energy it intercepts. In other words, Earth has an albedo with a value of $A = 0.3$. In terms of *S*, *A*, and *R*, what is the rate at which Earth absorbs energy from the Sun? (c) Find the temperature at which Earth radiates energy at the same rate. Assume that at the infrared wavelengths where it radiates, the emissivity *e* is 1. Does your result show that the greenhouse effect is important? (d) How does your answer depend on the the area of Earth?

122. Let's stop ignoring the greenhouse effect and incorporate it into the previous problem in a very rough way. Assume the atmosphere is a single layer, a spherical shell around Earth, with an emissivity $e = 0.77$ (chosen simply to give the right answer) at infrared wavelengths emitted by Earth and by the atmosphere. However, the atmosphere is transparent to the Sun's radiation (that is, assume the radiation is at visible wavelengths with no infrared), so the Sun's radiation reaches the surface. The greenhouse effect comes from the difference between the atmosphere's transmission of visible light and its rather strong absorption of infrared. Note that the atmosphere's radius is not significantly different from Earth's, but since the atmosphere is a layer above Earth, it emits radiation both upward and downward, so it has twice Earth's area. There are three radiative energy transfers in this problem: solar radiation absorbed by Earth's surface; infrared radiation from the surface, which is absorbed by the atmosphere according to its emissivity; and infrared radiation from the atmosphere, half of which is absorbed by Earth and half of which goes out into space. Apply the method of the previous problem to get an equation for Earth's surface and one for the atmosphere, and solve them for the two unknown temperatures, surface and atmosphere.

a. In terms of Earth's radius, the constant σ, and the unknown temperature T_s of the surface, what is the power of the infrared radiation from the surface?
b. What is the power of Earth's radiation absorbed by the atmosphere?
c. In terms of the unknown temperature T_e of the atmosphere, what is the power radiated from the atmosphere?
d. Write an equation that says the power of the radiation the atmosphere absorbs from Earth equals the power of the radiation it emits.
e. Half of the power radiated by the atmosphere hits Earth. Write an equation that says that the power Earth absorbs from the atmosphere and the Sun equals the power that it emits.
f. Solve your two equations for the unknown temperature of Earth.

For steps that make this model less crude, see for example the **lectures (https://openstax.org/l/21paulgormlec)** by Paul O'Gorman.

17 | KINETIC THEORY

Figure 17.1 A volcanic eruption releases tons of gas and dust into the atmosphere. Most of the gas is water vapor, but several other gases are common, including greenhouse gases such as carbon dioxide and acidic pollutants such as sulfur dioxide. However, the emission of volcanic gas is not all bad: Many geologists believe that in the earliest stages of Earth's formation, volcanic emissions formed the early atmosphere. (credit: modification of work by "Boaworm"/Wikimedia Commons)

Chapter Outline

Introduction

Gases are literally all around us—the air that we breathe is a mixture of gases. Other gases include those that make breads and cakes soft, those that make drinks fizzy, and those that burn to heat many homes. Engines and refrigerators depend on the behaviors of gases, as we will see in later chapters.

As we discussed in the preceding chapter, the study of heat and temperature is part of an area of physics known as thermodynamics, in which we require a system to be *macroscopic*, that is, to consist of a huge number (such as 10^{23}) of molecules. We begin by considering some macroscopic properties of gases: volume, pressure, and temperature. The simple model of a hypothetical "ideal gas" describes these properties of a gas very accurately under many conditions. We move from the ideal gas model to a more widely applicable approximation, called the Van der Waals model.

To understand gases even better, we must also look at them on the *microscopic* scale of molecules. In gases, the molecules interact weakly, so the microscopic behavior of gases is relatively simple, and they serve as a good introduction to systems of many molecules. The molecular model of gases is called the kinetic theory of gases and is one of the classic examples of a molecular model that explains everyday behavior.

17.1 | Fluids, Density, and Pressure

Learning Objectives

By the end of this section, you will be able to:

- State the different phases of matter
- Describe the characteristics of the phases of matter at the molecular or atomic level
- Distinguish between compressible and incompressible materials
- Define density and its related SI units
- Compare and contrast the densities of various substances
- Define pressure and its related SI units
- Explain the relationship between pressure and force
- Calculate force given pressure and area

Characteristics of Fluids

Liquids and gases are considered to be **fluids** because they yield to shearing forces, whereas solids resist them. Like solids, the molecules in a liquid are bonded to neighboring molecules, but possess many fewer of these bonds. The molecules in a liquid are not locked in place and can move with respect to each other. The distance between molecules is similar to the distances in a solid, and so liquids have definite volumes, but the shape of a liquid changes, depending on the shape of its container. Gases are not bonded to neighboring atoms and can have large separations between molecules. Gases have neither specific shapes nor definite volumes, since their molecules move to fill the container in which they are held (Figure 17.2).

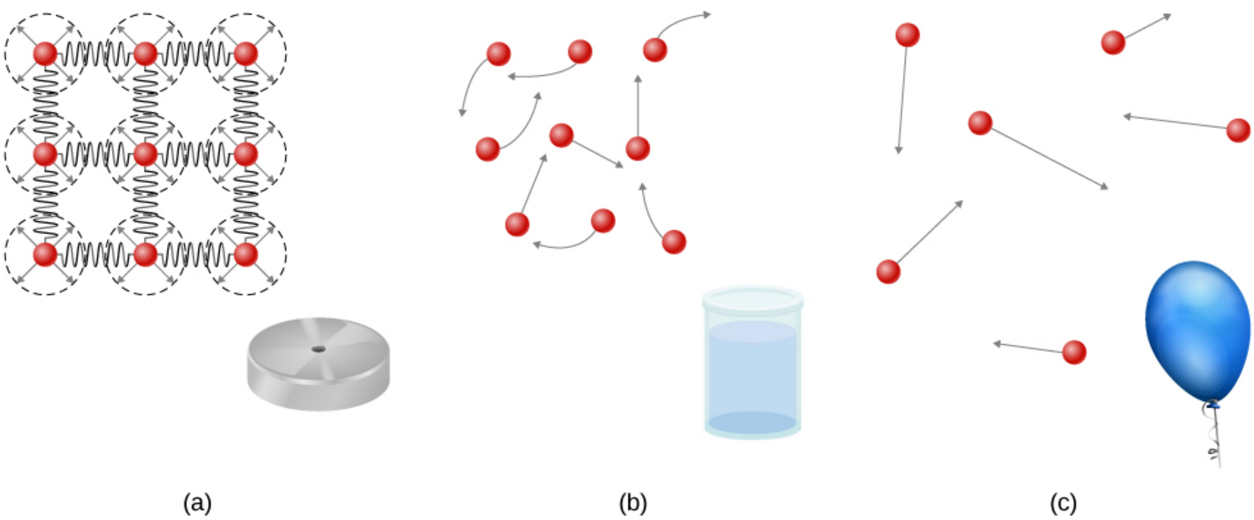
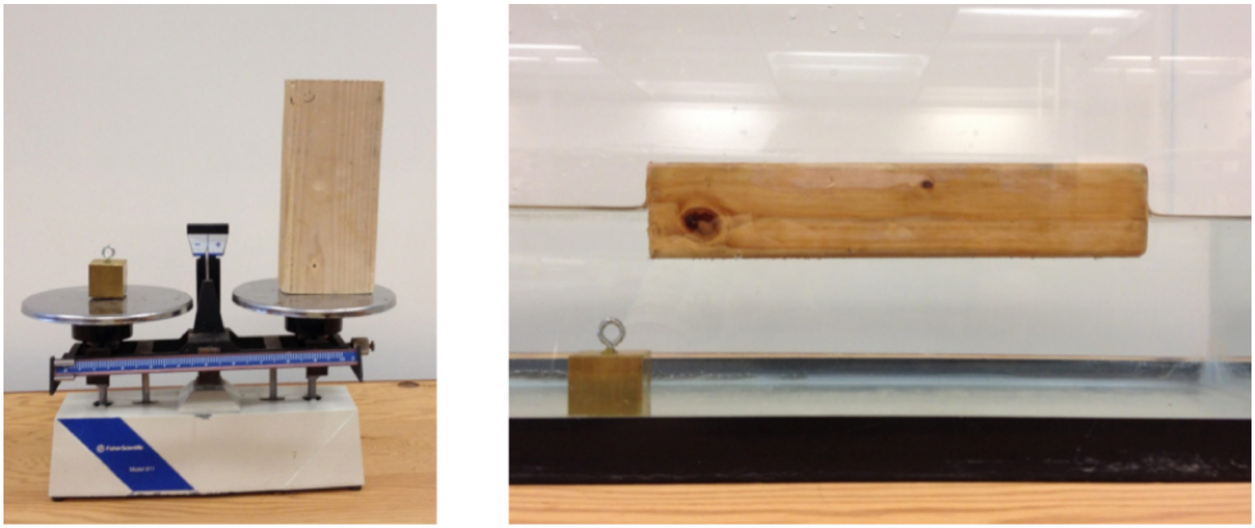


Figure 17.2 (a) Atoms in a solid are always in close contact with neighboring atoms, held in place by forces represented here by springs. (b) Atoms in a liquid are also in close contact but can slide over one another. Forces between the atoms strongly resist attempts to compress the atoms. (c) Atoms in a gas move about freely and are separated by large distances. A gas must be held in a closed container to prevent it from expanding freely and escaping.

Liquids deform easily when stressed and do not spring back to their original shape once a force is removed. This occurs because the atoms or molecules in a liquid are free to slide about and change neighbors. That is, liquids flow (so they are a type of fluid), with the molecules held together by mutual attraction. When a liquid is placed in a container with no lid, it remains in the container. Because the atoms are closely packed, liquids, like solids, resist compression; an extremely large force is necessary to change the volume of a liquid.

Density

Suppose a block of brass and a block of wood have exactly the same mass. If both blocks are dropped in a tank of water, why does the wood float and the brass sink (**Figure 17.3**)? This occurs because the brass has a greater density than water, whereas the wood has a lower density than water.



(a) (b)

Figure 17.3 (a) A block of brass and a block of wood both have the same weight and mass, but the block of wood has a much greater volume. (b) When placed in a fish tank filled with water, the cube of brass sinks and the block of wood floats. (The block of wood is the same in both pictures; it was turned on its side to fit on the scale.) (credit: modification of works by Joseph J. Trout, Stockton University)

Density is an important characteristic of substances. It is crucial, for example, in determining whether an object sinks or floats in a fluid.

Density

The average density of a substance or object is defined as its mass per unit volume,

$$\rho = \frac{m}{V} \tag{17.1}$$

where the Greek letter ρ (rho) is the symbol for density, *m* is the mass, and *V* is the volume.

The SI unit of density is kg/m^3. **Table 17.1** lists some representative values. The cgs unit of density is the gram per cubic centimeter, g/cm^3, where

$$1\ \text{g/cm}^3 = 1000\ \text{kg/m}^3.$$

The metric system was originally devised so that water would have a density of $1\ \text{g/cm}^3$, equivalent to $10^3\ \text{kg/m}^3$. Thus, the basic mass unit, the kilogram, was first devised to be the mass of 1000 mL of water, which has a volume of $1000\ \text{cm}^3$.

Solids (0.0°C)		Liquids (0.0°C)		Gases (0.0°C, 101.3 kPa)	
Substance	**$\rho(\text{kg/m}^3)$**	**Substance**	**$\rho(\text{kg/m}^3)$**	**Substance**	**$\rho(\text{kg/m}^3)$**
Aluminum	2.70×10^3	Benzene	8.79×10^2	Air	1.29×10^0

Table 17.1 Densities of Some Common Substances

Solids ($0.0°C$)		Liquids ($0.0°C$)		Gases ($0.0°C$, 101.3 kPa)	
Bone	1.90×10^3	Blood	1.05×10^3	Carbon dioxide	1.98×10^0
Brass	8.44×10^3	Ethyl alcohol	8.06×10^2	Carbon monoxide	1.25×10^0
Concrete	2.40×10^3	Gasoline	6.80×10^2	Helium	1.80×10^{-1}
Copper	8.92×10^3	Glycerin	1.26×10^3	Hydrogen	9.00×10^{-2}
Cork	2.40×10^2	Mercury	1.36×10^4	Methane	7.20×10^{-2}
Earth's crust	3.30×10^3	Olive oil	9.20×10^2	Nitrogen	1.25×10^0
Glass	2.60×10^3			Nitrous oxide	1.98×10^0
Gold	1.93×10^4			Oxygen	1.43×10^0
Granite	2.70×10^3				
Iron	7.86×10^3				
Lead	1.13×10^4				
Oak	7.10×10^2				
Pine	3.73×10^2				
Platinum	2.14×10^4				
Polystyrene	1.00×10^2				
Tungsten	1.93×10^4				
Uranium	1.87×10^3				

Table 17.1 Densities of Some Common Substances

As you can see by examining **Table 17.1**, the density of an object may help identify its composition. The density of gold, for example, is about 2.5 times the density of iron, which is about 2.5 times the density of aluminum. Density also reveals something about the phase of the matter and its substructure. Notice that the densities of liquids and solids are roughly comparable, consistent with the fact that their atoms are in close contact. The densities of gases are much less than those of liquids and solids, because the atoms in gases are separated by large amounts of empty space. The gases are displayed for a standard temperature of $0.0°C$ and a standard pressure of 101.3 kPa, and there is a strong dependence of the densities on temperature and pressure. The densities of the solids and liquids displayed are given for the standard temperature of $0.0°C$ and the densities of solids and liquids depend on the temperature. The density of solids and liquids normally increase with decreasing temperature.

Table 17.2 shows the density of water in various phases and temperature. The density of water increases with decreasing temperature, reaching a maximum at $4.0°C$, and then decreases as the temperature falls below $4.0°C$. This behavior of the density of water explains why ice forms at the top of a body of water.

Substance	$\rho(\text{kg/m}^3)$
Ice ($0°C$)	9.17×10^2

Table 17.2 Densities of Water

Substance	$\rho(\text{kg/m}^3)$
Water (0°C)	9.998×10^2
Water (4°C)	1.000×10^3
Water (20°C)	9.982×10^2
Water (100°C)	9.584×10^2
Steam (100°C, 101.3 kPa)	1.670×10^2
Sea water (0°C)	1.030×10^3

Table 17.2 Densities of Water

Since gases are free to expand and contract, the densities of the gases vary considerably with temperature, whereas the densities of liquids vary little with temperature. Therefore, the densities of liquids are often treated as constant, with the density equal to the average density.

Density is a dimensional property; therefore, when comparing the densities of two substances, the units must be taken into consideration. For this reason, a more convenient, dimensionless quantity called the **specific gravity** is often used to compare densities. Specific gravity is defined as the ratio of the density of the material to the density of water at $4.0\ °\text{C}$ and one atmosphere of pressure, which is $1000\ \text{kg/m}^3$:

$$\text{Specific gravity} = \frac{\text{Density of material}}{\text{Density of water}}.$$

The comparison uses water because the density of water is $1\ \text{g/cm}^3$, which was originally used to define the kilogram. Specific gravity, being dimensionless, provides a ready comparison among materials without having to worry about the unit of density. For instance, the density of aluminum is 2.7 in g/cm^3 (2700 in kg/m^3), but its specific gravity is 2.7, regardless of the unit of density. Specific gravity is a particularly useful quantity with regard to buoyancy, which we will discuss later in this chapter.

Pressure

You have no doubt heard the word ‘pressure’ used in relation to blood (high or low blood pressure) and in relation to weather (high- and low-pressure weather systems). These are only two of many examples of pressure in fluids.

Pressure

Pressure (p) is defined as the normal force F per unit area A over which the force is applied, or

$$p = \frac{F}{A}. \quad (17.2)$$

A given force can have a significantly different effect, depending on the area over which the force is exerted. For instance, a force applied to an area of $1\ \text{mm}^2$ has a pressure that is 100 times as great as the same force applied to an area of $1\ \text{cm}^2$. That is why a sharp needle is able to poke through skin when a small force is exerted, but applying the same force with a finger does not puncture the skin (**Figure 17.4**).

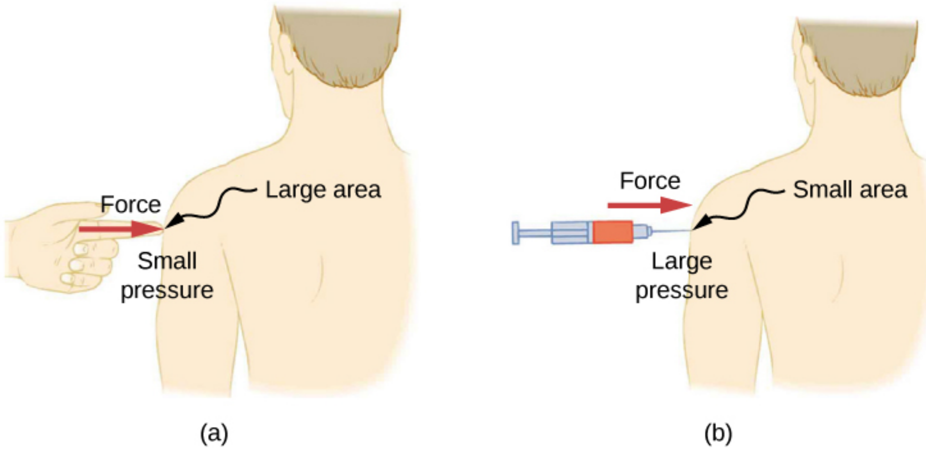


Figure 17.4 (a) A person being poked with a finger might be irritated, but the force has little lasting effect. (b) In contrast, the same force applied to an area the size of the sharp end of a needle is enough to break the skin.

Note that although force is a vector, pressure is a scalar. Pressure is a scalar quantity because it is defined to be proportional to the magnitude of the force acting perpendicular to the surface area. The SI unit for pressure is the *pascal* (Pa), named after the French mathematician and physicist Blaise Pascal (1623–1662), where

$$1\ \text{Pa} = 1\ \text{N/m}^2.$$

Several other units are used for pressure, which we discuss later in the chapter.

Direction of pressure in a fluid

Fluid pressure has no direction, being a scalar quantity, whereas the forces due to pressure have well-defined directions: They are always exerted perpendicular to any surface. Thus, in a static fluid enclosed in a tank, the force exerted on the walls of the tank is exerted perpendicular to the inside surface. Likewise, pressure is exerted perpendicular to the surfaces of any object within the fluid. **Figure 17.5** illustrates the pressure exerted by air on the walls of a tire and by water on the body of a swimmer.

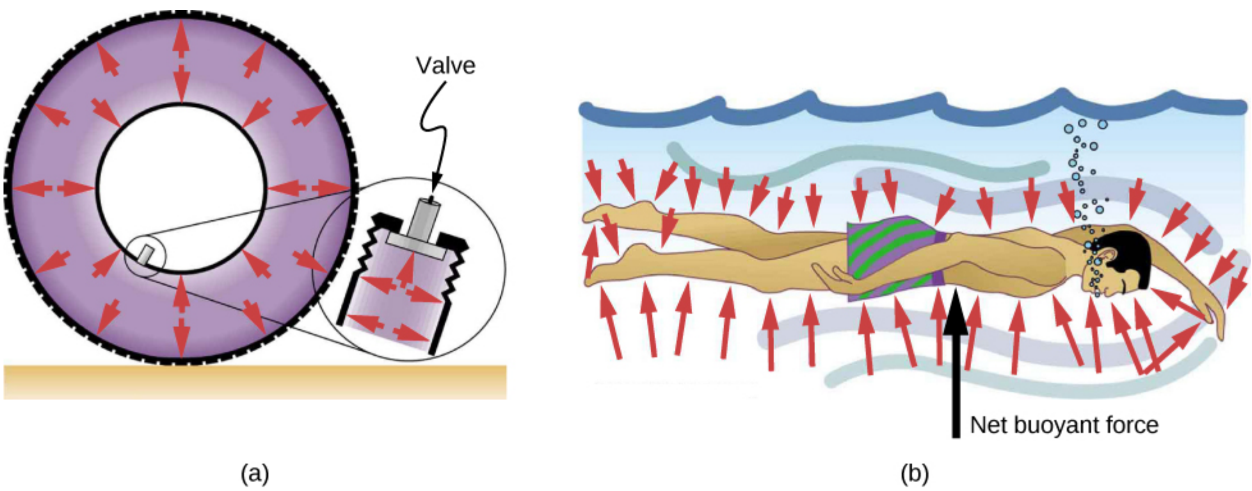


Figure 17.5 (a) Pressure inside this tire exerts forces perpendicular to all surfaces it contacts. The arrows represent directions and magnitudes of the forces exerted at various points. (b) Pressure is exerted perpendicular to all sides of this swimmer, since the water would flow into the space he occupies if he were not there. The arrows represent the directions and magnitudes of the forces exerted at various points on the swimmer. Note that the forces are larger underneath, due to greater depth, giving a net upward or buoyant force. The net vertical force on the swimmer is equal to the sum of the buoyant force and the weight of the swimmer.

17.2 | Molecular Model of an Ideal Gas

Learning Objectives

By the end of this section, you will be able to:

- Apply the ideal gas law to situations involving the pressure, volume, temperature, and the number of molecules of a gas
- Use the unit of moles in relation to numbers of molecules, and molecular and macroscopic masses
- Explain the ideal gas law in terms of moles rather than numbers of molecules

In this section, we explore the thermal behavior of gases. Our word “gas” comes from the Flemish word meaning “chaos,” first used for vapors by the seventeenth-century chemist J. B. van Helmont. The term was more appropriate than he knew, because gases consist of molecules moving and colliding with each other at random. This randomness makes the connection between the microscopic and macroscopic domains simpler for gases than for liquids or solids.

How do gases differ from solids and liquids? Under ordinary conditions, such as those of the air around us, the difference is that the molecules of gases are much farther apart than those of solids and liquids. Because the typical distances between molecules are large compared to the size of a molecule, as illustrated in **Figure 17.6**, the forces between them are considered negligible, except when they come into contact with each other during collisions. Also, at temperatures well above the boiling temperature, the motion of molecules is fast, and the gases expand rapidly to occupy all of the accessible volume. In contrast, in liquids and solids, molecules are closer together, and the behavior of molecules in liquids and solids is highly constrained by the molecules’ interactions with one another.

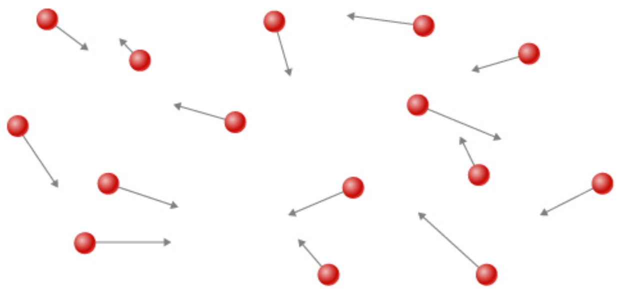


Figure 17.6 Atoms and molecules in a gas are typically widely separated. Because the forces between them are quite weak at these distances, the properties of a gas depend more on the number of atoms per unit volume and on temperature than on the type of atom.

The Gas Laws

In the previous chapter, we saw one consequence of the large intermolecular spacing in gases: Gases are easily compressed. **Table 16.2** shows that gases have larger coefficients of volume expansion than either solids or liquids. These large coefficients mean that gases expand and contract very rapidly with temperature changes. We also saw (in the section on thermal expansion) that most gases expand at the same rate or have the same coefficient of volume expansion, β. This raises a question: Why do all gases act in nearly the same way, when all the various liquids and solids have widely varying expansion rates?

To study how the pressure, temperature, and volume of a gas relate to one another, consider what happens when you pump air into a deflated car tire. The tire’s volume first increases in direct proportion to the amount of air injected, without much increase in the tire pressure. Once the tire has expanded to nearly its full size, the tire’s walls limit its volume expansion. If we continue to pump air into the tire, the pressure increases. When the car is driven and the tires flex, their temperature increases, and therefore the pressure increases even further (**Figure 17.7**).

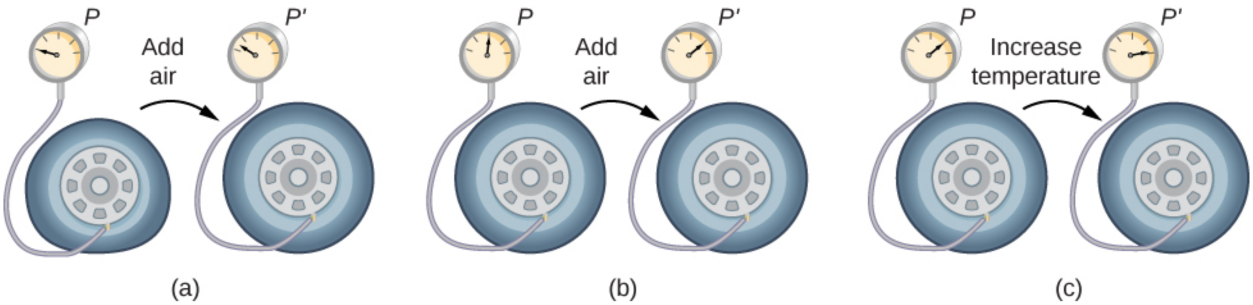


Figure 17.7 (a) When air is pumped into a deflated tire, its volume first increases without much increase in pressure. (b) When the tire is filled to a certain point, the tire walls resist further expansion, and the pressure increases with more air. (c) Once the tire is inflated, its pressure increases with temperature.

Figure 17.8 shows data from the experiments of Robert Boyle (1627–1691), illustrating what is now called Boyle's law: At constant temperature and number of molecules, the absolute pressure of a gas and its volume are inversely proportional. (The absolute pressure is the true pressure and the gauge pressure is the absolute pressure minus the ambient pressure, typically atmospheric pressure.) The graph in **Figure 17.8** displays this relationship as an inverse proportionality of volume to pressure.

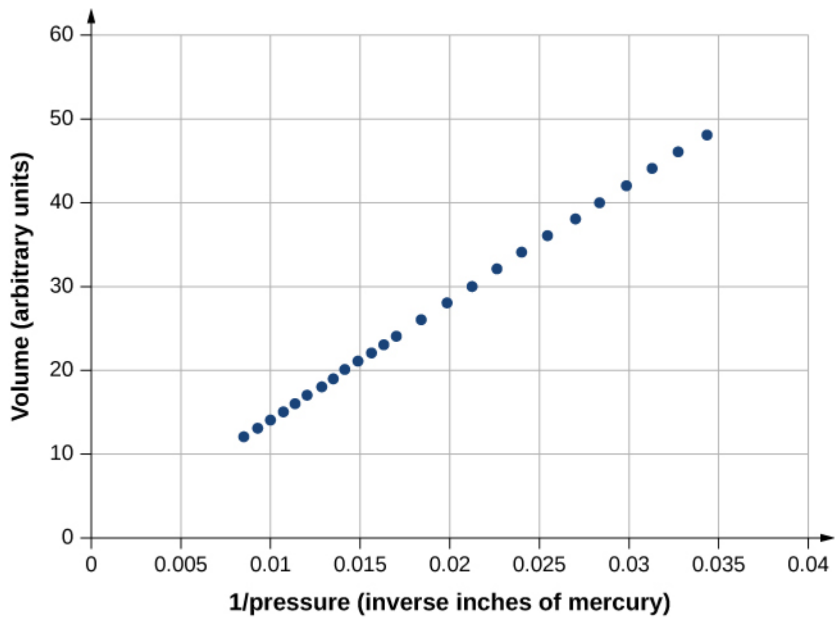


Figure 17.8 Robert Boyle and his assistant found that volume and pressure are inversely proportional. Here their data are plotted as V versus $1/p$; the linearity of the graph shows the inverse proportionality. The number shown as the volume is actually the height in inches of air in a cylindrical glass tube. The actual volume was that height multiplied by the cross-sectional area of the tube, which Boyle did not publish. The data are from Boyle's book *A Defence of the Doctrine Touching the Spring and Weight of the Air*…, p. 60.[1]

Figure 17.9 shows experimental data illustrating what is called Charles's law, after Jacques Charles (1746–1823). Charles's law states that at constant pressure and number of molecules, the volume of a gas is proportional to its absolute temperature.

1. http://bvpb.mcu.es/en/consulta/registro.cmd?id=406806

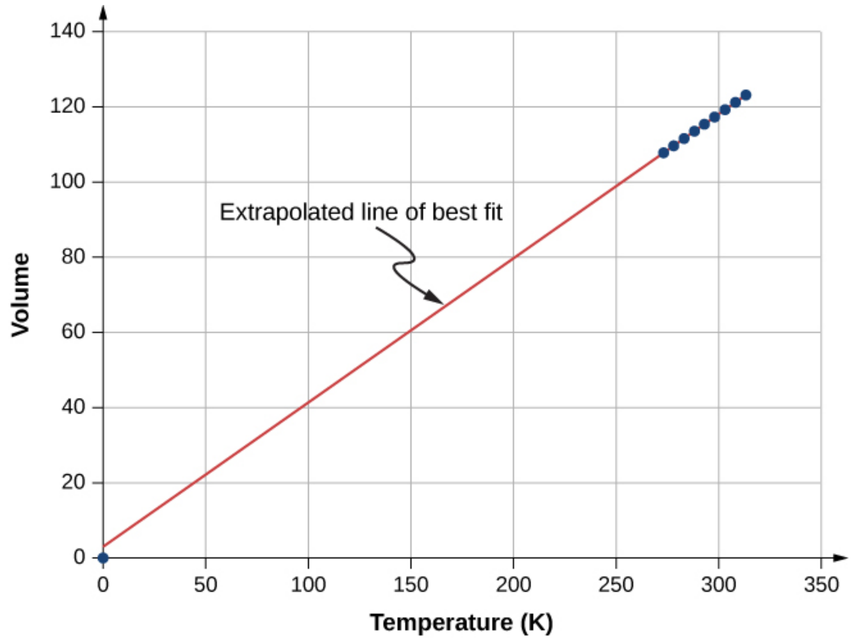


Figure 17.9 Experimental data showing that at constant pressure, volume is approximately proportional to temperature. The best-fit line passes approximately through the origin.[2]

Similar is Amonton's or Gay-Lussac's law, which states that at constant volume and number of molecules, the pressure is proportional to the temperature. That law is the basis of the constant-volume gas thermometer, discussed in the previous chapter. (The histories of these laws and the appropriate credit for them are more complicated than can be discussed here.)

It is known experimentally that for gases at low density (such that their molecules occupy a negligible fraction of the total volume) and at temperatures well above the boiling point, these proportionalities hold to a good approximation. Not surprisingly, with the other quantities held constant, either pressure or volume is proportional to the number of molecules. More surprisingly, when the proportionalities are combined into a single equation, the constant of proportionality is independent of the composition of the gas. The resulting equation for all gases applies in the limit of low density and high temperature; it's the same for oxygen as for helium or uranium hexafluoride. A gas at that limit is called an **ideal gas**; it obeys the **ideal gas law**, which is also called the equation of state of an ideal gas.

Ideal Gas Law

The ideal gas law states that

$$pV = Nk_B T, \tag{17.3}$$

where p is the absolute pressure of a gas, V is the volume it occupies, N is the number of molecules in the gas, and T is its absolute temperature.

The constant k_B is called the **Boltzmann constant** in honor of the Austrian physicist Ludwig Boltzmann (1844–1906) and has the value

$$k_B = 1.38 \times 10^{-23} \text{ J/K}.$$

The ideal gas law describes the behavior of any real gas when its density is low enough or its temperature high enough that it is far from liquefaction. This encompasses many practical situations. In the next section, we'll see why it's independent of the type of gas.

In many situations, the ideal gas law is applied to a sample of gas with a constant number of molecules; for instance, the gas may be in a sealed container. If N is constant, then solving for N shows that pV/T is constant. We can write that fact in a convenient form:

2. http://chemed.chem.purdue.edu/genchem/history/charles.html

$$\frac{p_1 V_1}{T_1} = \frac{p_2 V_2}{T_2}, \quad (17.4)$$

where the subscripts 1 and 2 refer to any two states of the gas at different times. Again, the temperature must be expressed in kelvin and the pressure must be absolute pressure, which is the sum of gauge pressure and atmospheric pressure.

Example 17.1

Calculating Pressure Changes Due to Temperature Changes

Suppose your bicycle tire is fully inflated, with an absolute pressure of 7.00×10^5 Pa (a gauge pressure of just under $90.0\ \text{lb/in.}^2$) at a temperature of $18.0\ °\text{C}$. What is the pressure after its temperature has risen to $35.0\ °\text{C}$ on a hot day? Assume there are no appreciable leaks or changes in volume.

Strategy

The pressure in the tire is changing only because of changes in temperature. We know the initial pressure $p_0 = 7.00 \times 10^5$ Pa, the initial temperature $T_0 = 18.0\ °\text{C}$, and the final temperature $T_\text{f} = 35.0\ °\text{C}$. We must find the final pressure p_f. Since the number of molecules is constant, we can use the equation

$$\frac{p_\text{f} V_\text{f}}{T_\text{f}} = \frac{p_0 V_0}{T_0}.$$

Since the volume is constant, V_f and V_0 are the same and they divide out. Therefore,

$$\frac{p_\text{f}}{T_\text{f}} = \frac{p_0}{T_0}.$$

We can then rearrange this to solve for p_f :

$$p_\text{f} = p_0 \frac{T_\text{f}}{T_0},$$

where the temperature must be in kelvin.

Solution

1. Convert temperatures from degrees Celsius to kelvin

$$T_0 = (18.0 + 273)\text{K} = 291\ \text{K},$$

$$T_\text{f} = (35.0 + 273)\text{K} = 308\ \text{K}.$$

2. Substitute the known values into the equation,

$$p_\text{f} = p_0 \frac{T_\text{f}}{T_0} = 7.00 \times 10^5\ \text{Pa} \left(\frac{308\ \text{K}}{291\ \text{K}}\right) = 7.41 \times 10^5\ \text{Pa}.$$

Significance

The final temperature is about 6% greater than the original temperature, so the final pressure is about 6% greater as well. Note that *absolute pressure* and *absolute temperature* (see **Thermometers and Temperature and Scales**) must be used in the ideal gas law.

Example 17.2

Calculating the Number of Molecules in a Cubic Meter of Gas

How many molecules are in a typical object, such as gas in a tire or water in a glass? This calculation can give us an idea of how large N typically is. Let's calculate the number of molecules in the air that a typical healthy young adult inhales in one breath, with a volume of 500 mL, at *standard temperature and pressure* (STP), which is defined as $0\,^\circ\text{C}$ and atmospheric pressure. (Our young adult is apparently outside in winter.)

Strategy

Because pressure, volume, and temperature are all specified, we can use the ideal gas law, $pV = Nk_\text{B}T,$ to find N.

Solution

1. Identify the knowns.

$$T = 0\,^\circ\text{C} = 273\text{ K},\ \ p = 1.01 \times 10^5\text{ Pa},\ \ V = 500\text{ mL} = 5 \times 10^{-4}\text{ m}^3,\ \ k_\text{B} = 1.38 \times 10^{-23}\text{ J/K}$$

2. Substitute the known values into the equation and solve for N.

$$N = \frac{pV}{k_\text{B}T} = \frac{(1.01 \times 10^5\text{ Pa})(5 \times 10^{-4}\text{ m}^3)}{(1.38 \times 10^{-23}\text{ J/K})(273\text{ K})} = 1.34 \times 10^{22}\text{ molecules}$$

Significance

N is huge, even in small volumes. For example, 1 cm^3 of a gas at STP contains 2.68×10^{19} molecules. Once again, note that our result for N is the same for all types of gases, including mixtures.

As we observed in the chapter on fluid mechanics, pascals are N/m^2, so $\text{Pa} \cdot \text{m}^3 = \text{N} \cdot \text{m} = \text{J}$. Thus, our result for N is dimensionless, a pure number that could be obtained by counting (in principle) rather than measuring. As it is the number of molecules, we put "molecules" after the number, keeping in mind that it is an aid to communication rather than a unit.

Moles and Avogadro's Number

It is often convenient to measure the amount of substance with a unit on a more human scale than molecules. The SI unit for this purpose was developed by the Italian scientist Amedeo Avogadro (1776–1856). (He worked from the hypothesis that equal volumes of gas at equal pressure and temperature contain equal numbers of molecules, independent of the type of gas. As mentioned above, this hypothesis has been confirmed when the ideal gas approximation applies.) A **mole** (abbreviated mol) is defined as the amount of any substance that contains as many molecules as there are atoms in exactly 12 grams (0.012 kg) of carbon-12. (Technically, we should say "formula units," not "molecules," but this distinction is irrelevant for our purposes.) The number of molecules in one mole is called **Avogadro's number** $(N_\text{A}),$ and the value of Avogadro's number is now known to be

$$N_\text{A} = 6.02 \times 10^{23}\text{ mol}^{-1}.$$

We can now write $N = N_A n$, where n represents the number of moles of a substance.

Avogadro's number relates the mass of an amount of substance in grams to the number of protons and neutrons in an atom or molecule (12 for a carbon-12 atom), which roughly determine its mass. It's natural to define a unit of mass such that the mass of an atom is approximately equal to its number of neutrons and protons. The unit of that kind accepted for use with the SI is the *unified atomic mass unit* (u), also called the *dalton*. Specifically, a carbon-12 atom has a mass of exactly 12 u, so that its molar mass M in grams per mole is numerically equal to the mass of one carbon-12 atom in u. That equality holds for any substance. In other words, N_A is not only the conversion from numbers of molecules to moles, but it is also the conversion from u to grams: $6.02 \times 10^{23}\text{ u} = 1\text{ g}$. See **Figure 17.10**.

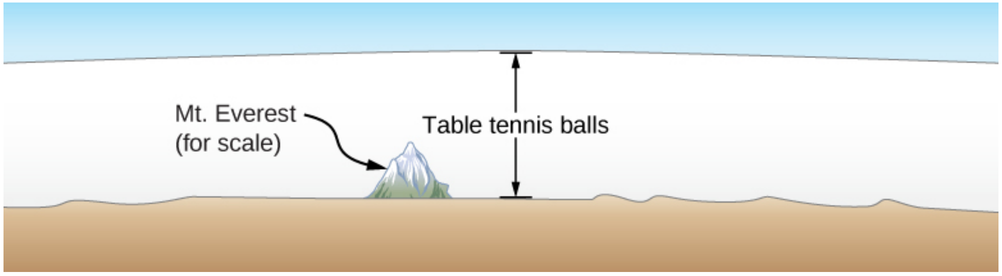


Figure 17.10 How big is a mole? On a macroscopic level, Avogadro's number of table tennis balls would cover Earth to a depth of about 40 km.

Now letting m_s stand for the mass of a sample of a substance, we have $m_s = nM$. Letting m stand for the mass of a molecule, we have $M = N_A m$.

17.1 Check Your Understanding The recommended daily amount of vitamin B_3 or niacin, $C_6NH_5O_2$, for women who are not pregnant or nursing, is 14 mg. Find the number of molecules of niacin in that amount.

17.2 Check Your Understanding The density of air in a classroom $(p = 1.00\,\text{atm}$ and $T = 20\,°\text{C})$ is $1.28\,\text{kg/m}^3$. At what pressure is the density $0.600\,\text{kg/m}^3$ if the temperature is kept constant?

The Ideal Gas Law Restated using Moles

A very common expression of the ideal gas law uses the number of moles in a sample, *n*, rather than the number of molecules, *N*. We start from the ideal gas law,

$$pV = Nk_B T,$$

and multiply and divide the right-hand side of the equation by Avogadro's number N_A. This gives us

$$pV = \frac{N}{N_A} N_A\, k_B T.$$

Note that $n = N/N_A$ is the number of moles. We define the **universal gas constant** as $R = N_A k_B$, and obtain the ideal gas law in terms of moles.

Ideal Gas Law (in terms of moles)

In terms of number of moles *n*, the ideal gas law is written as

$$pV = nRT. \quad (17.5)$$

In SI units,

$$R = N_A k_B = \left(6.02 \times 10^{23}\ \text{mol}^{-1}\right)\left(1.38 \times 10^{-23}\ \frac{\text{J}}{\text{K}}\right) = 8.31\ \frac{\text{J}}{\text{mol} \cdot \text{K}}.$$

In other units,

$$R = 1.99 \frac{\text{cal}}{\text{mol} \cdot \text{K}} = 0.0821 \frac{\text{L} \cdot \text{atm}}{\text{mol} \cdot \text{K}}.$$

You can use whichever value of *R* is most convenient for a particular problem.

Example 17.3

Density of Air at STP and in a Hot Air Balloon

Calculate the density of dry air (a) under standard conditions and (b) in a hot air balloon at a temperature of $120\,^\circ\text{C}$. Dry air is approximately $78\%\ \text{N}_2,\ 21\%\ \text{O}_2,$ and $1\%\ \text{Ar}$.

Strategy and Solution

a. We are asked to find the density, or mass per cubic meter. We can begin by finding the molar mass. If we have a hundred molecules, of which 78 are nitrogen, 21 are oxygen, and 1 is argon, the average molecular mass is $\frac{78\,m_{\text{N}_2} + 21\,m_{\text{O}_2} + m_{\text{Ar}}}{100}$, or the mass of each constituent multiplied by its percentage. The same applies to the molar mass, which therefore is

$$M = 0.78\,M_{\text{N}_2} + 0.21\,M_{\text{O}_2} + 0.01\,M_{\text{Ar}} = 29.0\ \text{g/mol}.$$

Now we can find the number of moles per cubic meter. We use the ideal gas law in terms of moles, $pV = nRT$, with $p = 1.00\ \text{atm}$, $T = 273\ \text{K}$, $V = 1\ \text{m}^3$, and $R = 8.31\ \text{J/mol}\cdot\text{K}$. The most convenient choice for R in this case is $R = 8.31\ \text{J/mol}\cdot\text{K}$ because the known quantities are in SI units:

$$n = \frac{pV}{RT} = \frac{(1.01\times10^5\ \text{Pa})\,(1\ \text{m}^3)}{(8.31\ \text{J/mol}\cdot\text{K})\,(273\ \text{K})} = 44.5\ \text{mol}.$$

Then, the mass m_s of that air is

$$m_\text{s} = nM = (44.5\ \text{mol})(29.0\ \text{g/mol}) = 1290\ \text{g} = 1.29\ \text{kg}.$$

Finally the density of air at STP is

$$\rho = \frac{m_\text{s}}{V} = \frac{1.29\ \text{kg}}{1\ \text{m}^3} = 1.29\ \text{kg/m}^3.$$

b. The air pressure inside the balloon is still 1 atm because the bottom of the balloon is open to the atmosphere. The calculation is the same except that we use a temperature of $120\,^\circ\text{C}$, which is 393 K. We can repeat the calculation in (a), or simply observe that the density is proportional to the number of moles, which is inversely proportional to the temperature. Then using the subscripts 1 for air at STP and 2 for the hot air, we have

$$\rho_2 = \frac{T_1}{T_2}\rho_1 = \frac{273\ \text{K}}{393\ \text{K}}(1.29\ \text{kg/m}^3) = 0.896\ \text{kg/m}^3.$$

17.3 Check Your Understanding Liquids and solids have densities on the order of 1000 times greater than gases. Explain how this implies that the distances between molecules in gases are on the order of 10 times greater than the size of their molecules.

The ideal gas law is closely related to energy: The units on both sides of the equation are joules. The right-hand side of the ideal gas law equation is $Nk_\text{B}T$. This term is roughly the total translational kinetic energy (which, when discussing gases, refers to the energy of translation of a molecule, not that of vibration of its atoms or rotation) of N molecules at an absolute temperature T, as we will see formally in the next section. The left-hand side of the ideal gas law equation is pV. As mentioned in the example on the number of molecules in an ideal gas, pressure multiplied by volume has units of energy. The energy of a gas can be changed when the gas does work as it increases in volume, something we explored in the preceding chapter, and the amount of work is related to the pressure. This is the process that occurs in gasoline or steam engines and turbines, as we'll see in the next chapter.

Problem-Solving Strategy: The Ideal Gas Law

Step 1. Examine the situation to determine that an ideal gas is involved. Most gases are nearly ideal unless they are close to the boiling point or at pressures far above atmospheric pressure.

Step 2. Make a list of what quantities are given or can be inferred from the problem as stated (identify the known quantities).

Step 3. Identify exactly what needs to be determined in the problem (identify the unknown quantities). A written list is useful.

Step 4. Determine whether the number of molecules or the number of moles is known or asked for to decide whether to use the ideal gas law as $pV = Nk_B T$, where *N* is the number of molecules, or $pV = nRT$, where *n* is the number of moles.

Step 5. Convert known values into proper SI units (K for temperature, Pa for pressure, m^3 for volume, molecules for *N*, and moles for *n*). If the units of the knowns are consistent with one of the non-SI values of *R*, you can leave them in those units. Be sure to use absolute temperature and absolute pressure.

Step 6. Solve the ideal gas law for the quantity to be determined (the unknown quantity). You may need to take a ratio of final states to initial states to eliminate the unknown quantities that are kept fixed.

Step 7. Substitute the known quantities, along with their units, into the appropriate equation and obtain numerical solutions complete with units.

Step 8. Check the answer to see if it is reasonable: Does it make sense?

17.3 | Pressure, Temperature, and RMS Speed

Learning Objectives

By the end of this section, you will be able to:

- Explain the relations between microscopic and macroscopic quantities in a gas
- Solve problems involving the distance and time between a gas molecule's collisions

We have examined pressure and temperature based on their macroscopic definitions. Pressure is the force divided by the area on which the force is exerted, and temperature is measured with a thermometer. We can gain a better understanding of pressure and temperature from the **kinetic theory of gases**, the theory that relates the macroscopic properties of gases to the motion of the molecules they consist of. First, we make two assumptions about molecules in an ideal gas.

1. There is a very large number *N* of molecules, all identical and each having mass *m*.
2. The molecules obey Newton's laws and are in continuous motion, which is random and isotropic, that is, the same in all directions.

To derive the ideal gas law and the connection between microscopic quantities such as the energy of a typical molecule and macroscopic quantities such as temperature, we analyze a sample of an ideal gas in a rigid container, about which we make two further assumptions:

3. The molecules are much smaller than the average distance between them, so their total volume is much less than that of their container (which has volume *V*). In other words, we take the Van der Waals constant *b*, the volume of a mole of gas molecules, to be negligible compared to the volume of a mole of gas in the container.
4. The molecules make perfectly elastic collisions with the walls of the container and with each other. Other forces on them, including gravity and the attractions represented by the Van der Waals constant *a*, are negligible (as is necessary for the assumption of isotropy).

The collisions between molecules do not appear in the derivation of the ideal gas law. They do not disturb the derivation either, since collisions between molecules moving with random velocities give new random velocities. Furthermore, if the velocities of gas molecules in a container are initially not random and isotropic, molecular collisions are what make them random and isotropic.

We make still further assumptions that simplify the calculations but do not affect the result. First, we let the container be a rectangular box. Second, we begin by considering *monatomic* gases, those whose molecules consist of single atoms, such as helium. Then, we can assume that the atoms have no energy except their translational kinetic energy; for instance, they have neither rotational nor vibrational energy. (Later, we discuss the validity of this assumption for real monatomic gases and dispense with it to consider diatomic and polyatomic gases.)

Figure 17.11 shows a collision of a gas molecule with the wall of a container, so that it exerts a force on the wall (by Newton's third law). These collisions are the source of pressure in a gas. As the number of molecules increases, the number of collisions, and thus the pressure, increases. Similarly, if the average velocity of the molecules is higher, the gas pressure is higher.

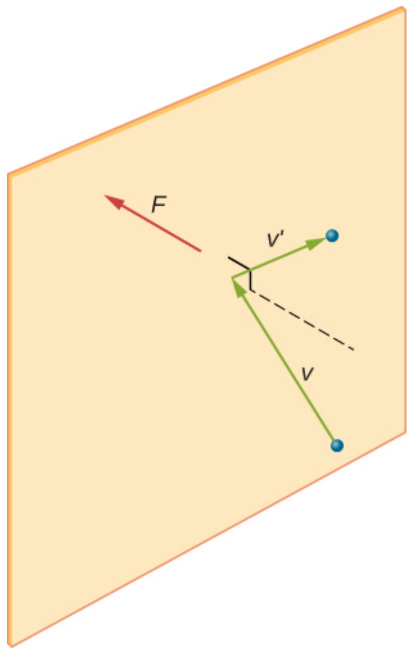


Figure 17.11 When a molecule collides with a rigid wall, the component of its momentum perpendicular to the wall is reversed. A force is thus exerted on the wall, creating pressure.

In a sample of gas in a container, the randomness of the molecular motion causes the number of collisions of molecules with any part of the wall in a given time to fluctuate. However, because a huge number of molecules collide with the wall in a short time, the number of collisions on the scales of time and space we measure fluctuates by only a tiny, usually unobservable fraction from the average. We can compare this situation to that of a casino, where the outcomes of the bets are random and the casino's takings fluctuate by the minute and the hour. However, over long times such as a year, the casino's takings are very close to the averages expected from the odds. A tank of gas has enormously more molecules than a casino has bettors in a year, and the molecules make enormously more collisions in a second than a casino has bets.

A calculation of the average force exerted by molecules on the walls of the box leads us to the ideal gas law and to the connection between temperature and molecular kinetic energy. (In fact, we will take two averages: one over time to get the average force exerted by one molecule with a given velocity, and then another average over molecules with different velocities.) This approach was developed by Daniel Bernoulli (1700–1782), who is best known in physics for his work on fluid flow (hydrodynamics). Remarkably, Bernoulli did this work before Dalton established the view of matter as consisting of atoms.

Figure 17.12 shows a container full of gas and an expanded view of an elastic collision of a gas molecule with a wall of the container, broken down into components. We have assumed that a molecule is small compared with the separation of molecules in the gas, and that its interaction with other molecules can be ignored. Under these conditions, the ideal gas law is experimentally valid. Because we have also assumed the wall is rigid and the particles are points, the collision is elastic (by conservation of energy—there's nowhere for a particle's kinetic energy to go). Therefore, the molecule's kinetic energy remains constant, and hence, its speed and the magnitude of its momentum remain constant as well. This assumption is not always valid, but the results in the rest of this module are also obtained in models that let the molecules exchange energy and momentum with the wall.

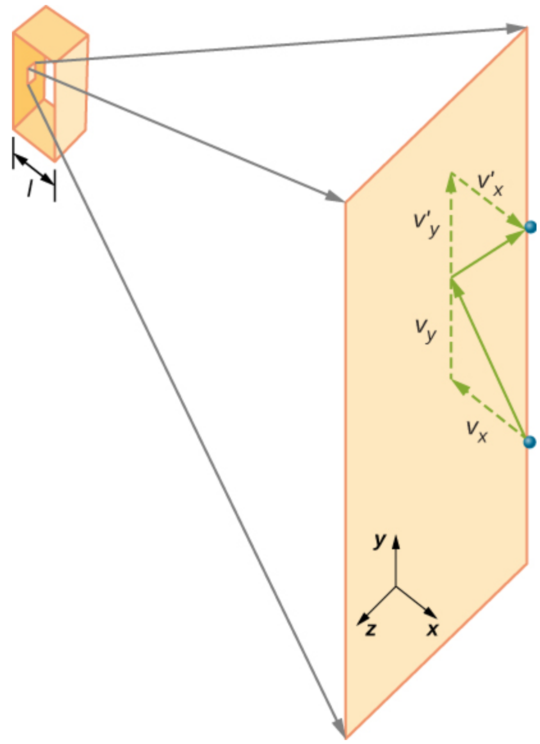


Figure 17.12 Gas in a box exerts an outward pressure on its walls. A molecule colliding with a rigid wall has its velocity and momentum in the *x*-direction reversed. This direction is perpendicular to the wall. The components of its velocity momentum in the *y*- and *z*-directions are not changed, which means there is no force parallel to the wall.

If the molecule's velocity changes in the *x*-direction, its momentum changes from $-mv_x$ to $+mv_x$. Thus, its change in momentum is $\Delta mv = +mv_x - (-mv_x) = 2mv_x$. According to the impulse-momentum theorem given in the chapter on linear momentum and collisions, the force exerted on the *i*th molecule, where *i* labels the molecules from 1 to *N*, is given by

$$F_i = \frac{\Delta p_i}{\Delta t} = \frac{2mv_{ix}}{\Delta t}.$$

(In this equation alone, *p* represents momentum, not pressure.) There is no force between the wall and the molecule except while the molecule is touching the wall. During the short time of the collision, the force between the molecule and wall is relatively large, but that is not the force we are looking for. We are looking for the average force, so we take Δt to be the average time between collisions of the given molecule with this wall, which is the time in which we expect to find one collision. Let *l* represent the length of the box in the *x*-direction. Then Δt is the time the molecule would take to go across the box and back, a distance 2*l*, at a speed of v_x. Thus $\Delta t = 2l/v_x$, and the expression for the force becomes

$$F_i = \frac{2mv_{ix}}{2l/v_{ix}} = \frac{mv_{ix}^2}{l}.$$

This force is due to *one* molecule. To find the total force on the wall, *F*, we need to add the contributions of all *N* molecules:

$$F = \sum_{i=1}^{N} F_i = \sum_{i=1}^{N} \frac{mv_{ix}^2}{l} = \frac{m}{l}\sum_{i=1}^{N} v_{ix}^2.$$

We now use the definition of the average, which we denote with a bar, to find the force:

$$F = N\frac{m}{l}\left(\frac{1}{N}\sum_{i=1}^{N} v_{ix}^2\right) = N\frac{m\overline{v_x^2}}{l}.$$

We want the force in terms of the speed *v*, rather than the *x*-component of the velocity. Note that the total velocity squared is the sum of the squares of its components, so that

$$\overline{v^2} = \overline{v_x^2} + \overline{v_y^2} + \overline{v_z^2}.$$

With the assumption of isotropy, the three averages on the right side are equal, so

$$\overline{v^2} = 3\overline{v_{ix}^2}.$$

Substituting this into the expression for *F* gives

$$F = N\frac{m\overline{v^2}}{3l}.$$

The pressure is *F*/*A*, so we obtain

$$p = \frac{F}{A} = N\frac{m\overline{v^2}}{3Al} = \frac{Nm\overline{v^2}}{3V},$$

where we used $V = Al$ for the volume. This gives the important result

$$pV = \frac{1}{3}Nm\overline{v^2}. \quad (17.6)$$

Combining this equation with $pV = Nk_BT$ gives

$$\frac{1}{3}Nm\overline{v^2} = Nk_BT.$$

We can get the average kinetic energy of a molecule, $\frac{1}{2}m\overline{v^2}$, from the left-hand side of the equation by dividing out *N* and multiplying by 3/2.

Average Kinetic Energy per Molecule

The average kinetic energy of a molecule is directly proportional to its absolute temperature:

$$\overline{K} = \frac{1}{2}m\overline{v^2} = \frac{3}{2}k_BT. \quad (17.7)$$

The equation $\overline{K} = \frac{3}{2}k_BT$ is the average kinetic energy per molecule. Note in particular that nothing in this equation depends on the molecular mass (or any other property) of the gas, the pressure, or anything but the temperature. If samples of helium and xenon gas, with very different molecular masses, are at the same temperature, the molecules have the same average kinetic energy.

The **internal energy** of a thermodynamic system is the sum of the mechanical energies of all of the molecules in it. We can now give an equation for the internal energy of a monatomic ideal gas. In such a gas, the molecules' only energy is their translational kinetic energy. Therefore, denoting the internal energy by E_{int}, we simply have $E_{int} = N\overline{K}$, or

$$E_{int} = \frac{3}{2}Nk_BT. \quad (17.8)$$

Often we would like to use this equation in terms of moles:

$$E_{int} = \frac{3}{2}nRT.$$

We can solve $\bar{K} = \frac{1}{2} m\overline{v^2} = \frac{3}{2} k_B T$ for a typical speed of a molecule in an ideal gas in terms of temperature to determine what is known as the *root-mean-square* (*rms*) *speed* of a molecule.

RMS Speed of a Molecule

The **root-mean-square (rms) speed** of a molecule, or the square root of the average of the square of the speed $\overline{v^2}$, is

$$v_{\text{rms}} = \sqrt{\overline{v^2}} = \sqrt{\frac{3k_B T}{m}}. \tag{17.9}$$

The rms speed is not the average or the most likely speed of molecules, as we will see in **Distribution of Molecular Speeds**, but it provides an easily calculated estimate of the molecules' speed that is related to their kinetic energy. Again we can write this equation in terms of the gas constant R and the molar mass M in kg/mol:

$$v_{\text{rms}} = \sqrt{\frac{3RT}{M}}. \tag{17.10}$$

We digress for a moment to answer a question that may have occurred to you: When we apply the model to atoms instead of theoretical point particles, does rotational kinetic energy change our results? To answer this question, we have to appeal to quantum mechanics. In quantum mechanics, rotational kinetic energy cannot take on just any value; it's limited to a discrete set of values, and the smallest value is inversely proportional to the rotational inertia. The rotational inertia of an atom is tiny because almost all of its mass is in the nucleus, which typically has a radius less than 10^{-14} m. Thus the minimum rotational energy of an atom is much more than $\frac{1}{2} k_B T$ for any attainable temperature, and the energy available is not enough to make an atom rotate. We will return to this point when discussing diatomic and polyatomic gases in the next section.

Example 17.4

Calculating Kinetic Energy and Speed of a Gas Molecule

(a) What is the average kinetic energy of a gas molecule at $20.0\,°\text{C}$ (room temperature)? (b) Find the rms speed of a nitrogen molecule (N_2) at this temperature.

Strategy

(a) The known in the equation for the average kinetic energy is the temperature:

$$\bar{K} = \frac{1}{2} m\overline{v^2} = \frac{3}{2} k_B T.$$

Before substituting values into this equation, we must convert the given temperature into kelvin: $T = (20.0 + 273)\text{ K} = 293\text{ K}$. We can find the rms speed of a nitrogen molecule by using the equation

$$v_{\text{rms}} = \sqrt{\overline{v^2}} = \sqrt{\frac{3k_B T}{m}},$$

but we must first find the mass of a nitrogen molecule. Obtaining the molar mass of nitrogen N_2 from the periodic table, we find

$$m = \frac{M}{N_A} = \frac{2\,(14.0067) \times 10^{-3}\text{ kg/mol})}{6.02 \times 10^{23}\text{ mol}^{-1}} = 4.65 \times 10^{-26}\text{ kg}.$$

Solution

a. The temperature alone is sufficient for us to find the average translational kinetic energy. Substituting the temperature into the translational kinetic energy equation gives

$$\bar{K} = \frac{3}{2}k_B T = \frac{3}{2}(1.38 \times 10^{-23}\ \text{J/K})(293\ \text{K}) = 6.07 \times 10^{-21}\ \text{J}.$$

b. Substituting this mass and the value for k_B into the equation for v_{rms} yields

$$v_{rms} = \sqrt{\frac{3k_B T}{m}} = \sqrt{\frac{3(1.38 \times 10^{-23}\ \text{J/K})(293\ \text{K})}{4.65 \times 10^{-26}\ \text{kg}}} = 511\ \text{m/s}.$$

Significance

Note that the average kinetic energy of the molecule is independent of the type of molecule. The average translational kinetic energy depends only on absolute temperature. The kinetic energy is very small compared to macroscopic energies, so that we do not feel when an air molecule is hitting our skin. On the other hand, it is much greater than the typical difference in gravitational potential energy when a molecule moves from, say, the top to the bottom of a room, so our neglect of gravitation is justified in typical real-world situations. The rms speed of the nitrogen molecule is surprisingly large. These large molecular velocities do not yield macroscopic movement of air, since the molecules move in all directions with equal likelihood. The *mean free path* (the distance a molecule moves on average between collisions, discussed a bit later in this section) of molecules in air is very small, so the molecules move rapidly but do not get very far in a second. The high value for rms speed is reflected in the speed of sound, which is about 340 m/s at room temperature. The higher the rms speed of air molecules, the faster sound vibrations can be transferred through the air. The speed of sound increases with temperature and is greater in gases with small molecular masses, such as helium (see **Figure 17.13**).

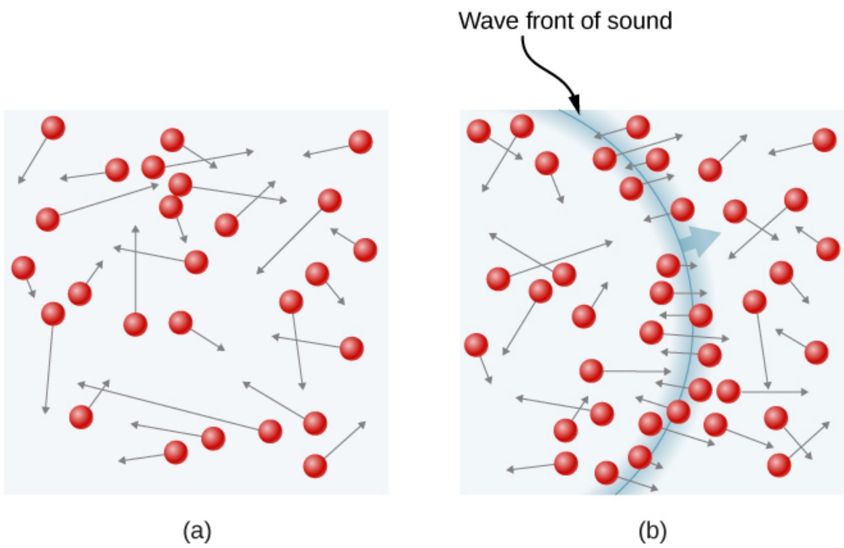


Figure 17.13 (a) In an ordinary gas, so many molecules move so fast that they collide billions of times every second. (b) Individual molecules do not move very far in a small amount of time, but disturbances like sound waves are transmitted at speeds related to the molecular speeds.

Example 17.5

Calculating Temperature: Escape Velocity of Helium Atoms

To escape Earth's gravity, an object near the top of the atmosphere (at an altitude of 100 km) must travel away from Earth at 11.1 km/s. This speed is called the *escape velocity*. At what temperature would helium atoms have an rms speed equal to the escape velocity?

Strategy

Identify the knowns and unknowns and determine which equations to use to solve the problem.

Solution

1. Identify the knowns: v is the escape velocity, 11.1 km/s.

2. Identify the unknowns: We need to solve for temperature, T. We also need to solve for the mass m of the helium atom.
3. Determine which equations are needed.
   - To get the mass m of the helium atom, we can use information from the periodic table:
   $$m = \frac{M}{N_A}.$$
   - To solve for temperature T, we can rearrange
   $$\frac{1}{2}m\overline{v^2} = \frac{3}{2}k_B T$$
   to yield
   $$T = \frac{m\overline{v^2}}{3k_B}.$$
4. Substitute the known values into the equations and solve for the unknowns,
$$m = \frac{M}{N_A} = \frac{4.0026 \times 10^{-3}\ \text{kg/mol}}{6.02 \times 10^{23}\ \text{mol}} = 6.65 \times 10^{-27}\ \text{kg}$$
and
$$T = \frac{(6.65 \times 10^{-27}\ \text{kg})\left(11.1 \times 10^{3}\ \text{m/s}\right)^2}{3\,(1.38 \times 10^{-23}\ \text{J/K})} = 1.98 \times 10^{4}\ \text{K}.$$

Significance

This temperature is much higher than atmospheric temperature, which is approximately 250 K $(-25\ °\text{C or } -10\ °\text{F})$ at high elevation. Very few helium atoms are left in the atmosphere, but many were present when the atmosphere was formed, and more are always being created by radioactive decay (see the chapter on nuclear physics). The reason for the loss of helium atoms is that a small number of helium atoms have speeds higher than Earth's escape velocity even at normal temperatures. The speed of a helium atom changes from one collision to the next, so that at any instant, there is a small but nonzero chance that the atom's speed is greater than the escape velocity. The chance is high enough that over the lifetime of Earth, almost all the helium atoms that have been in the atmosphere have reached escape velocity at high altitudes and escaped from Earth's gravitational pull. Heavier molecules, such as oxygen, nitrogen, and water, have smaller rms speeds, and so it is much less likely that any of them will have speeds greater than the escape velocity. In fact, the likelihood is so small that billions of years are required to lose significant amounts of heavier molecules from the atmosphere. **Figure 17.14** shows the effect of a lack of an atmosphere on the Moon. Because the gravitational pull of the Moon is much weaker, it has lost almost its entire atmosphere. The atmospheres of Earth and other bodies are compared in this chapter's exercises.

Figure 17.14 This photograph of Apollo 17 Commander Eugene Cernan driving the lunar rover on the Moon in 1972 looks as though it was taken at night with a large spotlight. In fact, the light is coming from the Sun. Because the acceleration due to gravity on the Moon is so low (about 1/6 that of Earth), the Moon's escape velocity is much smaller. As a result, gas molecules escape very easily from the Moon, leaving it with virtually no atmosphere. Even during the daytime, the sky is black because there is no gas to scatter sunlight. (credit: Harrison H. Schmitt/NASA)

17.4 Check Your Understanding If you consider a very small object, such as a grain of pollen, in a gas, then the number of molecules striking its surface would also be relatively small. Would you expect the grain of pollen to experience any fluctuations in pressure due to statistical fluctuations in the number of gas molecules striking it in a given amount of time?

Mean Free Path and Mean Free Time

We now consider collisions explicitly. The usual first step (which is all we'll take) is to calculate the **mean free path**, λ, the average distance a molecule travels between collisions with other molecules, and the *mean free time* τ, the average time between the collisions of a molecule. If we assume all the molecules are spheres with a radius *r*, then a molecule will collide with another if their centers are within a distance 2*r* of each other. For a given particle, we say that the area of a circle with that radius, $4\pi r^2$, is the "cross-section" for collisions. As the particle moves, it traces a cylinder with that cross-sectional area. The mean free path is the length λ such that the expected number of other molecules in a cylinder of length λ and cross-section $4\pi r^2$ is 1. If we temporarily ignore the motion of the molecules other than the one we're looking at, the expected number is the number density of molecules, *N*/*V*, times the volume, and the volume is $4\pi r^2 \lambda$, so we have $(N/V)4\pi r^2 \lambda = 1$, or

$$\lambda = \frac{V}{4\pi r^2 N}.$$

Taking the motion of all the molecules into account makes the calculation much harder, but the only change is a factor of $\sqrt{2}$. The result is

$$\lambda = \frac{V}{4\sqrt{2}\pi r^2 N}. \quad (17.11)$$

In an ideal gas, we can substitute $V/N = k_B T/p$ to obtain

$$\lambda = \frac{k_B T}{4\sqrt{2}\pi r^2 p}. \quad (17.12)$$

The **mean free time** τ is simply the mean free path divided by a typical speed, and the usual choice is the rms speed. Then

$$\tau = \frac{k_B T}{4\sqrt{2}\pi r^2 p v_{rms}}. \quad (17.13)$$

Example 17.6

Calculating Mean Free Time

Find the mean free time for argon atoms $(M = 39.9 \text{ g/mol})$ at a temperature of $0\,°\text{C}$ and a pressure of 1.00 atm. Take the radius of an argon atom to be 1.70×10^{-10} m.

Solution

1. Identify the knowns and convert into SI units. We know the molar mass is 0.0399 kg/mol, the temperature is 273 K, the pressure is 1.01×10^5 Pa, and the radius is 1.70×10^{-10} m.
2. Find the rms speed: $v_{rms} = \sqrt{\frac{3RT}{M}} = 413 \frac{\text{m}}{\text{s}}$.
3. Substitute into the equation for the mean free time:

$$\tau = \frac{k_B T}{4\sqrt{2}\pi r^2 p v_{rms}} = \frac{(1.38 \times 10^{-23} \text{ J/K})\,(273 \text{ K})}{4\sqrt{2}\pi(1.70 \times 10^{-10} \text{ m})^2(1.01 \times 10^5 \text{ Pa})(413 \text{ m/s})} = 1.76 \times 10^{-10} \text{ s}.$$

Significance

We can hardly compare this result with our intuition about gas molecules, but it gives us a picture of molecules colliding with extremely high frequency.

17.5 Check Your Understanding Which has a longer mean free path, liquid water or water vapor in the air?

17.4 | Distribution of Molecular Speeds

Learning Objectives

By the end of this section, you will be able to:

- Describe the distribution of molecular speeds in an ideal gas
- Find the average and most probable molecular speeds in an ideal gas

Particles in an ideal gas all travel at relatively high speeds, but they do not travel at the same speed. The rms speed is one kind of average, but many particles move faster and many move slower. The actual distribution of speeds has several interesting implications for other areas of physics, as we will see in later chapters.

The Maxwell-Boltzmann Distribution

The motion of molecules in a gas is random in magnitude and direction for individual molecules, but a gas of many molecules has a predictable distribution of molecular speeds. This predictable distribution of molecular speeds is known as the **Maxwell-Boltzmann distribution**, after its originators, who calculated it based on kinetic theory, and it has since been confirmed experimentally (**Figure 17.15**).

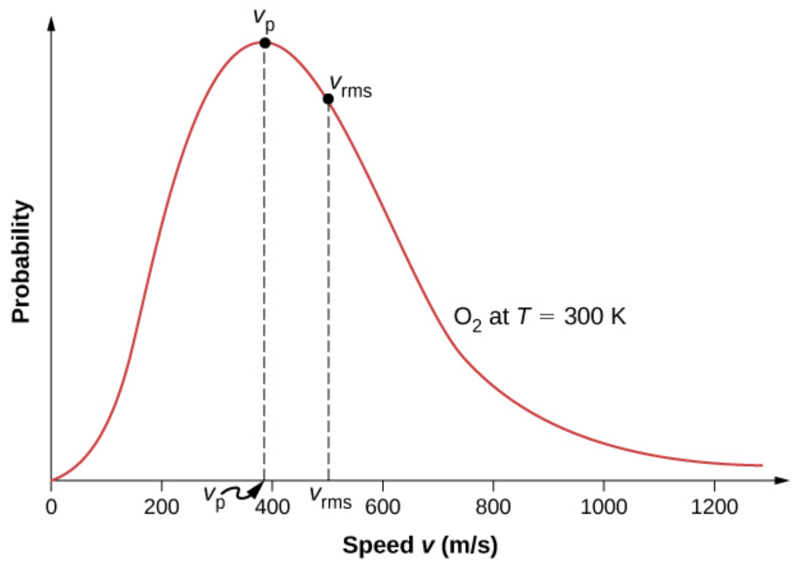


Figure 17.15 The Maxwell-Boltzmann distribution of molecular speeds in an ideal gas. The most likely speed v_p is less than the rms speed v_{rms}. Although very high speeds are possible, only a tiny fraction of the molecules have speeds that are an order of magnitude greater than v_{rms}.

We will quote Maxwell's result, although the proof is beyond our scope.

Maxwell-Boltzmann Distribution of Speeds

The distribution function for speeds of particles in an ideal gas at temperature T is

$$f(v) = \frac{4}{\sqrt{\pi}}\left(\frac{m}{2k_B T}\right)^{3/2} v^2 e^{-mv^2/2k_B T}. \quad (17.14)$$

Figure 17.16 shows that the curve is shifted to higher speeds at higher temperatures, with a broader range of speeds.

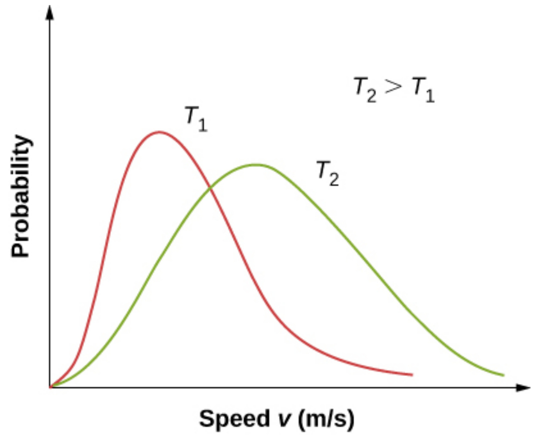


Figure 17.16 The Maxwell-Boltzmann distribution is shifted to higher speeds and broadened at higher temperatures.

With only a relatively small number of molecules, the distribution of speeds fluctuates around the Maxwell-Boltzmann distribution. However, you can view this **simulation (https://openstax.org/l/21maxboltzdisim)** to see the essential features that more massive molecules move slower and have a narrower distribution. Use the set-up “2 Gases, Random Speeds”. Note the display at the bottom comparing histograms of the speed distributions with the theoretical curves.

In fact, the rms speed is greater than both the most probable speed and the average speed.

The peak speed provides a sometimes more convenient way to write the Maxwell-Boltzmann distribution function:

$$f(v) = \frac{4v^2}{\sqrt{\pi}v_p^3}e^{-v^2/v_p^2} \tag{17.15}$$

In the factor $e^{-mv^2/2k_\text{B}T}$, it is easy to recognize the translational kinetic energy. Thus, that expression is equal to $e^{-K/k_\text{B}T}$. Boltzmann showed that the resulting formula is much more generally applicable if we replace the kinetic energy of translation with the total mechanical energy E. Boltzmann's result is

$$f(E) = \frac{2}{\sqrt{\pi}}(k_\text{B}T)^{-3/2}\sqrt{E}e^{-E/k_\text{B}T} = \frac{2}{\sqrt{\pi}(k_\text{B}T)^{3/2}}\frac{\sqrt{E}}{e^{E/k_\text{B}T}}.$$

The first part of this equation, with the negative exponential, is the usual way to write it. We give the second part only to remark that $e^{E/k_\text{B}T}$ in the denominator is ubiquitous in quantum as well as classical statistical mechanics.

Problem-Solving Strategy: Speed Distribution

Step 1. Examine the situation to determine that it relates to the distribution of molecular speeds.

Step 2. Make a list of what quantities are given or can be inferred from the problem as stated (identify the known quantities).

Step 3. Identify exactly what needs to be determined in the problem (identify the unknown quantities). A written list is useful.

Step 4. Convert known values into proper SI units (K for temperature, Pa for pressure, m^3 for volume, molecules for N, and moles for n). In many cases, though, using R and the molar mass will be more convenient than using k_B and the molecular mass.

Step 5. Determine whether you need the distribution function for velocity or the one for energy, and whether you are using a formula for one of the characteristic speeds (average, most probably, or rms), finding a ratio of values of the distribution function, or approximating an integral.

Step 6. Solve the appropriate equation for the ideal gas law for the quantity to be determined (the unknown quantity). Note that if you are taking a ratio of values of the distribution function, the normalization factors divide out. Or if approximating an integral, use the method asked for in the problem.

Step 7. Substitute the known quantities, along with their units, into the appropriate equation and obtain numerical solutions complete with units.

We can now gain a qualitative understanding of a puzzle about the composition of Earth's atmosphere. Hydrogen is by far the most common element in the universe, and helium is by far the second-most common. Moreover, helium is constantly produced on Earth by radioactive decay. Why are those elements so rare in our atmosphere? The answer is that gas molecules that reach speeds above Earth's escape velocity, about 11 km/s, can escape from the atmosphere into space. Because of the lower mass of hydrogen and helium molecules, they move at higher speeds than other gas molecules, such as nitrogen and oxygen. Only a few exceed escape velocity, but far fewer heavier molecules do. Thus, over the billions of years that Earth has existed, far more hydrogen and helium molecules have escaped from the atmosphere than other molecules, and hardly any of either is now present.

We can also now take another look at evaporative cooling, which we discussed in the chapter on temperature and heat. Liquids, like gases, have a distribution of molecular energies. The highest-energy molecules are those that can escape from the intermolecular attractions of the liquid. Thus, when some liquid evaporates, the molecules left behind have a lower average energy, and the liquid has a lower temperature.

CHAPTER 17 REVIEW

KEY TERMS

Avogadro's number N_A, the number of molecules in one mole of a substance; $N_A = 6.02 \times 10^{23}$ particles/mole

Boltzmann constant k_B, a physical constant that relates energy to temperature and appears in the ideal gas law; $k_B = 1.38 \times 10^{-23}$ J/K

density mass per unit volume of a substance or object

fluids liquids and gases; a fluid is a state of matter that yields to shearing forces

gauge pressure the absolute pressure minus the ambient pressure

ideal gas gas at the limit of low density and high temperature

ideal gas law physical law that relates the pressure and volume of a gas, far from liquefaction, to the number of gas molecules or number of moles of gas and the temperature of the gas

internal energy sum of the mechanical energies of all of the molecules in it

kinetic theory of gases theory that derives the macroscopic properties of gases from the motion of the molecules they consist of

Maxwell-Boltzmann distribution function that can be integrated to give the probability of finding ideal gas molecules with speeds in the range between the limits of integration

mean free path average distance between collisions of a particle

mean free time average time between collisions of a particle

mole quantity of a substance whose mass (in grams) is equal to its molecular mass

most probable speed speed near which the speeds of most molecules are found, the peak of the speed distribution function

peak speed same as "most probable speed"

pressure force per unit area exerted perpendicular to the area over which the force acts

root-mean-square (rms) speed square root of the average of the square (of a quantity)

specific gravity ratio of the density of an object to a fluid (usually water)

universal gas constant R, the constant that appears in the ideal gas law expressed in terms of moles, given by $R = N_A k_B$

KEY EQUATIONS

Ideal gas law in terms of molecules	$pV = Nk_B T$
Ideal gas law ratios if the amount of gas is constant	$\frac{p_1 V_1}{T_1} = \frac{p_2 V_2}{T_2}$
Ideal gas law in terms of moles	$pV = nRT$
Pressure, volume, and molecular speed	$pV = \frac{1}{3}Nm\overline{v^2}$
Root-mean-square speed	$v_{\text{rms}} = \sqrt{\frac{3RT}{M}} = \sqrt{\frac{3k_B T}{m}}$

Mean free path	$\lambda = \frac{V}{4\sqrt{2}\pi r^2 N} = \frac{k_B T}{4\sqrt{2}\pi r^2 p}$
Mean free time	$\tau = \frac{k_B T}{4\sqrt{2}\pi r^2 p v_{\text{rms}}}$

SUMMARY

17.1 Fluids, Density, and Pressure

- A fluid is a state of matter that yields to sideways or shearing forces. Liquids and gases are both fluids.
- Density is the mass per unit volume of a substance or object, defined as $\rho = m/V$. The SI unit of density is kg/m^3.
- Pressure is the force per unit perpendicular area over which the force is applied, $p = F/A$. The SI unit of pressure is the pascal: $1 \text{ Pa} = 1 \text{ N/m}^2$.

17.2 Molecular Model of an Ideal Gas

- The ideal gas law relates the pressure and volume of a gas to the number of gas molecules and the temperature of the gas.
- A mole of any substance has a number of molecules equal to the number of atoms in a 12-g sample of carbon-12. The number of molecules in a mole is called Avogadro's number N_A,

$$N_A = 6.02 \times 10^{23} \text{ mol}^{-1}.$$

- A mole of any substance has a mass in grams numerically equal to its molecular mass in unified mass units, which can be determined from the periodic table of elements. The ideal gas law can also be written and solved in terms of the number of moles of gas:

$$pV = nRT,$$

where n is the number of moles and R is the universal gas constant,

$$R = 8.31 \text{ J/mol} \cdot \text{K}.$$

- The ideal gas law is generally valid at temperatures well above the boiling temperature.

17.3 Pressure, Temperature, and RMS Speed

- Kinetic theory is the atomic description of gases as well as liquids and solids. It models the properties of matter in terms of continuous random motion of molecules.
- The ideal gas law can be expressed in terms of the mass of the gas's molecules and $\overline{v^2}$, the average of the molecular speed squared, instead of the temperature.
- The temperature of gases is proportional to the average translational kinetic energy of molecules. Hence, the typical speed of gas molecules v_{rms} is proportional to the square root of the temperature and inversely proportional to the square root of the molecular mass.
- In a mixture of gases, each gas exerts a pressure equal to the total pressure times the fraction of the mixture that the gas makes up.
- The mean free path (the average distance between collisions) and the mean free time of gas molecules are proportional to the temperature and inversely proportional to the molar density and the molecules' cross-sectional area.

17.4 Distribution of Molecular Speeds

- The motion of individual molecules in a gas is random in magnitude and direction. However, a gas of many molecules has a predictable distribution of molecular speeds, known as the Maxwell-Boltzmann distribution.

- The average and most probable velocities of molecules having the Maxwell-Boltzmann speed distribution, as well as the rms velocity, can be calculated from the temperature and molecular mass.

CONCEPTUAL QUESTIONS

17.1 Fluids, Density, and Pressure

1. Which of the following substances are fluids at room temperature and atmospheric pressure: air, mercury, water, glass?

2. Why are gases easier to compress than liquids and solids?

3. How is pressure related to the sharpness of a knife and its ability to cut?

4. Why is a force exerted by a static fluid on a surface always perpendicular to the surface?

5. Imagine that in a remote location near the North Pole, a chunk of ice floats in a lake. Next to the lake, a glacier with the same volume as the floating ice sits on land. If both chunks of ice should melt due to rising global temperatures, and the melted ice all goes into the lake, which one would cause the level of the lake to rise the most? Explain.

6. In ballet, dancing *en pointe* (on the tips of the toes) is much harder on the toes than normal dancing or walking. Explain why, in terms of pressure.

7. Atmospheric pressure exerts a large force (equal to the weight of the atmosphere above your body—about 10 tons) on the top of your body when you are lying on the beach sunbathing. Why are you able to get up?

8. You can break a strong wine bottle by pounding a cork into it with your fist, but the cork must press directly against the liquid filling the bottle—there can be no air between the cork and liquid. Explain why the bottle breaks only if there is no air between the cork and liquid.

17.2 Molecular Model of an Ideal Gas

9. Two H_2 molecules can react with one O_2 molecule to produce two H_2O molecules. How many moles of hydrogen molecules are needed to react with one mole of oxygen molecules?

10. Under what circumstances would you expect a gas to behave significantly differently than predicted by the ideal gas law?

11. A constant-volume gas thermometer contains a fixed amount of gas. What property of the gas is measured to indicate its temperature?

12. Inflate a balloon at room temperature. Leave the inflated balloon in the refrigerator overnight. What happens to the balloon, and why?

13. In the last chapter, free convection was explained as the result of buoyant forces on hot fluids. Explain the upward motion of air in flames based on the ideal gas law.

17.3 Pressure, Temperature, and RMS Speed

14. How is momentum related to the pressure exerted by a gas? Explain on the molecular level, considering the behavior of molecules.

15. If one kind of molecule has double the radius of another and eight times the mass, how do their mean free paths under the same conditions compare? How do their mean free times compare?

16. What is the average *velocity* of the air molecules in the room where you are right now?

17. Why do the atmospheres of Jupiter, Saturn, Uranus, and Neptune, which are much more massive and farther from the Sun than Earth is, contain large amounts of hydrogen and helium?

18. Statistical mechanics says that in a gas maintained at a constant temperature through thermal contact with a bigger system (a "reservoir") at that temperature, the fluctuations in internal energy are typically a fraction $1/\sqrt{N}$ of the internal energy. As a fraction of the total internal energy of a mole of gas, how big are the fluctuations in the internal energy? Are we justified in ignoring them?

17.4 Distribution of Molecular Speeds

19. One cylinder contains helium gas and another contains krypton gas at the same temperature. Mark each of these statements true, false, or impossible to determine from the given information. (a) The rms speeds of atoms in the two gases are the same. (b) The average kinetic energies of atoms in the two gases are the same. (c) The internal energies of 1 mole of gas in each cylinder are the same. (d) The pressures in the two cylinders are the same.

20. Repeat the previous question if one gas is still helium but the other is changed to fluorine, F_2.

21. An ideal gas is at a temperature of 300 K. To double the average speed of its molecules, what does the temperature need to be changed to?

PROBLEMS

17.1 Fluids, Density, and Pressure

22. Gold is sold by the troy ounce (31.103 g). What is the volume of 1 troy ounce of pure gold?

23. Mercury is commonly supplied in flasks containing 34.5 kg (about 76 lb.). What is the volume in liters of this much mercury?

24. What is the mass of a deep breath of air having a volume of 2.00 L? Discuss the effect taking such a breath has on your body's volume and density.

25. A straightforward method of finding the density of an object is to measure its mass and then measure its volume by submerging it in a graduated cylinder. What is the density of a 240-g rock that displaces $89.0\,\text{cm}^3$ of water? (Note that the accuracy and practical applications of this technique are more limited than a variety of others that are based on Archimedes' principle.)

26. Suppose you have a coffee mug with a circular cross-section and vertical sides (uniform radius). What is its inside radius if it holds 375 g of coffee when filled to a depth of 7.50 cm? Assume coffee has the same density as water.

27. A rectangular gasoline tank can hold 50.0 kg of gasoline when full. What is the depth of the tank if it is 0.500-m wide by 0.900-m long? (b) Discuss whether this gas tank has a reasonable volume for a passenger car.

28. A trash compactor can compress its contents to 0.350 times their original volume. Neglecting the mass of air expelled, by what factor is the density of the rubbish increased?

29. A 2.50-kg steel gasoline can holds 20.0 L of gasoline when full. What is the average density of the full gas can, taking into account the volume occupied by steel as well as by gasoline?

30. The tip of a nail exerts tremendous pressure when hit by a hammer because it exerts a large force over a small area. What force must be exerted on a nail with a circular tip of 1.00-mm diameter to create a pressure of $3.00 \times 10^9\ \text{N/m}^2$? (This high pressure is possible because the hammer striking the nail is brought to rest in such a short distance.)

17.2 Molecular Model of an Ideal Gas

31. The gauge pressure in your car tires is $2.50 \times 10^5\ \text{N/m}^2$ at a temperature of $35.0\ °\text{C}$ when you drive it onto a ship in Los Angeles to be sent to Alaska. What is their gauge pressure on a night in Alaska when their temperature has dropped to $-40.0\ °\text{C}$? Assume the tires have not gained or lost any air.

32. Suppose a gas-filled incandescent light bulb is manufactured so that the gas inside the bulb is at atmospheric pressure when the bulb has a temperature of $20.0\ °\text{C}$. (a) Find the gauge pressure inside such a bulb when it is hot, assuming its average temperature is $60.0\ °\text{C}$ (an approximation) and neglecting any change in volume due to thermal expansion or gas leaks. (b) The actual final pressure for the light bulb will be less than calculated in part (a) because the glass bulb will expand. Is this effect significant?

33. People buying food in sealed bags at high elevations often notice that the bags are puffed up because the air inside has expanded. A bag of pretzels was packed at a pressure of 1.00 atm and a temperature of $22.0\ °\text{C}$. When opened at a summer picnic in Santa Fe, New Mexico, at a temperature of $32.0\ °\text{C}$, the volume of the air in the bag is 1.38 times its original volume. What is the pressure of the air?

34. How many moles are there in (a) 0.0500 g of N_2 gas $(M = 28.0\ \text{g/mol})$? (b) 10.0 g of CO_2 gas $(M = 44.0\ \text{g/mol})$? (c) How many molecules are present in each case?

35. A cubic container of volume 2.00 L holds 0.500 mol of nitrogen gas at a temperature of $25.0\ °\text{C}$. What is the net force due to the nitrogen on one wall of the container? Compare that force to the sample's weight.

36. Calculate the number of moles in the 2.00-L volume of air in the lungs of the average person. Note that the air is at $37.0\ °\text{C}$ (body temperature) and that the total volume in the lungs is several times the amount inhaled in a typical breath as given in **Example 17.2**.

37. An airplane passenger has $100\,\text{cm}^3$ of air in his stomach just before the plane takes off from a sea-level airport. What volume will the air have at cruising altitude if cabin pressure drops to $7.50 \times 10^4\ \text{N/m}^2$?

38. A company advertises that it delivers helium at a gauge pressure of $1.72 \times 10^7\ \text{Pa}$ in a cylinder of volume 43.8 L. How many balloons can be inflated to a volume of 4.00 L with that amount of helium? Assume the pressure inside the balloons is $1.01 \times 10^5\ \text{Pa}$ and the temperature in the cylinder and the balloons is $25.0\,°\text{C}$.

39. According to http://hyperphysics.phy-astr.gsu.edu/hbase/solar/venusenv.html, the atmosphere of Venus is approximately $96.5\%\ CO_2$ and $3.5\%\ N_2$ by volume. On the surface, where the temperature is about 750 K and the pressure is about 90 atm, what is the density of the atmosphere?

40. An expensive vacuum system can achieve a pressure as low as $1.00 \times 10^{-7}\ \text{N/m}^2$ at $20.0\,°\text{C}$. How many molecules are there in a cubic centimeter at this pressure and temperature?

41. The number density *N*/*V* of gas molecules at a certain location in the space above our planet is about $1.00 \times 10^{11}\ \text{m}^{-3}$, and the pressure is $2.75 \times 10^{-10}\ \text{N/m}^2$ in this space. What is the temperature there?

42. A bicycle tire contains 2.00 L of gas at an absolute pressure of $7.00 \times 10^5\ \text{N/m}^2$ and a temperature of $18.0\,°\text{C}$. What will its pressure be if you let out an amount of air that has a volume of $100\,\text{cm}^3$ at atmospheric pressure? Assume tire temperature and volume remain constant.

43. In a common demonstration, a bottle is heated and stoppered with a hard-boiled egg that's a little bigger than the bottle's neck. When the bottle is cooled, the pressure difference between inside and outside forces the egg into the bottle. Suppose the bottle has a volume of 0.500 L and the temperature inside it is raised to $80.0\,°\text{C}$ while the pressure remains constant at 1.00 atm because the bottle is open. (a) How many moles of air are inside? (b) Now the egg is put in place, sealing the bottle. What is the gauge pressure inside after the air cools back to the ambient temperature of $25\,°\text{C}$ but before the egg is forced into the bottle?

44. A high-pressure gas cylinder contains 50.0 L of toxic gas at a pressure of $1.40 \times 10^7\ \text{N/m}^2$ and a temperature of $25.0\,°\text{C}$. The cylinder is cooled to dry ice temperature $(-78.5\,°\text{C})$ to reduce the leak rate and pressure so that it can be safely repaired. (a) What is the final pressure in the tank, assuming a negligible amount of gas leaks while being cooled and that there is no phase change? (b) What is the final pressure if one-tenth of the gas escapes? (c) To what temperature must the tank be cooled to reduce the pressure to 1.00 atm (assuming the gas does not change phase and that there is no leakage during cooling)? (d) Does cooling the tank as in part (c) appear to be a practical solution?

45. Find the number of moles in 2.00 L of gas at $35.0\,°\text{C}$ and under $7.41 \times 10^7\ \text{N/m}^2$ of pressure.

46. Calculate the depth to which Avogadro's number of table tennis balls would cover Earth. Each ball has a diameter of 3.75 cm. Assume the space between balls adds an extra 25.0% to their volume and assume they are not crushed by their own weight.

47. (a) What is the gauge pressure in a $25.0\,°\text{C}$ car tire containing 3.60 mol of gas in a 30.0-L volume? (b) What will its gauge pressure be if you add 1.00 L of gas originally at atmospheric pressure and $25.0\,°\text{C}$? Assume the temperature remains at $25.0\,°\text{C}$ and the volume remains constant.

17.3 Pressure, Temperature, and RMS Speed

In the problems in this section, assume all gases are ideal.

48. A person hits a tennis ball with a mass of 0.058 kg against a wall. The average component of the ball's velocity perpendicular to the wall is 11 m/s, and the ball hits the wall every 2.1 s on average, rebounding with the opposite perpendicular velocity component. (a) What is the average force exerted on the wall? (b) If the part of the wall the person hits has an area of $3.0\,\text{m}^2$, what is the average pressure on that area?

49. A person is in a closed room (a racquetball court) with $V = 453\,\text{m}^3$ hitting a ball $(m = 42.0\,\text{g})$ around at random without any pauses. The average kinetic energy of the ball is 2.30 J. (a) What is the average value of v_x^2? Does it matter which direction you take to be *x*? (b) Applying the methods of this chapter, find the average pressure on the walls? (c) Aside from the presence of only one "molecule" in this problem, what is the main assumption of kinetic theory that does not apply here?

50. Five bicyclists are riding at the following speeds: 5.4 m/s, 5.7 m/s, 5.8 m/s, 6.0 m/s, and 6.5 m/s. (a) What is their average speed? (b) What is their rms speed?

51. Some incandescent light bulbs are filled with argon gas. What is v_{rms} for argon atoms near the filament, assuming their temperature is 2500 K?

52. Typical molecular speeds (v_{rms}) are large, even at low temperatures. What is v_{rms} for helium atoms at 5.00 K, less than one degree above helium's liquefaction temperature?

53. What is the average kinetic energy in joules of hydrogen atoms on the $5500\,°\text{C}$ surface of the Sun? (b) What is the average kinetic energy of helium atoms in a region of the solar corona where the temperature is $6.00 \times 10^5\ \text{K}$?

54. What is the ratio of the average translational kinetic energy of a nitrogen molecule at a temperature of 300 K to the gravitational potential energy of a nitrogen-molecule−Earth system at the ceiling of a 3-m-tall room with respect to the same system with the molecule at the floor?

55. What is the total translational kinetic energy of the air molecules in a room of volume $23\ \text{m}^3$ if the pressure is $9.5 \times 10^4\ \text{Pa}$ (the room is at fairly high elevation) and the temperature is $21\ °\text{C}$? Is any item of data unnecessary for the solution?

56. The product of the pressure and volume of a sample of hydrogen gas at $0.00\ °\text{C}$ is 80.0 J. (a) How many moles of hydrogen are present? (b) What is the average translational kinetic energy of the hydrogen molecules? (c) What is the value of the product of pressure and volume at $200\ °\text{C}$?

57. What is the gauge pressure inside a tank of $4.86 \times 10^4\ \text{mol}$ of compressed nitrogen with a volume of $6.56\ \text{m}^3$ if the rms speed is 514 m/s?

58. The escape velocity of any object from Earth is 11.1 km/s. At what temperature would oxygen molecules (molar mass is equal to 32.0 g/mol) have root-mean-square velocity v_{rms} equal to Earth's escape velocity of 11.1 km/s?

59. The escape velocity from the Moon is much smaller than that from the Earth, only 2.38 km/s. At what temperature would hydrogen molecules (molar mass is equal to 2.016 g/mol) have a root-mean-square velocity v_{rms} equal to the Moon's escape velocity?

60. Nuclear fusion, the energy source of the Sun, hydrogen bombs, and fusion reactors, occurs much more readily when the average kinetic energy of the atoms is high—that is, at high temperatures. Suppose you want the atoms in your fusion experiment to have average kinetic energies of $6.40 \times 10^{-14}\ \text{J}$. What temperature is needed?

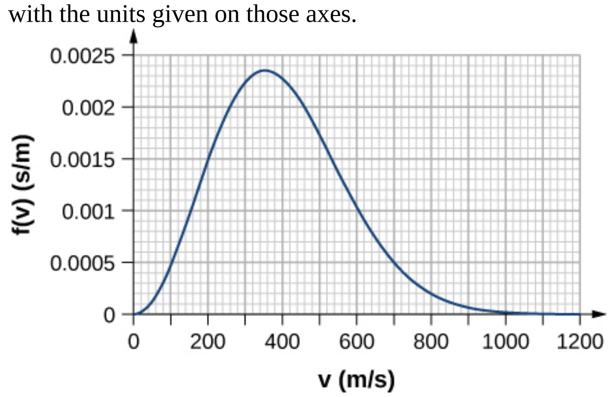
61. Suppose that the typical speed (v_{rms}) of carbon dioxide molecules (molar mass is 44.0 g/mol) in a flame is found to be 1350 m/s. What temperature does this indicate?

62. (a) Hydrogen molecules (molar mass is equal to 2.016 g/mol) have v_{rms} equal to 193 m/s. What is the temperature? (b) Much of the gas near the Sun is atomic hydrogen (H rather than H_2). Its temperature would have to be $1.5 \times 10^7\ \text{K}$ for the rms speed v_{rms} to equal the escape velocity from the Sun. What is that velocity?

63. There are two important isotopes of uranium, ^{235}U and ^{238}U ; these isotopes are nearly identical chemically but have different atomic masses. Only ^{235}U is very useful in nuclear reactors. Separating the isotopes is called uranium enrichment (and is often in the news as of this writing, because of concerns that some countries are enriching uranium with the goal of making nuclear weapons.) One of the techniques for enrichment, gas diffusion, is based on the different molecular speeds of uranium hexafluoride gas, UF_6 . (a) The molar masses of ^{235}U and $^{238}\text{UF}_6$ are 349.0 g/mol and 352.0 g/mol, respectively. What is the ratio of their typical speeds v_{rms} ? (b) At what temperature would their typical speeds differ by 1.00 m/s? (c) Do your answers in this problem imply that this technique may be difficult?

17.4 Distribution of Molecular Speeds

64. By counting squares in the following figure, estimate the fraction of argon atoms at $T = 300\ \text{K}$ that have speeds between 600 m/s and 800 m/s. The curve is correctly normalized. The value of a square is its length as measured on the *x*-axis times its height as measured on the *y*-axis, with the units given on those axes.



65. Find (a) the most probable speed, (b) the average speed, and (c) the rms speed for nitrogen molecules at 295 K.

66. Repeat the preceding problem for nitrogen molecules at 2950 K.

67. At what temperature is the average speed of carbon dioxide molecules $(M = 44.0 \text{ g/mol})$ 510 m/s?

68. The most probable speed for molecules of a gas at 296 K is 263 m/s. What is the molar mass of the gas? (You might like to figure out what the gas is likely to be.)

69. a) At what temperature do oxygen molecules have the same average speed as helium atoms $(M = 4.00 \text{ g/mol})$ have at 300 K? b) What is the answer to the same question about most probable speeds? c) What is the answer to the same question about rms speeds?

ADDITIONAL PROBLEMS

70. In the deep space between galaxies, the density of molecules (which are mostly single atoms) can be as low as 10^6 atoms/m^3, and the temperature is a frigid 2.7 K. What is the pressure? (b) What volume (in m^3) is occupied by 1 mol of gas? (c) If this volume is a cube, what is the length of its sides in kilometers?

71. The mean free path for helium at a certain temperature and pressure is 2.10×10^{-7} m. The radius of a helium atom can be taken as 1.10×10^{-11} m. What is the measure of the density of helium under those conditions (a) in molecules per cubic meter and (b) in moles per cubic meter?

72. The mean free path for methane at a temperature of 269 K and a pressure of 1.11×10^5 Pa is 4.81×10^{-8} m. Find the effective radius *r* of the methane molecule.

73. Find the total number of collisions between molecules in 1.00 s in 1.00 L of nitrogen gas at standard temperature and pressure ($0\,°\text{C}$, 1.00 atm). Use 1.88×10^{-10} m as the effective radius of a nitrogen molecule. (The number of collisions per second is the reciprocal of the collision time.) Keep in mind that each collision involves two molecules, so if one molecule collides once in a certain period of time, the collision of the molecule it hit cannot be counted.

74. A sealed, perfectly insulated container contains 0.630 mol of air at $20.0\,°\text{C}$ and an iron stirring bar of mass 40.0 g. The stirring bar is magnetically driven to a kinetic energy of 50.0 J and allowed to slow down by air resistance. What is the equilibrium temperature?

75. **Unreasonable results**. (a) Find the temperature of 0.360 kg of water, modeled as an ideal gas, at a pressure of 1.01×10^5 Pa if it has a volume of 0.615 m^3. (b) What is unreasonable about this answer? How could you get a better answer?

CHALLENGE PROBLEMS

76. Eight bumper cars, each with a mass of 322 kg, are running in a room 21.0 m long and 13.0 m wide. They have no drivers, so they just bounce around on their own. The rms speed of the cars is 2.50 m/s. Repeating the arguments of **Pressure, Temperature, and RMS Speed**, find the average force per unit length (analogous to pressure) that the cars exert on the walls.

77. Verify that $v_p = \sqrt{\frac{2k_\text{B}T}{m}}$.

18 | NUCLEAR ENERGY AND THE SOLAR SYSTEM

Chapter Outline

Introduction

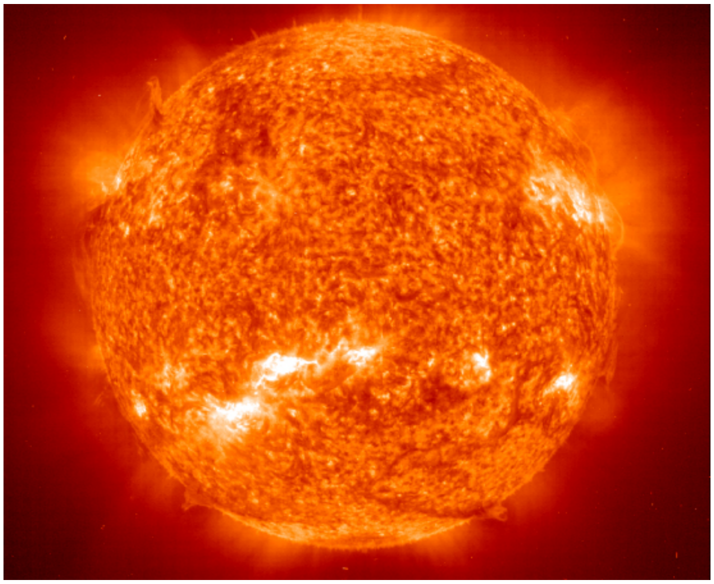


Figure 18.1 The Sun is powered by nuclear fusion in its core. The core converts approximately 10^{38} protons/second into helium at a temperature of 14 million K. This process releases energy in the form of photons, neutrinos, and other particles. (credit: modification of work by EIT SOHO Consortium, ESA, NASA)

Thermal energy plays many important roles in our solar system. In this chapter we will examine its sources and their consequences.

The Sun is the main source of energy in the solar system. The Sun is 109 Earth diameters across, and accounts for more than 99% of the total mass of the solar system. The Sun shines by fusing hydrogen nuclei—protons—deep inside its interior. Because of this, we must study some properties of the atomic nucleus. The nucleus lies at the center of an atom, and consists of protons and neutrons. A deep understanding of the nucleus also leads to numerous valuable technologies, including devices to date ancient rocks, map the galactic arms of the Milky Way, and generate electrical power.

18.1 | Formation of the Solar System

Learning Objectives

By the end of this section, you will be able to:

- Describe the motion, chemical, and age constraints that must be met by any theory of solar system formation
- Summarize the physical and chemical changes during the solar nebula stage of solar system formation
- Explain the formation process of the terrestrial and giant planets
- Describe the main events of the further evolution of the solar system

The comets, asteroids, and meteorites are the last surviving remnants from the processes that formed the solar system. The planets, moons, and the Sun, of course, also are the products of the formation process, although the material in them has undergone a wide range of changes. We are now ready to put together the information from all these objects, along with the physics of energy and thermodynamics that we have studied, to discuss what is known about the origin of the solar system.

Observational Constraints

There are certain basic properties of the planetary system that any theory of its formation must explain. These may be summarized under three categories: motion constraints, chemical constraints, and age constraints. We call them *constraints* because they place restrictions on our theories; unless a theory can explain the observed facts, it will not survive in the competitive marketplace of ideas that characterizes the endeavor of science. Let's take a look at these constraints one by one.

There are many regularities to the motions in the solar system. We saw that the planets all revolve around the Sun in the same direction and approximately in the plane of the Sun's own rotation. In addition, most of the planets rotate in the same direction as they revolve, and most of the moons also move in counterclockwise orbits (when seen from the north). With the exception of the comets and other trans-neptunian objects, the motions of the system members define a disk or Frisbee shape. Nevertheless, a full theory must also be prepared to deal with the exceptions to these trends, such as the *retrograde rotation* (not revolution) of Venus.

In the realm of chemistry, we saw that Jupiter and Saturn have approximately the same composition—dominated by hydrogen and helium. These are the two largest planets, with sufficient gravity to hold on to any gas present when and where they formed; thus, we might expect them to be representative of the original material out of which the solar system formed. Each of the other members of the planetary system is, to some degree, lacking in the light elements. A careful examination of the composition of solid solar-system objects shows a striking progression from the metal-rich inner planets, through those made predominantly of rocky materials, out to objects with ice-dominated compositions in the outer solar system. The comets in the Oort cloud and the trans-neptunian objects in the Kuiper belt are also icy objects, whereas the asteroids represent a transitional rocky composition with abundant dark, carbon-rich material.

As we saw in the section on the **Origin of the Solar System**, this general chemical pattern can be interpreted as a temperature sequence: hot near the Sun and cooler as we move outward. The inner parts of the system are generally missing those materials that could not condense (form a solid) at the high temperatures found near the Sun. However, there are (again) important exceptions to the general pattern. For example, it is difficult to explain the presence of water on Earth and Mars if these planets formed in a region where the temperature was too hot for ice to condense, unless the ice or water was brought in later from cooler regions. The extreme example is the observation that there are polar deposits of ice on both Mercury and the Moon; these are almost certainly formed and maintained by occasional comet impacts.

As far as age is concerned, we will see that radioactive dating demonstrates that some rocks on the surface of Earth have been present for at least 3.8 billion years, and that certain lunar samples are 4.4 billion years old. The primitive meteorites all have radioactive ages near 4.5 billion years. The age of these unaltered building blocks is considered the age of the planetary system. The similarity of the measured ages tells us that planets formed and their crusts cooled within a few tens of millions of years (at most) of the beginning of the solar system. Further, detailed examination of primitive meteorites indicates that they are made primarily from material that condensed or coagulated out of a hot gas; few identifiable fragments appear to have survived from before this hot-vapor stage 4.5 billion years ago.

The Solar Nebula

All the foregoing constraints are consistent with the general idea, introduced in **Origin of the Solar System**, that the solar system formed 4.5 billion years ago out of a rotating cloud of vapor and dust—which we call the solar nebula—with an initial composition similar to that of the Sun today. As the solar nebula collapsed under its own gravity, material fell toward the center, where things became more and more concentrated and hot. Increasing temperatures in the shrinking nebula vaporized most of the solid material that was originally present.

At the same time, the collapsing nebula began to rotate faster through the conservation of angular momentum (see the **Angular Momentum**. Like a figure skater pulling her arms in to spin faster, the shrinking cloud spun more quickly as time went on. Now, think about how a round object spins. Close to the poles, the spin rate is slow, and it gets faster as you get closer to the equator. In the same way, near the poles of the nebula, where orbits were slow, the nebular material fell directly into the center. Faster moving material, on the other hand, collapsed into a flat disk revolving around the central object (**Figure 18.2**). The existence of this disk-shaped rotating nebula explains the primary motions in the solar system that we discussed in the previous section. And since they formed from a rotating disk, the planets all orbit the same way.

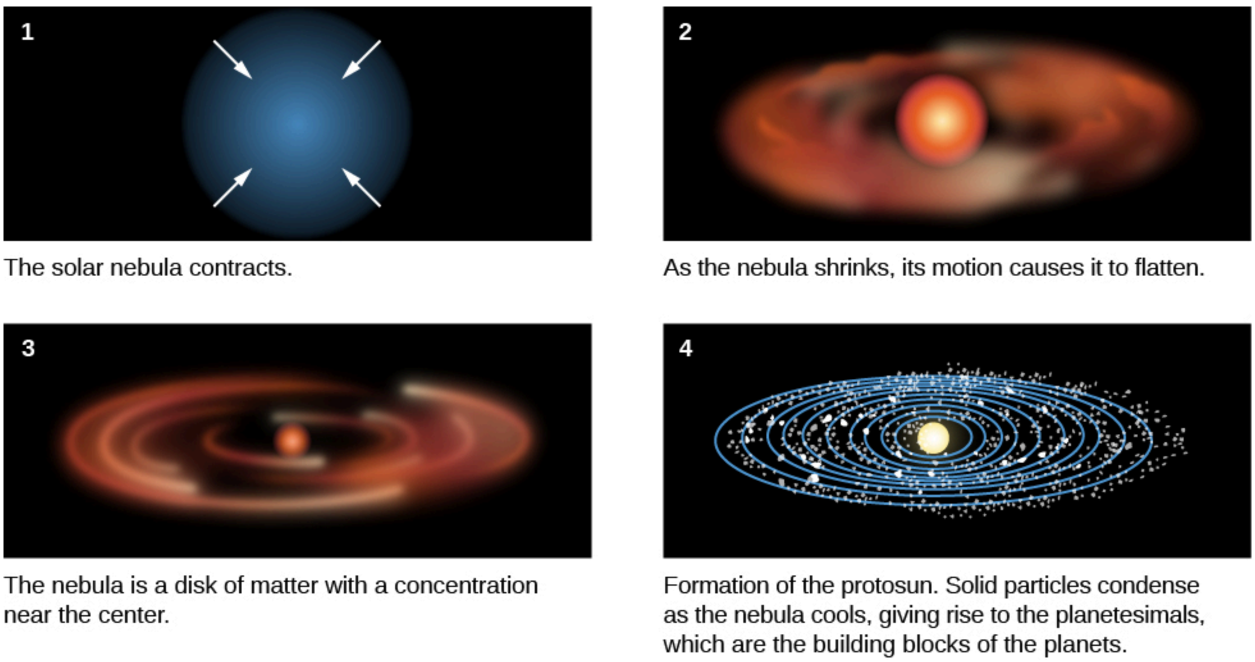


Figure 18.2 This illustration shows the steps in the formation of the solar system from the solar nebula. As the nebula shrinks, its rotation causes it to flatten into a disk. Much of the material is concentrated in the hot center, which will ultimately become a star. Away from the center, solid particles can condense as the nebula cools, giving rise to planetesimals, the building blocks of the planets and moons.

Picture the solar nebula at the end of the collapse phase, when it was at its hottest. With no more gravitational energy (from material falling in) to heat it, most of the nebula began to cool. The material in the center, however, where it was hottest and most crowded, formed a *star* that maintained high temperatures in its immediate neighborhood by producing its own energy. Turbulent motions and magnetic fields within the disk can drain away angular momentum, robbing the disk material of some of its spin. This allowed some material to continue to fall into the growing star, while the rest of the disk gradually stabilized.

The temperature within the disk decreased with increasing distance from the Sun, much as the planets' temperatures vary with position today. As the disk cooled, the gases interacted chemically to produce compounds; eventually these compounds condensed into liquid droplets or solid grains. This is similar to the process by which raindrops on Earth condense from moist air as it rises over a mountain.

Let's look in more detail at how material condensed at different places in the maturing disk (**Figure 18.3**). The first materials to form solid grains were the metals and various rock-forming silicates. As the temperature dropped, these were joined throughout much of the solar nebula by sulfur compounds and by carbon- and water-rich silicates, such as those now found abundantly among the asteroids. However, in the inner parts of the disk, the temperature never dropped low enough for such materials as ice or carbonaceous organic compounds to condense, so they were lacking on the innermost planets.

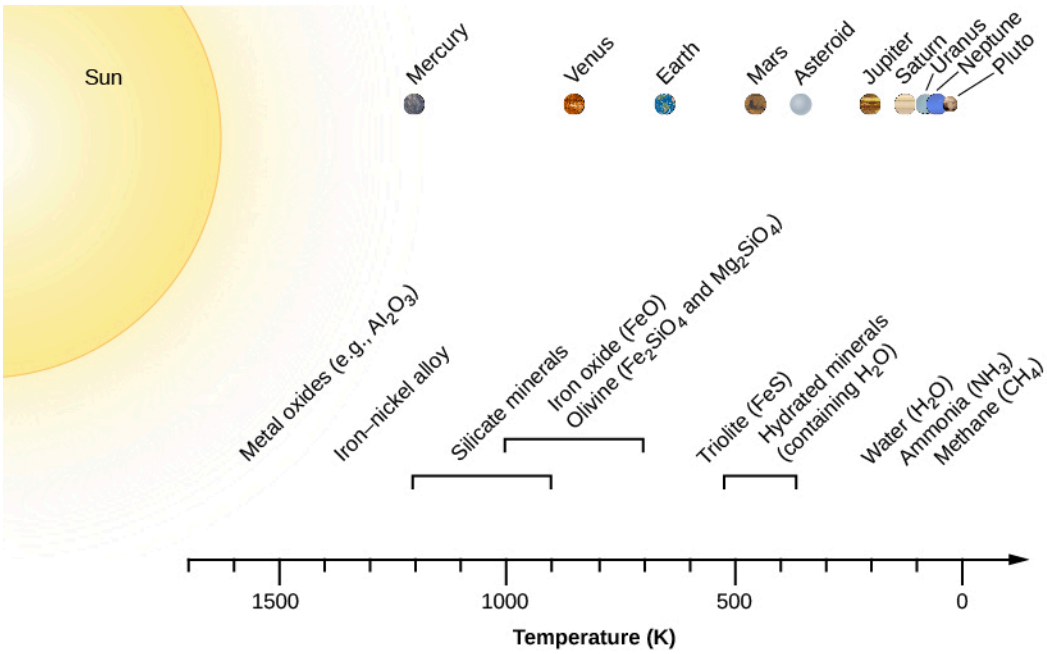


Figure 18.3 The scale along the bottom shows temperature; above are the materials that would condense out at each temperature under the conditions expected to prevail in the nebula.

Far from the Sun, cooler temperatures allowed the oxygen to combine with hydrogen and condense in the form of water (H_2O) ice. Beyond the orbit of Saturn, carbon and nitrogen combined with hydrogen to make ices such as methane (CH_4) and ammonia (NH_3). This sequence of events explains the basic chemical composition differences among various regions of the solar system.

Example 18.1

Rotation of the Solar Nebula

We can use the concept of angular momentum to trace the evolution of the collapsing solar nebula. From **Equation 12.3** we know how to find the angular momentum of a rotating body. Now, no matter what its specific shape is, a body's moment of inertia is always proportional to the square of its radius (see **Equation 9.2**) and therefore to the square of its diameter. Furthermore, its angular velocity is inversely proportional to its period of rotation. So, this means that its angular momentum is proportional to the square of its diameter divided by its period of rotation (D^2/T). If angular momentum is conserved, then any change in the size of a nebula must be compensated for by a proportional change in period, in order to keep D^2/T constant. Suppose the solar nebula began with a diameter of 10,000 AU and a rotation period of 1 million years. What is its rotation period when it has shrunk to the size of Pluto's orbit, which **Appendix D** tells us has a radius of about 40 AU?

Solution

We are given that the final diameter of the solar nebula is about 80 AU. Noting the initial state before the collapse and the final state at Pluto's orbit, then

$$\frac{T_{\text{final}}}{T_{\text{initial}}} = \left(\frac{D_{\text{final}}}{D_{\text{initial}}}\right)^2 = \left(\frac{80}{10{,}000}\right)^2 = (0.008)^2 = 0.000064$$

With T_{initial} equal to 1,000,000 years, T_{final}, the new rotation period, is 64 years. This is a lot shorter than the actual time Pluto takes to go around the Sun, but it gives you a sense of the kind of speeding up the conservation of angular momentum can produce. As we noted earlier, other mechanisms helped the material in the disk lose angular momentum before the planets fully formed.

Check Your Learning

What would the rotation period of the nebula in our example be when it had shrunk to the size of Jupiter's orbit?

Answer:

The period of the rotating nebula is inversely proportional to D^2. As we have just seen, $\frac{T_{\text{final}}}{T_{\text{initial}}} = \left(\frac{D_{\text{final}}}{D_{\text{initial}}}\right)^2$. Initially, we have $T_{\text{initial}} = 10^6$ yr and $D_{\text{initial}} = 10^4$ AU. Then, if D_{final} is in AU, T_{final} (in years) is given by $T_{\text{final}} = 0.01 D_{\text{final}}^2$. If Jupiter's orbit has a radius of 5.2 AU, then the diameter is 10.4 AU. The period is then 1.08 years.

Formation of the Terrestrial Planets

The grains that condensed in the solar nebula rather quickly joined into larger and larger chunks, until most of the solid material was in the form of *planetesimals*, chunks a few kilometers to a few tens of kilometers in diameter. Some planetesimals still survive today as comets and asteroids. Others have left their imprint on the cratered surfaces of many of the worlds we studied in earlier chapters. A substantial step up in size is required, however, to go from planetesimal to planet.

Some planetesimals were large enough to attract their neighbors gravitationally and thus to grow by the process called **accretion**. While the intermediate steps are not well understood, ultimately several dozen centers of accretion seem to have grown in the inner solar system. Each of these attracted surrounding planetesimals until it had acquired a mass similar to that of Mercury or Mars. At this stage, we may think of these objects as *protoplanets*—"not quite ready for prime time" planets.

Each of these protoplanets continued to grow by the accretion of planetesimals. Every incoming planetesimal was accelerated by the gravity of the protoplanet, striking with enough energy to melt both the projectile and a part of the impact area. Soon the entire protoplanet was heated to above the melting temperature of rocks. The result was *planetary differentiation*, with heavier metals sinking toward the core and lighter silicates rising toward the surface. As they were heated, the inner protoplanets lost some of their more volatile constituents (the lighter gases), leaving more of the heavier elements and compounds behind.

Formation of the Giant Planets

In the outer solar system, where the available raw materials included ices as well as rocks, the protoplanets grew to be much larger, with masses ten times greater than Earth. These protoplanets of the outer solar system were so large that they were able to attract and hold the surrounding gas. As the hydrogen and helium rapidly collapsed onto their cores, the giant planets were heated by the energy of contraction. But although these giant planets got hotter than their terrestrial siblings, they were far too small to raise their central temperatures and pressures to the point where nuclear reactions could begin (and it is such reactions that give us our definition of a star). After glowing dull red for a few thousand years, the giant planets gradually cooled to their present state (**Figure 18.4**).

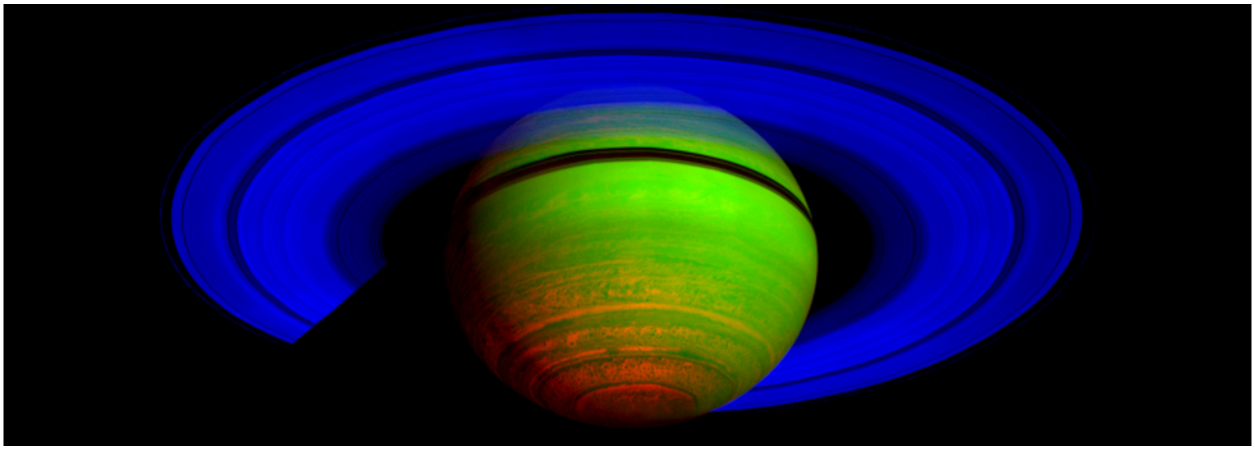


Figure 18.4 This image from the Cassini spacecraft is stitched together from 65 individual observations. Sunlight reflected at a wavelength of 2 micrometers is shown as blue, sunlight reflected at 3 micrometers is shown as green, and heat radiated from Saturn's interior at 5 micrometers is red. For example, Saturn's rings reflect sunlight at 2 micrometers, but not at 3 and 5 micrometers, so they appear blue. Saturn's south polar regions are seen glowing with internal heat. (credit: modification of work by NASA/JPL/University of Arizona)

The collapse of gas from the nebula onto the cores of the giant planets explains how these objects acquired nearly the same hydrogen-rich composition as the Sun. The process was most efficient for Jupiter and Saturn; hence, their compositions are most nearly "cosmic." Much less gas was captured by Uranus and Neptune, which is why these two planets have compositions dominated by the icy and rocky building blocks that made up their large cores rather than by hydrogen and helium. The initial formation period ended when much of the available raw material was used up and the solar wind (the flow of atomic particles) from the young Sun blew away the remaining supply of lighter gases.

Further Evolution of the System

All the processes we have just described, from the collapse of the solar nebula to the formation of protoplanets, took place within a few million years. However, the story of the formation of the solar system was not complete at this stage; there were many planetesimals and other debris that did not initially accumulate to form the planets. What was their fate?

The comets visible to us today are merely the tip of the cosmic iceberg (if you'll pardon the pun). Most comets are believed to be in the Oort cloud, far from the region of the planets. Additional comets and icy dwarf planets are in the Kuiper belt, which stretches beyond the orbit of Neptune. These icy pieces probably formed near the present orbits of Uranus and Neptune but were ejected from their initial orbits by the gravitational influence of the giant planets.

In the inner parts of the system, remnant planetesimals and perhaps several dozen protoplanets continued to whiz about. Over the vast span of time we are discussing, collisions among these objects were inevitable. Giant impacts at this stage probably stripped Mercury of part of its mantle and crust, reversed the rotation of Venus, and broke off part of Earth to create the Moon (all events we discussed in other chapters).

Smaller-scale impacts also added mass to the inner protoplanets. Because the gravity of the giant planets could "stir up" the orbits of the planetesimals, the material impacting on the inner protoplanets could have come from almost anywhere within the solar system. In contrast to the previous stage of accretion, therefore, this new material did not represent just a narrow range of compositions.

As a result, much of the debris striking the inner planets was ice-rich material that had condensed in the outer part of the solar nebula. As this comet-like bombardment progressed, Earth accumulated the water and various organic compounds that would later be critical to the formation of life. Mars and Venus probably also acquired abundant water and organic materials from the same source, as Mercury and the Moon are still doing to form their icy polar caps.

Gradually, as the planets swept up or ejected the remaining debris, most of the planetesimals disappeared. In two regions, however, stable orbits are possible where leftover planetesimals could avoid impacting the planets or being ejected from the system. These regions are the asteroid belt between Mars and Jupiter and the Kuiper belt beyond Neptune. The planetesimals (and their fragments) that survive in these special locations are what we now call asteroids, comets, and trans-neptunian objects.

Astronomers used to think that the solar system that emerged from this early evolution was similar to what we see today. Detailed recent studies of the orbits of the planets and asteroids, however, suggest that there were more violent events soon afterward, perhaps involving substantial changes in the orbits of Jupiter and Saturn. These two giant planets control, through their gravity, the distribution of asteroids. Working backward from our present solar system, it appears that orbital changes took place during the first few hundred million years. One consequence may have been scattering of asteroids into the inner solar system, causing the period of “heavy bombardment” recorded in the oldest lunar craters.

18.2 | Properties of Nuclei

Learning Objectives

By the end of this section, you will be able to:

- Describe the composition and size of an atomic nucleus
- Use a nuclear symbol to express the composition of an atomic nucleus
- Explain why the number of neutrons is greater than protons in heavy nuclei
- Calculate the atomic mass of an element given its isotopes

The **atomic nucleus** is composed of **protons** and **neutrons** (**Figure 18.5**). Protons and neutrons have approximately the same mass, but protons carry one unit of positive charge $(+e)$, and neutrons carry no charge. These particles are packed together into an extremely small space at the center of an atom. According to scattering experiments, the nucleus is spherical or ellipsoidal in shape, and about 1/100,000th the size of a hydrogen atom. If an atom were the size of a major league baseball stadium, the nucleus would be roughly the size of the baseball. Protons and neutrons within the nucleus are called **nucleons**.

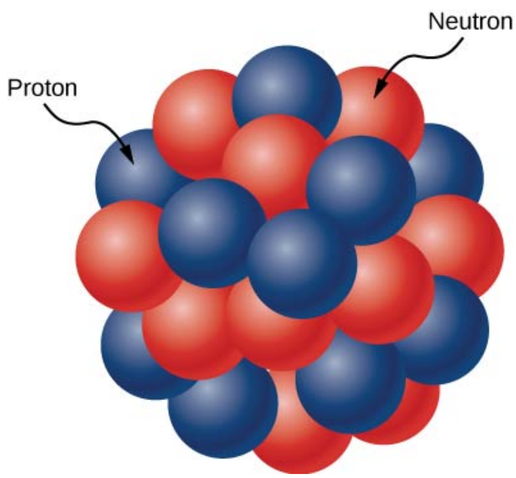


Figure 18.5 The atomic nucleus is composed of protons and neutrons. Protons are shown in blue, and neutrons are shown in red.

Counts of Nucleons

The number of protons in the nucleus is given by the **atomic number**, Z. The number of neutrons in the nucleus is the **neutron number**, N. The total number of nucleons is the **mass number**, A. These numbers are related by

$$A = Z + N. \tag{18.1}$$

A nucleus is represented symbolically by

$$^{A}_{Z}\text{X}, \tag{18.2}$$

where X represents the chemical element, A is the mass number, and Z is the atomic number. For example, $^{12}_{6}\text{C}$ represents the carbon nucleus with six protons and six neutrons (or 12 nucleons).

A graph of the number N of neutrons versus the number Z of protons for a range of stable nuclei (**nuclides**) is shown in **Figure 18.6**. For a given value of Z, multiple values of N (blue points) are possible. For small values of Z, the number of neutrons equals the number of protons $(N = P),$ and the data fall on the red line. For large values of Z, the number of neutrons is greater than the number of protons $(N > P),$ and the data points fall above the red line. The number of neutrons is generally greater than the number of protons for $Z > 15.$

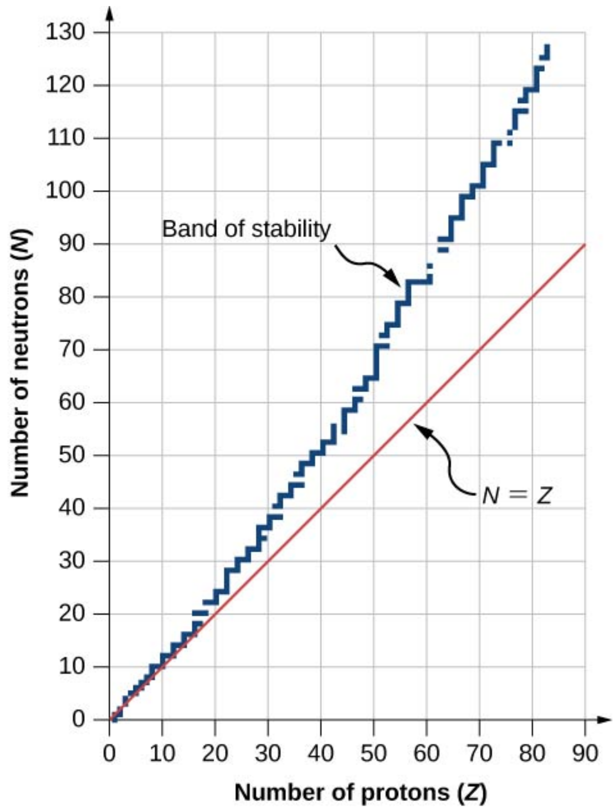


Figure 18.6 This graph plots the number of neutrons N against the number of protons Z for stable atomic nuclei. Larger nuclei, have more neutrons than protons.

A chart based on this graph that provides more detailed information about each nucleus is given in **Figure 18.7**. This chart is called a **chart of the nuclides**. Each cell or tile represents a separate nucleus. The nuclei are arranged in order of ascending Z (along the horizontal direction) and ascending N (along the vertical direction).

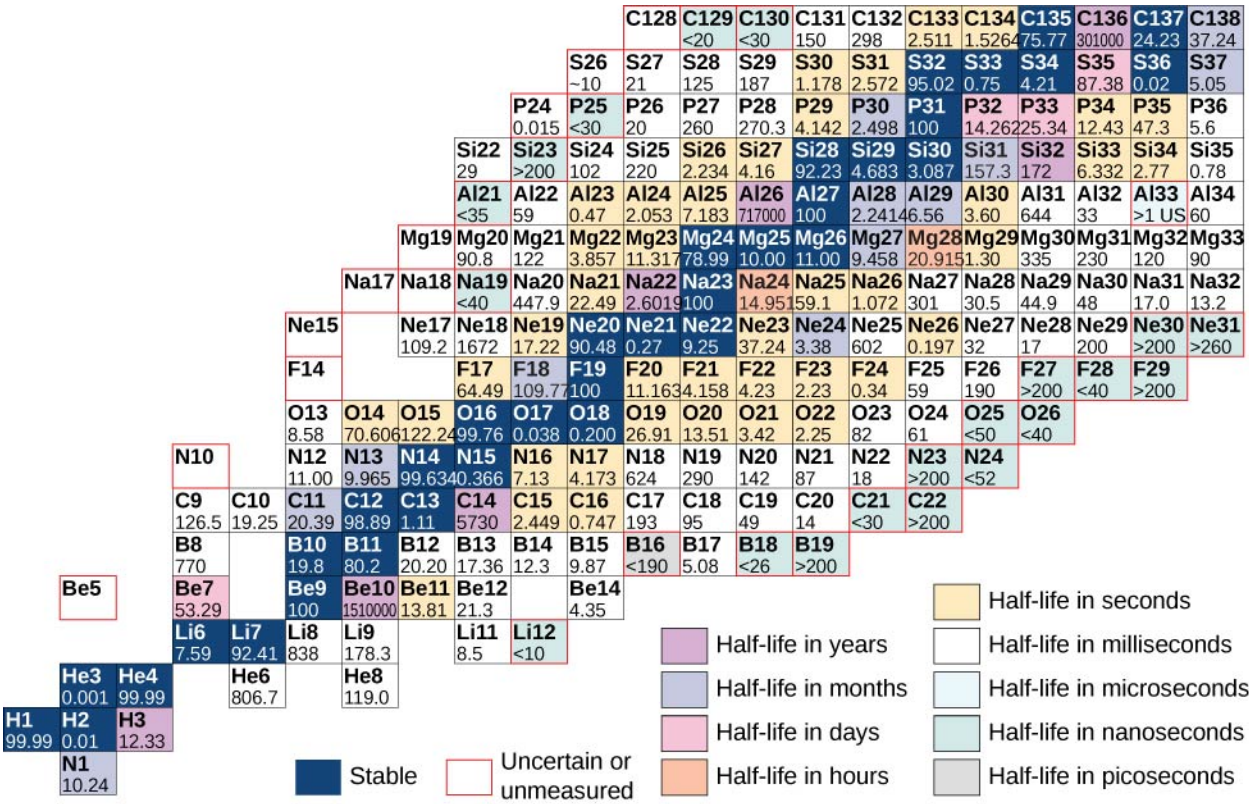


Figure 18.7 Partial chart of the nuclides. For stable nuclei (dark blue backgrounds), cell values represent the percentage of nuclei found on Earth with the same atomic number (percent abundance). For the unstable nuclei, the number represents the half-life.

Atoms that contain nuclei with the same number of protons (*Z*) and different numbers of neutrons (*N*) are called **isotopes**. For example, hydrogen has three isotopes: normal hydrogen (1 proton, no neutrons), deuterium (one proton and one neutron), and tritium (one proton and two neutrons). Isotopes of a given atom share the same chemical properties, since these properties are determined by interactions between the outer electrons of the atom, and not the nucleons. For example, water that contains deuterium rather than hydrogen ("heavy water") looks and tastes like normal water. The following table shows a list of common isotopes.

Element	Symbol	Mass Number	Mass (Atomic Mass Units)	Percent Abundance*	Half-life**
Hydrogen	H	1	1.0078	99.99	stable
	^{2}H or D	2	2.0141	0.01	stable
	^{3}H	3	3.0160	–	12.32 y
Carbon	^{12}C	12	12.0000	98.91	stable
	^{13}C	13	13.0034	1.1	stable
	^{14}C	14	14.0032	–	5730 y
Nitrogen	^{14}N	14	14.0031	99.6	stable

Table 18.1 Common Isotopes *No entry if less than 0.001 (trace amount).
**Stable if half-life > 10 seconds.

Element	Symbol	Mass Number	Mass (Atomic Mass Units)	Percent Abundance*	Half-life**
	^{15}N	15	15.0001	0.4	stable
	^{16}N	16	16.0061	–	7.13 s
Oxygen	^{16}O	16	15.9949	99.76	stable
	^{17}O	17	16.9991	0.04	stable
	^{18}O	18	17.9992	0.20	stable
	^{19}O	19	19.0035	–	26.46 s

Table 18.1 Common Isotopes *No entry if less than 0.001 (trace amount).
**Stable if half-life > 10 seconds.

Why do neutrons outnumber protons in heavier nuclei (**Figure 18.8**)? The answer to this question requires an understanding of forces inside the nucleus. Two types of forces exist: (1) the long-range electrostatic (Coulomb) force that makes the positively charged protons repel one another; and (2) the short-range **strong nuclear force** that makes all nucleons in the nucleus attract one another. You may also have heard of a "weak" nuclear force. This force is responsible for some nuclear decays, but as the name implies, it does not play a role in stabilizing the nucleus against the strong Coulomb repulsion it experiences. We discuss strong nuclear force in more detail in the next chapter when we cover particle physics. Nuclear stability occurs when the attractive forces between nucleons compensate for the repulsive, long-range electrostatic forces between all protons in the nucleus. For heavy nuclei $(Z > 15)$, excess neutrons are necessary to keep the electrostatic interactions from breaking the nucleus apart, as shown in **Figure 18.6**.

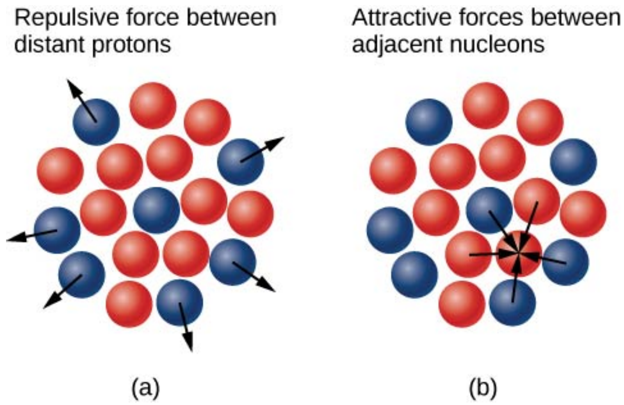


Figure 18.8 (a) The electrostatic force is repulsive and has long range. The arrows represent outward forces on protons (in blue) at the nuclear surface by a proton (also in blue) at the center. (b) The strong nuclear force acts between neighboring nucleons. The arrows represent attractive forces exerted by a neutron (in red) on its nearest neighbors.

Because of the existence of stable isotopes, we must take special care when quoting the mass of an element. For example, Copper (Cu) has two stable isotopes:

$$^{63}_{29}\text{Cu}\left(62.929595\text{ g/mol}\right)\text{ with an abundance of }69.09\%$$

$$^{65}_{29}\text{Cu}\left(64.927786\text{ g/mol}\right)\text{ with an abundance of }30.91\%$$

Given these two "versions" of Cu, what is the mass of this element? The **atomic mass** of an element is defined as the weighted average of the masses of its isotopes. Thus, the atomic mass of Cu is

$m_{Cu} = (62.929595)(0.6909) + (64.927786)(0.3091) = 63.55\ \text{g/mol}$. The mass of an individual nucleus is often expressed in **atomic mass units** (u), where $\text{u} = 1.66054 \times 10^{-27}\ \text{kg}$. (An atomic mass unit is defined as 1/12th the mass of a ^{12}C nucleus.) In atomic mass units, the mass of a helium nucleus (*A* = 4) is approximately 4 u. A helium nucleus is also called an alpha (α) particle.

Nuclear Size

The simplest model of the nucleus is a densely packed sphere of nucleons. The volume *V* of the nucleus is therefore proportional to the number of nucleons *A*, expressed by

$$V = \frac{4}{3}\pi r^3 = kA,$$

where *r* is the **radius of a nucleus** and *k* is a constant with units of volume. Solving for *r*, we have

$$r = r_0 A^{1/3} \tag{18.3}$$

where r_0 is a constant. For hydrogen ($A = 1$), r_0 corresponds to the radius of a single proton. Scattering experiments support this general relationship for a wide range of nuclei, and they imply that neutrons have approximately the same radius as protons. The experimentally measured value for r_0 is approximately 1.2 femtometer (recall that $1\ \text{fm} = 10^{-15}\ \text{m}$).

Example 18.2

The Iron Nucleus

Find the radius (*r*) and approximate density (ρ) of a Fe-56 nucleus. Assume the mass of the Fe-56 nucleus is approximately 56 u.

Strategy

(a) Finding the radius of ^{56}Fe is a straightforward application of $r = r_0 A^{1/3}$, given $A = 56$. (b) To find the approximate density of this nucleus, assume the nucleus is spherical. Calculate its volume using the radius found in part (a), and then find its density from $\rho = m/V$.

Solution

a. The radius of a nucleus is given by

$$r = r_0 A^{1/3}.$$

Substituting the values for r_0 and *A* yields

$$\begin{aligned} r &= (1.2\ \text{fm})(56)^{1/3} = (1.2\ \text{fm})(3.83) \\ &= 4.6\ \text{fm}. \end{aligned}$$

b. Density is defined to be $\rho = m/V$, which for a sphere of radius *r* is

$$\rho = \frac{m}{V} = \frac{m}{(4/3)\pi r^3}.$$

Substituting known values gives

$$\rho = \frac{56\ \text{u}}{(1.33)(3.14)(4.6\ \text{fm})^3} = 0.138\ \text{u/fm}^3.$$

Converting to units of kg/m^3, we find

$$\rho = (0.138\ \text{u/fm}^3)(1.66 \times 10^{-27}\ \text{kg/u})\left(\frac{1\ \text{fm}}{10^{-15}\ \text{m}}\right) = 2.3 \times 10^{17}\ \text{kg/m}^3.$$

Significance

a. The radius of the Fe-56 nucleus is found to be approximately 5 fm, so its diameter is about 10 fm, or 10^{-14} m. In previous discussions of Rutherford's scattering experiments, a light nucleus was estimated to be 10^{-15} m in diameter. Therefore, the result shown for a mid-sized nucleus is reasonable.

b. The density found here may seem incredible. However, it is consistent with earlier comments about the nucleus containing nearly all of the mass of the atom in a tiny region of space. One cubic meter of nuclear matter has the same mass as a cube of water 61 km on each side.

18.1 Check Your Understanding Nucleus X is two times larger than nucleus Y. What is the ratio of their atomic masses?

18.3 | Nuclear Binding Energy

Learning Objectives

By the end of this section, you will be able to:

- Calculate the mass defect and binding energy for a wide range of nuclei
- Use a graph of binding energy per nucleon (BEN) versus mass number(A) to assess the relative stability of a nucleus
- Compare the binding energy of a nucleon in a nucleus to the ionization energy of an electron in an atom

The forces that bind nucleons together in an atomic nucleus are much greater than those that bind an electron to an atom through electrostatic attraction. This is evident by the relative sizes of the atomic nucleus and the atom $(10^{-15}$ and 10^{-10} m, respectively). The energy required to pry a nucleon from the nucleus is therefore much larger than that required to remove (or ionize) an electron in an atom. In general, all nuclear changes involve large amounts of energy per particle undergoing the reaction. This has numerous practical applications.

Mass Defect

According to nuclear particle experiments, the total mass of a nucleus (m_{nuc}) is *less* than the sum of the masses of its constituent nucleons (protons and neutrons). The mass difference, or **mass defect**, is given by

$$\Delta m = Zm_p + (A - Z)m_n - m_{\text{nuc}} \quad (18.4)$$

where Zm_p is the total mass of the protons, $(A - Z)m_n$ is the total mass of the neutrons, and m_{nuc} is the mass of the nucleus. According to Einstein's special theory of relativity, mass is a measure of the total energy of a system ($E = mc^2$). Thus, the total energy of a nucleus is less than the sum of the energies of its constituent nucleons. The formation of a nucleus from a system of isolated protons and neutrons is therefore an exothermic reaction—meaning that it releases energy. The energy emitted, or radiated, in this process is $(\Delta m)c^2$.

Now imagine this process occurs in reverse. Instead of forming a nucleus, energy is put into the system to break apart the nucleus (Figure 18.9). The amount of energy required is called the total **binding energy (BE)**, E_{b}.

Binding Energy

The binding energy is equal to the amount of energy released in forming the nucleus, and is therefore given by

$$E_b = (\Delta m)c^2. \quad (18.5)$$

Experimental results indicate that the binding energy for a nucleus with mass number $A > 8$ is roughly proportional to the total number of nucleons in the nucleus, A. The binding energy of a magnesium nucleus (^{24}Mg), for example, is approximately two times greater than for the carbon nucleus (^{12}C).

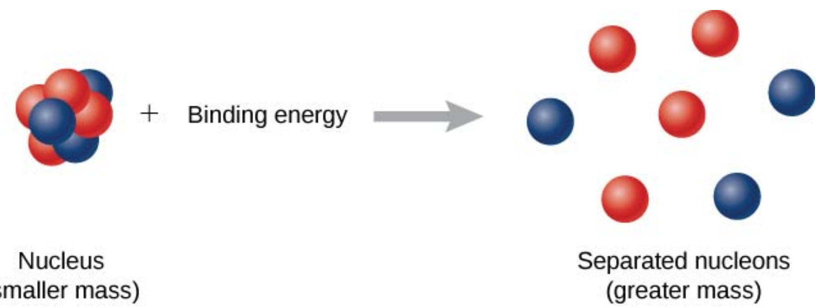


Figure 18.9 The binding energy is the energy required to break a nucleus into its constituent protons and neutrons. A system of separated nucleons has a greater mass than a system of bound nucleons.

Example 18.3

Mass Defect and Binding Energy of the Deuteron

Calculate the mass defect and the binding energy of the deuteron. The mass of the deuteron is $m_D = 3.34359 \times 10^{-27}\,\text{kg}$ or $1875.61\,\text{MeV}/c^2$.

Solution

From **Equation 18.4**, the mass defect for the deuteron is

$$\begin{aligned}\Delta m &= m_p + m_n - m_D \\ &= 938.28\,\text{MeV}/c^2 + 939.57\,\text{MeV}/c^2 - 1875.61\,\text{MeV}/c^2 \\ &= 2.24\,\text{MeV}/c^2.\end{aligned}$$

The binding energy of the deuteron is then

$$E_b = (\Delta m)c^2 = \left(2.24\,\text{MeV}/c^2\right)\left(c^2\right) = 2.24\,\text{MeV}.$$

Over two million electron volts are needed to break apart a deuteron into a proton and a neutron. This very large value indicates the great strength of the nuclear force. By comparison, the greatest amount of energy required to liberate an electron bound to a hydrogen atom by an attractive Coulomb force (an electromagnetic force) is about 10 eV.

Graph of Binding Energy per Nucleon

In nuclear physics, one of the most important experimental quantities is the **binding energy per nucleon (BEN)**, which is defined by

$$BEN = \frac{E_b}{A} \quad (18.6)$$

This quantity is the average energy required to remove an individual nucleon from a nucleus—analogous to the ionization energy of an electron in an atom. If the BEN is relatively large, the nucleus is relatively stable. BEN values are estimated from nuclear scattering experiments.

A graph of binding energy per nucleon versus atomic number *A* is given in **Figure 18.10**. This graph is considered by many physicists to be one of the most important graphs in physics. Two notes are in order. First, typical BEN values range from 6–10 MeV, with an average value of about 8 MeV. In other words, it takes several million electron volts to pry a nucleon from a typical nucleus, as compared to just 13.6 eV to ionize an electron in the ground state of hydrogen. This is why nuclear force is referred to as the "strong" nuclear force.

Second, the graph rises at low *A*, peaks very near iron $(\text{Fe},\ A = 56)$, and then tapers off at high *A*. The peak value suggests that the iron nucleus is the most stable nucleus in nature (it is also why nuclear fusion in the cores of stars ends with Fe). The reason the graph rises and tapers off has to do with competing forces in the nucleus. At low values of *A*, attractive nuclear forces between nucleons dominate over repulsive electrostatic forces between protons. But at high values of *A*, repulsive electrostatic forces between forces begin to dominate, and these forces tend to break apart the nucleus rather than hold it together.

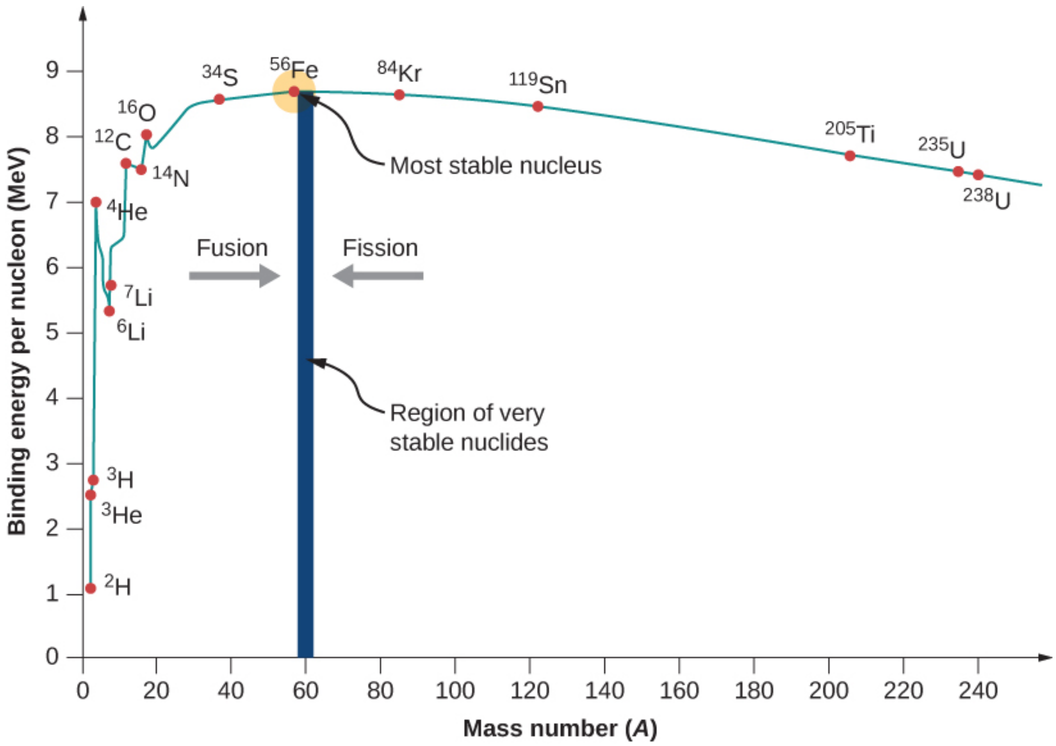


Figure 18.10 In this graph of binding energy per nucleon for stable nuclei, the BEN is greatest for nuclei with a mass near ^{56}Fe. Therefore, fusion of nuclei with mass numbers much less than that of Fe, and fission of nuclei with mass numbers greater than that of Fe, are exothermic processes.

As we will see, the BEN-versus-*A* graph implies that nuclei divided or combined release an enormous amount of energy. This is the basis for a wide range of phenomena, from the production of electricity at a nuclear power plant to sunlight.

Example 18.4

Tightly Bound Alpha Nuclides

Calculate the binding energy per nucleon of an $^4\text{He}\,(\alpha\text{ particle})$.

Strategy

Determine the total binding energy (BE) using the equation $\text{BE} = (\Delta m)c^2$, where Δm is the mass defect. The binding energy per nucleon (BEN) is BE divided by *A*.

Solution

For ^4He, we have $Z = N = 2$. The total binding energy is

$$\text{BE} = \{[2m_p + 2m_n] - m(^4\text{He})\}c^2.$$

These masses are $m(^4\text{He}) = 4.002602\ \text{u}$, $m_p = 1.007825\ \text{u}$, and $m_n = 1.008665\ \text{u}$. Thus we have,

$$\text{BE} = (0.030378\ \text{u})c^2.$$

Noting that $1\ \text{u} = 931.5\ \text{MeV}/c^2$, we find

$$\begin{aligned}\text{BE} &= (0.030378)(931.5\ \text{MeV}/c^2)c^2 \\ &= 28.3\ \text{MeV}.\end{aligned}$$

Since $A = 4$, the total binding energy per nucleon is

$$\text{BEN} = 7.07\ \text{MeV/nucleon}.$$

Significance

Notice that the binding energy per nucleon for ^4He is much greater than for the hydrogen isotopes $(\text{only} \approx 3\ \text{MeV/nucleon})$. Therefore, helium nuclei cannot break down hydrogen isotopes without energy being put into the system.

18.2 Check Your Understanding If the binding energy per nucleon is large, does this make it harder or easier to strip off a nucleon from a nucleus?

Applications in Astrophysics

As we continue to study the solar system, the production of energy in the Sun, and eventually the production of energy in other stars, we will see that the process of nuclear fusion, whereby two lighter nuclei combine to form one heavier nucleus, causes a release of energy. This is because there is an increase in the overall binding energy of all of the nucleons involved in the fusion process. (See **Source of Sunshine: Nuclear Fusion!**.)

For now, however, we turn to briefly examine the phenomenon of **radioactivity**, another aspect of nuclear physics that has direct applications to the study of the solar system.

18.4 | Radioactive Decay

Learning Objectives

By the end of this section, you will be able to:

- Describe the decay of a radioactive substance in terms of its decay constant and half-life
- Use the radioactive decay law to estimate the age of a substance
- Explain the natural processes that allow the dating of living tissue using ^{14}C

In 1896, Henri Becquerel discovered that a uranium-rich rock emits invisible rays that can darken a photographic plate in an enclosed container. Scientists offer three arguments for the nuclear origin of these rays. First, the effects of the radiation do not vary with chemical state; that is, whether the emitting material is in the form of an element or compound. Second, the radiation does not vary with changes in temperature or pressure—both factors that in sufficient degree can affect electrons in an atom. Third, the very large energy of the invisible rays (up to hundreds of eV) is not consistent with atomic electron transitions (only a few eV). Today, this radiation is explained by the conversion of mass into energy deep within the nucleus of an atom. The spontaneous emission of radiation from nuclei is called nuclear **radioactivity** (**Figure 18.11**).

Figure 18.11 The international ionizing radiation symbol is universally recognized as the warning symbol for nuclear radiation.

Radioactive Decay Law

When an individual nucleus transforms into another with the emission of radiation, the nucleus is said to **decay**. Radioactive decay occurs for all nuclei with $Z > 82,$ and also for some unstable isotopes with $Z < 83.$ The decay rate is proportional to the number of original (undecayed) nuclei N in a substance. The number of nuclei lost to decay per unit time, also called the **activity**, **A**, is written

$$A = \lambda N \tag{18.7}$$

where λ is called the **decay constant**. In words, the more nuclei available to decay, the more that do decay in a given unit of time. This equation can be rewritten as a **differential equation**:

$$\frac{dN}{N} = -\lambda dt. \tag{18.8}$$

By integrating this relationship, we can derive the **radioactive decay law**:

Radioactive Decay Law

The total number N of radioactive nuclei remaining after time t is

$$N = N_0 e^{-\lambda t} \tag{18.9}$$

where λ is the decay constant for the particular nucleus.

The total number of nuclei drops very rapidly at first, and then more slowly (**Figure 18.12**).

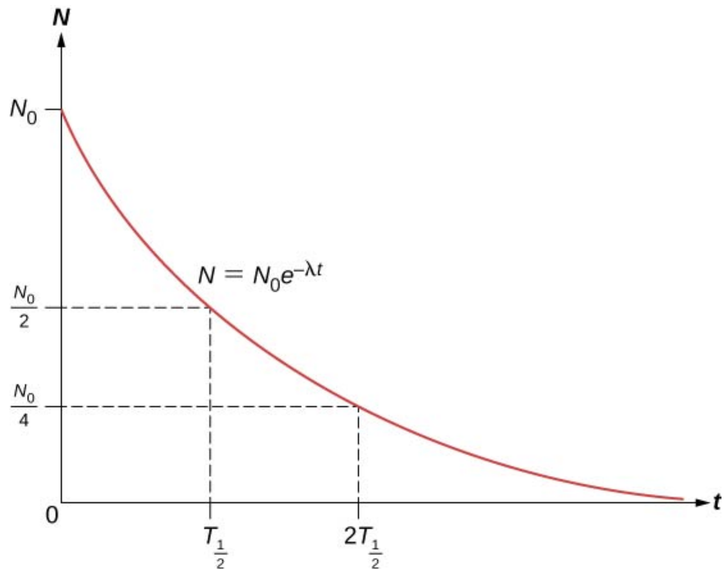


Figure 18.12 A plot of the radioactive decay law demonstrates that the number of nuclei remaining in a decay sample drops dramatically during the first moments of decay.

The **half-life** $(T_{1/2})$ of a radioactive substance is defined as the time for half of the original nuclei to decay (or the time at which half of the original nuclei remain). The half-lives of unstable isotopes are shown in the chart of nuclides in **Figure 18.7**. The number of radioactive nuclei remaining after an integer (n) number of half-lives is therefore

$$N = \frac{N_0}{2^n} \quad (18.10)$$

If the decay constant (λ) is large, the half-life is small, and vice versa. To determine the relationship between these quantities, note that when $t = T_{1/2}$, then $N = N_0/2$. This means that

$$\lambda = \frac{0.693}{T_{1/2}}. \quad (18.11)$$

Thus, if we know the half-life $T_{1/2}$ of a radioactive substance, we can find its decay constant. The **lifetime** $\bar{T}$ of a radioactive substance is defined as the average amount of time that a nucleus exists before decaying. The lifetime of a substance is just the reciprocal of the decay constant, written as

$$\bar{T} = \frac{1}{\lambda}. \quad (18.12)$$

Since the **activity** A is proportional to the decay rate, it follows simply that

$$A = A_0 e^{-\lambda t}. \quad (18.13)$$

where we have defined the initial activity as $A_0 = \lambda N_0$. Thus, the activity A of a radioactive substance also decreases exponentially with time (**Figure 18.13**).

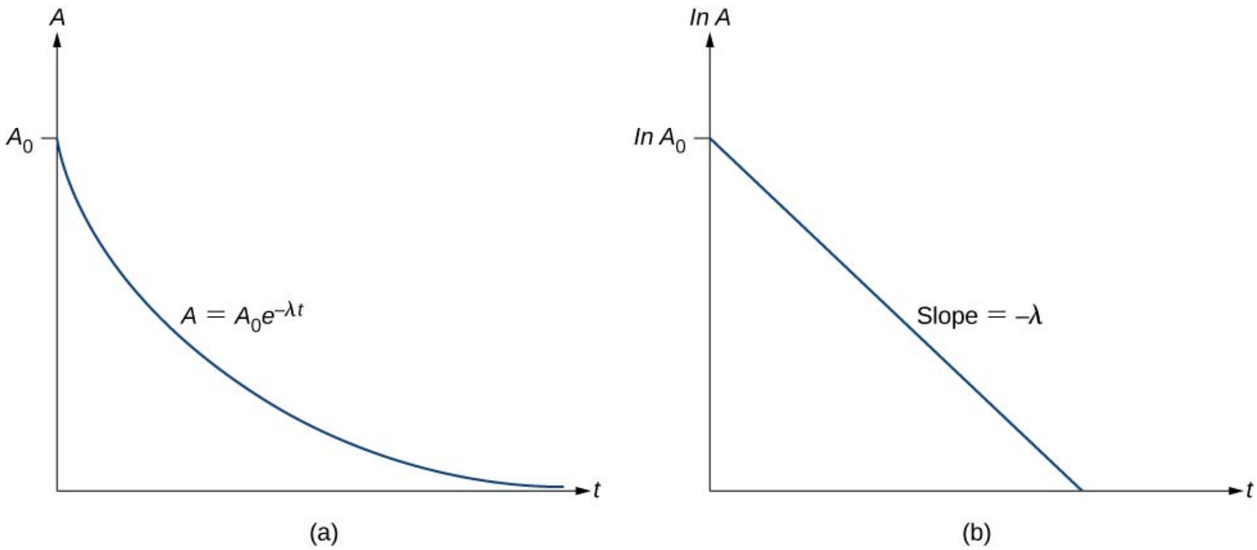


Figure 18.13 (a) A plot of the activity as a function of time (b) If we measure the activity at different times, we can plot ln *A* versus *t*, and obtain a straight line.

Example 18.5

Decay Constant and Activity of Strontium-90

The half-life of strontium-90, ${}^{90}_{38}\text{Sr}$, is 28.8 y. Find (a) its decay constant and (b) the initial activity of 1.00 g of the material.

Strategy

We can find the decay constant directly from **Equation 18.11**. To determine the activity, we first need to find the number of nuclei present.

Solution

a. The decay constant is found to be

$$\lambda = \frac{0.693}{T_{1/2}} = \left(\frac{0.693}{T_{1/2}}\right)\left(\frac{1\text{ yr}}{3.16 \times 10^7\text{ s}}\right) = 7.61 \times 10^{-10}\text{ s}^{-1}.$$

b. The atomic mass of ${}^{90}_{38}\text{Sr}$ is 89.91 g. Using Avogadro's number $N_A = 6.022 \times 10^{23}$ atoms/mol, we find the initial number of nuclei in 1.00 g of the material:

$$N_0 = \frac{1.00\text{ g}}{89.91\text{ g}}(N_A) = 6.70 \times 10^{21}\text{ nuclei.}$$

From this, we find that the activity A_0 at $t = 0$ for 1.00 g of strontium-90 is

$$\begin{aligned} A_0 &= \lambda N_0 \\ &= (7.61 \times 10^{-10}\text{ s}^{-1})(6.70 \times 10^{21}\text{ nuclei}) \\ &= 5.10 \times 10^{12}\text{ decays/s.} \end{aligned}$$

Expressing λ in terms of the half-life of the substance, we get

$$A = A_0 e^{-(0.693/T_{1/2})T_{1/2}} = A_0 e^{-0.693} = A_0/2. \quad \textbf{(18.14)}$$

Therefore, the activity is halved after one half-life. We can determine the decay constant λ by measuring the activity as a function of time. Taking the natural logarithm of the left and right sides of **???**, we get

$$\ln A = -\lambda t + \ln A_0. \tag{18.15}$$

This equation follows the linear form $y = mx + b$. If we plot ln *A* versus *t*, we expect a straight line with slope $-\lambda$ and *y*-intercept $\ln A_0$ (Figure 18.13(b)). Activity *A* is expressed in units of **becquerels** (Bq), where one $1\ \text{Bq} = 1\ \text{decay per second}$. This quantity can also be expressed in decays per minute or decays per year. One of the most common units for activity is the **curie (Ci)**, defined to be the activity of 1 g of ^{226}Ra. The relationship between the Bq and Ci is

$$1\ \text{Ci} = 3.70 \times 10^{10}\,\text{Bq}.$$

Example 18.6

What is ^{14}C Activity in Living Tissue?

Approximately 20% of the human body by mass is carbon. Calculate the activity due to ^{14}C in 1.00 kg of carbon found in a living organism. Express the activity in units of Bq and Ci.

Strategy

The activity of ^{14}C is determined using the equation $A_0 = \lambda N_0$, where λ is the decay constant and N_0 is the number of radioactive nuclei. The number of ^{14}C nuclei in a 1.00-kg sample is determined in two steps. First, we determine the number of ^{12}C nuclei using the concept of a mole. Second, we multiply this value by 1.3×10^{-12} (the known abundance of ^{14}C in a carbon sample from a living organism) to determine the number of ^{14}C nuclei in a living organism. The decay constant is determined from the known half-life of ^{14}C (available from Figure 18.7).

Solution

One mole of carbon has a mass of 12.0 g, since it is nearly pure ^{12}C. Thus, the number of carbon nuclei in a kilogram is

$$N\left(^{12}\text{C}\right) = \frac{6.02 \times 10^{23}\ \text{mol}^{-1}}{12.0\ \text{g/mol}} \times (1000\ \text{g}) = 5.02 \times 10^{25}.$$

The number of ^{14}C nuclei in 1 kg of carbon is therefore

$$N\left(^{14}\text{C}\right) = \left(5.02 \times 10^{25}\right)\left(1.3 \times 10^{-12}\right) = 6.52 \times 10^{13}.$$

Now we can find the activity *A* by using the equation $A = \frac{0.693\,N}{t_{1/2}}$. Entering known values gives us

$$A = \frac{0.693\left(6.52 \times 10^{13}\right)}{5730\ \text{y}} = 7.89 \times 10^{9}\ \text{y}^{-1}$$

or 7.89×10^9 decays per year. To convert this to the unit Bq, we simply convert years to seconds. Thus,

$$A = \left(7.89 \times 10^{9}\ \text{y}^{-1}\right)\frac{1.00\ \text{y}}{3.16 \times 10^{7}\ \text{s}} = 250\ \text{Bq},$$

or 250 decays per second. To express *A* in curies, we use the definition of a curie,

$$A = \frac{250\ \text{Bq}}{3.7 \times 10^{10}\ \text{Bq/Ci}} = 6.76 \times 10^{-9}\ \text{Ci}.$$

Thus,

$$A = 6.76\,\text{nCi}.$$

Significance

Approximately 20% of the human body by weight is carbon. Hundreds of ^{14}C decays take place in the human body every second. Carbon-14 and other naturally occurring radioactive substances in the body compose a person's background exposure to nuclear radiation. As we will see later in this chapter, this activity level is well below the maximum recommended dosages.

Radioactive Dating

Radioactive dating is a technique that uses naturally occurring radioactivity to determine the age of a material, such as a rock or an ancient artifact. The basic approach is to estimate the original number of nuclei in a material and the present number of nuclei in the material (after decay), and then use the known value of the decay constant λ and **???** to calculate the total time of the decay, *t*.

An important method of radioactive dating is **carbon-14 dating**. Carbon-14 nuclei are produced when high-energy solar radiation strikes ^{14}N nuclei in the upper atmosphere and subsequently decay with a half-life of 5730 years. Radioactive carbon has the same chemistry as stable carbon, so it combines with the ecosphere and eventually becomes part of every living organism. Carbon-14 has an abundance of 1.3 parts per trillion of normal carbon. Therefore, if you know the number of carbon nuclei in an object, you multiply that number by 1.3×10^{-12} to find the number of ^{14}C nuclei in that object. When an organism dies, carbon exchange with the environment ceases, and ^{14}C is not replenished as it decays.

By comparing the abundance of ^{14}C in an artifact, such as mummy wrappings, with the normal abundance in living tissue, it is possible to determine the mummy's age (or the time since the person's death). Carbon-14 dating can be used for biological tissues as old as 50,000 years, but is generally most accurate for younger samples, since the abundance of ^{14}C nuclei in them is greater. Very old biological materials contain no ^{14}C at all. The validity of carbon dating can be checked by other means, such as by historical knowledge or by tree-ring counting.

Example 18.7

An Ancient Burial Cave

In an ancient burial cave, your team of archaeologists discovers ancient wood furniture. Only 80% of the original ^{14}C remains in the wood. How old is the furniture?

Strategy

The problem statement implies that $N/N_0 = 0.80.$ Therefore, the equation $N = N_0 e^{-\lambda t}$ can be used to find the product, λt. We know the half-life of ^{14}C is 5730 y, so we also know the decay constant, and therefore the total decay time *t*.

Solution

Solving the equation $N = N_0 e^{-\lambda t}$ for N/N_0 gives us

$$\frac{N}{N_0} = e^{-\lambda t}.$$

Thus,

$$0.80 = e^{-\lambda t}.$$

Taking the natural logarithm of both sides of the equation yields

$$\ln 0.80 = -\lambda t,$$

so that

$$-0.223 = -\lambda t.$$

Rearranging the equation to isolate *t* gives us

$$t = \frac{0.223}{\lambda},$$

where

$$\lambda = \frac{0.693}{t_{1/2}} = \frac{0.693}{5730\ \text{y}}.$$

Combining this information yields

$$t = \frac{0.223}{\left(\frac{0.693}{5730\ \text{y}}\right)} = 1844\ \text{y}.$$

Significance

The furniture is almost 2000 years old—an impressive discovery. The typical uncertainty on carbon-14 dating is about 5%, so the furniture is anywhere between 1750 and 1950 years old. This date range must be confirmed by other evidence, such as historical records.

18.3 Check Your Understanding A radioactive nuclide has a high decay rate. What does this mean for its half-life and activity?

Visit the **Radioactive Dating Game (https://openstax.org/l/21raddatgame)** to learn about the types of radiometric dating and try your hand at dating some ancient objects.

Applications in Astrophysics

While the carbon-dating of organic materials may be interesting, perhaps even more important is the use of very long half-life isotopes to date objects in our solar system. In fact, our best calculations of the age of the solar system itself come from using these same kind of radioactive dating techniques on minerals and rocks obtained from the Earth, Moon, comets and meteorites.

In those cases, instead of an isotope like ^{14}C with a half-life of thousands of years, we use isotopes with half-lives on the order of millions or even billions of years.

CHAPTER 18 REVIEW

KEY TERMS

accretion the gradual accumulation of mass, as by a planet forming from colliding particles in the solar nebula

activity magnitude of the decay rate for radioactive nuclides

atomic mass total mass of the protons, neutrons, and electrons in a single atom

atomic mass unit unit used to express the mass of an individual nucleus, where $1\text{u} = 1.66054 \times 10^{-27}$ kg

atomic nucleus tightly packed group of nucleons at the center of an atom

atomic number number of protons in a nucleus

becquerel (Bq) SI unit for the decay rate of a radioactive material, equal to 1 decay/second

binding energy (BE) energy needed to break a nucleus into its constituent protons and neutrons

binding energy per nucleon (BEN) energy need to remove a nucleon from a nucleus

carbon-14 dating method to determine the age of formerly living tissue using the ratio $^{14}\text{C}/^{12}\text{C}$

chart of the nuclides graph comprising stable and unstable nuclei

curie (Ci) unit of decay rate, or the activity of 1 g of ^{226}Ra, equal to 3.70×10^{10} Bq

decay process by which an individual atomic nucleus of an unstable atom loses mass and energy by emitting ionizing particles

decay constant quantity that is inversely proportional to the half-life and that is used in equation for number of nuclei as a function of time

half-life time for half of the original nuclei to decay (or half of the original nuclei remain)

isotopes nuclei having the same number of protons but different numbers of neutrons

lifetime average time that a nucleus exists before decaying

mass defect difference between the mass of a nucleus and the total mass of its constituent nucleons

mass number number of nucleons in a nucleus

neutron number number of neutrons in a nucleus

nucleons protons and neutrons found inside the nucleus of an atom

nuclide nucleus

radioactive dating application of radioactive decay in which the age of a material is determined by the amount of radioactivity of a particular type that occurs

radioactive decay law describes the exponential decrease of parent nuclei in a radioactive sample

radioactivity spontaneous emission of radiation from nuclei

radius of a nucleus radius of a nucleus is defined as $r = r_0 A^{1/3}$

strong nuclear force force that binds nucleons together in the nucleus

KEY EQUATIONS

Atomic mass number	$A = Z + N$
Standard format for expressing an isotope	${}^{A}_{Z}\text{X}$

Nuclear radius, where r_0 is the radius of a single proton	$r = r_0 A^{1/3}$
Mass defect	$\Delta m = Zm_p + (A - Z)m_n - m_{\text{nuc}}$
Binding energy	$E = (\Delta m)c^2$
Binding energy per nucleon	$BEN = \frac{E_b}{A}$
Radioactive decay rate	$-\frac{dN}{dt} = \lambda N$
Radioactive decay law	$N = N_0 e^{-\lambda t}$
Decay constant	$\lambda = \frac{0.693}{T_{1/2}}$
Lifetime of a substance	$\bar{T} = \frac{1}{\lambda}$
Activity of a radioactive substance	$A = A_0 e^{-\lambda t}$
Activity of a radioactive substance (linear form)	$\ln A = -\lambda t + \ln A_0$

SUMMARY

18.1 Formation of the Solar System

- A viable theory of solar system formation must take into account motion constraints, chemical constraints, and age constraints.
- Meteorites, comets, and asteroids are survivors of the solar nebula out of which the solar system formed.
- This nebula was the result of the collapse of an interstellar cloud of gas and dust, which contracted (conserving its angular momentum) to form our star, the Sun, surrounded by a thin, spinning disk of dust and vapor.
- Condensation in the disk led to the formation of planetesimals, which became the building blocks of the planets.
- Accretion of infalling materials heated the planets, leading to their differentiation. The giant planets were also able to attract and hold gas from the solar nebula. After a few million years of violent impacts, most of the debris was swept up or ejected, leaving only the asteroids and cometary remnants surviving to the present.

18.2 Properties of Nuclei

- The atomic nucleus is composed of protons and neutrons.
- The number of protons in the nucleus is given by the atomic number, Z. The number of neutrons in the nucleus is the neutron number, N. The number of nucleons is mass number, A.
- Atomic nuclei with the same atomic number, Z, but different neutron numbers, N, are isotopes of the same element.
- The atomic mass of an element is the weighted average of the masses of its isotopes.

18.3 Nuclear Binding Energy

- The mass defect of a nucleus is the difference between the total mass of a nucleus and the sum of the masses of all its constituent nucleons.
- The binding energy (BE) of a nucleus is equal to the amount of energy released in forming the nucleus, or the mass defect multiplied by the speed of light squared.
- A graph of binding energy per nucleon (BEN) versus atomic number A implies that nuclei divided or combined release an enormous amount of energy.

- The binding energy of a nucleon in a nucleus is analogous to the ionization energy of an electron in an atom.

18.4 Radioactive Decay

- In the decay of a radioactive substance, if the decay constant (λ) is large, the half-life is small, and vice versa.
- The radioactive decay law, $N = N_0 e^{-\lambda t}$, uses the properties of radioactive substances to estimate the age of a substance.
- Radioactive carbon has the same chemistry as stable carbon, so it mixes into the ecosphere and eventually becomes part of every living organism. By comparing the abundance of ^{14}C in an artifact with the normal abundance in living tissue, it is possible to determine the artifact's age.

CONCEPTUAL QUESTIONS

18.1 Formation of the Solar System

1. Describe the solar nebula, and outline the sequence of events within the nebula that gave rise to the planetesimals.

2. Why do the giant planets and their moons have compositions different from those of the terrestrial planets?

18.2 Properties of Nuclei

3. Define and make clear distinctions between the terms neutron, nucleon, nucleus, and nuclide.

4. What are isotopes? Why do isotopes of the same atom share the same chemical properties?

18.3 Nuclear Binding Energy

5. Explain why a bound system should have less mass than its components. Why is this not observed traditionally, say, for a building made of bricks?

6. Why is the number of neutrons greater than the number of protons in stable nuclei that have an *A* greater than about 40? Why is this effect more pronounced for the heaviest nuclei?

7. To obtain the most precise value of the binding energy per nucleon, it is important to take into account forces between nucleons at the surface of the nucleus. Will surface effects increase or decrease estimates of BEN?

18.4 Radioactive Decay

8. How is the initial activity rate of a radioactive substance related to its half-life?

9. For the carbon dating described in this chapter, what important assumption is made about the time variation in the intensity of cosmic rays?

PROBLEMS

18.1 Formation of the Solar System

10. How long would material take to go around if the solar nebula in **Example 18.1** became the size of Earth's orbit?

18.2 Properties of Nuclei

11. Find the atomic numbers, mass numbers, and neutron numbers for (a) $^{58}_{29}\text{Cu}$, (b) $^{24}_{11}\text{Na}$, (c) $^{210}_{84}\text{Po}$, (d) $^{45}_{20}\text{Ca}$, and (e) $^{206}_{82}\text{Pb}$.

12. Silver has two stable isotopes. The nucleus, $^{107}_{47}\text{Ag}$, has atomic mass 106.905095 g/mol with an abundance of 51.83%; whereas $^{109}_{47}\text{Ag}$ has atomic mass 108.904754 g/mol with an abundance of 48.17%. Find the atomic mass of the element silver.

13. The mass (*M*) and the radius (*r*) of a nucleus can be expressed in terms of the mass number, *A*. (a) Show that the density of a nucleus is independent of *A*. (b) Calculate the density of a gold (Au) nucleus. Compare your answer to that for iron (Fe).

14. A particle has a mass equal to 10 u. If this mass is converted completely into energy, how much energy is released? Express your answer in mega-electron volts (MeV). (Recall that $1\text{ eV} = 1.6 \times 10^{-19}\text{ J}$.)

15. Find the length of a side of a cube having a mass of 1.0 kg and the density of nuclear matter.

16. The detail that you can observe using a probe is limited by its wavelength. Calculate the energy of a particle that has a wavelength of $1 \times 10^{-16}\,\text{m}$, small enough to detect details about one-tenth the size of a nucleon.

18.3 Nuclear Binding Energy

17. How much energy would be released if six hydrogen atoms and six neutrons were combined to form ${}^{12}_{6}\text{C}$?

18. Find the mass defect and the binding energy for the helium-4 nucleus.

19. ${}^{56}\text{Fe}$ is among the most tightly bound of all nuclides. It makes up more than 90% of natural iron. Note that ${}^{56}\text{Fe}$ has even numbers of protons and neutrons. Calculate the binding energy per nucleon for ${}^{56}\text{Fe}$ and compare it with the approximate value obtained from the graph in **Figure 18.10**.

20. ${}^{209}\text{Bi}$ is the heaviest stable nuclide, and its BEN is low compared with medium-mass nuclides. Calculate BEN for this nucleus and compare it with the approximate value obtained from the graph in **Figure 18.10**.

21. (a) Calculate BEN for ${}^{235}\text{U}$, the rarer of the two most common uranium isotopes; (b) Calculate BEN for ${}^{238}\text{U}$. (Most of uranium is ${}^{238}\text{U}$.)

22. The fact that BEN peaks at roughly $A = 60$ implies that the *range* of the strong nuclear force is about the diameter of this nucleus.

(a) Calculate the diameter of $A = 60$ nucleus.

(b) Compare BEN for ${}^{58}\text{Ni}$ and ${}^{90}\text{Sr}$. The first is one of the most tightly bound nuclides, whereas the second is larger and less tightly bound.

18.4 Radioactive Decay

23. A sample of radioactive material is obtained from a very old rock. A plot ln*A* verses *t* yields a slope value of $-10^{-9}\,\text{s}^{-1}$ (see **Figure 18.13**(b)). What is the half-life of this material?

24. Show that: $\bar{T} = \frac{1}{\lambda}$.

25. The half-life of strontium-91, ${}^{91}_{38}\text{Sr}$ is 9.70 h. Find (a) its decay constant and (b) for an initial 1.00-g sample, the activity after 15 hours.

26. A sample of pure carbon-14 $(T_{1/2} = 5730\,\text{y})$ has an activity of $1.0\,\mu\,\text{Ci}$. What is the mass of the sample?

27. A radioactive sample initially contains 2.40×10^{-2} mol of a radioactive material whose half-life is 6.00 h. How many moles of the radioactive material remain after 6.00 h? After 12.0 h? After 36.0 h?

28. An old campfire is uncovered during an archaeological dig. Its charcoal is found to contain less than 1/1000 the normal amount of ${}^{14}\text{C}$. Estimate the minimum age of the charcoal, noting that $2^{10} = 1024$.

29. Calculate the activity R, in curies of 1.00 g of ${}^{226}\text{Ra}$. (b) Explain why your answer is not exactly 1.00 Ci, given that the curie was originally supposed to be exactly the activity of a gram of radium.

30. Natural uranium consists of ${}^{235}\text{U}$ (percent abundance $= 0.7200\%$, $\lambda = 3.12 \times 10^{-17}/\text{s}$) and ${}^{238}\text{U}$ (percent abundance $= 99.27\%$, $\lambda = 4.92 \times 10^{-18}/\text{s}$). What were the values for percent abundance of ${}^{235}\text{U}$ and ${}^{238}\text{U}$ when Earth formed 4.5×10^{9} years ago?

31. World War II aircraft had instruments with glowing radium-painted dials. The activity of one such instrument was 1.0×10^{5} Bq when new. (a) What mass of ${}^{226}\text{Ra}$ was present? (b) After some years, the phosphors on the dials deteriorated chemically, but the radium did not escape. What is the activity of this instrument 57.0 years after it was made?

32. The ${}^{210}\text{Po}$ source used in a physics laboratory is labeled as having an activity of $1.0\,\mu\text{Ci}$ on the date it was prepared. A student measures the radioactivity of this source with a Geiger counter and observes 1500 counts per minute. She notices that the source was prepared 120 days before her lab. What fraction of the decays is she observing with her apparatus?

33. Armor-piercing shells with depleted uranium cores are fired by aircraft at tanks. (The high density of the uranium makes them effective.) The uranium is called depleted because it has had its ^{235}U removed for reactor use and is nearly pure ^{238}U. Depleted uranium has been erroneously called nonradioactive. To demonstrate that this is wrong: (a) Calculate the activity of 60.0 g of pure ^{238}U. (b) Calculate the activity of 60.0 g of natural uranium, neglecting the ^{234}U and all daughter nuclides.

34. A radioactive nucleus has a half-life of 5×10^8 years. Assuming that a sample of rock (say, in an asteroid) solidified right after the solar system formed, approximately what fraction of the radioactive element should be left in the rock today?

19 | COMPARATIVE PLANETOLOGY

Chapter Outline

Introduction

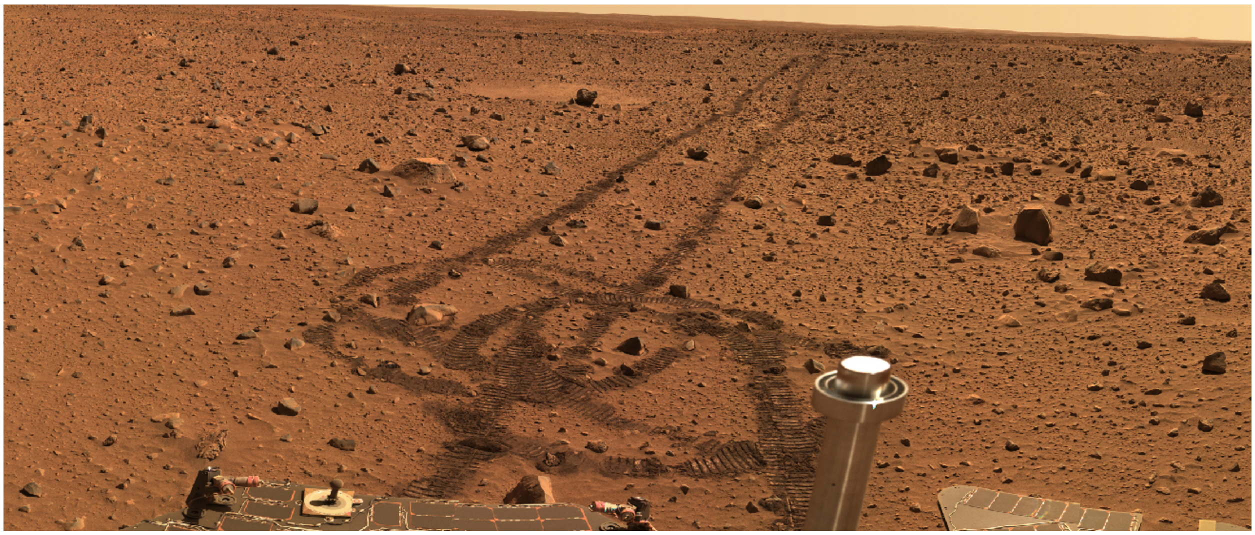


Figure 19.1 This May 2004 image shows the tracks made by the Mars Exploration *Spirit* rover on the surface of the red planet. *Spirit* was active on Mars between 2004 and 2010, twenty times longer than its planners had expected. It "drove" over 7.73 kilometers in the process of examining the martian landscape. (credit: modification of work by NASA/JPL/Cornell)

Introduction

From the details of The Formation of the Solar System we understand the basic reasons behind the creation of two very different kinds of planets, which we call **terrestrial** and **jovian**. But, even within the family of terrestrial planets, much can be learned about the details of planetary formation and evolution by comparing what we observe today on each planet. Such a study is called **comparative planetology**.

The Moon and Mercury are geologically dead. In contrast, the larger terrestrial planets—Earth, Venus, and Mars—are more active and interesting worlds. We will briefly discuss Earth, Venus and Mars. The latter two are the nearest planets and the most accessible to spacecraft. Not surprisingly, the greatest effort in planetary exploration has been devoted to these fascinating worlds. In the chapter, we discuss some of the results of more than four decades of scientific exploration of Mars and Venus. Mars is exceptionally interesting, with evidence that points to habitable conditions in the past. Even today, we are discovering things about Mars that make it the most likely place where humans might set up a habitat in the future. However, our robot explorers have clearly shown that neither Venus nor Mars has conditions similar to Earth. How did it happen that these three neighboring terrestrial planets have diverged so dramatically in their evolution?

19.1 | Composition and Structure of Planets

Learning Objectives

By the end of this section, you will be able to:

- Describe the characteristics of the giant planets, terrestrial planets, and small bodies in the solar system
- Explain what influences the temperature of a planet's surface
- Explain why there is geological activity on some planets and not on others

The fact that there are two distinct kinds of planets—the rocky terrestrial planets and the gas-rich jovian planets—leads us to believe that they formed under different conditions. Certainly their compositions are dominated by different elements. Let us look at each type in more detail.

The Giant Planets

The two largest planets, Jupiter and Saturn, have nearly the same chemical makeup as the Sun; they are composed primarily of the two elements hydrogen and helium, with 75% of their mass being hydrogen and 25% helium. On Earth, both hydrogen and helium are gases, so Jupiter and Saturn are sometimes called gas planets. But, this name is misleading. Jupiter and Saturn are so large that the gas is compressed in their interior until the hydrogen becomes a liquid. Because the bulk of both planets consists of compressed, liquefied hydrogen, we should really call them liquid planets.

Under the force of gravity, the heavier elements sink toward the inner parts of a liquid or gaseous planet. Both Jupiter and Saturn, therefore, have cores composed of heavier rock, metal, and ice, but we cannot see these regions directly. In fact, when we look down from above, all we see is the atmosphere with its swirling clouds (**Figure 19.2**). We must infer the existence of the denser core inside these planets from studies of each planet's gravity.

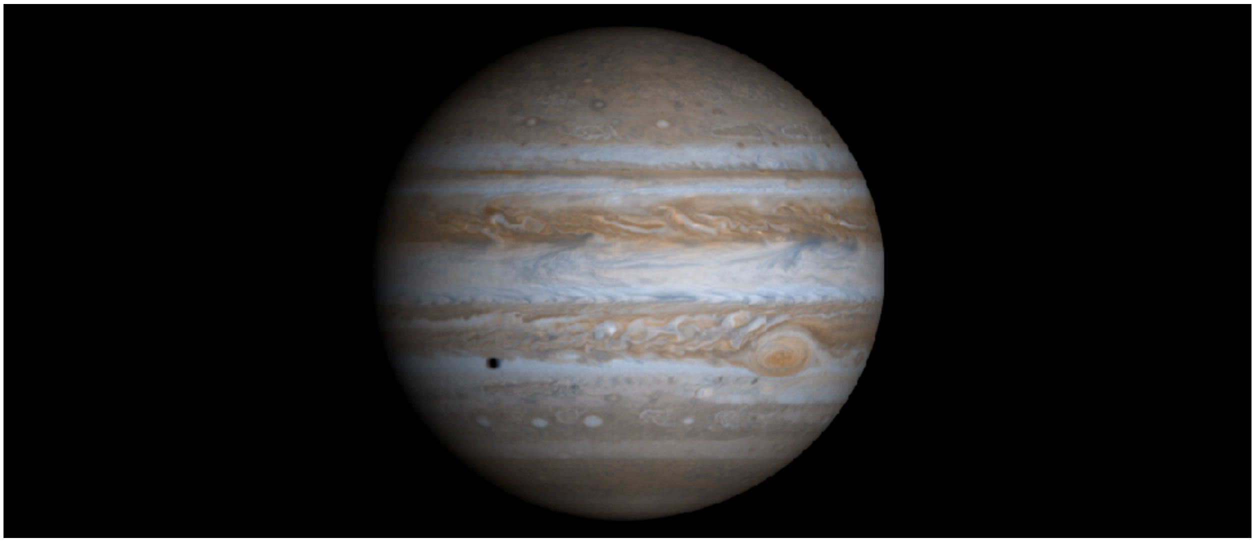


Figure 19.2 This true-color image of Jupiter was taken from the Cassini spacecraft in 2000. (credit: modification of work by NASA/JPL/University of Arizona)

Uranus and Neptune are much smaller than Jupiter and Saturn, but each also has a core of rock, metal, and ice. Uranus and Neptune were less efficient at attracting hydrogen and helium gas, so they have much smaller atmospheres in proportion to their cores.

Chemically, each giant planet is dominated by hydrogen and its many compounds. Nearly all the oxygen present is combined chemically with hydrogen to form water (H_2O). Chemists call such a hydrogen-dominated composition *reduced*. Throughout the outer solar system, we find abundant water (mostly in the form of ice) and reducing chemistry.

The Terrestrial Planets

The terrestrial planets are quite different from the giants. In addition to being much smaller, they are composed primarily of rocks and metals. These, in turn, are made of elements that are less common in the universe as a whole. The most abundant rocks, called silicates, are made of silicon and oxygen, and the most common metal is iron. We can tell from their densities (see **Table 7.2**) that Mercury has the greatest proportion of metals (which are denser) and the Moon has the lowest. Earth, Venus, and Mars all have roughly similar bulk compositions: about one third of their mass consists of iron-nickel or iron-sulfur combinations; two thirds is made of silicates. Because these planets are largely composed of oxygen compounds (such as the silicate minerals of their crusts), their chemistry is said to be *oxidized*.

When we look at the internal structure of each of the terrestrial planets, we find that the densest metals are in a central core, with the lighter silicates near the surface. If these planets were liquid, like the giant planets, we could understand this effect as the result the sinking of heavier elements due to the pull of gravity. This leads us to conclude that, although the terrestrial planets are solid today, at one time they must have been hot enough to melt.

Differentiation is the process by which gravity helps separate a planet's interior into layers of different compositions and densities. The heavier metals sink to form a core, while the lightest minerals float to the surface to form a crust. Later, when the planet cools, this layered structure is preserved. In order for a rocky planet to differentiate, it must be heated to the melting point of rocks, which is typically more than 1300 K.

Moons, Asteroids, and Comets

Chemically and structurally, Earth's Moon is like the terrestrial planets, but most moons are in the outer solar system, and they have compositions similar to the cores of the giant planets around which they orbit. The three largest moons—Ganymede and Callisto in the jovian system, and Titan in the saturnian system—are composed half of frozen water, and half of rocks and metals. Most of these moons differentiated during formation, and today they have cores of rock and metal, with upper layers and crusts of very cold and—thus very hard—ice (**Figure 19.3**).

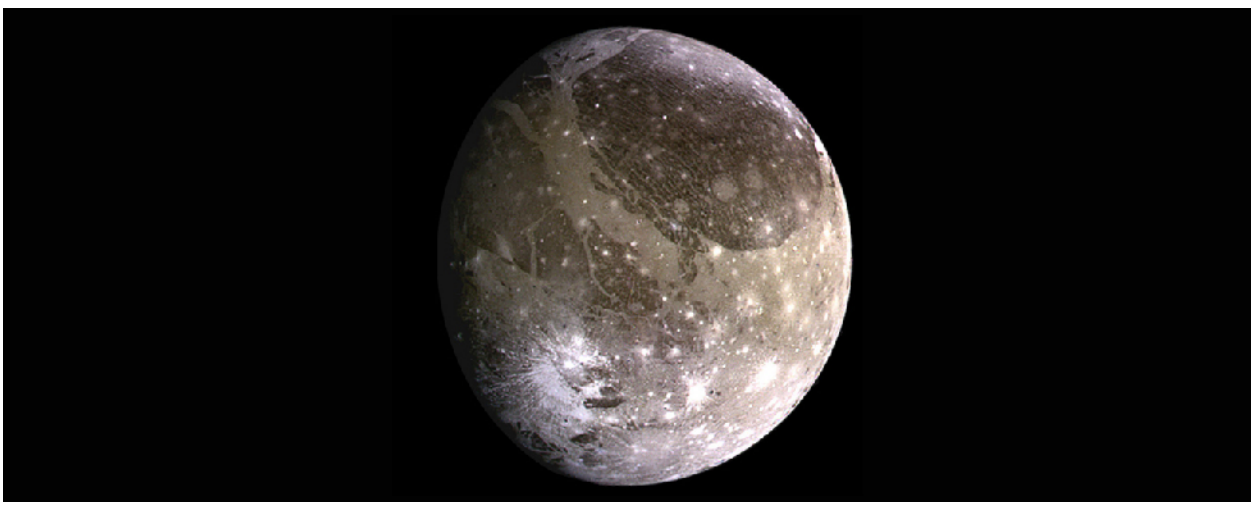


Figure 19.3 This view of Jupiter's moon Ganymede was taken in June 1996 by the Galileo spacecraft. The brownish gray color of the surface indicates a dusty mixture of rocky material and ice. The bright spots are places where recent impacts have uncovered fresh ice from underneath. (credit: modification of work by NASA/JPL)

Most of the asteroids and comets, as well as the smallest moons, were probably never heated to the melting point. However, some of the largest asteroids, such as Vesta, appear to be differentiated; others are fragments from differentiated bodies. Because most asteroids and comets retain their original composition, they represent relatively unmodified material dating back to the time of the formation of the solar system. In a sense, they act as chemical fossils, helping us to learn about a time long ago whose traces have been erased on larger worlds.

Temperatures: Going to Extremes

Generally speaking, the farther a planet or moon is from the Sun, the cooler its surface. The planets are heated by the radiant energy of the Sun, which gets weaker with the square of the distance. You know how rapidly the heating effect of a fireplace or an outdoor radiant heater diminishes as you walk away from it; the same effect applies to the Sun. Mercury, the closest planet to the Sun, has a blistering surface temperature that ranges from 280–430 °C on its sunlit side, whereas the surface temperature on Pluto is only about –220 °C, colder than liquid air.

Mathematically, the temperatures decrease approximately in proportion to the square root of the distance from the Sun. Pluto is about 30 AU at its closest to the Sun (or 100 times the distance of Mercury) and about 49 AU at its farthest from the Sun. Thus, Pluto's temperature is less than that of Mercury by the square root of 100, or a factor of 10: from 500 K to 50 K. Let's see why this is so.

No-Greenhouse Temperatures

A planet is in thermal equilibrium with its surroundings. Recall from **Mechanisms of Heat Transfer** that this implies a balance between the incoming radiation absorbed from the Sun and the outgoing radiation from the planet into space.

The luminosity (power) of the sun, $\mathbf{L_{sun}} = 3.85 \times 10^{26}$ Watts. A planet of radius **R** and emissivity **e** located at a distance **D** from the Sun will absorb energy at a rate

$$P_{\text{in}} = \left(\frac{L_{\text{Sun}}}{4\pi D^2}\right)e\left(\pi R^2\right) \tag{19.1}$$

Here $\frac{L_{\text{Sun}}}{4\pi D^2}$ is the intensity of the incoming solar radiation (the power per unit area) at a distance **D** away from the Sun.

The area of the planet available to absorb solar radiation is just the area of a circle with the planet's radius, πR^2.

At the same time, the planet at temperature **T** will be emitting energy at a rate

$$P_{\text{out}} = \sigma T^4\left(4\pi R^2\right) \tag{19.2}$$

Note that this last equation assumes that the emissivity (in the infrared) of the planet is essentially 1.

Equating the power in and out, as must be the case for thermal equilibrium, and combining all of the values for the various constants in appropriate units, we come up with an expression for what is refereed to as the **no-greenhouse temperature** for a planet:

$$T_{\text{no-greenhouse}} = 279\text{ K}\sqrt[4]{\frac{e}{D^2}} \tag{19.3}$$

where the distance, D, is expressed in AU. Since the fourth root of the reciprocal of D^2 is the same as the reciprocal of the square root of D, we have proven our previous assertion. e is the average emissivity of the planet's surface in the visible portion of the spectrum, determined primarily by the reflectivity of its surface to sunlight.

Both the **emissivity**, e, and the **reflectivity** (sometimes referred to as the **albedo**), are numbers with a range between 0 and 1. They represent the fraction of light absorbed and reflected, respectively, by a planet. Their sum is always 1 (or 100%).

Example 19.1 Earth with No Greenhouse Effect

Let's apply this equation to our own planet. Obviously, Earth is located 1 AU from the Sun. Its average reflectivity to sunlight is about 29% or 0.29. Therefore, its emissivity $e = 1 - 0.29 = 0.71$

So, its no-greenhouse temperature is $T_{\text{no-greenhouse}} = 279\text{ K}\sqrt[4]{\frac{0.71}{1^2}} = 256\text{ K}$

This means that, in the absence of a greenhouse effect, Earth's average surface temperature would be about -17°C.

In addition to its distance from the Sun, the surface temperature of a planet can be influenced strongly by its atmosphere. Without our atmospheric insulation (the greenhouse effect, which keeps the heat in), the oceans of Earth would be permanently frozen. Conversely, if Mars once had a larger atmosphere in the past, it could have supported a more temperate climate than it has today. Venus is an even more extreme example, where its thick atmosphere of carbon dioxide acts as insulation, reducing the escape of heat built up at the surface, resulting in temperatures greater than those on Mercury. Today, Earth is the only planet where surface temperatures generally lie between the freezing and boiling points of water. As far as we know, Earth is the only planet to support life.

There's No Place Like Home

In the classic film *The Wizard of Oz*, Dorothy, the heroine, concludes after her many adventures in "alien" environments that "there's no place like home." The same can be said of the other worlds in our solar system. There are many fascinating places, large and small, that we might like to visit, but humans could not survive on any without a great deal of artificial assistance.

A thick carbon dioxide atmosphere keeps the surface temperature on our neighbor Venus at a sizzling 700 K (near 900 °F). Mars, on the other hand, has temperatures generally below freezing, with air (also mostly carbon dioxide) so thin that it resembles that found at an altitude of 30 kilometers (100,000 feet) in Earth's atmosphere. And the red planet is so dry that it has not had any rain for billions of years.

The outer layers of the jovian planets are neither warm enough nor solid enough for human habitation. Any bases we build in the systems of the giant planets may well have to be in space or one of their moons—none of which is particularly hospitable to a luxury hotel with a swimming pool and palm trees. Perhaps we will find warmer havens deep inside the clouds of Jupiter or in the ocean under the frozen ice of its moon Europa.

All of this suggests that we had better take good care of Earth because it is the only site where life as we know it could survive. Recent human activity may be reducing the habitability of our planet by adding pollutants to the atmosphere, especially the potent greenhouse gas carbon dioxide. Human civilization is changing our planet dramatically, and these changes are not necessarily for the better. In a solar system that seems unready to receive us, making Earth less hospitable to life may be a grave mistake.

Geological Activity

The crusts of all of the terrestrial planets, as well as of the larger moons, have been modified over their histories by both internal and external forces. Externally, each has been battered by a slow rain of projectiles from space, leaving their surfaces pockmarked by impact craters of all sizes (see **Figure 7.4**). We have good evidence that this bombardment was far greater in the early history of the solar system, but it certainly continues to this day, even if at a lower rate. The collision of more than 20 large pieces of Comet Shoemaker–Levy 9 with Jupiter in the summer of 1994 (see **Figure 19.4**) is one dramatic example of this process.

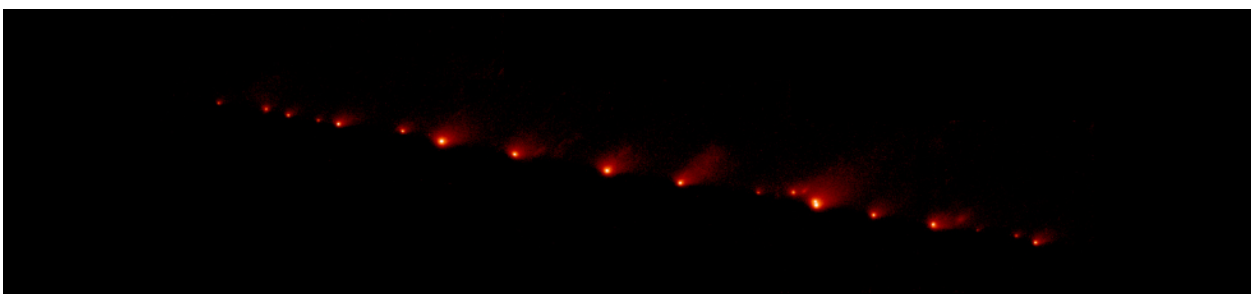


Figure 19.4 In this image of Comet Shoemaker–Levy 9 taken on May 17, 1994, by NASA's Hubble Space Telescope, you can see about 20 icy fragments into which the comet broke. The comet was approximately 660 million kilometers from Earth, heading on a collision course with Jupiter. (credit: modification of work by NASA, ESA, H. Weaver (STScl), E. Smith (STScl))

Figure 19.5 shows the aftermath of these collisions, when debris clouds larger than Earth could be seen in Jupiter's atmosphere.

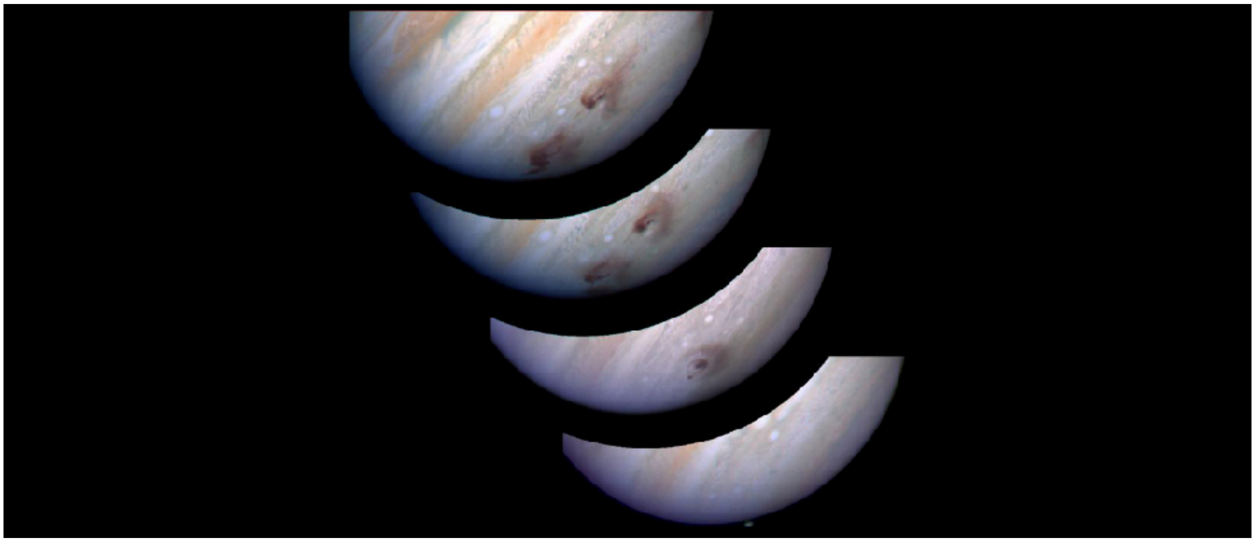


Figure 19.5 The Hubble Space Telescope took this sequence of images of Jupiter in summer 1994, when fragments of Comet Shoemaker–Levy 9 collided with the giant planet. Here we see the site hit by fragment G, from five minutes to five days after impact. Several of the dust clouds generated by the collisions became larger than Earth. (credit: modification of work by H. Hammel, NASA)

During the time all the planets have been subject to such impacts, internal forces on the terrestrial planets have buckled and twisted their crusts, built up mountain ranges, erupted as volcanoes, and generally reshaped the surfaces in what we call geological activity. (The prefix *geo* means "Earth," so this is a bit of an "Earth-chauvinist" term, but it is so widely used that we bow to tradition.) Among the terrestrial planets, Earth and Venus have experienced the most geological activity over their histories, although some of the moons in the outer solar system are also surprisingly active. In contrast, our own Moon is a dead world where geological activity ceased billions of years ago.

Geological activity on a planet is the result of a hot interior. The forces of volcanism and mountain building are driven by heat escaping from the interiors of planets. As we will see, each of the planets was heated at the time of its birth, and this primordial heat initially powered extensive volcanic activity, even on our Moon. But, small objects such as the Moon soon cooled off. The larger the planet or moon, the longer it retains its internal heat, and therefore the more we expect to see surface evidence of continuing geological activity. The effect is similar to our own experience with a hot baked potato: the larger the potato, the more slowly it cools. If we want a potato to cool quickly, we cut it into small pieces.

For the most part, the history of volcanic activity on the terrestrial planets conforms to the predictions of this simple theory. The Moon, the smallest of these objects, is a geologically dead world. Although we know less about Mercury, it seems likely that this planet, too, ceased most volcanic activity about the same time the Moon did. Mars represents an intermediate case. It has been much more active than the Moon, but less so than Earth. Earth and Venus, the largest terrestrial planets, still have molten interiors even today, some 4.5 billion years after their birth.

Challenge Problems

Exercise 19.1

Starting with **Equation 19.1** and **Equation 19.2**, derived the numerical value given in **Equation 19.3**.

19.2 | Dating Planetary Surfaces

Learning Objectives

By the end of this section, you will be able to:

- Explain how astronomers can tell whether a planetary surface is geologically young or old
- Describe different methods for dating planets

How do we know the age of the surfaces we see on planets and moons? If a world has a surface (as opposed to being mostly gas and liquid), astronomers have developed some techniques for estimating how long ago that surface solidified. Note that the age of these surfaces is not necessarily the age of the planet as a whole. On geologically active objects (including Earth), vast outpourings of molten rock or the erosive effects of water and ice, which we call planet weathering, have erased evidence of earlier epochs and present us with only a relatively young surface for investigation.

Counting the Craters

One way to estimate the age of a surface is by counting the number of impact craters. This technique works because the rate at which impacts have occurred in the solar system has been roughly constant for several billion years. Thus, in the absence of forces to eliminate craters, the number of craters is simply proportional to the length of time the surface has been exposed. This technique has been applied successfully to many solid planets and moons (**Figure 19.6**).

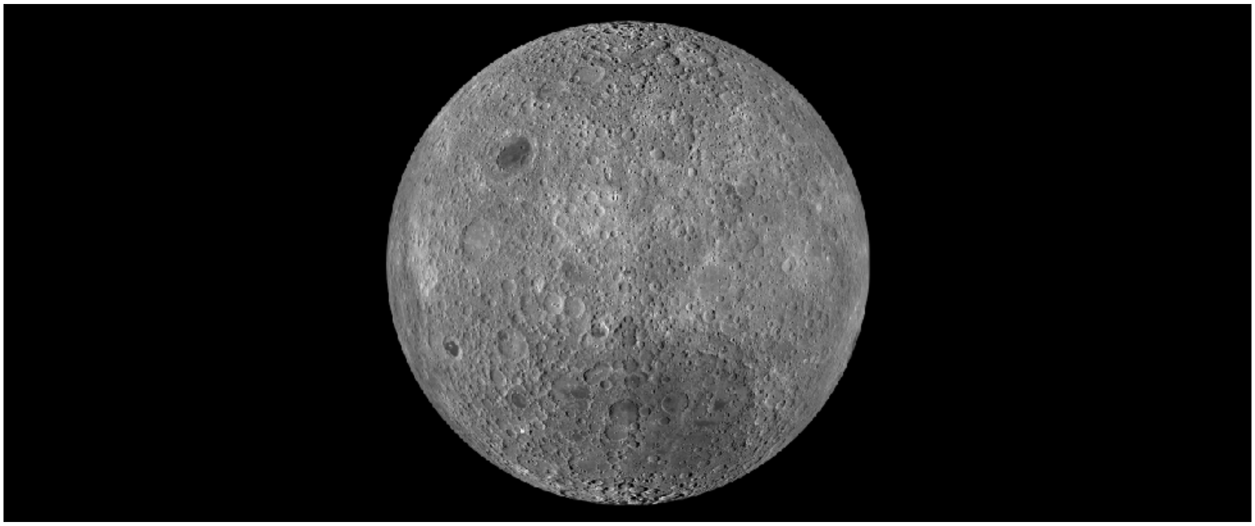


Figure 19.6 This composite image of the Moon's surface was made from many smaller images taken between November 2009 and February 2011 by the Lunar Reconnaissance Orbiter (LRO) and shows craters of many different sizes. (credit: modification of work by NASA/GSFC/Arizona State University)

Bear in mind that crater counts can tell us only the time since the surface experienced a major change that could modify or erase preexisting craters. Estimating ages from crater counts is a little like walking along a sidewalk in a snowstorm after the snow has been falling steadily for a day or more. You may notice that in front of one house the snow is deep, while next door the sidewalk may be almost clear. Do you conclude that less snow has fallen in front of Ms. Jones' house than Mr. Smith's? More likely, you conclude that Jones has recently swept the walk clean and Smith has not. Similarly, the numbers of craters indicate how long it has been since a planetary surface was last "swept clean" by ongoing lava flows or by molten materials ejected when a large impact happened nearby.

Still, astronomers can use the numbers of craters on different parts of the same world to provide important clues about how regions on that world evolved. On a given planet or moon, the more heavily cratered terrain will generally be older (that is, more time will have elapsed there since something swept the region clean).

Radioactive Rocks

Another way to trace the history of a solid world is to measure the age of individual rocks. After samples were brought back from the Moon by Apollo astronauts, the techniques that had been developed to date rocks on Earth were applied to rock samples from the Moon to establish a geological chronology for the Moon. Furthermore, a few samples of material from the Moon, Mars, and the large asteroid Vesta have fallen to Earth as meteorites and can be examined directly.

Scientists measure the age of rocks using the properties of natural **radioactivity** which we discussed in **Section 18.4**. Around the beginning of the twentieth century, physicists began to understand that some atomic nuclei are not stable but can split apart (decay) spontaneously into smaller nuclei. The process of radioactive decay involves the emission of particles such as electrons, or of radiation in the form of gamma rays (see the chapter on **Spectroscopy**).

For any one radioactive nucleus, it is not possible to predict when the decay process will happen. Such decay is random in nature, like the throw of dice: as gamblers have found all too often, it is impossible to say just when the dice will come up 7 or 11. But, for a very large number of dice tosses, we can calculate the odds that 7 or 11 will come up. Similarly, if we have a very large number of radioactive atoms of one type (say, uranium), there is a specific time period, called its **half-life**, during which the chances are fifty-fifty that decay will occur for any of the nuclei.

A particular nucleus may last a shorter or longer time than its half-life, but in a large sample, almost exactly half of the nuclei will have decayed after a time equal to one half-life. Half of the remaining nuclei will have decayed after two half-lives pass, leaving only one half of a half—or one quarter—of the original sample (**Figure 19.7**).

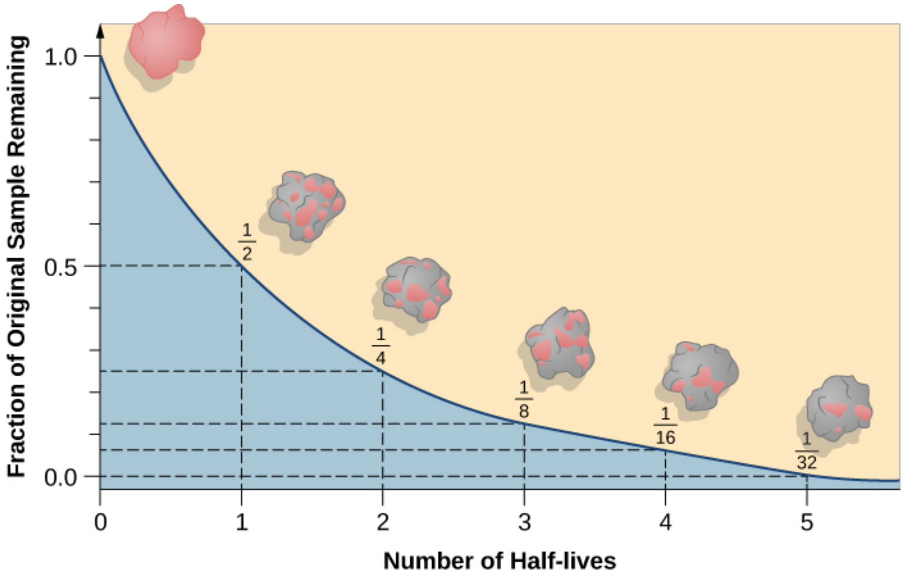


Figure 19.7 This graph shows (in pink) the amount of a radioactive sample that remains after several half-lives have passed. After one half-life, half the sample is left; after two half-lives, one half of the remainder (or one quarter) is left; and after three half-lives, one half of that (or one eighth) is left. Note that, in reality, the decay of radioactive elements in a rock sample would not cause any visible change in the appearance of the rock; the splashes of color are shown here for conceptual purposes only.

If you had 1 gram of pure radioactive nuclei with a half-life of 100 years, then after 100 years you would have 1/2 gram; after 200 years, 1/4 gram; after 300 years, only 1/8 gram; and so forth. However, the material does not disappear. Instead, the radioactive atoms are replaced with their decay products. Sometimes the radioactive atoms are called *parents* and the decay products are called *daughter* elements.

In this way, radioactive elements with half-lives we have determined can provide accurate nuclear clocks. By comparing how much of a radioactive parent element is left in a rock to how much of its daughter products have accumulated, we can learn how long the decay process has been going on and hence how long ago the rock formed. **Table 19.1** summarizes the decay reactions used most often to date lunar and terrestrial rocks.

Radioactive Decay Reaction Used to Date Rocks[1]

Parent	Daughter	Half-Life (billions of years)
Samarium-147	Neodymium-143	106
Rubidium-87	Strontium-87	48.8
Thorium-232	Lead-208	14.0
Uranium-238	Lead-206	4.47
Potassium-40	Argon-40	1.31

Table 19.1

1. The number after each element is its atomic weight, equal to the number of protons plus neutrons in its nucleus. This specifies the *isotope* of the element; different isotopes of the same element differ in the number of neutrons.

PBS provides an **evolution series excerpt (https://openstaxcollege.org/l/30pbsradiomat)** that explains how we use radioactive elements to date Earth.

This **Science Channel video (https://openstaxcollege.org/l/30billnyevideo)** features Bill Nye the Science Guy showing how scientists have used radioactive dating to determine the age of Earth.

When astronauts first flew to the Moon, one of their most important tasks was to bring back lunar rocks for radioactive age-dating. Until then, astronomers and geologists had no reliable way to measure the age of the lunar surface. Counting craters had let us calculate relative ages (for example, the heavily cratered lunar highlands were older than the dark lava plains), but scientists could not measure the actual age in years. Some thought that the ages were as young as those of Earth's surface, which has been resurfaced by many geological events. For the Moon's surface to be so young would imply active geology on our satellite. Only in 1969, when the first Apollo samples were dated, did we learn that the Moon is an ancient, geologically dead world. Using such dating techniques, we have been able to determine the ages of both Earth and the Moon: each was formed about 4.5 billion years ago (although, as we shall see, Earth probably formed earlier).

We should also note that the decay of radioactive nuclei generally releases energy in the form of heat. Although the energy from a single nucleus is not very large (in human terms), the enormous numbers of radioactive nuclei in a planet or moon (especially early in its existence) can be a significant source of internal energy for that world. Geologists estimate that about half of Earth's current internal heat budget comes from the decay of radioactive isotopes in its interior.

19.3 | Earth's Atmosphere

Learning Objectives

By the end of this section, you will be able to:

- Differentiate between Earth's various atmospheric layers
- Describe the chemical composition and possible origins of our atmosphere
- Explain the difference between weather and climate
- Describe the causes and effects of the atmospheric greenhouse effect and global warming
- Describe the impact of human activity on our planet's atmosphere and ecology

We live at the bottom of the ocean of air that envelops our planet. The atmosphere, weighing down upon Earth's surface under the force of gravity, exerts a pressure at sea level that scientists define as 1 **bar** (a term that comes from the same root as *barometer*, an instrument used to measure atmospheric pressure). A bar of pressure means that each square centimeter of Earth's surface has a weight equivalent to 1.03 kilograms pressing down on it. Humans have evolved to live at this pressure; make the pressure a lot lower or higher and we do not function well.

The total mass of Earth's atmosphere is about 5×10^{18} kilograms. This sounds like a large number, but it is only about a millionth of the total mass of Earth. The atmosphere represents a smaller fraction of Earth than the fraction of your mass represented by the hair on your head.

Structure of the Atmosphere

The structure of the atmosphere is illustrated in **Figure 19.8**. Most of the atmosphere is concentrated near the surface of Earth, within about the bottom 10 kilometers where clouds form and airplanes fly. Within this region—called the **troposphere**—warm air, heated by the surface, rises and is replaced by descending currents of cooler air; this is an example of convection. This circulation generates clouds and wind. Within the troposphere, temperature decreases rapidly with increasing elevation to values near 50 °C below freezing at its upper boundary, where the **stratosphere** begins. Most of the stratosphere, which extends to about 50 kilometers above the surface, is cold and free of clouds.

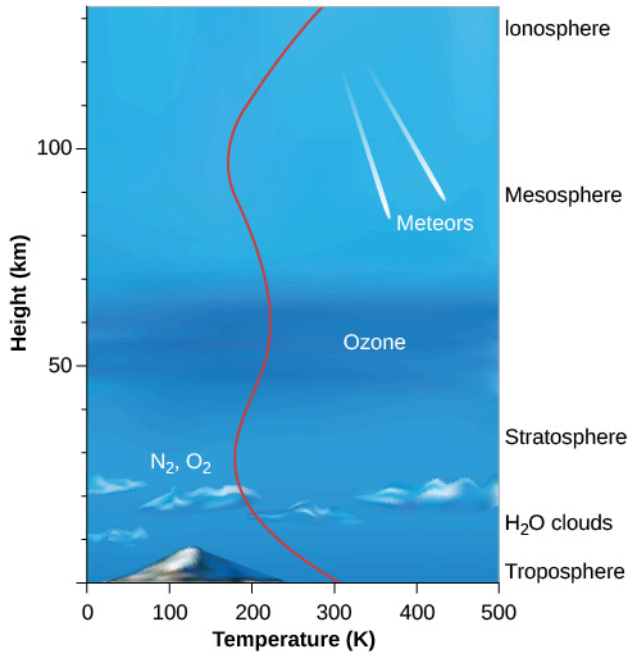


Figure 19.8 Height increases up the left side of the diagram, and the names of the different atmospheric layers are shown at the right. In the upper ionosphere, ultraviolet radiation from the Sun can strip electrons from their atoms, leaving the atmosphere ionized. The curving red line shows the temperature (see the scale on the *x*-axis).

Near the top of the stratosphere is a layer of **ozone** (O_3), a heavy form of oxygen with three atoms per molecule instead of the usual two. Because ozone is a good absorber of ultraviolet light, it protects the surface from some of the Sun's dangerous ultraviolet radiation, making it possible for life to exist on Earth. The breakup of ozone adds heat to the stratosphere, reversing the decreasing temperature trend in the troposphere. Because ozone is essential to our survival, we reacted with justifiable concern to evidence that became clear in the 1980s that atmospheric ozone was being destroyed by human activities. By international agreement, the production of industrial chemicals that cause ozone depletion, called chlorofluorocarbons, or CFCs, has been phased out. As a result, ozone loss has stopped and the "ozone hole" over the Antarctic is shrinking gradually. This is an example of how concerted international action can help maintain the habitability of Earth.

Visit NASA's scientific visualization studio for a **short video (https://openstax.org/l/302065regcfc)** of what would have happened to Earth's ozone layer by 2065 if CFCs had not been regulated.

At heights above 100 kilometers, the atmosphere is so thin that orbiting satellites can pass through it with very little friction. Many of the atoms are ionized by the loss of an electron, and this region is often called the ionosphere. At these elevations, individual atoms can occasionally escape completely from the gravitational field of Earth. There is a continuous, slow leaking of atmosphere—especially of lightweight atoms, which move faster than heavy ones. Earth's atmosphere cannot, for example, hold on for long to hydrogen or helium, which escape into space. Earth is not the only planet to experience atmosphere leakage. Atmospheric leakage also created Mars' thin atmosphere. Venus' dry atmosphere evolved because its proximity to the Sun vaporized and dissociated any water, with the component gases lost to space.

Atmospheric Composition and Origin

At Earth's surface, the atmosphere consists of 78% nitrogen (N_2), 21% oxygen (O_2), and 1% argon (Ar), with traces of water vapor (H_2O), carbon dioxide (CO_2), and other gases. Variable amounts of dust particles and water droplets are also found suspended in the air.

A complete census of Earth's volatile materials, however, should look at more than the gas that is now present. *Volatile* materials are those that evaporate at a relatively low temperature. If Earth were just a little bit warmer, some materials that are now liquid or solid might become part of the atmosphere. Suppose, for example, that our planet were heated to above the boiling point of water (100 °C, or 373 K); that's a large change for humans, but a small change compared to the range of possible temperatures in the universe. At 100 °C, the oceans would boil and the resulting water vapor would become a part of the atmosphere.

To estimate how much water vapor would be released, note that there is enough water to cover the entire Earth to a depth of about 300 meters. Because the pressure exerted by 10 meters of water is equal to about 1 bar, the average pressure at the ocean floor is about 300 bars. Water weighs the same whether in liquid or vapor form, so if the oceans boiled away, the atmospheric pressure of the water would still be 300 bars. Water would therefore greatly dominate Earth's atmosphere, with nitrogen and oxygen reduced to the status of trace constituents.

On a warmer Earth, another source of additional atmosphere would be found in the sedimentary carbonate rocks of the crust. These minerals contain abundant carbon dioxide. If all these rocks were heated, they would release about 70 bars of CO_2, far more than the current CO_2 pressure of only 0.0005 bar. Thus, the atmosphere of a warm Earth would be dominated by water vapor and carbon dioxide, with a surface pressure nearing 400 bars.

Several lines of evidence show that the composition of Earth's atmosphere has changed over our planet's history. Scientists can infer the amount of atmospheric oxygen, for example, by studying the chemistry of minerals that formed at various times. We examine this issue in more detail later in this chapter.

Today we see that CO_2, H_2O, sulfur dioxide (SO_2), and other gases are released from deeper within Earth through the action of volcanoes. (For CO_2, the primary source today is the burning of fossil fuels, which releases far more CO_2 than that from volcanic eruptions.) Much of this apparently new gas, however, is recycled material that has been subducted through plate tectonics. But where did our planet's original atmosphere come from?

Three possibilities exist for the original source of Earth's atmosphere and oceans: (1) the atmosphere could have been formed with the rest of Earth as it accumulated from debris left over from the formation of the Sun; (2) it could have been released from the interior through volcanic activity, subsequent to the formation of Earth; or (3) it may have been derived from impacts by comets and asteroids from the outer parts of the solar system. Current evidence favors a combination of the interior and impact sources.

Weather and Climate

All planets with atmospheres have *weather*, which is the name we give to the circulation of the atmosphere. The energy that powers the weather is derived primarily from the sunlight that heats the surface. Both the rotation of the planet and slower seasonal changes cause variations in the amount of sunlight striking different parts of Earth. The atmosphere and oceans redistribute the heat from warmer to cooler areas. Weather on any planet represents the response of its atmosphere to changing inputs of energy from the Sun (see **Figure 19.9** for a dramatic example).

Figure 19.9 This satellite image shows Hurricane Irene in 2011, shortly before the storm hit land in New York City. The combination of Earth's tilted axis of rotation, moderately rapid rotation, and oceans of liquid water can lead to violent weather on our planet. (credit: NASA/NOAA GOES Project)

Climate is a term used to refer to the effects of the atmosphere that last through decades and centuries. Changes in climate (as opposed to the random variations in weather from one year to the next) are often difficult to detect over short time periods, but as they accumulate, their effect can be devastating. One saying is that “Climate is what you expect, and weather is what you get.” Modern farming is especially sensitive to temperature and rainfall; for example, calculations indicate that a drop of only 2 °C throughout the growing season would cut the wheat production by half in Canada and the United States. At the other extreme, an increase of 2 °C in the average temperature of Earth would be enough to melt many glaciers, including much of the ice cover of Greenland, raising sea level by as much as 10 meters, flooding many coastal cities and ports, and putting small islands completely under water.

The best documented changes in Earth’s climate are the great ice ages, which have lowered the temperature of the Northern Hemisphere periodically over the past half million years or so (**Figure 19.10**). The last ice age, which ended about 14,000 years ago, lasted some 20,000 years. At its height, the ice was almost 2 kilometers thick over Boston and stretched as far south as New York City.

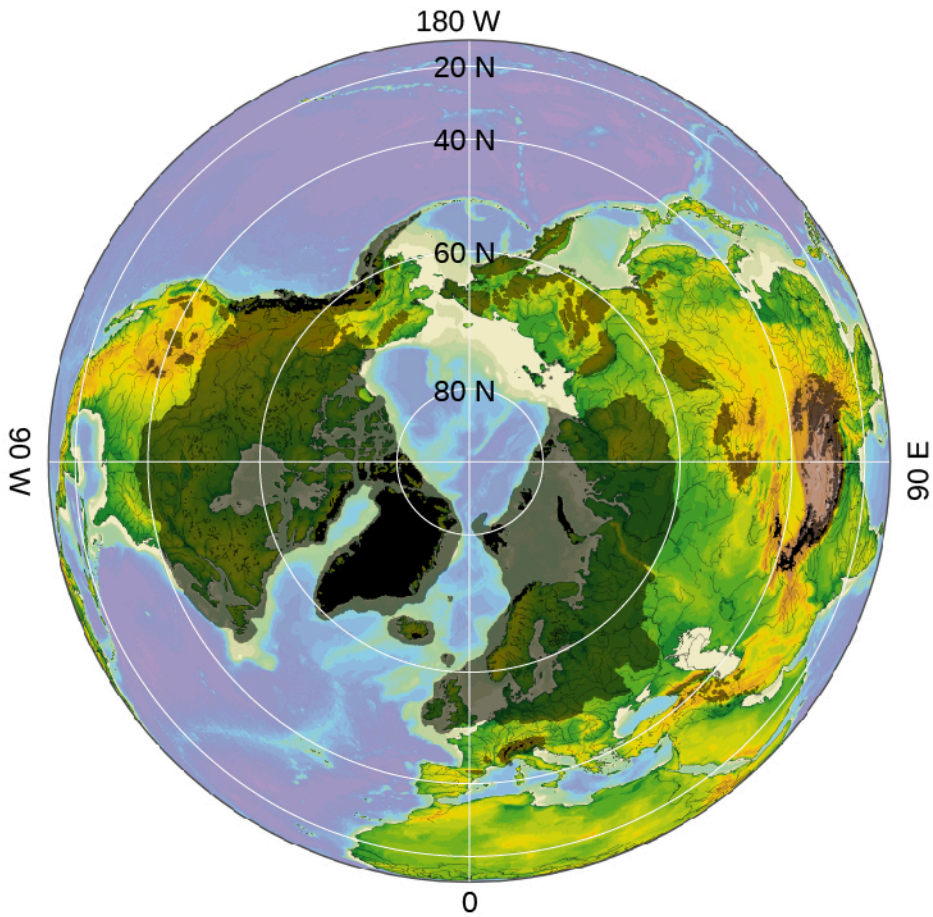


Figure 19.10 This computer-generated image shows the frozen areas of the Northern Hemisphere during past ice ages from the vantage point of looking down on the North Pole. The area in black indicates the most recent glaciation (coverage by glaciers), and the area in gray shows the maximum level of glaciation ever reached. (credit: modification of work by Hannes Grobe/AWI)

These ice ages were primarily the result of changes in the tilt of Earth’s rotational axis, produced by the gravitational effects of the other planets. We are less certain about evidence that at least once (and perhaps twice) about a billion years ago, the entire ocean froze over, a situation called *snowball Earth*.

The development and evolution of life on Earth has also produced changes in the composition and temperature of our planet’s atmosphere, as we shall see in the next section.

Watch this **short excerpt (https://openstax.org/l/30natgeoearth)** from the National Geographic documentary *Earth: The Biography*. In this segment, Dr. Iain Stewart explains the fluid nature of our atmosphere.

The Evolution of the Atmosphere

One of the key steps in the evolution of life on Earth was the development of blue-green algae, a very successful life-form that takes in carbon dioxide from the environment and releases oxygen as a waste product. These successful microorganisms proliferated, giving rise to all the lifeforms we call plants. Since the energy for making new plant material from chemical building blocks comes from sunlight, we call the process **photosynthesis**.

Studies of the chemistry of ancient rocks show that Earth's atmosphere lacked abundant free oxygen until about 2 billion years ago, despite the presence of plants releasing oxygen by photosynthesis. Apparently, chemical reactions with Earth's crust removed the oxygen gas as quickly as it formed. Slowly, however, the increasing evolutionary sophistication of life led to a growth in the plant population and thus increased oxygen production. At the same time, it appears that increased geological activity led to heavy erosion on our planet's surface. This buried much of the plant carbon before it could recombine with oxygen to form CO_2.

Free oxygen began accumulating in the atmosphere about 2 billion years ago, and the increased amount of this gas led to the formation of Earth's ozone layer (recall that ozone is a triple molecule of oxygen, O_3), which protects the surface from deadly solar ultraviolet light. Before that, it was unthinkable for life to venture outside the protective oceans, so the landmasses of Earth were barren.

The presence of oxygen, and hence ozone, thus allowed colonization of the land. It also made possible a tremendous proliferation of animals, which lived by taking in and using the organic materials produced by plants as their own energy source.

As animals evolved in an environment increasingly rich in oxygen, they were able to develop techniques for breathing oxygen directly from the atmosphere. We humans take it for granted that plenty of free oxygen is available in Earth's atmosphere, and we use it to release energy from the food we take in. Although it may seem funny to think of it this way, we are lifeforms that have evolved to breathe in the waste product of plants. It is plants and related microbes that are the primary producers, using sunlight to create energy-rich "food" for the rest of us.

On a planetary scale, one of the consequences of life has been a decrease in atmospheric carbon dioxide. In the absence of life, Earth would probably have an atmosphere dominated by CO_2, like Mars or Venus. But living things, in combination with high levels of geological activity, have effectively stripped our atmosphere of most of this gas.

The Greenhouse Effect and Global Warming

We have a special interest in the carbon dioxide content of the atmosphere because of the key role this gas plays in retaining heat from the Sun through a process called the **greenhouse effect**. To understand how the greenhouse effect works, consider the fate of sunlight that strikes the surface of Earth. The light penetrates our atmosphere, is absorbed by the ground, and heats the surface layers. At the temperature of Earth's surface, that energy is then reemitted as infrared or heat radiation (**Figure 19.11**). However, the molecules of our atmosphere, which allow visible light through, are good at absorbing infrared energy. As a result, CO_2 (along with methane and water vapor) acts like a blanket, trapping heat in the atmosphere and impeding its flow back to space. To maintain an energy balance, the temperature of the surface and lower atmosphere must increase until the total energy radiated by Earth to space equals the energy received from the Sun. The more CO_2 there is in our atmosphere, the higher the temperature at which Earth's surface reaches a new balance.

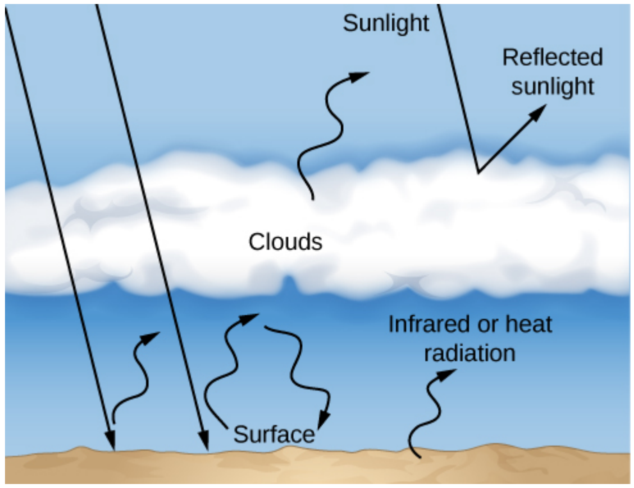


Figure 19.11 Sunlight that penetrates to Earth's lower atmosphere and surface is reradiated as infrared or heat radiation, which is trapped by greenhouse gases such as water vapor, methane, and CO_2 in the atmosphere. The result is a higher surface temperature for our planet.

The greenhouse effect in a planetary atmosphere is similar to the heating of a gardener's greenhouse or the inside of a car left out in the Sun with the windows rolled up. In these examples, the window glass plays the role of **greenhouse gases**, letting sunlight in but reducing the outward flow of heat radiation. As a result, a greenhouse or car interior winds up much hotter than would be expected from the heating of sunlight alone. On Earth, the current greenhouse effect elevates the surface temperature by about 23 °C. Without this greenhouse effect, the average surface temperature would be well below freezing and Earth would be locked in a global ice age.

That's the good news; the bad news is that the heating due to the greenhouse effect is increasing. Modern industrial society depends on energy extracted from burning fossil fuels. In effect, we are exploiting the energy-rich material created by photosynthesis tens of millions of years ago. As these ancient coal and oil deposits are oxidized (burned using oxygen), large quantities of carbon dioxide are released into the atmosphere. The problem is exacerbated by the widespread destruction of tropical forests, which we depend on to extract CO_2 from the atmosphere and replenish our supply of oxygen. In the past century of increased industrial and agricultural development, the amount of CO_2 in the atmosphere increased by about 30% and continues to rise at more than 0.5% per year.

Before the end of the present century, Earth's CO_2 level is predicted to reach twice the value it had before the industrial revolution (**Figure 19.12**). The consequences of such an increase for Earth's surface and atmosphere (and the creatures who live there) are likely to be complex changes in climate, and may be catastrophic for many species. Many groups of scientists are now studying the effects of such global warming with elaborate computer models, and climate change has emerged as the greatest known threat (barring nuclear war) to both industrial civilization and the ecology of our planet.

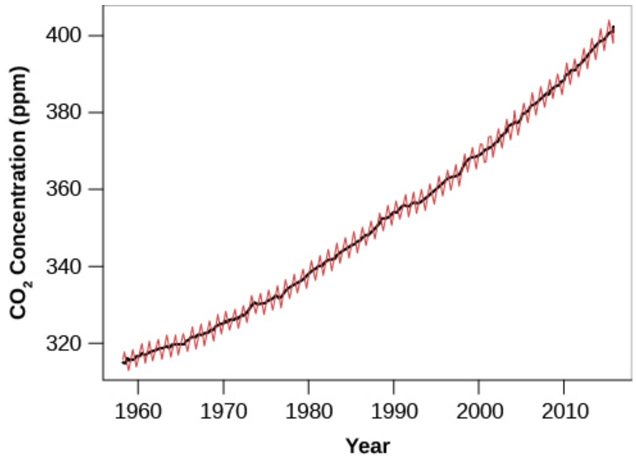


Figure 19.12 Scientists expect that the amount of CO_2 will double its preindustrial level before the end of the twenty-first century. Measurements of the isotopic signatures of this added CO_2 demonstrate that it is mostly coming from burning fossil fuels. (credit: modification of work by NOAA)

This **short PBS video (https://openstax.org/l/30pbsgreengas)** explains the physics of the greenhouse effect.

Already climate change is widely apparent. Around the world, temperature records are constantly set and broken; all but one of the hottest recorded years have taken place since 2000. Glaciers are retreating, and the Arctic Sea ice is now much thinner than when it was first explored with nuclear submarines in the 1950s. Rising sea levels (from both melting glaciers and expansion of the water as its temperature rises) pose one of the most immediate threats, and many coastal cities have plans to build dikes or seawalls to hold back the expected flooding. The rate of temperature increase is without historical precedent, and we are rapidly entering “unknown territory” where human activities are leading to the highest temperatures on Earth in more than 50 million years.

Human Impacts on Our Planet

Earth is so large and has been here for so long that some people have trouble accepting that humans are really changing the planet, its atmosphere, and its climate. They are surprised to learn, for example, that the carbon dioxide released from burning fossil fuels is 100 times greater than that emitted by volcanoes. But, the data clearly tell the story that our climate is changing rapidly, and that almost all of the change is a result of human activity.

This is not the first time that humans have altered our environment dramatically. Some of the greatest changes were caused by our ancestors, before the development of modern industrial society. If aliens had visited Earth 50,000 years ago, they would have seen much of the planet supporting large animals of the sort that now survive only in Africa. The plains of Australia were occupied by giant marsupials such as diprododon and zygomaturus (the size of our elephants today), and a species of kangaroo that stood 10 feet high. North America and North Asia hosted mammoths, saber tooth cats, mastodons, giant sloths, and even camels. The Islands of the Pacific teemed with large birds, and vast forests covered what are now the farms of Europe and China. Early human hunters killed many large mammals and marsupials, early farmers cut down most of the forests, and the Polynesian expansion across the Pacific doomed the population of large birds.

An even greater mass extinction is underway as a result of rapid climate change. In recognition of our impact on the environment, scientists have proposed giving a new name to the current epoch, the *anthropocine*, when human activity started to have a significant global impact. Although not an officially approved name, the concept of “anthropocine” is useful for recognizing that we humans now represent the dominant influence on our planet’s atmosphere and ecology, for better or for worse.

19.4 | The Massive Atmosphere of Venus

Learning Objectives

By the end of this section, you will be able to:

- Describe the general composition and structure of the atmosphere on Venus
- Explain how the greenhouse effect has led to high temperatures on Venus

The thick atmosphere of Venus produces the high surface temperature and shrouds the surface in a perpetual red twilight. Sunlight does not penetrate directly through the heavy clouds, but the surface is fairly well lit by diffuse light (about the same as the light on Earth under a heavy overcast). The weather at the bottom of this deep atmosphere remains perpetually hot and dry, with calm winds. Because of the heavy blanket of clouds and atmosphere, one spot on the surface of Venus is similar to any other as far as weather is concerned.

Composition and Structure of the Atmosphere

The most abundant gas on Venus is carbon dioxide (CO_2), which accounts for 96% of the atmosphere. The second most common gas is nitrogen. The predominance of carbon dioxide over nitrogen is not surprising when you recall that Earth's atmosphere would also be mostly carbon dioxide if this gas were not locked up in marine sediments (see the discussion of Earth's atmosphere in **Earth's Atmosphere**).

Table 19.2 compares the compositions of the atmospheres of Venus, Mars, and Earth. Expressed in this way, as percentages, the proportions of the major gases are very similar for Venus and Mars, but in total quantity, their atmospheres are dramatically different. With its surface pressure of 90 bars, the venusian atmosphere is more than 10,000 times more massive than its martian counterpart. Overall, the atmosphere of Venus is very dry; the absence of water is one of the important ways that Venus differs from Earth.

Atmospheric Composition of Earth, Venus, and Mars

Gas	Earth	Venus	Mars
Carbon dioxide (CO_2)	0.03%	96%	95.3%
Nitrogen (N_2)	78.1%	3.5%	2.7%
Argon (Ar)	0.93%	0.006%	1.6%
Oxygen (O_2)	21.0%	0.003%	0.15%
Neon (Ne)	0.002%	0.001%	0.0003%

Table 19.2

The atmosphere of Venus has a huge troposphere (region of convection) that extends up to at least 50 kilometers above the surface (**Figure 19.13**). Within the troposphere, the gas is heated from below and circulates slowly, rising near the equator and descending over the poles. Being at the base of the atmosphere of Venus is something like being a kilometer or more below the ocean surface on Earth. There, the mass of water evens out temperature variations and results in a uniform environment—the same effect the thick atmosphere has on Venus.

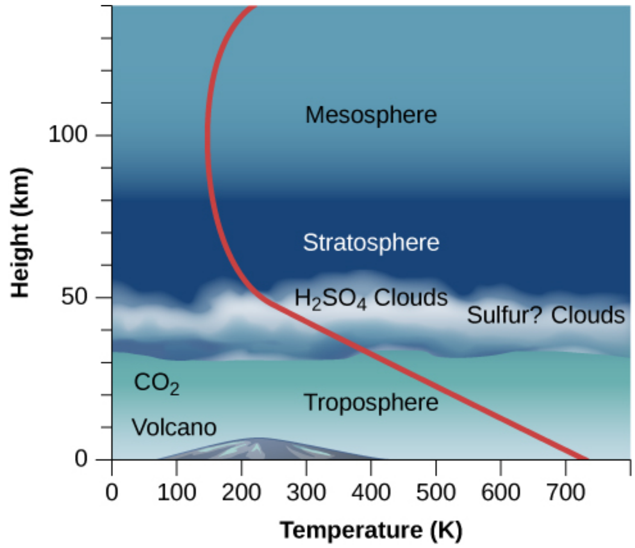


Figure 19.13 The layers of the massive atmosphere of Venus shown here are based on data from the Pioneer and Venera entry probes. Height is measured along the left axis, the bottom scale shows temperature, and the red line allows you to read off the temperature at each height. Notice how steeply the temperature rises below the clouds, thanks to the planet's huge greenhouse effect.

In the upper troposphere, between 30 and 60 kilometers above the surface, a thick cloud layer is composed primarily of sulfuric acid droplets. Sulfuric acid (H_2SO_4) is formed from the chemical combination of sulfur dioxide (SO_2) and water (H_2O). In the atmosphere of Earth, sulfur dioxide is one of the primary gases emitted by volcanoes, but it is quickly diluted and washed out by rainfall. In the dry atmosphere of Venus, this unpleasant substance is apparently stable. Below 30 kilometers, the Venus atmosphere is clear of clouds.

Surface Temperature on Venus

The high surface temperature of Venus was discovered by radio astronomers in the late 1950s and confirmed by the Mariner and Venera probes. How can our neighbor planet be so hot? Although Venus is somewhat closer to the Sun than is Earth, its surface is hundreds of degrees hotter than you would expect from the extra sunlight it receives. Scientists wondered what could be heating the surface of Venus to a temperature above 700 K. The answer turned out to be the *greenhouse effect*.

The greenhouse effect works on Venus just as it does on Earth, but since Venus has so much more CO_2—almost a million times more—the effect is much stronger. The thick CO_2 acts as a blanket, making it very difficult for the infrared (heat) radiation from the ground to get back into space. As a result, the surface heats up. The energy balance is only restored when the planet is radiating as much energy as it receives from the Sun, but this can happen only when the temperature of the lower atmosphere is very high. One way of thinking of greenhouse heating is that it must raise the surface temperature of Venus until this energy balance is achieved.

Has Venus always had such a massive atmosphere and high surface temperature, or might it have evolved to such conditions from a climate that was once more nearly earthlike? The answer to this question is of particular interest to us as we look at the increasing levels of CO_2 in Earth's atmosphere. As the greenhouse effect becomes stronger on Earth, are we in any danger of transforming our own planet into a hellish place like Venus?

Let us try to reconstruct the possible evolution of Venus from an earthlike beginning to its present state. Venus may once have had a climate similar to that of Earth, with moderate temperatures, water oceans, and much of its CO_2 dissolved in the ocean or chemically combined with the surface rocks. Then we allow for modest additional heating—by gradual increase in the energy output of the Sun, for example. When we calculate how Venus' atmosphere would respond to such effects, it turns out that even a small amount of extra heat can lead to increased evaporation of water from the oceans and the release of gas from surface rocks.

This in turn means a further increase in the atmospheric CO_2 and H_2O, gases that would amplify the greenhouse effect in Venus' atmosphere. That would lead to still more heat near Venus' surface and the release of further CO_2 and H_2O. Unless some other processes intervene, the temperature thus continues to rise. Such a situation is called the **runaway greenhouse**

effect.

We want to emphasize that the runaway greenhouse effect is not just a large greenhouse effect; it is an evolutionary *process*. The atmosphere evolves from having a small greenhouse effect, such as on Earth, to a situation where greenhouse warming is a major factor, as we see today on Venus. Once the large greenhouse conditions develop, the planet establishes a new, much hotter equilibrium near its surface.

Reversing the situation is difficult because of the role water plays. On Earth, most of the CO_2 is either chemically bound in the rocks of our crust or dissolved by the water in our oceans. As Venus got hotter and hotter, its oceans evaporated, eliminating that safety valve. But the water vapor in the planet's atmosphere will not last forever in the presence of ultraviolet light from the Sun. The light element hydrogen can escape from the atmosphere, leaving the oxygen behind to combine chemically with surface rock. The loss of water is therefore an irreversible process: once the water is gone, it cannot be restored. There is evidence that this is just what happened to the water once present on Venus.

We don't know if the same runaway greenhouse effect could one day happen on Earth. Although we are uncertain about the point at which a stable greenhouse effect breaks down and turns into a runaway greenhouse effect, Venus stands as clear testament to the fact that a planet cannot continue heating indefinitely without a major change in its oceans and atmosphere. It is a conclusion that we and our descendants will surely want to pay close attention to.

19.5 | Water and Life on Mars

Learning Objectives

By the end of this section, you will be able to:

- Describe the general composition of the atmosphere on Mars
- Explain what we know about the polar ice caps on Mars and how we know it
- Describe the evidence for the presence of water in the past history of Mars
- Summarize the evidence for and against the possibility of life on Mars

Of all the planets and moons in the solar system, Mars seems to be the most promising place to look for life, both fossil microbes and (we hope) some forms of life deeper underground that still survive today. But where (and how) should we look for life? We know that the one requirement shared by all life on Earth is liquid water. Therefore, the guiding principle in assessing habitability on Mars and elsewhere has been to "follow the water." That is the perspective we take in this section, to follow the water on the red planet and hope it will lead us to life.

Atmosphere and Clouds on Mars

The atmosphere of Mars today has an average surface pressure of only 0.007 bar, less than 1% that of Earth. (This is how thin the air is about 30 kilometers above Earth's surface.) Martian air is composed primarily of carbon dioxide (95%), with about 3% nitrogen and 2% argon. The proportions of different gases are similar to those in the atmosphere of Venus (see **Table 19.2**), but a lot less of each gas is found in the thin air on Mars.

While winds on Mars can reach high speeds, they exert much less force than wind of the same velocity would on Earth because the atmosphere is so thin. The wind is able, however, to loft very fine dust particles, which can sometimes develop planet-wide dust storms. It is this fine dust that coats almost all the surface, giving Mars its distinctive red color. In the absence of surface water, wind erosion plays a major role in sculpting the martian surface (**Figure 19.14**).

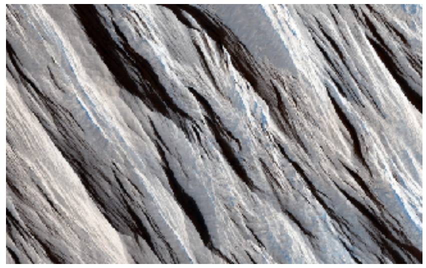


Figure 19.14 These long straight ridges, called yardangs, are aligned with the dominant wind direction. This is a high-resolution image from the *Mars Reconnaissance Orbiter* and is about 1 kilometer wide. (credit: NASA/JPL-Caltech/University of Arizona)

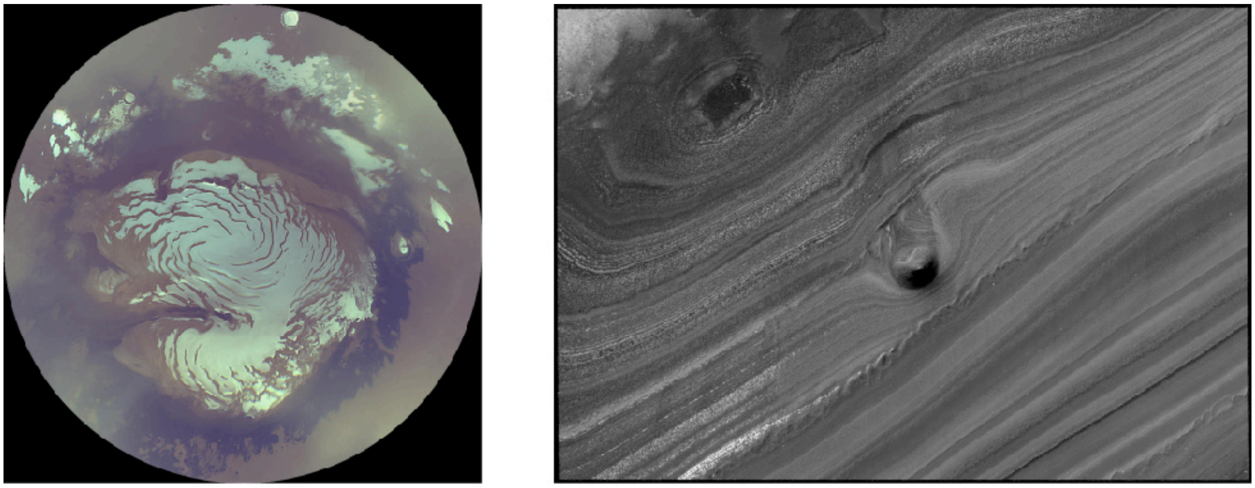
The issue of how strong the winds on Mars can be plays a big role in the **2015 hit movie The Martian (https://openstax.org/l/30TheMartian)** in which the main character is stranded on Mars after being buried in the sand in a windstorm so great that his fellow astronauts have to leave the planet so their ship is not damaged. Astronomers have noted that the martian winds could not possibly be as forceful as depicted in the film. In most ways, however, the depiction of Mars in this movie is remarkably accurate.

Although the atmosphere contains small amounts of water vapor and occasional clouds of water ice, liquid water is not stable under present conditions on Mars. Part of the problem is the low temperatures on the planet. But even if the temperature on a sunny summer day rises above the freezing point, the low pressure means that liquid water still cannot exist on the surface, except at the lowest elevations. At a pressure of less than 0.006 bar, the boiling point is as low or lower than the freezing point, and water changes directly from solid to vapor without an intermediate liquid state (as does “dry ice,” carbon dioxide, on Earth). However, salts dissolved in water lower its freezing point, as we know from the way salt is used to thaw roads after snow and ice forms during winter on Earth. Salty water is therefore sometimes able to exist in liquid form on the martian surface, under the right conditions.

Several types of clouds can form in the martian atmosphere. First there are dust clouds, discussed above. Second are water-ice clouds similar to those on Earth. These often form around mountains, just as happens on our planet. Finally, the CO_2 of the atmosphere can itself condense at high altitudes to form hazes of dry ice crystals. The CO_2 clouds have no counterpart on Earth, since on our planet temperatures never drop low enough (down to about 150 K or about –125 °C) for this gas to condense.

The Polar Caps

Through a telescope, the most prominent surface features on Mars are the bright polar caps, which change with the seasons, similar to the seasonal snow cover on Earth. We do not usually think of the winter snow in northern latitudes as a part of our polar caps, but seen from space, the thin winter snow merges with Earth’s thick, permanent ice caps to create an impression much like that seen on Mars (**Figure 19.15**).



(a) (b)

Figure 19.15 (a) This is a composite image of the north pole in summer, obtained in October 2006 by the *Mars Reconnaissance Orbiter*. It shows the mostly water-ice residual cap sitting atop light, tan-colored, layered sediments. Note that although the border of this photo is circular, it shows only a small part of the planet. (b) Here we see a small section of the layered terrain near the martian north pole. There is a mound about 40 meters high that is sticking out of a trough in the center of the picture. (credit a: modification of work by NASA/JPL/MSSS; credit b: modification of work by NASA/JPL-Caltech/ University of Arizona)

The *seasonal caps* on Mars are composed not of ordinary snow but of frozen CO_2 (dry ice). These deposits condense directly from the atmosphere when the surface temperature drops below about 150 K. The caps develop during the cold martian winters and extend down to about 50° latitude by the start of spring.

Quite distinct from these thin seasonal caps of CO_2 are the *permanent* or *residual caps* that are always present near the poles. The southern permanent cap has a diameter of 350 kilometers and is composed of frozen CO_2 deposits together with a great deal of water ice. Throughout the southern summer, it remains at the freezing point of CO_2, 150 K, and this cold reservoir is thick enough to survive the summer heat intact.

The northern permanent cap is different. It is much larger, never shrinking to a diameter less than 1000 kilometers, and is composed of water ice. Summer temperatures in the north are too high for the frozen CO_2 to be retained. Measurements from the *Mars Global Surveyor* have established the exact elevations in the north polar region of Mars, showing that it is a large basin about the size of our own Arctic Ocean basin. The ice cap itself is about 3 kilometers thick, with a total volume of about 10 million km^3 (similar to that of Earth's Mediterranean Sea). If Mars ever had extensive liquid water, this north polar basin would have contained a shallow sea. There is some indication of ancient shorelines visible, but better images will be required to verify this suggestion.

Images taken from orbit also show a distinctive type of terrain surrounding the permanent polar caps, as shown in **Figure 19.15**. At latitudes above 80° in both hemispheres, the surface consists of recent layered deposits that cover the older cratered ground below. Individual layers are typically ten to a few tens of meters thick, marked by alternating light and dark bands of sediment. Probably the material in the polar deposits includes dust carried by wind from the equatorial regions of Mars.

What do these terraced layers tell us about Mars? Some cyclic process is depositing dust and ice over periods of time. The time scales represented by the polar layers are tens of thousands of years. Apparently the martian climate experiences periodic changes at intervals similar to those between ice ages on Earth. Calculations indicate that the causes are probably also similar: the gravitational pull of the other planets produces variations in Mars' orbit and tilt as the great clockwork of the solar system goes through its paces.

The *Phoenix* spacecraft landed near the north polar cap in summer (**Figure 19.16**). Controllers knew that it would not be able to survive a polar winter, but directly measuring the characteristics of the polar region was deemed important enough to send a dedicated mission. The most exciting discovery came when the spacecraft tried to dig a shallow trench under the spacecraft. When the overlying dust was stripped off, they saw bright white material, apparently some kind of ice. From the way this ice sublimated over the next few days, it was clear that it was frozen water.

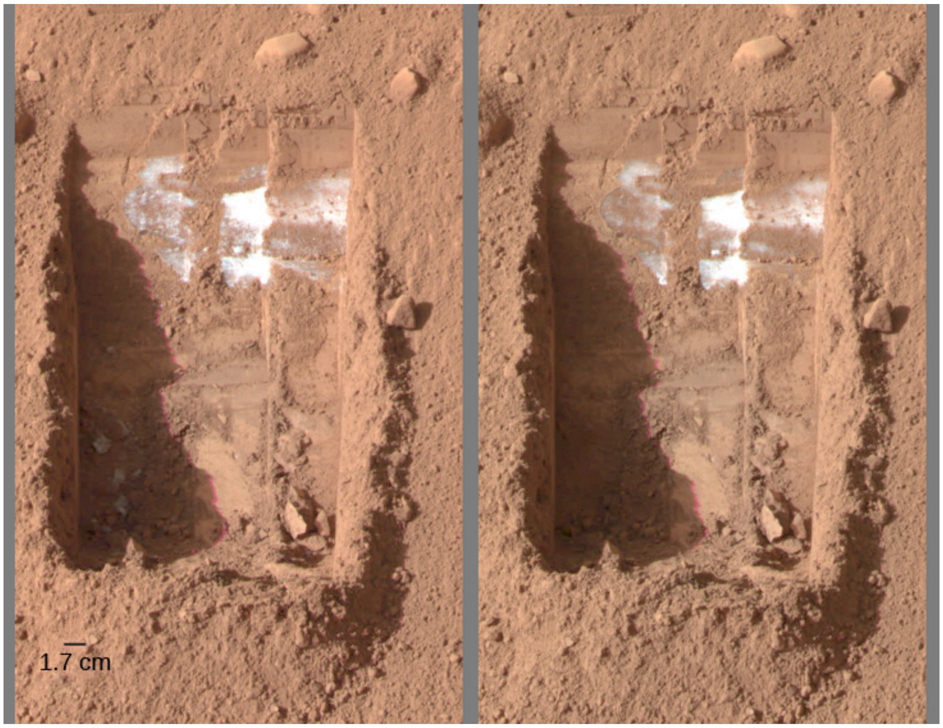


Figure 19.16 We see a trench dug by the *Phoenix* lander in the north polar region four martian days apart in June 2008. If you look at the shadowed region in the bottom left of the trench, you can see three spots of ice in the left image which have sublimated away in the right image. (credit: modification of work by NASA/JPL-Caltech/University of Arizona/Texas A&M University)

Example 19.2

Comparing the Amount of Water on Mars and Earth

It is interesting to estimate the amount of water (in the form of ice) on Mars and to compare this with the amount of water on Earth. In each case, we can find the total volume of a layer on a sphere by multiplying the area of the sphere ($4\pi R^2$) by the thickness of the layer. For Earth, the ocean water is equivalent to a layer 3 km thick spread over the entire planet, and the radius of Earth is 6.378×10^6 m (see **Appendix D**). For Mars, most of the water we are sure of is in the form of ice near the poles. We can calculate the amount of ice in one of the residual polar caps if it is (for example) 2 km thick and has a radius of 400 km (the area of a circle is πR^2).

Solution

The volume of Earth's water is therefore the area $4\pi R^2$

$$4\pi\left(6.378 \times 10^6 \text{ m}\right)^2 = 5.1 \times 10^{14} \text{ m}^2$$

multiplied by the thickness of 3000 m:

$$5.1 \times 10^{14} \text{ m}^2 \times 3000 \text{ m} = 1.5 \times 10^{18} \text{ m}^3$$

This gives 1.5×10^{18} m^3 of water. Since water has a density of 1 ton per cubic meter (1000 kg/m^3), we can calculate the mass:

$$1.5 \times 10^{18} \text{ m}^3 \times 1 \text{ ton/m}^3 = 1.5 \times 10^{18} \text{ tons}$$

For Mars, the ice doesn't cover the whole planet, only the caps; the polar cap area is

$$\pi R^2 = \pi(4 \times 10^5 \text{ m})^2 = 5 \times 10^{11} \text{ m}^2$$

(Note that we converted kilometers to meters.)

The volume = area × height, so we have:

$$(2 \times 10^3 \text{ m})(5 \times 10^{11} \text{ m}^2) = 1 \times 10^{15} \text{ m}^3 = 10^{15} \text{ m}^3$$

Therefore, the mass is:

$$10^{15} \text{ m}^3 \times 1 \text{ ton/m}^3 = 10^{15} \text{ tons}$$

This is about 0.1% that of Earth’s oceans.

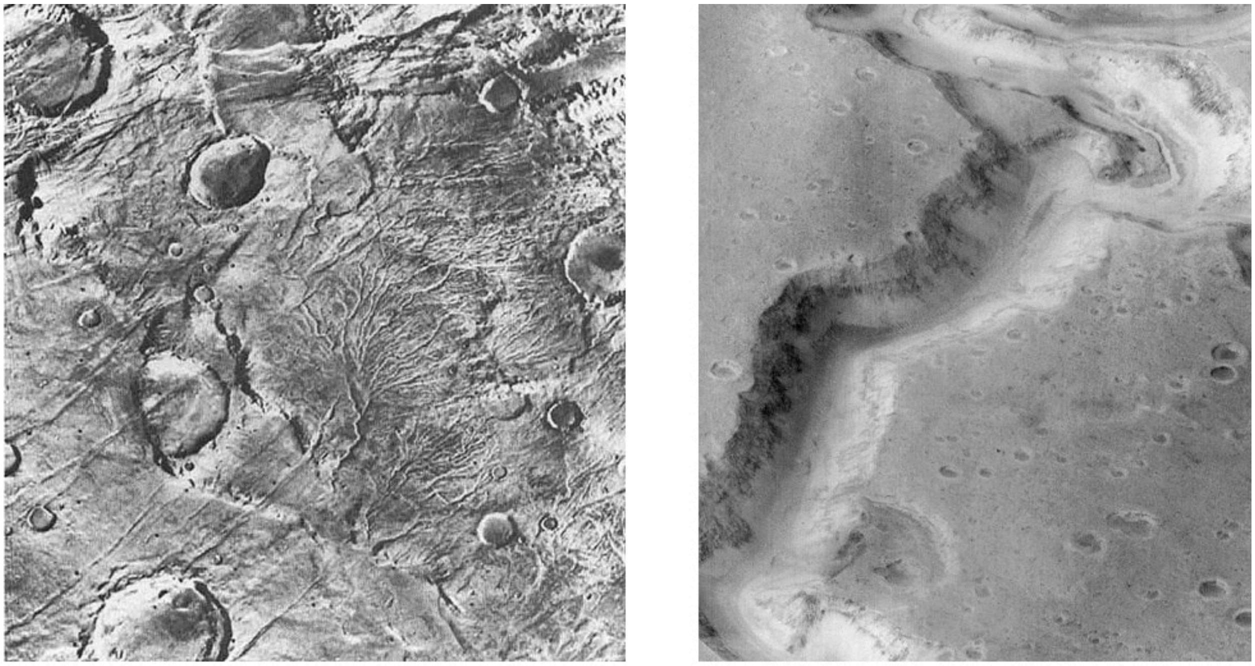
19.1 A better comparison might be to compare the amount of ice in the Mars polar ice caps to the amount of ice in the Greenland ice sheet on Earth, which has been estimated as 2.85×10^{15} m^3. How does this compare with the ice on Mars?

Channels and Gullies on Mars

Although no bodies of liquid water exist on Mars today, evidence has accumulated that rivers flowed on the red planet long ago. Two kinds of geological features appear to be remnants of ancient watercourses, while a third class—smaller gullies—suggests intermittent outbreaks of liquid water even today. We will examine each of these features in turn.

In the highland equatorial plains, there are multitudes of small, sinuous (twisting) channels—typically a few meters deep, some tens of meters wide, and perhaps 10 or 20 kilometers long (**Figure 19.17**). They are called runoff channels because they look like what geologists would expect from the surface runoff of ancient rain storms. These runoff channels seem to be telling us that the planet had a very different climate long ago. To estimate the age of these channels, we look at the cratering record. Crater counts show that this part of the planet is more cratered than the lunar maria but less cratered than the lunar highlands. Thus, the runoff channels are probably older than the lunar maria, presumably about 4 billion years old.

The second set of water-related features we see are *outflow channels* (**Figure 19.17**) are much larger than the runoff channels. The largest of these, which drain into the Chryse basin where Pathfinder landed, are 10 kilometers or more wide and hundreds of kilometers long. Many features of these outflow channels have convinced geologists that they were carved by huge volumes of running water, far too great to be produced by ordinary rainfall. Where could such floodwater have come from on Mars?



(a) (b)

Figure 19.17 (a) These runoff channels in the old martian highlands are interpreted as the valleys of ancient rivers fed by either rain or underground springs. The width of this image is about 200 kilometers. (b) This intriguing channel, called Nanedi Valles, resembles Earth riverbeds in some (but not all) ways. The tight curves and terraces seen in the channel certainly suggest the sustained flow of a fluid like water. The channel is about 2.5 kilometers across. (credit a: modification of work by Jim Secosky/NASA; credit b: modification of work by Jim Secosky/NASA)

As far we can tell, the regions where the outflow channels originate contained abundant water frozen in the soil as permafrost. Some local source of heating must have released this water, leading to a period of rapid and catastrophic flooding. Perhaps this heating was associated with the formation of the volcanic plains on Mars, which date back to roughly the same time as the outflow channels.

Note that neither the runoff channels nor the outflow channels are wide enough to be visible from Earth, nor do they follow straight lines. They could not have been the “canals” Percival Lowell imagined seeing on the red planet.

The third type of water feature, the smaller *gullies*, was discovered by the *Mars Global Surveyor* (**Figure 19.18**). The *Mars Global Surveyor's* camera images achieved a resolution of a few meters, good enough to see something as small as a truck or bus on the surface. On the steep walls of valleys and craters at high latitudes, there are many erosional features that look like gullies carved by flowing water. These gullies are very young: not only are there no superimposed impact craters, but in some instances, the gullies seem to cut across recent wind-deposited dunes. Perhaps there is liquid water underground that can occasionally break out to produce short-lived surface flows before the water can freeze or evaporate.

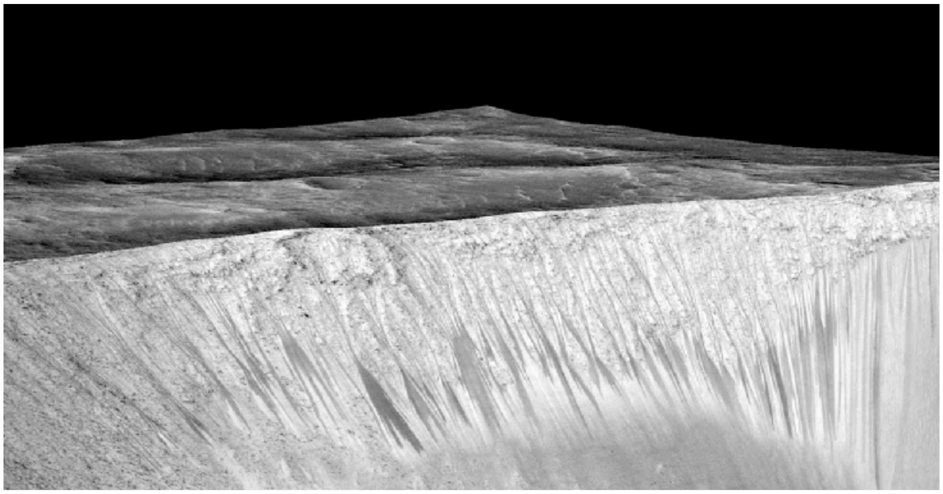


Figure 19.18 This high-resolution image is from the *Mars Reconnaissance Orbiter*. The dark streaks, which are each several hundred meters long, change in a seasonal pattern that suggests they are caused by the temporary flow of surface water. (credit: NASA/JPL-Caltech/University of Arizona)

The gullies also have the remarkable property of changing regularly with the martian seasons. Many of the dark streaks (visible in **Figure 19.18**) elongate within a period of a few days, indicating that something is flowing downhill—either water or dark sediment. If it is water, it requires a continuing source, either from the atmosphere or from springs that tap underground water layers (aquifers.) Underground water would be the most exciting possibility, but this explanation seems inconsistent with the fact that many of the dark streaks start at high elevations on the walls of craters.

Additional evidence that the dark streaks (called by the scientists *recurring slope lineae*) are caused by water was found in 2015 when spectra were obtained of the dark streaks (**Figure 19.19**). These showed the presence of hydrated salts produced by the evaporation of salty water. If the water is salty, it could remain liquid long enough to flow downstream for distances of a hundred meters or more, before it either evaporates or soaks into the ground. However, this discovery still does not identify the ultimate source of the water.

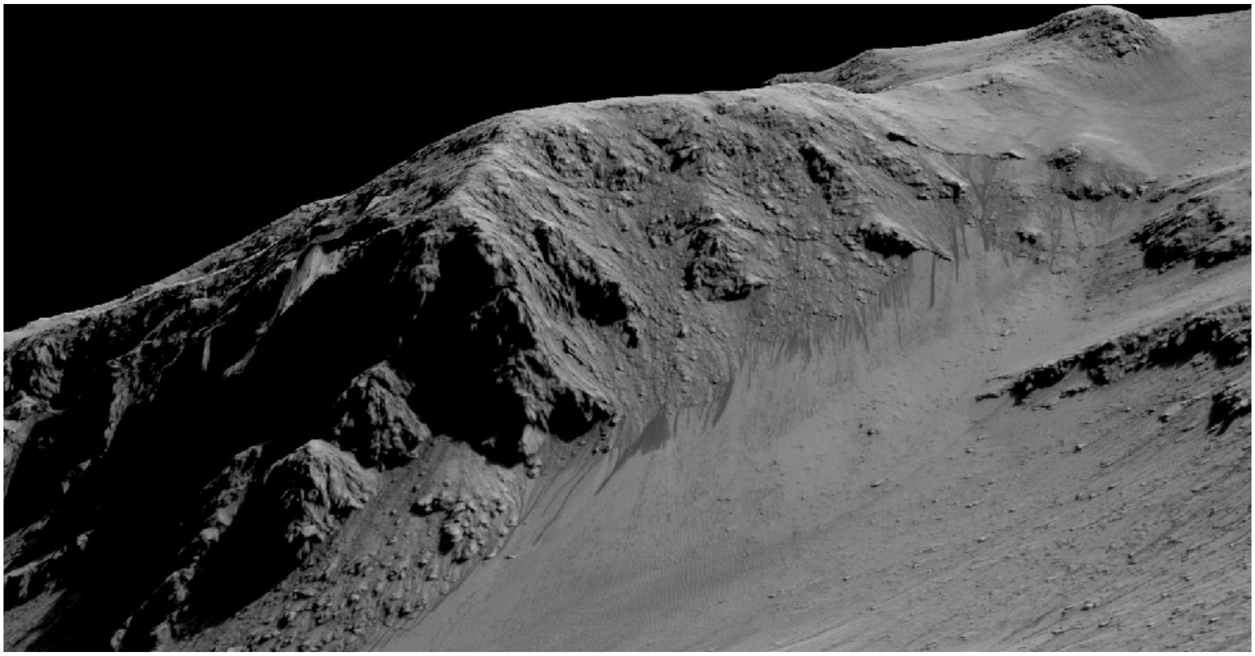


Figure 19.19 The dark streaks in Horowitz crater, which move downslope, have been called recurring slope lineae. The streaks in the center of the image go down the wall of the crater for about a distance of 100 meters. Spectra taken of this region indicate that these are locations where salty liquid water flows on or just below the surface of Mars. (The vertical dimension is exaggerated by a factor of 1.5 compared to horizontal dimensions.) (credit: NASA/JPL-Caltech/University of Arizona)

Ancient Lakes and Glaciers

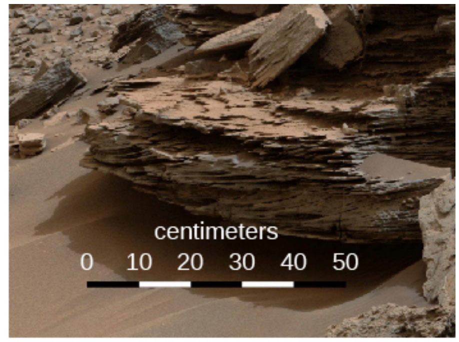
The rovers (*Spirit*, *Opportunity*, and *Curiosity*) that have operated on the surface of Mars have been used to hunt for additional evidence of water. They could not reach the most interesting sites, such as the gullies, which are located on steep slopes. Instead, they explored sites that might be dried-out lake beds, dating back to a time when the climate on Mars was warmer and the atmosphere thicker—allowing water to be liquid on the surface.

Spirit was specifically targeted to explore what looked like an ancient lake-bed in Gusev crater, with an outflow channel emptying into it. However, when the spacecraft landed, it found that the former lakebed had been covered by thin lava flows, blocking the rover from access to the sedimentary rocks it had hoped to find. However, *Opportunity* had better luck. Peering at the walls of a small crater, it detected layered sedimentary rock. These rocks contained chemical evidence of evaporation, suggesting there had been a shallow salty lake in that location. In these sedimentary rocks were also small spheres that were rich in the mineral hematite, which forms only in watery environments. Apparently this very large basin had once been underwater.

The small spherical rocks were nicknamed "blueberries" by the science team and the discovery of a whole "berry-bowl" of them was announced in this **interesting news release (https://openstax.org/l/30berrybowl)** from NASA.

The *Curiosity* rover landed inside Gale crater, where photos taken from orbit also suggested past water erosion. It discovered numerous sedimentary rocks, some in the form of mudstones from an ancient lakebed; it also found indications of rocks formed by the action of shallow water at the time the sediment formed (**Figure 19.20**).

Even today there is evidence of large quantities of ice just below the surface of Mars. In the mid-latitudes, high-resolution photos from orbit have revealed glaciers covered with dirt and dust. In some cliffs, the ice is observed directly (see **Figure 19.20**). These glaciers are thought to have formed during warm periods, when the atmospheric pressure was greater and snow and ice could precipitate. They also suggest readily available frozen water that could support future human exploration of the planet.



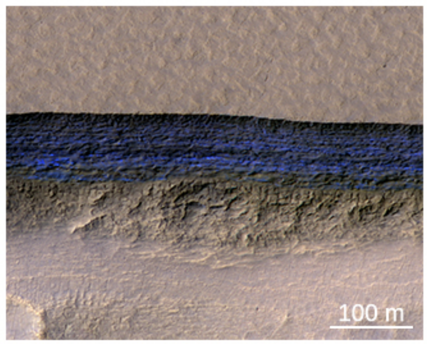


Figure 19.20 (a) This scene, photographed by the *Curiosity* rover, shows an ancient lakebed of cracked mudstones. (b) Geologists working with the *Curiosity* rover interpret this image of cross-bedded sandstone in Gale crater as evidence of liquid water passing over a loose bed of sediment at the time this rock formed. (c) Ice bands a hundred meters tall are visible in blue in a cliff-face on Mars, suggesting large deposits of frozen water buried just a few meters below the surface. Note that the blue color has been exaggerated in this photo, taken by the Mars Reconnaissance Orbiter spacecraft. (credit a: modification of work by NASA/JPL-Caltech/MSSS; credit b: modification of work by NASA/JPL-Caltech/MSSS; credit c: modification of work by NASA/JPL-Caltech/UA/USGS)

Astronomy and Pseudoscience: The "Face on Mars"

People like human faces. We humans have developed great skill in recognizing people and interpreting facial expressions. We also have a tendency to see faces in many natural formations, from clouds to the man in the Moon. One of the curiosities that emerged from the Viking orbiters' global mapping of Mars was the discovery of a strangely shaped mesa in the Cydonia region that resembled a human face. Despite later rumors of a cover-up, the "Face on Mars" was, in fact, recognized by Viking scientists and included in one of the early mission press releases. At the low resolution and oblique lighting under which the Viking image was obtained, the mile-wide mesa had something of a Sphinx-like appearance.

Unfortunately, a small band of individuals decided that this formation was an artificial, carved sculpture of a human face placed on Mars by an ancient civilization that thrived there hundreds of thousands of years ago. A band of "true believers" grew around the face and tried to deduce the nature of the "sculptors" who made it. This group also linked the face to a variety of other pseudoscientific phenomena such as crop circles (patterns in fields of grain, mostly in Britain, now known to be the work of pranksters).

Members of this group accused NASA of covering up evidence of intelligent life on Mars, and they received a great deal of help in publicizing their perspective from tabloid media. Some of the believers picketed the Jet Propulsion Laboratory at the time of the failure of the *Mars Observer* spacecraft, circulating stories that the "failure" of the *Mars Observer* was itself a fake, and that its true (secret) mission was to photograph the face.

The high-resolution *Mars Observer* camera (MOC) was reflown on the *Mars Global Surveyor* mission, which arrived at Mars in 1997. On April 5, 1998, in Orbit 220, the MOC obtained an oblique image of the face at a resolution of 4 meters per pixel, a factor-of-10 improvement in resolution over the Viking image. Another image in 2001 had even higher resolution. Immediately released by NASA, the new images showed a low mesa-like hill cut crossways by several roughly linear ridges and depressions, which were misidentified in the 1976 photo as the eyes and mouth of a face. Only with an enormous dose of imagination can any resemblance to a face be seen in the new images, demonstrating how dramatically our interpretation of geology can change with large improvements in resolution. The original and the higher resolution images can be seen in **Figure 19.21**.

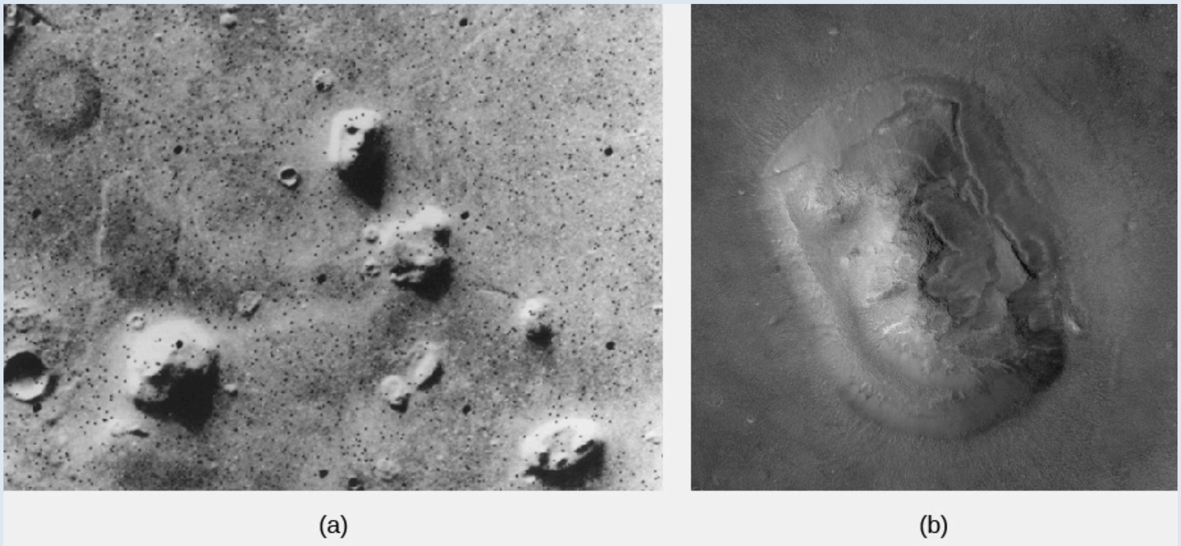


Figure 19.21 The so-called "Face on Mars" is seen (a) in low resolution from Viking (the "face" is in the upper part of the picture) and (b) with 20 times better resolution from the *Mars Global Surveyor*. (credit a: modification of work NASA/JPL; credit b: modification of work by NASA/JPL/MSSS)

After 20 years of promoting pseudoscientific interpretations and various conspiracy theories, can the "Face on Mars" believers now accept reality? Unfortunately, it does not seem so. They have accused NASA of faking the new picture. They also suggest that the secret mission of the *Mars Observer* included a nuclear bomb used to destroy the face before it could be photographed in greater detail by the *Mars Global Surveyor*.

Space scientists find these suggestions incredible. NASA is spending increasing sums for research on life in the universe, and a major objective of current and upcoming Mars missions is to search for evidence of past microbial life on Mars. Conclusive evidence of extraterrestrial life would be one of the great discoveries of science and incidentally might well lead to increased funding for NASA. The idea that NASA or other government agencies would (or could) mount a conspiracy to suppress such welcome evidence is truly bizarre.

Alas, the "Face on Mars" story is only one example of a whole series of conspiracy theories that are kept before the public by dedicated believers, by people out to make a fast buck, and by irresponsible media attention. Others include the "urban legend" that the Air Force has the bodies of extraterrestrials at a secret base, the widely circulated report that UFOs crashed near Roswell, New Mexico (actually it was a balloon carrying scientific instruments to find evidence of Soviet nuclear tests), or the notion that alien astronauts helped build the Egyptian pyramids and many other ancient monuments because our ancestors were too stupid to do it alone.

In response to the increase in publicity given to these "fiction science" ideas, a group of scientists, educators, scholars, and magicians (who know a good hoax when they see one) have formed the Committee for Skeptical Inquiry. Two of the original authors of your book are active on the committee. For more information about its work delving into the rational explanations for paranormal claims, see their excellent magazine, *The Skeptical Inquirer*, or check out their website at www.csicop.org/.

Climate Change on Mars

The evidence about ancient rivers and lakes of water on Mars discussed so far suggests that, billions of years ago, martian temperatures must have been warmer and the atmosphere must have been more substantial than it is today. But what could have changed the climate on Mars so dramatically?

We presume that, like Earth and Venus, Mars probably formed with a higher surface temperature thanks to the greenhouse effect. But Mars is a smaller planet, and its lower gravity means that atmospheric gases could escape more easily than from Earth and Venus. As more and more of the atmosphere escaped into space, the temperature on the surface gradually fell.

Eventually Mars became so cold that most of the water froze out of the atmosphere, further reducing its ability to retain heat. The planet experienced a sort of *runaway refrigerator effect*, just the opposite of the runaway greenhouse effect that occurred on Venus. Probably, this loss of atmosphere took place within less than a billion years after Mars formed. The result is the cold, dry Mars we see today.

Conditions a few meters below the martian surface, however, may be much different. There, liquid water (especially salty water) might persist, kept warm by the internal heat of Mars or the insulating layers solid and rock. Even on the surface, there may be ways to change the martian atmosphere temporarily.

Mars is likely to experience long-term climate cycles, which may be caused by the changing orbit and tilt of the planet. At times, one or both of the polar caps might melt, releasing a great deal of water vapor into the atmosphere. Perhaps an occasional impact by a comet might produce a temporary atmosphere that is thick enough to permit liquid water on the surface for a few weeks or months. Some have even suggested that future technology might allow us to *terraform* Mars—that is, to engineer its atmosphere and climate in ways that might make the planet more hospitable for long-term human habitation.

The Search for Life on Mars

If there was running water on Mars in the past, perhaps there was life as well. Could life, in some form, remain in the martian soil today? Testing this possibility, however unlikely, was one of the primary objectives of the Viking landers in 1976. These landers carried miniature biological laboratories to test for microorganisms in the martian soil. Martian soil was scooped up by the spacecraft's long arm and placed into the experimental chambers, where it was isolated and incubated in contact with a variety of gases, radioactive isotopes, and nutrients to see what would happen. The experiments looked for evidence of *respiration* by living animals, *absorption* of *nutrients* offered to organisms that might be present, and an *exchange of gases* between the soil and its surroundings for any reason whatsoever. A fourth instrument pulverized the soil and analyzed it carefully to determine what organic (carbon-bearing) material it contained.

The Viking experiments were so sensitive that, had one of the spacecraft landed anywhere on Earth (with the possible exception of Antarctica), it would easily have detected life. But, to the disappointment of many scientists and members of the public, no life was detected on Mars. The soil tests for absorption of nutrients and gas exchange did show some activity, but this was most likely caused by chemical reactions that began as water was added to the soil and had nothing to do with life. In fact, these experiments showed that martian soil seems much more chemically active than terrestrial soils because of its exposure to solar ultraviolet radiation (since Mars has no ozone layer).

The organic chemistry experiment showed no trace of organic material, which is apparently destroyed on the martian surface by the sterilizing effect of this ultraviolet light. While the possibility of life on the surface has not been eliminated, most experts consider it negligible. Although Mars has the most earthlike environment of any planet in the solar system, the sad fact is that nobody seems to be home today, at least on the surface.

However, there is no reason to think that life could not have begun on Mars about 4 billion years ago, at the same time it started on Earth. The two planets had very similar surface conditions then. Thus, the attention of scientists has shifted to the search for *fossil* life on Mars. One of the primary questions to be addressed by future spacecraft is whether Mars once supported its own life forms and, if so, how this martian life compared with that on our own planet. Future missions will include the return of martian samples selected from sedimentary rocks at sites that once held water and thus perhaps ancient life. The most powerful searches for martian life (past or present) will thus be carried out in our laboratories here on Earth.

Planetary Protection

When scientists begin to search for life on another planet, they must make sure that we do not contaminate the other world with life carried from Earth. At the very beginning of spacecraft exploration on Mars, an international agreement specified that all landers were to be carefully sterilized to avoid accidentally transplanting terrestrial microbes to Mars. In the case of Viking, we know the sterilization was successful. Viking's failure to detect martian organisms also implies that these experiments did not detect hitchhiking terrestrial microbes.

As we have learned more about the harsh conditions on the martian surface, the sterilization requirements have been somewhat relaxed. It is evident that no terrestrial microbes could grow on the martian surface, with its low temperature, absence of water, and intense ultraviolet radiation. Microbes from Earth might survive in a dormant, dried state, but they cannot grow and proliferate on Mars.

The problem of contaminating Mars will become more serious, however, as we begin to search for life below the surface, where temperatures are higher and no ultraviolet light penetrates. The situation will be even more daunting if we consider human flights to Mars. Any humans will carry with them a multitude of terrestrial microbes of all kinds, and it is hard to imagine how we can effectively keep the two biospheres isolated from each other if Mars has indigenous life. Perhaps the best situation could be one in which the two life-forms are so different that each is effectively invisible to the other—not recognized on a chemical level as living or as potential food.

The most immediate issue of public concern is not with the contamination of Mars but with any dangers associated with returning Mars samples to Earth. NASA is committed to the complete biological isolation of returned samples until they are demonstrated to be safe. Even though the chances of contamination are extremely low, it is better to be safe than sorry.

Most likely there is no danger, even if there is life on Mars and alien microbes hitch a ride to Earth inside some of the returned samples. In fact, Mars is sending samples to Earth all the time in the form of the Mars meteorites. Since some of these microbes (if they exist) could probably survive the trip to Earth inside their rocky home, we may have been exposed many times over to martian microbes. Either they do not interact with our terrestrial life, or in effect our planet has already been inoculated against such alien bugs.

More than any other planet, Mars has inspired science fiction writers over the years. You can find scientifically reasonable stories about Mars in a subject index of such stories online. If you click on **Mars (https://openstax.org/l/30MarsStories)** as a topic, you will find stories by a number of space scientists, including William Hartmann, Geoffrey Landis, and Ludek Pesek.

19.6 | Divergent Planetary Evolution

Learning Objectives

By the end of this section, you will be able to:

- Compare the planetary evolution of Venus, Earth, and Mars

As we have seen, Venus, Mars, and our own planet Earth form a remarkably diverse triad of worlds. Although all three orbit in roughly the same inner zone around the Sun and all apparently started with about the same chemical mix of silicates and metals, their evolutionary paths have diverged. As a result, Venus became hot and dry, Mars became cold and dry, and only Earth ended up with what we consider a hospitable climate.

Planetary Cooling

Because of their differing sizes, even if all of the terrestrial planets began their existence as similar balls of molten silicates and metals, they cooled at different rates. Suppose that a hot sphere of radius **R** begins to radiate heat into outer space. The total amount of (hot) mass initially present inside the planet is proportional to the volume of the sphere, i.e.

$$m \propto \frac{4\pi R^3}{3} \tag{19.4}$$

Now, the only heat transfer method available to cool this planet is radiation (see **Mechanisms of Heat Transfer**). The net power, or rate of heat transfer, is proportional to the surface area of the sphere:

$$P = \frac{\mathrm{dQ}}{\mathrm{dt}} \propto A = 4\pi R^2 \tag{19.5}$$

The rate of cooling

$$\frac{\mathrm{dT}}{\mathrm{dt}} = \frac{\frac{\mathrm{dQ}}{\mathrm{dt}}}{\frac{\mathrm{dQ}}{\mathrm{dT}}} = \frac{P}{mc} \tag{19.6}$$

where **c** is the specific heat of the planetary material.

The cooling rate is thus proportional to the ratio of the surface area to the volume:

$$\frac{\mathrm{dT}}{\mathrm{dt}} = \frac{P}{mc} \propto \frac{4\pi R^2}{\frac{4}{3}\pi R^3} \propto \frac{1}{R} \qquad (19.7)$$

Thus, we discover that the smaller the radius of the planet, the faster it cooled off.

Which of the terrestrial objects had the smallest radii? Mercury and the Moon, substantially smaller than Venus or Earth, cooled off very fast. They are currently cold, dead worlds. Mars, slightly smaller than Venus and Earth, also cooled faster than the larger two planets did.

Thermal Escape

One thing that led to the very different atmospheric compositions for the terrestrial planets was the **thermal escape** of certain species of gas molecules. Recall from the discussion of the **Pressure, Temperature and rms Speed** that the temperature of a gas is actually a measure of the average kinetic energy of its molecules. From that idea, we saw that the rms speed of a gas molecule is inversely proportional to the square root of its mass:

$$v_{\text{rms}} = \sqrt{\overline{v^2}} = \sqrt{\frac{3k_{\text{B}}T}{m}}. \qquad (19.8)$$

And, if you will recall from our discussion of **Energy Conservation and Universal Gravitation**, the escape velocity from a planet of mass M and radius R is:

$$v_{\text{esc}} = \sqrt{\frac{2GM}{R}}. \qquad (19.9)$$

Now, it should be obvious that, if the upward velocity of an individual gas molecule exceeds the escape velocity, that molecule will escape from the planet into outer space. But the situation is more complicated because, as we know, at any temperature **T** the gas molecules have various speeds characterized by the **Maxwell-Boltzmann distribution**.

Further complicating matters is the fact that each planet is in a thermal equilibrium, based upon its distance from the Sun and its reflectivity.

Let's examine the process of thermal escape carefully. Suppose a particular gas molecule is in the high-velocity part of the Maxwell-Boltzmann distribution, and so it escapes the planet's atmosphere. Since the atmosphere just lost one of its most energetic molecules, the remaining molecules will have an average velocity (and temperature) that is slightly lower.

However, because the Sun continues to provide incoming energy to maintain the thermal equilibrium of the planet, very quickly the remaining gas molecules will return to their original temperature. This means that there will now be more molecules that exceed the escape velocity, and so more will escape. This process has by now played out, repeatedly, over the roughly 4.5 billion-year age of our solar system.

It is not precise, but we can estimate the probability that any particular species of gas has left a planetary atmosphere in these four billion years. A rule of thumb is that:

Thermal Escape Rule of Thumb

If the rms speed of a gas molecule exceeds $\frac{1}{6}$ of the planet's escape velocity, it is likely that none of that gas species still exists in the planet's atmosphere today.

Example 19.3 Hydrogen in Earth's Atmosphere

Earth has an average surface temperature of about 288 K. Hydrogen (H_2) molecules have a mass of about 3.34 x

10^{-27} kg. From **Equation 19.8** their rms speed is $v_{\text{rms}} = \sqrt{\frac{3\left(1.38\times10^{-23}\,\frac{\text{J}}{\text{K}}\right)(288\text{K})}{3.34\times10^{-27}\,\text{kg}}} = 1.89\times10^{3}\ \text{m/s}.$

From **Equation 19.9** the escape speed from Earth is

$$v_{\text{esc}} = \sqrt{\frac{2\left(6.67\times10^{-11}\,\frac{\text{N - m}^2}{\text{kg}^2}\right)\left(5.97\times10^{24}\,\text{kg}\right)}{6.34\times10^{6}\,\text{m}}} = 1.12\times10^{4}\ \text{m/s}.$$

In this case, the rms speed of H_2 molecules is about 17% of the escape velocity from Earth. So, 4.5 billion years after the formation of the planet, there is virtually no atmospheric hydrogen left in Earth's atmosphere.

We have discussed the runaway greenhouse effect on Venus and the runaway refrigerator effect on Mars, but we do not understand exactly what started these two planets down these separate evolutionary paths. Was Earth ever in danger of a similar fate? Or might it still be diverted onto one of these paths, perhaps due to stress on the atmosphere generated by human pollutants? One of the reasons for studying Venus and Mars is to seek insight into these questions.

Some people have even suggested that if we understood the evolution of Mars and Venus better, we could possibly reverse their evolution and restore more earthlike environments. While it seems unlikely that humans could ever make either Mars or Venus into a replica of Earth, considering such possibilities is a useful part of our more general quest to understand the delicate environmental balance that distinguishes our planet from its two neighbors.

For Further Exploration

Websites

Earth

Astronaut Photography of Earth from Space: **http://earth.jsc.nasa.gov/ (http://earth.jsc.nasa.gov/)** . A site with many images and good information.

Exploration of the Earth’s Magnetosphere: **http://phy6.org/Education/Intro.html (http://phy6.org/Education/Intro.html)** . An educational website by Dr. Daniel Stern.

NASA Goddard: Earth from Space: Fifteen Amazing Things in 15 Years: **https://www.nasa.gov/content/goddard/earth-from-space-15-amazing-things-in-15-years (https://www.nasa.gov/content/goddard/earth-from-space-15-amazing-things-in-15-years)** . Images and videos that reveal things about our planet and its atmosphere.

U.S. Geological Survey: Earthquake Information Center: **http://earthquake.usgs.gov/learn/ (http://earthquake.usgs.gov/learn/)**

Views of the Solar System: **http://www.solarviews.com/eng/earth.htm (http://www.solarviews.com/eng/earth.htm)** . Overview of Earth.

Venus

European Space Agency Venus Express Page: **http://www.esa.int/Our_Activities/Space_Science/Venus_Express (http://www.esa.int/Our_Activities/Space_Science/Venus_Express)** .

NASA Solar System Exploration Venus Page: **http://solarsystem.nasa.gov/planets/venus (http://solarsystem.nasa.gov/planets/venus)** .

NASA's apps about Mars for phones and tablets can be found at: **http://mars.nasa.gov/mobile/info/ (http://mars.nasa.gov/mobile/info/)** .

NASA's Magellan Mission to Venus: **http://www2.jpl.nasa.gov/magellan/ (http://www2.jpl.nasa.gov/magellan/)** .

Russian (Soviet) Venus Missions and Images: **http://mentallandscape.com/C_CatalogVenus.htm (http://mentallandscape.com/C_CatalogVenus.htm)** .

Venus Atlas app: **https://itunes.apple.com/us/app/venus-atlas/id317310503?mt=8 (https://itunes.apple.com/us/app/venus-atlas/id317310503?mt=8)** .

Venus Express Results Article: **http://www.mpg.de/798302/F002_Focus_026-033.pdf (http://www.mpg.de/798302/F002_Focus_026-033.pdf)** .

Mars

European Space Agency Mars Express Page: **http://www.esa.int/Our_Activities/Space_Science/Mars_Express (http://www.esa.int/Our_Activities/Space_Science/Mars_Express)** .

High Resolution Imaging Science Experiment: **http://hirise.lpl.arizona.edu/ (http://hirise.lpl.arizona.edu/)** .

Jet Propulsion Lab Mars Exploration Page: **http://mars.jpl.nasa.gov/ (http://mars.jpl.nasa.gov/)** .

Mars Globe HD app: **https://itunes.apple.com/us/app/mars-globe-hd/id376020224?mt=8 (https://itunes.apple.com/us/app/mars-globe-hd/id376020224?mt=8)** .

Mars Rover 360° Panorama: **http://www.360cities.net/image/curiosity-rover-martian-solar-day-2#171.10,26.50,70.0 (http://www.360cities.net/image/curiosity-rover-martian-solar-day-2#171.10,26.50,70.0)** . Interactive.

NASA Center for Mars Exploration: **http://www.nasa.gov/mission_pages/mars/main/index.html (http://www.nasa.gov/mission_pages/mars/main/index.html)** .

NASA Solar System Exploration Mars Page: **http://solarsystem.nasa.gov/planets/mars (http://solarsystem.nasa.gov/planets/mars)** .

Videos

50 Years of Mars Exploration: **http://www.jpl.nasa.gov/video/details.php?id=1395 (http://www.jpl.nasa.gov/video/details.php?id=1395)** . NASA's summary of all missions through *MAVEN*; good quick overview (4:08).

Being a Mars Rover: What It's Like to be an Interplanetary Explorer: **https://www.youtube.com/watch?v=nRpCOEsPD54 (https://www.youtube.com/watch?v=nRpCOEsPD54)** . 2013 talk by Dr. Lori Fenton about what it's like on the surface of Mars (1:07:24).

Magellan Maps Venus: **http://www.bbc.co.uk/science/space/solarsystem/space_missions/magellan_probe#p005y07s (http://www.bbc.co.uk/science/space/solarsystem/space_missions/magellan_probe#p005y07s)** . BBC clip with Dr. Ellen Stofan on the radar images of Venus and what they tell us (3:06).

Our *Curiosity*: **https://www.youtube.com/watch?v=XczKXWvokm4 (https://www.youtube.com/watch?v=XczKXWvokm4)** . Mars *Curiosity* rover 2-year anniversary video narrated by Neil deGrasse Tyson and Felicia Day (6:01).

Planet Venus: The Deadliest Planet, Venus Surface and Atmosphere: **https://www.youtube.com/watch?v=HqFVxWfVtoo (https://www.youtube.com/watch?v=HqFVxWfVtoo)** . Quick tour of Venus' atmosphere and surface (2:04).

Planetary Protection and Hitchhikers in the Solar System: The Danger of Mingling Microbes: **https://www.youtube.com/watch?v=6iGC3uO7jBI (https://www.youtube.com/watch?v=6iGC3uO7jBI)** . 2009 talk by Dr. Margaret Race on preventing contamination between worlds (1:28:50).

CHAPTER 19 REVIEW

KEY TERMS

bar a force of 100,000 Newtons acting on a surface area of 1 square meter; the average pressure of Earth's atmosphere at sea level is 1.013 bars

differentiation gravitational separation of materials of different density into layers in the interior of a planet or moon

half-life time required for half of the radioactive atoms in a sample to disintegrate

no-greenhouse temperature the average equilibrium surface temperature of a planet assuming that its incoming solar radiation and outgoing thermal emissions are in balance

ozone (O_3) a heavy molecule of oxygen that contains three atoms rather than the more normal two

radioactivity process by which certain kinds of atomic nuclei decay naturally, with the spontaneous emission of subatomic particles and gamma rays

runaway greenhouse effect the process by which the greenhouse effect, rather than remaining stable or being lessened through intervention, continues to grow at an increasing rate

stratosphere the layer of Earth's atmosphere above the troposphere and below the ionosphere

thermal escape The process whereby the thermal energy of certain gas molecules gives them a velocity sufficient to escape the gravity of a planet

troposphere the lowest level of Earth's atmosphere, where most weather takes place

SUMMARY

19.1 Composition and Structure of Planets

- The giant planets have dense cores roughly 10 times the mass of Earth, surrounded by layers of hydrogen and helium.
- The terrestrial planets consist mostly of rocks and metals. They were once molten, which allowed their structures to differentiate (that is, their denser materials sank to the center).
- The Moon resembles the terrestrial planets in composition, but most of the other moons—which orbit the giant planets—have larger quantities of frozen ice within them.
- In general, worlds closer to the Sun have higher surface temperatures.
- The surfaces of terrestrial planets have been modified by impacts from space and by varying degrees of geological activity.

19.2 Dating Planetary Surfaces

- The ages of the surfaces of objects in the solar system can be estimated by counting craters: on a given world, a more heavily cratered region will generally be older than one that is less cratered.
- We can also use samples of rocks with radioactive elements in them to obtain the time since the layer in which the rock formed last solidified.
- The half-life of a radioactive element is the time it takes for half the sample to decay; we determine how many half-lives have passed by how much of a sample remains the radioactive element and how much has become the decay product. In this way, we have estimated the age of the Moon and Earth to be roughly 4.5 billion years.

19.3 Earth's Atmosphere

- The atmosphere has a surface pressure of 1 bar and is composed primarily of N_2 and O_2, plus such important trace gases as H_2O, CO_2, and O_3.
- Its structure consists of the troposphere, stratosphere, mesosphere, and ionosphere.

- Changing the composition of the atmosphere also influences the temperature.
- Atmospheric circulation (weather) is driven by seasonally changing deposition of sunlight.
- Many longer term climate variations, such as the ice ages, are related to changes in the planet’s orbit and axial tilt.
- CO_2 and methane in the atmosphere heat the surface through the greenhouse effect; today, increasing amounts of atmospheric CO_2 are leading to the global warming of our planet.

19.4 The Massive Atmosphere of Venus

- The atmosphere of Venus is 96% CO_2.
- Thick clouds at altitudes of 30 to 60 kilometers are made of sulfuric acid, and a CO_2 greenhouse effect maintains the high surface temperature.
- Venus presumably reached its current state from more earthlike initial conditions as a result of a runaway greenhouse effect, which included the loss of large quantities of water.

19.5 Water and Life on Mars

- The martian atmosphere has a surface pressure of less than 0.01 bar and is 95% CO_2.
- It has dust clouds, water clouds, and carbon dioxide (dry ice) clouds.
- Liquid water on the surface is not possible today, but there is subsurface permafrost at high latitudes.
- Seasonal polar caps are made of dry ice; the northern residual cap is water ice, whereas the southern permanent ice cap is made predominantly of water ice with a covering of carbon dioxide ice.
- Evidence of a very different climate in the past is found in water erosion features: both runoff channels and outflow channels, the latter carved by catastrophic floods.
- Our rovers, exploring ancient lakebeds and places where sedimentary rock has formed, have found evidence for extensive surface water in the past.
- Even more exciting are the gullies that seem to show the presence of flowing salty water on the surface today, hinting at near-surface aquifers.
- The Viking landers searched for martian life in 1976, with negative results, but life might have flourished long ago.
- We have found evidence of water on Mars, but following the water has not yet led us to life on that planet.

19.6 Divergent Planetary Evolution

- Earth, Venus, and Mars have diverged in their evolution from what may have been similar beginnings.
- After the creation of the solar system, the planets cooled at different rates, which were inversely proportional to their radii.
- Over the age of our solar system, if the rms speed of a species of gas molecule exceeds $\frac{1}{6}$ of a planet's escape velocity, most of that gas will have escaped the planet's atmosphere.
- We need to understand why if we are to protect the environment of Earth.

CONCEPTUAL QUESTIONS

19.1 Composition and Structure of Planets

2. What is the difference between a differentiated body and an undifferentiated body, and how might that influence a body’s ability to retain heat for the age of the solar system?

3. Why are there so many craters on the Moon and so few on Earth?

4. How and why is Earth’s Moon different from the larger moons of the giant planets?

5. Explain why the planet Venus is differentiated, but asteroid Fraknoi, a very boring and small member of the asteroid belt, is not.

6. Would you expect as many impact craters per unit area on the surface of Venus as on the surface of Mars? Why or why not?

19.2 Dating Planetary Surfaces

7. Why are there so many craters on the Moon and so few on Earth?

8. Describe how we use radioactive elements and their decay products to find the age of a rock sample. Is this necessarily the age of the entire world from which the sample comes? Explain.

19.3 Earth's Atmosphere

9. What is the thickest interior layer of Earth? The thinnest?

10. List, in order of decreasing altitude, the principle layers of Earth's atmosphere.

11. In which atmospheric layer are almost all water-based clouds formed?

12. What is, by far, the most abundant component of Earth's atmosphere?

13. Briefly describe the greenhouse effect.

14. Why is a decrease in Earth's ozone harmful to life?

15. Why are we concerned about the increases in CO_2 and other gases that cause the greenhouse effect in Earth's atmosphere? What steps can we take in the future to reduce the levels of CO_2 in our atmosphere? What factors stand in the way of taking the steps you suggest? (You may include technological, economic, and political factors in your answer.)

19.4 The Massive Atmosphere of Venus

16. How might Venus' atmosphere have evolved to its present state through a runaway greenhouse effect?

17. What evidence is there that Venus was volcanically active about 300–600 million years ago?

18. Describe two anomalous features of the rotation of Venus and what might account for them.

19. Why is there so much more carbon dioxide in the atmosphere of Venus than in that of Earth? Why so much more carbon dioxide than on Mars?

20. Suppose that, decades from now, NASA is considering sending astronauts to Mars and Venus. In each case, describe what kind of protective gear they would have to carry, and what their chances for survival would be if their spacesuits ruptured.

21. In what way is the high surface temperature of Venus relevant to concerns about global warming on Earth today?

19.5 Water and Life on Mars

22. Describe the current atmosphere on Mars. What evidence suggests that it must have been different in the past?

23. Explain the runaway refrigerator effect and the role it may have played in the evolution of Mars.

24. What evidence do we have that there was running (liquid) water on Mars in the past? What evidence is there for water coming out of the ground even today?

25. Why is Mars red?

26. What is the composition of clouds on Mars?

27. What is the composition of the polar caps on Mars?

28. How was the *Mars Odyssey* spacecraft able to detect water on Mars without landing on it?

29. If the Viking missions were such a rich source of information about Mars, why have we sent the Pathfinder, *Global Surveyor*, and other more recent spacecraft to Mars? Make a list of questions about Mars that still puzzle astronomers.

30. One source of information about Mars has been the analysis of meteorites from Mars. Since no samples from Mars have ever been returned to Earth from any of the missions we sent there, how do we know these meteorites are from Mars? What information have they revealed about Mars?

19.6 Divergent Planetary Evolution

31. List several ways that Venus, Earth, and Mars are similar, and several ways they are different.

32. Which species of gas molecules are most likely to escape from a planet's atmosphere?

33. What specfic planetary properties would promote the thermal escape of gas molecules from its atmosphere?

34. After the formation of the solar system, which terrestrial planets cooled off most quickly? Why?

35. Compare the current atmospheres of Earth, Venus, and Mars in terms of composition, thickness (and pressure at the surface), and the greenhouse effect.

36. Venus and Earth are nearly the same size and distance from the Sun. What are the main differences in the geology of the two planets? What might be some of the reasons for these differences?

37. Why is there so much more carbon dioxide in the atmosphere of Venus than in that of Earth? Why so much more carbon dioxide than on Mars?

38. Contrast the mountains on Mars and Venus with those on Earth and the Moon.

39. Is it likely that life ever existed on either Venus or Mars? Justify your answer in each case.

40. Suppose that, decades from now, NASA is considering sending astronauts to Mars and Venus. In each case, describe what kind of protective gear they would have to carry, and what their chances for survival would be if their spacesuits ruptured.

41. We believe that Venus, Earth, and Mars all started with a significant supply of water. Explain where that water is now for each planet.

42. The runaway greenhouse effect and its inverse, the runaway refrigerator effect, have led to harsh, uninhabitable conditions on Venus and Mars. Does the greenhouse effect always cause climate changes leading to loss of water and life? Give a reason for your answer.

43. Near the martian equator, temperatures at the same spot can vary from an average of –135 °C at night to an average of 30 °C during the day. How can you explain such a wide difference in temperature compared to that on Earth?

PROBLEMS

19.2 Dating Planetary Surfaces

44. A radioactive nucleus has a half-life of 5×10^8 years. Assuming that a sample of rock (say, in an asteroid) solidified right after the solar system formed, approximately what fraction of the radioactive element should be left in the rock today?

19.3 Earth's Atmosphere

45. What is the percent increase of atmospheric CO_2 in the past 20 years?

19.4 The Massive Atmosphere of Venus

46. If you weigh 150 lbs. on the surface of Earth, how much would you weigh on Venus?

19.5 Water and Life on Mars

47. If you weigh 150 lbs. on the surface of Earth, how much would you weigh on Venus? On Mars?

48. The closest approach distance between Mars and Earth is about 56 million km. Assume you can travel in a spaceship at 58,000 km/h, which is the speed achieved by the New Horizons space probe that went to Pluto and is the fastest speed so far of any space vehicle launched from Earth. How long would it take to get to Mars at the time of closest approach?

19.6 Divergent Planetary Evolution

49. Estimate the amount of water there could be in a global (planet-wide) region of subsurface permafrost on Mars (do the calculations for two permafrost thicknesses, 1 and 10 km, and a concentration of ice in the permafrost of 10% by volume). Compare the two results you get with the amount of water in Earth's oceans calculated in **Example 19.2**.

50. Calculate the relative land area—that is, the amount of the surface not covered by liquids—of Earth, the Moon, Venus, and Mars. (Assume that 70% of Earth is covered with water.)

51. **Appendix D** lists the escape velocities for the terrestrial planets. What would the temperature of a He molecule need to be in order for its rms speed to exceed 20% of the escape velocity for each of these planets? (The mass of a He molecule is approximately 4 atomic mass units.)

CHALLENGE PROBLEMS

52. Venus has an average reflectivity of 75%, and is located 0.723 AU from the Sun. Calculate its no-greenhouse temperature.

20 | EXOPLANETS

Chapter Outline

Introduction

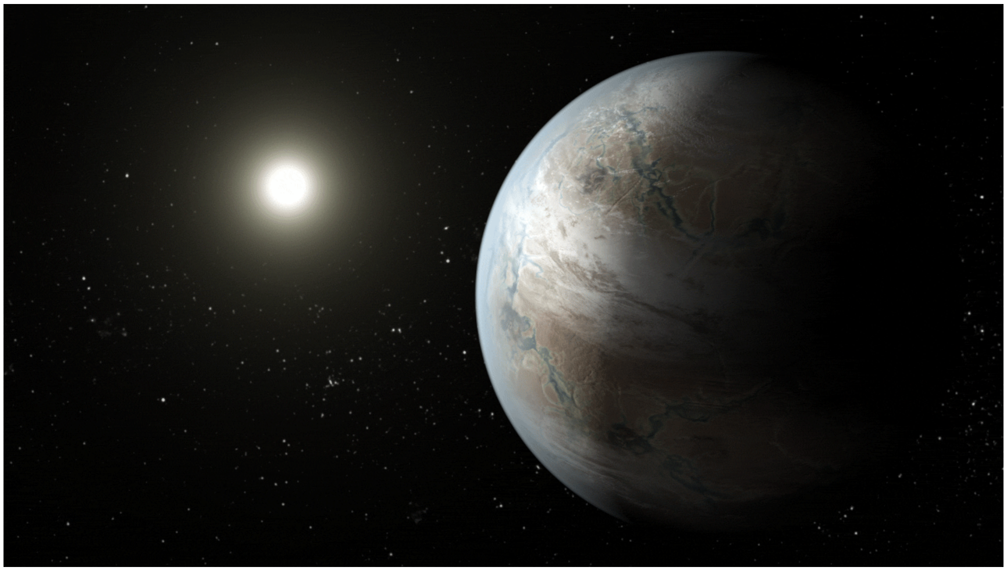


Figure 20.1 This artistic animation depicts one possible appearance of the planet Kepler-452b, the first near-Earth-size world to be found in the habitable zone of star that is similar to our sun. The star, Kepler-452, is a G2-type star like our sun, with nearly the same temperature and mass.(credit: NASA)

Until the middle 1990s, the practical study of the origin of planets focused on our single known example—the solar system. Although there had been a great deal of speculation about planets circling other stars, none had actually been detected. Logically enough, in the absence of data, most scientists assumed that our own system was likely to be typical. They were in for a big surprise.

Discovery of Other Planetary Systems

In **Formation of the Solar System**, we discussed the formation of stars and planets. Stars like our Sun are formed when dense regions in a molecular cloud (made of gas and dust) feel an extra gravitational force and begin to collapse. This is a runaway process: as the cloud collapses, the gravitational force gets stronger, concentrating material into a protostar. Roughly half of the time, the protostar will fragment or be gravitationally bound to other protostars, forming a binary or multiple star system—stars that are gravitationally bound and orbit each other. The rest of the time, the protostar collapses in isolation, as was the case for our Sun. In all cases, as we saw, conservation of angular momentum results in a spin-up of the collapsing protostar, with surrounding material flattened into a disk. Today, this kind of structure can actually be observed. The Hubble Space Telescope, as well as powerful new ground-based telescopes, enable astronomers to study directly the nearest of these *circumstellar disks* in regions of space where stars are being born today, such as the Orion Nebula (**Figure 20.2**) or the Taurus star-forming region.

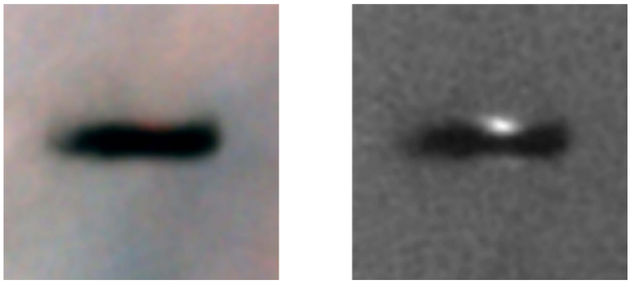
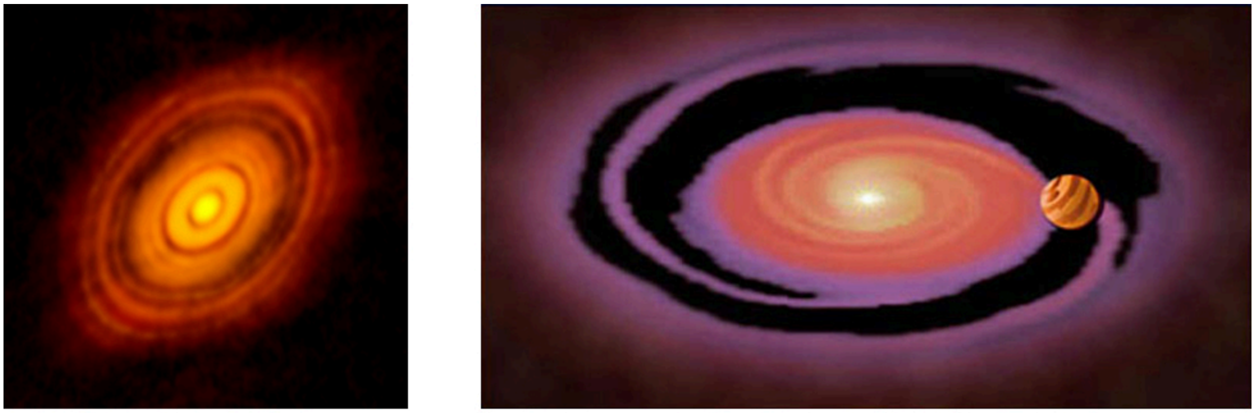


Figure 20.2 The Hubble Space Telescope imaged this protoplanetary disk in the Orion Nebula, a region of active star formation, using two different filters. The disk, about 17 times the size of our solar system, is in an edge-on orientation to us, and the newly formed star is shining at the center of the flattened dust cloud. The dark areas indicate absorption, not an absence of material. In the left image we see the light of the nebula and the dark cloud; in the right image, a special filter was used to block the light of the background nebula. You can see gas above and below the disk set to glow by the light of the newborn star hidden by the disk. (credit: modification of work by Mark McCaughrean (Max-Planck-Institute for Astronomy), C. Robert O'Dell (Rice University), and NASA)

Many of the circumstellar disks we have discovered show internal structure. The disks appear to be donut-shaped, with gaps close to the star. Such gaps indicate that the gas and dust in the disk have already collapsed to form large planets (**Figure 20.3**). The newly born protoplanets are too small and faint to be seen directly, but the depletion of raw materials in the gaps hints at the presence of something invisible in the inner part of the circumstellar disk—and that something is almost certainly one or more planets. Theoretical models of planet formation, like the one seen at right in **Figure 20.3**, have long supported the idea that planets would clear gaps as they form in disks.



(a) (b)

Figure 20.3 (a) This image of a protoplanetary disk around HL Tau was taken with the Atacama Large Millimeter/submillimeter Array (ALMA), which allows astronomers to construct radio images that rival those taken with visible light. (b) Newly formed planets that orbit the central star clear out dust lanes in their paths, just as our theoretical models predict. This computer simulation shows the empty lane and spiral density waves that result as a giant planet is forming within the disk. The planet is not shown to scale. (credit a: modification of work by ALMA (ESO/NAOJ/NRAO); credit b: modification of work by NASA/ESA and A. Feild (STScI))

Our figure shows HL Tau, a one-million-year-old "newborn" star in the Taurus star-forming region. The star is embedded in a shroud of dust and gas that obscures our visible-light view of a circumstellar disk around the star. In 2014 astronomers obtained a dramatic view of the HL Tau circumstellar disk using millimeter waves, which pierce the cocoon of dust around the star, showing dust lanes being carved out by several newly formed protoplanets. As the mass of the protoplanets increases, they travel in their orbits at speeds that are faster than the dust and gas in the circumstellar disk. As the protoplanets plow through the disk, their gravitational reach begins to exceed their cross-sectional area, and they become very efficient at sweeping up material and growing until they clear a gap in the disk. The image of **Figure 20.3** shows us that a number of protoplanets are forming in the disk and that they were able to form faster than our earlier ideas had suggested—all in the first million years of star formation.

For an explanation of ALMA's ground-breaking observations of HL Tau and what they reveal about plant formation, watch this **videocast (https://openstaxcollege.org/l/30eusoobhltavid)** from the European Southern Observatory.

Discovering Exoplanets

You might think that with the advanced telescopes and detectors astronomers have today, they could directly image planets around nearby stars (which we call **exoplanets**). This has proved extremely difficult, however, not only because the exoplanets are faint, but also because they are generally lost in the brilliant glare of the star they orbit. As we discuss in more detail in **Planets Beyond the Solar System**, the detection techniques that work best are indirect: they observe the effects of the planet on the star it orbits, rather than seeing the planet itself.

The first technique that yielded many planet detections is very high-resolution stellar spectroscopy. The *Doppler effect* lets astronomers measure the star's *radial velocity:* that is, the speed of the star, toward us or away from us, relative to the observer. If there is a massive planet in orbit around the star, the gravity of the planet causes the star to wobble, changing its radial velocity by a small but detectable amount. The distance of the star does not matter, as long as it is bright enough for us to take very high quality spectra.

Measurements of the variation in the star's radial velocity as the planet goes around the star can tell us the mass and orbital period of the planet. If there are several planets present, their effects on the radial velocity can be disentangled, so the entire planetary system can be deciphered—as long as the planets are massive enough to produce a measureable Doppler effect. This detection technique is most sensitive to large planets orbiting close to the star, since these produce the greatest wobble in their stars. It has been used on large ground-based telescopes to detect hundreds of planets, including one around Proxima Centauri, the nearest star to the Sun.

The second indirect technique is based on the slight dimming of a star when one of its planets *transits*, or crosses over the face of the star, as seen from Earth. Astronomers do not see the planet, but only detect its presence from careful measurements of a change in the brightness of the star over long periods of time. If the slight dips in brightness repeat at regular intervals, we can determine the orbital period of the planet. From the amount of starlight obscured, we can measure the planet's size.

While some transits have been measured from Earth, large-scale application of this transit technique requires a telescope in space, above the atmosphere and its distortions of the star images. It has been most successfully applied from the NASA Kepler space observatory, which was built for the sole purpose of "staring" for 5 years at a single part of the sky, continuously monitoring the light from more than 150,000 stars. The primary goal of Kepler was to determine the frequency of occurrence of exoplanets of different sizes around different classes of stars. Like the Doppler technique, the transit observations favor discovery of large planets and short-period orbits.

Recent detection of exoplanets using both the Doppler and transit techniques has been incredibly successful. Within two decades, we went from no knowledge of other planetary systems to a catalog of *thousands* of exoplanets. Most of the exoplanets found so far are more massive than or larger in size than Earth. It is not that Earth analogs do not exist. Rather, the shortage of small rocky planets is an observational bias: smaller planets are more difficult to detect.

Analyses of the data to correct for such biases or selection effects indicate that small planets (like the terrestrial planets in our system) are actually much more common than giant planets. Also relatively common are "super Earths," planets with two to ten times the mass of our planet (**Figure 20.4**). We don't have any of these in our solar system, but nature seems to have no trouble making them elsewhere. Overall, the Kepler data suggest that approximately one quarter of stars have exoplanet systems, implying the existence of at least 50 billion planets in our Galaxy alone.

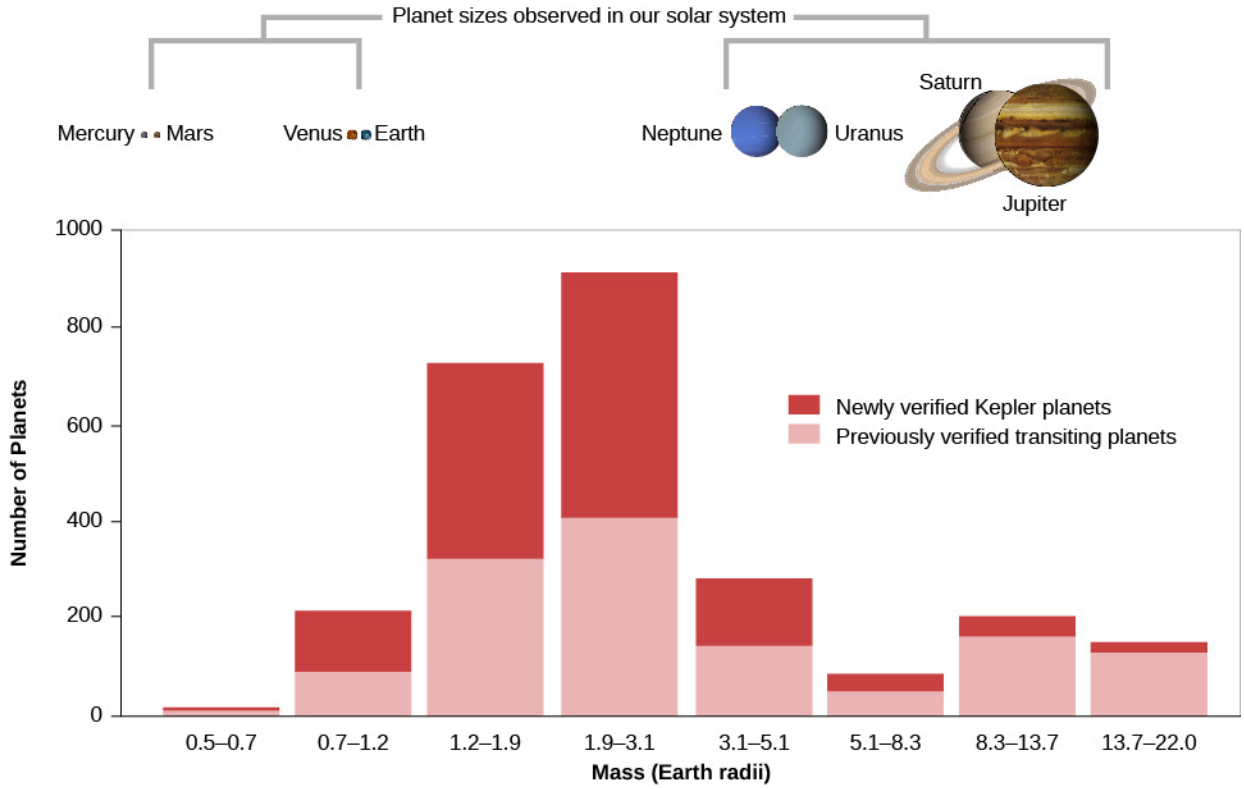


Figure 20.4 This bar graph shows the planets found so far using the transit method (the vast majority found by the Kepler mission). The orange parts of each bar indicate the planets announced by the Kepler team in May 2016. Note that the largest number of planets found so far are in two categories that we don't have in our own solar system—planets whose size is between Earth's and Neptune's. (credit: modification of work by NASA)

The Configurations of Other Planetary Systems

Let's look more closely at the progress in the detection of exoplanets. **Figure 20.5** shows the planets that were discovered each year by the two techniques we discussed. In the early years of exoplanet discovery, most of the planets were similar in mass to Jupiter. This is because, as mentioned above, the most massive planets were easiest to detect. In more recent years, planets smaller than Neptune and even close to the size of Earth have been detected.

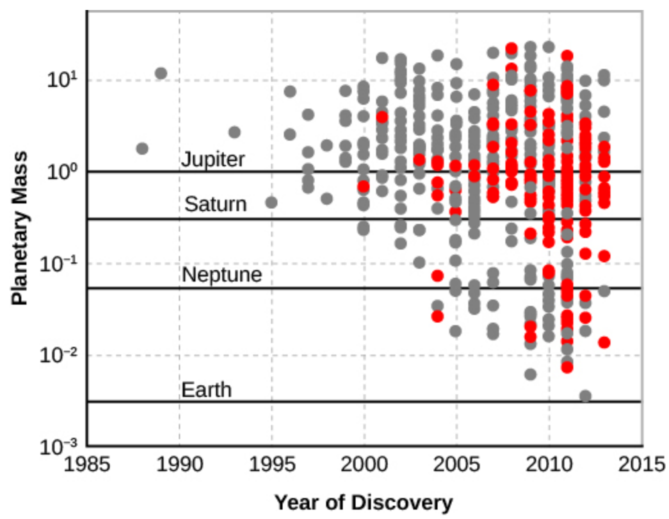


Figure 20.5 Horizontal lines are drawn to reference the masses of Jupiter, Saturn, Neptune, and Earth. The gray dots indicate planets discovered by measuring the radial velocity of the star, and the red dots are for planets that transit their stars. In the early years, the only planets that could be detected were similar in mass to Jupiter. Improvements in technology and observing strategies enabled the detection of lower mass planets as time went on, and now even smaller worlds are being found. (Note that this tally ends in 2014.)

We also know that many exoplanets are in multiplanet systems. This is one characteristic that our solar system shares with exosystems. Looking back at **Figure 20.3** and seeing how such large disks can give rise to more than one center of condensation, it is not too surprising that multiplanet systems are a typical outcome of planet formation. Astronomers have tried to measure whether multiple planet systems all lie in the same plane using astrometry. This is a difficult measurement to make with current technology, but it is an important measurement that could help us understand the origin and evolution of planetary systems.

Comparison between Theory and Data

Many of the planetary systems discovered so far do not resemble our own solar system. Consequently, we have had to reassess some aspects of the “standard models” for the formation of planetary systems. Science sometimes works in this way, with new data contradicting our expectations. The press often talks about a scientist making experiments to “confirm” a theory. Indeed, it is comforting when new data support a hypothesis or theory and increase our confidence in an earlier result. But the most exciting and productive moments in science often come when new data *don’t* support existing theories, forcing scientists to rethink their position and develop new and deeper insights into the way nature works.

Nothing about the new planetary systems contradicts the basic idea that planets form from the aggregation (clumping) of material within circumstellar disks. However, the existence of “ hot Jupiters”—planets of jovian mass that are closer to their stars than the orbit of Mercury—poses the biggest problem. As far as we know, a giant planet cannot be formed without the condensation of water ice, and water ice is not stable so close to the heat of a star. It seems likely that all the giant planets, “hot” or “normal,” formed at a distance of several astronomical units from the star, but we now see that they did not necessarily stay there. This discovery has led to a revision in our understanding of planet formation that now includes “planet migrations” within the protoplanetary disk, or later gravitational encounters between sibling planets that scatter one of the planets inward.

Many exoplanets have large orbital eccentricity (recall this means the orbits are not circular). High eccentricities were not expected for planets that form in a disk. This discovery provides further support for the scattering of planets when they interact gravitationally. When planets change each other’s motions, their orbits could become much more eccentric than the ones with which they began.

There are several suggestions for ways migration might have occurred. Most involve interactions between the giant planets and the remnant material in the circumstellar disk from which they formed. These interactions would have taken place when the system was very young, while material still remained in the disk. In such cases, the planet travels at a faster velocity than the gas and dust and feels a kind of “headwind” (or friction) that causes it to lose energy and spiral inward. It is still unclear how the spiraling planet stops before it plunges into the star. Our best guess is that this plunge into the star is the fate for many protoplanets; however, clearly some migrating planets can stop their inward motions and escape this destruction, since we find hot Jupiters in many mature planetary systems.

Summary

- The first planet circling a distant solar-type star was announced in 1995.
- Twenty years later, thousands of exoplanets have been identified, including planets with sizes and masses between Earth's and Neptune's, which we don't have in our own solar system.
- A few percent of exoplanet systems have "hot Jupiters," massive planets that orbit close to their stars, and many exoplanets are also in eccentric orbits.
- These two characteristics are fundamentally different from the attributes of gas giant planets in our own solar system and suggest that giant planets can migrate inward from their place of formation where it is cold enough for ice to form.
- Current data indicate that small (terrestrial type) rocky planets are common in our Galaxy; indeed, there must be tens of billions of such earthlike planets.

20.1 | Planets Beyond the Solar System: Search and Discovery

Learning Objectives

By the end of this section, you will be able to:

- Describe the orbital motion of planets in our solar system using Kepler's laws
- Compare the indirect and direct observational techniques for exoplanet detection

For centuries, astronomers have dreamed of finding planets around other stars, including other planets like Earth. Direct observations of such distant planets are very difficult, however. You might compare a planet orbiting a star to a mosquito flying around one of those giant spotlights at a shopping center opening. From close up, you might spot the mosquito. But imagine viewing the scene from some distance away—say, from an airplane. You could see the spotlight just fine, but what are your chances of catching the mosquito in that light? Instead of making direct images, astronomers have relied on indirect observations and have now succeeded in detecting a multitude of planets around other stars.

In 1995, after decades of effort, we found the first such **exoplanet** (a planet outside our solar system) orbiting a main-sequence star, and today we know that most stars form with planets. This is an example of how persistence and new methods of observation advance the knowledge of humanity. By studying exoplanets, astronomers hope to better understand our solar system in context of the rest of the universe. For instance, how does the arrangement of our solar system compare to planetary systems in the rest of the universe? What do exoplanets tell us about the process of planet formation? And how does knowing the frequency of exoplanets influence our estimates of whether there is life elsewhere?

Searching for Orbital Motion

Most exoplanet detections are made using techniques where we observe the *effect* that the planet exerts on the host star. For example, the gravitational tug of an unseen planet will cause a small wobble in the host star. Or, if its orbit is properly aligned, a planet will periodically cross in front of the star, causing the brightness of the star to dim.

To understand how a planet can move its host star, consider a single Jupiter-like planet. Both the planet and the star actually revolve about their *common center of mass*. Remember that gravity is a mutual attraction. The star and the planet each exert a force on the other, and we can find a stable point, the center of mass, between them about which both objects move. The smaller the mass of a body in such a system, the larger its orbit. A massive star barely swings around the center of mass, while a low-mass planet makes a much larger "tour."

Suppose the planet is like Jupiter and has a mass about one-thousandth that of its star; in this case, the size of the star's orbit is one-thousandth the size of the planet's. To get a sense of how difficult observing such motion might be, let's see how hard Jupiter would be to detect in this way from the distance of a nearby star. Consider an alien astronomer trying to observe our own system from Alpha Centauri, the closest star system to our own (about 4.3 light-years away). There are two ways this astronomer could try to detect the orbital motion of the Sun. One way would be to look for changes in the Sun's position on the sky. The second would be to use the Doppler effect to look for changes in its velocity. Let's discuss each of these in turn.

The diameter of Jupiter's apparent orbit viewed from Alpha Centauri is 10 seconds of arc, and that of the Sun's orbit is 0.010 seconds of arc. (Remember, 1 second of arc is 1/3600 degree.) If they could measure the apparent position of the Sun (which is bright and easy to detect) to sufficient precision, they would describe an orbit of diameter 0.010 seconds of arc with a period equal to that of Jupiter, which is 12 years.

In other words, if they watched the Sun for 12 years, they would see it wiggle back and forth in the sky by this minuscule fraction of a degree. From the observed motion and the period of the "wiggle," they could deduce the mass of Jupiter and its distance using Kepler's laws. (To refresh your memory about these laws, see the chapter on **Kepler's Laws of Planetary Motion**.)

Measuring positions in the sky this accurately is extremely difficult, and so far, astronomers have not made any confirmed detections of planets using this technique. However, we have been successful in using spectrometers to measure the changing velocity of stars with planets around them.

As the star and planet orbit each other, part of their motion will be in our line of sight (toward us or away from us). Such motion can be measured using the *Doppler effect* and the star's spectrum. As the star moves back and forth in orbit around the system's center of mass in response to the gravitational tug of an orbiting planet, the lines in its spectrum will shift back and forth.

Let's again consider the example of the Sun. Its *radial velocity* (motion toward or away from us) changes by about 13 meters per second with a period of 12 years because of the gravitational pull of Jupiter. This corresponds to about 30 miles per hour, roughly the speed at which many of us drive around town. Detecting motion at this level in a star's spectrum presents an enormous technical challenge, but several groups of astronomers around the world, using specialized spectrographs designed for this purpose, have succeeded. Note that the change in speed does not depend on the distance of the star from the observer. Using the Doppler effect to detect planets will work at any distance, as long as the star is bright enough to provide a good spectrum and a large telescope is available to make the observations (**Figure 20.6**).

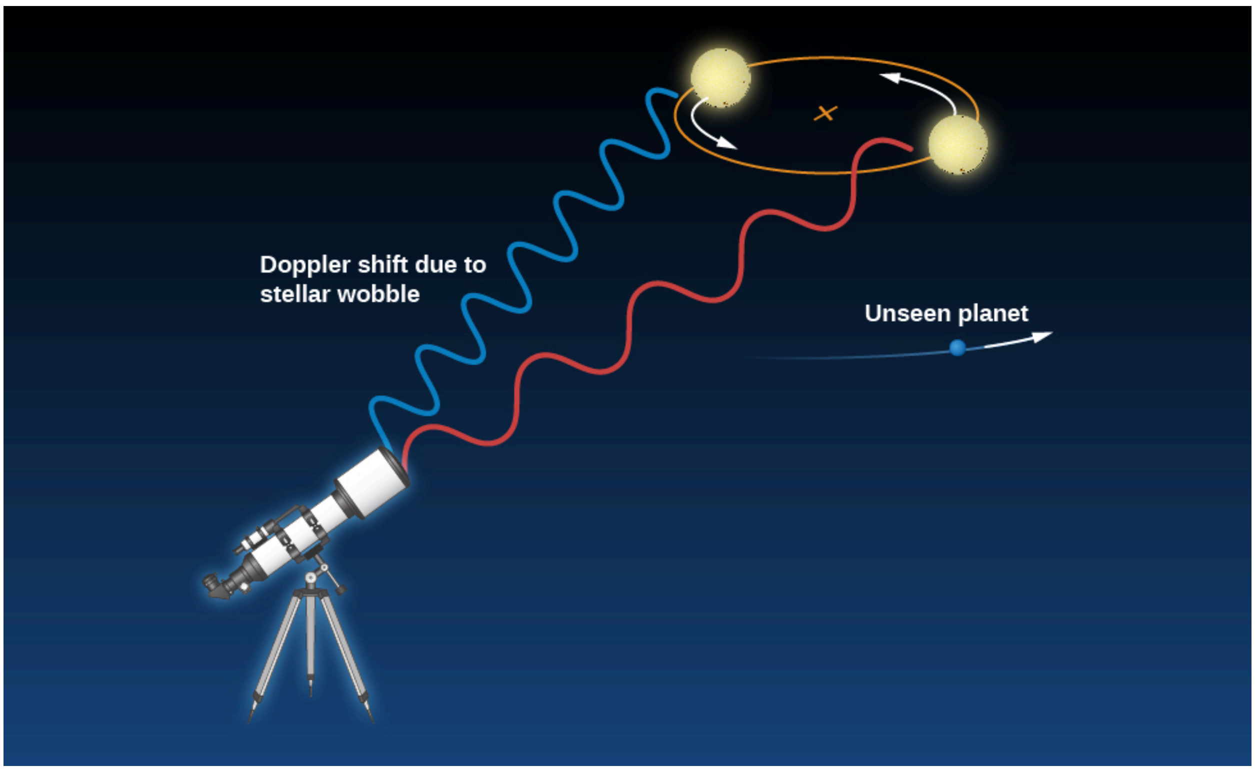


Figure 20.6 The motion of a star around a common center of mass with an orbiting planet can be detected by measuring the changing speed of the star. When the star is moving away from us, the lines in its spectrum show a tiny redshift; when it is moving toward us, they show a tiny blueshift. The change in color (wavelength) has been exaggerated here for illustrative purposes. In reality, the Doppler shifts we measure are extremely small and require sophisticated equipment to be detected.

From the Doppler formula (see **Section 14.2**, the radial velocity is proportional to the shift in wavelength. So, if the Doppler-shifted light from a wobbling star is studied over a long period of time, it is possible to construct a radial velocity graph of the star's motion as a function of time.

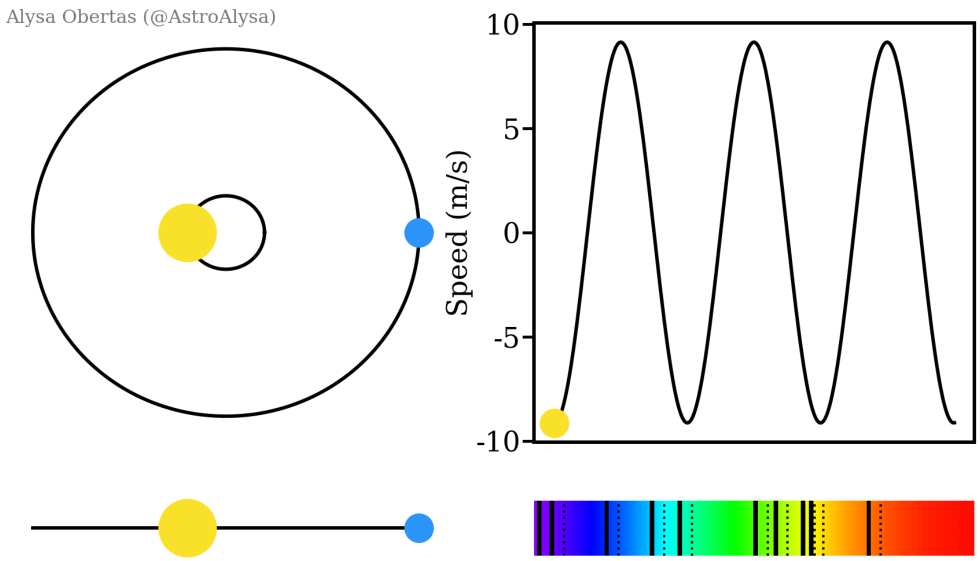


Figure 20.7 The wobbling motion of a star due to an orbiting planet can be detected by measuring the changing speed of the star. Here the Doppler measurements have been analyzed and converted into a radial velocity versus time graph. (Credit: Alysa Obertas [CC BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0)])

Let's assume that the amplitude of the star's wobble (found from the amplitude of the peaks and valleys in the graph) is v_{star}, and that the period of its wobble (from the same graph) is T. If we know the mass of the star, M_{star} (perhaps deduced from its temperature and luminosity), then **Kepler's Third Law** can be used to determine the orbital distance (semi-major axis) of the planet's motion:

$$a = \sqrt[3]{\frac{GM_{\text{star}}}{4\pi^2}T^2} \quad (20.1)$$

Now, since the planet and the star have exactly the same period of motion, T, we can easily find the orbital velocity of the planet by dividing the circumference of its motion by that period:

$$v_{\text{planet}} = \frac{2\pi a}{T} \quad (20.2)$$

Now, if we assume that the momentum of the center-of-mass of the star-planet system is zero, then the magnitude of the star's linear momentum is equal to that of the planet. (Their linear momenta are simply oriented in opposite directions.) As an equation:

$$M_{\text{planet}} v_{\text{planet}} = M_{\text{star}} v_{\text{star}} \quad (20.3)$$

Thus, we have the ability to calculate the planet's mass if we know the star's mass.

The first successful use of the Doppler effect to find a planet around another star was in 1995. Michel Mayor and Didier Queloz of the Geneva Observatory (**Figure 20.8**) used this technique to find a planet orbiting a star resembling our Sun called 51 Pegasi, about 40 light-years away. (The star can be found in the sky near the great square of Pegasus, the flying horse of Greek mythology, one of the easiest-to-find star patterns.) To everyone's surprise, the planet takes a mere 4.2 days to orbit around the star. (Remember that Mercury, the innermost planet in our solar system, takes 88 days to go once around the Sun, so 4.2 days seems fantastically short.)

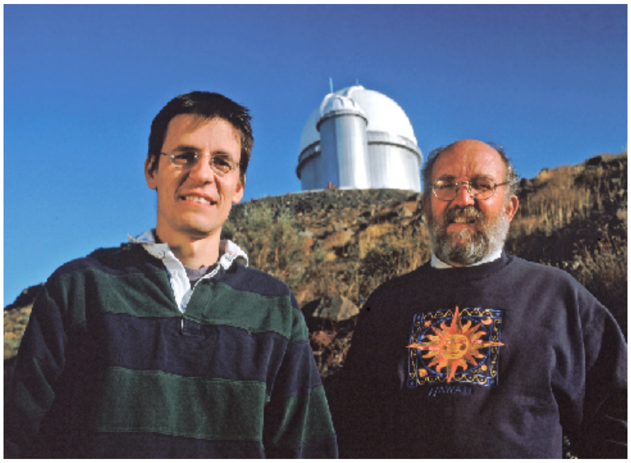


Figure 20.8 In 1995, Didier Queloz and Michel Mayor of the Geneva Observatory were the first to discover a planet around a regular star (51 Pegasi). They are seen here at an observatory in Chile where they are continuing their planet hunting. (credit: Weinstein/Ciel et Espace Photos)

Mayor and Queloz's findings mean the planet must be very close to 51 Pegasi, circling it about 7 million kilometers away (**Figure 20.9**). At that distance, the energy of the star should heat the planet's surface to a temperature of a few thousand degrees Celsius (a bit hot for future tourism). From its motion, astronomers calculate that it has at least half the mass of Jupiter[1], making it clearly a jovian and not a terrestrial-type planet.

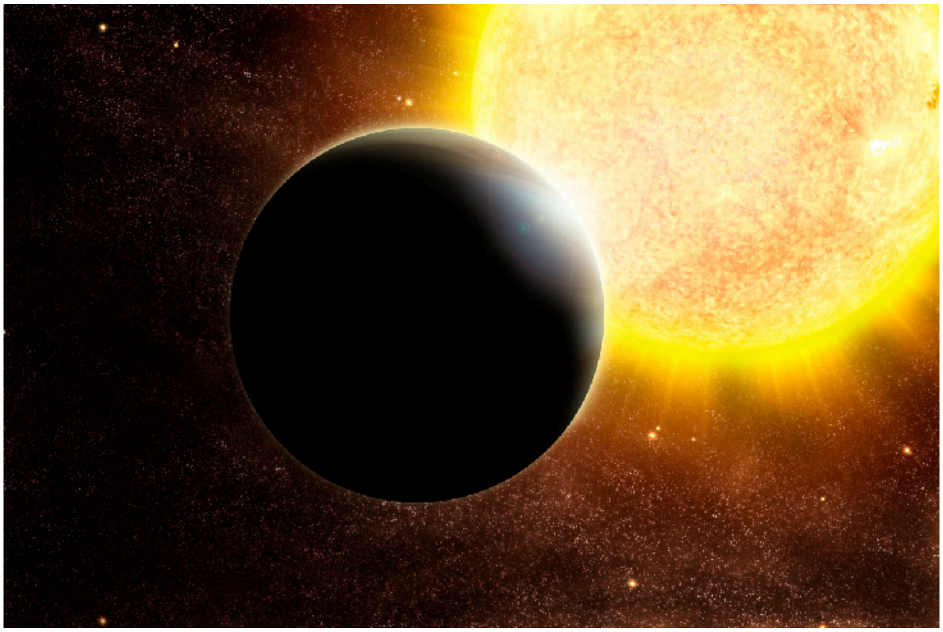


Figure 20.9 Artist Greg Bacon painted this impression of a hot, Jupiter-type planet orbiting close to a sunlike star. The artist shows bands on the planet like Jupiter, but we only estimate the mass of most hot, Jupiter-type planets from the Doppler method and don't know what conditions on the planet are like. (credit: ESO)

1. The Doppler method only allows us to find the minimum mass of a planet. To determine the exact mass using the Doppler shift and Kepler's laws, we must also have the angle at which the planet's orbit is oriented to our view—something we don't have any independent way of knowing in most cases. Still, if the minimum mass is half of Jupiter's, the actual mass can only be larger than that, and we are sure that we are dealing with a jovian planet.

Since that initial planet discovery, the rate of progress has been breathtaking. Hundreds of giant planets have been discovered using the Doppler technique. Many of these giant planets are orbiting close to their stars—astronomers have called these *hot Jupiters*.

The existence of giant planets so close to their stars was a surprise, and these discoveries have forced us to rethink our ideas about how planetary systems form. But for now, bear in mind that the Doppler-shift method—which relies on the pull of a planet making its star "wiggle" back and forth around the center of mass—is most effective at finding planets that are both close to their stars and massive. These planets cause the biggest "wiggles" in the motion of their stars and the biggest Doppler shifts in the spectrum. Plus, they will be found sooner, since astronomers like to monitor the star for at least one full orbit (and perhaps more) and hot Jupiters take the shortest time to complete their orbit.

So if such planets exist, we would expect to be finding this type first. Scientists call this a *selection effect*—where our technique of discovery selects certain kinds of objects as "easy finds." As an example of a selection effect in everyday life, imagine you decide you are ready for a new romantic relationship in your life. To begin with, you only attend social events on campus, all of which require a student ID to get in. Your selection of possible partners will then be limited to students at your college. That may not give you as diverse a group to choose from as you want. In the same way, when we first used the Doppler technique, it selected massive planets close to their stars as the most likely discoveries. As we spend longer times watching target stars and as our ability to measure smaller Doppler shifts improves, this technique can reveal more distant and less massive planets too.

View a **series of animations (https://openstaxcollege.org/l/30keplawsolarani)** demonstrating solar system motion and Kepler's laws, and select animation 1 (Kepler's laws) from the dropdown playlist. To view an animation demonstrating the radial velocity curve for an exoplanet, select animation 29 (radial velocity curve for an exoplanet) and animation 30 (radial velocity curve for an exoplanet—elliptical orbit) from the dropdown playlist.

Transiting Planets

The second method for indirect detection of exoplanets is based not on the motion of the star but on its brightness. When the orbital plane of the planet is tilted or inclined so that it is viewed edge-on, we will see the planet cross in front of the star once per orbit, causing the star to dim slightly; this event is known as **transit**. **Figure 20.10** shows a sketch of the transit at three time steps: (1) out of transit, (2) the start of transit, and (3) full transit, along with a sketch of the light curve, which shows the drop in the brightness of the host star. The amount of light blocked—the depth of the transit—depends on the area of the planet (its size) compared to the star. If we can determine the size of the star, the transit method tells us the size of the planet.

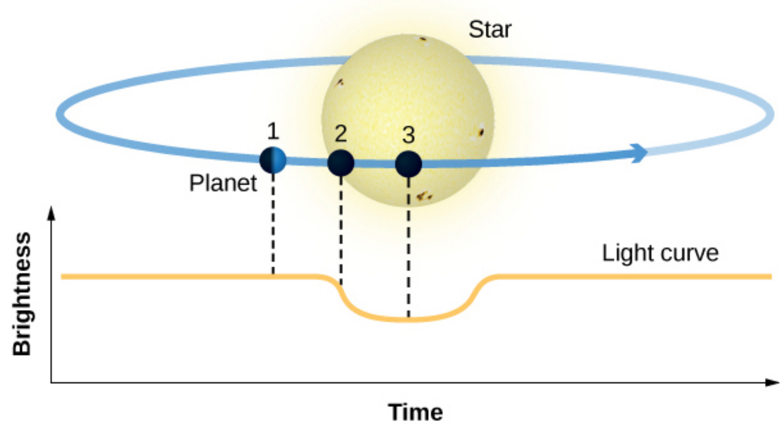


Figure 20.10 As the planet transits, it blocks out some of the light from the star, causing a temporary dimming in the brightness of the star. The top figure shows three moments during the transit event and the bottom panel shows the corresponding light curve: (1) out of transit, (2) transit ingress, and (3) the full drop in brightness.

The interval between successive transits is the length of the year for that planet, which can be used (again using Kepler's laws) to find its distance from the star. Larger planets like Jupiter block out more starlight than small earthlike planets, making transits by giant planets easier to detect, even from ground-based observatories. But by going into space, above the distorting effects of Earth's atmosphere, the transit technique has been extended to exoplanets as small as Mars.

Example 20.1

Transit Depth

In a transit, the planet's circular disk blocks the light of the star's circular disk. The area of a circle is πR^2. The amount of light the planet blocks, called the transit depth, is then given by

$$\frac{\pi R^2{}_{\text{planet}}}{\pi R^2{}_{\text{star}}} = \frac{R^2{}_{\text{planet}}}{R^2{}_{\text{star}}} = \left(\frac{R_{\text{planet}}}{R_{\text{star}}}\right)^2$$

Now calculate the transit depth for a star the size of the Sun with a gas giant planet the size of Jupiter.

Solution

The radius of Jupiter is 71,400 km, while the radius of the Sun is 695,700 km. Substituting into the equation, we get $\left(\frac{R_{\text{planet}}}{R_{\text{star}}}\right)^2 = \left(\frac{71{,}400\text{ km}}{695{,}700\text{ km}}\right)^2 = 0.01$ or 1%, which can easily be detected with the instruments on board the Kepler spacecraft.

20.1 What is the transit depth for a star half the size of the Sun with a much smaller planet, like the size of Earth?

The Doppler method allows us to estimate the mass of a planet. If the same object can be studied by both the Doppler and transit techniques, we can measure both the mass and the size of the exoplanet. This is a powerful combination that can be used to derive the average density (mass/volume) of the planet. In 1999, using measurements from ground-based telescopes, the first transiting planet was detected orbiting the star HD 209458. The planet transits its parent star for about 3 hours every 3.5 days as we view it from Earth. Doppler measurements showed that the planet around HD 209458 has about 70% the mass of Jupiter, but its radius is about 35% larger than Jupiter's. This was the first case where we could determine what an exoplanet was made of—with that mass and radius, HD 209458 must be a gas and liquid world like Jupiter or Saturn.

It is even possible to learn something about the planet's atmosphere. When the planet passes in front of HD 209458, the atoms in the planet's atmosphere absorb starlight. Observations of this absorption were first made at the wavelengths of yellow sodium lines and showed that the atmosphere of the planet contains sodium; now, other elements can be measured as well.

Try a **transit simulator (https://openstaxcollege.org/l/30transimul)** that demonstrates how a planet passing in front of its parent star can lead to the planet's detection. Follow the instructions to run the animation on your computer.

Transiting planets reveal such a wealth of information that the French Space Agency (CNES) and the European Space Agency (ESA) launched the CoRoT space telescope in 2007 to detect transiting exoplanets. CoRoT discovered 32 transiting exoplanets, including the first transiting planet with a size and density similar to Earth. In 2012, the spacecraft suffered an onboard computer failure, ending the mission. Meanwhile, NASA built a much more powerful transit observatory called Kepler.

In 2009, NASA launched the Kepler space telescope, dedicated to the discovery of transiting exoplanets. This spacecraft stared continuously at more than 150,000 stars in a small patch of sky near the constellation of Cygnus—just above the plane of our Milky Way Galaxy (**Figure 20.11**). Kepler's cameras and ability to measure small changes in brightness very precisely enabled the discovery of thousands of exoplanets, including many multi-planet systems. The spacecraft required three reaction wheels—a type of wheel used to help control slight rotation of the spacecraft—to stabilize the pointing of the telescope and monitor the brightness of the same group of stars over and over again. Kepler was launched with four reaction wheels (one a spare), but by May 2013, two wheels had failed and the telescope could no longer be accurately pointed toward the target area. Kepler had been designed to operate for 4 years, and ironically, the pointing failure occurred exactly 4 years and 1 day after it began observing.

However, this failure did not end the mission. The Kepler telescope continued to observe for two more years, looking for short-period transits in different parts of the sky. A new NASA mission called TESS (Transiting Exoplanet Survey Satellite) will carry out a survey all over the sky of the nearer (and therefore brighter) stars, starting in 2018.

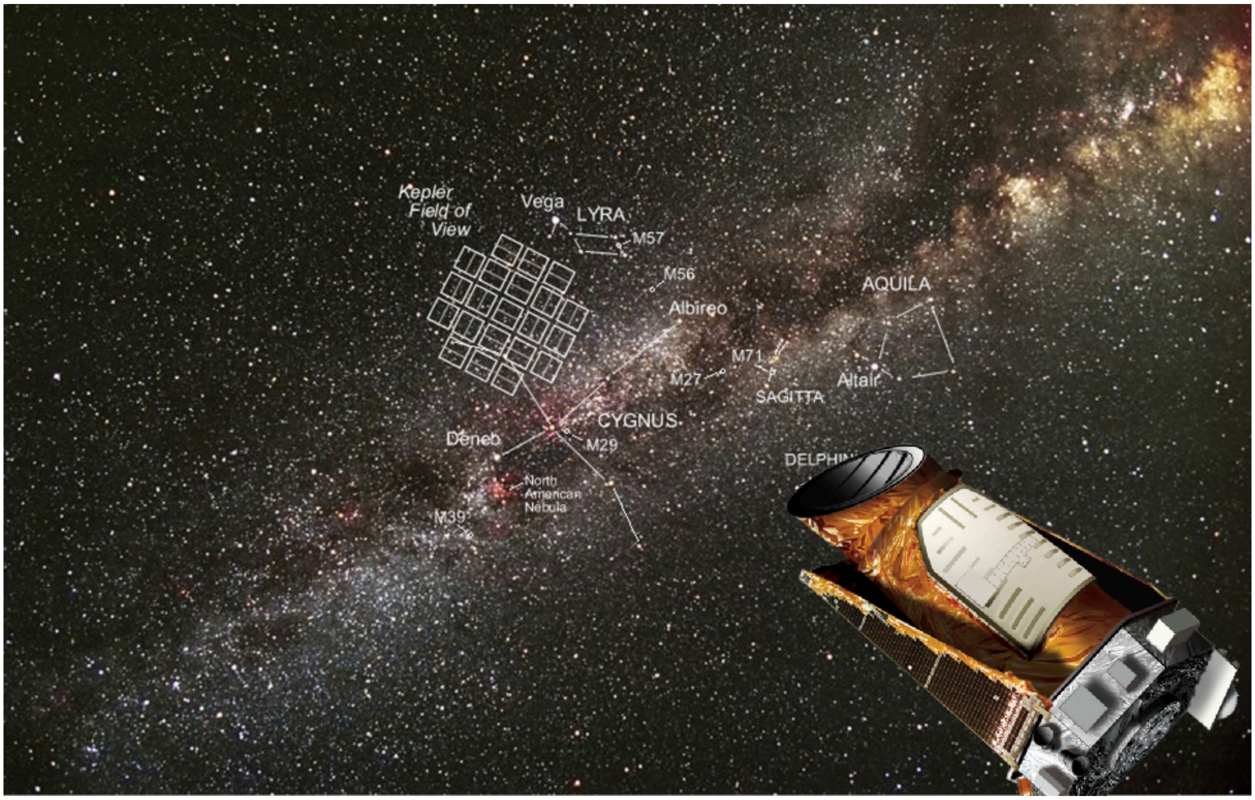


Figure 20.11 The boxes show the region where the Kepler spacecraft cameras took images of over 150,000 stars regularly, to find transiting planets. (credit “field of view”: modification of work by NASA/Kepler mission; credit “spacecraft”: modification of work by NASA/Kepler mission/Wendy Stenzel)

What do we mean, exactly, by “discovery” of transiting exoplanets? A single transit shows up as a very slight drop in the brightness of the star, lasting several hours. However, astronomers must be on guard against other factors that might produce a false transit, especially when working at the limit of precision of the telescope. We must wait for a second transit of similar depth. But when another transit is observed, we don’t initially know whether it might be due to another planet in a different orbit. The “discovery” occurs only when a third transit is found with similar depth and the same spacing in time as the first pair.

Computers normally conduct the analysis, which involves searching for tiny, periodic dips in the light from each star, extending over 4 years of observation. But the Kepler mission also has a program in which non-astronomers—citizen scientists—can examine the data. These dedicated volunteers have found several transits that were missed by the computer analyses, showing that the human eye and brain sometimes recognize unusual events that a computer was not programmed to look for.

Measuring three or four evenly spaced transits is normally enough to “discover” an exoplanet. But in a new field like exoplanet research, we would like to find further independent verification. The strongest confirmation happens when ground-based telescopes are also able to detect a Doppler shift with the same period as the transits. However, this is generally not possible for Earth-size planets. One of the most convincing ways to verify that a dip in brightness is due to a planet is to find more planets orbiting the same star—a *planetary system*. Multi-planet systems also provide alternative ways to estimate the masses of the planets, as we will discuss in the next section.

The selection effects (or biases) in the Kepler data are similar to those in Doppler observations. Large planets are easier to find than small ones, and short-period planets are easier than long-period planets. If we require three transits to establish the presence of a planet, we are of course limited to discovering planets with orbital periods less than one-third of the observing interval. Thus, it was only in its fourth and final year of operation that Kepler was able to find planets with orbits like Earth’s that require 1 year to go around their star.

Direct Detection

The best possible evidence for an earthlike planet elsewhere would be an image. After all, "seeing is believing" is a very human prejudice. But imaging a distant planet is a formidable challenge indeed. Suppose, for example, you were a great distance away and wished to detect reflected light from Earth. Earth intercepts and reflects less than one billionth of the Sun's radiation, so its apparent brightness in visible light is less than one billionth that of the Sun. Compounding the challenge of detecting such a faint speck of light, the planet is swamped by the blaze of radiation from its parent star.

Even today, the best telescope mirrors' optics have slight imperfections that prevent the star's light from coming into focus in a completely sharp point.

Direct imaging works best for young gas giant planets that emit infrared light and reside at large separations from their host stars. Young giant planets emit more infrared light because they have more internal energy, stored from the process of planet formation. Even then, clever techniques must be employed to subtract out the light from the host star. In 2008, three such young planets were discovered orbiting HR 8799, a star in the constellation of Pegasus (**Figure 20.12**). Two years later, a fourth planet was detected closer to the star. Additional planets may reside even closer to HR 8799, but if they exist, they are currently lost in the glare of the star.

Since then, a number of planets around other stars have been found using direct imaging. However, one challenge is to tell whether the objects we are seeing are indeed planets or if they are brown dwarfs (failed stars) in orbit around a star.

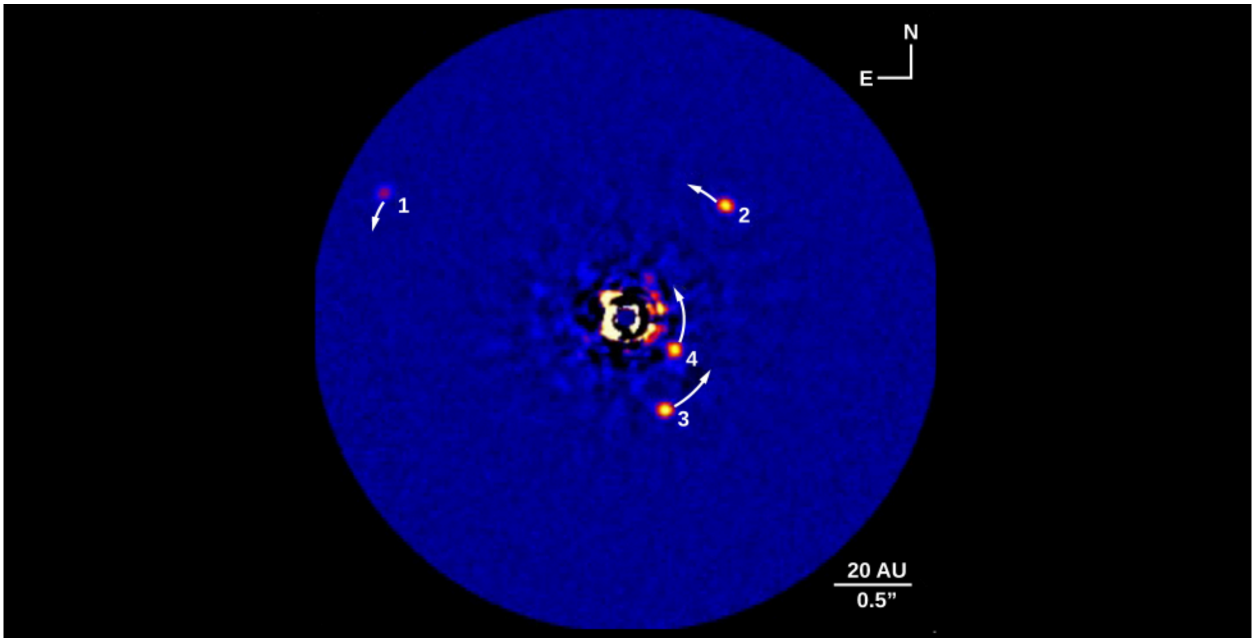


Figure 20.12 This image shows Keck telescope observations of four directly imaged planets orbiting HR 8799. A size scale for the system gives the distance in AU (remember that one astronomical unit is the distance between Earth and the Sun.) (credit: modification of work by Ben Zuckerman)

Direct imaging is an important technique for characterizing an exoplanet. The brightness of the planet can be measured at different wavelengths. These observations provide an estimate for the temperature of the planet's atmosphere; in the case of HR 8799 planet 1, the color suggests the presence of thick clouds. Spectra can also be obtained from the faint light to analyze the atmospheric constituents. A spectrum of HR 8799 planet 1 indicates a hydrogen-rich atmosphere, while the closer planet 4 shows evidence for methane in the atmosphere.

Another way to overcome the blurring effect of Earth's atmosphere is to observe from space. Infrared may be the optimal wavelength range in which to observe because planets get brighter in the infrared while stars like our Sun get fainter, thereby making it easier to detect a planet against the glare of its star. Special optical techniques can be used to suppress the light from the central star and make it easier to see the planet itself. However, even if we go into space, it will be difficult to obtain images of Earth-size planets.

20.2 | Exoplanets Everywhere: What We Are Learning

Learning Objectives

By the end of this section, you will be able to:

- Explain what we have learned from our discovery of exoplanets
- Identify which kind of exoplanets appear to be the most common in the Galaxy
- Discuss the kinds of planetary systems we are finding around other stars

Before the discovery of exoplanets, most astronomers expected that other planetary systems would be much like our own—planets following roughly circular orbits, with the most massive planets several AU from their parent star. Such systems do exist in large numbers, but many exoplanets and planetary systems are very different from those in our solar system. Another surprise is the existence of whole classes of exoplanets that we simply don't have in our solar system: planets with masses between the mass of Earth and Neptune, and planets that are several times more massive than Jupiter.

Kepler Results

The Kepler telescope has been responsible for the discovery of most exoplanets, especially at smaller sizes, as illustrated in **Figure 20.13**, where the Kepler discoveries are plotted in yellow. You can see the wide range of sizes, including planets substantially larger than Jupiter and smaller than Earth. The absence of Kepler-discovered exoplanets with orbital periods longer than a few hundred days is a consequence of the 4-year lifetime of the mission. (Remember that three evenly spaced transits must be observed to register a discovery.) At the smaller sizes, the absence of planets much smaller than one earth radius is due to the difficulty of detecting transits by very small planets. In effect, the "discovery space" for Kepler was limited to planets with orbital periods less than 400 days and sizes larger than Mars.

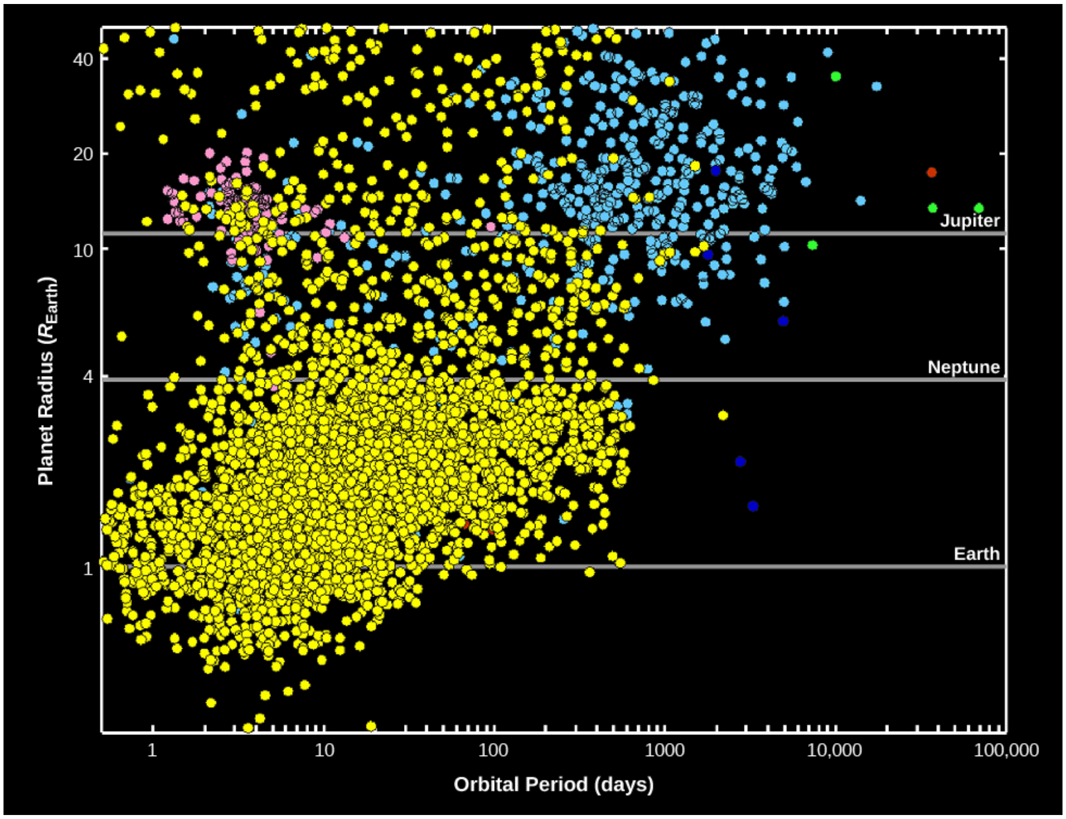


Figure 20.13 The vertical axis shows the radius of each planet compared to Earth. Horizontal lines show the size of Earth, Neptune, and Jupiter. The horizontal axis shows the time each planet takes to make one orbit (and is given in Earth days). Recall that Mercury takes 88 days and Earth takes a little more than 365 days to orbit the Sun. The yellow and red dots show planets discovered by transits, and the blue dots are the discoveries by the radial velocity (Doppler) technique. (credit: modification of work by NASA/Kepler mission)

One of the primary objectives of the Kepler mission was to find out how many stars hosted planets and especially to estimate the frequency of earthlike planets. Although Kepler looked at only a very tiny fraction of the stars in the Galaxy, the sample size was large enough to draw some interesting conclusions. While the observations apply only to the stars observed by Kepler, those stars are reasonably representative, and so astronomers can extrapolate to the entire Galaxy.

Figure 20.14 shows that the Kepler discoveries include many rocky, Earth-size planets, far more than Jupiter-size gas planets. This immediately tells us that the initial Doppler discovery of many hot Jupiters was a biased sample, in effect, finding the odd planetary systems because they were the easiest to detect. However, there is one huge difference between this observed size distribution and that of planets in our solar system. The most common planets have radii between 1.4 and 2.8 that of Earth, sizes for which we have no examples in the solar system. These have been nicknamed **super-Earths**, while the other large group with sizes between 2.8 and 4 that of Earth are often called **mini-Neptunes**.

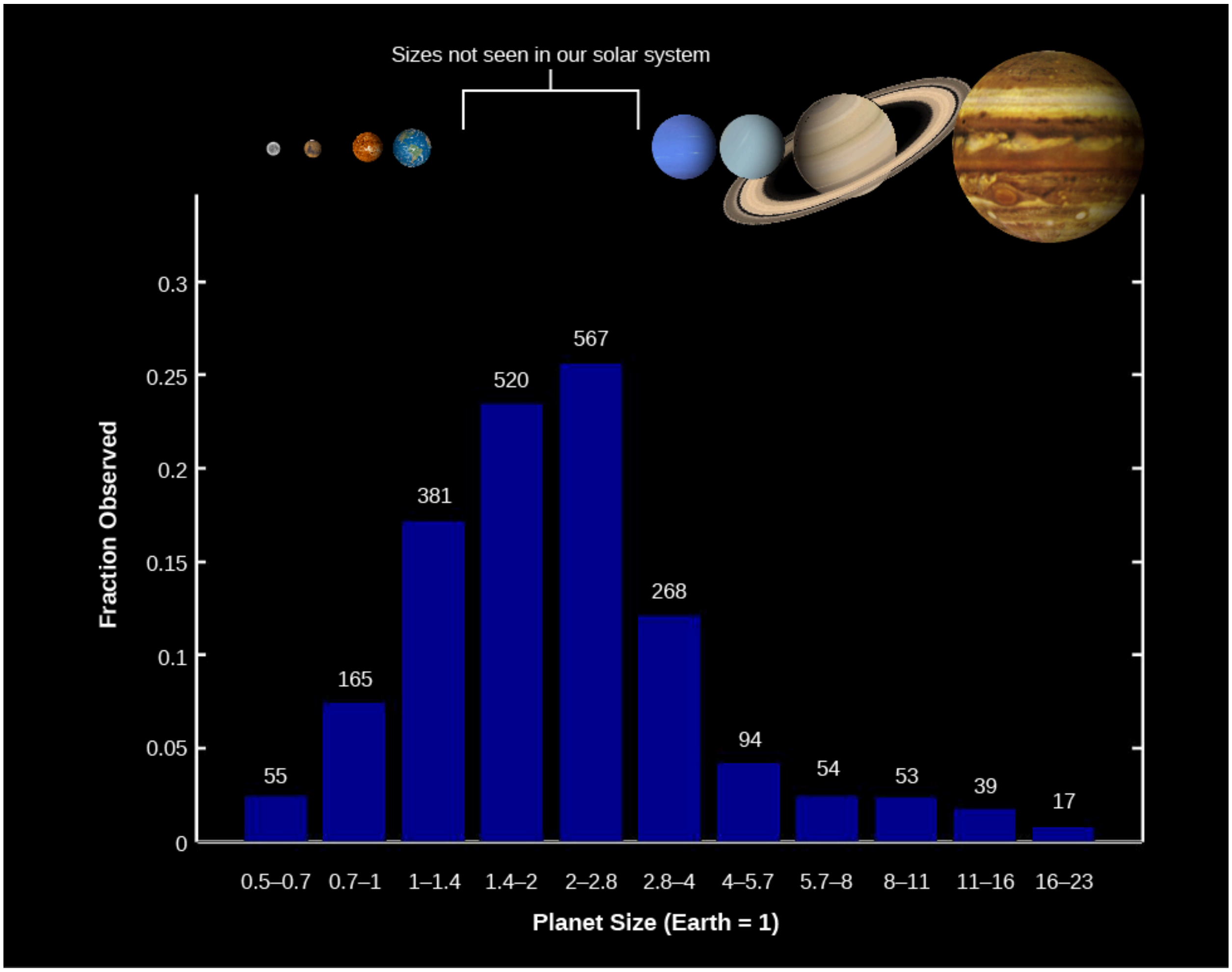


Figure 20.14 This bar graph shows the number of planets of each size range found among the first 2213 Kepler planet discoveries. Sizes range from half the size of Earth to 20 times that of Earth. On the vertical axis, you can see the fraction that each size range makes up of the total. Note that planets that are between 1.4 and 4 times the size of Earth make up the largest fractions, yet this size range is not represented among the planets in our solar system. (credit: modification of work by NASA/Kepler mission)

What a remarkable discovery it is that the most common types of planets in the Galaxy are completely absent from our solar system and were unknown until Kepler's survey. However, recall that really small planets were difficult for the Kepler instruments to find. So, to estimate the frequency of Earth-size exoplanets, we need to correct for this sampling bias. The result is the corrected size distribution shown in **Figure 20.15**. Notice that in this graph, we have also taken the step of showing not the number of Kepler detections but the average number of planets per star for solar-type stars (spectral types F, G, and K).

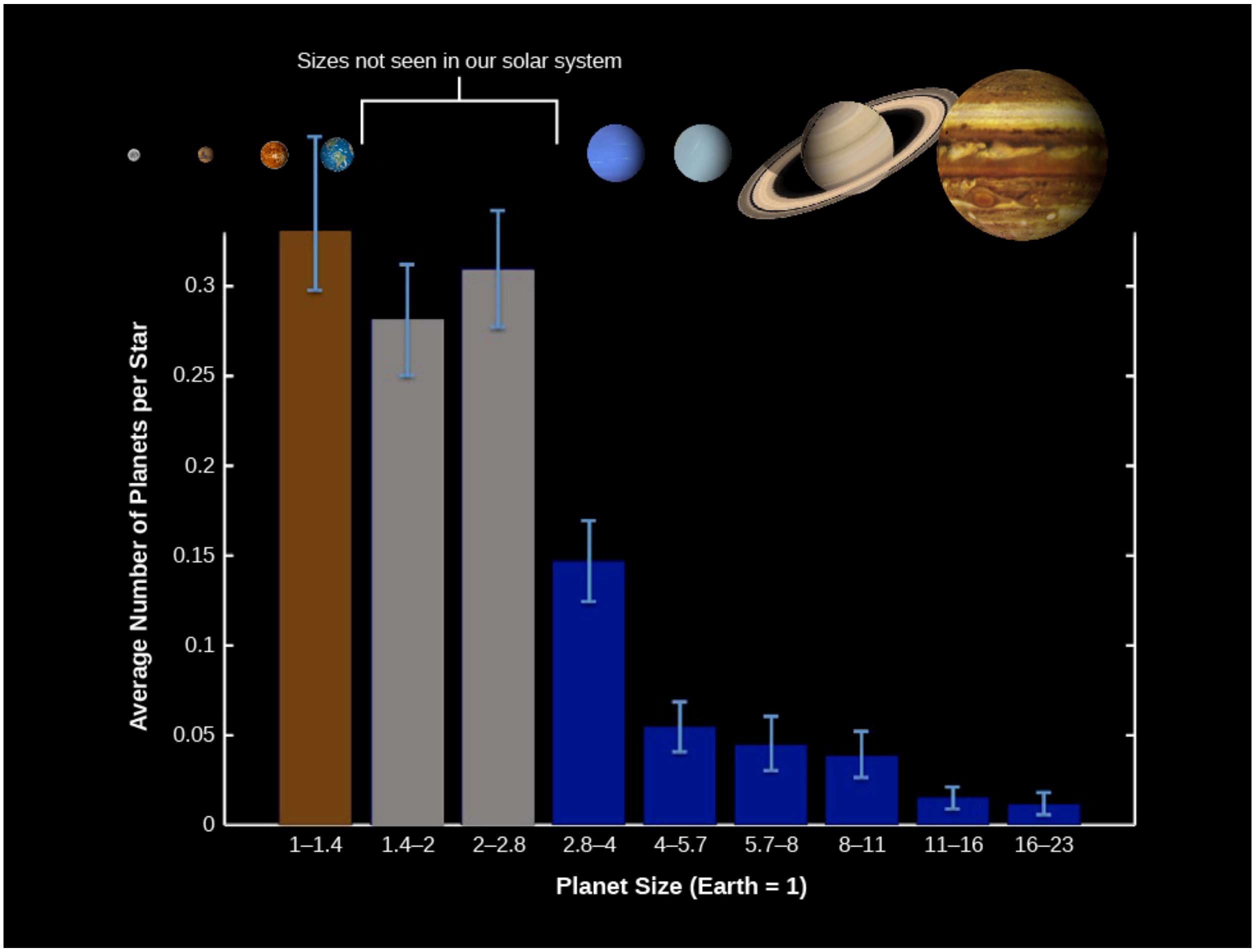


Figure 20.15 We show the average number of planets per star in each planet size range. (The average is less than one because some stars will have zero planets of that size range.) This distribution, corrected for biases in the Kepler data, shows that Earth-size planets may actually be the most common type of exoplanets. (credit: modification of work by NASA/Kepler mission)

We see that the most common planet sizes of are those with radii from 1 to 3 times that of Earth—what we have called "Earths" and "super-Earths." Each group occurs in about one-third to one-quarter of stars. In other words, if we group these sizes together, we can conclude there is nearly one such planet per star! And remember, this census includes primarily planets with orbital periods less than 2 years. We do not yet know how many undiscovered planets might exist at larger distances from their star.

To estimate the number of Earth-size planets in our Galaxy, we need to remember that there are approximately 100 billion stars of spectral types F, G, and K. Therefore, we estimate that there are about 30 billion Earth-size planets in our Galaxy. If we include the super-Earths too, then there could be one hundred billion in the whole Galaxy. This idea—that planets of roughly Earth's size are so numerous—is surely one of the most important discoveries of modern astronomy.

Planets with Known Densities

For several hundred exoplanets, we have been able to measure both the size of the planet from transit data and its mass from Doppler data, yielding an estimate of its density. Comparing the average density of exoplanets to the density of planets in our solar system helps us understand whether they are rocky or gaseous in nature. This has been particularly important for understanding the structure of the new categories of super-Earths and mini-Neptunes with masses between 3–10 times the mass of Earth. A key observation so far is that planets that are more than 10 times the mass of Earth have substantial gaseous envelopes (like Uranus and Neptune) whereas lower-mass planets are predominately rocky in nature (like the terrestrial planets).

Figure 20.16 compares all the exoplanets that have both mass and radius measurements. The dependence of the radius on planet mass is also shown for a few illustrative cases—hypothetical planets made of pure iron, rock, water, or hydrogen.

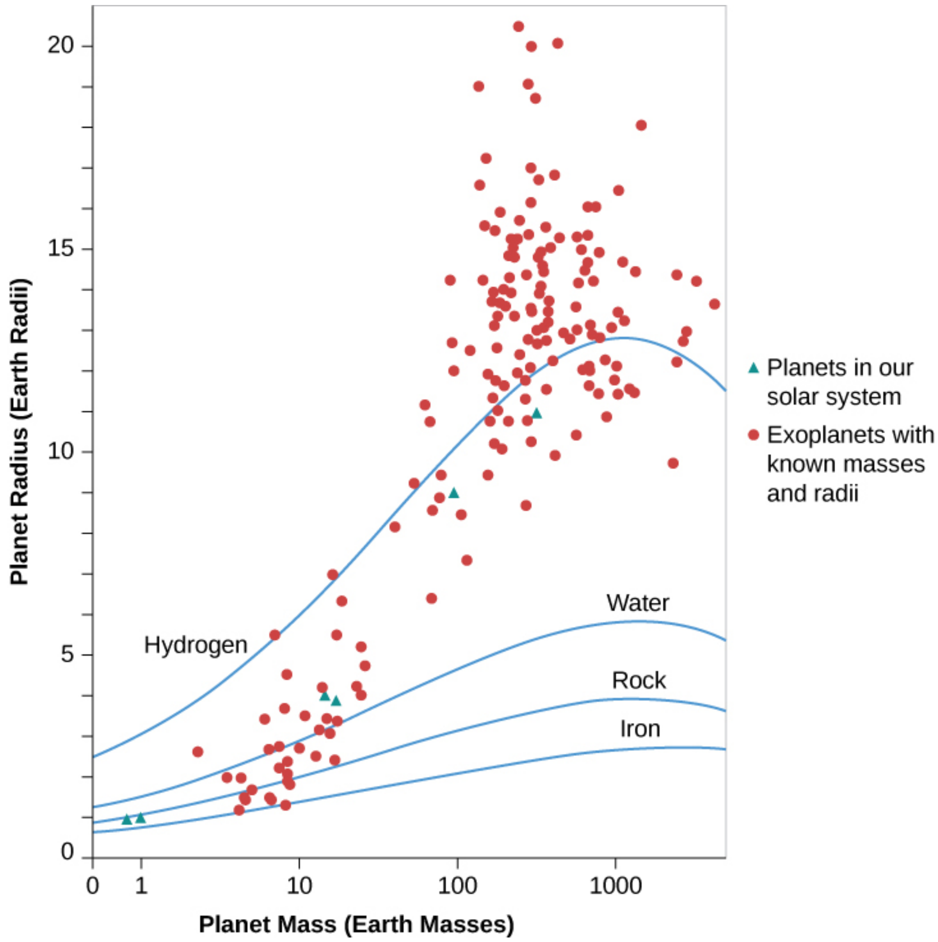


Figure 20.16 Exoplanets with known masses and radii (red circles) are plotted along with solid lines that show the theoretical size of pure iron, rock, water, and hydrogen planets with increasing mass. Masses are given in multiples of Earth's mass. (For comparison, Jupiter contains enough mass to make 320 Earths.) The green triangles indicate planets in our solar system.

At lower masses, notice that as the mass of these hypothetical planets increases, the radius also increases. That makes sense—if you were building a model of a planet out of clay, your toy planet would increase in size as you added more clay. However, for the highest mass planets ($M > 1000\ M_{\text{Earth}}$) in **Figure 20.16**, notice that the radius stops increasing and the planets with greater mass are actually smaller. This occurs because increasing the mass also increases the gravity of the planet, so that compressible materials (even rock is compressible) will become more tightly packed, shrinking the size of the more massive planet.

In reality, planets are not pure compositions like the hypothetical water or iron planet. Earth is composed of a solid iron core, an outer liquid-iron core, a rocky mantle and crust, and a relatively thin atmospheric layer. Exoplanets are similarly likely to be differentiated into compositional layers. The theoretical lines in **Figure 20.16** are simply guides that suggest a range of possible compositions.

Astronomers who work on the complex modeling of the interiors of rocky planets make the simplifying assumption that the planet consists of two or three layers. This is not perfect, but it is a reasonable approximation and another good example of how science works. Often, the first step in understanding something new is to narrow down the range of possibilities. This sets the stage for refining and deepening our knowledge. In **Figure 20.16**, the two green triangles with roughly 1 M_{Earth} and 1 R_{Earth} represent Venus and Earth. Notice that these planets fall between the models for a pure iron and a pure rock planet, consistent with what we would expect for the known mixed-chemical composition of Venus and Earth.

In the case of gaseous planets, the situation is more complex. Hydrogen is the lightest element in the periodic table, yet many of the detected exoplanets in **Figure 20.16** with masses greater than 100 M_{Earth} have radii that suggest they are lower in density than a pure hydrogen planet. Hydrogen is the lightest element, so what is happening here? Why do some gas giant planets have inflated radii that are larger than the fictitious pure hydrogen planet? Many of these planets reside in short-period orbits close to the host star where they intercept a significant amount of radiated energy. If this energy is trapped deep in the planet atmosphere, it can cause the planet to expand.

Planets that orbit close to their host stars in slightly eccentric orbits have another source of energy: the star will raise tides in these planets that tend to circularize the orbits. This process also results in tidal dissipation of energy that can inflate the atmosphere. It would be interesting to measure the size of gas giant planets in wider orbits where the planets should be cooler—the expectation is that unless they are very young, these cooler gas giant exoplanets (sometimes called “cold Jupiters”) should not be inflated. But we don’t yet have data on these more distant exoplanets.

Exoplanetary Systems

As we search for exoplanets, we don’t expect to find only one planet per star. Our solar system has eight major planets, half a dozen dwarf planets, and millions of smaller objects orbiting the Sun. The evidence we have of planetary systems in formation also suggest that they are likely to produce multi-planet systems.

The first planetary system was found around the star Upsilon Andromedae in 1999 using the Doppler method, and many others have been found since then (about 2600 as of 2016). If such exoplanetary system are common, let’s consider which systems we expect to find in the Kepler transit data.

A planet will transit its star only if Earth lies in the plane of the planet’s orbit. If the planets in other systems do not have orbits in the same plane, we are unlikely to see multiple transiting objects. Also, as we have noted before, Kepler was sensitive only to planets with orbital periods less than about 4 years. What we expect from Kepler data, then, is evidence of coplanar planetary systems confined to what would be the realm of the terrestrial planets in our solar system.

By 2018, astronomers gathered data on nearly 3000 such exoplanet systems. Many have only two known planets, but a few have as many as five, and one has eight (the same number of planets as our own solar system). For the most part, these are very compact systems with most of their planets closer to their star than Mercury is to the Sun. The figure below shows one of the largest exoplanet systems: that of the star called Kepler-62 (Figure 20.17). Our solar system is shown to the same scale, for comparison (note that the Kepler-62 planets are drawn with artistic license; we have no detailed images of any exoplanets).

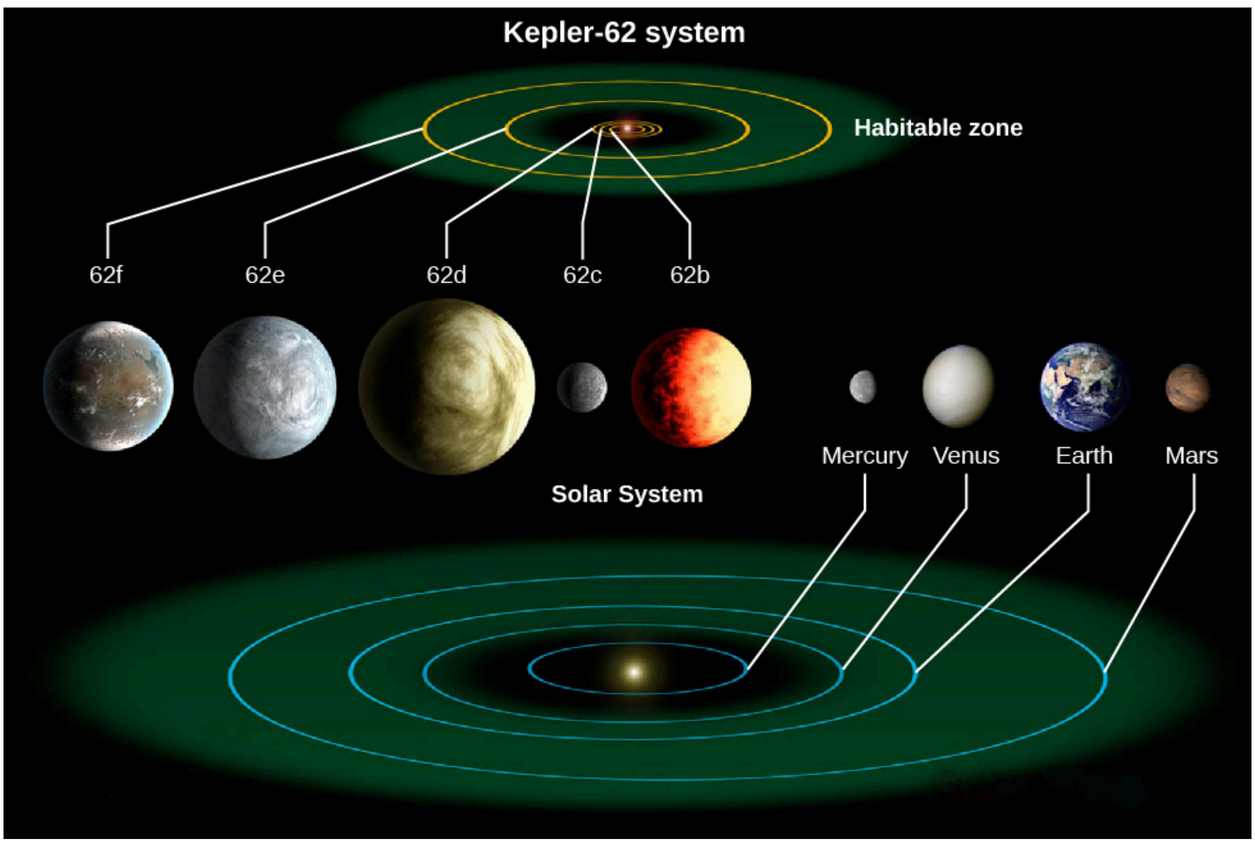


Figure 20.17 The green areas are the “habitable zones,” the range of distance from the star where surface temperatures are likely to be consistent with liquid water. (credit: modification of work by NASA/Ames/JPL-Caltech)

All but one of the planets in the K-62 system are larger than Earth. These are super-Earths, and one of them (62d) is in the size range of a mini-Neptune, where it is likely to be largely gaseous. The smallest planet in this system is about the size of Mars. The three inner planets orbit very close to their star, and only the outer two have orbits larger than Mercury in our system. The green areas represent each star's "habitable zone," which is the distance from the star where we calculate that surface temperatures would be consistent with liquid water. The Kepler-62 habitable zone is much smaller than that of the Sun because the star is intrinsically fainter.

With closely spaced systems like this, the planets can interact gravitationally with each other. The result is that the observed transits occur a few minutes earlier or later than would be predicted from simple orbits. These gravitational interactions have allowed the Kepler scientists to calculate masses for the planets, providing another way to learn about exoplanets.

Kepler has discovered some interesting and unusual planetary systems. For example, most astronomers expected planets to be limited to single stars. But we have found planets orbiting close double stars, so that the planet would see two suns in its sky, like those of the fictional planet Tatooine in the *Star Wars* films. At the opposite extreme, planets can orbit one star of a wide, double-star system without major interference from the second star.

20.3 | New Perspectives on Planet Formation

Learning Objectives

By the end of this section, you will be able to:

- Explain how exoplanet discoveries have revised our understanding of planet formation
- Discuss how planetary systems quite different from our solar system might have come about

Traditionally, astronomers have assumed that the planets in our solar system formed at about their current distances from the Sun and have remained there ever since. The first step in the formation of a giant planet is to build up a solid core, which happens when planetesimals collide and stick. Eventually, this core becomes massive enough to begin sweeping up gaseous material in the disk, thereby building the gas giants Jupiter and Saturn.

How to Make a Hot Jupiter

The traditional model for the formation of planets works only if the giant planets are formed far from the central star (about 5–10 AU), where the disk is cold enough to have a fairly high density of solid matter. It cannot explain the hot Jupiters, which are located very close to their stars where any rocky raw material would be completely vaporized. It also cannot explain the elliptical orbits we observe for some exoplanets because the orbit of a protoplanet, whatever its initial shape, will quickly become circular through interactions with the surrounding disk of material and will remain that way as the planet grows by sweeping up additional matter.

So we have two options: either we find a new model for forming planets close to the searing heat of the parent star, or we find a way to change the orbits of planets so that cold Jupiters can travel inward *after* they form. Most research now supports the latter explanation.

Calculations show that if a planet forms while a substantial amount of gas remains in the disk, then some of the planet's orbital angular momentum can be transferred to the disk. As it loses momentum (through a process that reminds us of the effects of friction), the planet will spiral inward. This process can transport giant planets, initially formed in cold regions of the disk, closer to the central star—thereby producing hot Jupiters. Gravitational interactions between planets in the chaotic early solar system can also cause planets to slingshot inward from large distances. But for this to work, the other planet has to carry away the angular momentum and move to a more distant orbit.

In some cases, we can use the combination of transit plus Doppler measurements to determine whether the planets orbit in the same plane and in the same direction as the star. For the first few cases, things seemed to work just as we anticipated: like the solar system, the gas giant planets orbited in their star's equatorial plane and in the same direction as the spinning star.

Then, some startling discoveries were made of gas giant planets that orbited at right angles or even in the opposite sense as the spin of the star. How could this happen? Again, there must have been interactions between planets. It's possible that before the system settled down, two planets came close together, so that one was kicked into an usual orbit. Or perhaps a passing star perturbed the system after the planets were newly formed.

Forming Planetary Systems

When the Milky Way Galaxy was young, the stars that formed did not contain many heavy elements like iron. Several generations of star formation and star death were required to enrich the interstellar medium for subsequent generations of stars. Since planets seem to form "inside out," starting with the accretion of the materials that can make the rocky cores with which planets start, astronomers wondered when in the history of the Galaxy, planet formation would turn on.

The star Kepler-444 has shed some light on this question. This is a tightly packed system of five planets—the smallest comparable in size to Mercury and the largest similar in size to Venus. All five planets were detected with the Kepler spacecraft as they transited their parent star. All five planets orbit their host star in less than the time it takes Mercury to complete one orbit about the Sun. Remarkably, the host star Kepler-444 is more than 11 billion years old and formed when the Milky Way was only 2 billion years old. So the heavier elements needed to make rocky planets must have already been available then. This ancient planetary system sets the clock on the beginning of rocky planet formation to be relatively soon after the formation of our Galaxy.

Kepler data demonstrate that while rocky planets inside Mercury's orbit are missing from our solar system, they are common around other stars, like Kepler-444. When the first systems packed with close-in rocky planets were discovered, we wondered why they were so different from our solar system. When many such systems were discovered, we began to wonder if it was our solar system that was different. This led to speculation that additional rocky planets might once have existed close to the Sun in our solar system.

There is some evidence from the motions in the outer solar system that Jupiter may have migrated inward long ago. If correct, then gravitational perturbations from Jupiter could have dislodged the orbits of close-in rocky planets, causing them to fall into the Sun. Consistent with this picture, astronomers now think that Uranus and Neptune probably did not form at their present distances from the Sun but rather closer to where Jupiter and Saturn are now. The reason for this idea is that density in the disk of matter surrounding the Sun at the time the planets formed was so low outside the orbit of Saturn that it would take several billion years to build up Uranus and Neptune. Yet we saw earlier in the chapter that the disks around protostars survive only a few million years.

Therefore, scientists have developed computer models demonstrating that Uranus and Neptune could have formed near the current locations of Jupiter and Saturn, and then been kicked out to larger distances through gravitational interactions with their neighbors. All these wonderful new observations illustrate how dangerous it can be to draw conclusions about a phenomenon in science (in this case, how planetary systems form and arrange themselves) when you are only working with a single example.

Exoplanets have given rise to a new picture of planetary system formation—one that is much more chaotic than we originally thought. If we think of the planets as being like skaters in a rink, our original model (with only our own solar system as a guide) assumed that the planets behaved like polite skaters, all obeying the rules of the rink and all moving in nearly the same direction, following roughly circular paths. The new picture corresponds more to a roller derby, where the skaters crash into one another, change directions, and sometimes are thrown entirely out of the rink.

Habitable Exoplanets

While thousands of exoplanets have been discovered in the past two decades, every observational technique has fallen short of finding more than a few candidates that resemble Earth (**Figure 20.18**). Astronomers are not sure exactly what properties would define another Earth. Do we need to find a planet that is *exactly* the same size and mass as Earth? That may be difficult and may not be important from the perspective of habitability. After all, we have no reason to think that life could not have arisen on Earth if our planet had been a little bit smaller or larger. And, remember that how habitable a planet is depends on both its distance from its star and the nature of its atmosphere. The greenhouse effect can make some planets warmer (as it did for Venus and is doing more and more for Earth).

Figure 20.18 This painting, commissioned by NASA, conveys the idea that there may be many planets resembling Earth out there as our methods for finding them improve. (credit: NASA/JPL-Caltech/R. Hurt (SSC-Caltech))

We can ask other questions to which we don't yet know the answers. Does this "twin" of Earth need to orbit a solar-type star, or can we consider as candidates the numerous exoplanets orbiting K- and M-class stars? (In the summer of 2016, astronomers reported the discovery of a planet with at least 1.3 times the mass of Earth around the nearest star, Proxima Centauri, which is spectral type M and located 4.2 light years from us.) We have a special interest in finding planets that could support life like ours, in which case, we need to find exoplanets within their star's habitable zone, where surface temperatures are consistent with liquid water on the surface. This is probably the most important characteristic defining an Earth-analog exoplanet.

The search for potentially habitable worlds is one of the prime drivers for exoplanet research in the next decade. Astronomers are beginning to develop realistic plans for new instruments that can even look for signs of life on distant worlds (examining their atmospheres for gases associated with life, for example). If we require telescopes in space to find such worlds, we need to recognize that years are required to plan, build, and launch such space observatories. The discovery of exoplanets and the knowledge that most stars have planetary systems are transforming our thinking about life beyond Earth. We are closer than ever to knowing whether habitable (and inhabited) planets are common. This work lends a new spirit of optimism to the search for life elsewhere in the universe.

Check out the habitability of various stars and planets by trying out the interactive **Circumstellar Habitable Zone Simulator (https://openstaxcollege.org/l/30cirhabzonsim)** and select a star system to investigate.

For Further Exploration

Websites

Exoplanet Exploration: **http://planetquest.jpl.nasa.gov/ (http://planetquest.jpl.nasa.gov/)** . PlanetQuest (from the Navigator Program at the Jet Propulsion Lab) is probably the best site for students and beginners, with introductory materials and nice illustrations; it focuses mostly on NASA work and missions.

Exoplanets: **http://www.planetary.org/exoplanets/ (http://www.planetary.org/exoplanets/)** . Planetary Society's exoplanets pages with a dynamic catalog of planets found and good explanations.

Exoplanets: The Search for Planets beyond Our Solar System: **http://www.iop.org/publications/iop/2010/page_42551.html (http://www.iop.org/publications/iop/2010/page_42551.html)** . From the British Institute of Physics in 2010.

Extrasolar Planets Encyclopedia: **http://exoplanet.eu/ (http://exoplanet.eu/)** . Maintained by Jean Schneider of the Paris Observatory, has the largest catalog of planet discoveries and useful background material (some of it more technical).

Kepler Mission: **http://kepler.nasa.gov/ (http://kepler.nasa.gov/)** . The public website for the remarkable telescope in space that is searching planets using the transit technique and is our best hope for finding earthlike planets.

Proxima Centauri Planet Discovery: **http://www.eso.org/public/news/eso1629/ (http://www.eso.org/public/news/eso1629/)** .

Apps

Exoplanet: **http://itunes.apple.com/us/app/exoplanet/id327702034?mt=8 (http://itunes.apple.com/us/app/exoplanet/id327702034?mt=8)** . Allows you to browse through a regularly updated visual catalog of exoplanets that have been found so far.

Journey to the Exoplanets: **http://itunes.apple.com/us/app/journey-to-the-exoplanets/id463532472?mt=8 (http://itunes.apple.com/us/app/journey-to-the-exoplanets/id463532472?mt=8)** . Produced by the staff of *Scientific American*, with input from scientists and space artists; gives background information and visual tours of the nearer star systems with planets.

Videos

Are We Alone: An Evening Dialogue with the Kepler Mission Leaders: **http://www.youtube.com/watch?v=O7ItAXfl0Lw (http://www.youtube.com/watch?v=O7ItAXfl0Lw)** . A non-technical panel discussion on Kepler results and ideas about planet formation with Bill Borucki, Natalie Batalha, and Gibor Basri (moderated by Andrew Fraknoi) at the University of California, Berkeley (2:07:01).

Finding the Next Earth: The Latest Results from Kepler: **https://www.youtube.com/watch?v=ZbijeR_AALo (https://www.youtube.com/watch?v=ZbijeR_AALo)** . Natalie Batalha (San Jose State University & NASA Ames) public talk in the Silicon Valley Astronomy Lecture Series (1:28:38).

From Hot Jupiters to Habitable Worlds: **https://vimeo.com/37696087 (https://vimeo.com/37696087)** (Part 1) and **https://vimeo.com/37700700 (https://vimeo.com/37700700)** (Part 2). Debra Fischer (Yale University) public talk in Hawaii sponsored by the Keck Observatory (15:20 Part 1, 21:32 Part 2).

Search for Habitable Exoplanets: **http://www.youtube.com/watch?v=RLWb_T9yaDU (http://www.youtube.com/watch?v=RLWb_T9yaDU)** . Sara Seeger (MIT) public talk at the SETI Institute, with Kepler results (1:10:35).

Strange Planetary Vistas: **http://www.youtube.com/watch?v=_8ww9eLRSCg (http://www.youtube.com/watch?v=_8ww9eLRSCg)** . Josh Carter (CfA) public talk at Harvard’s Center for Astrophysics with a friendly introduction to exoplanets for non-specialists (46:35).

CHAPTER 20 REVIEW

KEY TERMS

exoplanet a planet orbiting a star other than our Sun

exoplanet a planet orbiting a star other than our Sun

mini-Neptune a planet that is intermediate between the largest terrestrial planet in our solar system (Earth) and the smallest jovian planet (Neptune); generally, mini-Neptunes have sizes between 2.8 and 4 times Earth's size

super-Earth a planet larger than Earth, generally between 1.4 and 2.8 times the size of our planet

transit when one astronomical object moves in front of another

KEY EQUATIONS

Average orbital radius of planet	$a = \sqrt[3]{\frac{GM_{\text{star}}}{4\pi^2}T^2}$
Orbital speed of planet	$v_{\text{planet}} = \frac{2\pi a}{T}$
Total momentum of star-planet system is zero	$M_{\text{planet}}v_{\text{planet}} = M_{\text{star}}v_{\text{star}}$
Transit depth	$\left(\frac{R_{\text{planet}}}{R_{\text{star}}}\right)^2$

Insert paragraph text here.

SUMMARY

20.1 Planets Beyond the Solar System: Search and Discovery

- Several observational techniques have successfully detected planets orbiting other stars. These techniques fall into two general categories—direct and indirect detection.
- The Doppler and transit techniques are our most powerful indirect tools for finding exoplanets.
- Some planets are also being found by direct imaging.

20.2 Exoplanets Everywhere: What We Are Learning

- Although the Kepler mission is finding thousands of new exoplanets, these are limited to orbital periods of less than 400 days and sizes larger than Mars. Still, we can use the Kepler discoveries to extrapolate the distribution of planets in our Galaxy.
- The data so far imply that planets like Earth are the most common type of planet, and that there may be 100 billion Earth-size planets around Sun-like stars in the Galaxy.
- About 2600 planetary systems have been discovered around other stars. In many of them, planets are arranged differently than in our solar system.

20.3 New Perspectives on Planet Formation

- The ensemble of exoplanets is incredibly diverse and has led to a revision in our understanding of planet formation that includes the possibility of vigorous, chaotic interactions, with planet migration and scattering.
- It is possible that the solar system is unusual (and not representative) in how its planets are arranged. Many systems seem to have rocky planets farther inward than we do, for example, and some even have "hot Jupiters" very close to their star.

- Ambitious space experiments should make it possible to image earthlike planets outside the solar system and even to obtain information about their habitability as we search for life elsewhere.

CONCEPTUAL QUESTIONS

20.1 Planets Beyond the Solar System: Search and Discovery

1. Why did it take astronomers until 1995 to discover the first exoplanet orbiting another star like the Sun?

2. Which types of planets are most easily detected by Doppler measurements? By transits?

3. List three ways in which the exoplanets we have detected have been found to be different from planets in our solar system.

4. List any similarities between discovered exoplanets and planets in our solar system.

5. Suppose you wanted to observe a planet around another star with direct imaging. Would you try to observe in visible light or in the infrared? Why? Would the planet be easier to see if it were at 1 AU or 5 AU from its star?

6. Why were giant planets close to their stars the first ones to be discovered? Why has the same technique not been used yet to discover giant planets at the distance of Saturn?

7. Exoplanets in eccentric orbits experience large temperature swings during their orbits. Suppose you had to plan for a mission to such a planet. Based on Kepler's second law, does the planet spend more time closer or farther from the star? Explain.

20.2 Exoplanets Everywhere: What We Are Learning

8. List three ways in which the exoplanets we have detected have been found to be different from planets in our solar system.

9. List any similarities between discovered exoplanets and planets in our solar system.

10. What revisions to the theory of planet formation have astronomers had to make as a result of the discovery of exoplanets?

11. Why were giant planets close to their stars the first ones to be discovered? Why has the same technique not been used yet to discover giant planets at the distance of Saturn?

PROBLEMS

20.1 Planets Beyond the Solar System: Search and Discovery

12. When astronomers found the first giant planets with orbits of only a few days, they did not know whether those planets were gaseous and liquid like Jupiter or rocky like Mercury. The observations of HD 209458 settled this question because observations of the transit of the star by this planet made it possible to determine the radius of the planet. Use the data given in the text to estimate the density of this planet, and then use that information to explain why it must be a gas giant.

13. An exoplanetary system has two known planets. Planet X orbits in 290 days and Planet Y orbits in 145 days. Which planet is closest to its host star? If the star has the same mass as the Sun, what is the semi-major axis of the orbits for Planets X and Y?

14. Kepler's third law says that the orbital period (in years) is proportional to the square root of the cube of the mean distance (in AU) from the Sun ($P \propto a^{1.5}$). For mean distances from 0.1 to 32 AU, calculate and plot a curve showing the expected Keplerian period. For each planet in our solar system, look up the mean distance from the Sun in AU and the orbital period in years and overplot these data on the theoretical Keplerian curve.

15. Suppose that a new planet is found orbiting a distant star of mass $4M_{\text{sun}}$ with an orbital period of 200 days.

a. What is the semimajor axis of the planet's orbit?
b. If the peak Doppler shift detected for the star is 50 m/s, what is the planet's mass?

16. Calculate the transit depth for an M dwarf star that is 0.3 times the radius of the Sun with a gas giant planet the size of Jupiter.

17. If a transit depth of 0.00001 can be detected with the

Kepler spacecraft, what is the smallest planet that could be detected around a 0.3 R_{sun} M dwarf star?

20.2 Exoplanets Everywhere: What We Are Learning

18. The NASA Kepler mission discovered a planet orbiting the star HD219666. It has a mass of about 0.052 $M_{Jupiter}$ and a radius of 0.42 $R_{Jupiter}$. What is the density of this planet? Is it terrestrial or jovian in nature? (Youi can find the values of $M_{Jupiter}$ and $R_{Jupiter}$ in **Appendix D**).

21 | THE SUN

Figure 21.1 It takes an incredible amount of energy for the Sun to shine, as it has and will continue to do for billions of years. (credit: modification of work by Ed Dunens)

Chapter Outline

Introduction

The Sun puts out an incomprehensible amount of energy—so much that its ultraviolet radiation can cause sunburns from 93 million miles away. It is also very old. As you learned earlier, evidence shows that the Sun formed about 4.5 billion years ago and has been shining ever since. How can the Sun produce so much energy for so long?

The Sun's energy output is about 4×10^{26} watts. This is unimaginably bright: brighter than a trillion cities together each with a trillion 100-watt light bulbs. Most known methods of generating energy fall far short of the capacity of the Sun. The total amount of energy produced over the entire life of the Sun is staggering, since the Sun has been shining for billions of years. Scientists were unable to explain the seemingly unlimited energy of stars like the Sun prior to the twentieth century.

21.1 | The Structure and Composition of the Sun

Learning Objectives

By the end of this section, you will be able to:

- Explain how the composition of the Sun differs from that of Earth
- Describe the various layers of the Sun and their functions
- Explain what happens in the different parts of the Sun's atmosphere

The Sun, like all stars, is an enormous ball of extremely hot, largely ionized gas, shining under its own power. And we do mean enormous. The Sun could fit 109 Earths side-by-side across its diameter, and it has enough volume (takes up enough space) to hold about 1.3 million Earths.

The Sun does not have a solid surface or continents like Earth, nor does it have a solid core (**Figure 21.2**). However, it does have a lot of structure and can be discussed as a series of layers, not unlike an onion. In this section, we describe the huge changes that occur in the Sun's extensive interior and atmosphere, and the dynamic and violent eruptions that occur daily in its outer layers.

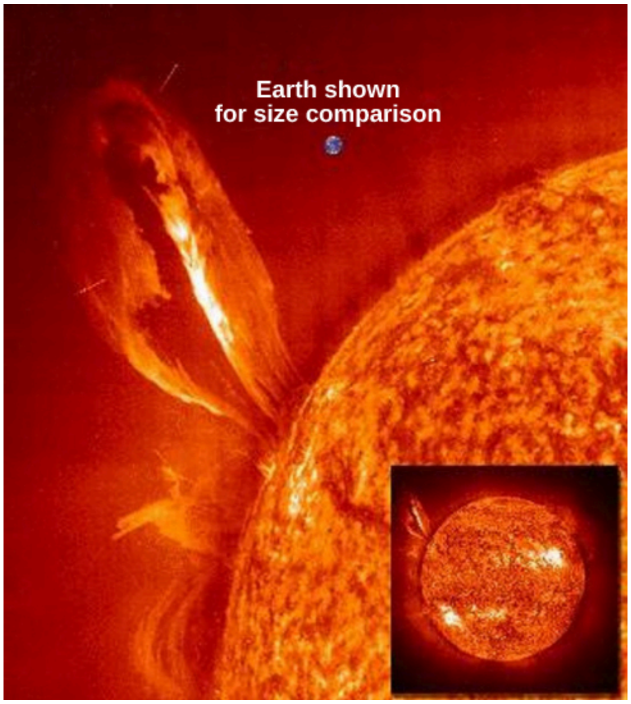


Figure 21.2 Here, Earth is shown to scale with part of the Sun and a giant loop of hot gas erupting from its surface. The inset shows the entire Sun, smaller. (credit: modification of work by SOHO/EIT/ESA)

Some of the basic characteristics of the Sun are listed in **Table 21.1**. Although some of the terms in that table may be unfamiliar to you right now, you will get to know them as you read further.

Characteristics of the Sun

Characteristic	How Found	Value
Mean distance	Radar reflection from planets	1 AU (149,597,892 km)
Maximum distance from Earth		1.521×10^8 km
Minimum distance from Earth		1.471×10^8 km
Mass	Orbit of Earth	333,400 Earth masses (1.99×10^{30} kg)
Mean angular diameter	Direct measure	31´59´´.3
Diameter of photosphere	Angular size and distance	109.3 × Earth diameter (1.39×10^6 km)
Mean density	Mass/volume	1.41 g/cm^3 (1400 kg/m^3)

Table 21.1

Characteristics of the Sun

Characteristic	How Found	Value
Gravitational acceleration at photosphere (surface gravity)	GM/R^2	27.9 × Earth surface gravity = 273 m/s^2
Solar constant	Instrument sensitive to radiation at all wavelengths	1370 W/m^2
Luminosity	Solar constant × area of spherical surface 1 AU in radius	3.8 × 10^{26} W
Spectral class	Spectrum	G2V
Effective temperature	Derived from luminosity and radius of the Sun	5800 K
Rotation period at equator	Sunspots and Doppler shift in spectra taken at the edge of the Sun	24 days 16 hours
Inclination of equator to ecliptic	Motions of sunspots	7°10´.5

Table 21.1

Composition of the Sun's Atmosphere

Let's begin by asking what the solar atmosphere is made of. As explained in **Spectroscopy**, we can use a star's *absorption line spectrum* to determine what elements are present. It turns out that the Sun contains the same elements as Earth but *not* in the same proportions. About 73% of the Sun's mass is hydrogen, and another 25% is helium. All the other chemical elements (including those we know and love in our own bodies, such as carbon, oxygen, and nitrogen) make up only 2% of our star. The 10 most abundant gases in the Sun's visible surface layer are listed in **Table 21.2**. Examine that table and notice that the composition of the Sun's outer layer is very different from Earth's crust, where we live. (In our planet's crust, the three most abundant elements are oxygen, silicon, and aluminum.) Although not like our planet's, the makeup of the Sun is quite typical of stars in general.

The Abundance of Elements in the Sun

Element	Percentage by Number of Atoms	Percentage By Mass
Hydrogen	92.0	73.4
Helium	7.8	25.0
Carbon	0.02	0.20
Nitrogen	0.008	0.09
Oxygen	0.06	0.80
Neon	0.01	0.16
Magnesium	0.003	0.06
Silicon	0.004	0.09
Sulfur	0.002	0.05
Iron	0.003	0.14

Table 21.2

The fact that our Sun and the stars all have similar compositions and are made up of mostly hydrogen and helium was first shown in a brilliant thesis in 1925 by Cecilia Payne-Gaposchkin, the first woman to get a PhD in astronomy in the United States (**Figure 21.3**). However, the idea that the simplest light gases—hydrogen and helium—were the most abundant elements in stars was so unexpected and so shocking that she assumed her analysis of the data must be wrong. At the time, she wrote, "The enormous abundance derived for these elements in the stellar atmosphere is almost certainly not real." Even scientists sometimes find it hard to accept new ideas that do not agree with what everyone "knows" to be right.

Figure 21.3 Her 1925 doctoral thesis laid the foundations for understanding the composition of the Sun and the stars. Yet, being a woman, she was not given a formal appointment at Harvard, where she worked, until 1938 and was not appointed a professor until 1956. (credit: Smithsonian Institution)

Before Payne-Gaposchkin's work, everyone assumed that the composition of the Sun and stars would be much like that of Earth. It was 3 years after her thesis that other studies proved beyond a doubt that the enormous abundance of hydrogen and helium in the Sun is indeed real. (And, as we will see, the composition of the Sun and the stars is much more typical of the makeup of the universe than the odd concentration of heavier elements that characterizes our planet.)

Most of the elements found in the Sun are in the form of atoms, with a small number of molecules, all in the form of gases: the Sun is so hot that no matter can survive as a liquid or a solid. In fact, the Sun is so hot that many of the atoms in it are *ionized*, that is, stripped of one or more of their electrons. This removal of electrons from their atoms means that there is a large quantity of free electrons and positively charged ions in the Sun, making it an electrically charged environment—quite different from the neutral one in which you are reading this text. (Scientists call such a hot ionized gas a **plasma**.)

In the nineteenth century, scientists observed a spectral line at 530.3 nanometers in the Sun's outer atmosphere, called the corona (a layer we will discuss in a minute.) This line had never been seen before, and so it was assumed that this line was the result of a new element found in the corona, quickly named coronium. It was not until 60 years later that astronomers discovered that this emission was in fact due to highly ionized iron—iron with 13 of its electrons stripped off. This is how we first discovered that the Sun's atmosphere had a temperature of more than a million degrees.

The Layers of the Sun beneath the Visible Surface

Figure 21.4 shows what the Sun would look like if we could see all parts of it from the center to its outer atmosphere; the terms in the figure will become familiar to you as you read on.

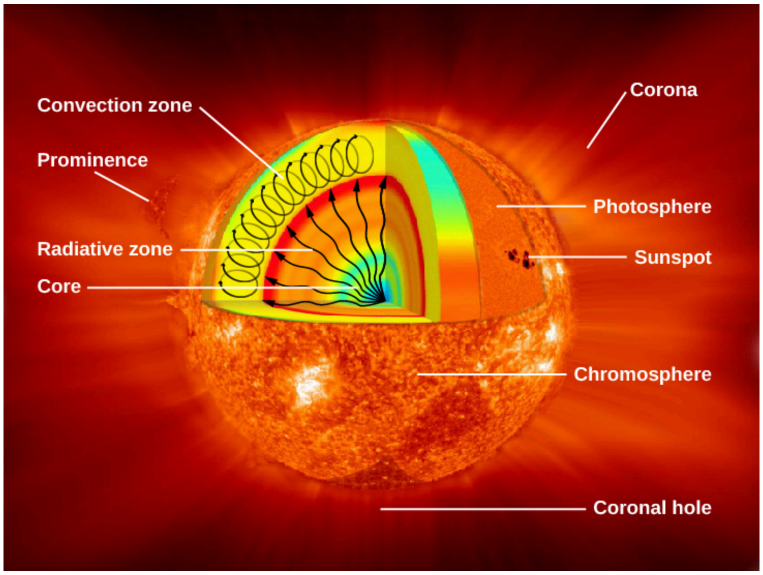


Figure 21.4 This illustration shows the different parts of the Sun, from the hot core where the energy is generated through regions where energy is transported outward, first by radiation, then by convection, and then out through the solar atmosphere. The parts of the atmosphere are also labeled the photosphere, chromosphere, and corona. Some typical features in the atmosphere are shown, such as coronal holes and prominences. (credit: modification of work by NASA/Goddard)

The Sun's layers are different from each other, and each plays a part in producing the energy that the Sun ultimately emits. We will begin with the core and work our way out through the layers. The Sun's *core* is extremely dense and is the source of all of its energy. Inside the core, nuclear energy is being released (as we discussed in **Nuclear Binding Energy**). The core is approximately 20% of the size of the solar interior and is thought to have a temperature of approximately 15 million K, making it the hottest part of the Sun.

Above the core is a region known as the *radiative zone*—named for the primary mode of transporting energy across it. This region starts at about 25% of the distance to the solar surface and extends up to about 70% of the way to the surface. The light generated in the core is transported through the radiative zone very slowly, since the high density of matter in this region means a photon cannot travel too far without encountering a particle, causing it to change direction and lose some energy.

The *convective zone* is the outermost layer of the solar interior. It is a thick layer approximately 200,000 kilometers deep that transports energy from the edge of the radiative zone to the surface through giant convection cells, similar to a pot of boiling oatmeal. The plasma at the bottom of the convective zone is extremely hot, and it bubbles to the surface where it loses its heat to space. Once the plasma cools, it sinks back to the bottom of the convective zone.

Now that we have given a quick overview of the structure of the whole Sun, in this section, we will embark on a journey through the visible layers of the Sun, beginning with the photosphere—the visible surface.

The Solar Photosphere

Earth's air is generally transparent. But on a smoggy day in many cities, it can become opaque, which prevents us from seeing through it past a certain point. Something similar happens in the Sun. Its outer atmosphere is transparent, allowing us to look a short distance through it. But when we try to look through the atmosphere deeper into the Sun, our view is blocked. The **photosphere** is the layer where the Sun becomes opaque and marks the boundary past which we cannot see (**Figure 21.5**).

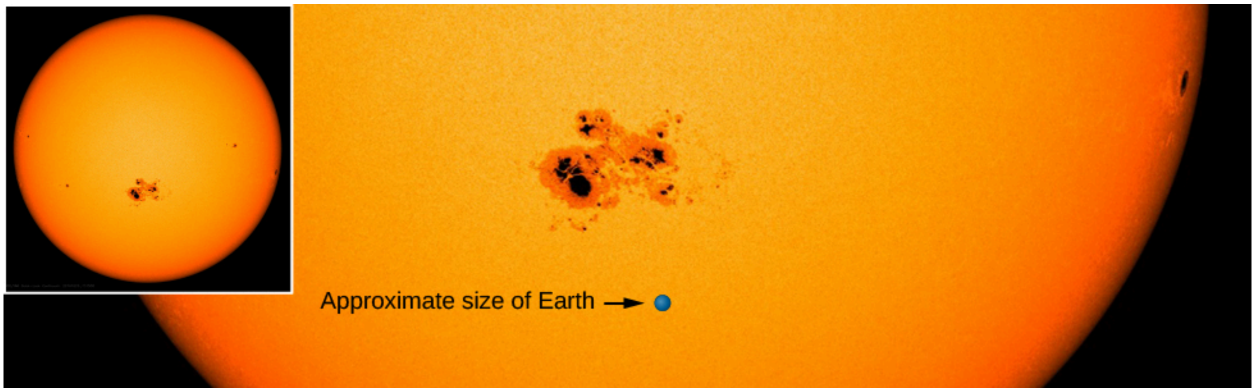


Figure 21.5 This photograph shows the photosphere—the visible surface of the Sun. Also shown is an enlarged image of a group of sunspots; the size of Earth is shown for comparison. Sunspots appear darker because they are cooler than their surroundings. The typical temperature at the center of a large sunspot is about 3800 K, whereas the photosphere has a temperature of about 5800 K. (credit: modification of work by NASA/SDO)

As we will see, the energy that emerges from the photosphere was originally generated deep inside the Sun (more on this in **Source of Sunshine: Nuclear Fusion!**). This energy is in the form of photons, which make their way slowly toward the solar surface. Outside the Sun, we can observe *only* those photons that are emitted into the solar photosphere, where the density of atoms is sufficiently low and the photons can finally escape from the Sun without colliding with another atom or ion.

As an analogy, imagine that you are attending a big campus rally and have found a prime spot near the center of the action. Your friend arrives late and calls you on your cell phone to ask you to join her at the edge of the crowd. You decide that friendship is worth more than a prime spot, and so you work your way out through the dense crowd to meet her. You can move only a short distance before bumping into someone, changing direction, and trying again, making your way slowly to the outside edge of the crowd. All this while, your efforts are not visible to your waiting friend at the edge. Your friend can't see you until you get very close to the edge because of all the bodies in the way. So too photons making their way through the Sun are constantly bumping into atoms, changing direction, working their way slowly outward, and becoming visible only when they reach the atmosphere of the Sun where the density of atoms is too low to block their outward progress.

Astronomers have found that the solar atmosphere changes from almost perfectly transparent to almost completely opaque in a distance of just over 400 kilometers; it is this thin region that we call the *photosphere*, a word that comes from the Greek for "light sphere." When astronomers speak of the "diameter" of the Sun, they mean the size of the region surrounded by the photosphere.

The photosphere looks sharp only from a distance. If you were falling into the Sun, you would not feel any surface but would just sense a gradual increase in the density of the gas surrounding you. It is much the same as falling through a cloud while skydiving. From far away, the cloud looks as if it has a sharp surface, but you do not feel a surface as you fall into it. (One big difference between these two scenarios, however, is temperature. The Sun is so hot that you would be vaporized long before you reached the photosphere. Skydiving in Earth's atmosphere is much safer.)

We might note that the atmosphere of the Sun is not a very dense layer compared to the air in the room where you are reading this text. At a typical point in the photosphere, the pressure is less than 10% of Earth's pressure at sea level, and the density is about one ten-thousandth of Earth's atmospheric density at sea level.

Observations with telescopes show that the photosphere has a mottled appearance, resembling grains of rice spilled on a dark tablecloth or a pot of boiling oatmeal. This structure of the photosphere is called **granulation** (see **Figure 21.6**). Granules, which are typically 700 to 1000 kilometers in diameter (about the width of Texas), appear as bright areas surrounded by narrow, darker (cooler) regions. The lifetime of an individual granule is only 5 to 10 minutes. Even larger are supergranules, which are about 35,000 kilometers across (about the size of two Earths) and last about 24 hours.

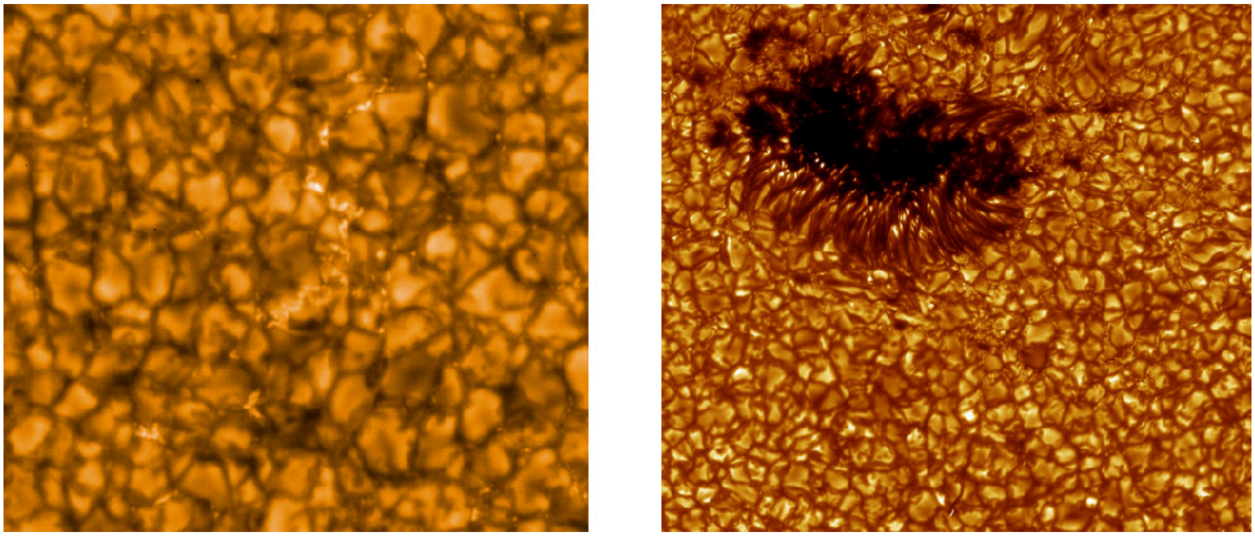


Figure 21.6 The surface markings of the convection cells create a granulation pattern on this dramatic image (left) taken from the Japanese Hinode spacecraft. You can see the same pattern when you heat up miso soup. The right image shows an irregular-shaped sunspot and granules on the Sun's surface, seen with the Swedish Solar Telescope on August 22, 2003. (credit left: modification of work by Hinode JAXA/NASA/PPARC; credit right: ISP/SST/Oddbjorn Engvold, Jun Elin Wiik, Luc Rouppe van der Voort)

The motions of the granules can be studied by examining the Doppler shifts in the spectra of gases just above them (see **The Doppler Effect**). The bright granules are columns of hotter gases rising at speeds of 2 to 3 kilometers per second from below the photosphere. As this rising gas reaches the photosphere, it spreads out, cools, and sinks down again into the darker regions between the granules. Measurements show that the centers of the granules are hotter than the intergranular regions by 50 to 100 K.

See the "boiling" action of granulation in this **30-second time-lapse video (https://openstax.org/l/30SolarGran)** from the Swedish Institute for Solar Physics.

The Chromosphere

The Sun's outer gases extend far beyond the photosphere (**Figure 21.7**). Because they are transparent to most visible radiation and emit only a small amount of light, these outer layers are difficult to observe. The region of the Sun's atmosphere that lies immediately above the photosphere is called the **chromosphere**. Until this century, the chromosphere was visible only when the photosphere was concealed by the Moon during a total solar eclipse. In the seventeenth century, several observers described what appeared to them as a narrow red "streak" or "fringe" around the edge of the Moon during a brief instant after the Sun's photosphere had been covered. The name *chromosphere*, from the Greek for "colored sphere," was given to this red streak.

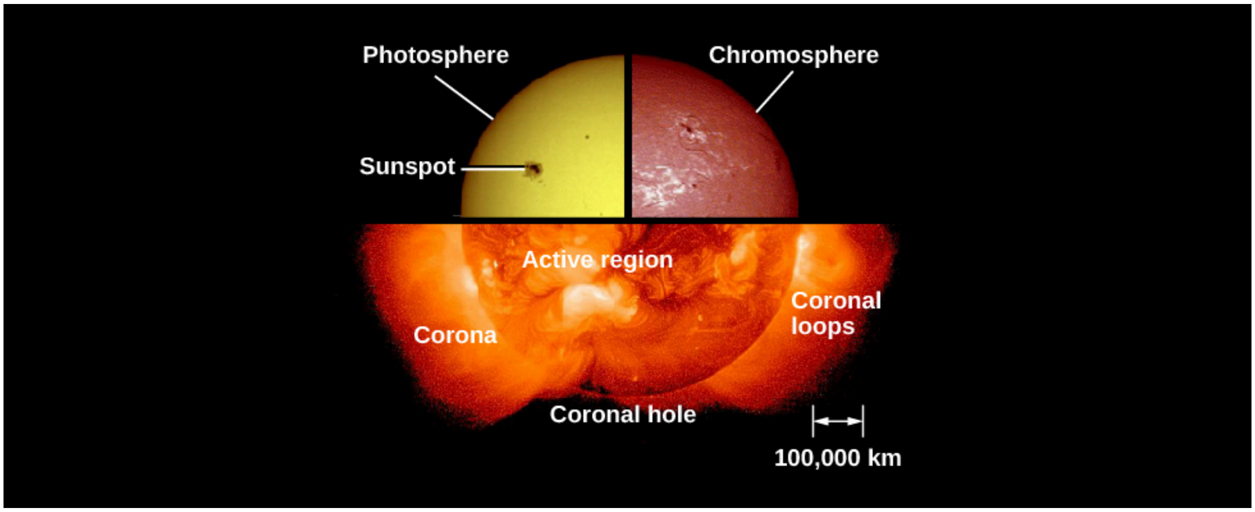


Figure 21.7 Composite image showing the three components of the solar atmosphere: the photosphere or surface of the Sun taken in ordinary light; the chromosphere, imaged in the light of the strong red spectral line of hydrogen (H-alpha); and the corona as seen with X-rays. (credit: modification of work by NASA)

Observations made during eclipses show that the chromosphere is about 2000 to 3000 kilometers thick, and its spectrum consists of bright emission lines, indicating that this layer is composed of hot gases emitting light at discrete wavelengths. The reddish color of the chromosphere arises from one of the strongest emission lines in the visible part of its spectrum—the bright red line caused by hydrogen, the element that, as we have already seen, dominates the composition of the Sun.

In 1868, observations of the chromospheric spectrum revealed a yellow emission line that did not correspond to any previously known element on Earth. Scientists quickly realized they had found a new element and named it *helium* (after *helios*, the Greek word for "Sun"). It took until 1895 for helium to be discovered on our planet. Today, students are probably most familiar with it as the light gas used to inflate balloons, although it turns out to be the second-most abundant element in the universe.

The temperature of the chromosphere is about 10,000 K. This means that the chromosphere is hotter than the photosphere, which should seem surprising. In all the situations we are familiar with, temperatures fall as one moves away from the source of heat, and the chromosphere is farther from the center of the Sun than the photosphere is.

The Transition Region

The increase in temperature does not stop with the chromosphere. Above it is a region in the solar atmosphere where the temperature changes from 10,000 K (typical of the chromosphere) to nearly a million degrees. The hottest part of the solar atmosphere, which has a temperature of a million degrees or more, is called the **corona**. Appropriately, the part of the Sun where the rapid temperature rise occurs is called the **transition region**. It is probably only a few tens of kilometers thick. **Figure 21.8** summarizes how the temperature of the solar atmosphere changes from the photosphere outward.

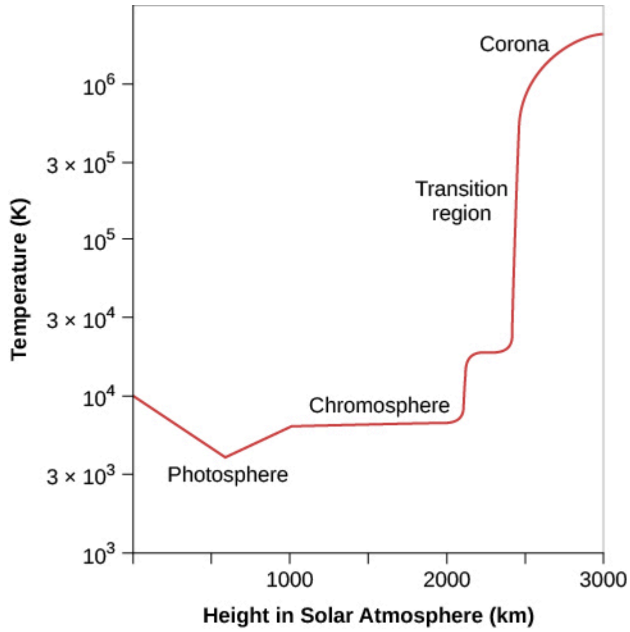


Figure 21.8 On this graph, temperature is shown increasing upward, and height above the photosphere is shown increasing to the right. Note the very rapid increase in temperature over a very short distance in the transition region between the chromosphere and the corona.

In 2013, NASA launched the Interface Region Imaging Spectrograph (IRIS) to study the transition region to understand better how and why this sharp temperature increase occurs. IRIS is the first space mission that is able to obtain high spatial resolution images of the different features produced over this wide temperature range and to see how they change with time and location (**Figure 21.9**).

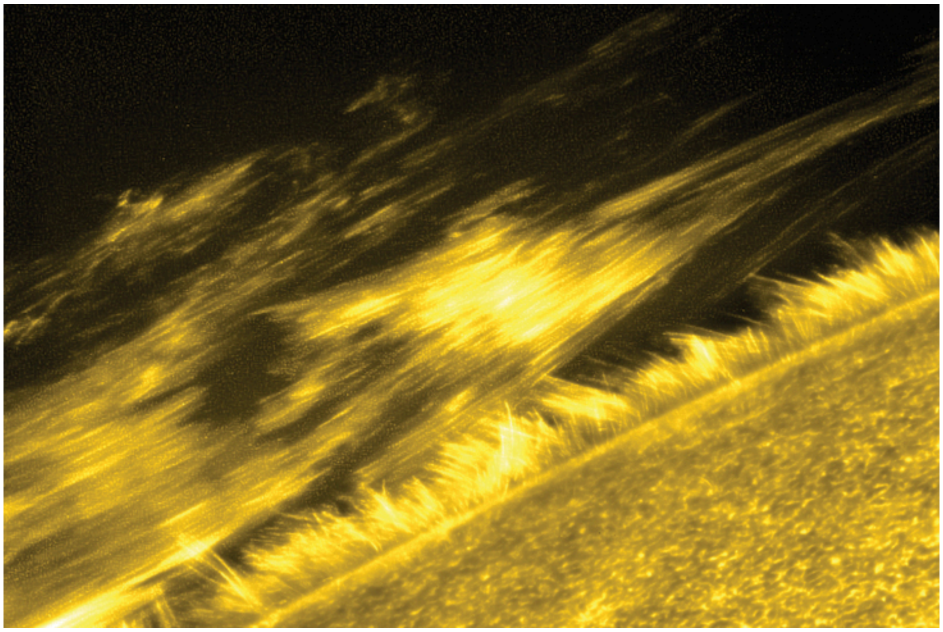


Figure 21.9 This image shows a giant ribbon of relatively cool gas threading through the lower portion of the hot corona. This ribbon (the technical term is filament) is made up of many individual threads. Time-lapse movies of this filament showed that it gradually heated as it moved through the corona. Scientists study events like this in order to try to understand what heats the chromosphere and corona to high temperatures. The “whiskers” at the edge of the Sun are spicules, jets of gas that shoot material up from the Sun’s surface and disappear after only a few minutes. This single image gives a hint of just how complicated it is to construct a model of the all the different structures and heating mechanisms in the solar atmosphere. (credit: JAXA/NASA/Hinode)

Figure 21.4 and the red graph in **Figure 21.8** make the Sun seem rather like an onion, with smooth spherical shells, each one with a different temperature. For a long time, astronomers did indeed think of the Sun this way. However, we now know that while this idea of layers—photosphere, chromosphere, transition region, corona—describes the big picture fairly well, the Sun’s atmosphere is really more complicated, with hot and cool regions intermixed. For example, clouds of carbon monoxide gas with temperatures colder than 4000 K have now been found at the same height above the photosphere as the much hotter gas of the chromosphere.

The Corona

The outermost part of the Sun’s atmosphere is called the *corona*. Like the chromosphere, the corona was first observed during total eclipses (**Figure 21.10**). Unlike the chromosphere, the corona has been known for many centuries: it was referred to by the Roman historian Plutarch and was discussed in some detail by Kepler.

The corona extends millions of kilometers above the photosphere and emits about half as much light as the full moon. The reason we don’t see this light until an eclipse occurs is the overpowering brilliance of the photosphere. Just as bright city lights make it difficult to see faint starlight, so too does the intense light from the photosphere hide the faint light from the corona. While the best time to see the corona from Earth is during a total solar eclipse, it can be observed easily from orbiting spacecraft. Its brighter parts can now be photographed with a special instrument—a coronagraph—that removes the Sun’s glare from the image with an occulting disk (a circular piece of material held so it is just in front of the Sun).

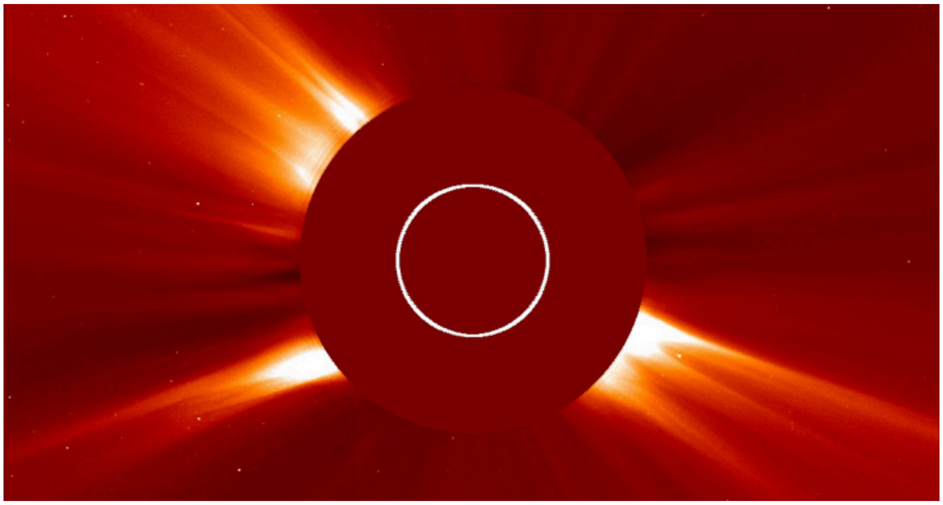


Figure 21.10 This image of the Sun was taken March 2, 2016. The larger dark circle in the center is the disk the blocks the Sun's glare, allowing us to see the corona. The smaller inner circle is where the Sun would be if it were visible in this image. (credit: modification of work by NASA/SOHO)

Studies of its spectrum show the corona to be very low in density. At the bottom of the corona, there are only about 10^9 atoms per cubic centimeter, compared with about 10^{16} atoms per cubic centimeter in the upper photosphere and 10^{19} molecules per cubic centimeter at sea level in Earth's atmosphere. The corona thins out very rapidly at greater heights, where it corresponds to a high vacuum by Earth laboratory standards. The corona extends so far into space—far past Earth—that here on our planet, we are technically living in the Sun's atmosphere.

The Solar Wind

One of the most remarkable discoveries about the Sun's atmosphere is that it produces a stream of charged particles (mainly protons and electrons) that we call the **solar wind**. These particles flow outward from the Sun into the solar system at a speed of about 400 kilometers per second (almost 1 million miles per hour)! The solar wind exists because the gases in the corona are so hot and moving so rapidly that they cannot be held back by solar gravity. (This wind was actually discovered by its effects on the charged tails of comets; in a sense, we can see the comet tails blow in the solar breeze the way wind socks at an airport or curtains in an open window flutter on Earth.)

Although the solar wind material is very, very rarified (i.e., *extremely* low density), the Sun has an enormous surface area. Astronomers estimate that the Sun is losing about 1–2 million tons of material each second through this wind. Although this sounds like a lot, it's so trivial compared to the enormous mass of the Sun that it can be neglected as we study the Sun.

From where in the Sun does the solar wind emerge? In visible photographs, the solar corona appears fairly uniform and smooth. X-ray and extreme ultraviolet pictures, however, show that the corona has loops, plumes, and both bright and dark regions. Large dark regions of the corona that are relatively cool and quiet are called **coronal holes** (Figure 21.11). In these regions, magnetic field lines stretch far out into space away from the Sun, rather than looping back to the surface. The solar wind comes predominantly from coronal holes, where gas can stream away from the Sun into space unhindered by magnetic fields. Hot coronal gas, on the other hand, is present mainly where magnetic fields have trapped and concentrated it.

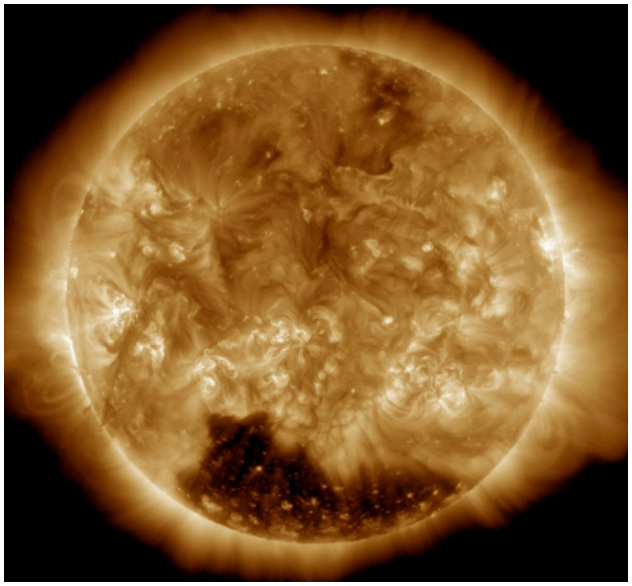


Figure 21.11 Coronal Hole. The dark area visible near the Sun's south pole on this Solar Dynamics Observer spacecraft image is a coronal hole. (credit: modification of work by NASA/SDO)

At the surface of Earth, we are protected to some degree from the solar wind by our atmosphere and Earth's magnetic field. However, the magnetic field lines come into Earth at the north and south magnetic poles. Here, charged particles accelerated by the solar wind can follow the field down into our atmosphere. As the particles strike molecules of air, they cause them to glow, producing beautiful curtains of light called the **auroras**, or the northern and southern lights (**Figure 21.12**).

Figure 21.12 The colorful glow in the sky results from charged particles in a solar wind interacting with Earth's magnetic fields. The stunning display captured here occurred over Jokulsarlon Lake in Iceland in 2013. (credit: Moyan Brenn)

This **NASA video (https://openstax.org/l/30Aurora)** explains and demonstrates the nature of the auroras and their relationship to Earth's magnetic field.

21.2 | Sources of Sunshine: Thermal and Gravitational

Energy?

Learning Objectives

By the end of this section, you will be able to:

- Identify different forms of energy
- Understand the law of conservation of energy
- Explain ways that energy can be transformed

Energy is a challenging concept to grasp because it exists in so many different forms that it defies any single simple explanation. In many ways, comprehending energy is like comprehending wealth: There are very different forms of wealth and they follow different rules, depending on if they are the stock market, real estate, a collection of old comic books, great piles of cash, or one of the many other ways to make and lose money. It is easier to discuss one or two forms of wealth—or energy—than to discuss that concept in general.

Of course today, since we understand the **Formation of the Solar System**, we know that the thermal energy increased as the gravitational potential energy diminished during the collapse of the solar nebula. And, we know that once sufficient temperature and density were reached at the center of the nebula, nuclear fusion began and a star was born. But it is interesting to examine the historical evolution of ideas about solar energy.

When striving to understand how the Sun can continue to put out so much energy for so long, scientists considered many different types of energy. Nineteenth-century scientists knew of two possible sources for the Sun's energy: chemical and gravitational energy. The source of chemical energy most familiar to them was the burning (the chemical term is *oxidation*) of wood, coal, gasoline, or other fuel. We know exactly how much energy the burning of these materials can produce. We can thus calculate that even if the immense mass of the Sun consisted of a burnable material like coal or wood, our star could not produce energy at its present rate for more than few thousand years. However, we know from geologic evidence that water was present on Earth's surface nearly 4 billion years ago, so the Sun must have been shining brightly (and making Earth warm) at least as long as that. Today, we also know that at the temperatures found in the Sun, nothing like solid wood or coal could survive.

Conservation of Energy

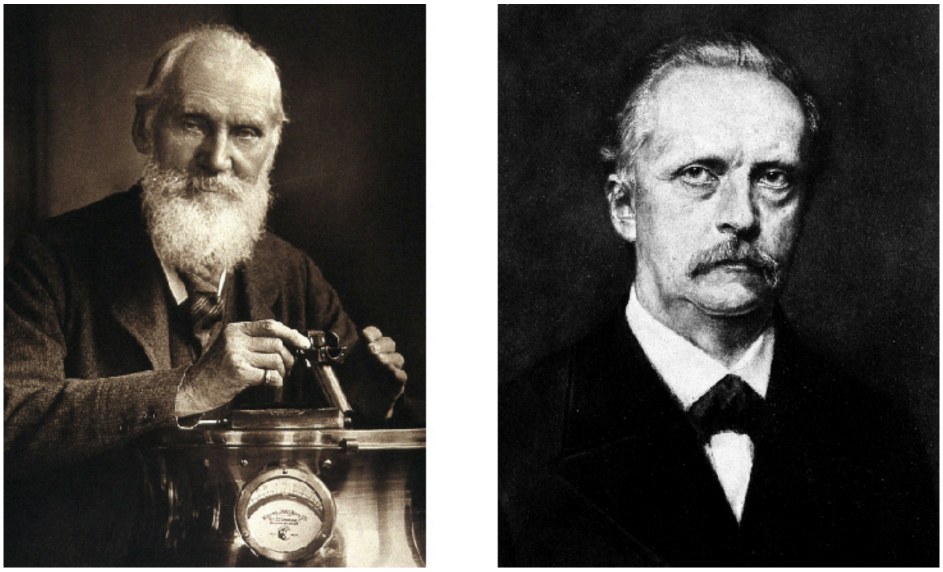
Other nineteenth-century attempts to determine what makes the Sun shine used the law of conservation of energy. Simply stated, this law says that energy cannot be created or destroyed, but can be transformed from one type to another, such as from heat to mechanical energy. The steam engine, which was key to the Industrial Revolution, provides a good example. In this type of engine, the hot steam from a boiler drives the movement of a piston, converting heat energy into motion energy.

Conversely, motion can be transformed into heat. If you clap your hands vigorously at the end of an especially good astronomy lecture, your palms become hotter. If you rub ice on the surface of a table, the heat produced by friction melts the ice. The brakes on cars use friction to reduce speed, and in the process, transform motion energy into heat energy. That is why after bringing a car to a stop, the brakes can be very hot; this also explains why brakes can overheat when used carelessly while descending long mountain roads.

In the nineteenth century, scientists thought that the source of the Sun's heat might be the mechanical motion of meteorites falling into it. Their calculations showed, however, that in order to produce the total amount of energy emitted by the Sun, the mass in meteorites that would have to fall into the Sun every 100 years would equal the mass of Earth. The resulting increase in the Sun's mass would, according to Kepler's third law, change the period of Earth's orbit by 2 seconds per year. Such a change would be easily measurable and was not, in fact, occurring. Scientists could then disprove this as the source of the Sun's energy.

Gravitational Contraction as a Source of Energy

Proposing an alternative explanation, British physicist Lord Kelvin and German scientist Hermann von Helmholtz (**Figure 21.13**), in about the middle of the nineteenth century, proposed that the Sun might produce energy by the conversion of gravitational energy into heat. They suggested that the outer layers of the Sun might be "falling" inward because of the force of gravity. In other words, they proposed that the Sun could be shrinking in size, staying hot and bright as a result.



(a) (b)

Figure 21.13 (a) British physicist William Thomson (Lord Kelvin) and (b) German scientist Hermann von Helmholtz proposed that the contraction of the Sun under its own gravity might account for its energy. (credit a: modification of work by Wellcome Library, London; credit b: modification of work by Wellcome Library, London)

To imagine what would happen if this hypothesis were true, picture the outer layer of the Sun starting to fall inward. This outer layer is a gas made up of individual atoms, all moving about in random directions. If a layer falls inward, the atoms acquire an additional speed because of falling motion. As the outer layer falls inward, it also contracts, moving the atoms closer together. Collisions become more likely, and some of them transfer the extra speed associated with the falling motion to other atoms. This, in turn, increases the speeds of those atoms. The temperature of a gas is a measure of the kinetic energy (motion) of the atoms within it; hence, the temperature of this layer of the Sun increases. Collisions also excite electrons within the atoms to higher-energy orbits. When these electrons return to their normal orbits, they emit photons, which can then escape from the Sun (see **Atomic Spectra**).

Kelvin and Helmholtz calculated that a contraction of the Sun at a rate of only about 40 meters per year would be enough to produce the amount of energy that it is now radiating. Over the span of human history, the decrease in the Sun's size from such a slow contraction would be undetectable.

If we assume that the Sun began its life as a large, diffuse cloud of gas, then we can calculate how much energy has been radiated by the Sun during its entire lifetime as it has contracted from a very large diameter to its present size. The amount of energy is on the order of 10^{42} joules. Since the solar luminosity is 4×10^{26} watts (joules/second) or about 10^{34} joules per year, contraction could keep the Sun shining at its present rate for roughly 100 million years.

In the nineteenth century, 100 million years at first seemed plenty long enough, since Earth was then widely thought to be much younger than this. But toward the end of that century and into the twentieth, geologists and physicists showed that Earth (and, hence, the Sun) is actually much older. Contraction therefore cannot be the primary source of solar energy (although, as we saw in **Formation of the Solar System**, contraction is an important source of energy for a while in stars that are just being born). Scientists were thus confronted with a puzzle of enormous proportions. Either an unknown type of energy was responsible for the most important energy source known to humanity, or estimates of the age of the solar system (and life on Earth) had to be seriously modified. Charles Darwin, whose theory of evolution required a longer time span than the theories of the Sun seemed to permit, was discouraged by these results and continued to worry about them until his death in 1882.

It was only in the twentieth century that the true source of the Sun's energy was identified. The two key pieces of information required to solve the puzzle were the structure of the nucleus of the atom and the fact that mass can be converted into energy.

21.3 | Source of Sunshine: Nuclear Fusion!

Learning Objectives

By the end of this section, you will be able to:

- Describe the process of nuclear fusion in terms of its products and reactants
- Calculate the energies of particles produced by a fusion reaction
- Explain the production of energy by the Sun, and nucleosynthesis

The process of combining lighter nuclei to make heavier nuclei is called **nuclear fusion**. As with fission reactions, fusion reactions are exothermic—they release energy. Suppose that we fuse a carbon and helium nuclei to produce oxygen:

$$^{12}_{6}\text{C} + {}^{4}_{2}\text{He} \rightarrow {}^{16}_{8}\text{O} + \gamma.$$

The energy changes in this reaction can be understood using a graph of binding energy per nucleon (**Figure 18.10**). Comparing the binding energy per nucleon for oxygen, carbon, and helium, the oxygen nucleus is much more tightly bound than the carbon and helium nuclei, indicating that the reaction produces a drop in the energy of the system. This energy is released in the form of gamma radiation. Fusion reactions are said to be exothermic when the amount of energy released (known as the *Q value*) in each reaction is greater than zero $(Q > 0)$.

An important example of nuclear fusion in nature is the production of energy in the Sun. In 1938, Hans Bethe proposed that the Sun produces energy when hydrogen nuclei (^{1}H) fuse into stable helium nuclei $\left(^{4}\text{He}\right)$ in the Sun's core (**Figure 21.14**). This process, called the **proton-proton chain**, is summarized by three reactions:

$$\begin{aligned} &{}^{1}_{1}\text{H} + {}^{1}_{1}\text{H} \rightarrow {}^{2}_{1}\text{H} + {}^{0}_{+1}\text{e} + v + Q, \\ &{}^{1}_{1}\text{H} + {}^{2}_{1}\text{H} \rightarrow {}^{3}_{2}\text{He} + \gamma + Q, \\ &{}^{3}_{2}\text{He} + {}^{3}_{2}\text{He} \rightarrow {}^{4}_{2}\text{He} + {}^{1}_{1}\text{H} + {}^{1}_{1}\text{H} + Q. \end{aligned}$$

Thus, a stable helium nucleus is formed from the fusion of the nuclei of the hydrogen atom. These three reactions can be summarized by

$$4\,{}^{1}_{1}\text{H} \rightarrow {}^{4}_{2}\text{He} + 2\,{}^{0}_{+1}\text{e} + 2\gamma + 2v + Q.$$

The net Q value is about 26 MeV. The release of this energy produces an outward thermal gas pressure that prevents the Sun from gravitational collapse. Astrophysicists find that hydrogen fusion supplies the energy stars require to maintain energy balance over most of a star's life span.

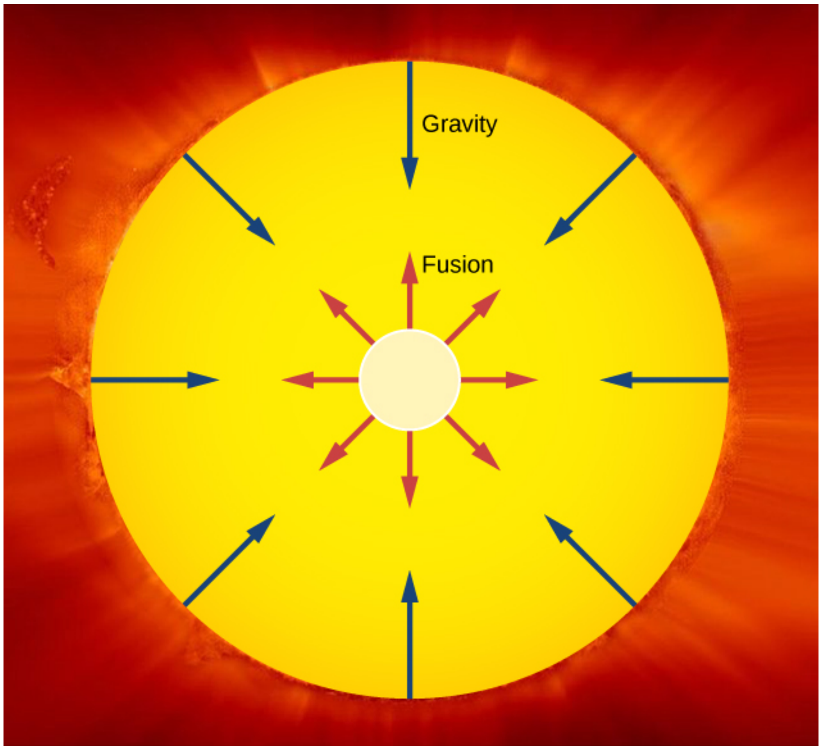


Figure 21.14 The Sun produces energy by fusing hydrogen into helium at the Sun's core. The red arrows show outward pressure due to thermal gas, which tends to make the Sun expand. The blue arrows show inward pressure due to gravity, which tends to make the Sun contract. These two influences balance each other.

Nucleosynthesis

Scientist now believe that many heavy elements found on Earth and throughout the universe were originally synthesized by fusion within the hot cores of the stars. This process is known as **nucleosynthesis**. For example, in lighter stars, hydrogen combines to form helium through the proton-proton chain. Once the hydrogen fuel is exhausted, the star enters the next stage of its life and fuses helium. An example of a nuclear reaction chain that can occur is:

$$\begin{aligned} {}_{2}^{4}\text{He} + {}_{2}^{4}\text{He} &\rightarrow {}_{4}^{8}\text{Be} + \gamma, \\ {}_{4}^{8}\text{Be} + {}_{2}^{4}\text{He} &\rightarrow {}_{6}^{12}\text{C} + \gamma, \\ {}_{6}^{12}\text{C} + {}_{2}^{4}\text{He} &\rightarrow {}_{8}^{16}\text{O} + \gamma. \end{aligned}$$

Carbon and oxygen nuclei produced in such processes eventually reach the star's surface by convection. Near the end of its lifetime, the star loses its outer layers into space, thus enriching the interstellar medium with the nuclei of heavier elements (**Figure 21.15**).

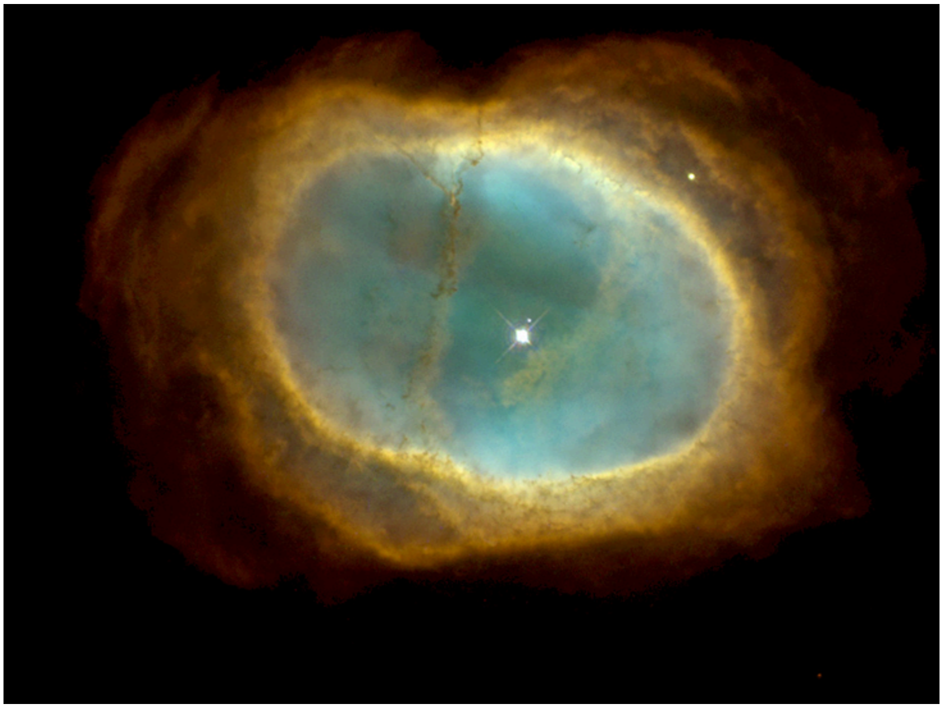


Figure 21.15 A planetary nebula is produced at the end of the life of a star. The greenish color of this planetary nebula comes from oxygen ions. (credit: Hubble Heritage Team (STScI/AURA/NASA/ESA))

Stars similar in mass to the Sun do not become hot enough to fuse nuclei as heavy (or heavier) than oxygen nuclei. However, in massive stars whose cores become much hotter $\left(T > 6 \times 10^8 \text{ K}\right)$, even more complex nuclei are produced. Some representative reactions are

$$\begin{aligned} {}^{12}_{6}\text{C} + {}^{12}_{6}\text{C} &\rightarrow {}^{23}_{11}\text{Na} + {}^{1}_{1}\text{H}, \\ {}^{12}_{6}\text{C} \;\rightarrow {}^{12}_{6}\text{C} &\rightarrow {}^{24}_{12}\text{Mg} + \gamma, \\ {}^{12}_{6}\text{C} + {}^{16}_{8}\text{O} &\rightarrow {}^{28}_{14}\text{Si} + \gamma. \end{aligned}$$

Nucleosynthesis continues until the core is primarily iron-nickel metal. Now, iron has the peculiar property that any fusion or fission reaction involving the iron nucleus is endothermic, meaning that energy is absorbed rather than produced. Hence, nuclear energy cannot be generated in an iron-rich core. Lacking an outward pressure from fusion reactions, the star begins to contract due to gravity. This process heats the core to a temperature on the order of $5 \times 10^9 \text{ K}$. Expanding shock waves generated within the star due to the collapse cause the star to quickly explode. The luminosity of the star can increase temporarily to nearly that of an entire galaxy. During this event, the flood of energetic neutrons reacts with iron and the other nuclei to produce elements heavier than iron. These elements, along with much of the star, are ejected into space by the explosion. Supernovae and the formation of planetary nebulas together play a major role in the dispersal of chemical elements into space.

Eventually, much of the material lost by stars is pulled together through the gravitational force, and it condenses into a new generation of stars and accompanying planets. Recent images from the Hubble Space Telescope provide a glimpse of this magnificent process taking place in the constellation Serpens (**Figure 21.16**). The new generation of stars begins the nucleosynthesis process anew, with a higher percentage of heavier elements. Thus, stars are “factories” for the chemical elements, and many of the atoms in our bodies were once a part of stars.

Figure 21.16 This image taken by NASA's Spitzer Space Telescope and the Two Micron All Sky Survey (2MASS), shows the Serpens Cloud Core, a star-forming region in the constellation Serpens (the "Serpent"). Located about 750 light-years away, this cluster of stars is formed from cooling dust and gases. Infrared light has been used to reveal the youngest stars in orange and yellow. (credit: NASA/JPL-Caltech/2MASS)

Example 21.1

Energy of the Sun

The power output of the Sun is approximately 3.8×10^{26} J/s. Most of this energy is produced in the Sun's core by the proton-proton chain. This energy is transmitted outward by the processes of convection and radiation. (a) How many of these fusion reactions per second must occur to supply the power radiated by the Sun? (b) What is the rate at which the mass of the Sun decreases? (c) In about five billion years, the central core of the Sun will be depleted of hydrogen. By what percentage will the mass of the Sun have decreased from its present value when the core is depleted of hydrogen?

Strategy

The total energy output per second is given in the problem statement. If we know the energy released in each fusion reaction, we can determine the rate of the fusion reactions. If the mass loss per fusion reaction is known, the mass loss rate is known. Multiplying this rate by five billion years gives the total mass lost by the Sun. This value is divided by the original mass of the Sun to determine the percentage of the Sun's mass that has been lost when the hydrogen fuel is depleted.

Solution

a. The decrease in mass for the fusion reaction is

$$\begin{aligned}\Delta m &= 4m\left({}^{1}_{1}\text{H}\right) - m\left({}^{4}_{2}\text{He}\right) - 2m\left({}^{\;\;0}_{+1}\text{e}\right) \\ &= 4(1.007825\text{ u}) - 4.002603\text{ u} - 2(0.000549\text{ u}) \\ &= 0.0276\text{ u}.\end{aligned}$$

The energy released per fusion reaction is

$$Q = (0.0276\text{ u})(931.49\text{ MeV/u}) = 25.7\text{ MeV}.$$

Thus, to supply $3.8 \times 10^{26}\text{ J/s} = 2.38 \times 10^{39}\text{ MeV/s},$ there must be

$$\frac{2.38 \times 10^{39}\text{ MeV/s}}{25.7\text{ MeV/reaction}} = 9.26 \times 10^{37}\text{ reaction/s}.$$

b. The Sun's mass decreases by $0.0276\text{ u} = 4.58 \times 10^{-29}\text{ kg}$ per fusion reaction, so the rate at which its mass decreases is

$$\left(9.26 \times 10^{37} \text{ reaction/s}\right)\left(4.58 \times 10^{-29} \text{ kg/reaction}\right) = 4.24 \times 10^{9} \text{ kg/s}.$$

c. In $5 \times 10^{9} \text{ y} = 1.6 \times 10^{17} \text{ s},$ the Sun's mass will therefore decrease by

$$\Delta M = \left(4.24 \times 10^{9} \text{ kg/s}\right)\left(1.6 \times 10^{17} \text{ s}\right) = 6.8 \times 10^{26} \text{ kg}.$$

The current mass of the Sun is about 2.0×10^{30} kg, so the percentage decrease in its mass when its hydrogen fuel is depleted will be

$$\left(\frac{6.8 \times 10^{26} \text{ kg}}{2.0 \times 10^{30} \text{ kg}}\right) \times 100\% = 0.034\%.$$

Significance

After five billion years, the Sun is very nearly the same mass as it is now. Hydrogen burning does very little to change the mass of the Sun. This calculation assumes that only the proton-proton decay change is responsible for the power output of the Sun.

21.1 Check Your Understanding Where does the energy from the Sun originate?

21.4 | The Solar Interior: Theory

Learning Objectives

By the end of this section, you will be able to:

- Describe the state of equilibrium of the Sun
- Understand the energy balance of the Sun
- Explain how energy moves outward through the Sun
- Describe the structure of the solar interior

Fusion of protons can occur in the center of the Sun only if the temperature exceeds 12 million K. How do we know that the Sun is actually this hot? To determine what the interior of the Sun might be like, it is necessary to resort to complex calculations. Since we can't see the interior of the Sun, we have to use our understanding of physics, combined with what we see at the surface, to construct a mathematical model of what must be happening in the interior. Astronomers use observations to build a computer program containing everything they think they know about the physical processes going on in the Sun's interior. The computer then calculates the temperature and pressure at every point inside the Sun and determines what nuclear reactions, if any, are taking place. For some calculations, we can use observations to determine whether the computer program is producing results that match what we see. In this way, the program evolves with ever-improving observations.

The computer program can also calculate how the Sun will change with time. After all, the Sun must change. In its center, the Sun is slowly depleting its supply of hydrogen and creating helium instead. Will the Sun get hotter? Cooler? Larger? Smaller? Brighter? Fainter? Ultimately, the changes in the center could be catastrophic, since eventually all the hydrogen fuel hot enough for fusion will be exhausted. Either a new source of energy must be found, or the Sun will cease to shine. We will describe the ultimate fate of the Sun in later chapters. For now, let's look at some of the things we must teach the computer about the Sun in order to carry out such calculations.

The Sun Is a Plasma

The Sun is so hot that all of the material in it is in the form of an ionized gas, called a plasma. Plasma acts much like a hot gas, which is easier to describe mathematically than either liquids or solids. The particles that constitute a gas are in rapid motion, frequently colliding with one another. This constant bombardment is the *pressure* of the gas (**Figure 21.17**).

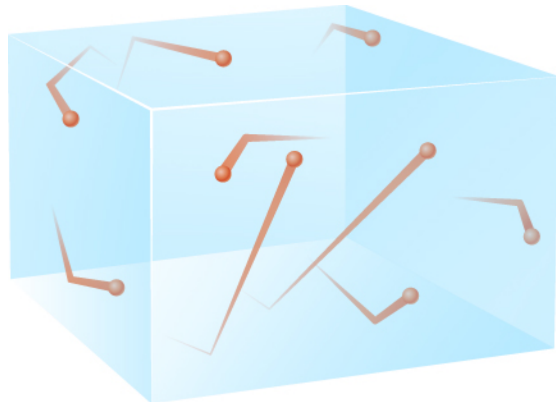


Figure 21.17 The particles in a gas are in rapid motion and produce pressure through collisions with the surrounding material. Here, particles are shown bombarding the sides of an imaginary container.

More particles within a given volume of gas produce more pressure because the combined impact of the moving particles increases with their number. The pressure is also greater when the molecules or atoms are moving faster. Since the molecules move faster when the temperature is hotter, higher temperatures produce higher pressure.

The Sun Is Stable

The Sun, like the majority of other stars, is stable; it is neither expanding nor contracting. Such a star is said to be in a condition of *equilibrium*. All the forces within it are balanced, so that at each point within the star, the temperature, pressure, density, and so on are maintained at constant values. We will see in later chapters that even these stable stars, including the Sun, are changing as they evolve, but such evolutionary changes are so gradual that, for all intents and purposes, the stars are still in a state of equilibrium at any given time.

The mutual gravitational attraction between the masses of various regions within the Sun produces tremendous forces that tend to collapse the Sun toward its center. Yet we know from the history of Earth that the Sun has been emitting roughly the same amount of energy for billions of years, so clearly it has managed to resist collapse for a very long time. The gravitational forces must therefore be counterbalanced by some other force. That force is due to the pressure of gases within the Sun (**Figure 21.18**). Calculations show that, in order to exert enough pressure to prevent the Sun from collapsing due to the force of gravity, the gases at its center must be maintained at a temperature of 15 million K. Think about what this tells us. Just from the fact that the Sun is not contracting, we can conclude that its temperature must indeed be high enough at the center for protons to undergo fusion.

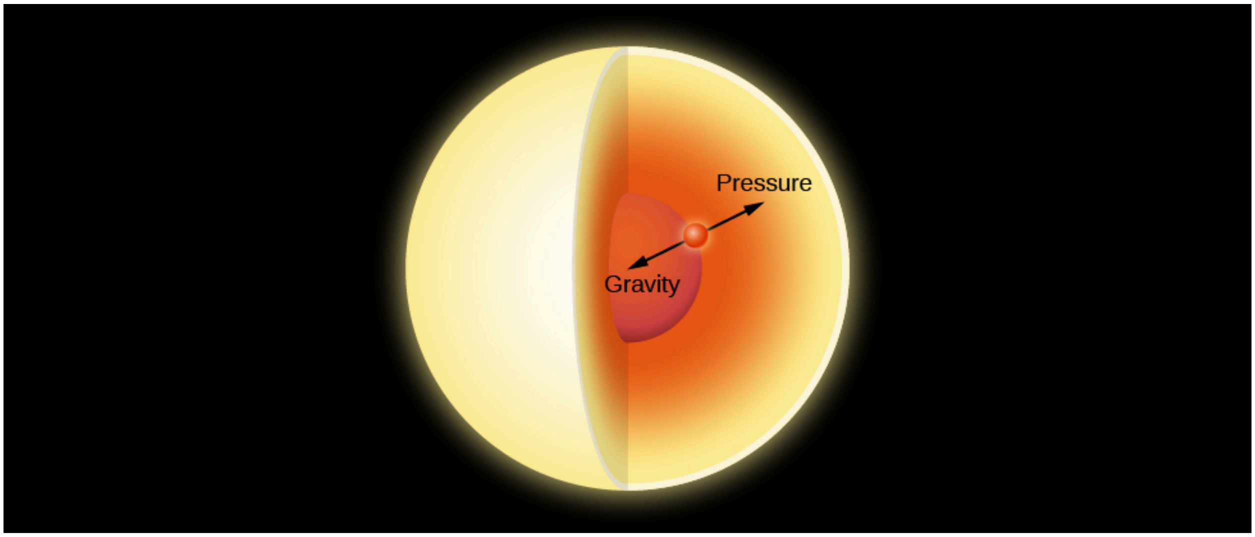


Figure 21.18 In the interior of a star, the inward force of gravity is exactly balanced at each point by the outward force of gas pressure.

The Sun maintains its stability in the following way. If the internal pressure in such a star were not great enough to balance the weight of its outer parts, the star would collapse somewhat, contracting and building up the pressure inside. On the other hand, if the pressure were greater than the weight of the overlying layers, the star would expand, thus decreasing the internal pressure. Expansion would stop, and equilibrium would again be reached when the pressure at every internal point equaled the weight of the stellar layers above that point. An analogy is an inflated balloon, which will expand or contract until an equilibrium is reached between the pressure of the air inside and outside. The technical term for this condition is **hydrostatic equilibrium**. Stable stars are all in hydrostatic equilibrium; so are the oceans of Earth as well as Earth's atmosphere. The air's own pressure keeps it from falling to the ground.

The Sun Is Not Cooling Down

As everyone who has ever left a window open on a cold winter night knows, heat always flows from hotter to cooler regions. As energy filters outward toward the surface of a star, it must be flowing from inner, hotter regions. The temperature cannot ordinarily get cooler as we go inward in a star, or energy would flow in and heat up those regions until they were at least as hot as the outer ones. Scientists conclude that the temperature is highest at the center of a star, dropping to lower and lower values toward the stellar surface. (The high temperature of the Sun's chromosphere and corona may therefore appear to be a paradox. But remember from **The Structure and Composition of the Sun** that these high temperatures are maintained by magnetic effects, which occur in the Sun's atmosphere.)

The outward flow of energy through a star robs it of its internal heat, and the star would cool down if that energy were not replaced. Similarly, a hot iron begins to cool as soon as it is unplugged from its source of electric energy. Therefore, a source of fresh energy must exist within each star. In the Sun's case, we have seen that this energy source is the ongoing fusion of hydrogen to form helium.

Heat Transfer in a Star

Since the nuclear reactions that generate the Sun's energy occur deep within it, the energy must be transported from the center of the Sun to its surface—where we see it in the form of both heat and light. There are three ways in which energy can be transferred from one place to another. In **conduction**, atoms or molecules pass on their energy by colliding with others nearby. This happens, for example, when the handle of a metal spoon heats up as you stir a cup of hot coffee. In **convection**, currents of warm material rise, carrying their energy with them to cooler layers. A good example is hot air rising from a fireplace. In **radiation**, energetic photons move away from hot material and are absorbed by some material to which they convey some or all of their energy. You can feel this when you put your hand close to the coils of an electric heater, allowing infrared photons to heat up your hand. Conduction and convection are both important in the interiors of planets. In stars, which are much more transparent, radiation and convection are important, whereas conduction can usually be ignored.

Stellar *convection* occurs as currents of hot gas flow up and down through the star (**Figure 21.19**). Such currents travel at moderate speeds and do not upset the overall stability of the star. They don't even result in a net transfer of mass either inward or outward because, as hot material rises, cool material falls and replaces it. This results in a convective circulation of rising and falling cells as seen in **Figure 21.19**. In much the same way, heat from a fireplace can stir up air currents in a room, some rising and some falling, without driving any air into or out the room. Convection currents carry heat very efficiently outward through a star. In the Sun, convection turns out to be important in the central regions and near the surface.

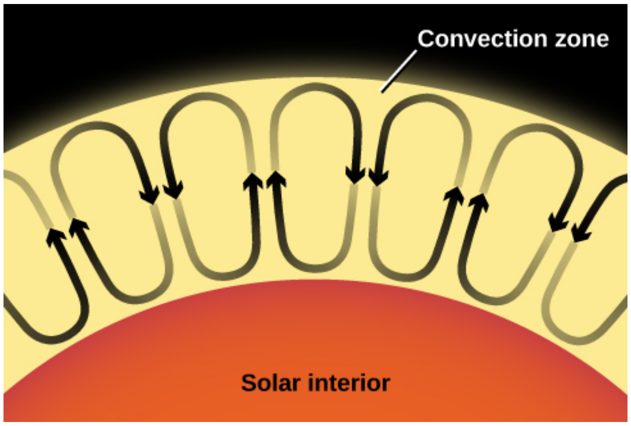


Figure 21.19 Rising convection currents carry heat from the Sun's interior to its surface, whereas cooler material sinks downward. Of course, nothing in a real star is as simple as diagrams in textbooks suggest.

Unless convection occurs, the only significant mode of energy transport through a star is by electromagnetic radiation. Radiation is not an efficient means of energy transport in stars because gases in stellar interiors are very opaque, that is, a photon does not go far (in the Sun, typically about 0.01 meter) before it is absorbed. (The processes by which atoms and ions can interrupt the outward flow of photons—such as becoming ionized—were discussed in the section on the **The Bohr Model**.) The absorbed energy is always reemitted, but it can be reemitted in any direction. A photon absorbed when traveling outward in a star has almost as good a chance of being radiated back toward the center of the star as toward its surface.

A particular quantity of energy, therefore, zigzags around in an almost random manner and takes a long time to work its way from the center of a star to its surface (**Figure 21.20**). Estimates are somewhat uncertain, but in the Sun, as we saw, the time required is probably between 100,000 and 1,000,000 years. If the photons were not absorbed and reemitted along the way, they would travel at the speed of light and could reach the surface in a little over 2 seconds, just as neutrinos do (**Figure 21.21**).

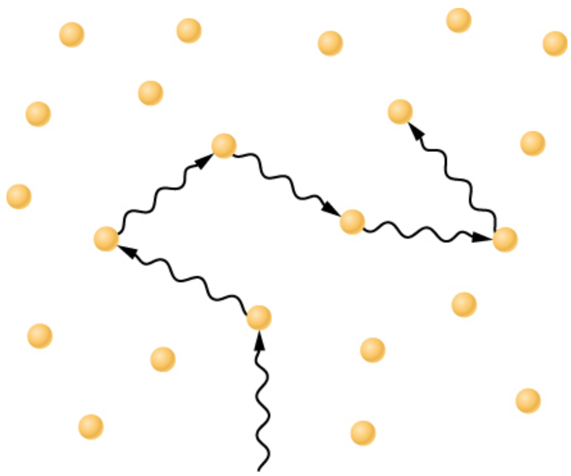


Figure 21.20 A photon moving through the dense gases in the solar interior travels only a short distance before it interacts with one of the surrounding atoms. The resulting photon usually has a lower energy after each interaction and may then travel in any random direction.

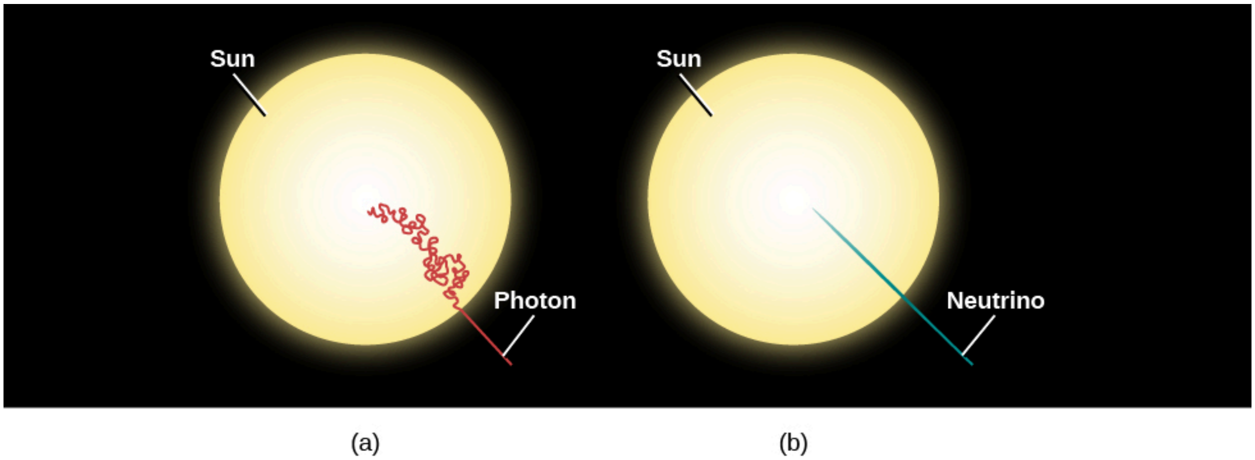


Figure 21.21 (a) Because photons generated by fusion reactions in the solar interior travel only a short distance before being absorbed or scattered by atoms and sent off in random directions, estimates are that it takes between 100,000 and 1,000,000 years for energy to make its way from the center of the Sun to its surface. (b) In contrast, neutrinos do not interact with matter but traverse straight through the Sun at the speed of light, reaching the surface in only a little more than 2 seconds.

Heat Transfer and Cooking

The three ways that heat energy moves from higher-temperature regions to cooler regions are all used in cooking, and this is important to all of us who enjoy making or eating food. (We introduced all three of them in the section on **Heat Transfer**.)

Conduction is heat transfer by physical contact during which the energetic motion of particles in one region spread to other regions and even to adjacent objects in close contact. A tasty example of this is cooking a steak on a hot iron skillet. When a flame makes the bottom of a skillet hot, the particles in it vibrate actively and collide with neighboring particles, spreading the heat energy throughout the skillet (the ability to spread heat uniformly is a key criterion for selecting materials for cookware). A steak sitting on the surface of the skillet picks up heat energy by the particles in the surface of the skillet colliding with particles on the surface of the steak. Many cooks will put a little oil on the pan, and this layer of oil, besides preventing sticking, increases heat transfer by filling in gaps and increasing the contact surface area.

Convection is heat transfer by the motion of matter that rises because it is hot and less dense. Heating a fluid makes it expand, which makes it less dense, so it rises. An oven is a great example of this: the fire is at the bottom of the oven and heats the air down there, causing it to expand (becoming less dense), so it rises up to where the food is. The rising hot air carries the heat from the fire to the food by convection. This is how conventional ovens work. You may also be familiar with convection ovens that use a fan to circulate hot air for more even cooking. A scientist would object to that name because normal non-fan ovens that rely on hot air rising to circulate the heat are convection ovens; technically, the ovens that use fans to help move heat are “advection” ovens. (You may not have heard about this because the scientists who complain loudly about misusing the terms convection and advection don’t get out much.)

Radiation is the transfer of heat energy by electromagnetic radiation. Although microwave ovens are an obvious example of using radiation to heat food, a simpler example is a toy oven. Toy ovens are powered by a very bright light bulb. The child-chefs prepare a mix for brownies or cookies, put it into a tray, and place it in the toy oven under the bright light bulb. The light and heat from the bulb hit the brownie mix and cook it. If you have ever put your hand near a bright light, you have undoubtedly noticed your hand getting warmed by the light.

Model Stars

Scientists use the principles we have just described to calculate what the Sun’s interior is like. These physical ideas are expressed as mathematical equations that are solved to determine the values of temperature, pressure, density, the efficiency with which photons are absorbed, and other physical quantities throughout the Sun. The solutions obtained, based on a specific set of physical assumptions, provide a theoretical model for the interior of the Sun.

Figure 21.22 schematically illustrates the predictions of a theoretical model for the Sun’s interior. Energy is generated through fusion in the core of the Sun, which extends only about one-quarter of the way to the surface but contains about one-third of the total mass of the Sun. At the center, the temperature reaches a maximum of approximately 15 million K, and the density is nearly 150 times that of water. The energy generated in the core is transported toward the surface by radiation until it reaches a point about 70% of the distance from the center to the surface. At this point, convection begins, and energy is transported the rest of the way, primarily by rising columns of hot gas.

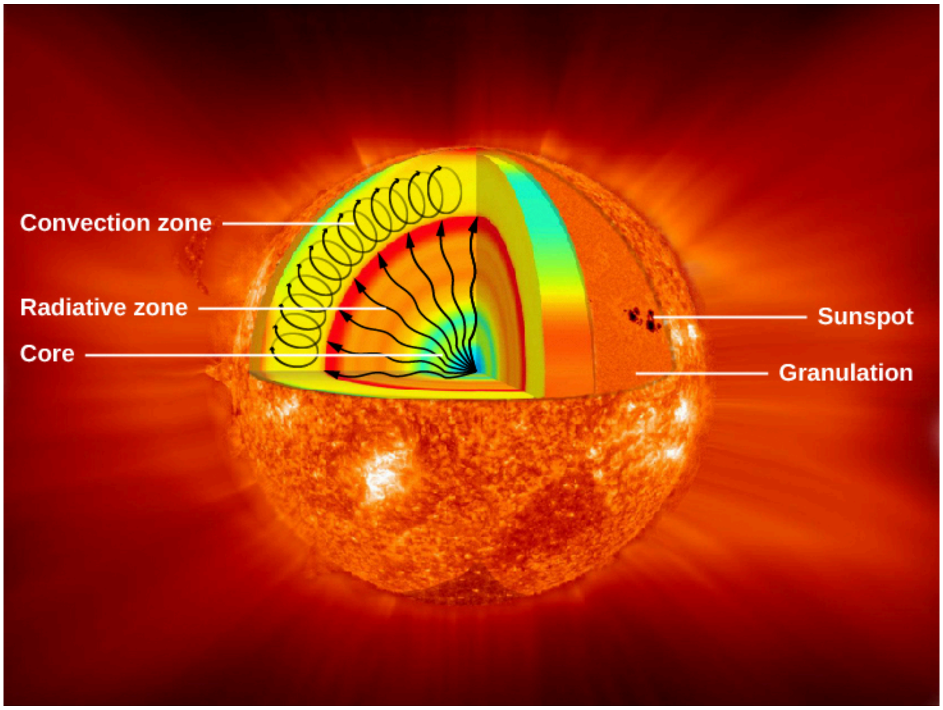


Figure 21.22 Energy is generated in the core by the fusion of hydrogen to form helium. This energy is transmitted outward by radiation—that is, by the absorption and reemission of photons. In the outermost layers, energy is transported mainly by convection. (credit: modification of work by NASA/Goddard)

Figure 21.23 shows how the temperature, density, rate of energy generation, and composition vary from the center of the Sun to its surface.

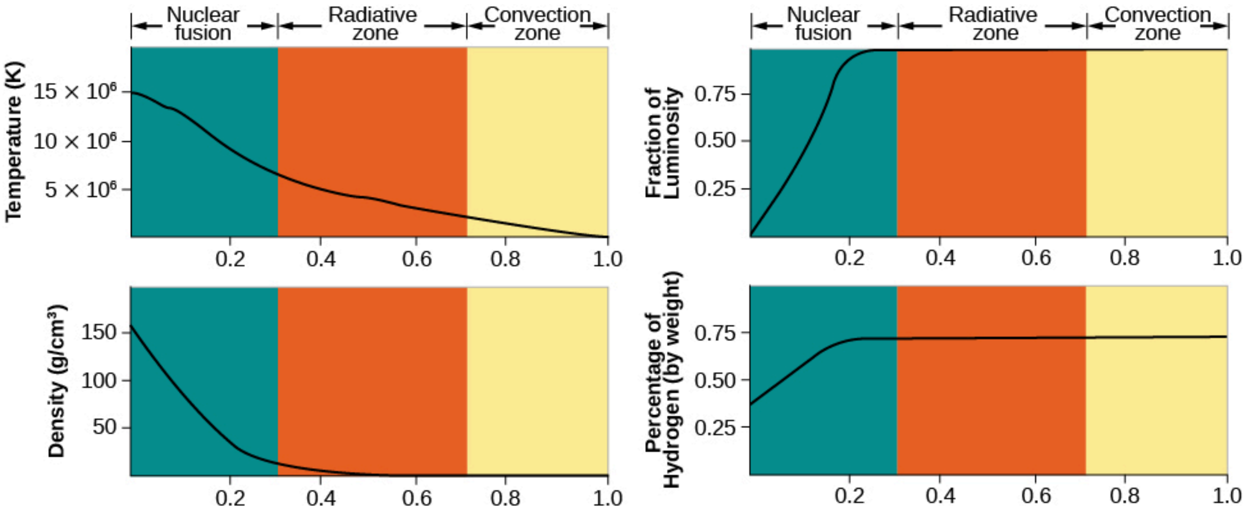


Figure 21.23 Diagrams showing how temperature, density, rate of energy generation, and the percentage (by mass) abundance of hydrogen vary inside the Sun. The horizontal scale shows the fraction of the Sun's radius: the left edge is the very center, and the right edge is the visible surface of the Sun, which is called the photosphere.

21.5 | The Solar Interior: Observations

Learning Objectives

By the end of this section, you will be able to:

- Explain how the Sun pulsates
- Explain what helioseismology is and what it can tell us about the solar interior
- Discuss how studying neutrinos from the Sun has helped understand neutrinos

Recall that when we observe the Sun's photosphere (the surface layer we see from the outside), we are not seeing very deeply into our star, certainly not into the regions where energy is generated. That's why the title of this section—observations of the solar interior—should seem very surprising. However, astronomers have indeed devised two types of measurements that can be used to obtain information about the inner parts of the Sun. One technique involves the analysis of tiny changes in the motion of small regions at the Sun's surface. The other relies on the measurement of the neutrinos emitted by the Sun.

Solar Pulsations

Astronomers discovered that the Sun pulsates—that is, it alternately expands and contracts—just as your chest expands and contracts as you breathe. This pulsation is very slight, but it can be detected by measuring the *radial velocity* of the solar surface—the speed with which it moves toward or away from us. The velocities of small regions on the Sun are observed to change in a regular way, first toward Earth, then away, then toward, and so on. It is as if the Sun were "breathing" through thousands of individual lungs, each having a size in the range of 4000 to 15,000 kilometers, each fluctuating back and forth (**Figure 21.24**).

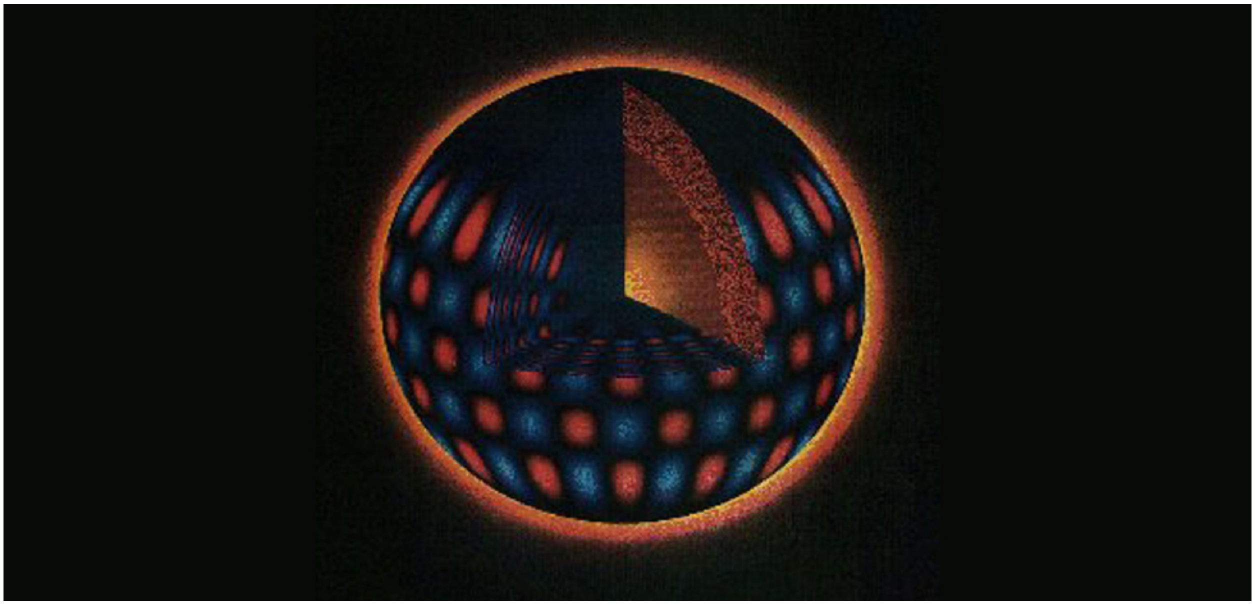


Figure 21.24 New observational techniques permit astronomers to measure small differences in velocity at the Sun's surface to infer what the deep solar interior is like. In this computer simulation, red shows surface regions that are moving away from the observer (inward motion); blue marks regions moving toward the observer (outward motion). Note that the velocity changes penetrate deep into the Sun's interior. (credit: modification of work by GONG, NOAO)

The typical velocity of one of the oscillating regions on the Sun is only a few hundred meters per second, and it takes about 5 minutes to complete a full cycle from maximum to minimum velocity and back again. The change in the size of the Sun measured at any given point is no more than a few kilometers.

The remarkable thing is that these small velocity variations can be used to determine what the interior of the Sun is like. The motion of the Sun's surface is caused by waves that reach it from deep in the interior. Study of the amplitude and cycle length of velocity changes provides information about the temperature, density, and composition of the layers through which the waves passed before they reached the surface. The situation is somewhat analogous to the use of seismic waves generated by earthquakes to infer the properties of Earth's interior. For this reason, studies of solar oscillations (back-and-forth motions) are referred to as **helioseismology**.

It takes a little over an hour for waves to traverse the Sun from center to surface, so the waves, like neutrinos, provide information about what the solar interior is like at the present time. In contrast, remember that the sunlight we see today emerging from the Sun was actually generated in the core several hundred thousand years ago.

Helioseismology has shown that convection extends inward from the surface 30% of the way toward the center; we have used this information in drawing **Figure 21.22**. Pulsation measurements also show that the *differential rotation* that we see at the Sun's surface, with the fastest rotation occurring at the equator, persists down through the convection zone. Below the convection zone, however, the Sun, even though it is gaseous throughout, rotates as if it were a solid body like a bowling ball. Another finding from helioseismology is that the abundance of helium inside the Sun, except in the center where nuclear reactions have converted hydrogen into helium, is about the same as at its surface. That result is important to astronomers because it means we are correct when we use the abundance of the elements measured in the solar atmosphere to construct models of the solar interior.

Helioseismology also allows scientists to look beneath a sunspot and see how it works. Sunspots are cool because strong magnetic fields block the outward flow of energy. **Figure 21.25** shows how gas moves around underneath a sunspot. Cool material from the sunspot flows downward, and material surrounding the sunspot is pulled inward, carrying magnetic field with it and thus maintaining the strong field that is necessary to form a sunspot. As the new material enters the sunspot region, it too cools, becomes denser, and sinks, thus setting up a self-perpetuating cycle that can last for weeks.

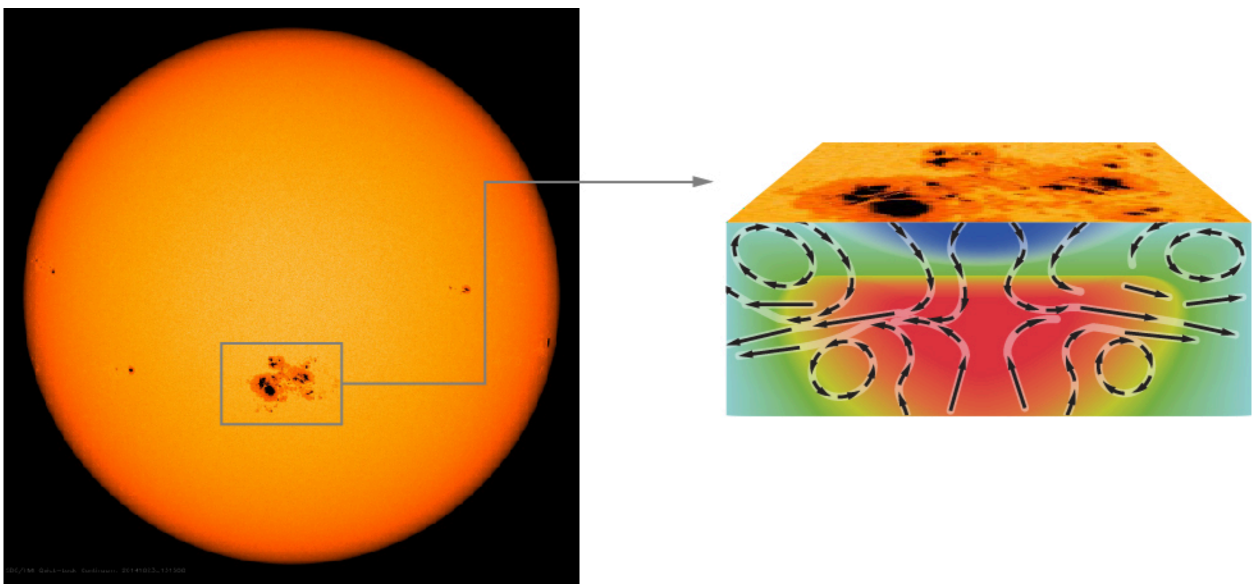


Figure 21.25 This drawing shows our new understanding, from helioseismology, of what lies beneath a sunspot. The black arrows show the direction of the flow of material. The intense magnetic field associated with the sunspot stops the upward flow of hot material and creates a kind of plug that blocks the hot gas. As the material above the plug cools (shown in blue), it becomes denser and plunges inward, drawing more gas and more magnetic field behind it into the spot. The concentrated magnetic field causes more cooling, thereby setting up a self-perpetuating cycle that allows a spot to survive for several weeks. Since the plug keeps hot material from flowing up into the sunspot, the region below the plug, represented by red in this picture, becomes hotter. This material flows sideways and then upward, eventually reaching the solar surface in the area surrounding the sunspot. (credit: modification of work by NASA, SDO)

The downward-flowing cool material acts as a kind of plug that block the upward flow of hot material, which is then diverted sideways and eventually reaches the solar surface in the region around the sunspot. This outward flow of hot material accounts for the paradox that the Sun emits slightly more energy when more of its surface is covered by cool sunspots.

Helioseismology has become an important tool for predicting solar storms that might impact Earth. Active regions can appear and grow large in only a few days. The solar rotation period is about 28 days. Therefore, regions capable of producing solar flares and coronal mass ejections can develop on the far side of the Sun, where, for a long time, we couldn't see them directly.

Fortunately, we now have space telescopes monitoring the Sun from all angles, so we know if there are sunspots forming on the opposite side of the Sun. Moreover, sound waves travel slightly faster in regions of high magnetic field, and waves generated in active regions traverse the Sun about 6 seconds faster than waves generated in quiet regions. By detecting this subtle difference, scientists can provide warnings of a week or more to operators of electric utilities and satellites about when a potentially dangerous active region might rotate into view. With this warning, it is possible to plan for disruptions, put key instruments into safe mode, or reschedule spacewalks in order to protect astronauts.

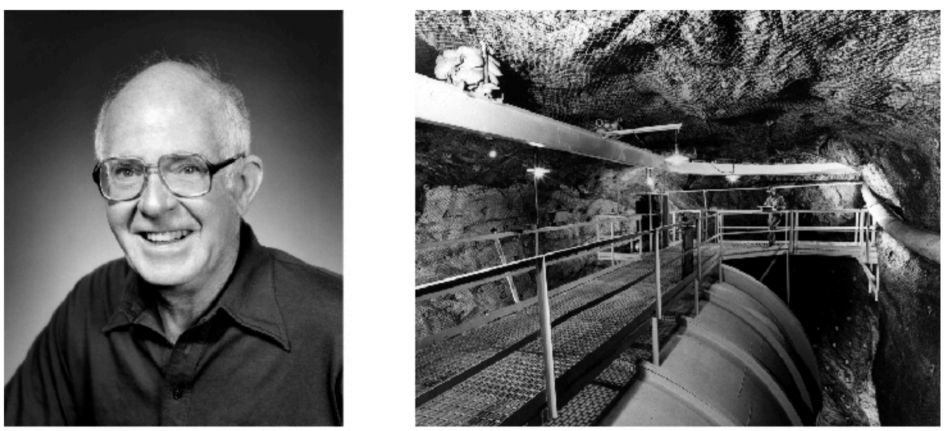
Solar Neutrinos

The second technique for obtaining information about the Sun's interior involves the detection of a few of those elusive neutrinos created during nuclear fusion. Recall from our earlier discussion that neutrinos created in the center of the Sun make their way directly out of the Sun and travel to Earth at nearly the speed of light. As far as neutrinos are concerned, the Sun is transparent.

About 3% of the total energy generated by nuclear fusion in the Sun is carried away by neutrinos. So many protons react and form neutrinos inside the Sun's core that, scientists calculate, 35 million billion (3.5×10^{16}) solar neutrinos pass through each square meter of Earth's surface every second. If we can devise a way to detect even a few of these solar neutrinos, then we can obtain information directly about what is going on in the center of the Sun. Unfortunately for those trying to "catch" some neutrinos, Earth and everything on it are also nearly transparent to passing neutrinos, just like the Sun.

On very, very rare occasions, however, one of the billions and billions of solar neutrinos will interact with another atom. The first successful detection of solar neutrinos made use of cleaning fluid (C_2Cl_4), which is the least expensive way to get a lot of chlorine atoms together. The nucleus of a chlorine (Cl) atom in the cleaning fluid can be turned into a radioactive argon nucleus by an interaction with a neutrino. Because the argon is radioactive, its presence can be detected. However, since the interaction of a neutrino with chlorine happens so rarely, a huge amount of chlorine is needed.

Raymond Davis, Jr. (**Figure 21.26**) and his colleagues at Brookhaven National Laboratory, placed a tank containing nearly 400,000 liters of cleaning fluid 1.5 kilometers beneath Earth's surface in a gold mine at Lead, South Dakota. A mine was chosen so that the surrounding material of Earth would keep cosmic rays (high-energy particles from space) from reaching the cleaning fluid and creating false signals. (Cosmic-ray particles are stopped by thick layers of Earth, but neutrinos find them of no significance.) Calculations show that solar neutrinos should produce about one atom of radioactive argon in the tank each day.



(a) (b)

Figure 21.26 (a) Raymond Davis received the Nobel Prize in physics in 2002. (b) Davis' experiment at the bottom of an abandoned gold mine first revealed problems with our understanding of neutrinos. (credit a: modification of work by Brookhaven National Laboratory; credit b: modification of work by the United States Department of Energy)

This was an amazing project: they counted argon atoms about once per month—and remember, they were looking for a tiny handful of argon atoms in a massive tank of chlorine atoms. When all was said and done, Davis' experiment, begun in 1970, detected only about one-third as many neutrinos as predicted by solar models! This was a shocking result because astronomers thought they had a pretty good understanding of both neutrinos and the Sun's interior. For many years, astronomers and physicists wrestled with Davis' results, trying to find a way out of the dilemma of the "missing" neutrinos.

Eventually Davis' result was explained by the surprising discovery that there are actually three types of neutrinos. Solar fusion produces only one type of neutrino, the so-called electron neutrino, and the initial experiments to detect solar neutrinos were designed to detect this one type. Subsequent experiments showed that these neutrinos change to a different type during their journey from the center of the Sun through space to Earth in a process called *neutrino oscillation*.

An experiment, conducted at the Sudbury Neutrino Observatory in Canada, was the first one designed to capture all three types of neutrinos (**Figure 21.27**). The experiment was located in a mine 2 kilometers underground. The neutrino detector consisted of a 12-meter-diameter transparent acrylic plastic sphere, which contained 1000 metric tons of heavy water. Remember that an ordinary water nucleus contains two hydrogen atoms and one oxygen atom. Heavy water instead contains two deuterium atoms and one oxygen atom, and incoming neutrinos can occasionally break up the loosely bound proton and neutron that make up the deuterium nucleus. The sphere of heavy water was surrounded by a shield of 1700 metric tons of very pure water, which in turn was surrounded by 9600 photomultipliers, devices that detect flashes of light produced after neutrinos interact with the heavy water.

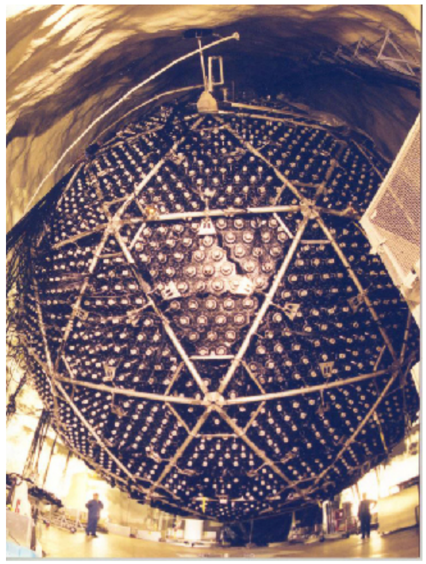


Figure 21.27 The 12-meter sphere of the Sudbury Neutrino Detector lies more than 2 kilometers underground and holds 1000 metric tons of heavy water. (credit: A.B. McDonald (Queen's University) et al., The Sudbury Neutrino Observatory Institute)

To the enormous relief of astronomers who make models of the Sun, the Sudbury experiment detected about 1 neutrino per hour and has shown that the *total* number of neutrinos reaching the heavy water is just what solar models predict. Only one-third of these, however, are electron neutrinos. It appears that two-thirds of the electron neutrinos produced by the Sun transform themselves into one of the other types of neutrinos as they make their way from the core of the Sun to Earth. This is why the earlier experiments saw only one-third the number of neutrinos expected.

Although it is not intuitively obvious, such neutrino oscillations can happen only if the mass of the electron neutrino is not zero. Other experiments indicate that its mass is tiny (even compared to the electron). The 2015 Nobel Prize in physics was awarded to researchers Takaaki Kajita and Arthur B. McDonald for their work establishing the changeable nature of neutrinos. (Raymond Davis shared the 2002 Nobel Prize with Japan's Masatoshi Koshiba for the experiments that led to our understanding of the neutrino problem in the first place.) But the fact that the neutrino has mass at all has deep implications for both physics and astronomy. For example, we will look at the role that neutrinos play in the inventory of the mass of the universe in the chapter on **Big Bang Cosmology**.

The Borexino experiment, an international experiment conducted in Italy, detected neutrinos coming from the Sun that were identified as coming from different reactions. Whereas the p-p chain is the reaction producing most of the Sun's energy, it is not the only nuclear reaction occurring in the Sun's core. There are side reactions involving nuclei of such elements as beryllium and boron. By probing the number of neutrinos that come from each reaction, the Borexino experiment has helped us confirm in detail our understanding of nuclear fusion in the Sun. In 2014, the Borexino experiment also identified neutrinos that were produced by the first step in the p-p chain, confirming the models of solar astronomers.

It's amazing that a series of experiments that began with enough cleaning fluid to fill a swimming pool brought down the shafts of an old gold mine is now teaching us about the energy source of the Sun and the properties of matter! This is a good example of how experiments in astronomy and physics, coupled with the best theoretical models we can devise, continue to lead to fundamental changes in our understanding of nature.

For Further Exploration

Websites

Albert Einstein Online: **http://www.westegg.com/einstein/ (http://www.westegg.com/einstein/)** .

Ghost Particle: **http://www.pbs.org/wgbh/nova/neutrino/ (http://www.pbs.org/wgbh/nova/neutrino/)** .

GONG Project Site: **http://gong.nso.edu/ (http://gong.nso.edu/)** .

Helioseismology: **http://solar-center.stanford.edu/about/helioseismology.html (http://solar-center.stanford.edu/about/helioseismology.html)** .

Princeton Plasma Physics Lab: **http://www.pppl.gov/ (http://www.pppl.gov/)** .

Solving the Mystery of the Solar Neutrinos: **http://www.nobelprize.org/nobel_prizes/themes/physics/bahcall/ (http://www.nobelprize.org/nobel_prizes/themes/physics/bahcall/)** .

Super Kamiokande Neutrino Mass Page: **http://www.ps.uci.edu/~superk/ (http://www.ps.uci.edu/~superk/)** .

Videos

Deep Secrets of the Neutrino: Physics Underground: **https://www.youtube.com/watch?v=Ar9ydagYkYg (https://www.youtube.com/watch?v=Ar9ydagYkYg)** . 2010 Public Lecture by Peter Rowson at the Stanford Linear Accelerator Center (1:22:00).

The Elusive Neutrino and the Nature of Physics: **https://www.youtube.com/watch?v=CBfUHzkcaHQ (https://www.youtube.com/watch?v=CBfUHzkcaHQ)** . Panel at the 2014 World Science Festival (1:30:00).

The Ghost Particle: **http://www.dailymotion.com/video/x20rn7s_nova-the-ghost-particle-discovery-science-universe-documentary_tv (http://www.dailymotion.com/video/x20rn7s_nova-the-ghost-particle-discovery-science-universe-documentary_tv)** . 2006 NOVA episode (52:49).

CHAPTER 21 REVIEW

KEY TERMS

aurora light radiated by atoms and ions in the ionosphere excited by charged particles from the Sun, mostly seen in the magnetic polar regions

chromosphere the part of the solar atmosphere that lies immediately above the photospheric layers

conduction process by which heat is directly transmitted through a substance when there is a difference of temperature between adjoining regions caused by atomic or molecular collisions

convection movement caused within a gas or liquid by the tendency of hotter, and therefore less dense material, to rise and colder, denser material to sink under the influence of gravity, which consequently results in transfer of heat

corona (of the Sun) the outer (hot) atmosphere of the Sun

coronal hole a region in the Sun's outer atmosphere that appears darker because there is less hot gas there

granulation the rice-grain-like structure of the solar photosphere; granulation is produced by upwelling currents of gas that are slightly hotter, and therefore brighter, than the surrounding regions, which are flowing downward into the Sun

helioseismology study of pulsations or oscillations of the Sun in order to determine the characteristics of the solar interior

hydrostatic equilibrium balance between the weights of various layers, as in a star or Earth's atmosphere, and the pressures that support them

nuclear fusion process of combining lighter nuclei to make heavier nuclei

nucleosynthesis process of fusion by which all elements on Earth are believed to have been created

photosphere the region of the solar (or stellar) atmosphere from which continuous radiation escapes into space

plasma a hot ionized gas

proton-proton chain combined reactions that fuse hydrogen nuclei to produce He nuclei

radiation emission of energy as electromagnetic waves or photons also the transmitted energy itself

solar wind a flow of hot, charged particles leaving the Sun

transition region the region in the Sun's atmosphere where the temperature rises very rapidly from the relatively low temperatures that characterize the chromosphere to the high temperatures of the corona

SUMMARY

21.1 The Structure and Composition of the Sun

- The Sun, our star, has several layers beneath the visible surface: the core, radiative zone, and convective zone.
- These, in turn, are surrounded by a number of layers that make up the solar atmosphere.
- In order of increasing distance from the center of the Sun, they are the photosphere, with a temperature that ranges from 4500 K to about 6800 K; the chromosphere, with a typical temperature of 10^4 K; the transition region, a zone that may be only a few kilometers thick, where the temperature increases rapidly from 10^4 K to 10^6 K; and the corona, with temperatures of a few million K.
- The Sun's surface is mottled with upwelling convection currents seen as hot, bright granules.
- Solar wind particles stream out into the solar system through coronal holes. When such particles reach the vicinity of Earth, they produce auroras, which are strongest near Earth's magnetic poles.
- Hydrogen and helium together make up 98% of the mass of the Sun, whose composition is much more characteristic of the universe at large than is the composition of Earth.

21.2 Sources of Sunshine: Thermal and Gravitational Energy?

- The Sun produces an enormous amount of energy every second.
- Since Earth and the solar system are roughly 4.5 billion years old, this means that the Sun has been producing vast amounts for energy for a very, very long time.
- Neither chemical burning nor gravitational contraction can account for the total amount of energy radiated by the Sun during all this time.

21.3 Source of Sunshine: Nuclear Fusion!

- Nuclear fusion is a reaction in which two nuclei are combined to form a larger nucleus; energy is released when light nuclei are fused to form medium-mass nuclei.
- The amount of energy released by a fusion reaction is known as the Q value.

21.4 The Solar Interior: Theory

- Even though we cannot see inside the Sun, it is possible to calculate what its interior must be like. As input for these calculations, we use what we know about the Sun.
- It is made entirely of hot gas.
- Apart from some very tiny changes, the Sun is neither expanding nor contracting (it is in hydrostatic equilibrium) and puts out energy at a constant rate.
- Fusion of hydrogen occurs in the center of the Sun, and the energy generated is carried to the surface by radiation and then convection.
- A solar model describes the structure of the Sun's interior. Specifically, it describes how pressure, temperature, mass, and luminosity depend on the distance from the center of the Sun.

21.5 The Solar Interior: Observations

- Studies of solar oscillations (helioseismology) and neutrinos can provide observational data about the Sun's interior.
- The technique of helioseismology has so far shown that the composition of the interior is much like that of the surface (except in the core, where some of the original hydrogen has been converted into helium), and that the convection zone extends about 30% of the way from the Sun's surface to its center.
- Helioseismology can also detect active regions on the far side of the Sun and provide better predictions of solar storms that may affect Earth.
- Neutrinos from the Sun call tell us about what is happening in the solar interior.
- A recent experiment has shown that solar models do predict accurately the number of electron neutrinos produced by nuclear reactions in the core of the Sun. However, two-thirds of these neutrinos are converted into different types of neutrinos during their long journey from the Sun to Earth, a result that also indicates that neutrinos are not massless particles.

CONCEPTUAL QUESTIONS

21.2 Sources of Sunshine: Thermal and Gravitational Energy?

1. Explain how we know that the Sun's energy is not supplied either by chemical burning, as in fires here on Earth, or by gravitational contraction (shrinking).

2. What is the ultimate source of energy that makes the Sun shine?

3. A friend who has not had the benefit of an astronomy course suggests that the Sun must be full of burning coal to shine as brightly as it does. List as many arguments as you can against this hypothesis.

21.3 Source of Sunshine: Nuclear Fusion!

4. Explain the difference between nuclear fission and nuclear fusion.

5. Why does the fusion of light nuclei into heavier nuclei release energy?

6. What are the formulas for the three steps in the proton-proton chain?

7. What conditions are required before proton-proton chain fusion can start in the Sun?

8. Which of the following transformations is (are) fusion and which is (are) fission: helium to carbon, carbon to iron, uranium to lead, boron to carbon, oxygen to neon? (See **Appendix F** for a list of the elements.)

9. Why is a higher temperature required to fuse hydrogen to helium by means of the CNO cycle than is required by the process that occurs in the Sun, which involves only isotopes of hydrogen and helium?

10. Do you think that nuclear fusion takes place in the atmospheres of stars? Why or why not?

11. Why is fission not an important energy source in the Sun?

21.4 The Solar Interior: Theory

12. Describe in your own words what is meant by the statement that the Sun is in hydrostatic equilibrium.

13. Describe the two main ways that energy travels through the Sun.

14. Neutrinos produced in the core of the Sun carry energy to its exterior. Is the mechanism for this energy transport conduction, convection, or radiation?

15. Earth's atmosphere is in hydrostatic equilibrium. What this means is that the pressure at any point in the atmosphere must be high enough to support the weight of air above it. How would you expect the pressure on Mt. Everest to differ from the pressure in your classroom? Explain why.

16. Explain what it means when we say that Earth's oceans are in hydrostatic equilibrium. Now suppose you are a scuba diver. Would you expect the pressure to increase or decrease as you dive below the surface to a depth of 200 feet? Why?

17. What mechanism transfers heat away from the surface of the Moon? If the Moon is losing energy in this way, why does it not simply become colder and colder?

18. Suppose you are standing a few feet away from a bonfire on a cold fall evening. Your face begins to feel hot. What is the mechanism that transfers heat from the fire to your face? (Hint: Is the air between you and the fire hotter or cooler than your face?)

19. Give some everyday examples of the transport of heat by convection and by radiation.

20. Why do you suppose so great a fraction of the Sun's energy comes from its central regions? Within what fraction of the Sun's radius does practically all of the Sun's luminosity originate (see **Figure 21.23**)? Within what radius of the Sun has its original hydrogen been partially used up? Discuss what relationship the answers to these questions bear to one another.

21. Explain how mathematical computer models allow us to understand what is going on inside of the Sun.

21.5 The Solar Interior: Observations

22. How do we know the age of the Sun?

23. How is a neutrino different from a neutron? List all the ways you can think of.

24. Two astronomy students travel to South Dakota. One stands on Earth's surface and enjoys some sunshine. At the same time, the other descends into a gold mine where neutrinos are detected, arriving in time to detect the creation of a new radioactive argon nucleus. Although the photon at the surface and the neutrinos in the mine arrive at the same time, they have had very different histories. Describe the differences.

25. What do measurements of the number of neutrinos emitted by the Sun tell us about conditions deep in the solar interior?

26. Do neutrinos have mass? Describe how the answer to this question has changed over time and why.

27. Someone suggests that astronomers build a special gamma-ray detector to detect gamma rays produced during the proton-proton chain in the core of the Sun, just like they built a neutrino detector. Explain why this would be a fruitless effort.

28. Earth contains radioactive elements whose decay produces neutrinos. How might we use neutrinos to determine how these elements are distributed in Earth's interior?

29. The Sun is much larger and more massive than Earth. Do you think the average density of the Sun is larger or smaller than that of Earth? Write down your answer before you look up the densities. Now find the values of the densities elsewhere in this text. Were you right? Explain clearly the meanings of density and mass.

30. Suppose the proton-proton cycle in the Sun were to slow down suddenly and generate energy at only 95% of its current rate. Would an observer on Earth see an immediate decrease in the Sun's brightness? Would she immediately see a decrease in the number of neutrinos emitted by the Sun?

PROBLEMS

21.3 Source of Sunshine: Nuclear Fusion!

31. Verify that the total number of nucleons, and total charge are conserved for each of the following fusion reactions in the proton-proton chain.

(i) $^{1}\text{H} + {}^{1}\text{H} \rightarrow {}^{2}\text{H} + e^{+} + \nu_{\text{e}}$,

(ii) $^{1}\text{H} + {}^{2}\text{H} \rightarrow {}^{3}\text{He} + \gamma$, and (iii) $^{3}\text{He} + {}^{3}\text{He} \rightarrow {}^{4}\text{He} + {}^{1}\text{H} + {}^{1}\text{H}$.

(List the value of each of the conserved quantities before and after each of the reactions.)

32. Calculate the energy output in each of the fusion reactions in the proton-proton chain, and verify the values determined in the preceding problem.

33. Show that the total energy released in the proton-proton chain is 26.7 MeV, considering the overall effect in $^{1}\text{H} + {}^{1}\text{H} \rightarrow {}^{2}\text{H} + e^{+} + \nu_{\text{e}}$, $^{1}\text{H} + {}^{2}\text{H} \rightarrow {}^{3}\text{He} + \gamma$, and $^{3}\text{He} + {}^{3}\text{He} \rightarrow {}^{4}\text{He} + {}^{1}\text{H} + {}^{1}\text{H}$. Be sure to include the annihilation energy.

34. Two fusion reactions mentioned in the text are $n + {}^{3}\text{He} \rightarrow {}^{4}\text{He} + \gamma$ and $n + {}^{1}\text{H} \rightarrow {}^{2}\text{H} + \gamma$. Both reactions release energy, but the second also creates more fuel. Confirm that the energies produced in the reactions are 20.58 and 2.22 MeV, respectively. Comment on which product nuclide is most tightly bound, ^{4}He or ^{2}H.

35. The power output of the Sun is $4 \times 10^{26}\,\text{W}$. (a) If 90% of this energy is supplied by the proton-proton chain, how many protons are consumed per second? (b) How many neutrinos per second should there be per square meter at the surface of Earth from this process?

36. Another set of reactions that fuses hydrogen into helium in the Sun and especially in hotter stars is called the CNO cycle:

$$^{12}\text{C} + {}^{1}\text{H} \rightarrow {}^{13}\text{N} + \gamma$$

$$^{13}\text{N} \rightarrow {}^{13}\text{C} + e^{+} + \nu_{\text{e}}$$

$$^{13}\text{C} + {}^{1}\text{H} \rightarrow {}^{14}\text{N} + \gamma$$

$$^{14}\text{N} + {}^{1}\text{H} \rightarrow {}^{15}\text{O} + \gamma$$

$$^{15}\text{O} \rightarrow {}^{15}\text{N} + e^{+} + \nu_{\text{e}}$$

$$^{15}\text{N} + {}^{1}\text{H} \rightarrow {}^{12}\text{C} + {}^{4}\text{He}$$

This process is a "cycle" because ^{12}C appears at the beginning and end of these reactions. Write down the overall effect of this cycle (as done for the proton-proton chain in $2e^{-} + 4\,{}^{1}\text{H} \rightarrow {}^{4}\text{He} + 2\nu_{\text{e}} + 6\gamma$). Assume that the positrons annihilate electrons to form more γ rays.

37. Estimate the amount of mass that is converted to energy when a proton combines with a deuterium nucleus to form ^{3}He.

38. How much energy is released when a proton combines with a deuterium nucleus to produce ^{3}He?

39. The Sun converts 4×10^{9} kg of mass to energy every second. How many years would it take the Sun to convert a mass equal to the mass of Earth to energy?

40. Assume that the mass of the Sun is 75% hydrogen and that all of this mass could be converted to energy according to Einstein's equation $E = mc^2$. How much total energy could the Sun generate? If m is in kg and c is in m/s, then E will be expressed in J. (The mass of the Sun is given in **Appendix D**.)

41. In fact, the conversion of mass to energy in the Sun is not 100% efficient. As we have seen in the text, the conversion of four hydrogen atoms to one helium atom results in the conversion of about 0.02862 times the mass of a proton to energy. How much energy in joules does one such reaction produce? (See **Appendix C** for the mass of the hydrogen atom, which, for all practical purposes, is the mass of a proton.)

42. Now suppose that all of the hydrogen atoms in the Sun were converted into helium. How much total energy would be produced? (To calculate the answer, you will have to estimate how many hydrogen atoms are in the Sun. This will give you good practice with scientific notation, since the numbers involved are very large!)

43. Show that the statement in the text is correct: namely, that roughly 600 million tons of hydrogen must be converted to helium in the Sun each second to explain its energy output. (Hint: Recall Einstein's most famous formula, and remember that for each kg of hydrogen, 0.0071 kg of mass is converted into energy.) How long will it be before 10% of the hydrogen is converted into helium?

44. Every second, the Sun converts 4 million tons of matter to energy. How long will it take the Sun to reduce its mass by 1% (the mass of the Sun is 2×10^{30} kg)? Compare your answer with the lifetime of the Sun so far.

21.4 The Solar Interior: Theory

45. Models of the Sun indicate that only about 10% of the total hydrogen in the Sun will participate in nuclear reactions, since it is only the hydrogen in the central regions that is at a high enough temperature. Use the total energy radiated per second by the Sun, 3.8×10^{26} watts, alongside the exercises and information given here to estimate the lifetime of the Sun. (Hint: Make sure you keep track of the units: if the luminosity is the energy radiated per second, your answer will also be in seconds. You should convert the answer to something more meaningful, such as years.)

21.5 The Solar Interior: Observations

46. Raymond Davis Jr.'s neutrino detector contained approximately 10^{30} chlorine atoms. During his experiment, he found that one neutrino reacted with a chlorine atom to produce one argon atom each day.

A. How many days would he have to run the experiment for 1% of his tank to be filled with argon atoms?
B. Convert your answer from A. into years.
C. Compare this answer to the age of the universe, which is approximately 14 billion years (1.4×10^{10} y).
D. What does this tell you about how frequently neutrinos interact with matter?

22 | STELLAR PROPERTIES

Figure 22.1 Stars come in a variety of sizes, masses, temperatures, and luminosities. This image shows part of a cluster of stars in the Small Magellanic Cloud (catalog number NGC 290). Located about 200,000 light-years away, NGC 290 is about 65 light-years across. Because the stars in this cluster are all at about the same distance from us, the differences in apparent brightness correspond to differences in luminosity; differences in temperature account for the differences in color. The various colors and luminosities of these stars provide clues about their life stories. (credit: modification of work by E. Olszewski (University of Arizona), European Space Agency, NASA)

Chapter Outline

Introduction

How do stars form? How long do they live? And how do they die? Stop and think how hard it is to answer these questions.

Stars live such a long time that nothing much can be gained from staring at one for a human lifetime. To discover how stars evolve from birth to death, it was necessary to measure the characteristics of many stars (to take a celestial census, in effect) and then determine which characteristics help us understand the stars' life stories. Astronomers tried a variety of hypotheses about stars until they came up with the right approach to understanding their development. But the key was first making a thorough census of the stars around us.

22.1 | Colors of Stars

Learning Objectives

By the end of this section, you will be able to:

- Compare the relative temperatures of stars based on their colors
- Understand how astronomers use color indexes to measure the temperatures of stars

Look at the beautiful picture of the stars in the Sagittarius Star Cloud shown in **Figure 22.2**. The stars show a multitude of colors, including red, orange, yellow, white, and blue. As we have seen, stars are not all the same color because they do not all have identical temperatures. To define *color* precisely, astronomers have devised quantitative methods for characterizing the color of a star and then using those colors to determine stellar temperatures. In the chapters that follow, we will provide the temperature of the stars we are describing, and this section tells you how those temperatures are determined from the colors of light the stars give off.

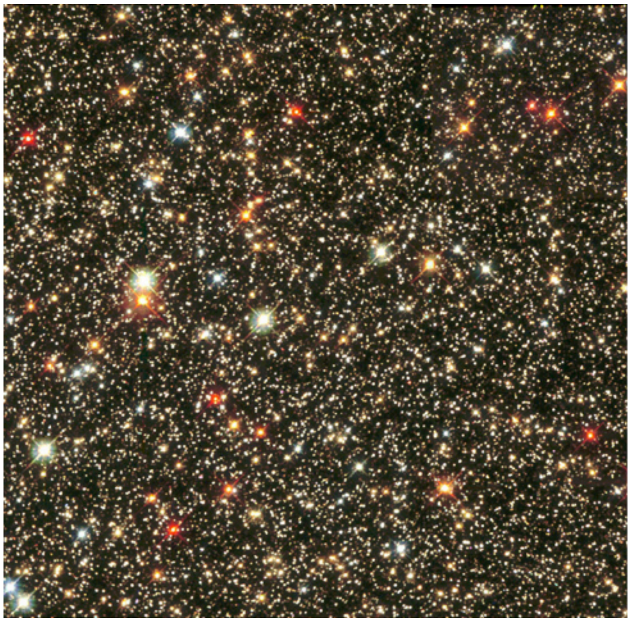


Figure 22.2 This image, which was taken by the Hubble Space Telescope, shows stars in the direction toward the center of the Milky Way Galaxy. The bright stars glitter like colored jewels on a black velvet background. The color of a star indicates its temperature. Blue-white stars are much hotter than the Sun, whereas red stars are cooler. On average, the stars in this field are at a distance of about 25,000 light-years (which means it takes light 25,000 years to traverse the distance from them to us) and the width of the field is about 13.3 light-years. (credit: Hubble Heritage Team (AURA/STScI/NASA))

Color and Temperature

As we learned in the section on **Blackbody Radiation** section, Wien's law relates stellar color to stellar temperature. Blue colors dominate the visible light output of very hot stars (with much additional radiation in the ultraviolet). On the other hand, cool stars emit most of their visible light energy at red wavelengths (with more radiation coming off in the infrared) (**Table 22.1**). The color of a star therefore provides a measure of its intrinsic or true surface temperature (apart from the effects of reddening by interstellar dust). Color does not depend on the distance to the object. This should be familiar to you from everyday experience. The color of a traffic signal, for example, appears the same no matter how far away it is. If we could somehow take a star, observe it, and then move it much farther away, its apparent brightness (magnitude) would change. But this change in brightness is the same for all wavelengths, and so its color would remain the same.

Example Star Colors and Corresponding Approximate Temperatures

Star Color	Approximate Temperature	Example
Blue	25,000 K	Spica
White	10,000 K	Vega
Yellow	6000 K	Sun
Orange	4000 K	Aldebaran
Red	3000 K	Betelgeuse

Table 22.1

Go to this **interactive simulation from the University of Colorado (https://openstax.org/l/30UofCsimstar)** to see the color of a star changing as the temperature is changed.

The hottest stars have temperatures of over 40,000 K, and the coolest stars have temperatures of about 2000 K. Our Sun's surface temperature is about 6000 K; its peak wavelength color is a slightly greenish-yellow. In space, the Sun would look white, shining with about equal amounts of reddish and bluish wavelengths of light. It looks somewhat yellow as seen from Earth's surface because our planet's nitrogen molecules scatter some of the shorter (i.e., blue) wavelengths out of the beams of sunlight that reach us, leaving more long wavelength light behind. This also explains why the sky is blue: the blue sky is sunlight scattered by Earth's atmosphere.

Color Indices

In order to specify the exact color of a star, astronomers normally measure a star's apparent brightness (discussed in **The Brightness of Stars**) through filters, each of which transmits only the light from a particular narrow band of wavelengths (colors). A crude example of a filter in everyday life is a green-colored, plastic, soft drink bottle, which, when held in front of your eyes, lets only the green colors of light through.

One commonly used set of filters in astronomy measures stellar brightness at three wavelengths corresponding to ultraviolet, blue, and yellow light. The filters are named: U (ultraviolet), B (blue), and V (visual, for yellow). These filters transmit light near the wavelengths of 360 nanometers (nm), 420 nm, and 540 nm, respectively. The brightness measured through each filter is usually expressed in magnitudes. The difference between any two of these magnitudes—say, between the blue and the visual magnitudes (B–V)—is called a **color index**.

Go to this **light and filters simulator (https://openstax.org/l/30lightfiltsim)** for a demonstration of how different light sources and filters can combine to determine the observed spectrum. You can also see how the perceived colors are associated with the spectrum.

By agreement among astronomers, the ultraviolet, blue, and visual magnitudes of the UBV system are adjusted to give a color index of 0 to a star with a surface temperature of about 10,000 K, such as Vega. The B–V color indexes of stars range from −0.4 for the bluest stars, with temperatures of about 40,000 K, to +2.0 for the reddest stars, with temperatures of about 2000 K. The B–V index for the Sun is about +0.65. Note that, by convention, the B–V index is always the "bluer" minus the "redder" color.

Why use a color index if it ultimately implies temperature? Because the brightness of a star through a filter is what astronomers actually measure, and we are always more comfortable when our statements have to do with measurable quantities.

22.2 | The Spectra of Stars

Learning Objectives

By the end of this section, you will be able to:

- Describe how astronomers use spectral classes to characterize stars
- Explain the difference between a star and a brown dwarf

Measuring colors is only one way of analyzing starlight. Another way is to use a spectrograph to spread out the light into a spectrum (see the **Spectroscopy** chapter). In 1814, the German physicist Joseph Fraunhofer observed that the spectrum of the Sun shows dark lines crossing a continuous band of colors. In the 1860s, English astronomers Sir William Huggins and Lady Margaret Huggins (**Figure 22.3**) succeeded in identifying some of the lines in stellar spectra as those of known elements on Earth, showing that the same chemical elements found in the Sun and planets exist in the stars. Since then, astronomers have worked hard to perfect experimental techniques for obtaining and measuring spectra, and they have developed a theoretical understanding of what can be learned from spectra. Today, spectroscopic analysis is one of the cornerstones of astronomical research.

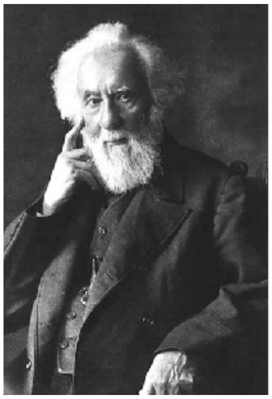



Figure 22.3 William and Margaret Huggins were the first to identify the lines in the spectrum of a star other than the Sun; they also took the first spectrogram, or photograph of a stellar spectrum.

Formation of Stellar Spectra

When the spectra of different stars were first observed, astronomers found that they were not all identical. Since the dark lines are produced by the chemical elements present in the stars, astronomers first thought that the spectra differ from one another because stars are not all made of the same chemical elements. This hypothesis turned out to be wrong. *The primary reason that stellar spectra look different is because the stars have different temperatures*. Most stars have nearly the same composition as the Sun, with only a few exceptions.

Hydrogen, for example, is by far the most abundant element in most stars. However, lines of hydrogen are not seen in the spectra of the hottest and the coolest stars. In the atmospheres of the hottest stars, hydrogen atoms are completely ionized. Because the electron and the proton are separated, ionized hydrogen cannot produce absorption lines. (Recall from the section on **Atomic Spectra** that the lines are the result of electrons in orbit around a nucleus changing energy levels.)

In the atmospheres of the coolest stars, hydrogen atoms have their electrons attached and can switch energy levels to produce lines. However, practically all of the hydrogen atoms are in the lowest energy state (unexcited) in these stars and thus can absorb only those photons able to lift an electron from that first energy level to a higher level. Photons with enough energy to do this lie in the ultraviolet part of the electromagnetic spectrum, and there are very few ultraviolet photons in the radiation from a cool star. What this means is that if you observe the spectrum of a very hot or very cool star with a typical telescope on the surface of Earth, the most common element in that star, hydrogen, will show very weak spectral lines or none at all.

The hydrogen lines in the visible part of the spectrum (called *Balmer lines*) are strongest in stars with intermediate temperatures—not too hot and not too cold. Calculations show that the optimum temperature for producing visible hydrogen lines is about 10,000 K. At this temperature, an appreciable number of hydrogen atoms are excited to the second energy level. They can then absorb additional photons, rise to still-higher levels of excitation, and produce a dark absorption line. Similarly, every other chemical element, in each of its possible stages of ionization, has a characteristic temperature at which it is most effective in producing absorption lines in any particular part of the spectrum.

Classification of Stellar Spectra

Astronomers use the patterns of lines observed in stellar spectra to sort stars into a **spectral class**. Because a star's temperature determines which absorption lines are present in its spectrum, these spectral classes are a measure of its surface temperature. There are seven standard spectral classes. From hottest to coldest, these seven spectral classes are designated O, B, A, F, G, K, and M. Recently, astronomers have added three additional classes for even cooler objects—L, T, and Y.

At this point, you may be looking at these letters with wonder and asking yourself why astronomers didn't call the spectral types A, B, C, and so on. You will see, as we tell you the history, that it's an instance where tradition won out over common sense.

In the 1880s, Williamina Fleming devised a system to classify stars based on the strength of hydrogen absorption lines. Spectra with the strongest lines were classified as "A" stars, the next strongest "B," and so on down the alphabet to "O" stars, in which the hydrogen lines were very weak. But we saw above that hydrogen lines alone are not a good indicator for classifying stars, since their lines disappear from the visible light spectrum when the stars get too hot or too cold.

In the 1890s, Annie Jump Cannon revised this classification system, focusing on just a few letters from the original system: A, B, F, G, K, M, and O. Instead of starting over, Cannon also rearranged the existing classes—in order of decreasing temperature—into the sequence we have learned: O, B, A, F, G, K, M. As you can read in the feature on **Annie Cannon: Classifier of the Stars** in this chapter, she classified around 500,000 stars over her lifetime, classifying up to three stars per minute by looking at the stellar spectra.

For a deep dive into spectral types, explore the interactive project at the **Sloan Digital Sky Survey (https://openstax.org/l/30sloandigsky)** in which you can practice classifying stars yourself.

To help astronomers remember this crazy order of letters, Cannon created a mnemonic, "Oh Be A Fine Girl, Kiss Me." (If you prefer, you can easily substitute "Guy" for "Girl.") The 21st-century version of this mnemonic might be "Only Boys Accepting Feminism Get Kissed Meaningfully." Other mnemonics, which we hope will not be relevant for you, include "Oh Brother, Astronomers Frequently Give Killer Midterms" and "Oh Boy, An F Grade Kills Me!" With the new L, T, and Y spectral classes, the mnemonic might be expanded to "Oh Be A Fine Girl (Guy), Kiss Me Like That, Yo!"

Each of these spectral classes, except possibly for the Y class which is still being defined, is further subdivided into 10 subclasses designated by the numbers 0 through 9. A B0 star is the hottest type of B star; a B9 star is the coolest type of B star and is only slightly hotter than an A0 star.

And just one more item of vocabulary: for historical reasons, astronomers call all the elements heavier than helium *metals*, even though most of them do not show metallic properties. (If you are getting annoyed at the peculiar jargon that astronomers use, just bear in mind that every field of human activity tends to develop its own specialized vocabulary. Just try reading a credit card or social media agreement form these days without training in law!)

Let's take a look at some of the details of how the spectra of the stars change with temperature. (It is these details that allowed Annie Cannon to identify the spectral types of stars as quickly as three per minute!) As **Figure 22.4** shows, in the hottest O stars (those with temperatures over 28,000 K), only lines of ionized helium and highly ionized atoms of other elements are conspicuous. Hydrogen lines are strongest in A stars with atmospheric temperatures of about 10,000 K. Ionized metals provide the most conspicuous lines in stars with temperatures from 6000 to 7500 K (spectral type F). In the coolest M stars (below 3500 K), absorption bands of titanium oxide and other molecules are very strong. By the way, the spectral class assigned to the Sun is G2. The sequence of spectral classes is summarized in **Table 22.2**.

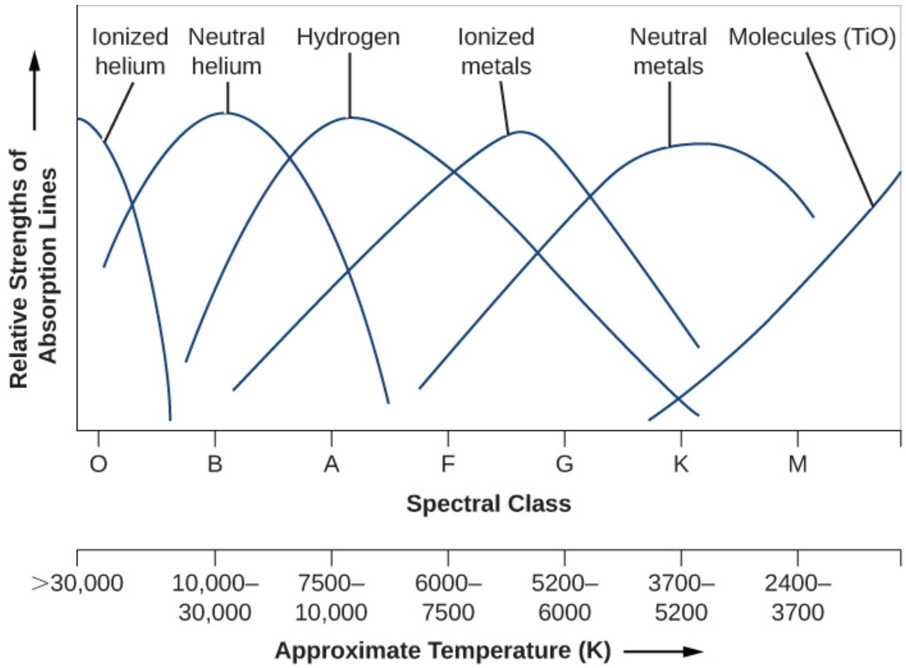


Figure 22.4 This graph shows the strengths of absorption lines of different chemical species (atoms, ions, molecules) as we move from hot (left) to cool (right) stars. The sequence of spectral types is also shown.

Spectral Classes for Stars

Spectral Class	Color	Approximate Temperature (K)	Principal Features	Examples
O	Blue	> 30,000	Neutral and ionized helium lines, weak hydrogen lines	10 Lacertae
B	Blue-white	10,000–30,000	Neutral helium lines, strong hydrogen lines	Rigel, Spica
A	White	7500–10,000	Strongest hydrogen lines, weak ionized calcium lines, weak ionized metal (e.g., iron, magnesium) lines	Sirius, Vega
F	Yellow-white	6000–7500	Strong hydrogen lines, strong ionized calcium lines, weak sodium lines, many ionized metal lines	Canopus, Procyon
G	Yellow	5200–6000	Weaker hydrogen lines, strong ionized calcium lines, strong sodium lines, many lines of ionized and neutral metals	Sun, Capella
K	Orange	3700–5200	Very weak hydrogen lines, strong ionized calcium lines, strong sodium lines, many lines of neutral metals	Arcturus, Aldebaran
M	Red	2400–3700	Strong lines of neutral metals and molecular bands of titanium oxide dominate	Betelgeuse, Antares
L	Red	1300–2400	Metal hydride lines, alkali metal lines (e.g., sodium, potassium, rubidium)	Teide 1

Table 22.2

Spectral Classes for Stars

Spectral Class	Color	Approximate Temperature (K)	Principal Features	Examples
T	Magenta	700–1300	Methane lines	Gliese 229B
Y	Infrared[1]	< 700	Ammonia lines	WISE 1828+2650

Table 22.2

To see how spectral classification works, let's use **Figure 22.4**. Suppose you have a spectrum in which the hydrogen lines are about half as strong as those seen in an A star. Looking at the lines in our figure, you see that the star could be either a B star or a G star. But if the spectrum also contains helium lines, then it is a B star, whereas if it contains lines of ionized iron and other metals, it must be a G star.

If you look at **Figure 22.5**, you can see that you, too, could assign a spectral class to a star whose type was not already known. All you have to do is match the pattern of spectral lines to a standard star (like the ones shown in the figure) whose type has already been determined.

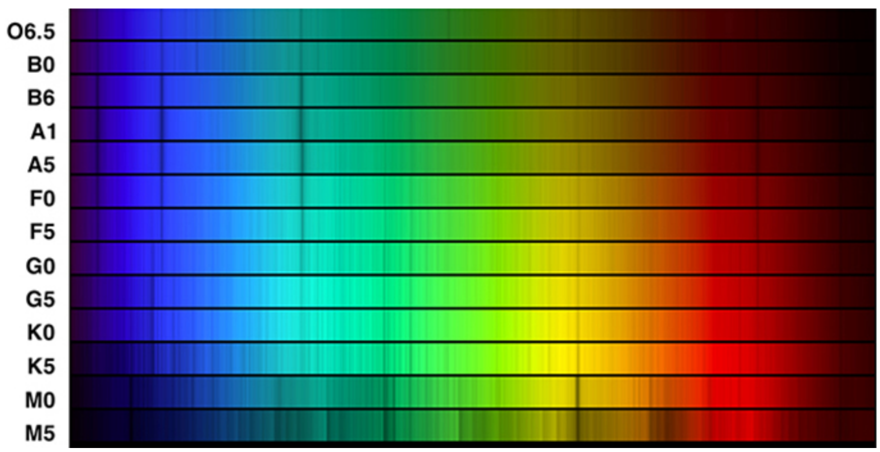


Figure 22.5 This image compares the spectra of the different spectral classes. The spectral class assigned to each of these stellar spectra is listed at the left of the picture. The strongest four lines seen at spectral type A1 (one in the red, one in the blue-green, and two in the blue) are Balmer lines of hydrogen. Note how these lines weaken at both higher and lower temperatures, as **Figure 22.4** also indicates. The strong pair of closely spaced lines in the yellow in the cool stars is due to neutral sodium (one of the neutral metals in **Figure 22.4**). (Credit: modification of work by NOAO/AURA/NSF)

Both colors and spectral classes can be used to estimate the temperature of a star. Spectra are harder to measure because the light has to be bright enough to be spread out into all colors of the rainbow, and detectors must be sensitive enough to respond to individual wavelengths. In order to measure colors, the detectors need only respond to the many wavelengths that pass simultaneously through the colored filters that have been chosen—that is, to *all* the blue light or *all* the yellow-green light.

Annie Cannon: Classifier of the Stars

Annie Jump Cannon was born in Delaware in 1863 (**Figure 22.6**). In 1880, she went to Wellesley College, one of the new breed of US colleges opening up to educate young women. Wellesley, only 5 years old at the time, had the second student physics lab in the country and provided excellent training in basic science. After college, Cannon spent a decade with her parents but was very dissatisfied, longing to do scientific work. After her mother's death in 1893,

1. Absorption by sodium and potassium atoms makes Y dwarfs appear a bit less red than L dwarfs.

she returned to Wellesley as a teaching assistant and also to take courses at Radcliffe, the women's college associated with Harvard.

Figure 22.6 Cannon is well-known for her classifications of stellar spectra. (credit: modification of work by Smithsonian Institution)

In the late 1800s, the director of the Harvard Observatory, Edward C. Pickering, needed lots of help with his ambitious program of classifying stellar spectra. The basis for these studies was a monumental collection of nearly a million photographic spectra of stars, obtained from many years of observations made at Harvard College Observatory in Massachusetts as well as at its remote observing stations in South America and South Africa. Pickering quickly discovered that educated young women could be hired as assistants for one-third or one-fourth the salary paid to men, and they would often put up with working conditions and repetitive tasks that men with the same education would not tolerate. These women became known as the Harvard Computers. (We should emphasize that astronomers were not alone in reaching such conclusions about the relatively new idea of upper-class, educated women working outside the home: women were exploited and undervalued in many fields. This is a legacy from which our society is just beginning to emerge.)

Cannon was hired by Pickering as one of the "computers" to help with the classification of spectra. She became so good at it that she could visually examine and determine the spectral types of several hundred stars per hour (dictating her conclusions to an assistant). She made many discoveries while investigating the Harvard photographic plates, including 300 variable stars (stars whose luminosity changes periodically). But her main legacy is a marvelous catalog of spectral types for hundreds of thousands of stars, which served as a foundation for much of twentieth-century astronomy.

In 1911, a visiting committee of astronomers reported that "she is the one person in the world who can do this work quickly and accurately" and urged Harvard to give Cannon an official appointment in keeping with her skill and renown. Not until 1938, however, did Harvard appoint her an astronomer at the university; she was then 75 years old.

Cannon received the first honorary degree Oxford awarded to a woman, and she became the first woman to be elected an officer of the American Astronomical Society, the main professional organization of astronomers in the US. She generously donated the money from one of the major prizes she had won to found a special award for women in astronomy, now known as the Annie Jump Cannon Prize. True to form, she continued classifying stellar spectra almost to the very end of her life in 1941.

Spectral Classes L, T, and Y

The scheme devised by Cannon worked well until 1988, when astronomers began to discover objects even cooler than M9-type stars. We use the word *object* because many of the new discoveries are not true stars. A star is defined as an object that during some part of its lifetime derives 100% of its energy from the same process that makes the Sun shine—the fusion of hydrogen nuclei (protons) into helium. Objects with masses less than about 7.5% of the mass of our Sun (about 0.075 M_{Sun}) do not become hot enough for hydrogen fusion to take place. Even before the first such "failed star" was found, this class of objects, with masses intermediate between stars and planets, was given the name **brown dwarfs**.

Brown dwarfs are very difficult to observe because they are extremely faint and cool, and they put out most of their light in the infrared part of the spectrum. It was only after the construction of very large telescopes, like the Keck telescopes in Hawaii, and the development of very sensitive infrared detectors, that the search for brown dwarfs succeeded. The first brown dwarf was discovered in 1988, and, as of the summer of 2015, there are more than 2200 known brown dwarfs.

Initially, brown dwarfs were given spectral classes like $M10^+$ or "much cooler than M9," but so many are now known that it is possible to begin assigning spectral types. The hottest brown dwarfs are given types L0–L9 (temperatures in the range 2400–1300 K), whereas still cooler (1300–700 K) objects are given types T0–T9 (see **Figure 22.7**). In class L brown dwarfs, the lines of titanium oxide, which are strong in M stars, have disappeared. This is because the L dwarfs are so cool that atoms and molecules can gather together into dust particles in their atmospheres; the titanium is locked up in the dust grains rather than being available to form molecules of titanium oxide. Lines of steam (hot water vapor) are present, along with lines of carbon monoxide and neutral sodium, potassium, cesium, and rubidium. Methane (CH_4) lines are strong in class-T brown dwarfs, as methane exists in the atmosphere of the giant planets in our own solar system.

In 2009, astronomers discovered ultra-cool brown dwarfs with temperatures of 500–600 K. These objects exhibited absorption lines due to ammonia (NH_3), which are not seen in T dwarfs. A new spectral class, Y, was created for these objects. As of 2015, over two dozen brown dwarfs belonging to spectral class Y have been discovered, some with temperatures comparable to that of the human body (about 300 K).

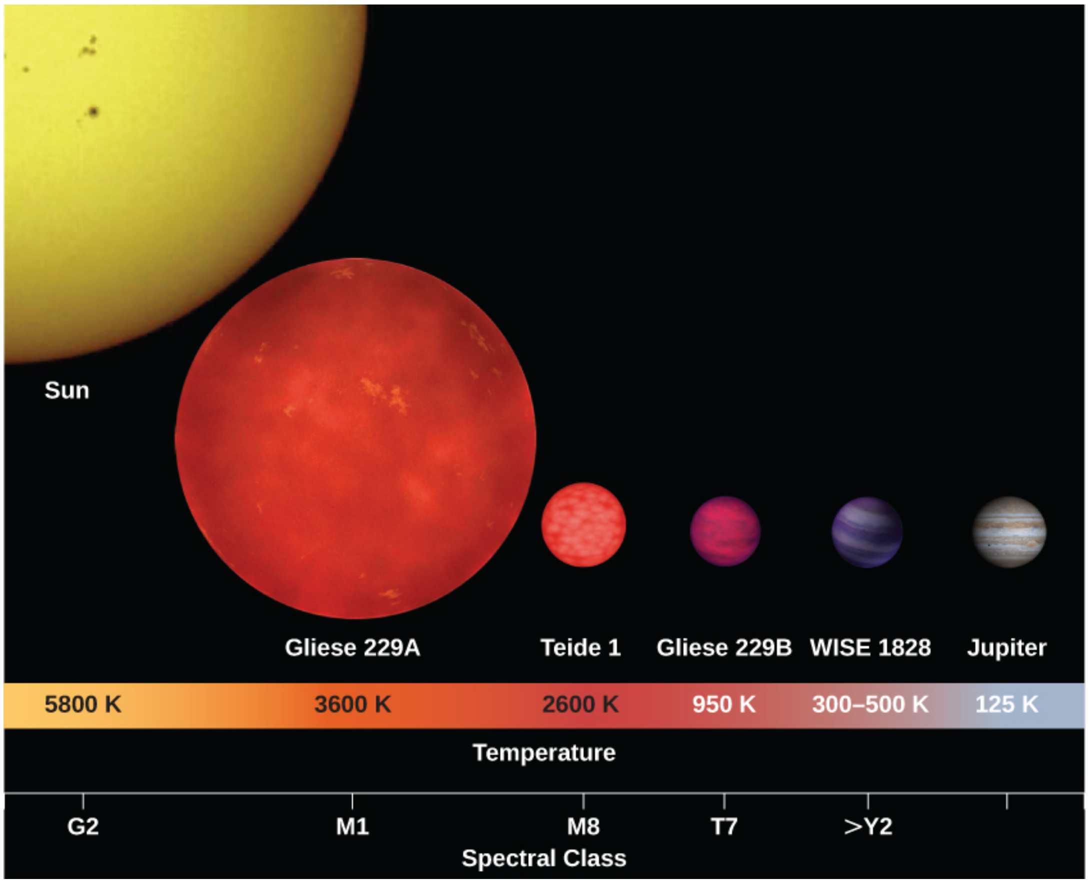


Figure 22.7 This illustration shows the sizes and surface temperatures of brown dwarfs Teide 1, Gliese 229B, and WISE1828 in relation to the Sun, a red dwarf star (Gliese 229A), and Jupiter. (credit: modification of work by MPIA/V. Joergens)

Most brown dwarfs start out with atmospheric temperatures and spectra like those of true stars with spectral classes of M6.5 and later, even though the brown dwarfs are not hot and dense enough in their interiors to fuse hydrogen. In fact, the spectra of brown dwarfs and true stars are so similar from spectral types late M through L that it is not possible to distinguish the two types of objects based on spectra alone. An independent measure of mass is required to determine whether a specific object is a brown dwarf or a very low mass star. Since brown dwarfs cool steadily throughout their lifetimes, the spectral type of a given brown dwarf changes with time over a billion years or more from late M through L, T, and Y spectral types.

Low-Mass Brown Dwarfs vs. High-Mass Planets

An interesting property of brown dwarfs is that they are all about the same radius as Jupiter, regardless of their masses. Amazingly, this covers a range of masses from about 13 to 80 times the mass of Jupiter (M_J). This can make distinguishing a low-mass brown dwarf from a high-mass planet very difficult.

So, what is the difference between a low-mass brown dwarf and a high-mass planet? The International Astronomical Union considers the distinctive feature to be *deuterium fusion*. Although brown dwarfs do not sustain regular (proton-proton) hydrogen fusion, they are capable of fusing deuterium (a rare form of hydrogen with one proton and one neutron in its nucleus). The fusion of deuterium can happen at a lower temperature than the fusion of hydrogen. If an object has enough mass to fuse deuterium (about 13 M_J or 0.012 M_{Sun}), it is a brown dwarf. Objects with less than 13 M_J do not fuse deuterium and are usually considered planets.

22.3 | Using Spectra to Measure Stellar Radius, Composition, and Motion

Learning Objectives

By the end of this section, you will be able to:

- Understand how astronomers can learn about a star's radius and composition by studying its spectrum
- Explain how astronomers can measure the motion and rotation of a star using the Doppler effect
- Describe the proper motion of a star and how it relates to a star's space velocity

Analyzing the spectrum of a star can teach us all kinds of things in addition to its temperature. We can measure its detailed chemical composition as well as the pressure in its atmosphere. From the pressure, we get clues about its size. We can also measure its motion toward or away from us and estimate its rotation.

Clues to the Size of a Star

As we shall see in **A Stellar Census**, stars come in a wide variety of sizes. At some periods in their lives, stars can expand to enormous dimensions. Stars of such exaggerated size are called **giants**. Luckily for the astronomer, stellar spectra can be used to distinguish giants from run-of-the-mill stars (such as our Sun).

Suppose you want to determine whether a star is a giant. A giant star has a large, extended photosphere. Because it is so large, a giant star's atoms are spread over a great volume, which means that the density of particles in the star's photosphere is low. As a result, the pressure in a giant star's photosphere is also low. This low pressure affects the spectrum in two ways. First, a star with a lower-pressure photosphere shows narrower spectral lines than a star of the same temperature with a higher-pressure photosphere (**Figure 22.8**). The difference is large enough that careful study of spectra can tell which of two stars at the same temperature has a higher pressure (and is thus more compressed) and which has a lower pressure (and thus must be extended). This effect is due to collisions between particles in the star's photosphere—more collisions lead to broader spectral lines. Collisions will, of course, be more frequent in a higher-density environment. Think about it like traffic—collisions are much more likely during rush hour, when the density of cars is high.

Second, more atoms are ionized in a giant star than in a star like the Sun with the same temperature. The ionization of atoms in a star's outer layers is caused mainly by photons, and the amount of energy carried by photons is determined by temperature. But how long atoms *stay* ionized depends in part on pressure. Compared with what happens in the Sun (with its relatively dense photosphere), ionized atoms in a giant star's photosphere are less likely to pass close enough to electrons to interact and combine with one or more of them, thereby becoming neutral again. Ionized atoms, as we discussed earlier, have different spectra from atoms that are neutral.

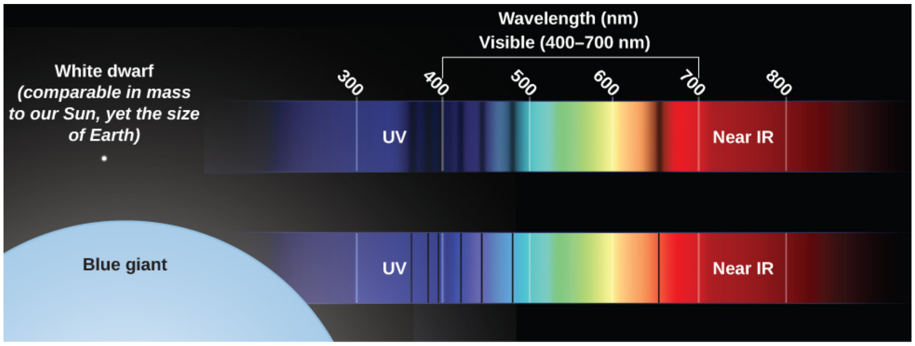


Figure 22.8 This figure illustrates one difference in the spectral lines from stars of the same temperature but different pressures. A giant star with a very-low-pressure photosphere shows very narrow spectral lines (bottom), whereas a smaller star with a higher-pressure photosphere shows much broader spectral lines (top). (credit: modification of work by NASA, ESA, A. Field, and J. Kalirai (STScI))

Abundances of the Elements

Absorption lines of a majority of the known chemical elements have now been identified in the spectra of the Sun and stars. If we see lines of iron in a star's spectrum, for example, then we know immediately that the star must contain iron.

Note that the *absence* of an element's spectral lines does not necessarily mean that the element itself is absent. As we saw, the temperature and pressure in a star's atmosphere will determine what types of atoms are able to produce absorption lines. Only if the physical conditions in a star's photosphere are such that lines of an element *should* (according to calculations) be there can we conclude that the absence of observable spectral lines implies low abundance of the element.

Suppose two stars have identical temperatures and pressures, but the lines of, say, sodium are stronger in one than in the other. Stronger lines mean that there are more atoms in the stellar photosphere absorbing light. Therefore, we know immediately that the star with stronger sodium lines contains more sodium. Complex calculations are required to determine exactly how much more, but those calculations can be done for any element observed in any star with any temperature and pressure.

Of course, astronomy textbooks such as ours always make these things sound a bit easier than they really are. If you look at the stellar spectra such as those in **Figure 22.5**, you may get some feeling for how hard it is to decode all of the information contained in the thousands of absorption lines. First of all, it has taken many years of careful laboratory work on Earth to determine the precise wavelengths at which hot gases of each element have their spectral lines. Long books and computer databases have been compiled to show the lines of each element that can be seen at each temperature. Second, stellar spectra usually have many lines from a number of elements, and we must be careful to sort them out correctly. Sometimes nature is unhelpful, and lines of different elements have identical wavelengths, thereby adding to the confusion. And third, as we saw in the section on **The Doppler Effect**, the motion of the star can change the observed wavelength of each of the lines. So, the observed wavelengths may not match laboratory measurements exactly. In practice, analyzing stellar spectra is a demanding, sometimes frustrating task that requires both training and skill.

Studies of stellar spectra have shown that hydrogen makes up about three-quarters of the mass of most stars. Helium is the second-most abundant element, making up almost a quarter of a star's mass. Together, hydrogen and helium make up from 96 to 99% of the mass; in some stars, they amount to more than 99.9%. Among the 4% or less of "heavy elements," oxygen, carbon, neon, iron, nitrogen, silicon, magnesium, and sulfur are among the most abundant. Generally, but not invariably, the elements of lower atomic weight are more abundant than those of higher atomic weight.

Take a careful look at the list of elements in the preceding paragraph. Two of the most abundant are hydrogen and oxygen (which make up water); add carbon and nitrogen and you are starting to write the prescription for the chemistry of an astronomy student. We are made of elements that are common in the universe—just mixed together in a far more sophisticated form (and a much cooler environment) than in a star.

As we mentioned in **The Spectra of Stars** section, astronomers use the term "metals" to refer to all elements heavier than hydrogen and helium. The fraction of a star's mass that is composed of these elements is referred to as the star's *metallicity*. The metallicity of the Sun, for example, is 0.02, since 2% of the Sun's mass is made of elements heavier than helium.

Appendix F lists how common each element is in the universe (compared to hydrogen); these estimates are based primarily on investigation of the Sun, which is a typical star. Some very rare elements, however, have not been detected in the Sun. Estimates of the amounts of these elements in the universe are based on laboratory measurements of their abundance in primitive meteorites, which are considered representative of unaltered material condensed from the solar nebula (see the **Formation of the Solar System** section).

Radial Velocity

When we measure the spectrum of a star, we determine the wavelength of each of its lines. If the star is not moving with respect to the Sun, then the wavelength corresponding to each element will be the same as those we measure in a laboratory here on Earth. But if stars are moving toward or away from us, we must consider the *Doppler effect* (see **The Doppler Effect** section). We should see all the spectral lines of moving stars shifted toward the red end of the spectrum if the star is moving away from us, or toward the blue (violet) end if it is moving toward us (**Figure 22.9**). The greater the shift, the faster the star is moving. Such motion, along the line of sight between the star and the observer, is called **radial velocity** and is usually measured in kilometers per second.

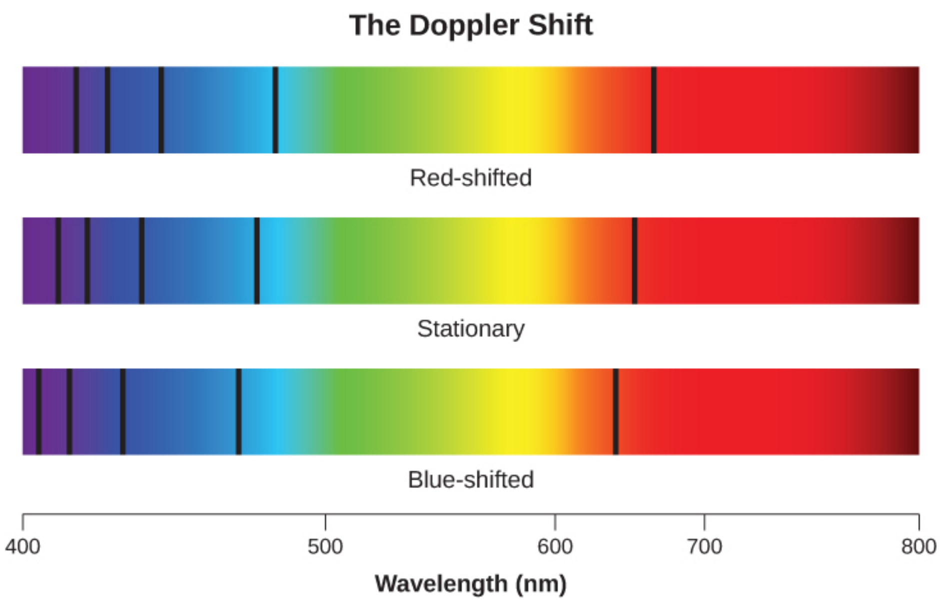


Figure 22.9 When the spectral lines of a moving star shift toward the red end of the spectrum, we know that the star is moving away from us. If they shift toward the blue end, the star is moving toward us.

William Huggins, pioneering yet again, in 1868 made the first radial velocity determination of a star. He observed the Doppler shift in one of the hydrogen lines in the spectrum of Sirius and found that this star is moving toward the solar system. Today, radial velocity can be measured for any star bright enough for its spectrum to be observed. As we will see in **A Stellar Census**, radial velocity measurements of double stars are crucial in deriving stellar masses.

Proper Motion

There is another type of motion stars can have that cannot be detected with stellar spectra. Unlike radial motion, which is along our line of sight (i.e., toward or away from Earth), this motion, called **proper motion**, is *transverse*: that is, across our line of sight. We see it as a change in the relative positions of the stars on the celestial sphere (**Figure 22.10**). These changes are very slow. Even the star with the largest proper motion takes 200 years to change its position in the sky by an amount equal to the width of the full Moon, and the motions of other stars are smaller yet.

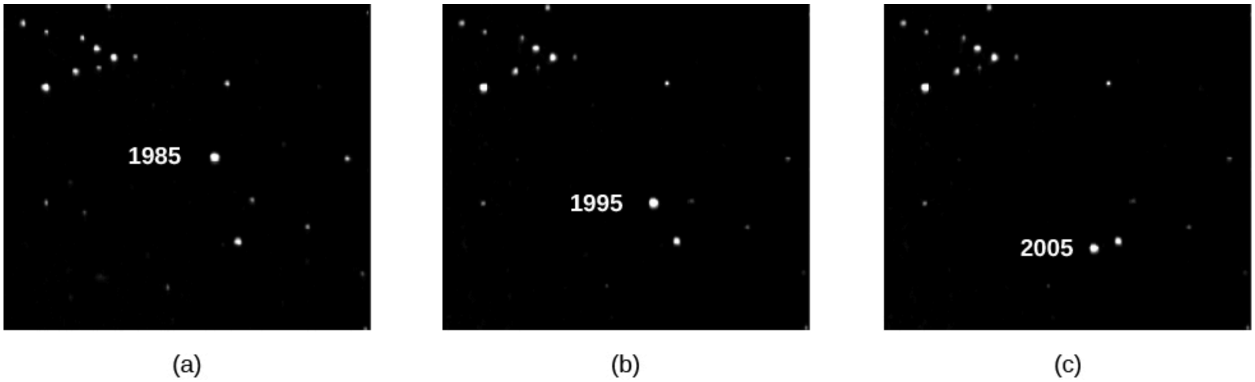


Figure 22.10 Three photographs of Barnard's star, the star with the largest known proper motion, show how this faint star has moved over a period of 20 years. (modification of work by Steve Quirk)

For this reason, with our naked eyes, we do not notice any change in the positions of the bright stars during the course of a human lifetime. If we could live long enough, however, the changes would become obvious. For example, some 50,000 years from now, terrestrial observers will find the handle of the Big Dipper unmistakably more bent than it is now (**Figure 22.11**).

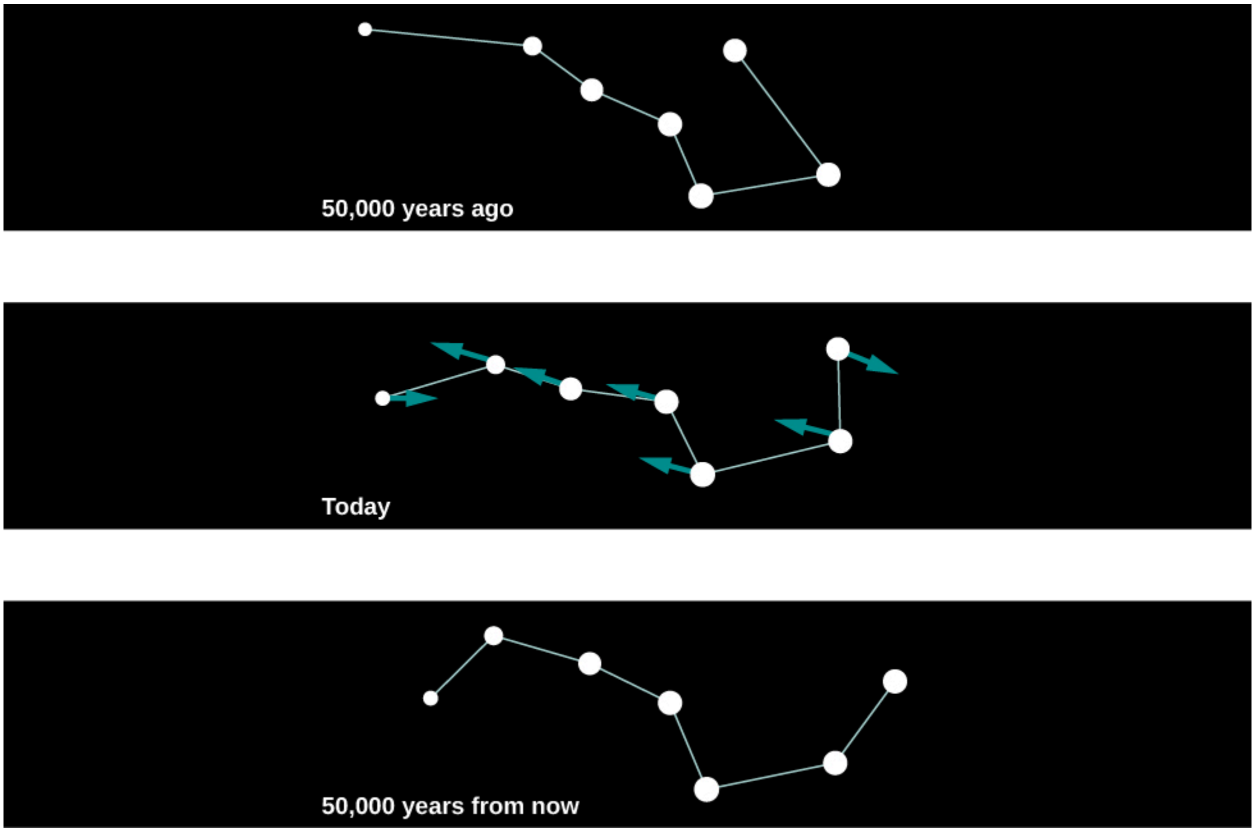


Figure 22.11 This figure shows changes in the appearance of the Big Dipper due to proper motion of the stars over 100,000 years.

We measure the proper motion of a star in arcseconds (1/3600 of a degree) per year. That is, the measurement of proper motion tells us only by how much of an angle a star has changed its position on the celestial sphere. If two stars at different distances are moving at the same velocity perpendicular to our line of sight, the closer one will show a larger shift in its position on the celestial sphere in a year's time. As an analogy, imagine you are standing at the side of a freeway. Cars will appear to whiz past you. If you then watch the traffic from a vantage point half a mile away, the cars will move much more slowly across your field of vision. In order to convert this angular motion to a velocity, we need to know how far away the star is.

To know the true **space velocity** of a star—that is, its total speed and the direction in which it is moving through space relative to the Sun—we must know its radial velocity, proper motion, and distance (Figure 22.12). A star's space velocity can also, over time, cause its distance from the Sun to change significantly. Over several hundred thousand years, these changes can be large enough to affect the apparent brightnesses of nearby stars. Today, Sirius, in the constellation Canis Major (the Big Dog) is the brightest star in the sky, but 100,000 years ago, the star Canopus in the constellation Carina (the Keel) was the brightest one. A little over 200,000 years from now, Sirius will have moved away and faded somewhat, and Vega, the bright blue star in Lyra, will take over its place of honor as the brightest star in Earth's skies.

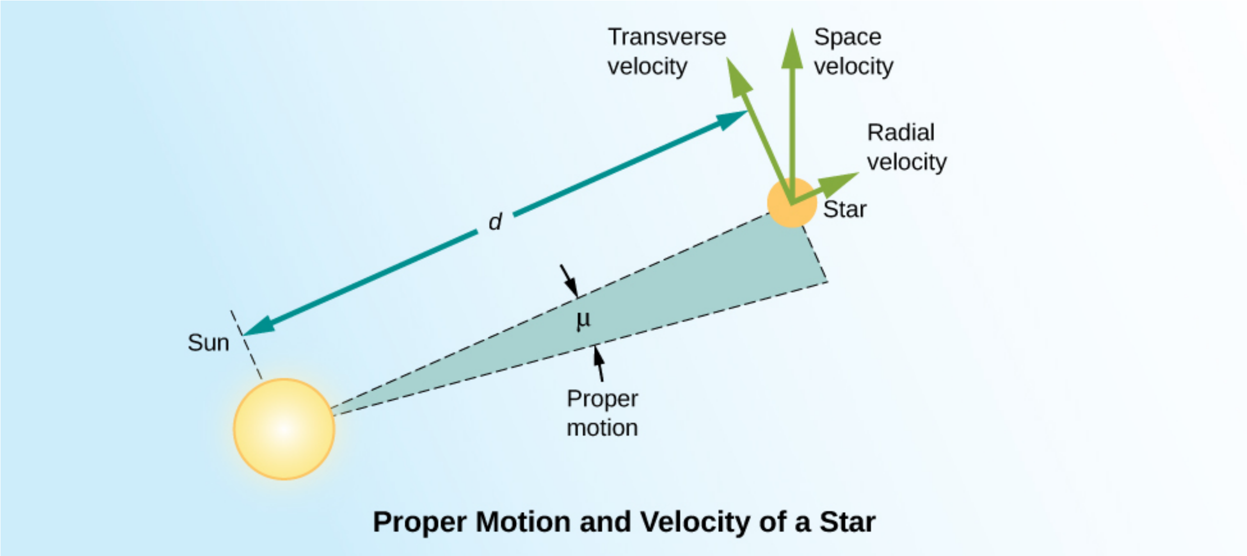


Figure 22.12 This figure shows the true space velocity of a star. The radial velocity is the component of the space velocity projected along the line of sight from the Sun to a star. The transverse velocity is a component of the space velocity projected on the sky. What astronomers measure is proper motion (μ), which is the change in the apparent direction on the sky measured in fractions of a degree. To convert this change in direction to a speed in, say, kilometers per second, it is necessary to also know the distance (*d*) from the Sun to the star.

Rotation

We can also use the Doppler effect to measure how fast a star rotates. If an object is rotating, then one of its sides is approaching us while the other is receding (unless its axis of rotation happens to be pointed exactly toward us). This is clearly the case for the Sun or a planet; we can observe the light from either the approaching or receding edge of these nearby objects and directly measure the Doppler shifts that arise from the rotation.

Stars, however, are so far away that they all appear as unresolved points. The best we can do is to analyze the light from the entire star at once. Due to the Doppler effect, the lines in the light that come from the side of the star rotating toward us are shifted to shorter wavelengths and the lines in the light from the opposite edge of the star are shifted to longer wavelengths. You can think of each spectral line that we observe as the sum or composite of spectral lines originating from different speeds with respect to us. Each point on the star has its own Doppler shift, so the absorption line we see from the whole star is actually much wider than it would be if the star were not rotating. If a star is rotating rapidly, there will be a greater spread of Doppler shifts and all its spectral lines should be quite broad. In fact, astronomers call this effect *line broadening*, and the amount of broadening can tell us the speed at which the star rotates (**Figure 22.13**).

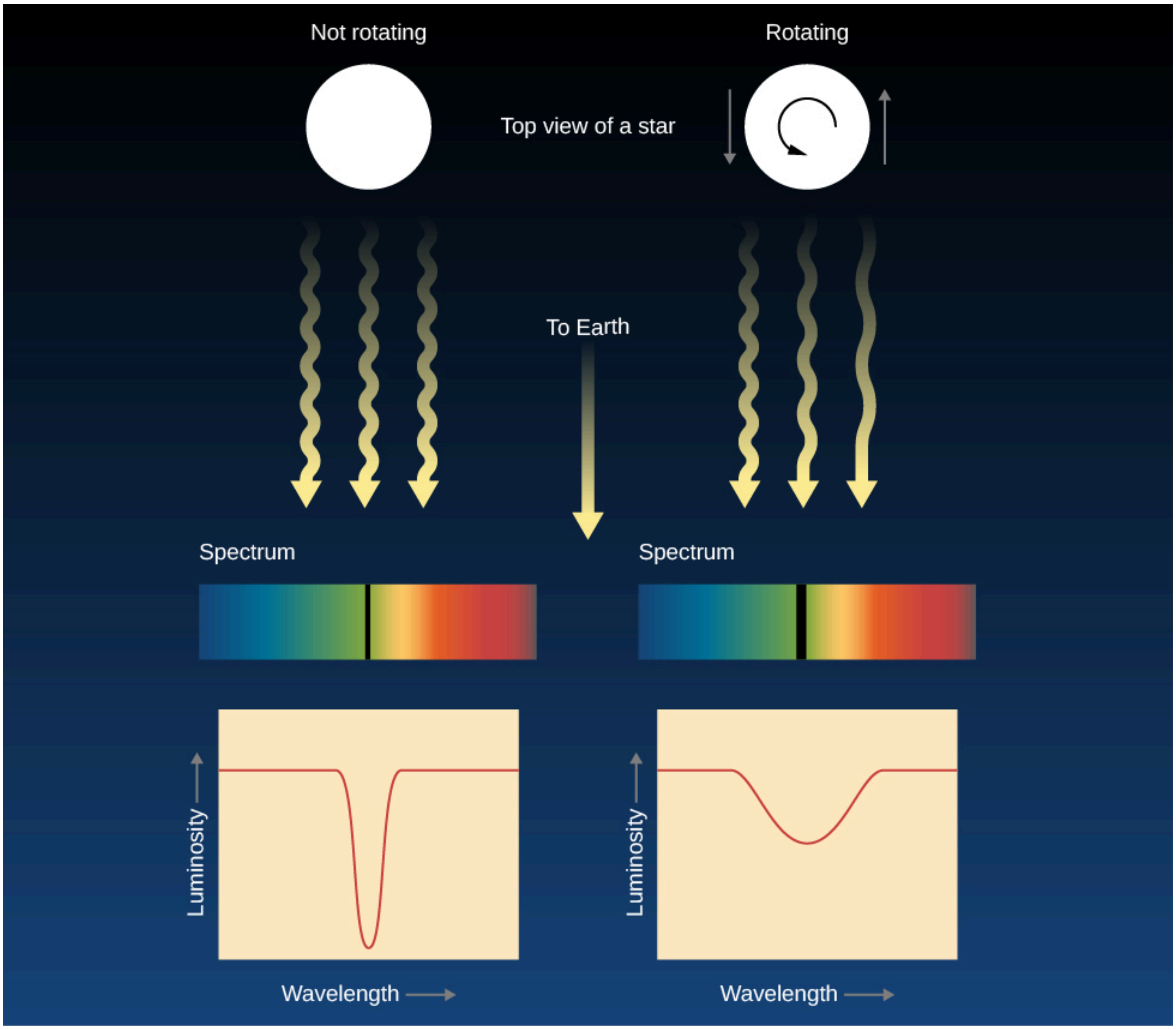


Figure 22.13 A rotating star will show broader spectral lines than a nonrotating star.

Measurements of the widths of spectral lines show that many stars rotate faster than the Sun, some with periods of less than a day! These rapid rotators spin so fast that their shapes are "flattened" into what we call *oblate spheroids*. An example of this is the star Vega, which rotates once every 12.5 hours. Vega's rotation flattens its shape so much that its diameter at the equator is 23% wider than its diameter at the poles (**Figure 22.14**). The Sun, with its rotation period of about a month, rotates rather slowly. Studies have shown that stars decrease their rotational speed as they age. Young stars rotate very quickly, with rotational periods of days or less. Very old stars can have rotation periods of several months.

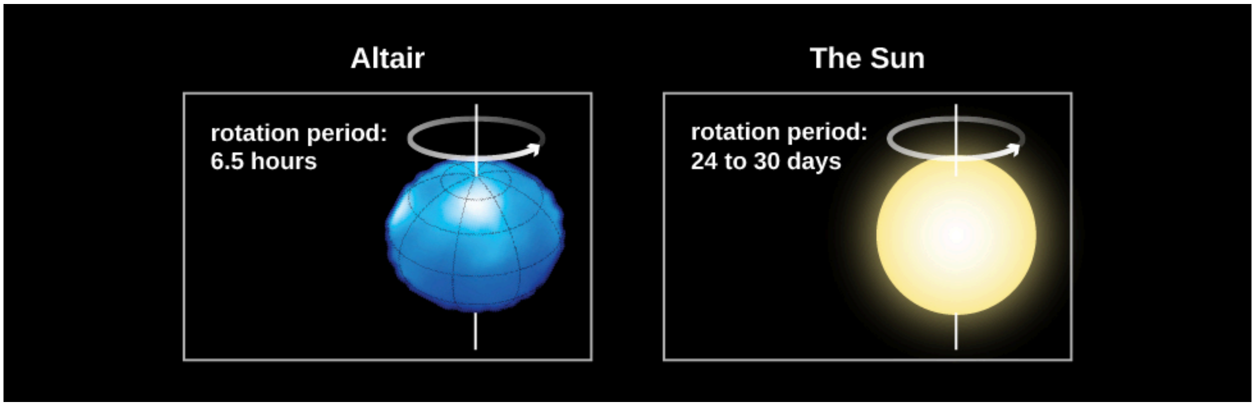


Figure 22.14 This illustration compares the more rapidly rotating star Altair to the slower rotating Sun.

As you can see, spectroscopy is an extremely powerful technique that helps us learn all kinds of information about stars that we simply could not gather any other way. We will see in later chapters that these same techniques can also teach us about galaxies, which are the most distant objects that can we observe. Without spectroscopy, we would know next to nothing about the universe beyond the solar system.

Astronomy and Philanthropy

Throughout the history of astronomy, contributions from wealthy patrons of the science have made an enormous difference in building new instruments and carrying out long-term research projects. Edward Pickering's stellar classification project, which was to stretch over several decades, was made possible by major donations from Anna Draper. She was the widow of Henry Draper, a physician who was one of the most accomplished amateur astronomers of the nineteenth century and the first person to successfully photograph the spectrum of a star. Anna Draper gave several hundred thousand dollars to Harvard Observatory. As a result, the great spectroscopic survey is still known as the Henry Draper Memorial, and many stars are still referred to by their "HD" numbers in that catalog (such as HD 209458).

In the 1870s, the eccentric piano builder and real estate magnate James Lick (**Figure 22.15**) decided to leave some of his fortune to build the world's largest telescope. When, in 1887, the pier to house the telescope was finished, Lick's body was entombed in it. Atop the foundation rose a 36-inch refractor, which for many years was the main instrument at the Lick Observatory near San Jose.

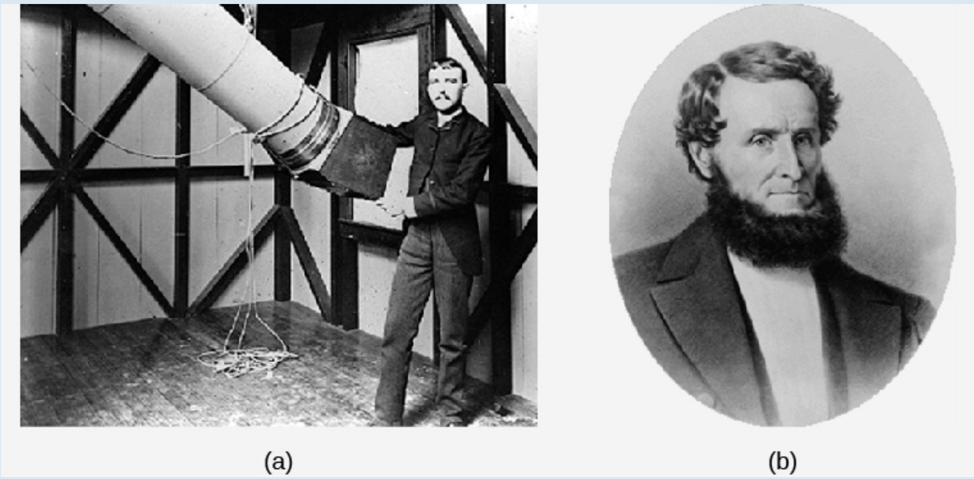


Figure 22.15 (a) Draper stands next to a telescope used for photography. After his death, his widow funded further astronomy work in his name. (b) Lick was a philanthropist who provided funds to build a 36-inch refractor not only as a memorial to himself but also to aid in further astronomical research.

The Lick telescope remained the largest in the world until 1897, when George Ellery Hale persuaded railroad millionaire Charles Yerkes to finance the construction of a 40-inch telescope near Chicago. More recently, Howard Keck, whose family made its fortune in the oil industry, gave $70 million from his family foundation to the California Institute of Technology to help build the world's largest telescope atop the 14,000-foot peak of Mauna Kea in Hawaii. The Keck Foundation was so pleased with what is now called the Keck telescope that they gave $74 million more to build Keck II, another 10-meter reflector on the same volcanic peak.

Now, if any of you become millionaires or billionaires, and astronomy has sparked your interest, do keep an astronomical instrument or project in mind as you plan your estate. But frankly, private philanthropy could not possibly support the full enterprise of scientific research in astronomy. Much of our exploration of the universe is financed by federal agencies such as the National Science Foundation and NASA in the United States, and by similar government agencies in the other countries. In this way, all of us, through a very small share of our tax dollars, are philanthropists for astronomy.

22.4 | A Stellar Census

Learning Objectives

By the end of this section, you will be able to:

- Explain why the stars visible to the unaided eye are not typical
- Describe the distribution of stellar masses found close to the Sun

Before we can make our own survey, we need to agree on a unit of distance appropriate to the objects we are studying. The stars are all so far away that kilometers (and even astronomical units) would be very cumbersome to use; so—as discussed in **The Universe at its Limits**—astronomers use a much larger "measuring stick" called the *light-year*. A light-year is the distance that light (the fastest signal we know) travels in 1 year. Since light covers an astounding 300,000 kilometers per second, and since there are a lot of seconds in 1 year, a light-year is a very large quantity: 9.5 trillion (9.5×10^{12}) kilometers to be exact. (Bear in mind that the light-year is a unit of *distance* even though the term *year* appears in it.) If you drove at the legal US speed limit without stopping for food or rest, you would not arrive at the end of a light-year in space until roughly 12 million years had passed. And the closest star is more than 4 light-years away.

Notice that we have not yet said much about how such enormous distances can be measured. That is a complicated question, to which we will return in **Celestial Distances**. For now, let us assume that distances have been measured for stars in our cosmic vicinity so that we can proceed with our census.

Small Is Beautiful—Or at Least More Common

When we do a census of people in the United States, we count the inhabitants by neighborhood. We can try the same approach for our stellar census and begin with our own immediate neighborhood. As we shall see, we run into two problems—just as we do with a census of human beings. First, it is hard to be sure we have counted *all* the inhabitants; second, our local neighborhood may not contain all possible types of people.

Table 22.3 shows an estimate of the number of stars of each spectral type[2] in our own local neighborhood—within 21 light-years of the Sun. (The Milky Way Galaxy, in which we live, is about 100,000 light-years in diameter, so this figure really applies to a *very* local neighborhood, one that contains a *tiny* fraction of all the billions of stars in the Milky Way.) You can see that there are many more low-luminosity (and hence low mass) stars than high-luminosity ones. Only three of the stars in our local neighborhood (one F type and two A types) are significantly more luminous and more massive than the Sun. This is truly a case where small triumphs over large—at least in terms of numbers. The Sun is more massive than the vast majority of stars in our vicinity.

2. The spectral types of stars were defined and discussed in Analyzing Starlight (https://legacy.cnx.org/content/m59889/latest/) .

Stars within 21 Light-Years of the Sun

Spectral Type	Number of Stars
A	2
F	1
G	7
K	17
M	94
White dwarfs	8
Brown dwarfs	33

Table 22.3

This table is based on data published through 2015, and it is likely that more faint objects remain to be discovered (see **Figure 22.16**). Along with the L and T brown dwarfs already observed in our neighborhood, astronomers expect to find perhaps hundreds of additional T dwarfs. Many of these are likely to be even cooler than the coolest currently known T dwarf. The reason the lowest-mass dwarfs are so hard to find is that they put out very little light—ten thousand to a million times less light than the Sun. Only recently has our technology progressed to the point that we can detect these dim, cool objects.

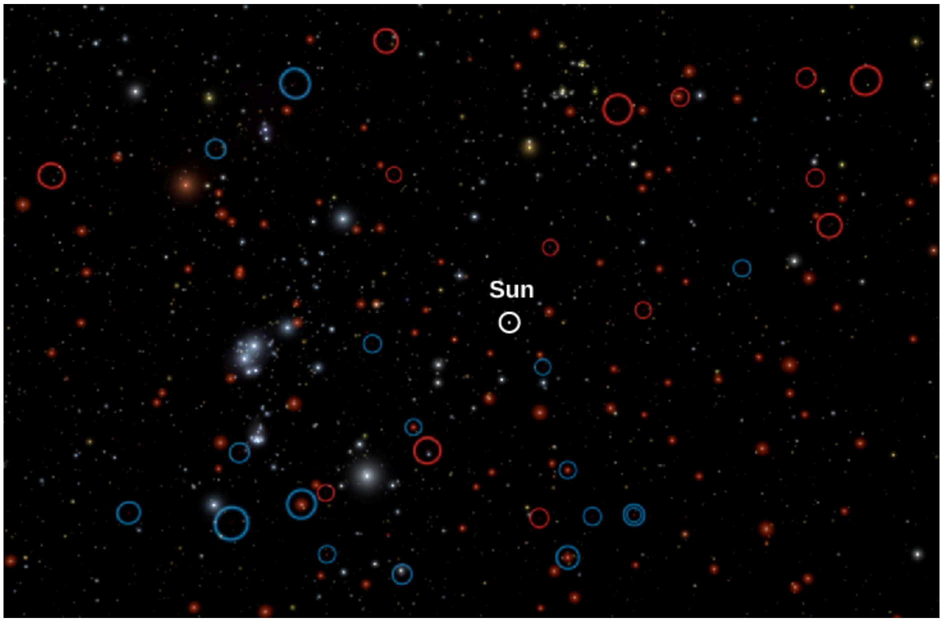


Figure 22.16 This computer simulation shows the stars in our neighborhood as they would be seen from a distance of 30 light-years away. The Sun is in the center. All the brown dwarfs are circled; those found earlier are circled in blue, the ones found recently with the WISE infrared telescope in space (whose scientists put this diagram together) are circled in red. The common M stars, which are red and faint, are made to look brighter than they really would be so that you can see them in the simulation. Note that luminous hot stars like our Sun are very rare. (credit: modification of work by NASA/ JPL-Caltech)

To put all this in perspective, we note that even though the stars counted in the table are our closest neighbors, you can't just look up at the night sky and see them without a telescope; stars fainter than the Sun cannot be seen with the unaided eye unless they are *very* nearby. For example, stars with luminosities ranging from 1/100 to 1/10,000 the luminosity of the Sun (L_{Sun}) are very common, but a star with a luminosity of 1/100 L_{Sun} would have to be within 5 light-years to be visible to the naked eye—and only three stars (all in one system) are this close to us. The nearest of these three stars, Proxima Centauri, still cannot be seen without a telescope because it has such a low luminosity.

Astronomers are working hard these days to complete the census of our local neighborhood by finding our faintest neighbors. Recent discoveries of nearby stars have relied heavily upon infrared telescopes that are able to find these many cool, low-mass stars. You should expect the number of known stars within 21 light-years of the Sun to keep increasing as more and better surveys are undertaken.

Remember: Bright Does Not Necessarily Mean Close

If we confine our census to the local neighborhood, we will miss many of the most interesting kinds of stars. After all, the neighborhood in which you live does not contain all the types of people—distinguished according to age, education, income, race, and so on—that live in the entire country. For example, a few people do live to be over 100 years old, but there may be no such individual within several miles of where you live. In order to sample the full range of the human population, you would have to extend your census to a much larger area. Similarly, some types of stars simply are not found nearby.

A clue that we are missing something in our stellar census comes from the fact that only six of the 20 stars that appear brightest in our sky— Sirius, Vega, Altair, Alpha Centauri, Fomalhaut, and Procyon—are found within 26 light-years of the Sun (**Figure 22.17**). Why are we missing most of the brightest stars when we take our census of the local neighborhood?

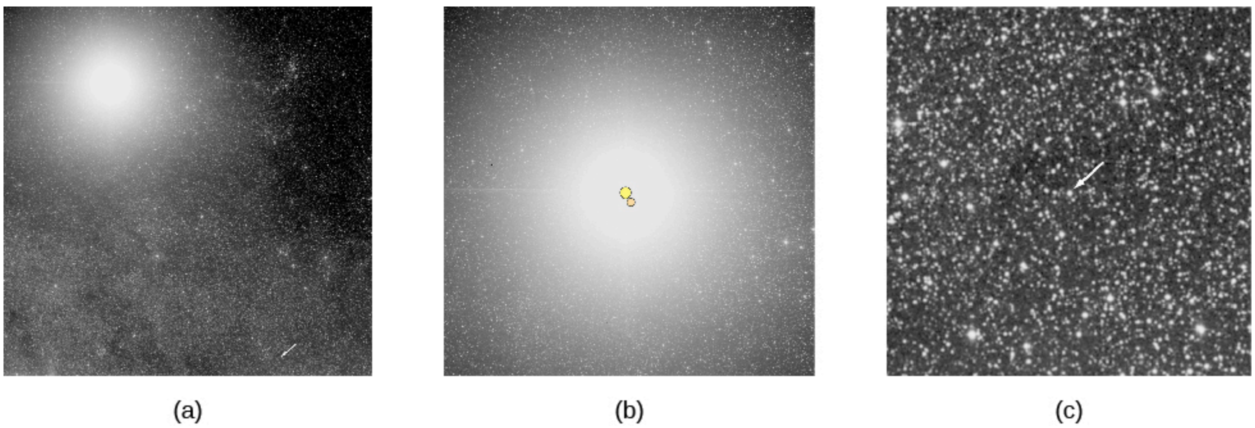


Figure 22.17 (a) This image, taken with a wide-angle telescope at the European Southern Observatory in Chile, shows the system of three stars that is our nearest neighbor. (b) Two bright stars that are close to each other (Alpha Centauri A and B) blend their light together. (c) Indicated with an arrow (since you'd hardly notice it otherwise) is the much fainter Proxima Centauri star, which is spectral type M. (credit: modification of work by ESO)

The answer, as we examined in **The Brightness of Stars**, is that the stars that appear brightest are *not* the ones closest to us. The brightest stars look the way they do because they emit a very large amount of energy—so much, in fact, that they do not have to be nearby to look brilliant. You can confirm this by looking at **Appendix D**, which gives distances for the 20 stars that appear brightest from Earth. The most distant of these stars is more than *1000 light-years* from us. In fact, it turns out that most of the stars visible without a telescope are hundreds of light-years away and many times more luminous than the Sun. Among the 9000 stars visible to the unaided eye, only about 50 are intrinsically fainter than the Sun. Note also that several of the stars in **Appendix D** are spectral type B, a type that is completely missing from **Table 22.3**.

The most luminous of the bright stars listed in **Appendix D** emit more than 50,000 times more energy than does the Sun. These highly luminous stars are missing from the solar neighborhood because they are very rare. None of them happens to be in the tiny volume of space immediately surrounding the Sun, and only this small volume was surveyed to get the data shown in **Table 22.3**.

For example, let's consider the most luminous stars—those 100 or more times as luminous as the Sun. Although such stars are rare, they are visible to the unaided eye, even when hundreds to thousands of light-years away. A star with a luminosity 10,000 times greater than that of the Sun can be seen without a telescope out to a distance of 5000 light-years. The volume of space included within a distance of 5000 light-years, however, is enormous; so even though highly luminous stars are intrinsically rare, many of them are readily visible to our unaided eye.

The contrast between these two samples of stars, those that are close to us and those that can be seen with the unaided eye, is an example of a **selection effect**. When a population of objects (stars in this example) includes a great variety of different types, we must be careful what conclusions we draw from an examination of any particular subgroup. Certainly we would be fooling ourselves if we assumed that the stars visible to the unaided eye are characteristic of the general stellar population; this subgroup is heavily weighted to the most luminous stars. It requires much more effort to assemble a complete data set for the nearest stars, since most are so faint that they can be observed only with a telescope. However, it is only by doing so that astronomers are able to know about the properties of the vast majority of the stars, which are actually much smaller and fainter than our own Sun. In the next section, we will look at how we measure some of these properties.

22.5 | Measuring Stellar Masses

Learning Objectives

By the end of this section, you will be able to:

- Distinguish the different types of binary star systems
- Understand how we can apply Newton's version of Kepler's third law to derive the sum of star masses in a binary star system
- Apply the relationship between stellar mass and stellar luminosity to determine the physical characteristics of a star

The mass of a star—how much material it contains—is one of its most important characteristics. If we know a star's mass, as we shall see, we can estimate how long it will shine and what its ultimate fate will be. Yet the mass of a star is very difficult to measure directly. Somehow, we need to put a star on the cosmic equivalent of a scale.

Luckily, not all stars live like the Sun, in isolation from other stars. About half the stars are **binary stars**—two stars that orbit each other, bound together by gravity. Masses of binary stars can be calculated from measurements of their orbits, just as the mass of the Sun can be derived by measuring the orbits of the planets around it.

Binary Stars

Before we discuss in more detail how mass can be measured, we will take a closer look at stars that come in pairs. The first binary star was discovered in 1650, less than half a century after Galileo began to observe the sky with a telescope. John Baptiste Riccioli (1598–1671), an Italian astronomer, noted that the star Mizar, in the middle of the Big Dipper's handle, appeared through his telescope as two stars. Since that discovery, thousands of binary stars have been cataloged. (Astronomers call any pair of stars that appear to be close to each other in the sky *double stars*, but not all of these form a true binary, that is, not all of them are physically associated. Some are just chance alignments of stars that are actually at different distances from us.) Although stars most commonly come in pairs, there are also triple and quadruple systems.

One well-known binary star is Castor, located in the constellation of Gemini. By 1804, astronomer William Herschel, who also discovered the planet Uranus, had noted that the fainter component of Castor had slightly changed its position relative to the brighter component. (We use the term "component" to mean a member of a star system.) Here was evidence that one star was moving around another. It was actually the first evidence that gravitational influences exist outside the solar system. The orbital motion of a binary star is shown in Figure 22.18. A binary star system in which both of the stars can be seen with a telescope is called a **visual binary**.

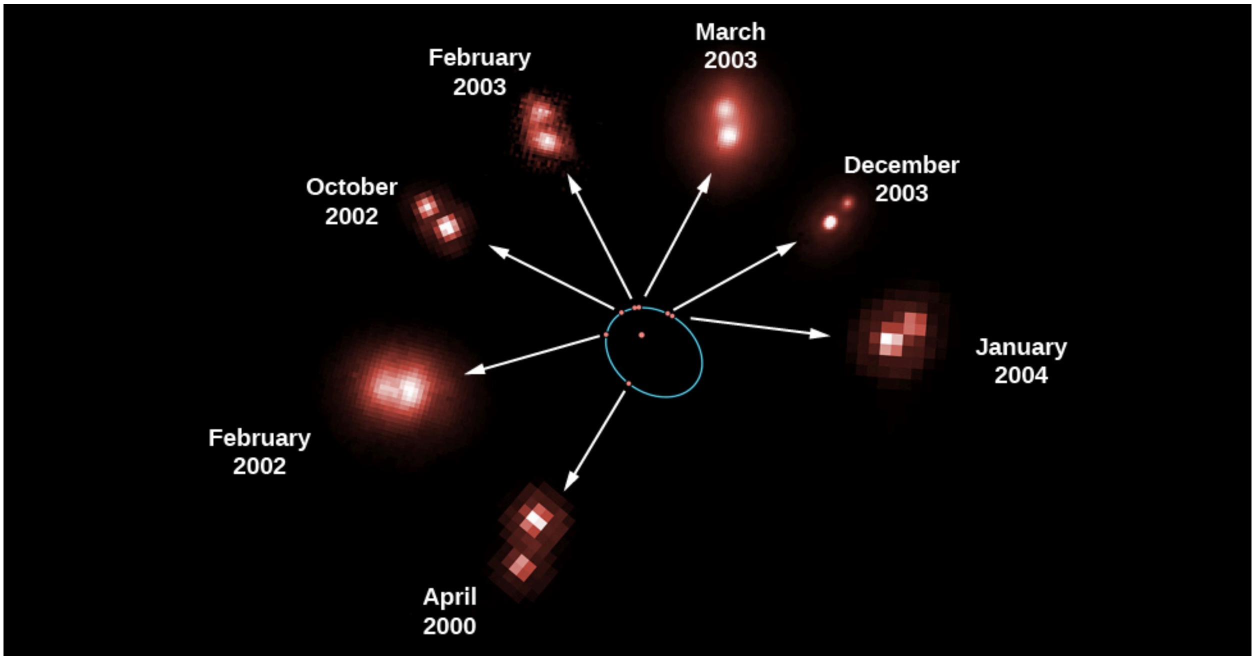


Figure 22.18 This figure shows seven observations of the mutual revolution of two stars, one a brown dwarf and one an ultra-cool L dwarf. Each red dot on the orbit, which is shown by the blue ellipse, corresponds to the position of one of the dwarfs relative to the other. The reason that the pair of stars looks different on the different dates is that some images were taken with the Hubble Space Telescope and others were taken from the ground. The arrows point to the actual observations that correspond to the positions of each red dot. From these observations, an international team of astronomers directly measured the mass of an ultra-cool brown dwarf star for the first time. Barely the size of the planet Jupiter, the dwarf star weighs in at just 8.5% of the mass of our Sun. (credit: modification of work by ESA/NASA and Herve Bouy (Max-Planck-Institut für Extraterrestrische Physik/ESO, Germany))

Edward C. Pickering (1846–1919), at Harvard, discovered a second class of binary stars in 1889—a class in which only one of the stars is actually seen directly. He was examining the spectrum of Mizar and found that the dark absorption lines in the brighter star's spectrum were usually double. Not only were there two lines where astronomers normally saw only one, but the spacing of the lines was constantly changing. At times, the lines even became single. Pickering correctly deduced that the brighter component of Mizar, called Mizar A, is itself really two stars that revolve about each other in a period of 104 days. A star like Mizar A, which appears as a single star when photographed or observed visually through the telescope, but which spectroscopy shows really to be a double star, is called a **spectroscopic binary**.

Mizar, by the way, is a good example of just how complex such star systems can be. Mizar has been known for centuries to have a faint companion called Alcor, which can be seen without a telescope. Mizar and Alcor form an *optical double*—a pair of stars that appear close together in the sky but do not orbit each other. Through a telescope, as Riccioli discovered in 1650, Mizar can be seen to have another, closer companion that does orbit it; Mizar is thus a visual binary. The two components that make up this visual binary, known as Mizar A and Mizar B, are both spectroscopic binaries. So, Mizar is really a quadruple system of stars.

Strictly speaking, it is not correct to describe the motion of a binary star system by saying that one star orbits the other. Gravity is a *mutual* attraction. Each star exerts a gravitational force on the other, with the result that both stars orbit a point between them called the *center of mass*. Imagine that the two stars are seated at either end of a seesaw. The point at which the fulcrum would have to be located in order for the seesaw to balance is the center of mass, and it is always closer to the more massive star (**Figure 22.19**).

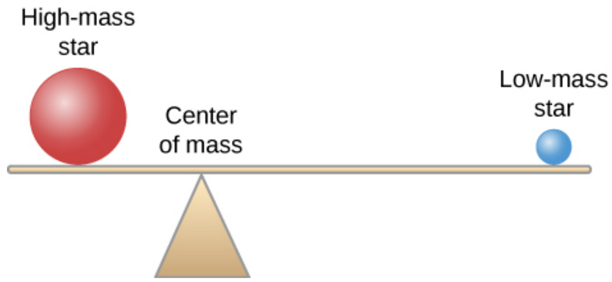


Figure 22.19 In a binary star system, both stars orbit their center of mass. The image shows the relative positions of two, different-mass stars from their center of mass, similar to how two masses would have to be located on a seesaw in order to keep it level. The star with the higher mass will be found closer to the center of mass, while the star with the lower mass will be farther from it.

Figure 22.20 shows two stars (A and B) moving around their center of mass, along with one line in the spectrum of each star that we observe from the system at different times. When one star is approaching us relative to the center of mass, the other star is receding from us. In the top left illustration, star A is moving toward us, so the line in its spectrum is Doppler-shifted toward the blue end of the spectrum. Star B is moving away from us, so its line shows a redshift. When we observe the composite spectrum of the two stars, the line appears double. When the two stars are both moving across our line of sight (neither away from nor toward us), they both have the same radial velocity (that of the pair's center of mass); hence, the spectral lines of the two stars come together. This is shown in the two bottom illustrations in **Figure 22.20**.

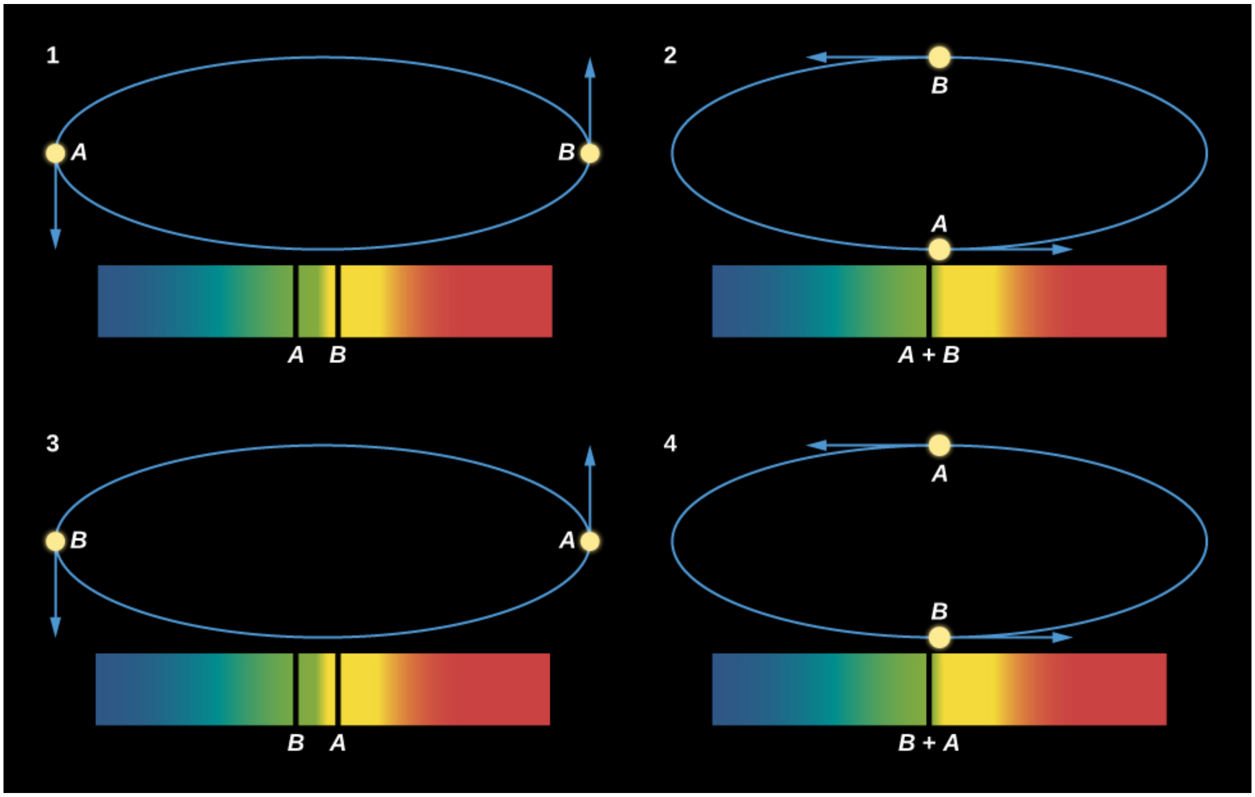


Figure 22.20 We see changes in velocity because when one star is moving toward Earth, the other is moving away; half a cycle later, the situation is reversed. Doppler shifts cause the spectral lines to move back and forth. In diagrams 1 and 3, lines from both stars can be seen well separated from each other. When the two stars are moving perpendicular to our line of sight (that is, they are not moving either toward or away from us), the two lines are exactly superimposed, and so in diagrams 2 and 4, we see only a single spectral line. Note that in the diagrams, the orbit of the star pair is tipped slightly with respect to the viewer (or if the viewer were looking at it in the sky, the orbit would be tilted with respect to the viewer's line of sight). If the orbit were exactly in the plane of the page or screen (or the sky), then it would look nearly circular, but we would see no change in radial velocity (no part of the motion would be toward us or away from us.) If the orbit were perpendicular to the plane of the page or screen, then the stars would appear to move back and forth in a straight line, and we would see the largest-possible radial velocity variations.

A plot showing how the velocities of the stars change with time is called a *radial velocity curve*; the curve for the binary system in **Figure 22.20** is shown in **Figure 22.21**.

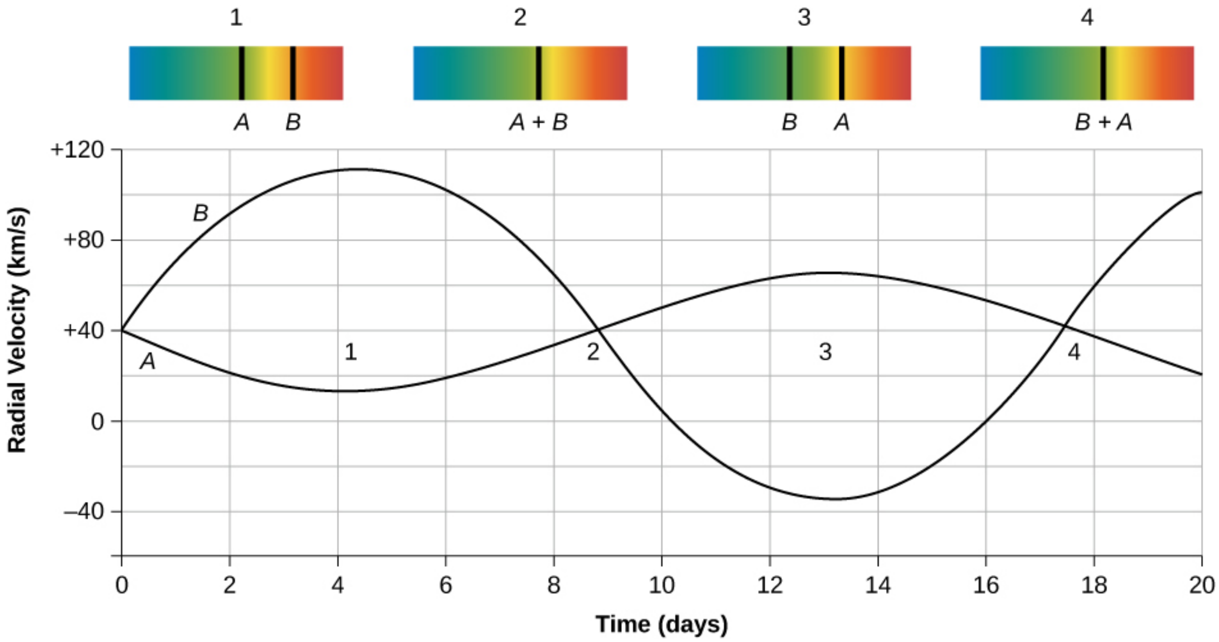


Figure 22.21 These curves plot the radial velocities of two stars in a spectroscopic binary system, showing how the stars alternately approach and recede from Earth. Note that positive velocity means the star is moving away from us relative to the center of mass of the system, which in this case is 40 kilometers per second. Negative velocity means the star is moving toward us relative to the center of mass. The positions on the curve corresponding to the illustrations in **Figure 22.20** are marked with the diagram number (1–4).

This **animation (https://openstax.org/l/30binstaranim)** lets you follow the orbits of a binary star system in various combinations of the masses of the two stars.

Masses from the Orbits of Binary Stars

We can estimate the masses of binary star systems using Newton's reformulation of Kepler's third law (discussed in **Kepler's Laws of Planetary Motion**). Kepler found that the time a planet takes to go around the Sun is related by a specific mathematical formula to its distance from the Sun. In our binary star situation, if two objects are in mutual revolution, then the period (T) with which they go around each other is related to the semimajor axis (a) of the orbit of one with respect to the other, according to this equation

$$a^3 = (M_1 + M_2)T^2 \quad (22.1)$$

where a is in astronomical units, T is measured in years, and $M_1 + M_2$ is the sum of the masses of the two stars in units of the Sun's mass. This is a very useful formula for astronomers; it says that if we can observe the size of the orbit and the period of mutual revolution of the stars in a binary system, we can calculate the sum of their masses.

Most spectroscopic binaries have periods ranging from a few days to a few months, with separations of usually less than 1 AU between their member stars. Recall that an AU is the distance from Earth to the Sun, so this is a small separation and very hard to see at the distances of stars. This is why many of these systems are known to be double only through careful study of their spectra.

We can analyze a radial velocity curve (such as the one in **Figure 22.21**) to determine the masses of the stars in a spectroscopic binary. This is complex in practice but not hard in principle. We measure the speeds of the stars from the **Doppler effect**. We then determine the period—how long the stars take to go through an orbital cycle—from the velocity curve. Knowing how fast the stars are moving and how long they take to go around tells us the circumference of the orbit and, hence, the separation of the stars in kilometers or astronomical units. From Kepler's law, the period and the separation allow us to calculate the sum of the stars' masses.

Of course, knowing the sum of the masses is not as useful as knowing the mass of each star separately. But the relative orbital speeds of the two stars can tell us how much of the total mass each star has. As we saw in our seesaw analogy, the more massive star is closer to the center of mass and therefore has a smaller orbit. Therefore, it moves more slowly to get around in the same time compared to the more distant, lower-mass star. If we sort out the speeds relative to each other, we can sort out the masses relative to each other. In practice, we also need to know how the binary system is oriented in the sky to our line of sight, but if we do, and the just-described steps are carried out carefully, the result is a calculation of the masses of each of the two stars in the system.

To summarize, a good measurement of the motion of two stars around a common center of mass, combined with the laws of gravity, allows us to determine the masses of stars in such systems. These mass measurements are absolutely crucial to developing a theory of how stars evolve. One of the best things about this method is that it is independent of the location of the binary system. It works as well for stars 100 light-years away from us as for those in our immediate neighborhood.

To take a specific example, Sirius is one of the few binary stars in **Appendix D** for which we have enough information to apply Kepler's third law:

$$a^3 = (M_1 + M_2)T^2$$

In this case, the two stars, the one we usually call Sirius and its very faint companion, are separated by about 20 AU and have an orbital period of about 50 years. If we place these values in the formula we would have

$$(20)^3 = (M_1 + M_2)(50)^2$$
$$8000 = (M_1 + M_2)(2500)$$

This can be solved for the sum of the masses:

$$M_1 + M_2 = \frac{8000}{2500} = 3.2$$

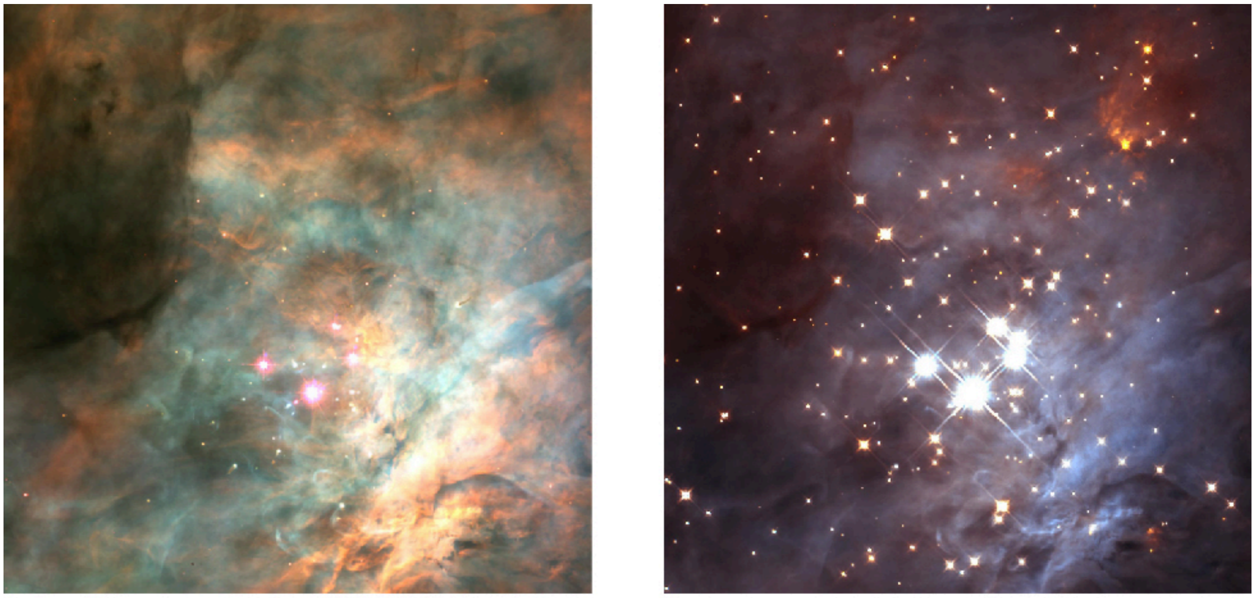
Therefore, the sum of masses of the two stars in the Sirius binary system is 3.2 times the Sun's mass. In order to determine the individual mass of each star, we would need the velocities of the two stars and the orientation of the orbit relative to our line of sight. If we kew those, we could apply the principle that their momenta must be equal in magnitude but opposite in direction (just as we did for a star-planet system in **Planets Beyond the Solar System**).

The Range of Stellar Masses

How large can the mass of a star be? Stars more massive than the Sun are rare. None of the stars within 30 light-years of the Sun has a mass greater than four times that of the Sun. Searches at large distances from the Sun have led to the discovery of a few stars with masses up to about 100 times that of the Sun, and a handful of stars (a few out of several billion) may have masses as large as 250 solar masses. However, most stars have less mass than the Sun.

According to theoretical calculations, the smallest mass that a true star can have is about 1/12 that of the Sun. By a "true" star, astronomers mean one that becomes hot enough to fuse protons to form helium (as discussed in **Source of Sunshine: Nuclear Fusion!**). Objects with masses between roughly 1/100 and 1/12 that of the Sun may produce energy for a brief time by means of nuclear reactions involving deuterium, but they do not become hot enough to fuse protons. Such objects are intermediate in mass between stars and planets and have been given the name **brown dwarfs** (**Figure 22.22**). Brown dwarfs are similar to Jupiter in radius but have masses from approximately 13 to 80 times larger than the mass of Jupiter.[3]

3. Exactly where to put the dividing line between planets and brown dwarfs is a subject of some debate among astronomers as we write this book (as is, in fact, the exact definition of each of these objects). Even those who accept deuterium fusion as the crucial issue for brown dwarfs concede that, depending on the composition of the star and other factors, the lowest mass for such a dwarf could be anywhere from 11 to 16 Jupiter masses.



(a) (b)

Figure 22.22 These images, taken with the Hubble Space Telescope, show the region surrounding the Trapezium star cluster inside the star-forming region called the Orion Nebula. (a) No brown dwarfs are seen in the visible light image, both because they put out very little light in the visible and because they are hidden within the clouds of dust in this region. (b) This image was taken in infrared light, which can make its way to us through the dust. The faintest objects in this image are brown dwarfs with masses between 13 and 80 times the mass of Jupiter. (credit a: NASA, C.R. O'Dell and S.K. Wong (Rice University); credit b: NASA; K.L. Luhman (Harvard-Smithsonian Center for Astrophysics) and G. Schneider, E. Young, G. Rieke, A. Cotera, H. Chen, M. Rieke, R. Thompson (Steward Observatory))

Still-smaller objects with masses less than about 1/100 the mass of the Sun (or 10 Jupiter masses) are called planets. They may radiate energy produced by the radioactive elements that they contain, and they may also radiate heat generated by slowly compressing under their own weight (a process called gravitational contraction). However, their interiors will never reach temperatures high enough for any nuclear reactions, to take place. Jupiter, whose mass is about 1/1000 the mass of the Sun, is unquestionably a planet, for example. Until the 1990s, we could only detect planets in our own solar system, but now we have thousands of them elsewhere as well. (We discussed these exciting observations in **Planets Beyond the Solar System**.)

The Mass-Luminosity Relation

Now that we have measurements of the characteristics of many different types of stars, we can search for relationships among the characteristics. For example, we can ask whether the mass and luminosity of a star are related. It turns out that for most stars, they are: The more massive stars are generally also the more luminous. This relationship, known as the **mass-luminosity relation**, is shown graphically in **Figure 22.23**. Each point represents a star whose mass and luminosity are both known. The horizontal position on the graph shows the star's mass, given in units of the Sun's mass, and the vertical position shows its luminosity in units of the Sun's luminosity.

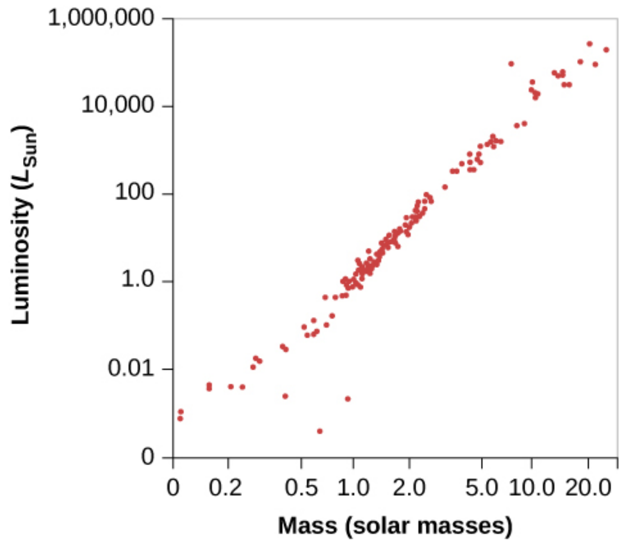


Figure 22.23 The plotted points show the masses and luminosities of stars. The three points lying below the sequence of points are all white dwarf stars.

We can also say this in mathematical terms.

$$L \propto M^{3.9} \quad (22.2)$$

It's a reasonably good approximation to say that luminosity (expressed in units of the Sun's luminosity) varies as the fourth power of the mass (in units of the Sun's mass). (The symbol ∝ means the two quantities are proportional.) If two stars differ in mass by a factor of 2, then the more massive one will be 2^4, or about 16 times brighter; if one star is 1/3 the mass of another, it will be approximately 81 times less luminous.

Example 22.1

Calculating the Mass from the Luminosity of a Star

The mass-luminosity formula can be rewritten so that a value of mass can be determined if the luminosity is known.

Solution

First, we must get our units right by expressing both the mass and the luminosity of a star in units of the Sun's mass and luminosity:

$$L/L_{\text{Sun}} = (M/M_{\text{Sun}})^4$$

Now we can take the 4th root of both sides, which is equivalent to taking both sides to the 1/4 = 0.25 power. The formula in this case would be:

$$M/M_{\text{Sun}} = (L/L_{\text{Sun}})^{0.25} = (L/L_{\text{Sun}})^{0.25}$$

22.1 In the previous section, we determined the sum of the masses of the two stars in the Sirius binary system (Sirius and its faint companion) using Kepler's third law to be 3.2 solar masses. Using the mass-luminosity relationship, calculate the mass of each individual star.

Notice how good this mass-luminosity relationship is. Most stars (see **Figure 22.23**) fall along a line running from the lower-left (low mass, low luminosity) corner of the diagram to the upper-right (high mass, high luminosity) corner. About

90% of all stars obey the mass-luminosity relation. Later, we will explore why such a relationship exists and what we can learn from the roughly 10% of stars that “disobey” it.

22.6 | Diameters of Stars

Learning Objectives

By the end of this section, you will be able to:

- Describe the methods used to determine star diameters
- Identify the parts of an eclipsing binary star light curve that correspond to the diameters of the individual components

It is easy to measure the diameter of the Sun. Its angular diameter—that is, its apparent size on the sky—is about 1/2°. If we know the angle the Sun takes up in the sky and how far away it is, we can calculate its true (linear) diameter, which is 1.39 million kilometers, or about 109 times the diameter of Earth.

Unfortunately, the Sun is the only star whose angular diameter is easily measured. All the other stars are so far away that they look like pinpoints of light through even the largest ground-based telescopes. (They often seem to be bigger, but that is merely distortion introduced by turbulence in Earth’s atmosphere.) Luckily, there are several techniques that astronomers can use to estimate the sizes of stars.

Stars Blocked by the Moon

One technique, which gives very precise diameters but can be used for only a few stars, is to observe the dimming of light that occurs when the Moon passes in front of a star. What astronomers measure (with great precision) is the time required for the star’s brightness to drop to zero as the edge of the Moon moves across the star’s disk. Since we know how rapidly the Moon moves in its orbit around Earth, it is possible to calculate the angular diameter of the star. If the distance to the star is also known, we can calculate its diameter in kilometers. This method works only for fairly bright stars that happen to lie along the zodiac, where the Moon (or, much more rarely, a planet) can pass in front of them as seen from Earth.

Eclipsing Binary Stars

Now, we have already examined a technique, currently in use in the NASA Kepler mission to discover exoplanets (see **Exoplanets Everywhere**) where the passage of a planet in front of a star allows us to analyze the resulting light curve to determine the diameter of the planet.

With slight modification, this technique can be used to determine the diameter of stars that form what is know as an **eclipsing binary system**.

Accurate sizes for a large number of stars come from measurements of eclipsing binary star systems, and so we must make a brief detour from our main story to examine this type of star system. Some binary stars are lined up in such a way that, when viewed from Earth, each star passes in front of the other during every revolution (**Figure 22.24**). When one star blocks the light of the other, preventing it from reaching Earth, the luminosity of the system decreases, and astronomers say that an eclipse has occurred.

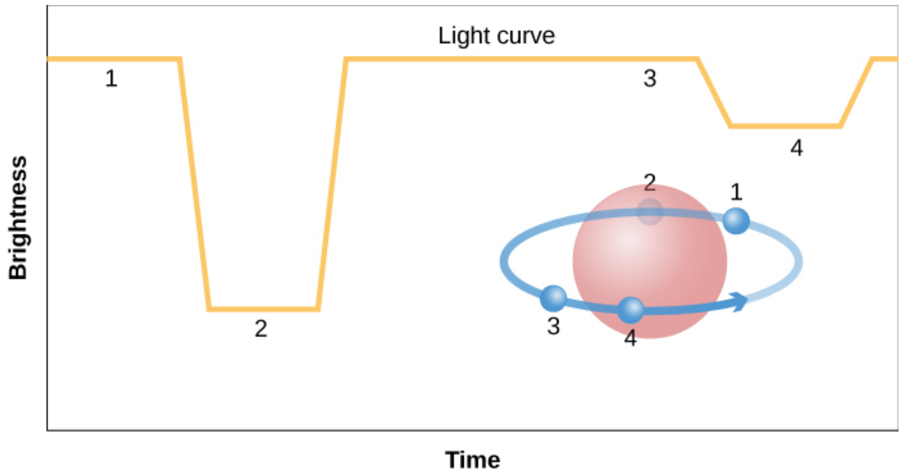


Figure 22.24 The light curve of an eclipsing binary star system shows how the combined light from both stars changes due to eclipses over the time span of an orbit. This light curve shows the behavior of a hypothetical eclipsing binary star with total eclipses (one star passes directly in front of and behind the other). The numbers indicate parts of the light curve corresponding to various positions of the smaller star in its orbit. In this diagram, we have assumed that the smaller star is also the hotter one so that it emits more flux (energy per second per square meter) than the larger one. When the smaller, hotter star goes behind the larger one, its light is completely blocked, and so there is a strong dip in the light curve. When the smaller star goes in front of the bigger one, a small amount of light from the bigger star is blocked, so there is a smaller dip in the light curve.

The discovery of the first eclipsing binary helped solve a long-standing puzzle in astronomy. The star Algol, in the constellation of Perseus, changes its brightness in an odd but regular way. Normally, Algol is a fairly bright star, but at intervals of 2 days, 20 hours, 49 minutes, it fades to one-third of its regular brightness. After a few hours, it brightens to normal again. This effect is easily seen, even without a telescope, if you know what to look for.

In 1783, a young English astronomer named John Goodricke (1764–1786) made a careful study of Algol (see the feature on **John Goodricke** for a discussion of his life and work). Even though Goodricke could neither hear nor speak, he made a number of major discoveries in the 21 years of his brief life. He suggested that Algol's unusual brightness variations might be due to an invisible companion that regularly passes in front of the brighter star and blocks its light. Unfortunately, Goodricke had no way to test this idea, since it was not until about a century later that equipment became good enough to measure Algol's spectrum.

In 1889, the German astronomer Hermann Vogel (1841–1907) demonstrated that, like Mizar, Algol is a spectroscopic binary. The spectral lines of Algol were not observed to be double because the fainter star of the pair gives off too-little light compared with the brighter star for its lines to be conspicuous in the composite spectrum. Nevertheless, the periodic shifting back and forth of the brighter star's lines gave evidence that it was revolving about an unseen companion. (The lines of both components need not be visible for a star to be recognized as a spectroscopic binary.)

The discovery that Algol is a spectroscopic binary verified Goodricke's hypothesis. The plane in which the stars revolve is turned nearly edgewise to our line of sight, and each star passes in front of the other during every revolution. (The eclipse of the fainter star in the Algol system is not very noticeable because the part of it that is covered contributes little to the total light of the system. This second eclipse can, however, be detected by careful measurements.)

Any binary star produces eclipses if viewed from the proper direction, near the plane of its orbit, so that one star passes in front of the other (see **Figure 22.24**). But from our vantage point on Earth, only a few binary star systems are oriented in this way.

Astronomy and Mythology: Algol the Demon Star and Perseus the Hero

The name Algol comes from the Arabic *Ras al Ghul*, meaning "the demon's head."[4] The word "ghoul" in English has the same derivation. Many of the bright stars have Arabic names because during the long dark ages in medieval Europe, it was Arabic astronomers who preserved and expanded the Greek and Roman knowledge of the skies. The reference

4. Fans of Batman comic books and movies will recognize that this name was given to an archvillain in the series.

to the demon is part of the ancient Greek legend of the hero Perseus, who is commemorated by the constellation in which we find Algol and whose adventures involve many of the characters associated with the northern constellations.

Perseus was one of the many half-god heroes fathered by Zeus (Jupiter in the Roman version), the king of the gods in Greek mythology. Zeus had, to put it delicately, a roving eye and was always fathering somebody or other with a human maiden who caught his fancy. (Perseus derives from *Per Zeus*, meaning “fathered by Zeus.”) Set adrift with his mother by an (understandably) upset stepfather, Perseus grew up on an island in the Aegean Sea. The king there, taking an interest in Perseus’ mother, tried to get rid of the young man by assigning him an extremely difficult task.

In a moment of overarching pride, a beautiful young woman named Medusa had compared her golden hair to that of the goddess Athena (Minerva for the Romans). The Greek gods did not take kindly to being compared to mere mortals, and Athena turned Medusa into a gorgon: a hideous, evil creature with writhing snakes for hair and a face that turned anyone who looked at it into stone. Perseus was given the task of slaying this demon, which seemed like a pretty sure way to get him out of the way forever.

But because Perseus had a god for a father, some of the other gods gave him tools for the job, including Athena’s reflective shield and the winged sandals of Hermes (Mercury in the Roman story). By flying over her and looking only at her reflection, Perseus was able to cut off Medusa’s head without ever looking at her directly. Taking her head (which, conveniently, could still turn onlookers to stone even without being attached to her body) with him, Perseus continued on to other adventures.

He next came to a rocky seashore, where boasting had gotten another family into serious trouble with the gods. Queen Cassiopeia had dared to compare her own beauty to that of the Nereids, sea nymphs who were daughters of Poseidon (Neptune in Roman mythology), the god of the sea. Poseidon was so offended that he created a sea-monster named Cetus to devastate the kingdom. King Cepheus, Cassiopeia’s beleaguered husband, consulted the oracle, who told him that he must sacrifice his beautiful daughter Andromeda to the monster.

When Perseus came along and found Andromeda chained to a rock near the sea, awaiting her fate, he rescued her by turning the monster to stone. (Scholars of mythology actually trace the essence of this story back to far-older legends from ancient Mesopotamia, in which the god-hero Marduk vanquishes a monster named Tiamat. Symbolically, a hero like Perseus or Marduk is usually associated with the Sun, the monster with the power of night, and the beautiful maiden with the fragile beauty of dawn, which the Sun releases after its nightly struggle with darkness.)

Many of the characters in these Greek legends can be found as constellations in the sky, not necessarily resembling their namesakes but serving as reminders of the story. For example, vain Cassiopeia is sentenced to be very close to the celestial pole, rotating perpetually around the sky and hanging upside down every winter. The ancients imagined Andromeda still chained to her rock (it is much easier to see the chain of stars than to recognize the beautiful maiden in this star grouping). Perseus is next to her with the head of Medusa swinging from his belt. Algol represents this gorgon head and has long been associated with evil and bad fortune in such tales. Some commentators have speculated that the star’s change in brightness (which can be observed with the unaided eye) may have contributed to its unpleasant reputation, with the ancients regarding such a change as a sort of evil “wink.”

Diameters of Eclipsing Binary Stars

We now turn back to the main thread of our story to discuss how all this can be used to measure the sizes of stars. The technique involves making a light curve of an eclipsing binary, a graph that plots how the brightness changes with time. Let us consider a hypothetical binary system in which the stars are very different in size, like those illustrated in **Figure 22.25**. To make life easy, we will assume that the orbit is viewed exactly edge-on.

Even though we cannot see the two stars separately in such a system, the light curve can tell us what is happening. When the smaller star just starts to pass behind the larger star (a point we call *first contact*), the brightness begins to drop. The eclipse becomes total (the smaller star is completely hidden) at the point called *second contact*. At the end of the total eclipse (*third contact*), the smaller star begins to emerge. When the smaller star has reached *last contact*, the eclipse is completely over.

To see how this allows us to measure diameters, look carefully at **Figure 22.25**. During the time interval between the first and second contacts, the smaller star has moved a distance equal to its own diameter. During the time interval from the first to third contacts, the smaller star has moved a distance equal to the diameter of the larger star. If the spectral lines of both stars are visible in the spectrum of the binary, then the speed of the smaller star with respect to the larger one can be measured from the Doppler shift. But knowing the speed with which the smaller star is moving and how long it took to cover some distance can tell the span of that distance—in this case, the diameters of the stars. The speed multiplied by the time interval from the first to second contact gives the diameter of the smaller star. We multiply the speed by the time between the first and third contacts to get the diameter of the larger star.

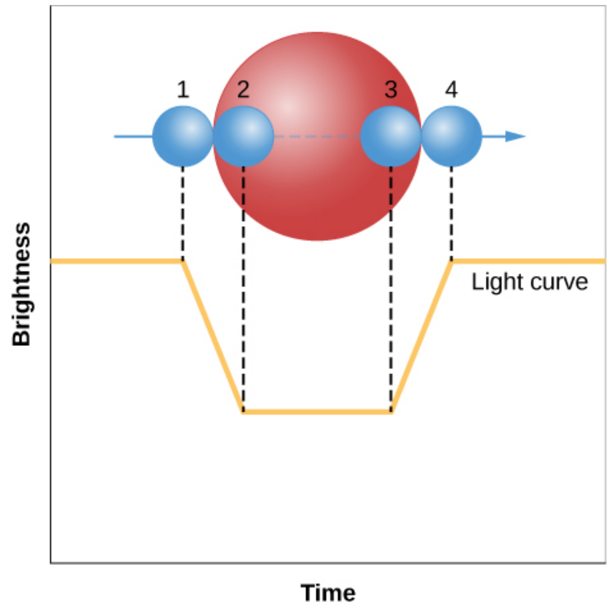


Figure 22.25 Here we see the light curve of a hypothetical eclipsing binary star whose orbit we view exactly edge-on, in which the two stars fully eclipse each other. From the time intervals between contacts, it is possible to estimate the diameters of the two stars.

In actuality, the situation with eclipsing binaries is often a bit more complicated: orbits are generally not seen exactly edge-on, and the light from each star may be only partially blocked by the other. Furthermore, binary star orbits, just like the orbits of the planets, are ellipses, not circles. However, all these effects can be sorted out from very careful measurements of the light curve.

Using the Radiation Law to Get the Diameter

Another method for measuring star diameters makes use of the Stefan-Boltzmann law for the relationship between energy radiated and temperature (see **Blackbody Radiation**). In this method, the *energy flux* (energy emitted per second per square meter by a blackbody, like the Sun) is given by

$$F = \sigma T^4$$

where σ is a constant and T is the temperature. The surface area of a sphere (like a star) is given by

$$A = 4\pi R^2$$

The luminosity (L) of a star is then given by its surface area in square meters times the energy flux:

$$L = (A \times F)$$

Previously, we determined the masses of the two stars in the Sirius binary system. Sirius gives off 8200 times more energy than its fainter companion star, although both stars have nearly identical temperatures. The extremely large difference in luminosity is due to the difference in radius, since the temperatures and hence the energy fluxes for the two stars are nearly the same. To determine the relative sizes of the two stars, we take the ratio of the corresponding luminosities:

$$\frac{L_{\text{Sirius}}}{L_{\text{companion}}} = \frac{\left(A_{\text{Sirius}} \times F_{\text{Sirius}}\right)}{\left(A_{\text{companion}} \times F_{\text{companion}}\right)}$$
$$= \frac{A_{\text{Sirius}}}{A_{\text{companion}}} = \frac{4\pi R^2{}_{\text{Sirius}}}{4\pi R^2{}_{\text{companion}}} = \frac{R^2{}_{\text{Sirius}}}{R^2{}_{\text{companion}}}$$

$$\frac{L_{\text{Sirius}}}{L_{\text{companion}}} = 8200 = \frac{R^2{}_{\text{Sirius}}}{R^2{}_{\text{companion}}}$$

Therefore, the relative sizes of the two stars can be found by taking the square root of the relative luminosity. Since $\sqrt{8200} = 91$, the radius of Sirius is 91 times larger than the radium of its faint companion.

The method for determining the radius shown here requires both stars be visible, which is not always the case.

Stellar Diameters

The results of many stellar size measurements over the years have shown that most nearby stars are roughly the size of the Sun, with typical diameters of a million kilometers or so. Faint stars, as we might have expected, are generally smaller than more luminous stars. However, there are some dramatic exceptions to this simple generalization.

A few of the very luminous stars, those that are also red (indicating relatively low surface temperatures), turn out to be truly enormous. These stars are called, appropriately enough, giant stars or supergiant stars. An example is Betelgeuse, the second brightest star in the constellation of Orion and one of the dozen brightest stars in our sky. Its diameter, remarkably, is greater than 10 AU (1.5 *billion* kilometers!), large enough to fill the entire inner solar system almost as far out as Jupiter. In **Stellar Life Cycles**, we will look in detail at the evolutionary process that leads to the formation of such giant and supergiant stars.

Watch this **star size comparison video (https://openstax.org/l/30starsizecomp)** for a striking visual that highlights the size of stars versus planets and the range of sizes among stars.

22.7 | The H-R Diagram

Learning Objectives

By the end of this section, you will be able to:

- Identify the physical characteristics of stars that are used to create an H-R diagram, and describe how those characteristics vary among groups of stars
- Discuss the physical properties of most stars found at different locations on the H–R diagram, such as radius, and for main sequence stars, mass

In this chapter, we have described some of the characteristics by which we might classify stars and how those are measured. These ideas are summarized in **Table 22.4**. We have also given an example of a relationship between two of these characteristics in the mass-luminosity relation. When the characteristics of large numbers of stars were measured at the beginning of the twentieth century, astronomers were able to begin a deeper search for patterns and relationships in these data.

Measuring the Characteristics of Stars

Characteristic	Technique
Surface temperature	1. Determine the color (very rough). 2. Measure the spectrum and get the spectral type.

Table 22.4

Measuring the Characteristics of Stars

Characteristic	Technique
Chemical composition	Determine which lines are present in the spectrum.
Luminosity	Measure the apparent brightness and compensate for distance.
Radial velocity	Measure the Doppler shift in the spectrum.
Rotation	Measure the width of spectral lines.
Mass	Measure the period and radial velocity curves of spectroscopic binary stars.
Diameter	1. Measure the way a star's light is blocked by the Moon. 2. Measure the light curves and Doppler shifts for eclipsing binary stars.

Table 22.4

To help understand what sorts of relationships might be found, let's look briefly at a range of data about human beings. If you want to understand humans by comparing and contrasting their characteristics—without assuming any previous knowledge of these strange creatures—you could try to determine which characteristics lead you in a fruitful direction. For example, you might plot the heights of a large sample of humans against their weights (which is a measure of their mass). Such a plot is shown in Figure 22.26 and it has some interesting features. In the way we have chosen to present our data, height increases upward, whereas weight increases to the left. Notice that humans are not randomly distributed in the graph. Most points fall along a sequence that goes from the upper left to the lower right.

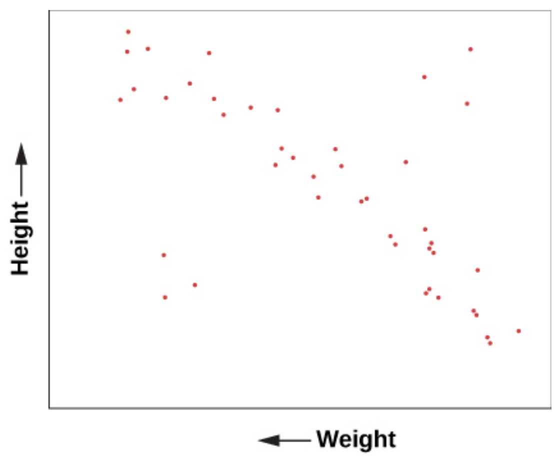


Figure 22.26 The plot of the heights and weights of a representative group of human beings. Most points lie along a "main sequence" representing most people, but there are a few exceptions.

We can conclude from this graph that human height and weight are related. Generally speaking, taller human beings weigh more, whereas shorter ones weigh less. This makes sense if you are familiar with the structure of human beings. Typically, if we have bigger bones, we have more flesh to fill out our larger frame. It's not mathematically exact—there is a wide range of variation—but it's not a bad overall rule. And, of course, there are some dramatic exceptions. You occasionally see a short human who is very overweight and would thus be more to the bottom left of our diagram than the average sequence of people. Or you might have a very tall, skinny fashion model with great height but relatively small weight, who would be found near the upper right.

A similar diagram has been found extremely useful for understanding the lives of stars. In 1913, American astronomer Henry Norris Russell plotted the luminosities of stars against their spectral classes (a way of denoting their surface temperatures). This investigation, and a similar independent study in 1911 by Danish astronomer Ejnar Hertzsprung, led to the extremely important discovery that the temperature and luminosity of stars are related (Figure 22.27).

(a) (b)

Figure 22.27 (a) Ejnar Hertzsprung and (b) Henry Norris Russell independently discovered the relationship between the luminosity and surface temperature of stars that is summarized in what is now called the H–R diagram.

Henry Norris Russell

When Henry Norris Russell graduated from Princeton University, his work had been so brilliant that the faculty decided to create a new level of honors degree beyond "summa cum laude" for him. His students later remembered him as a man whose thinking was three times faster than just about anybody else's. His memory was so phenomenal, he could correctly quote an enormous number of poems and limericks, the entire Bible, tables of mathematical functions, and almost anything he had learned about astronomy. He was nervous, active, competitive, critical, and very articulate; he tended to dominate every meeting he attended. In outward appearance, he was an old-fashioned product of the nineteenth century who wore high-top black shoes and high starched collars, and carried an umbrella every day of his life. His 264 papers were enormously influential in many areas of astronomy.

Born in 1877, the son of a Presbyterian minister, Russell showed early promise. When he was 12, his family sent him to live with an aunt in Princeton so he could attend a top preparatory school. He lived in the same house in that town until his death in 1957 (interrupted only by a brief stay in Europe for graduate work). He was fond of recounting that both his mother and his maternal grandmother had won prizes in mathematics, and that he probably inherited his talents in that field from their side of the family.

Before Russell, American astronomers devoted themselves mainly to surveying the stars and making impressive catalogs of their properties, especially their spectra (as described in **The Spectra of Stars**. Russell began to see that interpreting the spectra of stars required a much more sophisticated understanding of the physics of the atom, a subject that was being developed by European physicists in the 1910s and 1920s. Russell embarked on a lifelong quest to ascertain the physical conditions inside stars from the clues in their spectra; his work inspired, and was continued by, a generation of astronomers, many trained by Russell and his collaborators.

Russell also made important contributions in the study of binary stars and the measurement of star masses, the origin of the solar system, the atmospheres of planets, and the measurement of distances in astronomy, among other fields. He was an influential teacher and popularizer of astronomy, writing a column on astronomical topics for *Scientific American* magazine for more than 40 years. He and two colleagues wrote a textbook for college astronomy classes that helped train astronomers and astronomy enthusiasts over several decades. That book set the scene for the kind of textbook you are now reading, which not only lays out the facts of astronomy but also explains how they fit together. Russell gave lectures around the country, often emphasizing the importance of understanding modern physics in order to grasp what was happening in astronomy.

Harlow Shapley, director of the Harvard College Observatory, called Russell "the dean of American astronomers." Russell was certainly regarded as the leader of the field for many years and was consulted on many astronomical problems by colleagues from around the world. Today, one of the highest recognitions that an astronomer can receive is an award from the American Astronomical Society called the Russell Prize, set up in his memory.

Features of the H–R Diagram

Following Hertzsprung and Russell, let us plot the temperature (or spectral class) of a selected group of nearby stars against their luminosity and see what we find (**Figure 22.28**). Such a plot is frequently called the *Hertzsprung–Russell diagram*, abbreviated **H–R diagram**. It is one of the most important and widely used diagrams in astronomy, with applications that extend far beyond the purposes for which it was originally developed more than a century ago.

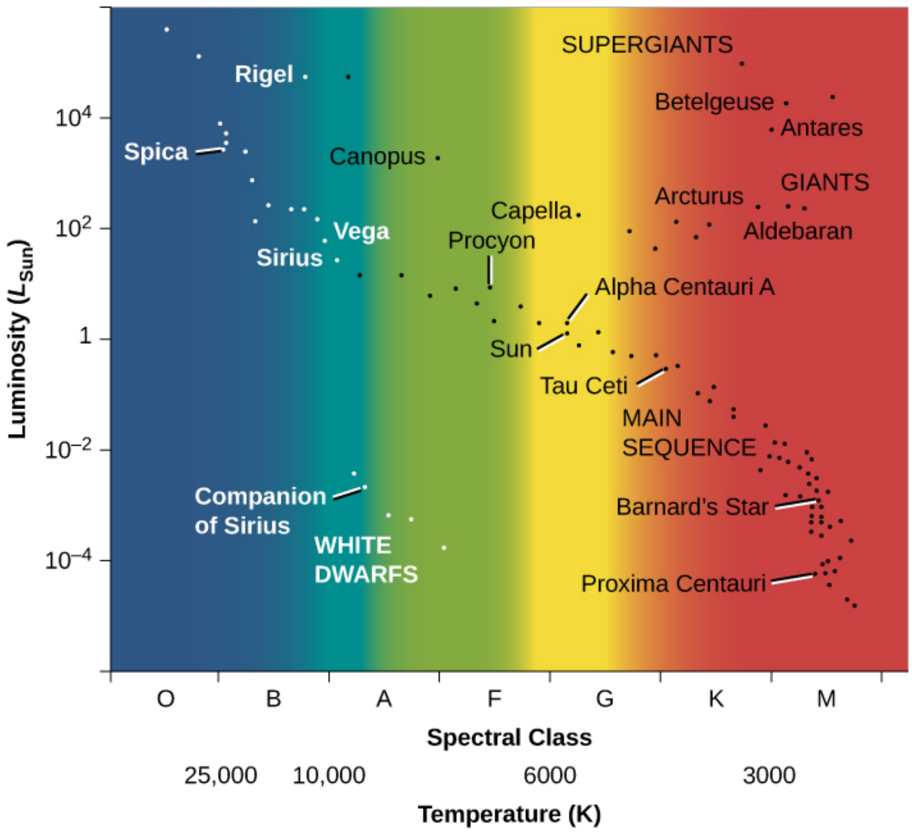


Figure 22.28 In such diagrams, luminosity is plotted along the vertical axis. Along the horizontal axis, we can plot either temperature or spectral type (also sometimes called spectral class). Several of the brightest stars are identified by name. Most stars fall on the main sequence.

It is customary to plot H–R diagrams in such a way that temperature increases toward the left and luminosity toward the top. Notice the similarity to our plot of height and weight for people (**Figure 22.26**). Stars, like people, are not distributed over the diagram at random, as they would be if they exhibited all combinations of luminosity and temperature. Instead, we see that the stars cluster into certain parts of the H–R diagram. The great majority are aligned along a narrow sequence running from the upper left (hot, highly luminous) to the lower right (cool, less luminous). This band of points is called the **main sequence**. It represents a relationship between *temperature* and *luminosity* that is followed by most stars. We can summarize this relationship by saying that hotter stars are more luminous than cooler ones.

A number of stars, however, lie above the main sequence on the H–R diagram, in the upper-right region, where stars have low temperature and high luminosity. How can a star be at once cool, meaning each square meter on the star does not put out all that much energy, and yet very luminous? The only way is for the star to be enormous—to have so many square meters on its surface that the *total* energy output is still large. These stars must be *giants* or *supergiants*, the stars of huge diameter we discussed earlier.

There are also some stars in the lower-left corner of the diagram, which have high temperature and low luminosity. If they have high surface temperatures, each square meter on that star puts out a lot of energy. How then can the overall star be dim? It must be that it has a very small total surface area; such stars are known as **white dwarfs** (white because, at these high temperatures, the colors of the electromagnetic radiation that they emit blend together to make them look bluish-white). We will say more about these puzzling objects in a moment. **Figure 22.29** is a schematic H–R diagram for a large sample of stars, drawn to make the different types more apparent.

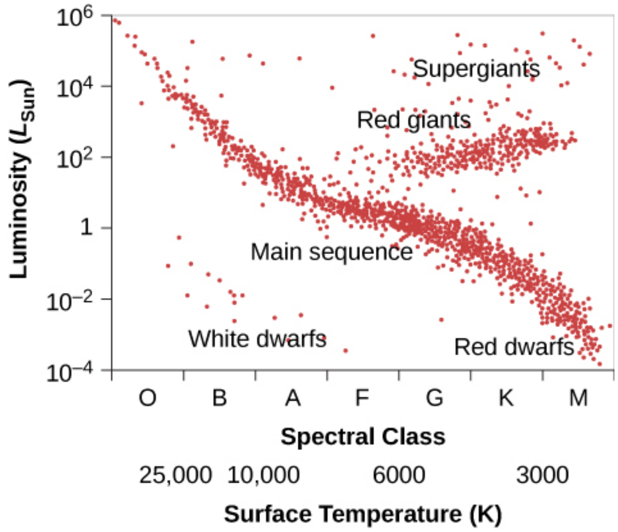


Figure 22.29 Ninety percent of all stars on such a diagram fall along a narrow band called the main sequence. A minority of stars are found in the upper right; they are both cool (and hence red) and bright, and must be giants. Some stars fall in the lower left of the diagram; they are both hot and dim, and must be white dwarfs.

Now, think back to our discussion of star surveys. It is difficult to plot an H–R diagram that is truly representative of all stars because most stars are so faint that we cannot see those outside our immediate neighborhood. The stars plotted in **Figure 22.28** were selected because their distances are known. This sample omits many intrinsically faint stars that are nearby but have not had their distances measured, so it shows fewer faint main-sequence stars than a "fair" diagram would. To be truly representative of the stellar population, an H–R diagram should be plotted for all stars within a certain distance. Unfortunately, our knowledge is reasonably complete only for stars within 10 to 20 light-years of the Sun, among which there are no giants or supergiants. Still, from many surveys (and more can now be done with new, more powerful telescopes), we estimate that about 90% of the true stars overall (excluding brown dwarfs) in our part of space are main-sequence stars, about 10% are white dwarfs, and fewer than 1% are giants or supergiants.

These estimates can be used directly to understand the lives of stars. Permit us another quick analogy with people. Suppose we survey people just like astronomers survey stars, but we want to focus our attention on the location of young people, ages 6 to 18 years. Survey teams fan out and take data about where such youngsters are found at all times during a 24-hour day. Some are found in the local pizza parlor, others are asleep at home, some are at the movies, and many are in school. After surveying a very large number of young people, one of the things that the teams determine is that, averaged over the course of the 24 hours, one-third of all youngsters are found in school.

How can they interpret this result? Does it mean that two-thirds of students are truants and the remaining one-third spend all their time in school? No, we must bear in mind that the survey teams counted youngsters throughout the full 24-hour day. Some survey teams worked at night, when most youngsters were at home asleep, and others worked in the late afternoon, when most youngsters were on their way home from school (and more likely to be enjoying a pizza). If the survey was truly representative, we *can* conclude, however, that if an average of one-third of all youngsters are found in school, then humans ages 6 to 18 years must spend about one-third of *their time* in school.

We can do something similar for stars. We find that, on average, 90% of all stars are located on the main sequence of the H–R diagram. If we can identify some activity or life stage with the main sequence, then it follows that stars must spend 90% of their lives in that activity or life stage.

Understanding the Main Sequence

In **The Structure and Composition of the Sun**, we discussed the Sun as a representative star. We saw that what stars such as the Sun "do for a living" is to convert protons into helium deep in their interiors via the process of nuclear fusion, thus producing energy. The fusion of protons to helium is an excellent, long-lasting source of energy for a star because the bulk of every star consists of hydrogen atoms, whose nuclei are protons.

Our computer models of how stars evolve over time show us that a typical star will spend about 90% of its life fusing the abundant hydrogen in its core into helium. This then is a good explanation of why 90% of all stars are found on the main sequence in the H–R diagram. But if all the stars on the main sequence are doing the same thing (fusing hydrogen), why are they distributed along a sequence of points? That is, why do they differ in luminosity and surface temperature (which is what we are plotting on the H–R diagram)?

To help us understand how main-sequence stars differ, we can use one of the most important results from our studies of model stars. Astrophysicists have been able to show that the structure of stars that are in equilibrium and derive all their energy from nuclear fusion is completely and uniquely determined by just two quantities: the *total mass* and the *composition* of the star. This fact provides an interpretation of many features of the H–R diagram.

Imagine a cluster of stars forming from a cloud of interstellar “raw material” whose chemical composition is similar to the Sun’s. (We’ll describe this process in more detail in **Star Formation**, but for now, the details will not concern us.) In such a cloud, all the clumps of gas and dust that become stars begin with the same chemical composition and differ from one another only in mass. Now suppose that we compute a model of each of these stars for the time at which it becomes stable and derives its energy from nuclear reactions, but before it has time to alter its composition appreciably as a result of these reactions.

The models calculated for these stars allow us to determine their luminosities, temperatures, and sizes. If we plot the results from the models—one point for each model star—on the H–R diagram, we get something that looks just like the main sequence we saw for real stars.

And here is what we find when we do this. The model stars with the largest masses are the hottest and most luminous, and they are located at the upper left of the diagram.

The least-massive model stars are the coolest and least luminous, and they are placed at the lower right of the plot. The other model stars all lie along a line running diagonally across the diagram. In other words, *the main sequence turns out to be a sequence of stellar masses*.

This makes sense if you think about it. The most massive stars have the most gravity and can thus compress their centers to the greatest degree. This means they are the hottest inside and the best at generating energy from nuclear reactions deep within. As a result, they shine with the greatest luminosity and have the hottest surface temperatures. The stars with lowest mass, in turn, are the coolest inside and least effective in generating energy. Thus, they are the least luminous and wind up being the coolest on the surface. Our Sun lies somewhere in the middle of these extremes (as you can see in **Figure 22.28**). The characteristics of representative main-sequence stars (excluding brown dwarfs, which are not true stars) are listed in **Table 22.5**.

Characteristics of Main-Sequence Stars

Spectral Type	Mass (Sun = 1)	Luminosity (Sun = 1)	Temperature	Radius (Sun = 1)
O5	40	7×10^5	40,000 K	18
B0	16	2.7×10^5	28,000 K	7
A0	3.3	55	10,000 K	2.5
F0	1.7	5	7500 K	1.4
G0	1.1	1.4	6000 K	1.1
K0	0.8	0.35	5000 K	0.8
M0	0.4	0.05	3500 K	0.6

Table 22.5

Note that we’ve seen this 90% figure come up before. This is exactly what we found earlier when we examined the mass-luminosity relation (**Figure 22.23**). We observed that 90% of all stars seem to follow the relationship; these are the 90% of all stars that lie on the main sequence in our H–R diagram. Our models and our observations agree.

What about the other stars on the H–R diagram—the giants and supergiants, and the white dwarfs? As we will see in the next few chapters, these are what main-sequence stars turn into as they age: They are the later stages in a star’s life. As a star consumes its nuclear fuel, its source of energy changes, as do its chemical composition and interior structure. These changes cause the star to alter its luminosity and surface temperature so that it no longer lies on the main sequence on our diagram. Because stars spend much less time in these later stages of their lives, we see fewer stars in those regions of the H–R diagram.

Extremes of Stellar Luminosities, Diameters, and Densities

We can use the H–R diagram to explore the extremes in size, luminosity, and density found among the stars. Such extreme stars are not only interesting to fans of the *Guinness Book of World Records*; they can teach us a lot about how stars work. For example, we saw that the most massive main-sequence stars are the most luminous ones. We know of a few extreme stars that are a million times more luminous than the Sun, with masses that exceed 100 times the Sun's mass. These superluminous stars, which are at the upper left of the H–R diagram, are exceedingly hot, very blue stars of spectral type O. These are the stars that would be the most conspicuous at vast distances in space.

The cool supergiants in the upper corner of the H–R diagram are as much as 10,000 times as luminous as the Sun. In addition, these stars have diameters very much larger than that of the Sun. As discussed above, some supergiants are so large that if the solar system could be centered in one, the star's surface would lie beyond the orbit of Mars (see **Figure 22.30**). We will have to ask, in coming chapters, what process can make a star swell up to such an enormous size, and how long these "swollen" stars can last in their distended state.

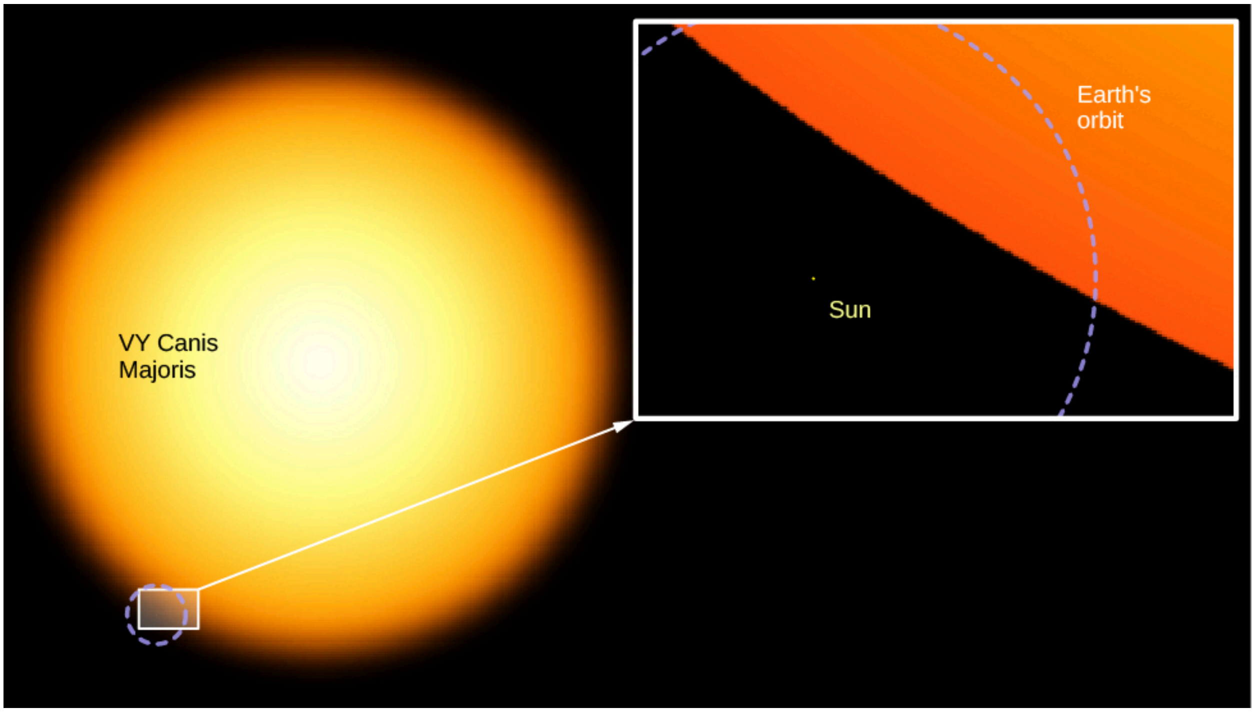


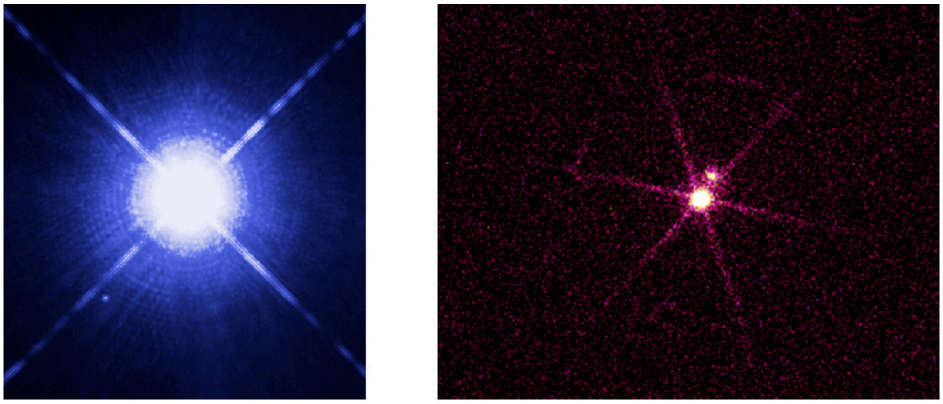
Figure 22.30 Here you see how small the Sun looks in comparison to one of the largest known stars: VY Canis Majoris, a supergiant.

In contrast, the very common red, cool, low-luminosity stars at the lower end of the main sequence are much smaller and more compact than the Sun. An example of such a red dwarf is Ross 614B, with a surface temperature of 2700 K and only 1/2000 of the Sun's luminosity. We call such a star a dwarf because its diameter is only 1/10 that of the Sun. A star with such a low luminosity also has a low mass (about 1/12 that of the Sun). This combination of mass and diameter means that it is so compressed that the star has an average density about 80 times that of the Sun. Its density must be higher, in fact, than that of any known solid found on the surface of Earth. (Despite this, the star is made of gas throughout because its center is so hot.)

The faint, red, main-sequence stars are not the stars of the most extreme densities, however. The white dwarfs, at the lower-left corner of the H–R diagram, have densities many times greater still.

The White Dwarfs

The first white dwarf star was detected in 1862. Called Sirius B, it forms a binary system with Sirius A, the brightest-appearing star in the sky. It eluded discovery and analysis for a long time because its faint light tends to be lost in the glare of nearby Sirius A (**Figure 22.31**). (Since Sirius is often called the Dog Star—being the brightest star in the constellation of Canis Major, the big dog—Sirius B is sometimes nicknamed the Pup.)



(a) (b)

Figure 22.31 (a) The (visible light) image, taken with the Hubble Space Telescope, shows bright Sirius A, and, below it and off to its left, faint Sirius B. (b) This image of the Sirius star system was taken with the Chandra X-Ray Telescope. Now, the bright object is the white dwarf companion, Sirius B. Sirius A is the faint object above it; what we are seeing from Sirius is probably not actually X-ray radiation but rather ultraviolet light that has leaked into the detector. Note that the ultraviolet intensities of these two objects are completely reversed from the situation in visible light because Sirius B is hotter and emits more higher-frequency radiation. (credit a: modification of work by NASA, H.E. Bond and E. Nelan (Space Telescope Science Institute), M. Barstow and M. Burleigh (University of Leicester) and J.B. Holberg (University of Arizona); credit b: modification of work by NASA/SAO/CXC)

We have now found thousands of white dwarfs. **Table 22.3** shows that about 7% of the true stars (spectral types O–M) in our local neighborhood are white dwarfs. A good example of a typical white dwarf is the nearby star 40 Eridani B. Its surface temperature is a relatively hot 12,000 K, but its luminosity is only 1/275 L_{Sun}. Calculations show that its radius is only 1.4% of the Sun's, or about the same as that of Earth, and its volume is 2.5×10^{-6} that of the Sun. Its mass, however, is 0.57 times the Sun's mass, just a little more than half. To fit such a substantial mass into so tiny a volume, the star's density must be about 210,000 times the density of the Sun, or more than 300,000 g/cm^3. A teaspoonful of this material would have a mass of some 1.6 tons! At such enormous densities, matter cannot exist in its usual state; we will examine the particular behavior of this type of matter in **The Death of Stars**. For now, we just note that white dwarfs are dying stars, reaching the end of their productive lives and ready for their stories to be over.

The British astrophysicist (and science popularizer) Arthur Eddington (1882–1944) described the first known white dwarf this way:

" The message of the companion of Sirius, when decoded, ran: "I am composed of material three thousand times denser than anything you've ever come across. A ton of my material would be a little nugget you could put in a matchbox." What reply could one make to something like that? Well, the reply most of us made in 1914 was, "Shut up; don't talk nonsense.""

Today, however, astronomers not only accept that stars as dense as white dwarfs exist but (as we will see) have found even denser and stranger objects in their quest to understand the evolution of different types of stars.

For Further Exploration

Websites

Discovery of Brown Dwarfs: **http://w.astro.berkeley.edu/~basri/bdwarfs/SciAm-book.pdf (http://w.astro.berkeley.edu/~basri/bdwarfs/SciAm-book.pdf)** .

Listing of Nearby Brown Dwarfs: **http://www.solstation.com/stars/pc10bd.htm (http://www.solstation.com/stars/pc10bd.htm)** .

Spectral Types of Stars: **http://www.skyandtelescope.com/astronomy-equipment/the-spectral-types-of-stars/ (http://www.skyandtelescope.com/astronomy-equipment/the-spectral-types-of-stars/)** .

Stellar Velocities **https://www.e-education.psu.edu/astro801/content/l4_p7.html (https://www.e-education.psu.edu/astro801/content/l4_p7.html)** .

Unheard Voices! The Contributions of Women to Astronomy: A Resource Guide: **http://multiverse.ssl.berkeley.edu/women (http://multiverse.ssl.berkeley.edu/women)** and **http://www.astrosociety.org/education/astronomy-resource-guides/women-in-astronomy-an-introductory-resource-guide/ (http://www.astrosociety.org/education/astronomy-resource-guides/women-in-astronomy-an-introductory-resource-guide/)** .

Eclipsing Binary Stars: **http://www.midnightkite.com/index.aspx?URL=Binary (http://www.midnightkite.com/index.aspx?URL=Binary)** . Dan Bruton at Austin State University has created this collection of animations, articles, and links showing how astronomers use eclipsing binary light curves.

Henry Norris Russell: **http://www.nasonline.org/publications/biographical-memoirs/memoir-pdfs/russell-henry-n.pdf (http://www.nasonline.org/publications/biographical-memoirs/memoir-pdfs/russell-henry-n.pdf)** . A biographic memoir by Harlow Shapley.

Henry Norris Russell: **http://www.phys-astro.sonoma.edu/brucemedalists/russell/RussellBio.pdf (http://www.phys-astro.sonoma.edu/brucemedalists/russell/RussellBio.pdf)** . A Bruce Medal profile of Russell.

Hertzsprung–Russell Diagram: **http://skyserver.sdss.org/dr1/en/proj/advanced/hr/ (http://skyserver.sdss.org/dr1/en/proj/advanced/hr/)** . This site from the Sloan Digital Sky Survey introduces the H–R diagram and gives you information for making your own. You can go step by step by using the menu at the left. Note that in the project instructions, the word “here” is a link and takes you to the data you need.

Stars of the Week: **http://stars.astro.illinois.edu/sow/sowlist.html (http://stars.astro.illinois.edu/sow/sowlist.html)** . Astronomer James Kaler does “biographical summaries” of famous stars—not the Hollywood type, but ones in the real sky.

Videos

When You Are Just Too Small to be a Star: **https://www.youtube.com/watch?v=zXCDsb4n4KU (https://www.youtube.com/watch?v=zXCDsb4n4KU)** . 2013 Public Talk on Brown Dwarfs and Planets by Dr. Gibor Basri of the University of California–Berkeley (1:32:52).

WISE Mission Surveys Nearby Stars: **http://www.jpl.nasa.gov/video/details.php?id=1089 (http://www.jpl.nasa.gov/video/details.php?id=1089)** . Short video about the WISE telescope survey of brown dwarfs and M dwarfs in our immediate neighborhood (1:21).

CHAPTER 22 REVIEW

KEY TERMS

binary stars two stars that revolve about each other

brown dwarf an object intermediate in size between a planet and a star; the approximate mass range is from about 1/100 of the mass of the Sun up to the lower mass limit for self-sustaining nuclear reactions, which is about 0.075 the mass of the Sun; brown dwarfs are capable of deuterium fusion, but not hydrogen fusion

brown dwarf an object intermediate in size between a planet and a star; the approximate mass range is from about 1/100 of the mass of the Sun up to the lower mass limit for self-sustaining nuclear reactions, which is about 1/12 the mass of the Sun

color index difference between the magnitudes of a star or other object measured in light of two different spectral regions—for example, blue minus visual (B–V) magnitudes

eclipsing binary a binary star in which the plane of revolution of the two stars is nearly edge-on to our line of sight, so that the light of one star is periodically diminished by the other passing in front of it

giant a star of exaggerated size with a large, extended photosphere

H–R diagram (Hertzsprung–Russell diagram) a plot of luminosity against surface temperature (or spectral type) for a group of stars

main sequence a sequence of stars on the Hertzsprung–Russell diagram, containing the majority of stars, that runs diagonally from the upper left to the lower right

mass-luminosity relation the observed relation between the masses and luminosities of many (90% of all) stars

proper motion the angular change per year in the direction of a star as seen from the Sun

radial velocity motion toward or away from the observer; the component of relative velocity that lies in the line of sight

selection effect the selection of sample data in a nonrandom way, causing the sample data to be unrepresentative of the entire data set

space velocity the total (three-dimensional) speed and direction with which an object is moving through space relative to the Sun

spectral class (or spectral type) the classification of stars according to their temperatures using the characteristics of their spectra; the types are O, B, A, F, G, K, and M with L, T, and Y added recently for cooler star-like objects that recent survey have revealed

spectroscopic binary a binary star in which the components are not resolved but whose binary nature is indicated by periodic variations in radial velocity, indicating orbital motion

visual binary a binary star in which the two components are telescopically resolved

white dwarf a low-mass star that has exhausted most or all of its nuclear fuel and has collapsed to a very small size; such a star is near its final state of life

KEY EQUATIONS

Kepler's Third Law for a binary system	$a^3 = (M_1 + M_2)T^2$
Mass-luminosity relation	$L \propto M^{3.9}$

SUMMARY

22.1 Colors of Stars

- Stars have different colors, which are indicators of temperature.

- The hottest stars tend to appear blue or blue-white, whereas the coolest stars are red.
- A color index of a star is the difference in the magnitudes measured at any two wavelengths and is one way that astronomers measure and express the temperature of stars.

22.2 The Spectra of Stars

- The differences in the spectra of stars are principally due to differences in temperature, not composition.
- The spectra of stars are described in terms of spectral classes.
- In order of decreasing temperature, these spectral classes are O, B, A, F, G, K, M, L, T, and Y.
- These are further divided into subclasses numbered from 0 to 9.
- The classes L, T, and Y have been added recently to describe newly discovered star-like objects—mainly brown dwarfs—that are cooler than M9.
- Our Sun has spectral type G2.

22.3 Using Spectra to Measure Stellar Radius, Composition, and Motion

- Spectra of stars of the same temperature but different atmospheric pressures have subtle differences, so spectra can be used to determine whether a star has a large radius and low atmospheric pressure (a giant star) or a small radius and high atmospheric pressure.
- Stellar spectra can also be used to determine the chemical composition of stars; hydrogen and helium make up most of the mass of all stars.
- Measurements of line shifts produced by the Doppler effect indicate the radial velocity of a star.
- Broadening of spectral lines by the Doppler effect is a measure of rotational velocity.
- A star can also show proper motion, due to the component of a star's space velocity across the line of sight.

22.4 A Stellar Census

- To understand the properties of stars, we must make wide-ranging surveys.
- We find the stars that appear brightest to our eyes are bright primarily because they are intrinsically very luminous, not because they are the closest to us.
- Most of the nearest stars are intrinsically so faint that they can be seen only with the aid of a telescope.
- Stars with low mass and low luminosity are much more common than stars with high mass and high luminosity.
- Most of the brown dwarfs in the local neighborhood have not yet been discovered.

22.5 Measuring Stellar Masses

- The masses of stars can be determined by analysis of the orbit of binary stars—two stars that orbit a common center of mass.
- In visual binaries, the two stars can be seen separately in a telescope, whereas in a spectroscopic binary, only the spectrum reveals the presence of two stars.
- Stellar masses range from about 1/12 to more than 100 times the mass of the Sun (in rare cases, going to 250 times the Sun's mass).
- Objects with masses between 1/12 and 1/100 that of the Sun are called brown dwarfs.
- Objects in which no nuclear reactions can take place are planets.
- The most massive stars are, in most cases, also the most luminous, and this correlation is known as the mass-luminosity relation.

22.6 Diameters of Stars

- The diameters of stars can be determined by measuring the time it takes an object (the Moon, a planet, or a companion star) to pass in front of it and block its light.

- Diameters of members of eclipsing binary systems (where the stars pass in front of each other) can be determined through analysis of their orbital motions.

22.7 The H-R Diagram

- The Hertzsprung–Russell diagram, or H–R diagram, is a plot of stellar luminosity against surface temperature.
- Most stars lie on the main sequence, which extends diagonally across the H–R diagram from high temperature and high luminosity to low temperature and low luminosity.
- The position of a star along the main sequence is determined by its mass.
- High-mass stars emit more energy and are hotter than low-mass stars on the main sequence.
- Main-sequence stars derive their energy from the fusion of protons to helium.
- About 90% of the stars lie on the main sequence.
- Only about 10% of the stars are white dwarfs, and fewer than 1% are giants or supergiants.

CONCEPTUAL QUESTIONS

22.1 Colors of Stars

1. Explain why color is a measure of a star's temperature.

2. How would two stars of equal luminosity—one blue and the other red—appear in an image taken through a filter that passes mainly blue light? How would their appearance change in an image taken through a filter that transmits mainly red light?

3. Suppose you are given the task of measuring the colors of the brightest stars, listed in **Appendix D**, through three filters: the first transmits blue light, the second transmits yellow light, and the third transmits red light. If you observe the star Vega, it will appear equally bright through each of the three filters. Which stars will appear brighter through the blue filter than through the red filter? Which stars will appear brighter through the red filter? Which star is likely to have colors most nearly like those of Vega?

4. Sam, a college student, just bought a new car. Sam's friend Adam, a graduate student in astronomy, asks Sam for a ride. In the car, Adam remarks that the colors on the temperature control are wrong. Why did he say that?

Figure 22.32 (credit: modification of work by Michael Sheehan)

22.2 The Spectra of Stars

5. What is the main reason that the spectra of all stars are not identical? Explain.

6. What elements are stars mostly made of? How do we know this?

7. What did Annie Cannon contribute to the understanding of stellar spectra?

8. Name five characteristics of a star that can be determined by measuring its spectrum. Explain how you would use a spectrum to determine these characteristics.

9. How do objects of spectral types L, T, and Y differ from those of the other spectral types?

10. Order the seven basic spectral types from hottest to coldest.

11. What is the defining difference between a brown dwarf and a true star?

12. Table 22.2 lists the temperature ranges that correspond to the different spectral types. What part of the star do these temperatures refer to? Why?

13. Star X has lines of ionized helium in its spectrum, and star Y has bands of titanium oxide. Which is hotter? Why? The spectrum of star Z shows lines of ionized helium and also molecular bands of titanium oxide. What is strange about this spectrum? Can you suggest an explanation?

14. The spectrum of the Sun has hundreds of strong lines of nonionized iron but only a few, very weak lines of helium. A star of spectral type B has very strong lines of helium but very weak iron lines. Do these differences mean that the Sun contains more iron and less helium than the B star? Explain.

15. What are the approximate spectral classes of stars with the following characteristics?

A. Balmer lines of hydrogen are very strong; some lines of ionized metals are present.
B. The strongest lines are those of ionized helium.
C. Lines of ionized calcium are the strongest in the spectrum; hydrogen lines show only moderate strength; lines of neutral and metals are present.
D. The strongest lines are those of neutral metals and bands of titanium oxide.

22.3 Using Spectra to Measure Stellar Radius, Composition, and Motion

16. Do stars that look brighter in the sky have larger or smaller magnitudes than fainter stars?

17. The star Antares has an apparent magnitude of 1.0, whereas the star Procyon has an apparent magnitude of 0.4. Which star appears brighter in the sky?

18. Based on their colors, which of the following stars is hottest? Which is coolest? Archenar (blue), Betelgeuse (red), Capella (yellow).

19. Look at the chemical elements in Appendix F. Can you identify any relationship between the abundance of an element and its atomic weight? Are there any obvious exceptions to this relationship?

20. Appendix D lists some of the nearest stars. Are most of these stars hotter or cooler than the Sun? Do any of them emit more energy than the Sun? If so, which ones?

21. Appendix D lists the stars that appear brightest in our sky. Are most of these hotter or cooler than the Sun? Can you suggest a reason for the difference between this answer and the answer to the previous question? (Hint: Look at the luminosities.) Is there any tendency for a correlation between temperature and luminosity? Are there exceptions to the correlation?

22. What star appears the brightest in the sky (other than the Sun)? The second brightest? What color is Betelgeuse? Use Appendix D to find the answers.

23. Why can only a lower limit to the rate of stellar rotation be determined from line broadening rather than the actual rotation rate? (Refer to Figure 22.13.)

24. Two stars have proper motions of one arcsecond per year. Star A is 20 light-years from Earth, and Star B is 10 light-years away from Earth. Which one has the faster velocity in space?

25. Suppose there are three stars in space, each moving at 100 km/s. Star A is moving across (i.e., perpendicular to) our line of sight, Star B is moving directly away from Earth, and Star C is moving away from Earth, but at a 30° angle to the line of sight. From which star will you observe the greatest Doppler shift? From which star will you observe the smallest Doppler shift?

22.4 A Stellar Census

26. Suppose you want to determine the average educational level of people throughout the nation. Since it would be a great deal of work to survey every citizen, you decide to make your task easier by asking only the people on your campus. Will you get an accurate answer? Will your survey be distorted by a selection effect? Explain.

22.5 Measuring Stellar Masses

27. Why do most known visual binaries have relatively long periods and most spectroscopic binaries have relatively short periods?

28. Figure 22.25 shows the light curve of a hypothetical eclipsing binary star in which the light of one star is completely blocked by another. What would the light curve look like for a system in which the light of the smaller star is only partially blocked by the larger one? Assume the smaller star is the hotter one. Sketch the relative positions of the two stars that correspond to various portions of the light curve.

29. There are fewer eclipsing binaries than spectroscopic binaries. Explain why.

30. Within 50 light-years of the Sun, visual binaries outnumber eclipsing binaries. Why?

31. Which is easier to observe at large distances—a spectroscopic binary or a visual binary?

32. The eclipsing binary Algol drops from maximum to minimum brightness in about 4 hours, remains at minimum brightness for 20 minutes, and then takes another 4 hours to return to maximum brightness. Assume that we view this system exactly edge-on, so that one star crosses directly in front of the other. Is one star much larger than the other, or are they fairly similar in size? (Hint: Refer to the diagrams of eclipsing binary light curves.)

33. If a visual binary system were to have two equal-mass stars, how would they be located relative to the center of the mass of the system? What would you observe as you watched these stars as they orbited the center of mass, assuming very circular orbits, and assuming the orbit was face on to your view?

34. Two stars are in a visual binary star system that we see face on. One star is very massive whereas the other is much less massive. Assuming circular orbits, describe their relative orbits in terms of orbit size, period, and orbital velocity.

35. Describe the spectra for a spectroscopic binary for a system comprised of an F-type and L-type star. Assume that the system is too far away to be able to easily observe the L-type star.

36. Figure 22.21 shows the velocity of two stars in a spectroscopic binary system. Which star is the most massive? Explain your reasoning.

22.6 Diameters of Stars

37. Describe two ways of determining the diameter of a star.

38. You are able to take spectra of both stars in an eclipsing binary system. List all properties of the stars that can be measured from their spectra and light curves.

39. One method to measure the diameter of a star is to use an object like the Moon or a planet to block out its light and to measure the time it takes to cover up the object. Why is this method used more often with the Moon rather than the planets, even though there are more planets?

22.7 The H-R Diagram

40. What are the largest- and smallest-known values of the mass, luminosity, surface temperature, and diameter of stars (roughly)?

41. Sketch an H–R diagram. Label the axes. Show where cool supergiants, white dwarfs, the Sun, and main-sequence stars are found.

42. Describe what a typical star in the Galaxy would be like compared to the Sun.

43. Describe how the mass, luminosity, surface temperature, and radius of main-sequence stars change in value going from the "bottom" to the "top" of the main sequence.

44. Is the Sun an average star? Why or why not?

45. Review this spectral data for five stars.

Table A

Star	Spectrum
1	G, main sequence
2	K, giant
3	K, main sequence
4	O, main sequence
5	M, main sequence

Which is the hottest? Coolest? Most luminous? Least luminous? In each case, give your reasoning.

46. Which changes by the largest factor along the main sequence from spectral types O to M—mass or luminosity?

47. Suppose you want to search for brown dwarfs using a space telescope. Will you design your telescope to detect light in the ultraviolet or the infrared part of the spectrum? Why?

48. An astronomer discovers a type-M star with a large luminosity. How is this possible? What kind of star is it?

49. Approximately 9000 stars are bright enough to be seen without a telescope. Are any of these white dwarfs? Use the information given in this chapter to explain your reasoning.

50. Use the data in Appendix D to plot an H–R diagram for the brightest stars. Use the data from Table 22.5 to show where the main sequence lies. Do 90% of the brightest stars lie on or near the main sequence? Explain why or why not.

51. Use the diagram you have drawn for Exercise 22.50 to answer the following questions: Which star is more massive—Sirius or Alpha Centauri? Rigel and Regulus have nearly the same spectral type. Which is larger? Rigel and Betelgeuse have nearly the same luminosity. Which is larger? Which is redder?

52. Use the data in **Appendix D** to plot an H–R diagram for a sample of nearby stars. How does this plot differ from the one for the brightest stars in **Exercise 22.50**? Why?

53. You go out stargazing one night, and someone asks you how far away the brightest stars we see in the sky without a telescope are. What would be a good, general response? (Use **Appendix D** for more information.)

54. If you were to compare three stars with the same surface temperature, with one star being a giant, another a supergiant, and the third a main-sequence star, how would their radii compare to one another?

55. Are supergiant stars also extremely massive? Explain the reasoning behind your answer.

56. Consider the following data on four stars:

Table B

Star	Luminosity (in L_{Sun})	Type
1	100	B, main sequence
2	1/100	B, white dwarf
3	1/100	M, main sequence
4	100	M, giant

Which star would have the largest radius? Which star would have the smallest radius? Which star is the most common in our area of the Galaxy? Which star is the least common?

PROBLEMS

22.2 The Spectra of Stars

57. You have enough information from this chapter to estimate the distance to Alpha Centauri, the second nearest star, which has an apparent magnitude of 0. Since it is a G2 star, like the Sun, assume it has the same luminosity as the Sun and the difference in magnitudes is a result only of the difference in distance. Estimate how far away Alpha Centauri is. Describe the necessary steps in words and then do the calculation. (As we will learn in the **Celestial Distances** chapter, this method—namely, assuming that stars with identical spectral types emit the same amount of energy—is actually used to estimate distances to stars.) If you assume the distance to the Sun is in AU, your answer will come out in AU.

58. Do the previous problem again, this time using the information that the Sun is 150,000,000 km away. You will get a very large number of km as your answer. To get a better feeling for how the distances compare, try calculating the time it takes light at a speed of 299,338 km/s to travel from the Sun to Earth and from Alpha Centauri to Earth. For Alpha Centauri, figure out how long the trip will take in years as well as in seconds.

59. Our Sun, a type G star, has a surface temperature of 5800 K. We know, therefore, that it is cooler than a type O star and hotter than a type M star. Given what you learned about the temperature ranges of these types of stars, how many times hotter than our Sun is the hottest type O star? How many times cooler than our Sun is the coolest type M star?

22.5 Measuring Stellar Masses

60. If two stars are in a binary system with a combined mass of 5.5 solar masses and an orbital period of 12 years, what is the average distance between the two stars?

61. We can estimate the masses of most of the stars in **Appendix D** from the mass-luminosity relationship in **Figure 22.23**. However, remember this relationship works only for main sequence stars. Determine which of the first 10 stars in **Appendix D** are main sequence stars. Use one of the figures in this chapter. Make a table of stars' masses.

22.6 Diameters of Stars

62. In this section, the relative diameters of the two stars in the Sirius system were determined. Let's use this value to explore other aspects of this system. This will be done through several steps, each in its own exercise. Assume the temperature of the Sun is 5800 K, and the temperature of Sirius A, the larger star of the binary, is
10,000 K. The luminosity of Sirius A can be found in **Appendix D**, and is given as about 23 times that of the Sun. Using the values provided, calculate the radius of Sirius A relative to that of the Sun.

63. Now calculate the radius of Sirius' white dwarf companion, Sirius B, to the Sun.

64. How does this radius of Sirius B compare with that of Earth?

65. From the previous calculations and the results from this section, it is possible to calculate the density of Sirius

B relative to the Sun. It is worth noting that the radius of the companion is very similar to that of Earth, whereas the mass is very similar to the Sun's. How does the companion's density compare to that of the Sun? Recall that density = mass/volume, and the volume of a sphere = $(4/3)\pi R^3$. How does this density compare with that of water and other materials discussed in this text? Can you see why astronomers were so surprised and puzzled when they first determined the orbit of the companion to Sirius?

22.7 The H-R Diagram

66. If a 100 solar mass star were to have a luminosity of 10^7 times the Sun's luminosity, how would such a star's density compare when it is on the main sequence as an O-type star, and when it is a cool supergiant (M-type)? Use values of temperature from **Figure 22.28** or **Figure 22.29** and the relationship between luminosity, radius, and temperature as given in **Exercise 22.73**.

67. If Betelgeuse had a mass that was 25 times that of the Sun, how would its average density compare to that of the Sun? Use the definition of $\text{density} = \frac{\text{mass}}{\text{volume}}$, where the volume is that of a sphere.

ADDITIONAL PROBLEMS

68. It is possible that stars as much as 200 times the Sun's mass or more exist. What is the luminosity of such a star based upon the mass-luminosity relation?

69. The lowest mass for a true star is 1/12 the mass of the Sun. What is the luminosity of such a star based upon the mass-luminosity relationship?

70. Spectral types are an indicator of temperature. For the first 10 stars in **Appendix D**, the list of the brightest stars in our skies, estimate their temperatures from their spectral types. Use information in the figures and/or tables in this chapter and describe how you made the estimates.

71. How much would you weigh if you were suddenly transported to the white dwarf Sirius B? You may use your own weight (or if don't want to own up to what it is, assume you weigh 70 kg or 150 lb). In this case, assume that the companion to Sirius has a mass equal to that of the Sun and a radius equal to that of Earth. Remember Newton's law of gravity:

$$F = GM_1 M_2 / R^2$$

and that your weight is proportional to the force that you feel. What kind of star should you travel to if you want to *lose* weight (and not gain it)?

72. The star Betelgeuse has a temperature of 3400 K and a luminosity of 13,200 L_{Sun}. Calculate the radius of Betelgeuse relative to the Sun.

73. Using the information provided in **Table 22.3**, what is the average stellar density in our part of the Galaxy? Use only the true stars (types O–M) and assume a spherical distribution with radius of 26 light-years.

74. Confirm that the angular diameter of the Sun of 1/2° corresponds to a linear diameter of 1.39 million km. Use the average distance of the Sun and Earth to derive the answer. (Hint: This can be solved using a trigonometric function.)

75. An eclipsing binary star system is observed with the following contact times for the main eclipse:

Table C

Contact	Time	Date
First contact	12:00 p.m.	March 12
Second contact	4:00 p.m.	March 13
Third contact	9:00 a.m.	March 18
Fourth contact	1:00 p.m.	March 19

The orbital velocity of the smaller star relative to the larger is 62,000 km/h. Determine the diameters for each star in the system.

23 | CELESTIAL DISTANCES

Figure 23.1 This beautiful image shows a giant cluster of stars called Messier 80, located about 28,000 light-years from Earth. Such crowded groups, which astronomers call globular clusters, contain hundreds of thousands of stars, including some of the RR Lyrae variables discussed in this chapter. Especially obvious in this picture are the bright red giants, which are stars similar to the Sun in mass that are nearing the ends of their lives. (credit: modification of work by The Hubble Heritage Team (AURA/ STScI/ NASA))

Chapter Outline

Introduction

How large is the universe? What is the most distant object we can see? These are among the most fundamental questions astronomers can ask. But just as babies must crawl before they can take their first halting steps, so too must we start with a more modest question: How far away are the stars? And even this question proves to be very hard to answer. After all, stars are mere points of light. Suppose you see a point of light in the darkness when you are driving on a country road late at night. How can you tell whether it is a nearby firefly, an oncoming motorcycle some distance away, or the porchlight of a house much farther down the road? It's not so easy, is it? Astronomers faced an even more difficult problem when they tried to estimate how far away the stars are.

In this chapter, we begin with the fundamental definitions of distances on Earth and then extend our reach outward to the stars. We will also examine the newest satellites that are surveying the night sky and discuss the special types of stars that can be used as trail markers to distant galaxies.

23.1 | Fundamental Units of Distance

Learning Objectives

By the end of this section, you will be able to:

- Understand the importance of defining a standard distance unit
- Explain how the meter was originally defined and how it has changed over time
- Discuss how radar is used to measure distances to the other members of the solar system

The first measures of distances were based on human dimensions—the inch as the distance between knuckles on the finger, or the yard as the span from the extended index finger to the nose of the British king. Later, the requirements of commerce led to some standardization of such units, but each nation tended to set up its own definitions. It was not until the middle of the eighteenth century that any real efforts were made to establish a uniform, international set of standards.

The Metric System

One of the enduring legacies of the era of the French emperor Napoleon is the establishment of the *metric system* of units, officially adopted in France in 1799 and now used in most countries around the world. The fundamental metric unit of length is the *meter*, originally defined as one ten-millionth of the distance along Earth's surface from the equator to the pole. French astronomers of the seventeenth and eighteenth centuries were pioneers in determining the dimensions of Earth, so it was logical to use their information as the foundation of the new system.

Practical problems exist with a definition expressed in terms of the size of Earth, since anyone wishing to determine the distance from one place to another can hardly be expected to go out and re-measure the planet. Therefore, an intermediate standard meter consisting of a bar of platinum-iridium metal was set up in Paris. In 1889, by international agreement, this bar was defined to be exactly one meter in length, and precise copies of the original meter bar were made to serve as standards for other nations.

Other units of length are derived from the meter. Thus, 1 kilometer (km) equals 1000 meters, 1 centimeter (cm) equals 1/100 meter, and so on. Even the old British and American units, such as the inch and the mile, are now defined in terms of the metric system.

Modern Redefinitions of the Meter

In 1960, the official definition of the meter was changed again. As a result of improved technology for generating spectral lines of precisely known wavelengths (see the chapter on **Spectroscopy**), the meter was redefined to equal 1,650,763.73 wavelengths of a particular atomic transition in the element krypton-86. The advantage of this redefinition is that anyone with a suitably equipped laboratory can reproduce a standard meter, without reference to any particular metal bar.

In 1983, the meter was defined once more, this time in terms of the velocity of light. Light in a vacuum can travel a distance of one meter in 1/299,792,458.6 second. Today, therefore, light travel time provides our basic unit of length. Put another way, a distance of *one light-second* (the amount of space light covers in one second) is defined to be 299,792,458.6 meters. That's almost 300 million meters that light covers in just one second; light really is *very* fast! We could just as well use the light-second as the fundamental unit of length, but for practical reasons (and to respect tradition), we have defined the meter as a small fraction of the light-second.

Distance within the Solar System

The work of Copernicus and Kepler established the *relative* distances of the planets—that is, how far from the Sun one planet is compared to another (see **Kepler's Laws of Planetary Motion** and **The Newtonian Synthesis**). But their work could not establish the *absolute* distances (in light-seconds or meters or other standard units of length). This is like knowing the height of all the students in your class only as compared to the height of your astronomy instructor, but not in inches or centimeters. Somebody's height has to be measured directly.

Similarly, to establish absolute distances, astronomers had to measure one distance in the solar system directly. Generally, the closer to us the object is, the easier such a measurement would be. Estimates of the distance to Venus were made as Venus crossed the face of the Sun in 1761 and 1769, and an international campaign was organized to estimate the distance to the asteroid Eros in the early 1930s, when its orbit brought it close to Earth. More recently, Venus crossed (or *transited*) the surface of the Sun in 2004 and 2012, and allowed us to make a modern distance estimate, although, as we will see below, by then it wasn't needed (**Figure 23.2**).

If you would like more information on just how the motion of Venus across the Sun helped us pin down distances in the solar system, you can turn to a **nice explanation (https://openstaxcollege.org/l/30VenusandSun)** by a NASA astronomer.

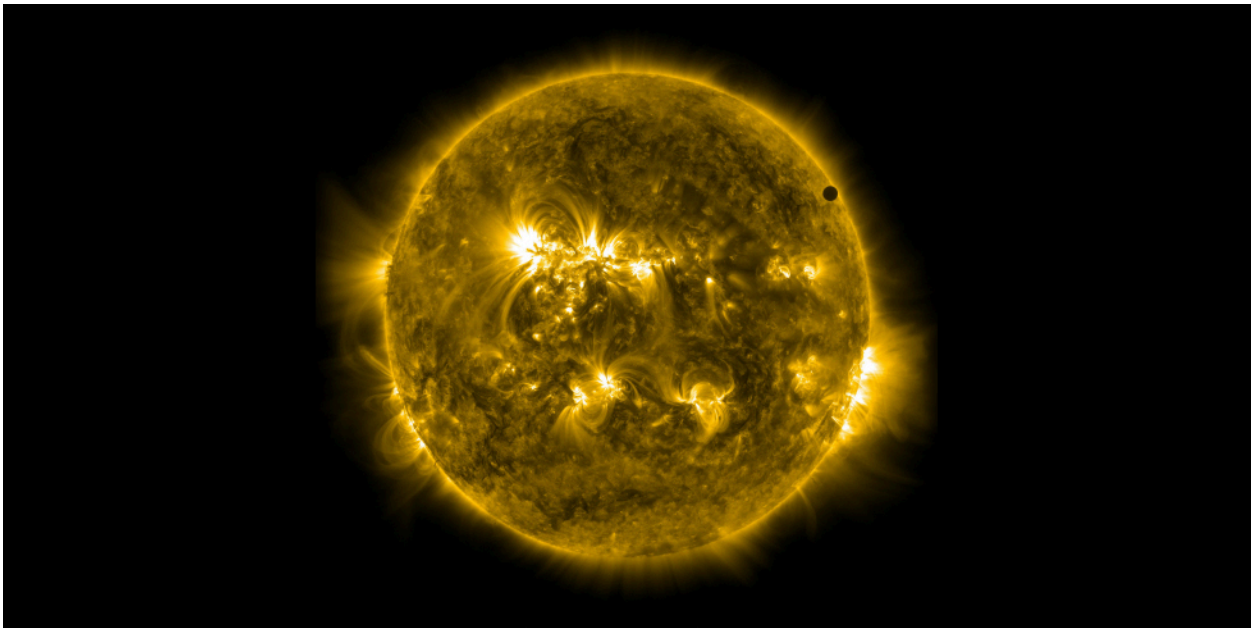


Figure 23.2 This striking "picture" of Venus crossing the face of the Sun (it's the black dot at about 2 o'clock) is more than just an impressive image. Taken with the Solar Dynamics Observatory spacecraft and special filters, it shows a modern transit of Venus. Such events allowed astronomers in the 1800s to estimate the distance to Venus. They measured the time it took Venus to cross the face of the Sun from different latitudes on Earth. The differences in times can be used to estimate the distance to the planet. Today, radar is used for much more precise distance estimates. (credit: modification of work by NASA/SDO, AIA)

The key to our modern determination of solar system dimensions is radar, a type of radio wave that can bounce off solid objects (**Figure 23.3**). As discussed in several earlier chapters, by timing how long a radar beam (traveling at the speed of light) takes to reach another world and return, we can measure the distance involved very accurately. In 1961, radar signals were bounced off Venus for the first time, providing a direct measurement of the distance from Earth to Venus in terms of light-seconds (from the roundtrip travel time of the radar signal).

Subsequently, radar has been used to determine the distances to Mercury, Mars, the satellites of Jupiter, the rings of Saturn, and several asteroids. Note, by the way, that it is not possible to use radar to measure the distance to the Sun directly because the Sun does not reflect radar very efficiently. But we can measure the distance to many other solar system objects and use Kepler's laws to give us the distance to the Sun.

Figure 23.3 This dish-shaped antenna, part of the NASA Deep Space Network in California's Mojave Desert, is 70 meters wide. Nicknamed the "Mars antenna," this radar telescope can send and receive radar waves, and thus measure the distances to planets, satellites, and asteroids. (credit: NASA/JPL-Caltech)

From the various (related) solar system distances, astronomers selected the average distance from Earth to the Sun as our standard "measuring stick" within the solar system. When Earth and the Sun are closest, they are about 147.1 million kilometers apart; when Earth and the Sun are farthest, they are about 152.1 million kilometers apart. The average of these two distances is called the astronomical unit (AU). We then express all the other distances in the solar system in terms of the AU. Years of painstaking analyses of radar measurements have led to a determination of the length of the AU to a precision of about one part in a billion. The length of 1 AU can be expressed in light travel time as 499.004854 light-seconds, or about 8.3 light-minutes. If we use the definition of the meter given previously, this is equivalent to 1 AU = 149,597,870,700 meters.

These distances are, of course, given here to a much higher level of precision than is normally needed. In this text, we are usually content to express numbers to a couple of significant places and leave it at that. For our purposes, it will be sufficient to round off these numbers:

$$\text{speed of light: } c = 3 \times 10^8 \text{ m/s} = 3 \times 10^5 \text{ km/s}$$
$$\text{length of light-second: ls} = 3 \times 10^8 \text{ m} = 3 \times 10^5 \text{ km}$$
$$\text{astronomical unit: AU} = 1.50 \times 10^{11} \text{ m} = 1.50 \times 10^8 \text{ km} = 500 \text{ light-seconds}$$

We now know the absolute distance scale within our own solar system with fantastic accuracy. This is the first link in the chain of cosmic distances.

The distances between the celestial bodies in our solar system are sometimes difficult to grasp or put into perspective. This **interactive website (https://openstaxcollege.org/l/30DistanceScale)** provides a "map" that shows the distances by using a scale at the bottom of the screen and allows you to scroll (using your arrow keys) through screens of "empty space" to get to the next planet—all while your current distance from the Sun is visible on the scale.

23.2 | Surveying the Stars

Learning Objectives

By the end of this section, you will be able to:

- Understand the concept of triangulating distances to distant objects, including stars
- Explain why space-based satellites deliver more precise distances than ground-based methods
- Discuss astronomers' efforts to study the stars closest to the Sun

It is an enormous step to go from the planets to the stars. For example, our Voyager 1 probe, which was launched in 1977, has now traveled farther from Earth than any other spacecraft. As this is written in 2016, Voyager 1 is 134 AU from the Sun.[1] The nearest star, however, is hundreds of thousands of AU from Earth. Even so, we can, in principle, survey distances to the stars using the same technique that a civil engineer employs to survey the distance to an inaccessible mountain or tree—the method of *triangulation*.

Triangulation in Space

A practical example of triangulation is your own depth perception. As you are pleased to discover every morning when you look in the mirror, your two eyes are located some distance apart. You therefore view the world from two different vantage points, and it is this dual perspective that allows you to get a general sense of how far away objects are.

To see what we mean, take a pen and hold it a few inches in front of your face. Look at it first with one eye (closing the other) and then switch eyes. Note how the pen seems to shift relative to objects across the room. Now hold the pen at arm's length: the shift is less. If you play with moving the pen for a while, you will notice that the farther away you hold it, the less it seems to shift. Your brain automatically performs such comparisons and gives you a pretty good sense of how far away things in your immediate neighborhood are.

If your arms were made of rubber, you could stretch the pen far enough away from your eyes that the shift would become imperceptible. This is because our depth perception fails for objects more than a few tens of meters away. In order to see the shift of an object a city block or more from you, your eyes would need to be spread apart a lot farther.

Let's see how surveyors take advantage of the same idea. Suppose you are trying to measure the distance to a tree across a deep river (**Figure 23.4**). You set up two observing stations some distance apart. That distance (line AB in **Figure 23.4**) is called the *baseline*. Now the direction to the tree (C in the figure) in relation to the baseline is observed from each station. Note that C appears in different directions from the two stations. This apparent change in direction of the remote object due to a change in vantage point of the observer is called **parallax**.

1. To have some basis for comparison, the dwarf planet Pluto orbits at an average distance of 40 AU from the Sun, and the dwarf planet Eris is currently roughly 96 AU from the Sun.

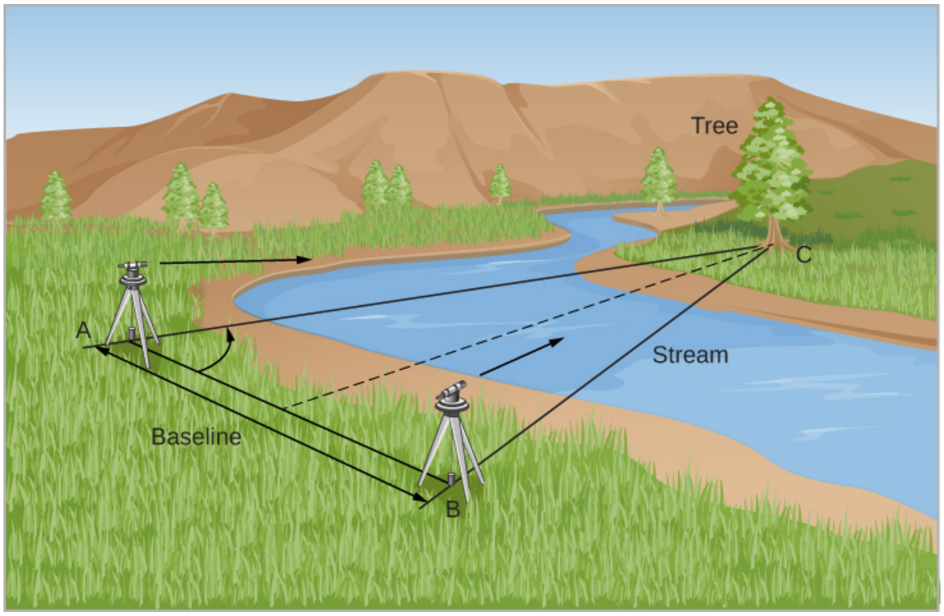


Figure 23.4 Triangulation allows us to measure distances to inaccessible objects. By getting the angle to a tree from two different vantage points, we can calculate the properties of the triangle they make and thus the distance to the tree.

The parallax is also the angle that lines AC and BC make—in mathematical terms, the angle subtended by the baseline. A knowledge of the angles at A and B and the length of the baseline, AB, allows the triangle ABC to be solved for any of its dimensions—say, the distance AC or BC. The solution could be reached by constructing a scale drawing or by using trigonometry to make a numerical calculation. If the tree were farther away, the whole triangle would be longer and skinnier, and the parallax angle would be smaller. Thus, we have the general rule that the smaller the parallax, the more distant the object we are measuring must be.

In practice, the kinds of baselines surveyors use for measuring distances on Earth are completely useless when we try to gauge distances in space. The farther away an astronomical object lies, the longer the baseline has to be to give us a reasonable chance of making a measurement. Unfortunately, nearly all astronomical objects are very far away. To measure their distances requires a very large baseline and highly precise angular measurements. The Moon is the only object near enough that its distance can be found fairly accurately with measurements made without a telescope. Ptolemy determined the distance to the Moon correctly to within a few percent. He used the turning Earth itself as a baseline, measuring the position of the Moon relative to the stars at two different times of night.

With the aid of telescopes, later astronomers were able to measure the distances to the nearer planets and asteroids using Earth's diameter as a baseline. This is how the AU was first established. To reach for the stars, however, requires a much longer baseline for triangulation and extremely sensitive measurements. Such a baseline is provided by Earth's annual trip around the Sun.

Distances to Stars

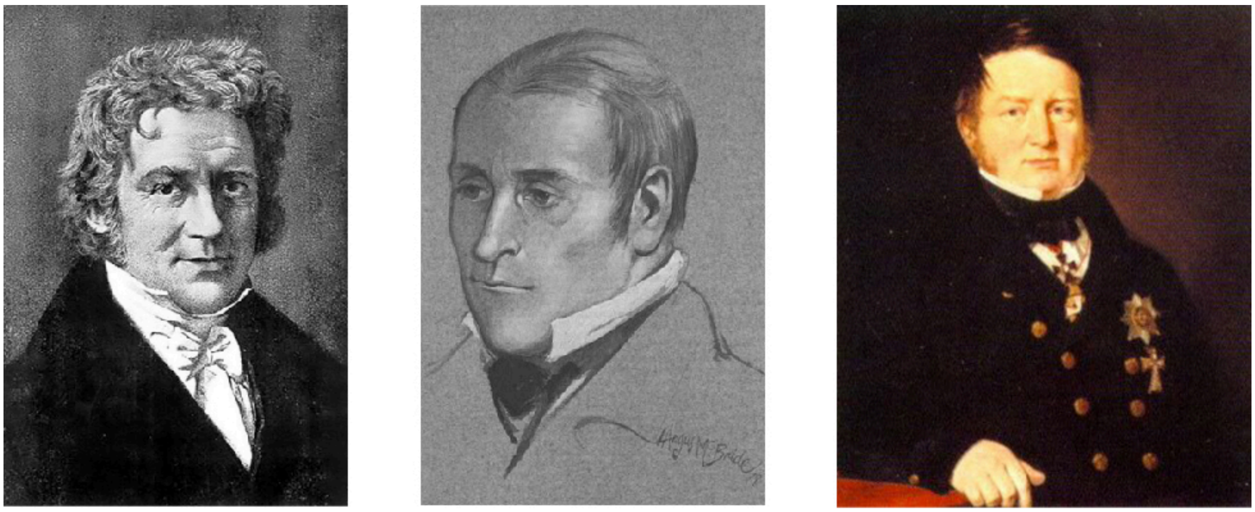
As Earth travels from one side of its orbit to the other, it graciously provides us with a baseline of 2 AU, or about 300 million kilometers. Although this is a much bigger baseline than the diameter of Earth, the stars are *so far away* that the resulting parallax shift is *still* not visible to the naked eye—not even for the closest stars.

This dilemma perplexed the ancient Greeks, some of whom had actually suggested that the Sun might be the center of the solar system, with Earth in motion around it. Aristotle and others argued, however, that Earth could not be revolving about the Sun. If it were, they said, we would surely observe the parallax of the nearer stars against the background of more distant objects as we viewed the sky from different parts of Earth's orbit (**Figure 23.6**). Tycho Brahe (1546–1601) advanced the same faulty argument nearly 2000 years later, when his careful measurements of stellar positions with the unaided eye revealed no such shift.

These early observers did not realize how truly distant the stars were and how small the change in their positions therefore was, even with the entire orbit of Earth as a baseline. The problem was that they did not have tools to measure parallax shifts too small to be seen with the human eye. By the eighteenth century, when there was no longer serious doubt about Earth's revolution, it became clear that the stars must be extremely distant. Astronomers equipped with telescopes began to devise instruments capable of measuring the tiny shifts of nearby stars relative to the background of more distant (and thus unshifting) celestial objects.

This was a significant technical challenge, since, even for the nearest stars, parallax angles are usually only a fraction of a second of arc. Recall that one second of arc (arcsec) is an angle of only 1/3600 of a degree. A coin the size of a US quarter would appear to have a diameter of 1 arcsecond if you were viewing it from a distance of about 5 kilometers (3 miles). Think about how small an angle that is. No wonder it took astronomers a long time before they could measure such tiny shifts.

The first successful detections of stellar parallax were in the year 1838, when Friedrich Bessel in Germany (**Figure 23.5**), Thomas Henderson, a Scottish astronomer working at the Cape of Good Hope, and Friedrich Struve in Russia independently measured the parallaxes of the stars 61 Cygni, Alpha Centauri, and Vega, respectively. Even the closest star, Alpha Centauri, showed a total displacement of only about 1.5 arcseconds during the course of a year.



(a) (b) (c)

Figure 23.5 (a) Bessel made the first authenticated measurement of the distance to a star (61 Cygni) in 1838, a feat that had eluded many dedicated astronomers for almost a century. But two others, (b) Scottish astronomer Thomas J. Henderson and (c) Friedrich Struve, in Russia, were close on his heels.

Figure 23.6 shows how such measurements work. Seen from opposite sides of Earth's orbit, a nearby star shifts position when compared to a pattern of more distant stars. Astronomers actually define parallax to be *one-half* the angle that a star shifts when seen from opposite sides of Earth's orbit (the angle labeled *P* in **Figure 23.6**). The reason for this definition is just that they prefer to deal with a baseline of 1 AU instead of 2 AU.

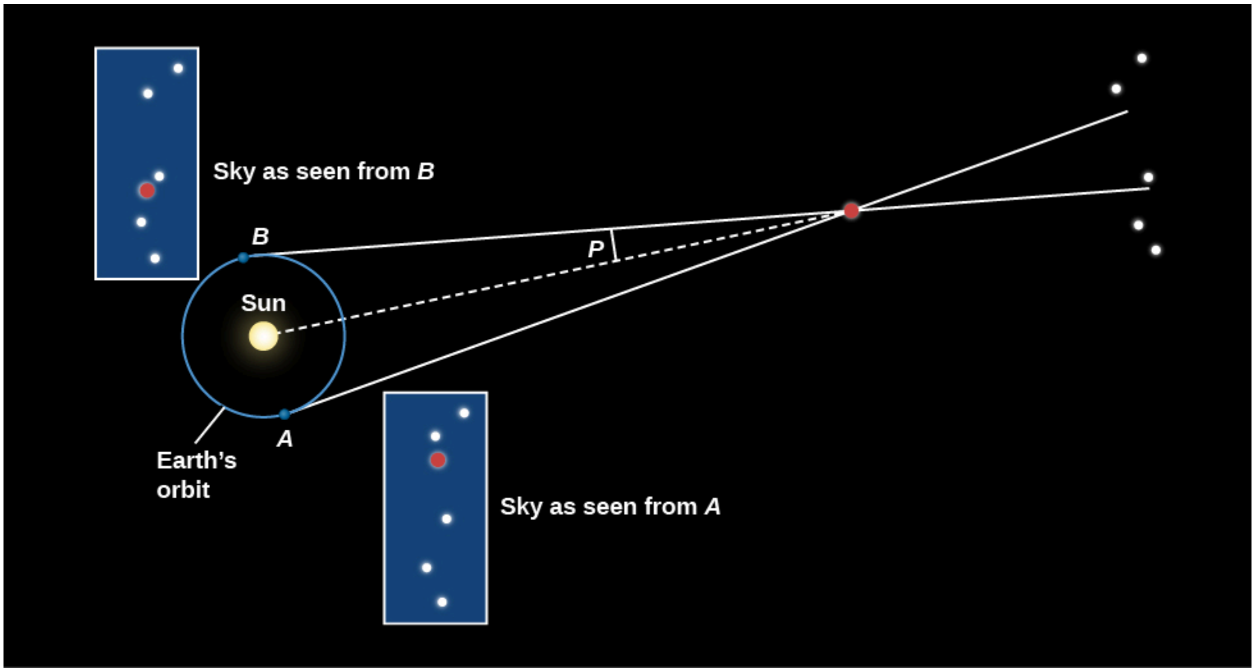


Figure 23.6 As Earth revolves around the Sun, the direction in which we see a nearby star varies with respect to distant stars. We define the parallax of the nearby star to be one half of the total change in direction, and we usually measure it in arcseconds.

Units of Stellar Distance

With a baseline of one AU, how far away would a star have to be to have a parallax of 1 arcsecond? The answer turns out to be 206,265 AU, or 3.26 light-years. This is equal to 3.1×10^{13} kilometers (in other words, 31 trillion kilometers). We give this unit a special name, the **parsec** (pc)—derived from "the distance at which we have a *par*allax of one *sec*ond." The distance (D) of a star in parsecs is just the reciprocal of its parallax (p) in arcseconds; that is,

$$D = \frac{1}{p} \tag{23.1}$$

Thus, a star with a parallax of 0.1 arcsecond would be found at a distance of 10 parsecs, and one with a parallax of 0.05 arcsecond would be 20 parsecs away.

Back in the days when most of our distances came from parallax measurements, a parsec was a useful unit of distance, but it is not as intuitive as the light-year. One advantage of the light-year as a unit is that it emphasizes the fact that, as we look out into space, we are also looking back into time. The light that we see from a star 100 light-years away left that star 100 years ago. What we study is not the star as it is now, but rather as it was in the past. The light that reaches our telescopes today from distant galaxies left them before Earth even existed.

In this text, we will use light-years as our unit of distance, but many astronomers still use parsecs when they write technical papers or talk with each other at meetings. To convert between the two distance units, just bear in mind: 1 parsec = 3.26 light-year, and 1 light-year = 0.31 parsec.

Example 23.1

How Far Is a Light-Year?

A light-year is the distance light travels in 1 year. Given that light travels at a speed of 300,000 km/s, how many kilometers are there in a light-year?

Solution

We learned earlier that speed = distance/time. We can rearrange this equation so that distance = velocity × time. Now, we need to determine the number of seconds in a year.

There are approximately 365 days in 1 year. To determine the number of seconds, we must estimate the number of seconds in 1 day.

We can change units as follows (notice how the units of time cancel out):

$$1 \text{ day} \times 24 \text{ hr/day} \times 60 \text{ min/hr} \times 60 \text{ s/min} = 86{,}400 \text{ s/day}$$

Next, to get the number of seconds per year:

$$365 \text{ days/year} \times 86{,}400 \text{ s/day} = 31{,}536{,}000 \text{ s/year}$$

Now we can multiply the speed of light by the number of seconds per year to get the distance traveled by light in 1 year:

$$\begin{aligned} \text{distance} &= \text{velocity} \times \text{time} \\ &= 300{,}000 \text{ km/s} \times 31{,}536{,}000 \text{ s} \\ &= 9.46 \times 10^{12} \text{ km} \end{aligned}$$

That's almost 10,000,000,000,000 km that light covers in a year. To help you imagine how long this distance is, we'll mention that a string 1 light-year long could fit around the circumference of Earth 236 million times.

23.1 The number above is really large. What happens if we put it in terms that might be a little more understandable, like the diameter of Earth? Earth's diameter is about 12,700 km.

Naming Stars

You may be wondering why stars have such a confusing assortment of names. Just look at the first three stars to have their parallaxes measured: 61 Cygni, Alpha Centauri, and Vega. Each of these names comes from a different tradition of designating stars.

The brightest stars have names that derive from the ancients. Some are from the Greek, such as Sirius, which means "the scorched one"—a reference to its brilliance. A few are from Latin, but many of the best-known names are from Arabic because much of Greek and Roman astronomy was "rediscovered" in Europe after the Dark Ages by means of Arabic translations. Vega, for example, means "swooping Eagle," and Betelgeuse (pronounced "Beetle-juice") means "right hand of the central one."

In 1603, German astronomer Johann Bayer (1572–1625) introduced a more systematic approach to naming stars. For each constellation, he assigned a Greek letter to the brightest stars, roughly in order of brightness. In the constellation of Orion, for example, Betelgeuse is the brightest star, so it got the first letter in the Greek alphabet—alpha—and is known as Alpha Orionis. ("Orionis" is the possessive form of Orion, so Alpha Orionis means "the first of Orion.") A star called Rigel, being the second brightest in that constellation, is called Beta Orionis (Figure 23.7). Since there are 24 letters in the Greek alphabet, this system allows the labeling of 24 stars in each constellation, but constellations have many more stars than that.

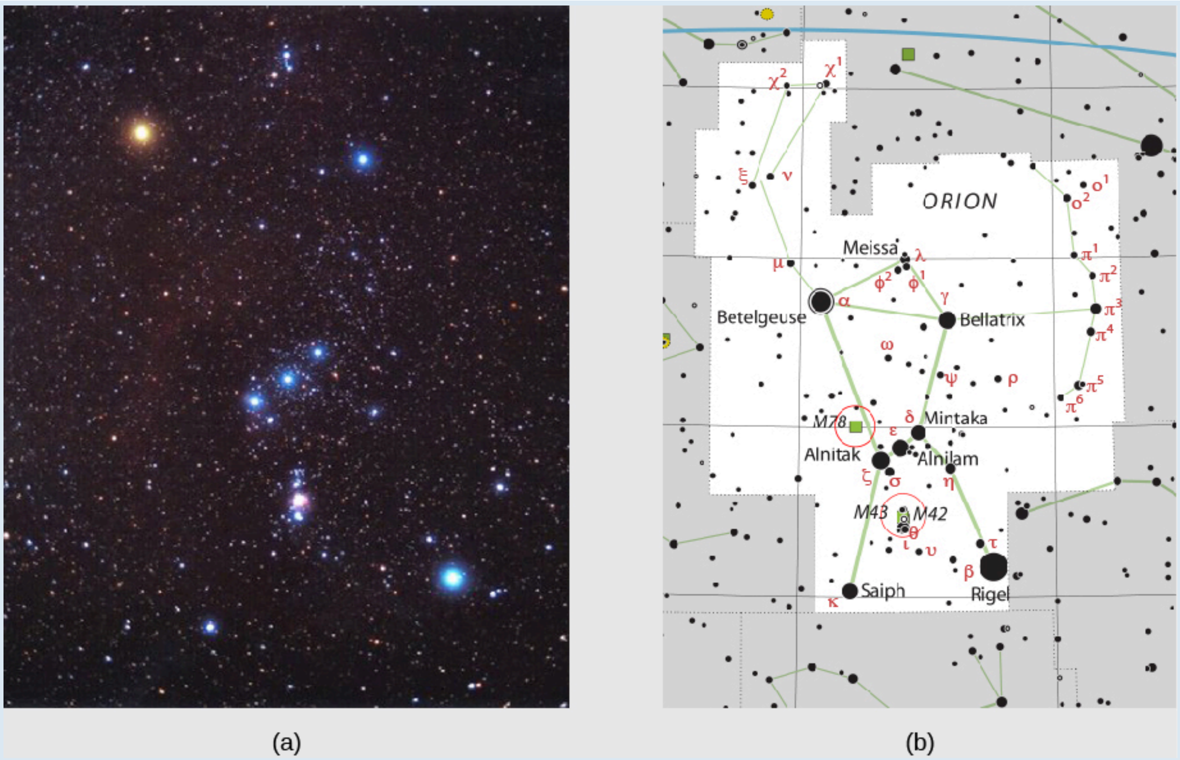


Figure 23.7 (a) This image shows the brightest objects in or near the star pattern of Orion, the hunter (of Greek mythology), in the constellation of Orion. (b) Note the Greek letters of Bayer's system in this diagram of the Orion constellation. The objects denoted M42, M43, and M78 are not stars but nebulae—clouds of gas and dust; these numbers come from a list of "fuzzy objects" made by Charles Messier in 1781. (credit a: modification of work by Matthew Spinelli; credit b: modification of work by ESO, IAU and *Sky & Telescope*)

In 1725, the English Astronomer Royal John Flamsteed introduced yet another system, in which the brighter stars eventually got a number in each constellation in order of their location in the sky or, more precisely, their right ascension. (The system of sky coordinates that includes right ascension is one in which a star's coordinates are given in such a way as to remove any dependence upon the Earth's rotation.) In this system, Betelgeuse is called 58 Orionis and 61 Cygni is the 61st star in the constellation of Cygnus, the swan.

It gets worse. As astronomers began to understand more and more about stars, they drew up a series of specialized star catalogs, and fans of those catalogs began calling stars by their catalog numbers. If you look at **Appendix D**—our list of the nearest stars (many of which are much too faint to get an ancient name, Bayer letter, or Flamsteed number)—you will see references to some of these catalogs. An example is a set of stars labeled with a BD number, for "Bonner Durchmusterung." This was a mammoth catalog of over 324,000 stars in a series of zones in the sky, organized at the Bonn Observatory in the 1850s and 1860s. Keep in mind that this catalog was made before photography or computers came into use, so the position of each star had to be measured (at least twice) by eye, a daunting undertaking.

There is also a completely different system for keeping track of stars whose luminosity varies, and another for stars that brighten explosively at unpredictable times. Astronomers have gotten used to the many different star-naming systems, but students often find them bewildering and wish astronomers would settle on one. Don't hold your breath: in astronomy, as in many fields of human thought, tradition holds a powerful attraction. Still, with high-speed computer databases to aid human memory, names may become less and less necessary. Today's astronomers often refer to stars by their precise locations in the sky rather than by their names or various catalog numbers.

The Nearest Stars

No known star (other than the Sun) is within 1 light-year or even 1 parsec of Earth. The stellar neighbors nearest the Sun are three stars in the constellation of Centaurus. To the unaided eye, the brightest of these three stars is Alpha Centauri, which is only 30° from the south celestial pole and hence not visible from the mainland United States. Alpha Centauri itself is a binary star—two stars in mutual revolution—too close together to be distinguished without a telescope. These two stars are 4.4 light-years from us. Nearby is a third faint star, known as Proxima Centauri. Proxima, with a distance of 4.3 light-years, is slightly closer to us than the other two stars. If Proxima Centauri is part of a triple star system with the binary Alpha Centauri, as seems likely, then its orbital period may be longer than 500,000 years.

Proxima Centauri is an example of the most common type of star, and our most common type of stellar neighbor (as we saw in Stars: A Celestial Census.) Low-mass red M dwarfs make up about 70% of all stars and dominate the census of stars within 10 parsecs (33 light-years) of the Sun. For example, a recent survey of the solar neighborhood counted 357 stars and brown dwarfs within 10 parsecs, and 248 of these are red dwarfs. Yet, if you wanted to see an M dwarf with your naked eye, you would be out of luck. These stars only produce a fraction of the Sun's light, and nearly all of them require a telescope to be detected.

The nearest star visible without a telescope from most of the United States is the brightest appearing of all the stars, Sirius, which has a distance of a little more than 8 light-years. It too is a binary system, composed of a faint white dwarf orbiting a bluish-white, main-sequence star. It is an interesting coincidence of numbers that light reaches us from the Sun in about 8 minutes and from the next brightest star in the sky in about 8 years.

Example 23.2

Calculating the Diameter of the Sun

For nearby stars, we can measure the apparent shift in their positions as Earth orbits the Sun. We wrote earlier that an object must be 206,265 AU distant to have a parallax of one second of arc. This must seem like a very strange number, but you can figure out why this is the right value. We will start by estimating the diameter of the Sun and then apply the same idea to a star with a parallax of 1 arcsecond. Make a sketch that has a round circle to represent the Sun, place Earth some distance away, and put an observer on it. Draw two lines from the point where the observer is standing, one to each side of the Sun. Sketch a circle centered at Earth with its circumference passing through the center of the Sun. Now think about proportions. The Sun spans about half a degree on the sky. A full circle has 360°. The circumference of the circle centered on Earth and passing through the Sun is given by:

$$\text{circumference} = 2\pi \times 93{,}000{,}000 \text{ miles}$$

Then, the following two ratios are equal:

$$\frac{0.5^\circ}{360^\circ} = \frac{\text{diameter of Sun}}{2\pi \times 93{,}000{,}000}$$

Calculate the diameter of the Sun. How does your answer compare to the actual diameter?

Solution

To solve for the diameter of the Sun, we can evaluate the expression above.

$$\begin{aligned}\text{diameter of the sun} &= \frac{0.5^\circ}{360^\circ} \times 2\pi \times 93{,}000{,}000 \text{ miles} \\ &= 811{,}577 \text{ miles}\end{aligned}$$

This is very close to the true value of about 848,000 miles.

23.2 Now apply this idea to calculating the distance to a star that has a parallax of 1 arcsec. Draw a picture similar to the one we suggested above and calculate the distance in AU. (Hint: Remember that the parallax angle is defined by 1 AU, not 2 AU, and that 3600 arcseconds = 1 degree.)

Measuring Parallaxes in Space

The measurements of stellar parallax were revolutionized by the launch of the spacecraft Hipparcos in 1989, which measured distances for thousands of stars out to about 300 light-years with an accuracy of 10 to 20% (see **Figure 23.8** and the feature on **Parallax and Space Astronomy**). However, even 300 light-years are less than 1% the size of our Galaxy's main disk.

In December 2013, the successor to Hipparcos, named *Gaia*, was launched by the European Space Agency. *Gaia* is measuring the position and distances to almost one billion stars with an accuracy of a few millionths of an arcsecond. *Gaia's* distance limit will extend well beyond Hipparcos, studying stars out to 30,000 light-years (100 times farther than Hipparcos, covering nearly 1/3 of the galactic disk). *Gaia* will also be able to measure proper motions[2] for thousands of stars in the

2. Proper motion (as discussed in Using Spectra to Measure Stellar Composition and Motion, is the motion of a star

halo of the Milky Way—something that can only be done for the brightest stars right now. At the end of *Gaia's* mission, we will not only have a three-dimensional map of a large fraction of our own Milky Way Galaxy, but we will also have a strong link in the chain of cosmic distances that we are discussing in this chapter. Yet, to extend this chain beyond *Gaia's* reach and explore distances to nearby galaxies, we need some completely new techniques.

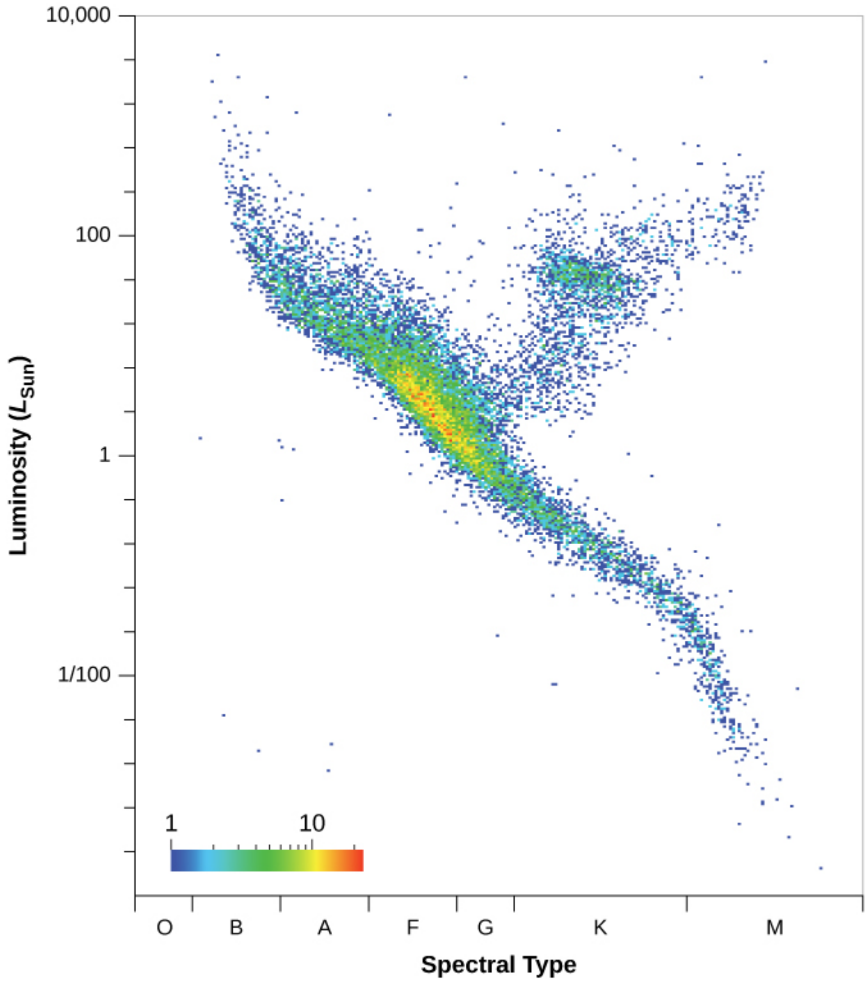


Figure 23.8 This plot includes 16,631 stars for which the parallaxes have an accuracy of 10% or better. The colors indicate the numbers of stars at each point of the diagram, with red corresponding to the largest number and blue to the lowest. Luminosity is plotted along the vertical axis, with luminosity increasing upward. An infrared color is plotted as a proxy for temperature, with temperature decreasing to the right. Most of the data points are distributed along the diagonal running from the top left corner (high luminosity, high temperature) to the bottom right (low temperature, low luminosity). These are main sequence stars. The large clump of data points above the main sequence on the right side of the diagram is composed of red giant stars. (credit: modification of work by the European Space Agency)

Parallax and Space Astronomy

One of the most difficult things about precisely measuring the tiny angles of parallax shifts from Earth is that you have to observe the stars through our planet's atmosphere. The effect of the atmosphere is to spread out the points of starlight into fuzzy disks, making exact measurements of their positions more difficult. Astronomers had long dreamed of being able to measure parallaxes from space, and two orbiting observatories have now turned this dream into reality.

The name of the Hipparcos satellite, launched in 1989 by the European Space Agency, is both an abbreviation for High Precision Parallax Collecting Satellite and a tribute to Hipparchus, the pioneering Greek astronomer. The satellite was

across the sky (perpendicular to our line of sight.)

designed to make the most accurate parallax measurements in history, from 36,000 kilometers above Earth. However, its onboard rocket motor failed to fire, which meant it did not get the needed boost to reach the desired altitude. Hipparcos ended up spending its 4-year life in an elliptical orbit that varied from 500 to 36,000 kilometers high. In this orbit, the satellite plunged into Earth's radiation belts every 5 hours or so, which finally took its toll on the solar panels that provided energy to power the instruments.

Nevertheless, the mission was successful, resulting in two catalogs. One gives positions of 120,000 stars to an accuracy of one-thousandth of an arcsecond—about the diameter of a golf ball in New York as viewed from Europe. The second catalog contains information for more than a million stars, whose positions have been measured to thirty-thousandths of an arcsecond. We now have accurate parallax measurements of stars out to distances of about 300 light-years. (With ground-based telescopes, accurate measurements were feasible out to only about 60 light-years.)

In order to build on the success of Hipparcos, in 2013, the European Space Agency launched a new satellite called *Gaia*. The Gaia mission is scheduled to last for 5 years. Because *Gaia* carries larger telescopes than Hipparcos, it can observe fainter stars and measure their positions 200 times more accurately. The main goal of the Gaia mission is to make an accurate three-dimensional map of that portion of the Galaxy within about 30,000 light-years by observing 1 billion stars 70 times each, measuring their positions and hence their parallaxes as well as their brightnesses.

For a long time, the measurement of parallaxes and accurate stellar positions was a backwater of astronomical research—mainly because the accuracy of measurements did not improve much for about 100 years. However, the ability to make measurements from space has revolutionized this field of astronomy and will continue to provide a critical link in our chain of cosmic distances.

The European Space Agency (ESA) maintains a **Gaia mission website (https://openstaxcollege.org/l/30GaiaMission)** where you can learn more about the Gaia mission and to get the latest news on *Gaia* observations.

To learn more about Hipparcos, explore this **European Space Agency webpage (https://openstaxcollege.org/l/30Hipparcos)** with an ESA vodcast *Charting the Galaxy—from Hipparcos to Gaia*.

23.3 | Variable Stars: One Key to Cosmic Distances

Learning Objectives

By the end of this section, you will be able to:

- Describe how some stars vary their light output and why such stars are important
- Explain the importance of pulsating variable stars, such as cepheids and RR Lyrae-type stars, to our study of the universe

Let's briefly review the key reasons that measuring distances to the stars is such a struggle. As discussed in **The Brightness of Stars**, our problem is that stars come in a bewildering variety of intrinsic luminosities. (If stars were light bulbs, we'd say they come in a wide range of wattages.) Suppose, instead, that all stars had the same "wattage" or luminosity. In that case, the more distant ones would always look dimmer, and we could tell how far away a star is simply by how dim it appeared. In the real universe, however, when we look at a star in our sky (with eye or telescope) and measure its apparent brightness, we cannot know whether it looks dim because it's a low-wattage bulb or because it is far away, or perhaps some of each.

Astronomers need to discover something else about the star that allows us to "read off" its intrinsic luminosity—in effect, to know what the star's true wattage is. With this information, we can then attribute how dim it looks from Earth to its distance. Recall that the apparent brightness of an object decreases with the square of the distance to that object. If two objects have the same luminosity but one is three times farther than the other, the more distant one will look nine times fainter. Therefore, if we know the luminosity of a star and its apparent brightness, we can calculate how far away it is. Astronomers have long searched for techniques that would somehow allow us to determine the luminosity of a star—and it is to these techniques that we turn next.

Variable Stars

The breakthrough in measuring distances to remote parts of our Galaxy, and to other galaxies as well, came from the study of variable stars. Most stars are constant in their luminosity, at least to within a percent or two. Like the Sun, they generate a steady flow of energy from their interiors. However, some stars are seen to vary in brightness and, for this reason, are called *variable stars*. Many such stars vary on a regular cycle, like the flashing bulbs that decorate stores and homes during the winter holidays.

Let's define some tools to help us keep track of how a star varies. A graph that shows how the brightness of a variable star changes with time is called a **light curve** (**Figure 23.9**). The *maximum* is the point of the light curve where the star has its greatest brightness; the *minimum* is the point where it is faintest. If the light variations repeat themselves periodically, the interval between the two maxima is called the *period* of the star. (If this kind of graph looks familiar, it is because we introduced it in **Diameters of Stars**.)

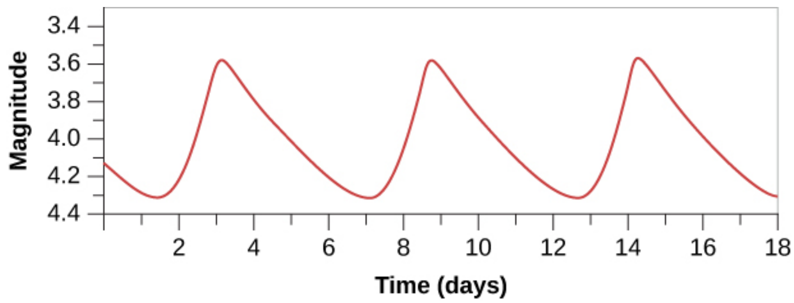


Figure 23.9 This light curve shows how the brightness changes with time for a typical cepheid variable, with a period of about 6 days.

Pulsating Variables

There are two special types of variable stars for which—as we will see—measurements of the light curve give us accurate distances. These are called **cepheid** and **RR Lyrae** variables, both of which are **pulsating variable stars**. Such a star actually changes its diameter with time—periodically expanding and contracting, as your chest does when you breathe. We now understand that these stars are going through a brief unstable stage late in their lives.

The expansion and contraction of pulsating variables can be measured by using the Doppler effect. The lines in the spectrum shift toward the blue as the surface of the star moves toward us and then shift to the red as the surface shrinks back. As the star pulsates, it also changes its overall color, indicating that its temperature is also varying. And, most important for our purposes, the luminosity of the pulsating variable also changes in a regular way as it expands and contracts.

Cepheid Variables

Cepheids are large, yellow, pulsating stars named for the first-known star of the group, Delta Cephei. This, by the way, is another example of how confusing naming conventions get in astronomy; here, a whole class of stars is named after the constellation in which the first one happened to be found. (We textbook authors can only apologize to our readers for the whole mess!)

The variability of Delta Cephei was discovered in 1784 by the young English astronomer John Goodricke (see **John Goodricke**). The star rises rather rapidly to maximum light and then falls more slowly to minimum light, taking a total of 5.4 days for one cycle. The curve in **Figure 23.9** represents a simplified version of the light curve of Delta Cephei.

Several hundred cepheid variables are known in our Galaxy. Most cepheids have periods in the range of 3 to 50 days and luminosities that are about 1000 to 10,000 times greater than that of the Sun. Their variations in luminosity range from a few percent to a factor of 10.

Polaris, the North Star, is a cepheid variable that, for a long time, varied by one tenth of a magnitude, or by about 10% in visual luminosity, in a period of just under 4 days. Recent measurements indicate that the amount by which the brightness of Polaris changes is decreasing and that, sometime in the future, this star will no longer be a pulsating variable. This is just one more piece of evidence that stars really do evolve and change in fundamental ways as they age, and that being a cepheid variable represents a stage in the life of the star.

The Period-Luminosity Relation

The importance of cepheid variables lies in the fact that their periods and average luminosities turn out to be directly related. The longer the period (the longer the star takes to vary), the greater the luminosity. This **period-luminosity relation** was a remarkable discovery, one for which astronomers still (pardon the expression) thank their lucky stars. The period of such a star is easy to measure: a good telescope and a good clock are all you need. Once you have the period, the relationship (which can be put into precise mathematical terms) will give you the luminosity of the star.

Let's be clear on what that means. The relation allows you to essentially "read off" how bright the star really is (how much energy it puts out). Astronomers can then compare this intrinsic brightness with the apparent brightness of the star. As we saw, the difference between the two allows them to calculate the distance.

The relation between period and luminosity was discovered in 1908 by Henrietta Leavitt (**Figure 23.10**), a staff member at the Harvard College Observatory (and one of a number of women working for low wages assisting Edward Pickering, the observatory's director; see **Annie Cannon: Classifier of the Stars**). Leavitt discovered hundreds of variable stars in the Large Magellanic Cloud and Small Magellanic Cloud, two great star systems that are actually neighboring galaxies (although they were not known to be galaxies then). A small fraction of these variables were cepheids (**Figure 23.11**).

Figure 23.10 Leavitt worked as an astronomer at the Harvard College Observatory. While studying photographs of the Magellanic Clouds, she found over 1700 variable stars, including 20 cepheids. Since all the cepheids in these systems were at roughly the same distance, she was able to compare their luminosities and periods of variation. She thus discovered a fundamental relationship between these characteristics that led to a new and much better way of estimating cosmic distances. (credit: modification of work by AIP)

These systems presented a wonderful opportunity to study the behavior of variable stars independent of their distance. For all practical purposes, the Magellanic Clouds are so far away that astronomers can assume that all the stars in them are at roughly the same distance from us. (In the same way, all the suburbs of Los Angeles are roughly the same distance from New York City. Of course, if you are *in* Los Angeles, you will notice annoying distances between the suburbs, but compared to how far away New York City is, the differences seem small.) If all the variable stars in the Magellanic Clouds are at roughly the same distance, then any difference in their apparent brightnesses must be caused by differences in their intrinsic luminosities.

Figure 23.11 The Large Magellanic Cloud (so named because Magellan's crew were the first Europeans to record it) is a small, irregularly shaped galaxy near our own Milky Way. It was in this galaxy that Henrietta Leavitt discovered the cepheid period-luminosity relation. (credit: ESO)

Leavitt found that the brighter-appearing cepheids always have the longer periods of light variation. Thus, she reasoned, the period must be related to the luminosity of the stars. When Leavitt did this work, the distance to the Magellanic Clouds was not known, so she was only able to show that luminosity was related to period. She could not determine exactly what the relationship is.

To define the period-luminosity relation with actual numbers (to *calibrate* it), astronomers first had to measure the actual distances to a few nearby cepheids in another way. (This was accomplished by finding cepheids associated in clusters with other stars whose distances could be estimated from their spectra, as discussed in the next section of this chapter.) But once the relation was thus defined, it could give us the distance to any cepheid, wherever it might be located (**Figure 23.12**).

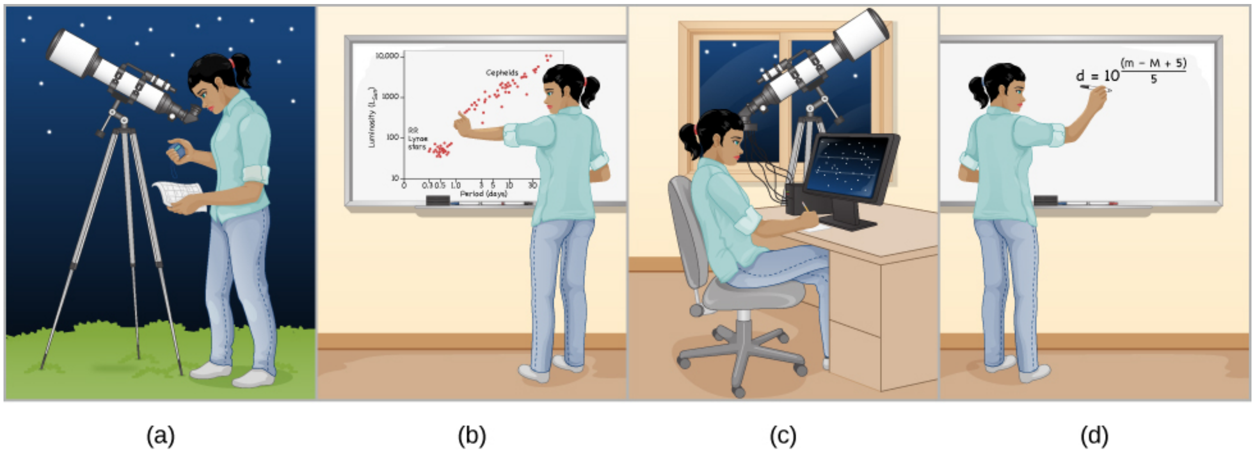


Figure 23.12 (a) Find a cepheid variable star and measure its period. (b) Use the period-luminosity relation to calculate the star's luminosity. (c) Measure the star's apparent brightness. (d) Compare the luminosity with the apparent brightness to calculate the distance.

Here at last was the technique astronomers had been searching for to break the confines of distance that parallax imposed on them. Cepheids can be observed and monitored, it turns out, in many parts of our own Galaxy and in other nearby galaxies as well. Astronomers, including Ejnar Hertzsprung and Harvard's Harlow Shapley, immediately saw the potential of the new technique; they and many others set to work exploring more distant reaches of space using cepheids as signposts. In the 1920s, Edwin Hubble made one of the most significant astronomical discoveries of all time using cepheids, when he observed them in nearby galaxies and discovered the expansion of the universe. As we will see, this work still continues, as the Hubble Space Telescope and other modern instruments try to identify and measure individual cepheids in galaxies farther and farther away. The most distant known variable stars are all cepheids, with some about 60 million light-years away.

John Goodricke

The brief life of John Goodricke (**Figure 23.13**) is a testament to the human spirit under adversity. Born deaf and unable to speak, Goodricke nevertheless made a number of pioneering discoveries in astronomy through patient and careful observations of the heavens.

Figure 23.13 This portrait of Goodricke by artist J. Scouler hangs in the Royal Astronomical Society in London. There is some controversy about whether this is actually what Goodricke looked like or whether the painting was much retouched to please his family. (credit: James Scouler)

Born in Holland, where his father was on a diplomatic mission, Goodricke was sent back to England at age eight to study at a special school for the deaf. He did sufficiently well to enter Warrington Academy, a secondary school that offered no special assistance for students with handicaps. His mathematics teacher there inspired an interest in astronomy, and in 1781, at age 17, Goodricke began observing the sky at his family home in York, England. Within a year, he had discovered the brightness variations of the star Algol (discussed in **Stellar Properties**) and suggested that an unseen companion star was causing the changes, a theory that waited over 100 years for proof. His paper on the subject was read before the Royal Society (the main British group of scientists) in 1783 and won him a medal from that distinguished group.

In the meantime, Goodricke had discovered two other stars that varied regularly, Beta Lyrae and Delta Cephei, both of which continued to interest astronomers for years to come. Goodricke shared his interest in observing with his older cousin, Edward Pigott, who went on to discover other variable stars during his much longer life. But Goodricke's time was quickly drawing to a close; at age 21, only 2 weeks after he was elected to the Royal Society, he caught a cold while making astronomical observations and never recovered.

Today, the University of York has a building named Goodricke Hall and a plaque that honors his contributions to science. Yet if you go to the churchyard cemetery where he is buried, an overgrown tombstone has only the initials "J. G." to show where he lies. Astronomer Zdenek Kopal, who looked carefully into Goodricke's life, speculated on why the marker is so modest: perhaps the rather staid Goodricke relatives were ashamed of having a "deaf-mute" in the family and could not sufficiently appreciate how much a man who could not hear could nevertheless see.

RR Lyrae Stars

A related group of stars, whose nature was understood somewhat later than that of the cepheids, are called RR Lyrae variables, named for the star RR Lyrae, the best-known member of the group. More common than the cepheids, but less luminous, thousands of these pulsating variables are known in our Galaxy. The periods of RR Lyrae stars are always less than 1 day, and their changes in brightness are typically less than about a factor of two.

Astronomers have observed that the RR Lyrae stars occurring in any particular cluster all have about the same apparent brightness. Since stars in a cluster are all at approximately the same distance, it follows that RR Lyrae variables must all have nearly the same intrinsic luminosity, which turns out to be about 50 L_{Sun}. In this sense, RR Lyrae stars are a little bit like standard light bulbs and can also be used to obtain distances, particularly within our Galaxy. **Figure 23.14** displays the ranges of periods and luminosities for both the cepheids and the RR Lyrae stars.

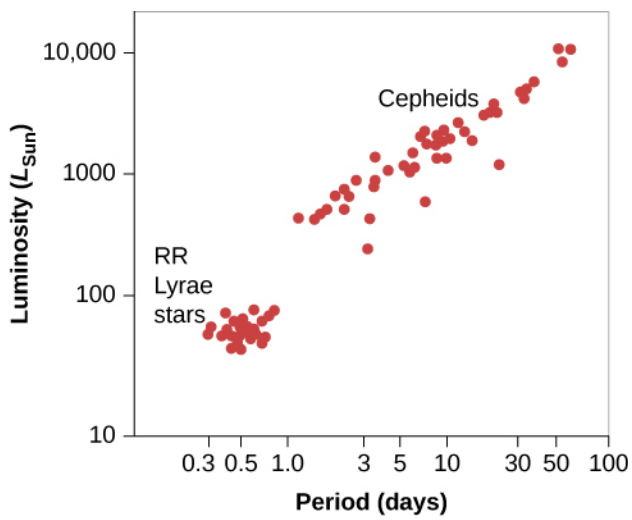


Figure 23.14 In this class of variable stars, the time the star takes to go through a cycle of luminosity changes is related to the average luminosity of the star. Also shown are the period and luminosity for RR Lyrae stars.

23.4 | The H–R Diagram and Cosmic Distances

Learning Objectives

By the end of this section, you will be able to:

- Understand how spectral types are used to estimate stellar luminosities
- Examine how these techniques are used by astronomers today

Variable stars are not the only way that we can estimate the luminosity of stars. Another way involves the H–R diagram, which shows that the intrinsic brightness of a star can be estimated if we know its spectral type.

Distances from Spectral Types

As satisfying and productive as variable stars have been for distance measurement, these stars are rare and are not found near all the objects to which we wish to measure distances. Suppose, for example, we need the distance to a star that is not varying, or to a group of stars, none of which is a variable. In this case, it turns out the H–R diagram can come to our rescue.

If we can observe the spectrum of a star, we can estimate its distance from our understanding of the H–R diagram. As discussed in **The Spectra of Stars**, a detailed examination of a stellar spectrum allows astronomers to classify the star into one of the *spectral types* indicating surface temperature. (The types are O, B, A, F, G, K, M, L, T, and Y; each of these can be divided into numbered subgroups.) In general, however, the spectral type alone is not enough to allow us to estimate luminosity. Look again at **Figure 22.29**. A G2 star could be a main-sequence star with a luminosity of 1 L_{Sun}, or it could be a giant with a luminosity of 100 L_{Sun}, or even a supergiant with a still higher luminosity.

We can learn more from a star's spectrum, however, than just its temperature. Remember, for example, that we can detect pressure differences in stars from the details of the spectrum. This knowledge is very useful because giant stars are larger (and have lower pressures) than main-sequence stars, and supergiants are still larger than giants. If we look in detail at the spectrum of a star, we can determine whether it is a main-sequence star, a giant, or a supergiant.

Suppose, to start with the simplest example, that the spectrum, color, and other properties of a distant G2 star match those of the Sun exactly. It is then reasonable to conclude that this distant star is likely to be a main-sequence star just like the Sun and to have the same luminosity as the Sun. But if there are subtle differences between the solar spectrum and the spectrum of the distant star, then the distant star may be a giant or even a supergiant.

The most widely used system of star classification divides stars of a given spectral class into six categories called **luminosity classes**. These luminosity classes are denoted by Roman numbers as follows:

- Ia: Brightest supergiants
- Ib: Less luminous supergiants
- II: Bright giants
- III: Giants
- IV: Subgiants (intermediate between giants and main-sequence stars)
- V: Main-sequence stars

The full spectral specification of a star includes its luminosity class. For example, a main-sequence star with spectral class F3 is written as F3 V. The specification for an M2 giant is M2 III. **Figure 23.15** illustrates the approximate position of stars of various luminosity classes on the H–R diagram. The dashed portions of the lines represent regions with very few or no stars.

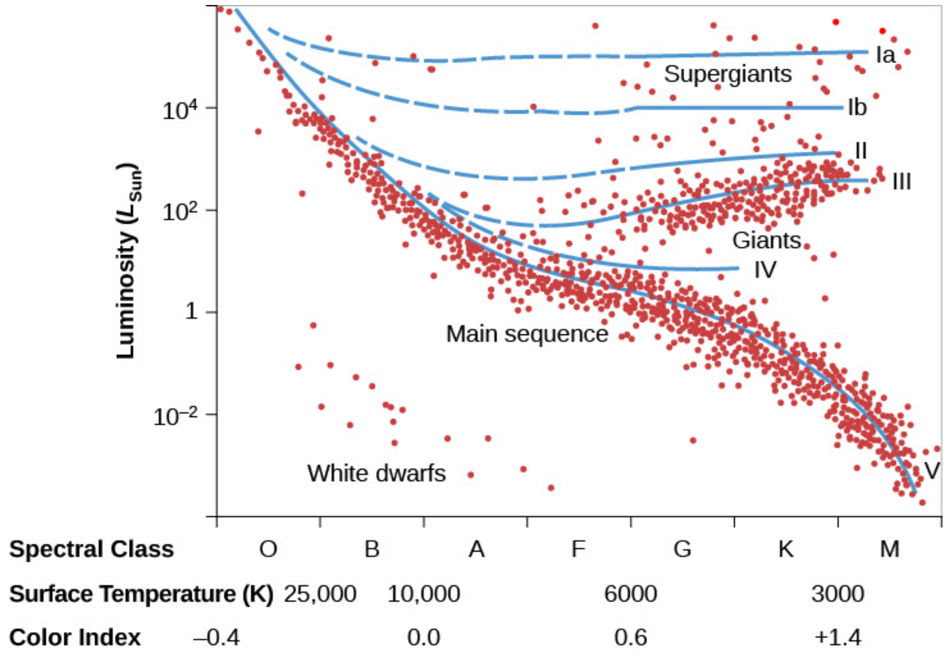


Figure 23.15 Stars of the same temperature (or spectral class) can fall into different luminosity classes on the Hertzsprung-Russell diagram. By studying details of the spectrum for each star, astronomers can determine which luminosity class they fall in (whether they are main-sequence stars, giant stars, or supergiant stars).

With both its spectral and luminosity classes known, a star's position on the H–R diagram is uniquely determined. Since the diagram plots luminosity versus temperature, this means we can now read off the star's luminosity (once its spectrum has helped us place it on the diagram). As before, if we know how luminous the star really is and see how dim it looks, the difference allows us to calculate its distance. (For historical reasons, astronomers sometimes call this method of distance determination *spectroscopic parallax*, even though the method has nothing to do with parallax.)

The H–R diagram method allows astronomers to estimate distances to nearby stars, as well as some of the most distant stars in our Galaxy, but it is anchored by measurements of parallax. The distances measured using parallax are the gold standard for distances: they rely on no assumptions, only geometry. Once astronomers take a spectrum of a nearby star for which we also know the parallax, we know the luminosity that corresponds to that spectral type. Nearby stars thus serve as benchmarks for more distant stars because we can assume that two stars with identical spectra have the same intrinsic luminosity.

Spectroscopic Parallax

Returning to the concepts of **absolute magnitude** and **apparent magnitude** that were introduced in the section on **The Brightness of Stars**, we can state a quantitative relationship that allows us to calculate the distance (in parsecs) to any object provided that we know the values of both those quantities. This relationship is known as the **spectroscopic parallax** formula:

$$d = 10 \times 10^{0.2(m - M)} \tag{23.2}$$

While we will not prove or derive this relationship, it is easy to see for the case where the absolute and apparent magnitudes of a star are identical $(M = m)$, that the distance would be $10 \times 10^0 = 10$ parsec. This is, of course, consistent with the definition of **absolute magnitude**.

Quite often the H-R diagram can be used to estimate the absolute magnitude, *M*, of a star, so a direct measurement of its apparent magnitude, *m* provides a determination of its distance from Earth.

23.3 As a numerical example, let's use the star Spica, whose absolute magnitude $M = -3.6$ and whose apparent magnitude $m = 0.9$. How far is Spica from Earth?

A Few Words about the Real World

Introductory textbooks such as ours work hard to present the material in a straightforward and simplified way. In doing so, we sometimes do our students a disservice by making scientific techniques seem too clean and painless. In the real world, the techniques we have just described turn out to be messy and difficult, and often give astronomers headaches that last long into the day.

For example, the relationships we have described such as the period-luminosity relation for certain variable stars aren't exactly straight lines on a graph. The points representing many stars scatter widely when plotted, and thus, the distances derived from them also have a certain built-in scatter or uncertainty.

The distances we measure with the methods we have discussed are therefore only accurate to within a certain percentage of error—sometimes 10%, sometimes 25%, sometimes as much as 50% or more. A 25% error for a star estimated to be 10,000 light-years away means it could be anywhere from 7500 to 12,500 light-years away. This would be an unacceptable uncertainty if you were loading fuel into a spaceship for a trip to the star, but it is not a bad first figure to work with if you are an astronomer stuck on planet Earth.

Nor is the construction of H–R diagrams as easy as you might think at first. To make a good diagram, one needs to measure the characteristics and distances of many stars, which can be a time-consuming task. Since our own solar neighborhood is already well mapped, the stars astronomers most want to study to advance our knowledge are likely to be far away and faint. It may take hours of observing to obtain a single spectrum. Observers may have to spend many nights at the telescope (and many days back home working with their data) before they get their distance measurement. Fortunately, this is changing because surveys like Gaia will study billions of stars, producing public datasets that all astronomers can use.

Despite these difficulties, the tools we have been discussing allow us to measure a remarkable range of distances—parallaxes for the nearest stars, RR Lyrae variable stars; the H–R diagram for clusters of stars in our own and nearby galaxies; and cepheids out to distances of 60 million light-years. **Table 23.1** describes the distance limits and overlap of each method.

Each technique described in this chapter builds on at least one other method, forming what many call the *cosmic distance ladder*. Parallaxes are the foundation of all stellar distance estimates, spectroscopic methods use nearby stars to calibrate their H–R diagrams, and RR Lyrae and cepheid distance estimates are grounded in H–R diagram distance estimates (and even in a parallax measurement to a nearby cepheid, Delta Cephei).

This chain of methods allows astronomers to push the limits when looking for even more distant stars. Recent work, for example, has used RR Lyrae stars to identify dim companion galaxies to our own Milky Way out at distances of 300,000 light-years. The H–R diagram method was recently used to identify the two most distant stars in the Galaxy: red giant stars way out in the halo of the Milky Way with distances of almost 1 million light-years.

We can combine the distances we find for stars with measurements of their composition, luminosity, and temperature—made with the techniques described in the chapter on **Stellar Properties**. Together, these make up the arsenal of information we need to trace the evolution of stars from birth to death, the subject to which we turn in the chapters that follow.

Distance Range of Celestial Measurement Methods

Method	Distance Range
Trigonometric parallax	4–30,000 light-years when the Gaia mission is complete
RR Lyrae stars	Out to 300,000 light-years
H–R diagram and spectroscopic distances	Out to 1,200,000 light-years
Cepheid stars	Out to 60,000,000 light-years

Table 23.1

For Further Exploration

Websites

ABCs of Distance: **http://www.astro.ucla.edu/~wright/distance.htm (http://www.astro.ucla.edu/~wright/distance.htm)** . Astronomer Ned Wright (UCLA) gives a concise primer on many different methods of obtaining distances. This site is at a higher level than our textbook, but is an excellent review for those with some background in astronomy.

American Association of Variable Star Observers (AAVSO): **https://www.aavso.org/ (https://www.aavso.org/)** . This organization of amateur astronomers helps to keep track of variable stars; its site has some background material, observing instructions, and links.

Friedrich Wilhelm Bessel: **http://messier.seds.org/xtra/Bios/bessel.html (http://messier.seds.org/xtra/Bios/bessel.html)** . A brief site about the first person to detect stellar parallax, with references and links.

Gaia: **http://sci.esa.int/gaia/ (http://sci.esa.int/gaia/)** . News from the Gaia mission, including images and a blog of the latest findings.

Hipparchos: **http://sci.esa.int/hipparcos/ (http://sci.esa.int/hipparcos/)** . Background, results, catalogs of data, and educational resources from the Hipparchos mission to observe parallaxes from space. Some sections are technical, but others are accessible to students.

John Goodricke: The Deaf Astronomer: **http://www.bbc.com/news/magazine-20725639 (http://www.bbc.com/news/magazine-20725639)** . A biographical article from the BBC.

Women in Astronomy: **http://www.astrosociety.org/education/astronomy-resource-guides/women-in-astronomy-an-introductory-resource-guide/ (http://www.astrosociety.org/education/astronomy-resource-guides/women-in-astronomy-an-introductory-resource-guide/)** . More about Henrietta Leavitt's and other women's contributions to astronomy and the obstacles they faced.

Videos

Gaia's Mission: Solving the Celestial Puzzle: **https://www.youtube.com/watch?v=oGri4YNggoc (https://www.youtube.com/watch?v=oGri4YNggoc)** . Describes the Gaia mission and what scientists hope to learn, from Cambridge University (19:58).

Hipparcos: Route Map to the Stars: **http://www.esa.int/spaceinvideos/Videos/1997/05/Hipparcos_Route_Maps_to_the_Stars_May_97 (http://www.esa.int/spaceinvideos/Videos/1997/05/Hipparcos_Route_Maps_to_the_Stars_May_97)** . This ESA video describes the mission to measure parallax and its results (14:32)

How Big Is the Universe: **https://www.youtube.com/watch?v=K_xZuopg4Sk (https://www.youtube.com/watch?v=K_xZuopg4Sk)** . Astronomer Pete Edwards from the British Institute of Physics discusses the size of the universe and gives a step-by-step introduction to the concepts of distances (6:22)

Search for Miss Leavitt: **http://perimeterinstitute.ca/videos/search-miss-leavitt (http://perimeterinstitute.ca/videos/search-miss-leavitt)** . Video of talk by George Johnson on his search for Miss Leavitt (55:09).

Women in Astronomy: **http://www.youtube.com/watch?v=5vMR7su4fi8 (http://www.youtube.com/watch?v=5vMR7su4fi8)** . Emily Rice (CUNY) gives a talk on the contributions of women to astronomy, with many historical and contemporary examples, and an analysis of modern trends (52:54).

CHAPTER 23 REVIEW

KEY TERMS

cepheid a star that belongs to a class of yellow supergiant pulsating stars; these stars vary periodically in brightness, and the relationship between their periods and luminosities is useful in deriving distances to them

light curve a graph that displays the time variation of the light from a variable or eclipsing binary star or, more generally, from any other object whose radiation output changes with time

luminosity class a classification of a star according to its luminosity within a given spectral class; our Sun, a G2V star, has luminosity class V, for example

parallax an apparent displacement of a nearby star that results from the motion of Earth around the Sun

parsec a unit of distance in astronomy, equal to 3.26 light-years; at a distance of 1 parsec, a star has a parallax of 1 arcsecond

period-luminosity relation an empirical relation between the periods and luminosities of certain variable stars

pulsating variable star a variable star that pulsates in size and luminosity

RR Lyrae one of a class of giant pulsating stars with periods shorter than 1 day, useful for finding distances

spectroscopic parallax using both the apparent magnitude and the absolute magnitude (usually obtained from an analysis of its spectrum) of a star to calculate its distance from Earth

KEY EQUATIONS

Stellar parallax	$D = \frac{1}{p}$
Spectroscopic Parallax	$d = 10 \times 10^{0.2(m - M)}$ [pc]

SUMMARY

23.1 Fundamental Units of Distance

- Early measurements of length were based on human dimensions, but today, we use worldwide standards that specify lengths in units such as the meter.
- Distances within the solar system are now determined by timing how long it takes radar signals to travel from Earth to the surface of a planet or other body and then return.

23.2 Surveying the Stars

- For stars that are relatively nearby, we can “triangulate” the distances from a baseline created by Earth’s annual motion around the Sun.
- Half the shift in a nearby star’s position relative to very distant background stars, as viewed from opposite sides of Earth’s orbit, is called the parallax of that star and is a measure of its distance.
- The units used to measure stellar distance are the light-year, the distance light travels in 1 year, and the parsec (pc), the distance of a star with a parallax of 1 arcsecond (1 parsec = 3.26 light-years).
- The closest star, a red dwarf, is over 1 parsec away.
- The first successful measurements of stellar parallaxes were reported in 1838.
- Parallax measurements are a fundamental link in the chain of cosmic distances.
- The Hipparcos satellite has allowed us to measure accurate parallaxes for stars out to about 300 light-years, and the Gaia mission will result in parallaxes out to 30,000 light-years.

23.3 Variable Stars: One Key to Cosmic Distances

- Cepheids and RR Lyrae stars are two types of pulsating variable stars.
- Light curves of these stars show that their luminosities vary with a regularly repeating period.
- RR Lyrae stars can be used as standard bulbs, and cepheid variables obey a period-luminosity relation, so measuring their periods can tell us their luminosities.
- Then, we can calculate their distances by comparing their luminosities with their apparent brightnesses, and this can allow us to measure distances to these stars out to over 60 million light-years.

23.4 The H–R Diagram and Cosmic Distances

- Stars with identical temperatures but different pressures (and diameters) have somewhat different spectra. Spectral classification can therefore be used to estimate the luminosity class of a star as well as its temperature.
- As a result, a spectrum can allow us to pinpoint where the star is located on an H–R diagram and establish its luminosity.
- This, with the star's apparent brightness, again yields its distance.
- The various distance methods can be used to check one against another and thus make a kind of distance ladder which allows us to find even larger distances.

CONCEPTUAL QUESTIONS

23.1 Fundamental Units of Distance

1. The meter was redefined as a reference to Earth, then to krypton, and finally to the speed of light. Why do you think the reference point for a meter continued to change?

2. While a meter is the fundamental unit of length, most distances traveled by humans are measured in miles or kilometers. Why do you think this is?

3. Most distances in the Galaxy are measured in light-years instead of meters. Why do you think this is the case?

4. The AU is defined as the *average* distance between Earth and the Sun, not the distance between Earth and the Sun. Why does this need to be the case?

23.2 Surveying the Stars

5. Explain how parallax measurements can be used to determine distances to stars. Why can we not make accurate measurements of parallax beyond a certain distance?

6. What would be the advantage of making parallax measurements from Pluto rather than from Earth? Would there be a disadvantage?

7. Parallaxes are measured in fractions of an arcsecond. One arcsecond equals 1/60 arcmin; an arcminute is, in turn, 1/60th of a degree (°). To get some idea of how big 1° is, go outside at night and find the Big Dipper. The two pointer stars at the ends of the bowl are 5.5° apart. The two stars across the top of the bowl are 10° apart. (Ten degrees is also about the width of your fist when held at arm's length and projected against the sky.) Mizar, the second star from the end of the Big Dipper's handle, appears double. The fainter star, Alcor, is about 12 arcmin from Mizar. For comparison, the diameter of the full moon is about 30 arcmin. The belt of Orion is about 3° long. Keeping all this in mind, why did it take until 1838 to make parallax measurements for even the nearest stars?

8. For centuries, astronomers wondered whether comets were true celestial objects, like the planets and stars, or a phenomenon that occurred in the atmosphere of Earth. Describe an experiment to determine which of these two possibilities is correct.

9. The Sun is much closer to Earth than are the nearest stars, yet it is not possible to measure accurately the diurnal parallax of the Sun relative to the stars by measuring its position relative to background objects in the sky directly. Explain why.

10. Parallaxes of stars are sometimes measured relative to the positions of galaxies or distant objects called quasars. Why is this a good technique?

11. What is the advantage of measuring a parallax distance to a star as compared to our other distance measuring methods?

12. What is the disadvantage of the parallax method, especially for studying distant parts of the Galaxy?

23.3 Variable Stars: One Key to Cosmic Distances

13. Suppose you have discovered a new cepheid variable star. What steps would you take to determine its distance?

14. Why would it be easier to measure the characteristics of intrinsically less luminous cepheids than more luminous ones?

15. When Henrietta Leavitt discovered the period-luminosity relationship, she used cepheid stars that were all located in the Small Magellanic Cloud. Why did she need to use stars in another galaxy and not cepheids located in the Milky Way?

16. **Figure 23.9** is the light curve for the prototype cepheid variable Delta Cephei. How does the luminosity of this star compare with that of the Sun?

23.4 The H–R Diagram and Cosmic Distances

17. Explain how you would use the spectrum of a star to estimate its distance.

18. Which method would you use to obtain the distance to each of the following?

A. An asteroid crossing Earth's orbit
B. A star astronomers believe to be no more than 50 light-years from the Sun
C. A tight group of stars in the Milky Way Galaxy that includes a significant number of variable stars
D. A star that is not variable but for which you can obtain a clearly defined spectrum

19. What are the luminosity class and spectral type of a star with an effective temperature of 5000 K and a luminosity of 100 L_{Sun}?

20. Luhman 16 and WISE 0720 are brown dwarfs, also known as failed stars, and are some of the new closest neighbors to Earth, but were only discovered in the last decade. Why do you think they took so long to be discovered?

21. Most stars close to the Sun are red dwarfs. What does this tell us about the average star formation event in our Galaxy?

22. Estimating the luminosity class of an M star is much more important than measuring it for an O star if you are determining the distance to that star. Why is that the case?

23. Which of the following can you determine about a star without knowing its distance, and which can you not determine: radial velocity, temperature, apparent brightness, or luminosity? Explain.

24. A G2 star has a luminosity 100 times that of the Sun. What kind of star is it? How does its radius compare with that of the Sun?

25. A star has a temperature of 10,000 K and a luminosity of 10^{-2} L_{Sun}. What kind of star is it?

26. Referring to **Table 13.2**, which of the stars listed is closest to Earth?

Which is farthest from Earth?

PROBLEMS

23.1 Fundamental Units of Distance

27. A radar astronomer who is new at the job claims she beamed radio waves to Jupiter and received an echo exactly 48 min later. Should you believe her? Why or why not?

28. The New Horizons probe flew past Pluto in July 2015. At the time, Pluto was about 32 AU from Earth. How long did it take for communication from the probe to reach Earth, given that the speed of light in km/hr is 1.08×10^9?

29. Estimate the maximum and minimum time it takes a radar signal to make the round trip between Earth and Venus, which has a semimajor axis of 0.72 AU.

30. The Apollo program (not the lunar missions with astronauts) being conducted at the Apache Point Observatory uses a 3.5-m telescope to direct lasers at retro-reflectors left on the Moon by the Apollo astronauts. If the Moon is 384,472 km away, approximately how long do the operators need to wait to see the laser light return to Earth?

31. In 1974, the Arecibo Radio telescope in Puerto Rico was used to transmit a signal to M13, a star cluster about 25,000 light-years away. How long will it take the message to reach M13, and how far has the message travelled so far (in light-years)?

23.2 Surveying the Stars

32. Demonstrate that 1 pc equals 3.09×10^{13} km and that it also equals 3.26 light-years. Show your calculations.

33. The best parallaxes obtained with Hipparcos have an accuracy of 0.001 arcsec. If you want to measure the distance to a star with an accuracy of 10%, its parallax must be 10 times larger than the typical error. How far away can you obtain a distance that is accurate to 10% with Hipparcos data? The disk of our Galaxy is 100,000 light-years in diameter. What fraction of the diameter of the Galaxy's disk is the distance for which we can measure accurate parallaxes?

34. Astronomers are always making comparisons between measurements in astronomy and something that might be more familiar. For example, the Hipparcos web pages tell us that the measurement accuracy of 0.001 arcsec is equivalent to the angle made by a golf ball viewed from across the Atlantic Ocean, or to the angle made by the height of a person on the Moon as viewed from Earth, or to the length of growth of a human hair in 10 sec as seen from 10 meters away. Use the ideas in **Example 23.2** to verify one of the first two comparisons.

35. *Gaia* will have greatly improved precision over the measurements of Hipparcos. The average uncertainty for most *Gaia* parallaxes will be about 50 microarcsec, or 0.00005 arcsec. How many times better than Hipparcos (see **Exercise 23.33**) is this precision?

36. Using the same techniques as used in **Exercise 23.33**, how far away can *Gaia* be used to measure distances with an uncertainty of 10%? What fraction of the Galactic disk does this correspond to?

37. The human eye is capable of an angular resolution of about one arcminute, and the average distance between eyes is approximately 2 in. If you blinked and saw something move about one arcmin across, how far away from you is it? (Hint: You can use the setup in **Example 23.2** as a guide.)

38. How much better is the resolution of the *Gaia* spacecraft compared to the human eye (which can resolve about 1 arcmin)?

39. The most recently discovered system close to Earth is a pair of brown dwarfs known as Luhman 16. It has a distance of 6.5 light-years. How many parsecs is this?

40. What would the parallax of Luhman 16 (see **Exercise 23.39**) be as measured from Earth?

41. The New Horizons probe that passed by Pluto during July 2015 is one of the fastest spacecraft ever assembled. It was moving at about 14 km/s when it went by Pluto. If it maintained this speed, how long would it take New Horizons to reach the nearest star, Proxima Centauri, which is about 4.3 light-years away? (Note: It isn't headed in that direction, but you can pretend that it is.)

23.4 The H–R Diagram and Cosmic Distances

42. What physical properties are different for an M giant with a luminosity of 1000 L_{Sun} and an M dwarf with a luminosity of 0.5 L_{Sun}? What physical properties are the same?

43. The brightest star in our sky, Sirius, has an apparent magnitude of -1.4. Its absolute magnitude is +1.4. Using these numbers, how far away is Sirius from Earth?

24 | STELLAR LIFE CYCLES

Figure 24.1 During the later phases of stellar evolution, stars expel some of their mass, which returns to the interstellar medium to form new stars. This Hubble Space Telescope image shows a star losing mass. Known as Menzel 3, or the Ant Nebula, this beautiful region of expelled gas is about 3000 light-years away from the Sun. We see a central star that has ejected mass preferentially in two opposite directions. The object is about 1.6 light-years long. The image is color coded—red corresponds to an emission line of sulfur, green to nitrogen, blue to hydrogen, and blue/violet to oxygen. (credit: modification of work by NASA, ESA and The Hubble Heritage Team (STScI/AURA))

Chapter Outline

Introduction

The Sun and other stars cannot last forever. Eventually they will exhaust their nuclear fuel and cease to shine. But how do they change during their long lifetimes? And what do these changes mean for the future of Earth?

We now turn to the life cycles of stars - from their birth, to the rest of their life stories, to their eventual death. This is not an easy task since stars live much longer than astronomers. Thus, we cannot hope to see the life story of any single star unfold before our eyes or telescopes. To learn about their lives, we must survey as many of the stellar inhabitants of the Galaxy as possible. With thoroughness and a little luck, we can catch at least a few of them in each stage of their lives. As you've learned, stars have many different characteristics, with the differences sometimes resulting from their different masses, temperatures, and luminosities, and at other times derived from changes that occur as they age. Through a combination of observation and theory, we can use these differences to piece together the life story of a star.

24.1 | Star Formation

Learning Objectives

By the end of this section, you will be able to:

- Identify the sometimes-violent processes by which parts of a molecular cloud collapse to produce stars
- Recognize some of the structures seen in images of molecular clouds like the one in Orion
- Explain how the environment of a molecular cloud enables the formation of stars
- Describe how advancing waves of star formation cause a molecular cloud to evolve

As we begin our exploration of how stars are formed, let's review some basics about stars discussed in earlier chapters:

- Stable (main-sequence) stars such as our Sun maintain equilibrium by producing energy through nuclear fusion in their cores. The ability to generate energy by fusion defines a star.
- Each second in the Sun, approximately 600 million tons of hydrogen undergo fusion into helium, with about 4 million tons turning into energy in the process. This rate of hydrogen use means that eventually the Sun (and all other stars) will run out of central fuel.
- Stars come with many different masses, ranging from 1/12 solar masses (M_{Sun}) to roughly 100–200 M_{Sun}. There are far more low-mass than high-mass stars.
- The most massive main-sequence stars (spectral type O) are also the most luminous and have the highest surface temperature. The lowest-mass stars on the main sequence (spectral type M or L) are the least luminous and the coolest.
- A galaxy of stars such as the Milky Way contains enormous amounts of gas and dust—enough to make billions of stars like the Sun.

If we want to find stars still in the process of formation, we must look in places that have plenty of the raw material from which stars are assembled. Since stars are made of gas, we focus our attention (and our telescopes) on the dense and cold clouds of gas and dust that dot the Milky Way (see **Figure 24.2**).

Figure 24.2 (a) This Hubble Space Telescope image of the central regions of M16 (also known as the Eagle Nebula) shows huge columns of cool gas, (including molecular hydrogen, H2) and dust. These columns are of higher density than the surrounding regions and have resisted evaporation by the ultraviolet radiation from a cluster of hot stars just beyond the upper-right corner of this image. The tallest pillar is about 1 light-year long, and the M16 region is about 7000 light-years away from us. (b) This close-up view of one of the pillars shows some very dense globules, many of which harbor embryonic stars. Astronomers coined the term *evaporating gas globules* (EGGs) for these structures, in part so that they could say we found EGGs inside the Eagle Nebula. It is possible that because these EGGs are exposed to the relentless action of the radiation from nearby hot stars, some may not yet have collected enough material to form a star. (credit a : modification of work by NASA, ESA, and the Hubble Heritage Team (STScI/AURA); credit b: modification of work by NASA, ESA, STScI, J. Hester and P. Scowen (Arizona State University))

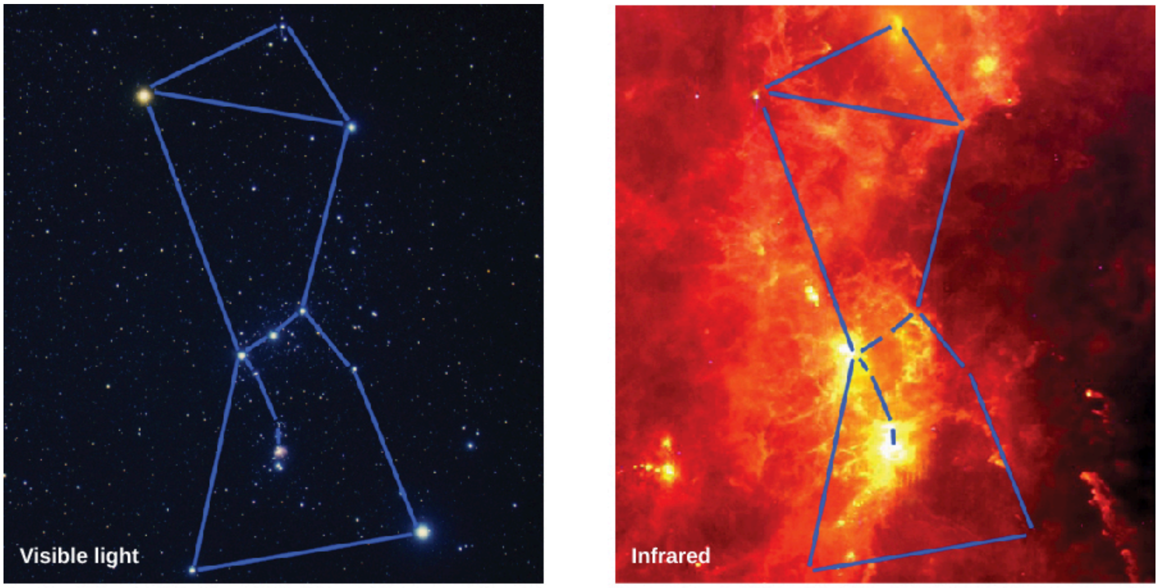
Molecular Clouds: Stellar Nurseries

The most massive reservoirs of interstellar matter—and some of the most massive objects in the Milky Way Galaxy—are the **giant molecular clouds**. These clouds have cold interiors with characteristic temperatures of only 10–20 K; most of their gas atoms are bound into molecules. These clouds turn out to be the birthplaces of most stars in our Galaxy.

The masses of molecular clouds range from a thousand times the mass of the Sun to about 3 million solar masses. Molecular clouds have a complex filamentary structure, similar to cirrus clouds in Earth's atmosphere, but much less dense. The molecular cloud filaments can be up to 1000 light-years long. Within the clouds are cold, dense regions with typical masses of 50 to 500 times the mass of the Sun; we give these regions the highly technical name *clumps*. Within these clumps, there are even denser, smaller regions called cores. The cores are the embryos of stars. The conditions in these cores—low temperature and high density—are just what is required to make stars. Remember that the essence of the life story of any star is the ongoing competition between two forces: *gravity* and *pressure*. The force of gravity, pulling inward, tries to make a star collapse. Internal pressure produced by the motions of the gas atoms, pushing outward, tries to force the star to expand. When a star is first forming, low temperature (and hence, low pressure) and high density (hence, greater gravitational attraction) both work to give gravity the advantage. In order to form a star—that is, a dense, hot ball of matter capable of starting nuclear reactions deep within—we need a typical core of interstellar atoms and molecules to shrink in radius and increase in density by a factor of nearly 10^{20}. It is the force of gravity that produces this drastic collapse.

The Orion Molecular Cloud

Let's discuss what happens in regions of star formation by considering a nearby site where stars are forming right now. One of the best-studied stellar nurseries is in the constellation of Orion, The Hunter, about 1500 light-years away (**Figure 24.3**). The pattern of the hunter is easy to recognize by the conspicuous "belt" of three stars that mark his waist. The Orion molecular cloud is much larger than the star pattern and is truly an impressive structure. In its long dimension, it stretches over a distance of about 100 light-years. The total quantity of molecular gas is about 200,000 times the mass of the Sun. Most of the cloud does not glow with visible light but betrays its presence by the radiation that the dusty gas gives off at infrared and radio wavelengths.



(a) (b)

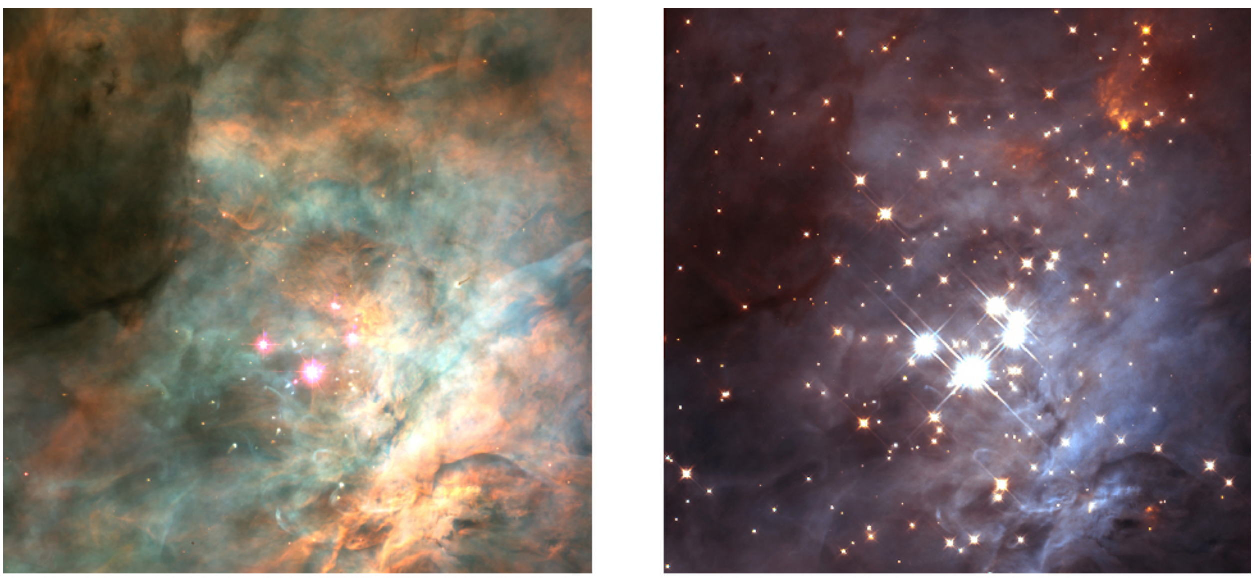
Figure 24.3 (a) The Orion star group was named after the legendary hunter in Greek mythology. Three stars close together in a link mark Orion's belt. The ancients imagined a sword hanging from the belt; the object at the end of the blue line in this sword is the Orion Nebula. (b) This wide-angle, infrared view of the same area was taken with the Infrared Astronomical Satellite. Heated dust clouds dominate in this false-color image, and many of the stars that stood out on part (a) are now invisible. An exception is the cool, red-giant star Betelgeuse, which can be seen as a yellowish point at the left vertex of the blue triangle (at Orion's left armpit). The large, yellow ring to the right of Betelgeuse is the remnant of an exploded star. The infrared image lets us see how large and full of cooler material the Orion molecular cloud really is. On the visible-light image at left, you see only two colorful regions of interstellar matter—the two, bright yellow splotches at the left end of and below Orion's belt. The lower one is the Orion Nebula and the higher one is the region of the Horsehead Nebula. (credit: modification of work by NASA, visible light: Akira Fujii; infrared: Infrared Astronomical Satellite)

The stars in Orion's belt are typically about 5 million years old, whereas the stars near the middle of the "sword" hanging from Orion's belt are only 300,000 to 1 million years old. The region about halfway down the sword where star formation is still taking place is called the Orion Nebula. About 2200 young stars are found in this region, which is only slightly larger than a dozen light-years in diameter. The Orion Nebula also contains a tight cluster of stars called the Trapezium (**Figure 24.5**). The brightest Trapezium stars can be seen easily with a small telescope.

(a) (b)

Figure 24.4 (a) The Orion Nebula is shown in visible light. (b) With near-infrared radiation, we can see more detail within the dusty nebula since infrared can penetrate dust more easily than can visible light. (credit a: modification of work by Filip Lolić; credit b: modification of work by NASA/JPL-Caltech/T. Megeath (University of Toledo, Ohio))

Compare this with our own solar neighborhood, where the typical spacing between stars is about 3 light-years. Only a small number of stars in the Orion cluster can be seen with visible light, but infrared images—which penetrate the dust better—detect the more than 2000 stars that are part of the group (**Figure 24.5**).



(a) (b)

Figure 24.5 The Orion Nebula harbors some of the youngest stars in the solar neighborhood. At the heart of the nebula is the Trapezium cluster, which includes four very bright stars that provide much of the energy that causes the nebula to glow so brightly. In these images, we see a section of the nebula in (a) visible light and (b) infrared. The four bright stars in the center of the visible-light image are the Trapezium stars. Notice that most of the stars seen in the infrared are completely hidden by dust in the visible-light image. (credit a: modification of work by NASA, C.R. O'Dell and S.K. Wong (Rice University); credit b: modification of work by NASA; K.L. Luhman (Harvard-Smithsonian Center for Astrophysics); and G. Schneider, E. Young, G. Rieke, A. Cotera, H. Chen, M. Rieke, R. Thompson (Steward Observatory, University of Arizona))

Studies of Orion and other star-forming regions show that star formation is not a very efficient process. In the region of the Orion Nebula, about 1% of the material in the cloud has been turned into stars. That is why we still see a substantial amount of gas and dust near the Trapezium stars. The leftover material is eventually heated, either by the radiation and winds from the hot stars that form or by explosions of the most massive stars. (We will see in later chapters that the most massive stars go through their lives very quickly and end by exploding.)

Take a **journey through the Orion Nebula (https://openstaxcollege.org/l/30OriNebula)** to view a nice narrated video tour of this region.

Whether gently or explosively, the material in the neighborhood of the new stars is blown away into interstellar space. Older groups or clusters of stars can now be easily observed in visible light because they are no longer shrouded in dust and gas (**Figure 24.6**).

Figure 24.6 This young cluster of stars known as Westerlund 2 formed within the Carina star-forming region about 2 million years ago. Stellar winds and pressure produced by the radiation from the hot stars within the cluster are blowing and sculpting the surrounding gas and dust. The nebula still contains many globules of dust. Stars are continuing to form within the denser globules and pillars of the nebula. This Hubble Space Telescope image includes near-infrared exposures of the star cluster and visible-light observations of the surrounding nebula. Colors in the nebula are dominated by the red glow of hydrogen gas, and blue-green emissions from glowing oxygen. (credit: NASA, ESA, the Hubble Heritage Team (STScI/AURA), A. Nota (ESA/STScI), and the Westerlund 2 Science Team)

Although we do not know what initially caused stars to begin forming in Orion, there is good evidence that the first generation of stars triggered the formation of additional stars, which in turn led to the formation of still more stars (**Figure 24.7**).

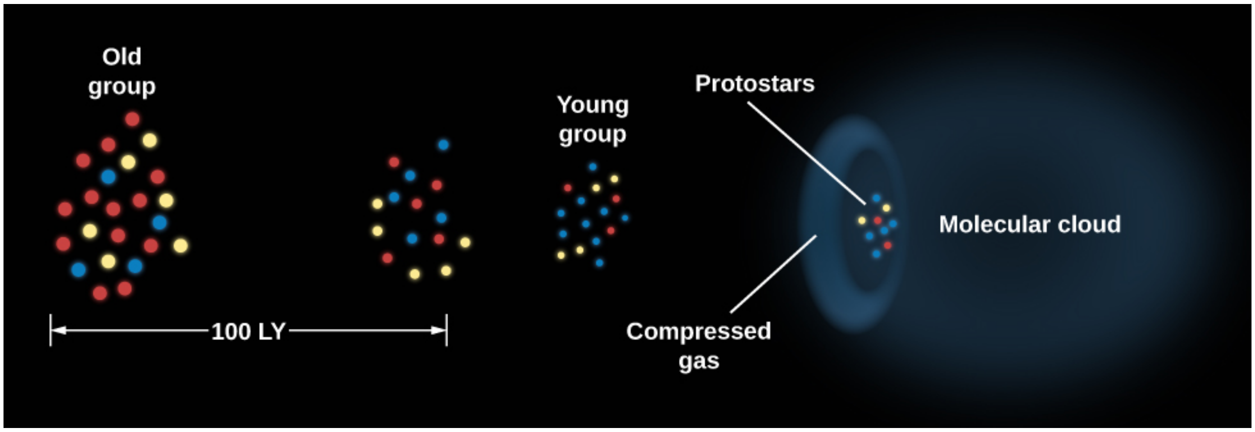


Figure 24.7 Star formation can move progressively through a molecular cloud. The oldest group of stars lies to the left of the diagram and has expanded because of the motions of individual stars. Eventually, the stars in the group will disperse and no longer be recognizable as a cluster. The youngest group of stars lies to the right, next to the molecular cloud. This group of stars is only 1 to 2 million years old. The pressure of the hot, ionized gas surrounding these stars compresses the material in the nearby edge of the molecular cloud and initiates the gravitational collapse that will lead to the formation of more stars.

The basic idea of triggered star formation is this: when a massive star is formed, it emits a large amount of ultraviolet radiation and ejects high-speed gas in the form of a stellar wind. This injection of energy heats the gas around the stars and causes it to expand. When massive stars exhaust their supply of fuel, they explode, and the energy of the explosion also heats the gas. The hot gases pile into the surrounding cold molecular cloud, compressing the material in it and increasing its density. If this increase in density is large enough, gravity will overcome pressure, and stars will begin to form in the compressed gas. Such a chain reaction—where the brightest and hottest stars of one area become the cause of star formation "next door"—seems to have occurred not only in Orion but also in many other molecular clouds.

There are many molecular clouds that form only (or mainly) low-mass stars. Because low-mass stars do not have strong winds and do not die by exploding, triggered star formation cannot occur in these clouds. There are also stars that form in relative isolation in small cores. Therefore, not all star formation is originally triggered by the death of massive stars. However, there are likely to be other possible triggers, such as spiral density waves and other processes we do not yet understand.

The Birth of a Star

Although regions such as Orion give us clues about how star formation begins, the subsequent stages are still shrouded in mystery (and a lot of dust). There is an enormous difference between the density of a molecular cloud core and the density of the youngest stars that can be detected. Direct observations of this collapse to higher density are nearly impossible for two reasons. First, the dust-shrouded interiors of molecular clouds where stellar births take place cannot be observed with visible light. Second, the timescale for the initial collapse—thousands of years—is very short, astronomically speaking. Since each star spends such a tiny fraction of its life in this stage, relatively few stars are going through the collapse process at any given time. Nevertheless, through a combination of theoretical calculations and the limited observations available, astronomers have pieced together a picture of what the earliest stages of stellar evolution are likely to be.

The first step in the process of creating stars is the formation of dense cores within a clump of gas and dust (**Figure 24.8**(a)). It is generally thought that all the material for the star comes from the core, the larger structure surrounding the forming star. Eventually, the gravitational force of the infalling gas becomes strong enough to overwhelm the pressure exerted by the cold material that forms the dense cores. The material then undergoes a rapid collapse, and the density of the core increases greatly as a result. During the time a dense core is contracting to become a true star, but before the fusion of protons to produce helium begins, we call the object a **protostar**.

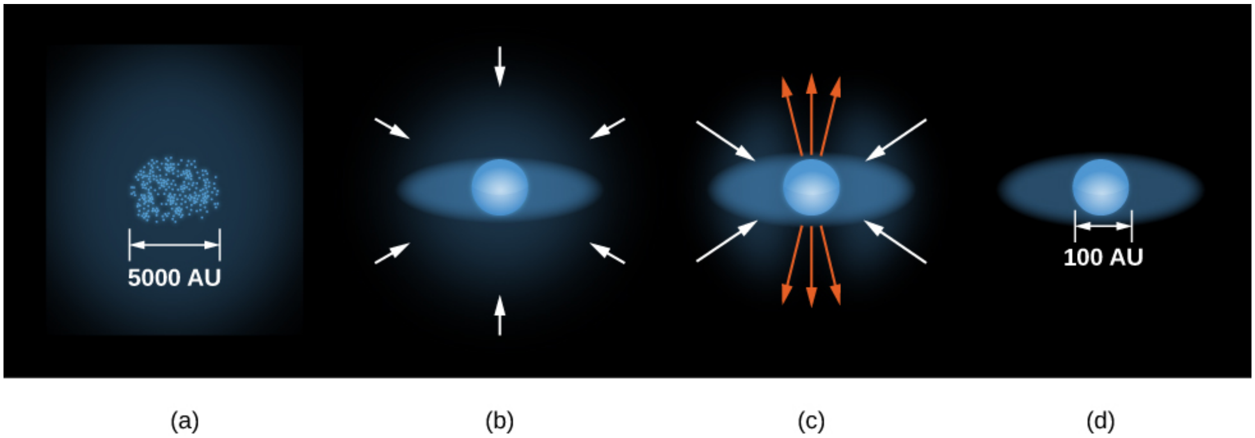


Figure 24.8 (a) Dense cores form within a molecular cloud. (b) A protostar with a surrounding disk of material forms at the center of a dense core, accumulating additional material from the molecular cloud through gravitational attraction. (c) A stellar wind breaks out but is confined by the disk to flow out along the two poles of the star. (d) Eventually, this wind sweeps away the cloud material and halts the accumulation of additional material, and a newly formed star, surrounded by a disk, becomes observable. These sketches are not drawn to the same scale. The diameter of a typical envelope that is supplying gas to the newly forming star is about 5000 AU. The typical diameter of the disk is about 100 AU or slightly larger than the diameter of the orbit of Pluto.

The natural turbulence inside a clump tends to give any portion of it some initial spinning motion (even if it is very slow). As a result, each collapsing core is expected to spin. According to the law of conservation of angular momentum (discussed in the chapter on **Conservation of Angular Momentum**), a rotating body spins more rapidly as it decreases in size. In other words, if the object can turn its material around a smaller circle, it can move that material more quickly—like a figure skater spinning more rapidly as she brings her arms in tight to her body. This is exactly what happens when a core contracts to form a protostar: as it shrinks, its rate of spin increases.

But all directions on a spinning sphere are not created equal. As the protostar rotates, it is much easier for material to fall right onto the poles (which spin most slowly) than onto the equator (where material moves around most rapidly). Therefore, gas and dust falling in toward the protostar's equator are "held back" by the rotation and form a whirling extended disk around the equator (part b in **Figure 24.8**). You may have observed this same "equator effect" on the amusement park ride in which you stand with your back to a cylinder that is spun faster and faster. As you spin really fast, you are pushed against the wall so strongly that you cannot possibly fall toward the center of the cylinder. Gas can, however, fall onto the protostar easily from directions away from the star's equator.

The protostar and disk at this stage are embedded in an envelope of dust and gas from which material is still falling onto the protostar. This dusty envelope blocks visible light, but infrared radiation can get through. As a result, in this phase of its evolution, the protostar itself is emitting infrared radiation and so is observable only in the infrared region of the spectrum. Once almost all of the available material has been accreted and the central protostar has reached nearly its final mass, it is given a special name: it is called a *T Tauri star*, named after one of the best studied and brightest members of this class of stars, which was discovered in the constellation of Taurus. (Astronomers have a tendency to name types of stars after the first example they discover or come to understand. It's not an elegant system, but it works.) Only stars with masses less than or similar to the mass of the Sun become T Tauri stars. Massive stars do not go through this stage, although they do appear to follow the formation scenario illustrated in **Figure 24.8**.

Winds and Jets

Recent observations suggest that T Tauri stars may actually be stars in a middle stage between protostars and hydrogen-fusing stars such as the Sun. High-resolution infrared images have revealed jets of material as well as *stellar winds* coming from some T Tauri stars, proof of interaction with their environment. A **stellar wind** consists mainly of protons (hydrogen nuclei) and electrons streaming away from the star at speeds of a few hundred kilometers per second (several hundred thousand miles per hour). When the wind first starts up, the disk of material around the star's equator blocks the wind in its direction. Where the wind particles *can* escape most effectively is in the direction of the star's poles.

Astronomers have actually seen evidence of these beams of particles shooting out in opposite directions from the polar regions of newly formed stars. In many cases, these beams point back to the location of a protostar that is still so completely shrouded in dust that we cannot yet see it (**Figure 24.9**).

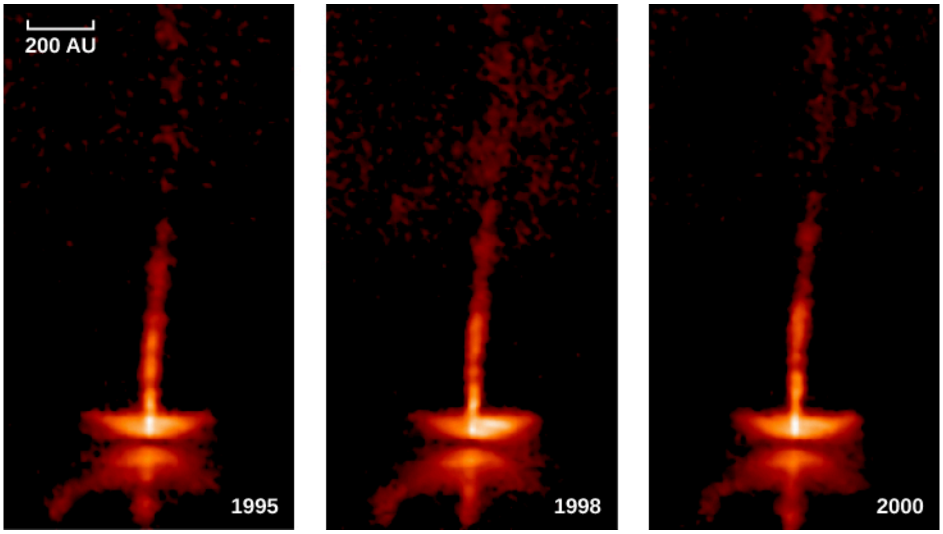


Figure 24.9 Here we see the neighborhood of a protostar, known to us as HH 34 because it is a Herbig-Haro object. The star is about 450 light-years away and only about 1 million years old. Light from the star itself is blocked by a disk, which is larger than 60 billion kilometers in diameter and is seen almost edge-on. Jets are seen emerging perpendicular to the disk. The material in these jets is flowing outward at speeds up to 580,000 kilometers per hour. The series of three images shows changes during a period of 5 years. Every few months, a compact clump of gas is ejected, and its motion outward can be followed. The changes in the brightness of the disk may be due to motions of clouds within the disk that alternately block some of the light and then let it through. This image corresponds to the stage in the life of a protostar shown in part (c) of **Figure 24.8**. (credit: modification of work by Hubble Space Telescope, NASA, ESA)

On occasion, the jets of high-speed particles streaming away from the protostar collide with a somewhat-denser lump of gas nearby, excite its atoms, and cause them to emit light. These glowing regions, each of which is known as a **Herbig-Haro (HH) object** after the two astronomers who first identified them, allow us to trace the progress of the jet to a distance of a light-year or more from the star that produced it. **Figure 24.10** shows two spectacular images of HH objects.

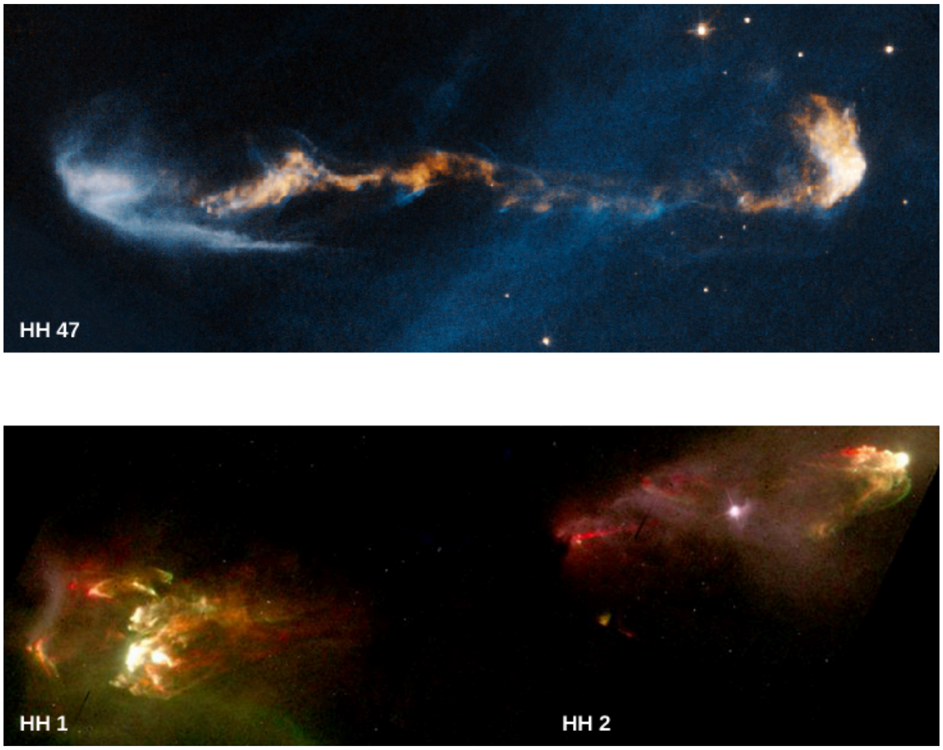


Figure 24.10 These images were taken with the Hubble Space Telescope and show jets flowing outward from newly formed stars. In the HH47 image, a protostar 1500 light-years away (invisible inside a dust disk at the left edge of the image) produces a very complicated jet. The star may actually be wobbling, perhaps because it has a companion. Light from the star illuminates the white region at the left because light can emerge perpendicular to the disk (just as the jet does). At right, the jet is plowing into existing clumps of interstellar gas, producing a shock wave that resembles an arrowhead. The HH1/2 image shows a double-beam jet emanating from a protostar (hidden in a dust disk in the center) in the constellation of Orion. Tip to tip, these jets are more than 1 light-year long. The bright regions (first identified by Herbig and Haro) are places where the jet is a slamming into a clump of interstellar gas and causing it to glow. (credit "HH 47": modification of work by NASA, ESA, and P. Hartigan (Rice University); credit "HH 1 and HH 2: modification of work by J. Hester, WFPC2 Team, NASA)

The wind from a forming star will ultimately sweep away the material that remains in the obscuring envelope of dust and gas, leaving behind the naked disk and protostar, which can then be seen with visible light. We should note that at this point, the protostar itself is still contracting slowly and has not yet reached the main-sequence stage on the H–R diagram (a concept introduced in the chapter **Stellar Properties**). The disk can be detected directly when observed at infrared wavelengths or when it is seen silhouetted against a bright background (**Figure 24.11**).

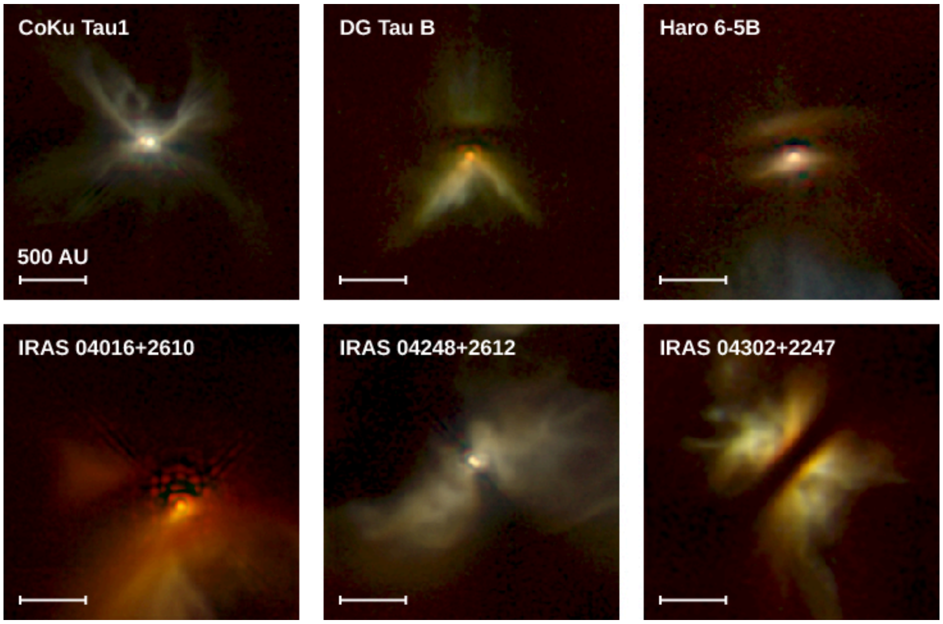


Figure 24.11 These Hubble Space Telescope infrared images show disks around young stars in the constellation of Taurus, in a region about 450 light-years away. In some cases, we can see the central star (or stars—some are binaries). In other cases, the dark, horizontal bands indicate regions where the dust disk is so thick that even infrared radiation from the star embedded within it cannot make its way through. The brightly glowing regions are starlight reflected from the upper and lower surfaces of the disk, which are less dense than the central, dark regions. (Credit: modification of work by D. Padgett (IPAC/Caltech), W. Brandner (IPAC), K. Stapelfeldt (JPL) and NASA)

This description of a protostar surrounded by a rotating disk of gas and dust sounds very much like what happened in our solar system when the Sun and planets formed. Indeed, one of the most important discoveries from the study of star formation in the last decade of the twentieth century was that disks are an inevitable byproduct of the process of creating stars. The next questions that astronomers set out to answer was: will the disks around protostars also form planets? And if so, how often? We will return to these questions later in this chapter.

To keep things simple, we have described the formation of single stars. Many stars, however, are members of binary or triple systems, where several stars are born together. In this case, the stars form in nearly the same way. Widely separated binaries may each have their own disk; close binaries may share a single disk.

24.2 | The H–R Diagram and the Study of Stellar Evolution

Learning Objectives

By the end of this section, you will be able to:

- Determine the age of a protostar using an H-R diagram and the protostar’s luminosity and temperature
- Explain the interplay between gravity and pressure, and how the contracting protostar changes its position in the H–R diagram as a result

One of the best ways to summarize all of these details about how a star or protostar changes with time is to use a Hertzsprung-Russell (H–R) diagram. Recall that, when looking at an **H–R diagram**, the temperature (the horizontal axis) is plotted increasing toward the left. As a star goes through the stages of its life, its luminosity and temperature change. Thus, its position on the H–R diagram, in which luminosity is plotted against temperature, also changes. As a star ages, we must replot it in different places on the diagram. Therefore, astronomers often speak of a star *moving* on the H–R diagram, or of its evolution tracing out a path on the diagram. In this context, “tracing out a path” has nothing to do with the star’s motion through space; this is just a shorthand way of saying that its temperature and luminosity change as it evolves.

Watch an **animation (https://openstaxcollege.org/l/30aniomelen)** of the stars in the Omega Centauri cluster as they rearrange according to luminosity and temperature, forming a Hertzsprung-Russell (H–R) diagram.

To estimate just how much the luminosity and temperature of a star change as it ages, we must resort to calculations. Theorists compute a series of models for a star, with each successive model representing a later point in time. Stars may change for a variety of reasons. Protostars, for example, change in size because they are contracting, and their temperature and luminosity change as they do so. After nuclear fusion begins in the star's core (see **Star Formation**), main-sequence stars change because they are using up their nuclear fuel.

Given a model that represents a star at one stage of its evolution, we can calculate what it will be like at a slightly later time. At each step, the model predicts the luminosity and size of the star, and from these values, we can figure out its surface temperature. A series of points on an H–R diagram, calculated in this way, allows us to follow the life changes of a star and hence is called its *evolutionary track*.

Evolutionary Tracks

Let's now use these ideas to follow the evolution of protostars that are on their way to becoming main-sequence stars. The evolutionary tracks of newly forming stars with a range of stellar masses are shown in **Figure 24.12**. These young stellar objects are not yet producing energy by nuclear reactions, but they derive energy from gravitational contraction—through the sort of process proposed for the Sun by Helmhotz and Kelvin in this last century (see the chapter on **Sources of Sunshine: Thermal and Gravitational?**).

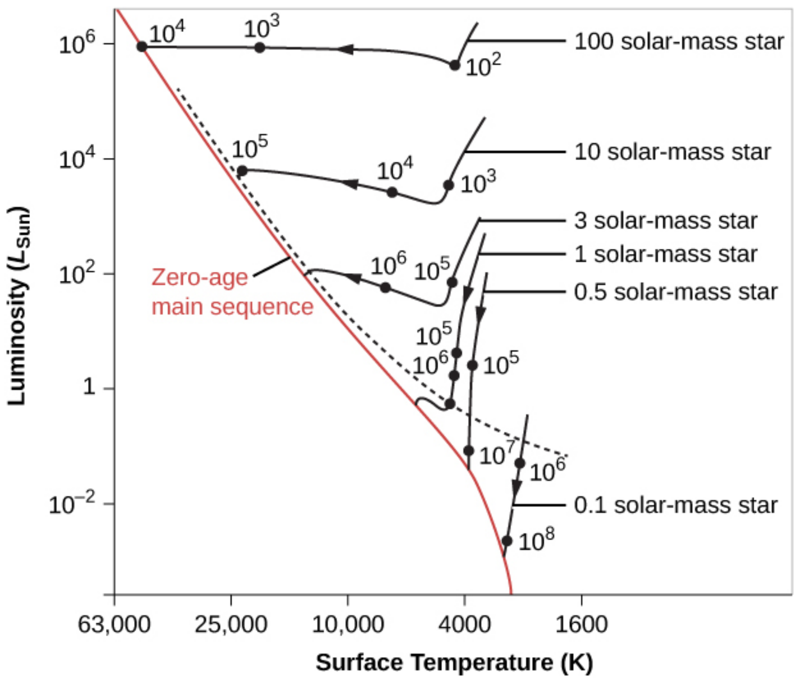


Figure 24.12 Tracks are plotted on the H–R diagram to show how stars of different masses change during the early parts of their lives. The number next to each dark point on a track is the rough number of years it takes an embryo star to reach that stage (the numbers are the result of computer models and are therefore not well known). Note that the surface temperature (K) on the horizontal axis increases toward the left. You can see that the more mass a star has, the shorter time it takes to go through each stage. Stars above the dashed line are typically still surrounded by infalling material and are hidden by it.

Initially, a protostar remains fairly cool with a very large radius and a very low density. It is transparent to infrared radiation, and the heat generated by gravitational contraction can be radiated away freely into space. Because heat builds up slowly inside the protostar, the gas pressure remains low, and the outer layers fall almost unhindered toward the center. Thus, the protostar undergoes very rapid collapse, a stage that corresponds to the roughly vertical lines at the right of **Figure 24.12**. As the star shrinks, its surface area gets smaller, and so its total luminosity decreases. The rapid contraction stops only when the protostar becomes dense and opaque enough to trap the heat released by gravitational contraction.

When the star begins to retain its heat, the contraction becomes much slower, and changes inside the contracting star keep the luminosity of stars like our Sun roughly constant. The surface temperatures start to build up, and the star "moves" to the left in the H–R diagram. Stars first become visible only after the stellar wind described earlier clears away the surrounding dust and gas. This can happen during the rapid-contraction phase for low-mass stars, but high-mass stars remain shrouded in dust until they end their early phase of gravitational contraction (see the dashed line in **Figure 24.12**).

To help you keep track of the various stages that stars go through in their lives, it can be useful to compare the development of a star to that of a human being. (Clearly, you will not find an exact correspondence, but thinking through the stages in human terms may help you remember some of the ideas we are trying to emphasize.) Protostars might be compared to human embryos—as yet unable to sustain themselves but drawing resources from their environment as they grow. Just as the birth of a child is the moment it is called upon to produce its own energy (through eating and breathing), so astronomers say that a star is born when it is able to sustain itself through nuclear reactions (by making its own energy.)

When the star's central temperature becomes high enough (about 12 million K) to fuse hydrogen into helium, we say that the star has reached the main sequence (a concept introduced in **Stellar Properties**). It is now a full-fledged star, more or less in equilibrium, and its rate of change slows dramatically. Only the gradual depletion of hydrogen as it is transformed into helium in the core slowly changes the star's properties.

The mass of a star determines exactly where it falls on the main sequence. As **Figure 24.12** shows, massive stars on the main sequence have high temperatures and high luminosities. Low-mass stars have low temperatures and low luminosities.

Objects of extremely low mass never achieve high-enough central temperatures to ignite nuclear reactions. The lower end of the main sequence stops where stars have a mass just barely great enough to sustain nuclear reactions at a sufficient rate to stop gravitational contraction. This critical mass is calculated to be about 0.075 times the mass of the Sun. As we discussed in the chapter on **Stellar Properties**, objects below this critical mass are called either brown dwarfs or planets. At the other extreme, the upper end of the main sequence terminates at the point where the energy radiated by the newly forming massive star becomes so great that it halts the accretion of additional matter. The upper limit of stellar mass is between 100 and 200 solar masses.

Will It Become a Star?

In discussing the process of star formation, it is important to realize that not every gas cloud that contracts will result in the formation of a star. The gravitational contraction of the cloud results in a rising temperature, because the loss of gravitational potential energy (see **Section 10.5**) leads to an increase in thermal energy. The rising temperature, in turn, leads to an increased thermodynamic pressure (recall the ideal gas law from **Section 17.2**).

The resulting outward force from this pressure opposes the inward gravitational force. (This is the same kind of **hydrostatic equilibrium** that we discussed in the chapter on **The Sun**). Only if the gravitational force can overcome the outward force of thermodynamic pressure can the protostar eventually become dense enough and hot enough to initiate nuclear fusion.

Two major factors determine the fate of the initial gas cloud. One is its particular molecular composition. If it contains certain molecules (like carbon monoxide) that can radiate away thermal energy in the infrared region of the spectrum (see the section on **Molecular Spectra**), then the thermal pressure can be reduced, allowing the force of gravitational collapse to proceed.

But by far the most crucial factor is the amount of mass contained in the original molecular cloud as compared to its initial temperature. A rule of thumb for whether the balance between forces tips in favor of star formation is whether or not the cloud exceeds the **Jeans mass**:

$$M_{\text{Jeans}} = 18M_{\text{Sun}}\sqrt{\frac{T^3}{n}} \tag{24.1}$$

Here the temperature, T, is in kelvins, and the quantity n is the density of the cloud in molecules per cm^3.

In general, molecular clouds exceeding the Jeans mass have enough mass for the gravitational force to overpower thermal pressure, and they eventually become hot and dense enough to begin nuclear fusion and form a star. Clouds with less than the Jeans mass will likely end up as a brown dwarf, not a star. Of course, as mentioned above, molecular composition also plays a role here. The fact that, as we shall see in **Big Bang Cosmology**, the early universe contained few molecules that could radiate away thermal energy during the cloud-collapse process means that the earliest stars that formed in the universe had to be considerably more massive in order to form at all.

Example 24.1

Typical Molecular Clouds

In the Orion nebula, a typical molecular cloud might have a temperature of 30 K and a number density of 300 particles per cm^3. How much mass would such a cloud need to have in order to collapse to form a star?

Solution

The Jeans mass for such a cloud would be

$$M_{\text{Jeans}} = 18M_{\text{Sun}}\sqrt{\frac{30^3}{300}} = 171M_{\text{Sun}} \quad (24.2)$$

So, such a cloud would need a mass larger than 171 M_{Sun} in order to collapse to form a star.

24.1 Suppose that the same cloud as in the previous example had already partially collapsed, and had successfully radiated away some of its thermal energy, so that its temperature had remained unchanged while its density had increased to 30,000 particles per cm^3. How massive would such a cloud need to be in order to form a star?

Evolutionary Timescales

How long it takes a star to form depends on its mass. The numbers that label the points on each track in **Figure 24.12** are the times, in years, required for the embryo stars to reach the stages we have been discussing. Stars of masses much higher than the Sun's reach the main sequence in a few thousand to a million years. The Sun required millions of years before it was born. Tens of millions of years are required for stars of lower mass to evolve to the lower main sequence. (We will see that this turns out to be a general principle: massive stars go through *all* stages of evolution faster than low-mass stars do.)

We will take up the subsequent stages in the life of a star in **Evolution from the Main Sequence to Red Giants**, examining what happens after stars arrive in the main sequence and begin a "prolonged adolescence" and "adulthood" of fusing hydrogen to form helium. But now we want to examine the connection between the formation of stars and planets.

24.3 | Evolution from the Main Sequence to Red Giants

Learning Objectives

By the end of this section, you will be able to:

- Explain the zero-age main sequence
- Describe what happens to main-sequence stars of various masses as they exhaust their hydrogen supply

One of the best ways to get a "snapshot" of a group of stars is by plotting their properties on an H–R diagram. We have already used the H–R diagram to follow the evolution of protostars up to the time they reach the main sequence. Now we'll see what happens next.

Once a star has reached the main-sequence stage of its life, it derives its energy almost entirely from the conversion of hydrogen to helium via the process of nuclear fusion in its core (see **Source of Sunshine: Nuclear Fusion!**). Since hydrogen is the most abundant element in stars, this process can maintain the star's equilibrium for a long time. Thus, all stars remain on the main sequence for most of their lives. Some astronomers like to call the main-sequence phase the star's "prolonged adolescence" or "adulthood" (continuing our analogy to the stages in a human life).

The left-hand edge of the main-sequence band in the H–R diagram is called the **zero-age main sequence** (see **Figure 22.29**). We use the term *zero-age* to mark the time when a star stops contracting, settles onto the main sequence, and begins to fuse hydrogen in its core. The zero-age main sequence is a continuous line in the H–R diagram that shows where stars of different masses but similar chemical composition can be found when they begin to fuse hydrogen.

Since only 0.7% of the hydrogen used in fusion reactions is converted into energy, fusion does not change the *total* mass of the star appreciably during this long period. It does, however, change the chemical composition in its central regions where nuclear reactions occur: hydrogen is gradually depleted, and helium accumulates. This change of composition changes the luminosity, temperature, size, and interior structure of the star. When a star's luminosity and temperature begin to change, the point that represents the star on the H–R diagram moves away from the zero-age main sequence.

Calculations show that the temperature and density in the inner region slowly increase as helium accumulates in the center of a star. As the temperature gets hotter, each proton acquires more energy of motion on average; this means it is more likely to interact with other protons, and as a result, the rate of fusion also increases. For the proton-proton cycle described in **Source of Sunshine: Nuclear Fusion!**, the rate of fusion goes up roughly as the temperature to the fourth power.

If the rate of fusion goes up, the rate at which energy is being generated also increases, and the luminosity of the star gradually rises. Initially, however, these changes are small, and stars remain within the main-sequence band on the H–R diagram for most of their lifetimes.

Example 24.2

Star Temperature and Rate of Fusion

If a star's temperature were to double, by what factor would its rate of fusion increase?

Solution

Since the rate of fusion (like temperature) goes up to the fourth power, it would increase by a factor of 2^4, or 16 times.

24.2 If the rate of fusion of a star increased 256 times, by what factor would the temperature increase?

Lifetimes on the Main Sequence

How many years a star remains in the main-sequence band depends on its mass. You might think that a more massive star, having more fuel, would last longer, but it's not that simple. The lifetime of a star in a particular stage of evolution depends on how much nuclear fuel it has and on *how quickly* it uses up that fuel. (In the same way, how long people can keep spending money depends not only on how much money they have but also on how quickly they spend it. This is why many lottery winners who go on spending sprees quickly wind up poor again.) In the case of stars, more massive ones use up their fuel much more quickly than stars of low mass.

The reason massive stars are such spendthrifts is that, as we saw above, the rate of fusion depends *very* strongly on the star's core temperature. And what determines how hot a star's central regions get? It is the *mass* of the star—the weight of the overlying layers determines how high the pressure in the core must be: higher mass requires higher pressure to balance it. Higher pressure, in turn, is produced by higher temperature. The higher the temperature in the central regions, the faster the star races through its storehouse of central hydrogen. Although massive stars have more fuel, they burn it so prodigiously that their lifetimes are much shorter than those of their low-mass counterparts. You can also understand now why the most massive main-sequence stars are also the most luminous. Like new rock stars with their first platinum album, they spend their resources at an astounding rate.

The main-sequence lifetimes of stars of different masses are listed in **Table 24.1**. This table shows that the most massive stars spend only a few million years on the main sequence. A star of 1 solar mass remains there for roughly 10 billion years, while a star of about 0.4 solar mass has a main-sequence lifetime of some 200 billion years, which is longer than the current age of the universe. (Bear in mind, however, that every star spends *most* of its total lifetime on the main sequence. Stars devote an average of 90% of their lives to peacefully fusing hydrogen into helium.)

Lifetimes of Main-Sequence Stars

Spectral Type	Surface Temperature (K)	Mass (Mass of Sun = 1)	Lifetime on Main Sequence (years)
O5	54,000	40	1 million

Table 24.1

Lifetimes of Main-Sequence Stars

Spectral Type	Surface Temperature (K)	Mass (Mass of Sun = 1)	Lifetime on Main Sequence (years)
B0	29,200	16	10 million
A0	9600	3.3	500 million
F0	7350	1.7	2.7 billion
G0	6050	1.1	9 billion
K0	5240	0.8	14 billion
M0	3750	0.4	200 billion

Table 24.1

These results are not merely of academic interest. Human beings developed on a planet around a G-type star. This means that the Sun's stable main-sequence lifetime is so long that it afforded life on Earth plenty of time to evolve. When searching for intelligent life like our own on planets around other stars, it would be a pretty big waste of time to search around O- or B-type stars. These stars remain stable for such a short time that the development of creatures complicated enough to take astronomy courses is very unlikely.

From Main-Sequence Star to Red Giant

Eventually, all the hydrogen in a star's core, where it is hot enough for fusion reactions, is used up. The core then contains only helium, "contaminated" by whatever small percentage of heavier elements the star had to begin with. The helium in the core can be thought of as the accumulated "ash" from the nuclear "burning" of hydrogen during the main-sequence stage.

Energy can no longer be generated by hydrogen fusion in the stellar core because the hydrogen is all gone and, as we will see, the fusion of helium requires much higher temperatures. Since the central temperature is not yet high enough to fuse helium, there is no nuclear energy source to supply heat to the central region of the star. The long period of stability now ends, gravity again takes over, and the core begins to contract. Once more, the star's energy is partially supplied by gravitational energy, in the way described by Kelvin and Helmholtz (see **Sources of Sunshine: Thermal and Gravitational Energy?**). As the star's core shrinks, the energy of the inward-falling material is converted to heat.

The heat generated in this way, like all heat, flows outward to where it is a bit cooler. In the process, the heat raises the temperature of a layer of hydrogen that spent the whole long main-sequence time just outside the core. Like an understudy waiting in the wings of a hit Broadway show for a chance at fame and glory, this hydrogen was almost (but not quite) hot enough to undergo fusion and take part in the main action that sustains the star. Now, the additional heat produced by the shrinking core puts this hydrogen "over the limit," and a shell of hydrogen nuclei just outside the core becomes hot enough for hydrogen fusion to begin.

New energy produced by fusion of this hydrogen now pours outward from this shell and begins to heat up layers of the star farther out, causing them to expand. Meanwhile, the helium core continues to contract, producing more heat right around it. This leads to more fusion in the shell of fresh hydrogen outside the core (**Figure 24.13**). The additional fusion produces still more energy, which also flows out into the upper layer of the star.

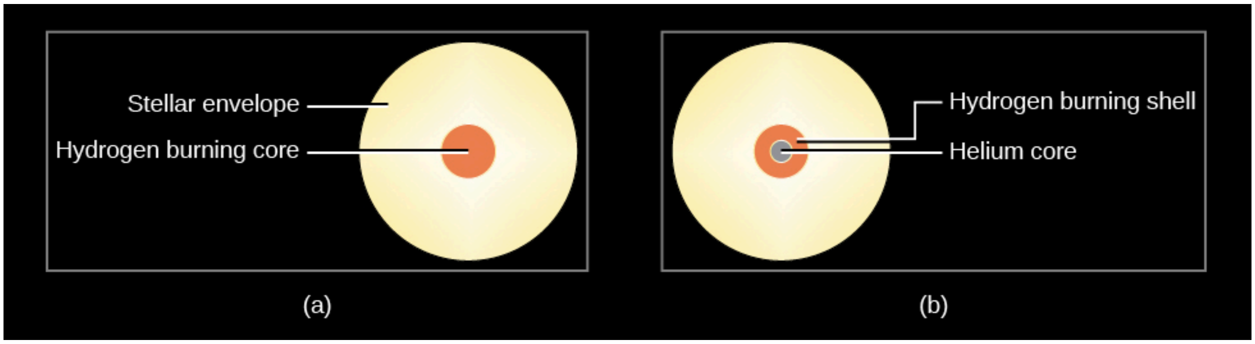


Figure 24.13 (a) During the main sequence, a star has a core where fusion takes place and a much larger envelope that is too cold for fusion. (b) When the hydrogen in the core is exhausted (made of helium, not hydrogen), the core is compressed by gravity and heats up. The additional heat starts hydrogen fusion in a layer just outside the core. Note that these parts of the Sun are not drawn to scale.

Most stars actually generate more energy each second when they are fusing hydrogen in the shell surrounding the helium core than they did when hydrogen fusion was confined to the central part of the star; thus, they increase in luminosity. With all the new energy pouring outward, the outer layers of the star begin to expand, and the star eventually grows and grows until it reaches enormous proportions (**Figure 24.14**).

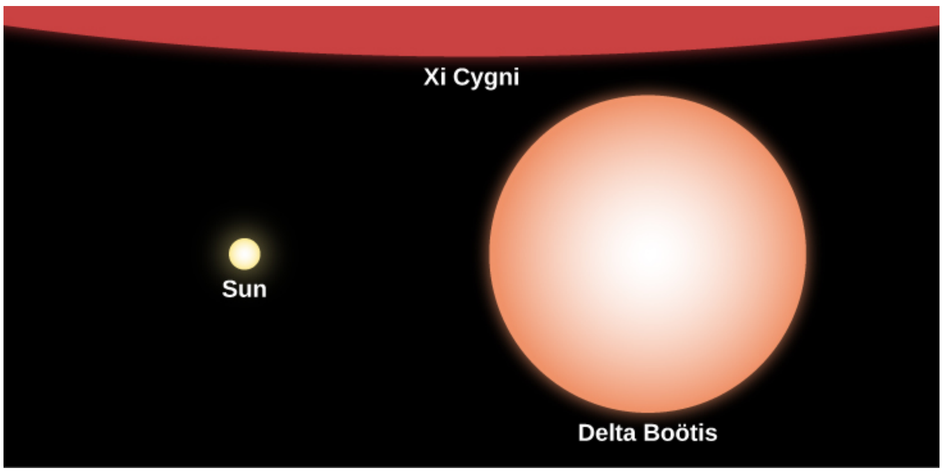


Figure 24.14 This image compares the size of the Sun to that of Delta Boötis, a giant star, and Xi Cygni, a supergiant. Note that Xi Cygni is so large in comparison to the other two stars that only a small portion of it is visible at the top of the frame.

When you take the lid off a pot of boiling water, the steam can expand and it cools down. In the same way, the expansion of a star's outer layers causes the temperature at the surface to decrease. As it cools, the star's overall color becomes redder. (We saw in **Colors of Stars** that a red color corresponds to cooler temperature.)

So the star becomes simultaneously more luminous and cooler. On the H–R diagram, the star therefore leaves the main-sequence band and moves upward (brighter) and to the right (cooler surface temperature). Over time, massive stars become red supergiants, and lower-mass stars like the Sun become red giants. (We first discussed such giant stars in **Stellar Properties**; here we see how such "swollen" stars originate.) You might also say that these stars have "split personalities": their cores are contracting while their outer layers are expanding. (Note that red giant stars do not actually look deep red; their colors are more like orange or orange-red.)

Just how different are these red giants and supergiants from a main-sequence star? **Table 24.2** compares the Sun with the red supergiant Betelgeuse, which is visible above Orion's belt as the bright red star that marks the hunter's armpit. Relative to the Sun, this supergiant has a much larger radius, a much lower average density, a cooler surface, and a much hotter core.

Comparing a Supergiant with the Sun

Property	Sun	Betelgeuse
Mass (2×10^{33} g)	1	16
Radius (km)	700,000	500,000,000
Surface temperature (K)	5,800	3,600
Core temperature (K)	15,000,000	160,000,000
Luminosity (4×10^{26} W)	1	46,000
Average density (g/cm^3)	1.4	1.3×10^{-7}
Age (millions of years)	4,500	10

Table 24.2

Red giants can become so large that if we were to replace the Sun with one of them, its outer atmosphere would extend to the orbit of Mars or even beyond (Figure 24.15). This is the next stage in the life of a star as it moves (to continue our analogy to human lives) from its long period of "youth" and "adulthood" to "old age." (After all, many human beings today also see their outer layers expand a bit as they get older.) By considering the relative ages of the Sun and Betelgeuse, we can also see that the idea that "bigger stars die faster" is indeed true here. Betelgeuse is a mere 10 million years old, which is relatively young compared with our Sun's 4.5 billion years, but it is already nearing its death throes as a red supergiant.

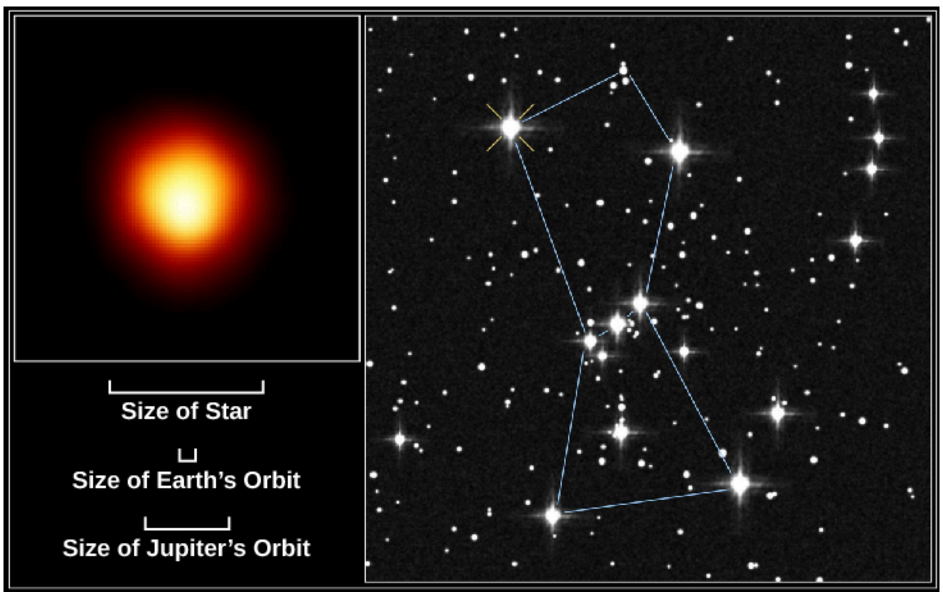


Figure 24.15 Betelgeuse is in the constellation Orion, the hunter; in the right image, it is marked with a yellow "X" near the top left. In the left image, we see it in ultraviolet with the Hubble Space Telescope, in the first direct image ever made of the surface of another star. As shown by the scale at the bottom, Betelgeuse has an extended atmosphere so large that, if it were at the center of our solar system, it would stretch past the orbit of Jupiter. (credit: Modification of work by Andrea Dupree (Harvard-Smithsonian CfA), Ronald Gilliland (STScI), NASA and ESA)

Models for Evolution to the Giant Stage

As we discussed earlier, astronomers can construct computer models of stars with different masses and compositions to see how stars change throughout their lives. Figure 24.16, which is based on theoretical calculations by University of Illinois astronomer Icko Iben, shows an H–R diagram with several tracks of evolution from the main sequence to the giant stage. Tracks are shown for stars with different masses (from 0.5 to 15 times the mass of our Sun) and with chemical compositions similar to that of the Sun. The red line is the initial or zero-age main sequence. The numbers along the tracks indicate the time, in years, required for each star to reach those points in their evolution after leaving the main sequence. Once again, you can see that the more massive a star is, the more quickly it goes through each stage in its life.

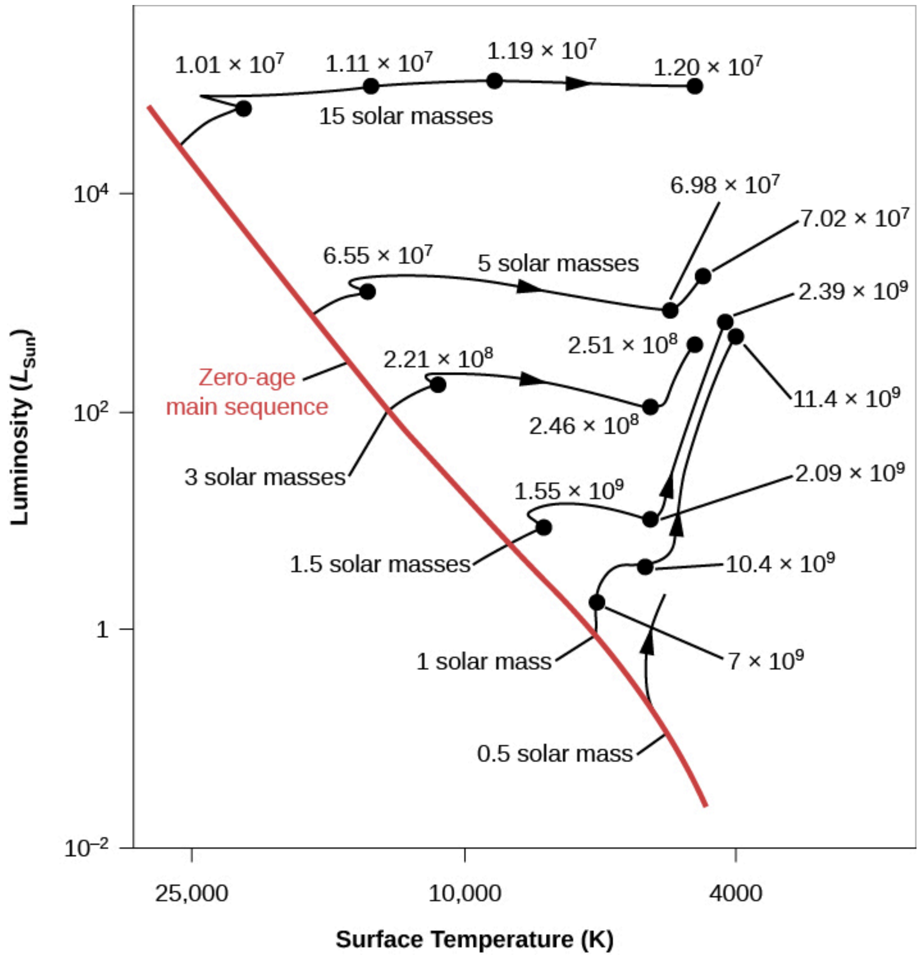


Figure 24.16 The solid black lines show the predicted evolution from the main sequence through the red giant or supergiant stage on the H–R diagram. Each track is labeled with the mass of the star it is describing. The numbers show how many years each star takes to become a giant. The red line is the zero-age main sequence. While theorists debate the exact number of years shown here, our main point should be clear. The more massive the star, the shorter time it takes for each stage in its life.

Note that the most massive star in this diagram has a mass similar to that of Betelgeuse, and so its evolutionary track shows approximately the history of Betelgeuse. The track for a 1-solar-mass star shows that the Sun is still in the main-sequence phase of evolution, since it is only about 4.5 billion years old. It will be billions of years before the Sun begins its own "climb" away from the main sequence—the expansion of its outer layers that will make it a red giant.

24.4 | Star Clusters

Learning Objectives

By the end of this section, you will be able to:

- Explain how star clusters help us understand the stages of stellar evolution
- List the different types of star clusters and describe how they differ in number of stars, structure, and age
- Explain why the chemical composition of globular clusters is different from that of open clusters

The preceding description of stellar evolution is based on calculations. However, no star completes its main-sequence lifetime or its evolution to a red giant quickly enough for us to observe these structural changes as they happen. Fortunately, nature has provided us with an indirect way to test our calculations.

Instead of observing the evolution of a single star, we can look at a group or *cluster* of stars. We look for a group of stars that is very close together in space, held together by gravity, often moving around a common center. Then it is reasonable to assume that the individual stars in the group all formed at nearly the same time, from the same cloud, and with the same composition. We expect that these stars will differ only in mass. And their masses determine how quickly they go through each stage of their lives.

Since stars with higher masses evolve more quickly, we can find clusters in which massive stars have already completed their main-sequence phase of evolution and become red giants, while stars of lower mass in the same cluster are still on the main sequence, or even—if the cluster is very young—undergoing pre-main-sequence gravitational contraction. We can see many stages of stellar evolution among the members of a single cluster, and we can see whether our models can explain why the H–R diagrams of clusters of different ages look the way they do.

The three basic types of clusters astronomers have discovered are globular clusters, open clusters, and stellar associations. Their properties are summarized in **Table 24.3**. As we will see in the next section of this chapter, globular clusters contain only very old stars, whereas open clusters and associations contain young stars.

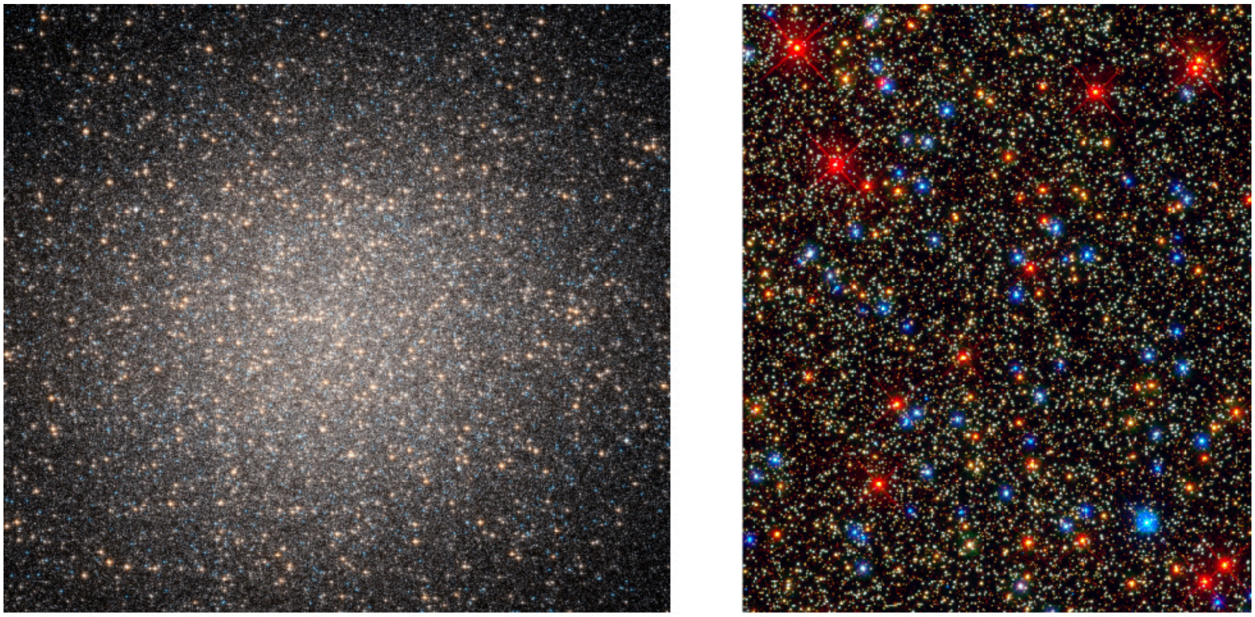
Characteristics of Star Clusters

Characteristic	Globular Clusters	Open Clusters	Associations
Number in the Galaxy	150	Thousands	Thousands
Location in the Galaxy	Halo and central bulge	Disk (and spiral arms)	Spiral arms
Diameter (in light-years)	50–450	<30	100–500
Mass M_{Sun}	10^4–10^6	10^2–10^3	10^2–10^3
Number of stars	10^4–10^6	50–1000	10^2–10^4
Color of brightest stars	Red	Red or blue	Blue
Luminosity of cluster (L_{Sun})	10^4–10^6	10^2–10^6	10^4–10^7
Typical ages	Billions of years	A few hundred million years to, in the case of unusually large clusters, more than a billion years	Up to about 10^7 years

Table 24.3

Globular Clusters

Globular clusters were given this name because they are nearly symmetrical round systems of, typically, hundreds of thousands of stars. The most massive globular cluster in our own Galaxy is Omega Centauri, which is about 16,000 light-years away and contains several million stars (**Figure 24.17**). Note that the brightest stars in this cluster, which are red giants that have already completed the main-sequence phase of their evolution, are red-orange in color. These stars have typical surface temperatures around 4000 K. As we will see, globular clusters are among the oldest parts of our Milky Way Galaxy.



(a) (b)

Figure 24.17 (a) Located at about 16,000 light-years away, Omega Centauri is the most massive globular cluster in our Galaxy. It contains several million stars. (b) This image, taken with the Hubble Space Telescope, zooms in near the center of Omega Centauri. The image is about 6.3 light-years wide. The most numerous stars in the image, which are yellow-white in color, are main-sequence stars similar to our Sun. The brightest stars are red giants that have begun to exhaust their hydrogen fuel and have expanded to about 100 times the diameter of our Sun. The blue stars have started helium fusion. (credit a: modification of work by NASA, ESA and the Hubble Heritage Team (STScI/AURA); credit b: modification of work by NASA, ESA, and the Hubble SM4 ERO Team)

What would it be like to live inside a globular cluster? In the dense central regions, the stars would be roughly a million times closer together than in our own neighborhood. If Earth orbited one of the inner stars in a globular cluster, the nearest stars would be light-months, not light-years, away. They would still appear as points of light, but would be brighter than any of the stars we see in our own sky. The Milky Way would probably be difficult to see through the bright haze of starlight produced by the cluster.

About 150 globular clusters are known in our Galaxy. Most of them are in a spherical halo (or cloud) surrounding the flat disk formed by the majority of our Galaxy's stars. All the globular clusters are very far from the Sun, and some are found at distances of 60,000 light-years or more from the main disk of the Milky Way. The diameters of globular star clusters range from 50 light-years to more than 450 light-years.

Open Clusters

Open clusters are found in the disk of the Galaxy. They have a range of ages, some as old as, or even older than, our Sun. The youngest open clusters are still associated with the interstellar matter from which they formed. Open clusters are smaller than globular clusters, usually having diameters of less than 30 light-years, and they typically contain only several dozen to several hundreds of stars (**Figure 24.18**). The stars in open clusters usually appear well separated from one another, even in the central regions, which explains why they are called "open." Our Galaxy contains thousands of open clusters, but we can see only a small fraction of them. Interstellar dust, which is also concentrated in the disk, dims the light of more distant clusters so much that they are undetectable.

Figure 24.18 This open cluster of young, bright stars is about 6400 light-years away from the Sun. Note the contrast in color between the bright yellow supergiant and the hot blue main-sequence stars. The name comes from John Herschel's nineteenth-century description of it as "a casket of variously colored precious stones." (credit: ESO/Y. Beletsky)

Although the individual stars in an open cluster can survive for billions of years, they typically remain together as a cluster for only a few million years, or at most, a few hundred million years. There are several reasons for this. In small open clusters, the average speed of the member stars within the cluster may be higher than the cluster's escape velocity,[1] and the stars will gradually "evaporate" from the cluster. Close encounters of member stars may also increase the velocity of one of the members beyond the escape velocity. Every few hundred million years or so, the cluster may have a close encounter with a giant molecular cloud, and the gravitational force exerted by the cloud may tear the cluster apart.

Several open clusters are visible to the unaided eye. Most famous among them is the Pleiades, which appears as a tiny group of six stars (some people can see even more than six, and the Pleiades is sometimes called the Seven Sisters). This cluster is arranged like a small dipping spoon and is seen in the constellation of Taurus, the bull. A good pair of binoculars shows dozens of stars in the cluster, and a telescope reveals hundreds. (A car company, Subaru, takes its name from the Japanese term for this cluster; you can see the star group on the Subaru logo.)

The Hyades is another famous open cluster in Taurus. To the naked eye, it appears as a V-shaped group of faint stars marking the face of the bull. Telescopes show that Hyades actually contains more than 200 stars.

Stellar Associations

An **association** is a group of extremely young stars, typically containing 5 to 50 hot, bright O and B stars scattered over a region of space some 100–500 light-years in diameter. As an example, most of the stars in the constellation Orion form one of the nearest stellar associations. Associations also contain hundreds to thousands of low-mass stars, but these are much fainter and less conspicuous. The presence of really hot, luminous stars indicates that star formation in the association has occurred in the last million years or so. Since O stars go through their entire lives in only about a million years, they would not still be around unless star formation has occurred recently. It is therefore not surprising that associations are found in regions rich in the gas and dust required to form new stars. It's like a brand new building still surrounded by some of the construction materials used to build it and with the landscape still showing signs of construction. On the other hand, because associations, like ordinary open clusters, lie in regions occupied by dusty interstellar matter, many are hidden from our view.

1. Escape velocity is the speed needed to overcome the gravity of some object or group of objects. The rockets we send up from Earth, for example, must travel faster than the escape velocity of our planet to be able to get to other worlds.

24.5 | Checking Out the Theory

Learning Objectives

By the end of this section, you will be able to:

- Explain how the H-R diagram of a star cluster can be related to the cluster's age and the stages of evolution of its stellar members
- Describe how the main-sequence turnoff of a cluster reveals its age

In the previous section, we indicated that open clusters are younger than globular clusters, and associations are typically even younger. In this section, we will show how we determine the ages of these star clusters. The key observation is that the stars in these different types of clusters are found in different places in the H–R diagram, and we can use their locations in the diagram in combination with theoretical calculations to estimate how long they have lived.

H–R Diagrams of Young Clusters

What does theory predict for the H–R diagram of a cluster whose stars have recently condensed from an interstellar cloud? Remember that at every stage of evolution, massive stars evolve more quickly than their lower-mass counterparts. After a few million years ("recently" for astronomers), the most massive stars should have completed their contraction phase and be on the main sequence, while the less massive ones should be off to the right, still on their way to the main sequence. These ideas are illustrated in **Figure 24.19**, which shows the H–R diagram calculated by R. Kippenhahn and his associates at Munich University for a hypothetical cluster with an age of 3 million years.

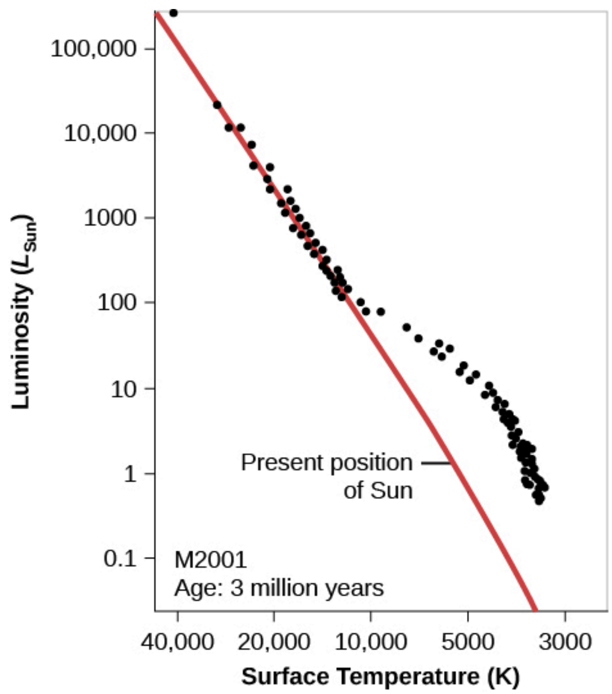


Figure 24.19 We see an H–R diagram for a hypothetical young cluster with an age of 3 million years. Note that the high-mass (high-luminosity) stars have already arrived at the main-sequence stage of their lives, while the lower-mass (lower-luminosity) stars are still contracting toward the zero-age main sequence (the red line) and are not yet hot enough to derive all of their energy from the fusion of hydrogen.

There are real star clusters that fit this description. The first to be studied (in about 1950) was NGC 2264, which is still associated with the region of gas and dust from which it was born (**Figure 24.20**).

Figure 24.20 Located about 2600 light-years from us, this region of newly formed stars, known as the Christmas Tree Cluster, is a complex mixture of hydrogen gas (which is ionized by hot embedded stars and shown in red), dark obscuring dust lanes, and brilliant young stars. The image shows a scene about 30 light-years across. (credit: ESO)

The NGC 2264 cluster's H–R diagram is shown in **Figure 24.21**. The cluster in the middle of the Orion Nebula (shown in **Figure 24.4** and **Figure 24.5**) is in a similar stage of evolution.

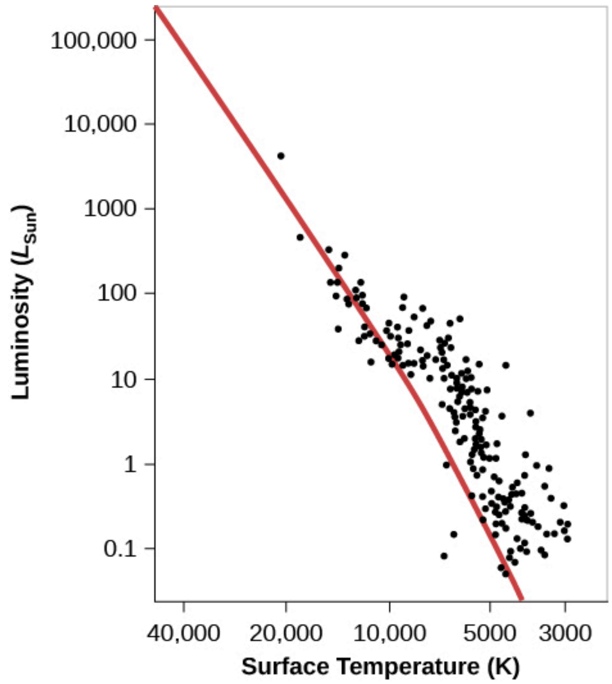


Figure 24.21 Compare this H–R diagram to that in Figure 24.19; although the points scatter a bit more here, the theoretical and observational diagrams are remarkably, and satisfyingly, similar.

As clusters get older, their H–R diagrams begin to change. After a short time (less than a million years after they reach the main sequence), the most massive stars use up the hydrogen in their cores and evolve off the main sequence to become red giants and supergiants. As more time passes, stars of lower mass begin to leave the main sequence and make their way to the upper right of the H–R diagram.

To see the evolution of a star cluster in a dwarf galaxy, you can watch this brief animation (https://openstax.org/l/30StarCluster) of how its H–R diagram changes.

Figure 24.22 is a photograph of NGC 3293, a cluster that is about 10 million years old. The dense clouds of gas and dust are gone. One massive star has evolved to become a red giant and stands out as an especially bright orange member of the cluster.

Figure 24.22 All the stars in an open star cluster like NGC 3293 form at about the same time. The most massive stars, however, exhaust their nuclear fuel more rapidly and hence evolve more quickly than stars of low mass. As stars evolve, they become redder. The bright orange star in NGC 3293 is the member of the cluster that has evolved most rapidly. (credit: ESO/G. Beccari)

Figure 24.23 shows the H–R diagram of the open cluster M41, which is roughly 100 million years old; by this time, a significant number of stars have moved off to the right and become red giants. Note the gap that appears in this H–R diagram between the stars near the main sequence and the red giants. A gap does not necessarily imply that stars avoid a region of certain temperatures and luminosities. In this case, it simply represents a domain of temperature and luminosity through which stars evolve very quickly. We see a gap for M41 because at this particular moment, we have not caught a star in the process of scurrying across this part of the diagram.

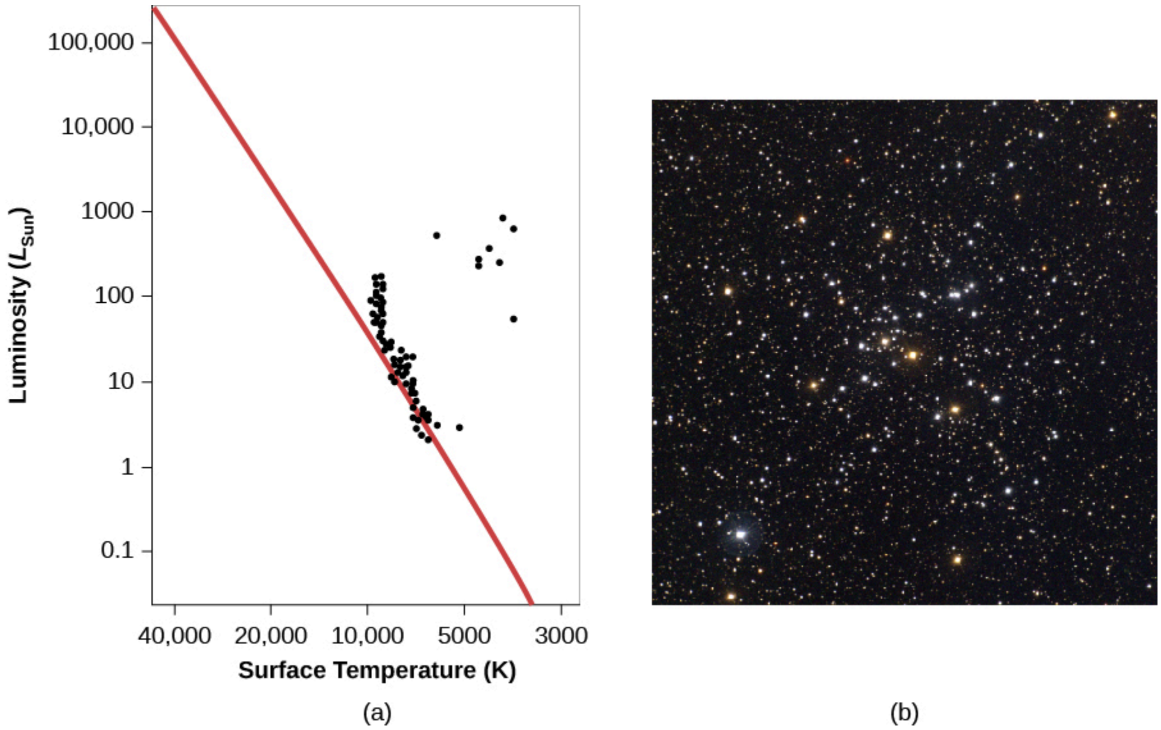


Figure 24.23 (a) Cluster M41 is older than NGC 2264 (see **Figure 24.21**) and contains several red giants. Some of its more massive stars are no longer close to the zero-age main sequence (red line). (b) This ground-based photograph shows the open cluster M41. Note that it contains several orange-color stars. These are stars that have exhausted hydrogen in their centers, and have swelled up to become red giants. (credit b: modification of work by NOAO/AURA/NSF)

H–R Diagrams of Older Clusters

After 4 billion years have passed, many more stars, including stars that are only a few times more massive than the Sun, have left the main sequence (**Figure 24.24**). This means that no stars are left near the top of the main sequence; only the low-mass stars near the bottom remain. The older the cluster, the lower the point on the main sequence (and the lower the mass of the stars) where stars begin to move toward the red giant region. The location in the H–R diagram where the stars have begun to leave the main sequence is called the **main-sequence turnoff**.

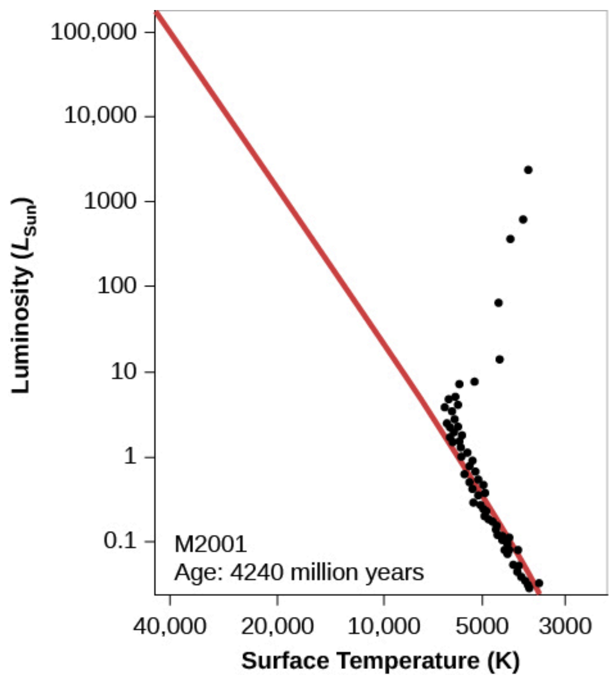


Figure 24.24 We see the H–R diagram for a hypothetical older cluster at an age of 4.24 billion years. Note that most of the stars on the upper part of the main sequence have turned off toward the red-giant region. And the most massive stars in the cluster have already died and are no longer on the diagram.

The oldest clusters of all are the globular clusters. **Figure 24.25** shows the H–R diagram of globular cluster 47 Tucanae. Notice that the luminosity and temperature scales are different from those of the other H–R diagrams in this chapter. In **Figure 24.24**, for example, the luminosity scale on the left side of the diagram goes from 0.1 to 100,000 times the Sun's luminosity. But in **Figure 24.25**, the luminosity scale has been significantly reduced in extent. So many stars in this old cluster have had time to turn off the main sequence that only the very bottom of the main sequence remains.

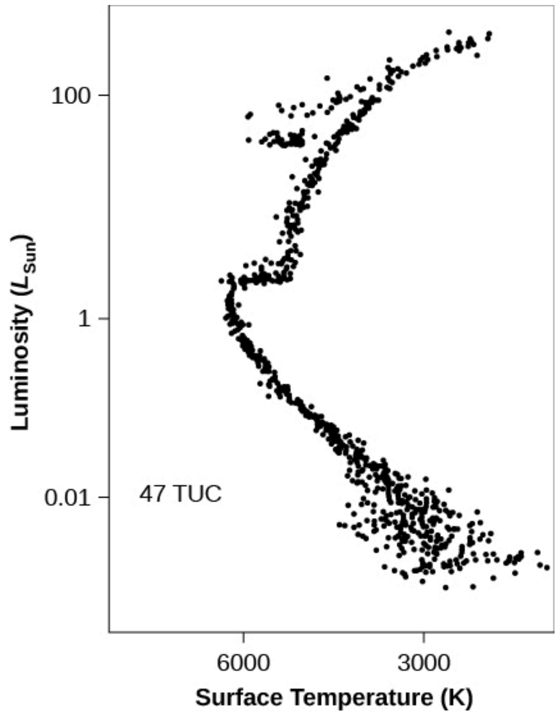


Figure 24.25 This H–R diagram is for the globular cluster 47. Note that the scale of luminosity differs from that of the other H–R diagrams in this chapter. We are only focusing on the lower portion of the main sequence, the only part where stars still remain in this old cluster.

Check out this brief NASA video with a **3-D visualization (https://openstax.org/l/30HRDiagram)** of how an H–R diagram is created for the globular cluster Omega Centauri.

Just how old are the different clusters we have been discussing? To get their actual ages (in years), we must compare the appearances of our *calculated* H–R diagrams of different ages to *observed* H–R diagrams of real clusters. In practice, astronomers use the position at the top of the main sequence (that is, the luminosity at which stars begin to move off the main sequence to become red giants) as a measure of the age of a cluster (the main-sequence turnoff we discussed previously). For example, we can compare the luminosities of the brightest stars that are still on the main sequence in **Figure 24.21** and **Figure 24.24**.

Using this method, some associations and open clusters turn out to be as young as 1 million years old, while others are several hundred million years old. Once all of the interstellar matter surrounding a cluster has been used to form stars or has dispersed and moved away from the cluster, star formation ceases, and stars of progressively lower mass move off the main sequence, as shown in **Figure 24.21**, **Figure 24.23**, and **Figure 24.24**.

To our surprise, even the youngest of the globular clusters in our Galaxy are found to be older than the oldest open cluster. All of the globular clusters have main sequences that turn off at a luminosity less than that of the Sun. Star formation in these crowded systems ceased billions of years ago, and no new stars are coming on to the main sequence to replace the ones that have turned off (see **Figure 24.26**).

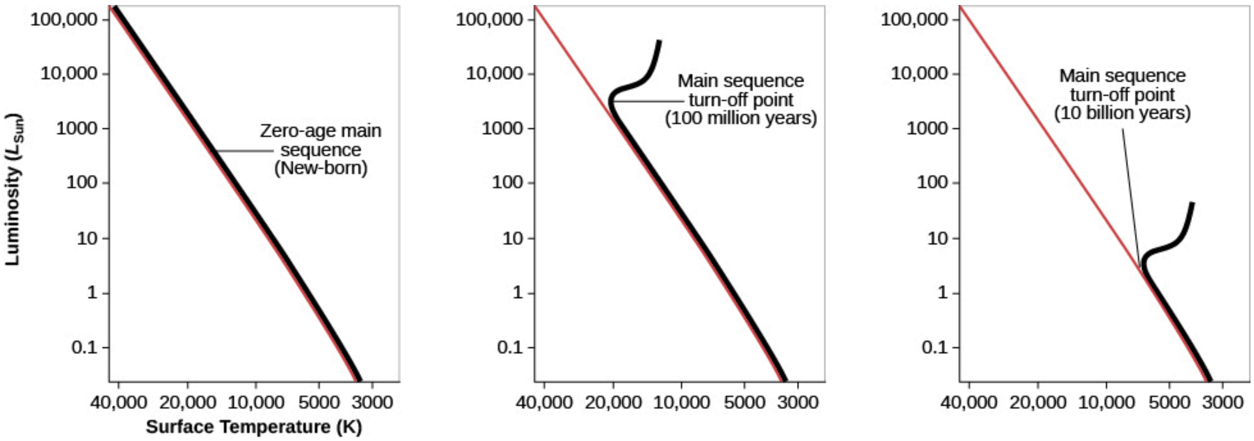


Figure 24.26 This sketch shows how the turn-off point from the main sequence gets lower as we make H–R diagrams for clusters that are older and older.

Indeed, the globular clusters are the oldest structures in our Galaxy (and in other galaxies as well). The youngest have ages of about 11 billion years and some appear to be even older. Since these are the oldest objects we know of, this estimate is one of the best limits we have on the age of the universe itself—it must be at least 11 billion years old. We will return to the fascinating question of determining the age of the entire universe in the chapter on **Big Bang Cosmology**.

H-R Diagrams of Star Clusters Using Magnitudes

The first important feature of a star cluster, which we have just discussed, it that all of its stars were born at about the same time, and are therefore of the same age. As we have seen, an H-R diagram of the cluster yields a main-sequence turnoff point that indicates the age of the cluster.

But stars in a cluster share one other very important property - they are all located at about the same distance from Earth. Recall from the chapter on **The Brightness of Stars** that apparent brightness depends upon the inverse square of its distance from Earth. For stars that are all situated at the same distance from Earth, then, there is a one-to-one correspondence between their apparent brightness and their luminosity.

Now, there is always a one-to-one (logarthmic, see the chapter on **The Brightness of Stars**) relationship between apparent visual magnitude V and brightness. Stars with a smaller V are brighter. (Remember, magnitude is a kind of "backwards" index.)

Furthermore, there is always a one-to-one relationship (again, logarithmic, see the chapter on **The Brightness of Stars)** between the visual absolute magnitude M_V of a star and its luminosity. Stars with a smaller M_V are more luminous.

H-R Diagram by Proxy

We can therfore use M_V and B-V as proxies for luminosity and temperature, and construct an H-R diagram for a particular star cluster. To ensure that more luminous stars are toward the top of the diagram, we must arrange the values of V on the vertical axis in **descending** order. Then, we can use the color index B-V along the horizontal axis, because we know that it has a lower value for bluer (hotter) stars and a higher value for redder (cooler) stars (see the section on **The Electromagnetic Spectrum** for a discussion of this color index).

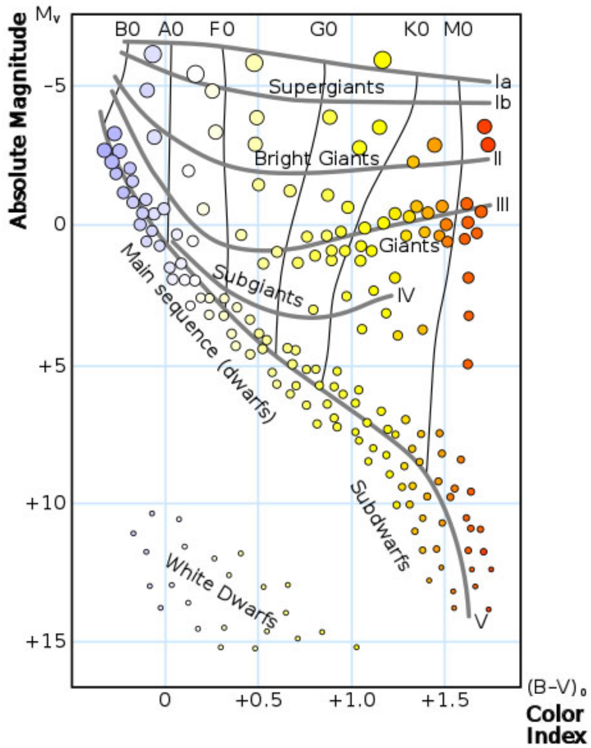


Figure 24.27 This H-R diagram was produced by plotting the absolute visual magnitude, M_V, in reverse order along the vertical axis, and the color index, *B-V*, along the horizontal axis. Credit: Rursus [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0/)]

Now imagine two such proxy H-R diagrams, but with one slight difference. On the vertical axis, we plot the stars' apparent visual magnitude, *V*, instead of M_V. The first H-R diagram is for a cluster located exactly 10 parsecs from Earth, the other from an actual cluster located at some unknown distance. The first diagram would be identical to **Figure 24.27**, because of the definition of absolute magnitude. (At a distance of 10 parsecs, $M_V = V$.) The second would be shifted (up or down) on the vertical axis by a difference of $V - M_V$ for every star in the second cluster.

If we could line up the two graphs horizontally, and determine their vertical shift along the magnitude scale, we would have a measure of the difference between the absolute and apparent magnitudes of the stars in the second cluster, i.e. *m-M*. Recall from the section on **spectroscopic parallax**, that the distance to an object can be found from this very quantity. *m-M* is sometimes called the **distance modulus**. And we know that the distance (in parsecs) is found from **Equation 23.2**

$$d = 10 \times 10^{0.2(m-M)}$$

This technique allows us to use an H-R diagram of any cluster (formed by using the proxies of V and B-V) to compare with a similar diagram for a theoretical cluster located exactly 10 parsecs from Earth, in order to determine the distance to the actual cluster. It's just another application of **spectroscopic parallax**.

Example 24.3

Distance to the Pleiades

An H-R diagram of the Pleiades cluster indicates that its apparent magnitude is larger than the absolute magnitude of a similar cluster. The distance modulus, *m-M* = 5.68. How far away is the Pleiades?

Solution

137 pc

24.3 If the distance modulus to some cluster (*m*-*M*) was a negative number, what would that tell us?

24.6 | Further Evolution of Stars

Learning Objectives

By the end of this section, you will be able to:

- Explain what happens in a star's core when all of the hydrogen has been used up
- Define "planetary nebulae" and discuss their origin
- Discuss the creation of new chemical elements during the late stages of stellar evolution

The "life story" we have related so far applies to almost all stars: each starts as a contracting protostar, then lives most of its life as a stable main-sequence star, and eventually moves off the main sequence toward the red-giant region.

As we have seen, the pace at which each star goes through these stages depends on its mass, with more massive stars evolving more quickly. But after this point, the life stories of stars of different masses diverge, with a wider range of possible behavior according to their masses, their compositions, and the presence of any nearby companion stars.

Because we have written this book for students taking their first astronomy course, we will recount a simplified version of what happens to stars as they move toward the final stages in their lives. We will (perhaps to your heartfelt relief) not delve into all the possible ways aging stars can behave and the strange things that happen when a star is orbited by a second star in a binary system. Instead, we will focus only on the key stages in the evolution of single stars and show how the evolution of high-mass stars differs from that of low-mass stars (such as our Sun).

Helium Fusion

Let's begin by considering stars with composition like that of the Sun and whose *initial* masses are comparatively low—no more than about twice the mass of our Sun. (Such mass may not seem too low, but stars with masses less than this all behave in a fairly similar fashion. We will see what happens to more massive stars in the next section.) Because there are much more low-mass stars than high-mass stars in the Milky Way, the vast majority of stars—including our Sun—follow the scenario we are about to relate. By the way, we carefully used the term *initial masses* of stars because, as we will see, stars can lose quite a bit of mass in the process of aging and dying.

Remember that red giants start out with a helium core where no energy generation is taking place, surrounded by a shell where hydrogen is undergoing fusion. The core, having no source of energy to oppose the inward pull of gravity, is shrinking and growing hotter. As time goes on, the temperature in the core can rise to much hotter values than it had in its main-sequence days. Once it reaches a temperature of 100 million K (but not before such point), three helium atoms can begin to fuse to form a single carbon nucleus. This process is called the **triple-alpha process**, so named because physicists call the nucleus of the helium atom an alpha particle.

When the triple-alpha process begins in low-mass (about 0.8 to 2.0 solar masses) stars, calculations show that the entire core is ignited in a quick burst of fusion called a **helium flash**. (More massive stars also ignite helium but more gradually and not with a flash.) As soon as the temperature at the center of the star becomes high enough to start the triple-alpha process, the extra energy released is transmitted quickly through the entire helium core, producing very rapid heating. The heating speeds up the nuclear reactions, which provide more heating, and which accelerates the nuclear reactions even more. We have runaway generation of energy, which reignites the entire helium core in a flash.

You might wonder why the next major step in nuclear fusion in stars involves three helium nuclei and not just two. Although it is a lot easier to get two helium nuclei to collide, the product of this collision is not stable and falls apart very quickly. It takes three helium nuclei coming together *simultaneously* to make a stable nuclear structure. Given that each helium nucleus has two positive protons and that such protons repel one another, you can begin to see the problem. It takes a temperature of 100 million K to slam three helium nuclei (six protons) together and make them stick. But when that happens, the star produces a carbon nucleus.

Stars in Your Little Finger

Stop reading for a moment and look at your little finger. It's full of carbon atoms because carbon is a fundamental chemical building block for life on Earth. Each of those carbon atoms was once inside a red giant star and was fused from helium nuclei in the triple-alpha process. All the carbon on Earth—in you, in the charcoal you use for barbecuing, and in the diamonds you might exchange with a loved one—was "cooked up" by previous generations of stars. How the carbon atoms (and other elements) made their way from inside some of those stars to become part of Earth is something we will discuss in the next chapter. For now, we want to emphasize that our description of stellar evolution is, in a very real sense, the story of our own cosmic "roots"—the history of how our own atoms originated among the stars. We are made of "star-stuff."

Becoming a Giant Again

After the helium flash, the star, having survived the "energy crisis" that followed the end of the main-sequence stage and the exhaustion of the hydrogen fuel at its center, finds its balance again. As the star readjusts to the release of energy from the triple-alpha process in its core, its internal structure changes once more: its surface temperature increases and its overall luminosity decreases. The point that represents the star on the H–R diagram thus moves to a new position to the left of and somewhat below its place as a red giant (Figure 24.28). The star then continues to fuse the helium in its core for a while, returning to the kind of equilibrium between pressure and gravity that characterized the main-sequence stage. During this time, a newly formed carbon nucleus at the center of the star can sometimes be joined by another helium nucleus to produce a nucleus of oxygen—another building block of life.

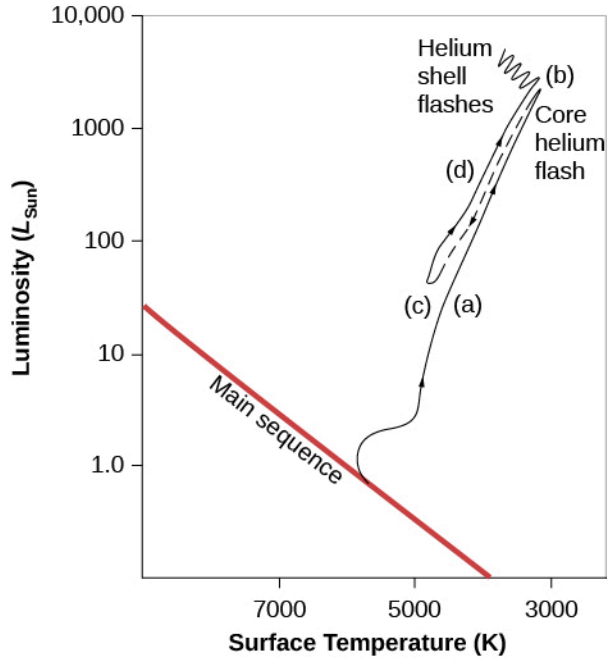


Figure 24.28 Each stage in the star's life is labeled. (a) The star evolves from the main sequence to be a red giant, decreasing in surface temperature and increasing in luminosity. (b) A helium flash occurs, leading to a readjustment of the star's internal structure and to (c) a brief period of stability during which helium is fused to carbon and oxygen in the core (in the process the star becomes hotter and less luminous than it was as a red giant). (d) After the central helium is exhausted, the star becomes a giant again and moves to higher luminosity and lower temperature. By this time, however, the star has exhausted its inner resources and will soon begin to die. Where the evolutionary track becomes a dashed line, the changes are so rapid that they are difficult to model.

However, at a temperature of 100 million K, the inner core is converting its helium fuel to carbon (and a bit of oxygen) at a rapid rate. Thus, the new period of stability cannot last very long: it is far shorter than the main-sequence stage. Soon, all the helium hot enough for fusion will be used up, just like the hot hydrogen that was used up earlier in the star's evolution. Once again, the inner core will not be able to generate energy via fusion. Once more, gravity will take over, and the core will start to shrink again. We can think of stellar evolution as a story of a constant struggle against gravitational collapse. A star can avoid collapsing as long as it can tap energy sources, but once any particular fuel is used up, it starts to collapse again.

The star's situation is analogous to the end of the main-sequence stage (when the central hydrogen got used up), but the star now has a somewhat more complicated structure. Again, the star's core begins to collapse under its own weight. Heat released by the shrinking of the carbon and oxygen core flows into a shell of helium just above the core. This helium, which had not been hot enough for fusion into carbon earlier, is heated just enough for fusion to begin and to generate a new flow of energy.

Farther out in the star, there is also a shell where fresh hydrogen has been heated enough to fuse helium. The star now has a multi-layered structure like an onion: a carbon-oxygen core, surrounded by a shell of helium fusion, a layer of helium, a shell of hydrogen fusion, and finally, the extended outer layers of the star (see **Figure 24.29**). As energy flows outward from the two fusion shells, once again the outer regions of the star begin to expand. Its brief period of stability is over; the star moves back to the red-giant domain on the H–R diagram for a short time (see **Figure 24.28**). But this is a brief and final burst of glory.

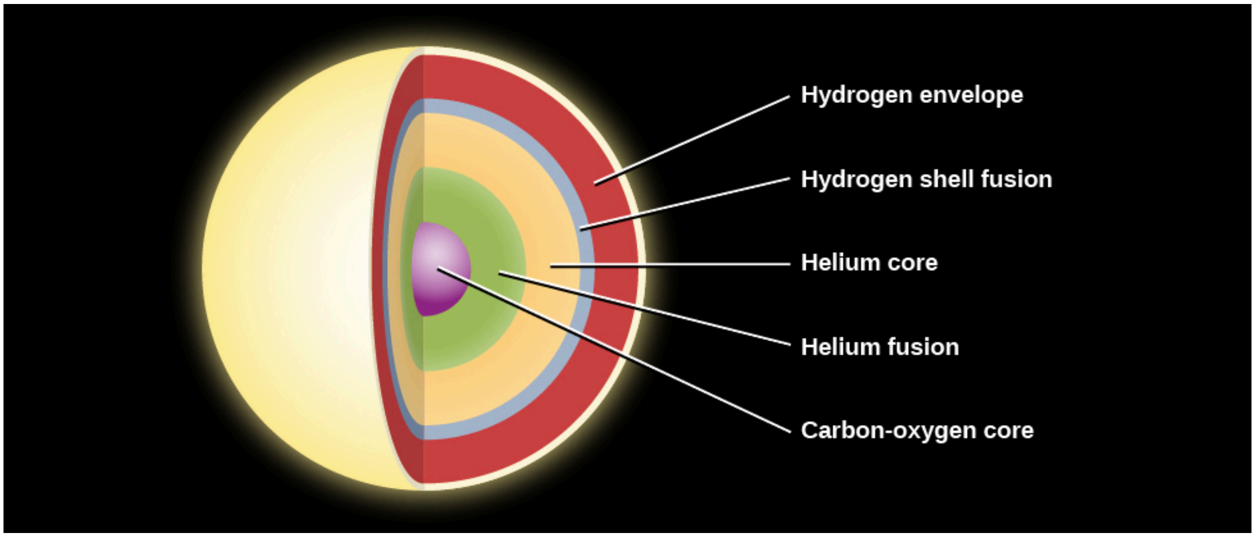


Figure 24.29 Here we see the layers inside a star with an initial mass that is less than twice the mass of the Sun. These include, from the center outward, the carbon-oxygen core, a layer of helium hot enough to fuse, a layer of cooler helium, a layer of hydrogen hot enough to fuse, and then cooler hydrogen beyond.

Recall that the last time the star was in this predicament, helium fusion came to its rescue. The temperature at the star's center eventually became hot enough for the *product* of the previous step of fusion (helium) to become the *fuel* for the next step (helium fusing into carbon). But the step after the fusion of helium nuclei requires a temperature so hot that the kinds of lower-mass stars (less than 2 solar masses) we are discussing simply cannot compress their cores to reach it. No further types of fusion are possible for such a star.

In a star with a mass similar to that of the Sun, the formation of a carbon-oxygen core thus marks the end of the generation of nuclear energy at the center of the star. The star must now confront the fact that its death is near. We will discuss how stars like this end their lives in **The Death of Stars**, but in the meantime, **Table 24.4** summarizes the stages discussed so far in the life of a star with the same mass as that of the Sun. One thing that gives us confidence in our calculations of stellar evolution is that when we make H–R diagrams of older clusters, we actually see stars in each of the stages that we have been discussing.

The Evolution of a Star with the Sun's Mass

Stage	Time in This Stage (years)	Surface Temperature (K)	Luminosity (L_{Sun})	Diameter (Sun = 1)
Main sequence	11 billion	6000	1	1
Becomes red giant	1.3 billion	3100 at minimum	2300 at maximum	165
Helium fusion	100 million	4800	50	10
Giant again	20 million	3100	5200	180

Table 24.4

Mass Loss from Red-Giant Stars and the Formation of Planetary Nebulae

When stars swell up to become red giants, they have very large radii and therefore a low escape velocity.[2] Radiation pressure, stellar pulsations, and violent events like the helium flash can all drive atoms in the outer atmosphere away from the star, and cause it to lose a substantial fraction of its mass into space. Astronomers estimate that by the time a star like the Sun reaches the point of the helium flash, for example, it will have lost as much as 25% of its mass. And it can lose still more mass when it ascends the red-giant branch for the second time. As a result, aging stars are surrounded by one or more expanding shells of gas, each containing as much as 10–20% of the Sun's mass (or 0.1–0.2 M_{Sun}).

When nuclear energy generation in the carbon-oxygen core ceases, the star's core begins to shrink again and to heat up as it gets more and more compressed. (Remember that this compression will not be halted by another type of fusion in these low-mass stars.) The whole star follows along, shrinking and also becoming very hot—reaching surface temperatures as high as 100,000 K. Such hot stars are very strong sources of stellar winds and ultraviolet radiation, which sweep outward into the shells of material ejected when the star was a red giant. The winds and the ultraviolet radiation heat the shells, ionize them, and set them aglow.

The result is the creation of some of the most beautiful objects in the cosmos (see the gallery in **Figure 24.30** and **Figure 24.1**). These objects were given an extremely misleading name when first found in the eighteenth century: **planetary nebulae**. The name is derived from the fact that a few planetary nebulae, when viewed through a small telescope, have a round shape bearing a superficial resemblance to planets. Actually, they have nothing to do with planets, but once names are put into regular use in astronomy, it is extremely difficult to change them. There are tens of thousands of planetary nebulae in our own Galaxy, although many are hidden from view because their light is absorbed by interstellar dust.

2. Recall that the force of gravity depends not only on the mass doing the pulling, but also on our distance from the center of gravity. As a red giant star gets a lot bigger, a point on the surface of the star is now farther from the center, and thus has less gravity. That's why the speed needed to escape the star goes down.

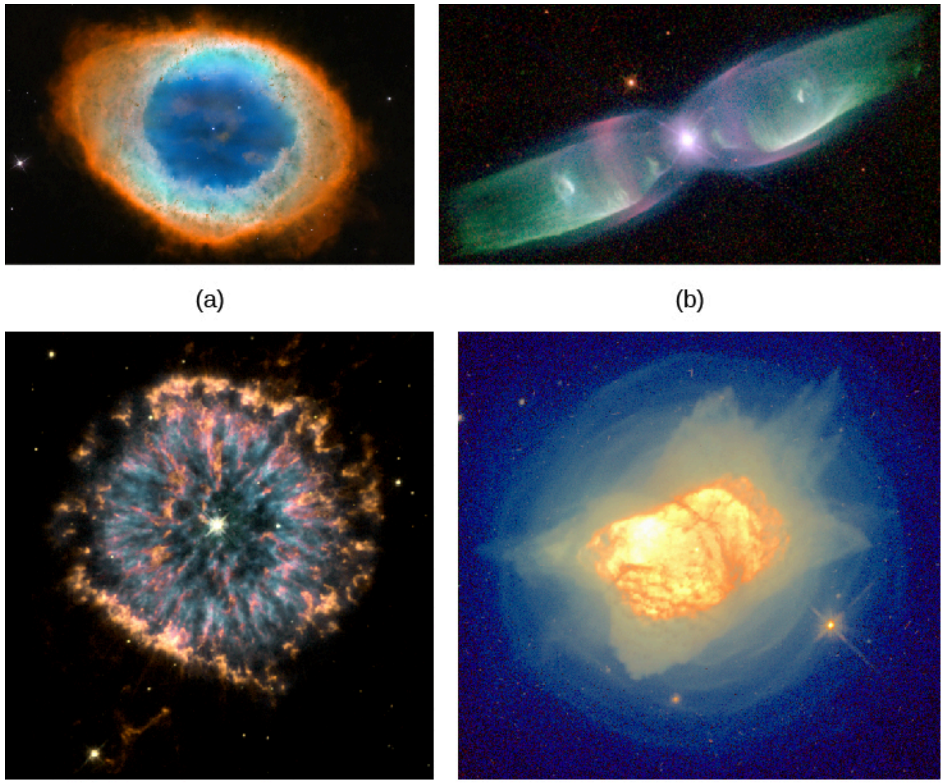


Figure 24.30 This series of beautiful images depicting some intriguing planetary nebulae highlights the capabilities of the Hubble Space Telescope. (a) Perhaps the best known planetary nebula is the Ring Nebula (M57), located about 2000 light-years away in the constellation of Lyra. The ring is about 1 light-year in diameter, and the central star has a temperature of about 120,000 °C. Careful study of this image has shown scientists that, instead of looking at a spherical shell around this dying star, we may be looking down the barrel of a tube or cone. The blue region shows emission from very hot helium, which is located very close to the star; the red region isolates emission from ionized nitrogen, which is radiated by the coolest gas farthest from the star; and the green region represents oxygen emission, which is produced at intermediate temperatures and is at an intermediate distance from the star. (b) This planetary nebula, M2-9, is an example of a butterfly nebula. The central star (which is part of a binary system) has ejected mass preferentially in two opposite directions. In other images, a disk, perpendicular to the two long streams of gas, can be seen around the two stars in the middle. The stellar outburst that resulted in the expulsion of matter occurred about 1200 years ago. Neutral oxygen is shown in red, once-ionized nitrogen in green, and twice-ionized oxygen in blue. The planetary nebula is about 2100 light-years away in the constellation of Ophiuchus. (c) In this image of the planetary nebula NGC 6751, the blue regions mark the hottest gas, which forms a ring around the central star. The orange and red regions show the locations of cooler gas. The origin of these cool streamers is not known, but their shapes indicate that they are affected by radiation and stellar winds from the hot star at the center. The temperature of the star is about 140,000 °C. The diameter of the nebula is about 600 times larger than the diameter of our solar system. The nebula is about 6500 light-years away in the constellation of Aquila. (d) This image of the planetary nebula NGC 7027 shows several stages of mass loss. The faint blue concentric shells surrounding the central region identify the mass that was shed slowly from the surface of the star when it became a red giant. Somewhat later, the remaining outer layers were ejected but not in a spherically symmetric way. The dense clouds formed by this late ejection produce the bright inner regions. The hot central star can be seen faintly near the center of the nebulosity. NGC 7027 is about 3000 light-years away in the direction of the constellation of Cygnus. (credit a: modification of work by NASA, ESA, and the Hubble Heritage (STScI/AURA)-ESA/Hubble Collaboration; credit b: modification of work by Bruce Balick (University of Washington), Vincent Icke (Leiden University, The Netherlands), Garrelt Mellema (Stockholm University), and NASA; credit c: modification of work by NASA, The Hubble Heritage Team (STScI/AURA); credit d: modification of work by H. Bond (STScI) and NASA)

As **Figure 24.30** shows, sometimes a planetary nebula appears to be a simple ring. Others have faint shells surrounding the bright ring, which is evidence that there were multiple episodes of mass loss when the star was a red giant (see image (d) in **Figure 24.30**). In a few cases, we see two lobes of matter flowing in opposite directions. Many astronomers think that a considerable number of planetary nebulae basically consist of the same structure, but that the shape we see depends on the viewing angle (**Figure 24.31**). According to this idea, the dying star is surrounded by a very dense, doughnut-shaped disk of gas. (Theorists do not yet have a definite explanation for why the dying star should produce this ring, but many believe that binary stars, which are common, are involved.)

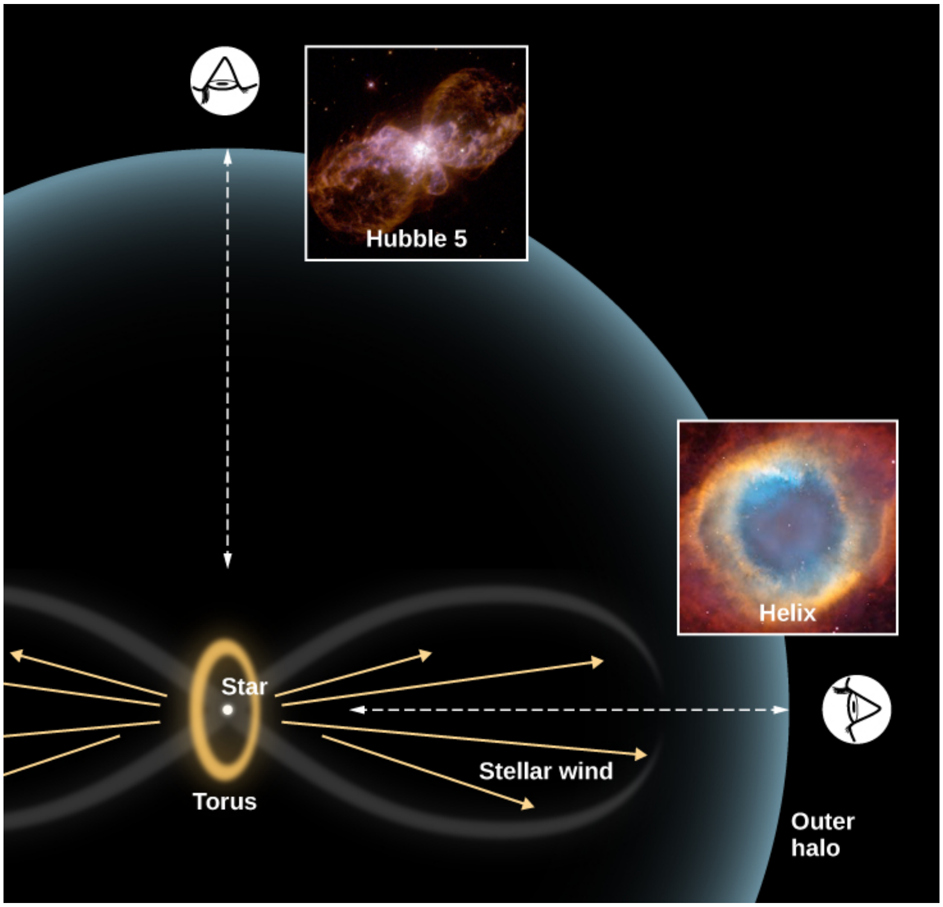


Figure 24.31 The range of different shapes that we see among planetary nebulae may, in many cases, arise from the same geometric shape, but seen from a variety of viewing directions. The basic shape is a hot central star surrounded by a thick torus (or doughnut-shaped disk) of gas. The star's wind cannot flow out into space very easily in the direction of the torus, but can escape more freely in the two directions perpendicular to it. If we view the nebula along the direction of the flow (Helix Nebula), it will appear nearly circular (like looking directly down into an empty ice-cream cone). If we look along the equator of the torus, we see both outflows and a very elongated shape (Hubble 5). Current research on planetary nebulae focuses on the reasons for having a torus around the star in the first place. Many astronomers suggest that the basic cause may be that many of the central stars are actually close binary stars, rather than single stars. (credit "Hubble 5": modification of work by Bruce Balick (University of Washington), Vincent Icke (Leiden University, The Netherlands), Garrelt Mellema (Stockholm University), and NASA/ESA; credit "Helix": modification of work by NASA, ESA, C.R. O'Dell (Vanderbilt University), and M. Meixner, P. McCullough)

As the star continues to lose mass, any less dense gas that leaves the star cannot penetrate the torus, but the gas *can* flow outward in directions perpendicular to the disk. If we look perpendicular to the direction of outflow, we see the disk and both of the outward flows. If we look "down the barrel" and into the flows, we see a ring. At intermediate angles, we may see wonderfully complex structures. Compare the viewpoints in **Figure 24.31** with the images in **Figure 24.30**.

Planetary nebula shells usually expand at speeds of 20–30 km/s, and a typical planetary nebula has a diameter of about

1 light-year. If we assume that the gas shell has expanded at a constant speed, we can calculate that the shells of all the planetary nebulae visible to us were ejected within the past 50,000 years at most. After this amount of time, the shells have expanded so much that they are too thin and tenuous to be seen. That's a pretty short time that each planetary nebula can be observed (when compared to the whole lifetime of the star). Given the number of such nebulae we nevertheless see, we must conclude that a large fraction of all stars evolve through the planetary nebula phase. Since we saw that low-mass stars are much more common than high-mass stars, this confirms our view of planetary nebulae as sort of "last gasp" of low-mass star evolution.

Cosmic Recycling

The loss of mass by dying stars is a key step in the gigantic cosmic recycling scheme. Remember that stars form from vast clouds of gas and dust. As they end their lives, stars return part of their gas to the galactic reservoirs of raw material. Eventually, some of the expelled material from aging stars will participate in the formation of new star systems.

However, the atoms returned to the Galaxy by an aging star are not necessarily the same ones it received initially. The star, after all, has fused hydrogen and helium to form new elements over the course of its life. And during the red-giant stage, material from the star's central regions is dredged up and mixed with its outer layers, which can cause further nuclear reactions and the creation of still more new elements. As a result, the winds that blow outward from such stars include atoms that were "newly minted" inside the stars' cores. (As we will see, this mechanism is even more effective for high-mass stars, but it does work for stars with masses like that of the Sun.) In this way, the raw material of the Galaxy is not only resupplied but also receives infusions of new elements. You might say this cosmic recycling plan allows the universe to get more "interesting" all the time.

The Red Giant Sun and the Fate of Earth

How will the evolution of the Sun affect conditions on Earth in the future? Although the Sun has appeared reasonably steady in size and luminosity over recorded human history, that brief span means nothing compared with the timescales we have been discussing. Let's examine the long-term prospects for our planet.

The Sun took its place on the zero-age main sequence approximately 4.5 billion years ago. At that time, it emitted only about 70% of the energy that it radiates today. One might expect that Earth would have been a lot colder than it is now, with the oceans frozen solid. But if this were the case, it would be hard to explain why simple life forms existed when Earth was less than a billion years old. Scientists now think that the explanation may be that much more carbon dioxide was present in Earth's atmosphere when it was young, and that a much stronger greenhouse effect kept Earth warm. (In the greenhouse effect, gases like carbon dioxide or water vapor allow the Sun's light to come in but do not allow the infrared radiation from the ground to escape back into space, so the temperature near Earth's surface increases.)

Carbon dioxide in Earth's atmosphere has steadily declined as the Sun has increased in luminosity. As the brighter Sun increases the temperature of Earth, rocks weather faster and react with carbon dioxide, removing it from the atmosphere. The warmer Sun and the weaker greenhouse effect have kept Earth at a nearly constant temperature for most of its life. This remarkable coincidence, which has resulted in fairly stable climatic conditions, has been the key in the development of complex life-forms on our planet.

As a result of changes caused by the buildup of helium in its core, the Sun will continue to increase in luminosity as it grows older, and more and more radiation will reach Earth. For a while, the amount of carbon dioxide will continue to decrease. (Note that this effect counteracts increases in carbon dioxide from human activities, but on a much-too-slow timescale to undo the changes in climate that are likely to occur in the next 100 years.)

Eventually, the heating of Earth will melt the polar caps and increase the evaporation of the oceans. Water vapor is also an efficient greenhouse gas and will more than compensate for the decrease in carbon dioxide. Sooner or later (atmospheric models are not yet good enough to say exactly when, but estimates range from 500 million to 2 billion years), the increased water vapor will cause a runaway greenhouse effect.

About 1 billion years from now, Earth will lose its water vapor. In the upper atmosphere, sunlight will break down water vapor into hydrogen, and the fast-moving hydrogen atoms will escape into outer space. Like Humpty Dumpty, the water molecules cannot be put back together again. Earth will start to resemble the Venus of today, and temperatures will become much too high for life as we know it.

All of this will happen before the Sun even becomes a red giant. Then the bad news really starts. The Sun, as it expands, will swallow Mercury and Venus, and friction with our star's outer atmosphere will make these planets spiral inward until they are completely vaporized. It is not completely clear whether Earth will escape a similar fate. As described in this chapter, the Sun will lose some of its mass as it becomes a red giant. The gravitational pull of the Sun decreases when it loses mass. The result would be that the diameter of Earth's orbit would increase (remember Kepler's third law). However, recent calculations also show that forces due to the tides raised on the Sun by Earth will act in the opposite direction, causing Earth's orbit to shrink. Thus, many astrophysicists conclude that Earth will be vaporized along with Mercury and Venus. Whether or not this dire prediction is true, there is little doubt that all life on Earth will surely be incinerated. But don't lose any sleep over this—we are talking about events that will occur billions of years from now.

What then are the prospects for preserving Earth life as we know it? The first strategy you might think of would be to move humanity to a more distant and cooler planet. However, calculations indicate that there are long periods of time (several hundred million years) when no planet is habitable. For example, Earth becomes far too warm for life long before Mars warms up enough.

A better alternative may be to move the entire Earth progressively farther from the Sun. The idea is to use gravity in the same way NASA has used it to send spacecraft to distant planets. When a spacecraft flies near a planet, the planet's motion can be used to speed up the spacecraft, slow it down, or redirect it. Calculations show that if we were to redirect an asteroid so that it follows just the right orbit between Earth and Jupiter, it could transfer orbital energy from Jupiter to Earth and move Earth slowly outward, pulling us away from the expanding Sun on each flyby. Since we have hundreds of millions of years to change Earth's orbit, the effect of each flyby need not be large. (Of course, the people directing the asteroid had better get the orbit exactly right and not cause the asteroid to hit Earth.)

It may seem crazy to think about projects to move an entire planet to a different orbit. But remember that we are talking about the distant future. If, by some miracle, human beings are able to get along for all that time and don't blow ourselves to bits, our technology is likely to be far more sophisticated than it is today. It may also be that if humans survive for hundreds of millions of years, we may spread to planets or habitats around other stars. Indeed, Earth, by then, might be a museum world to which youngsters from other planets return to learn about the origin of our species. It is also possible that evolution will by then have changed us in ways that allow us to survive in very different environments. Wouldn't it be exciting to see how the story of the story of the human race turns out after all those billions of years?

24.7 | The Evolution of More Massive Stars

Learning Objectives

By the end of this section, you will be able to:

- Explain how and why massive stars evolve much more rapidly than lower-mass stars like our Sun
- Discuss the origin of the elements heavier than carbon within stars

If what we have described so far were the whole story of the evolution of stars and elements, we would have a big problem on our hands. We will see in later chapters that in our best models of the first few minutes of the universe, everything starts with the two simplest elements—hydrogen and helium (plus a tiny bit of lithium). All the predictions of the models imply that no heavier elements were produced at the beginning of the universe. Yet when we look around us on Earth, we see lots of other elements besides hydrogen and helium. These elements must have been made (fused) somewhere in the universe, *and the only place hot enough to make them is inside stars*. One of the fundamental discoveries of twentieth-century astronomy is that the stars are the source of most of the chemical richness that characterizes our world and our lives.

We have already seen that carbon and some oxygen are manufactured inside the lower-mass stars that become red giants. But where do the heavier elements we know and love (such as the silicon and iron inside Earth, and the gold and silver in our jewelry) come from? The kinds of stars we have been discussing so far never get hot enough at their centers to make these elements. It turns out that such heavier elements can be formed only late in the lives of *more massive* stars.

Making New Elements in Massive Stars

Massive stars evolve in much the same way that the Sun does (but always more quickly)—up to the formation of a carbon-oxygen core. One difference is that for stars with more than about twice the mass of the Sun, helium begins fusion more gradually, rather than with a sudden flash. Also, when more massive stars become red giants, they become so bright and large that we call them *supergiants*. Such stars can expand until their outer regions become as large as the orbit of Jupiter, which is precisely what the Hubble Space Telescope has shown for the star Betelgeuse (see **Figure 24.15**). They also lose mass very effectively, producing dramatic winds and outbursts as they age. **Figure 24.32** shows a wonderful image of the very massive star Eta Carinae, with a great deal of ejected material clearly visible.

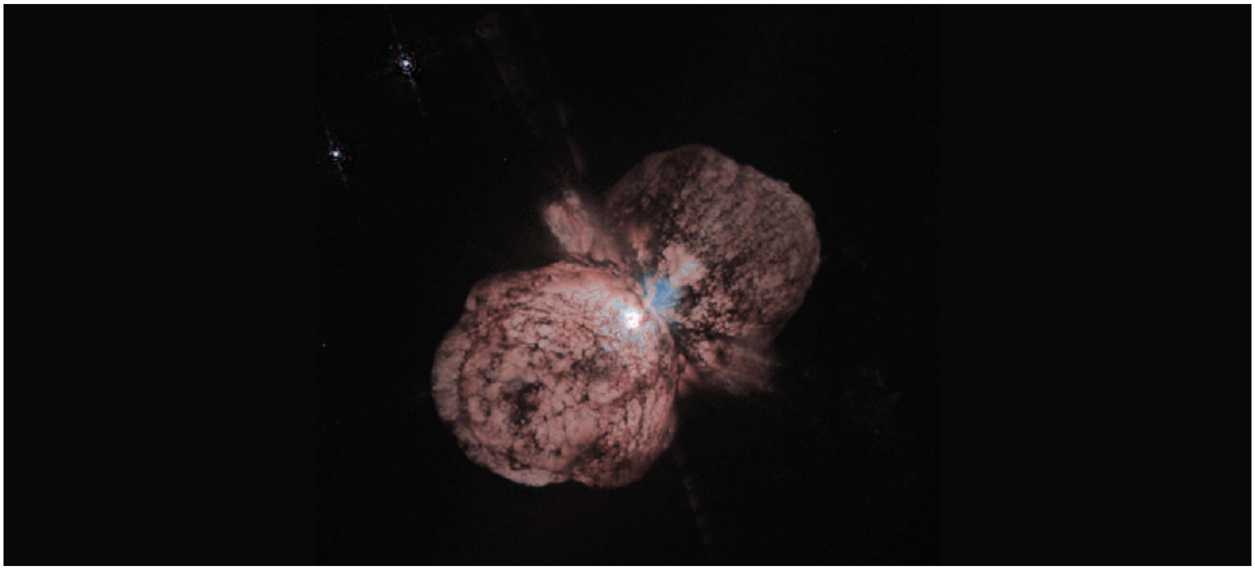


Figure 24.32 With a mass at least 100 times that of the Sun, the hot supergiant Eta Carinae is one of the most massive stars known. This Hubble Space Telescope image records the two giant lobes and equatorial disk of material it has ejected in the course of its evolution. The pink outer region is material ejected in an outburst seen in 1843, the largest of such mass loss event that any star is known to have survived. Moving away from the star at a speed of about 1000 km/s, the material is rich in nitrogen and other elements formed in the interior of the star. The inner blue-white region is the material ejected at lower speeds and is thus still closer to the star. It appears blue-white because it contains dust and reflects the light of Eta Carinae, whose luminosity is 4 million times that of our Sun. (credit: modification of work by Jon Morse (University of Colorado) & NASA)

But the crucial way that massive stars diverge from the story we have outlined is that they can start additional kinds of fusion in their centers and in the shells surrounding their central regions. The outer layers of a star with a mass greater than about 8 solar masses have a weight that is enough to compress the carbon-oxygen core until it becomes hot enough to ignite fusion of carbon nuclei. Carbon can fuse into still more oxygen, and at still higher temperatures, oxygen and then neon, magnesium, and finally silicon can build even heavier elements. Iron is, however, the endpoint of this process. The fusion of iron atoms produces products that are *more* massive than the nuclei that are being fused and therefore the process *requires* energy, as opposed to releasing energy, which all fusion reactions up to this point have done. This required energy comes at the expense of the star itself, which is now on the brink of death (**Figure 24.33**). What happens next will be described in the chapter on **The Death of Stars**.

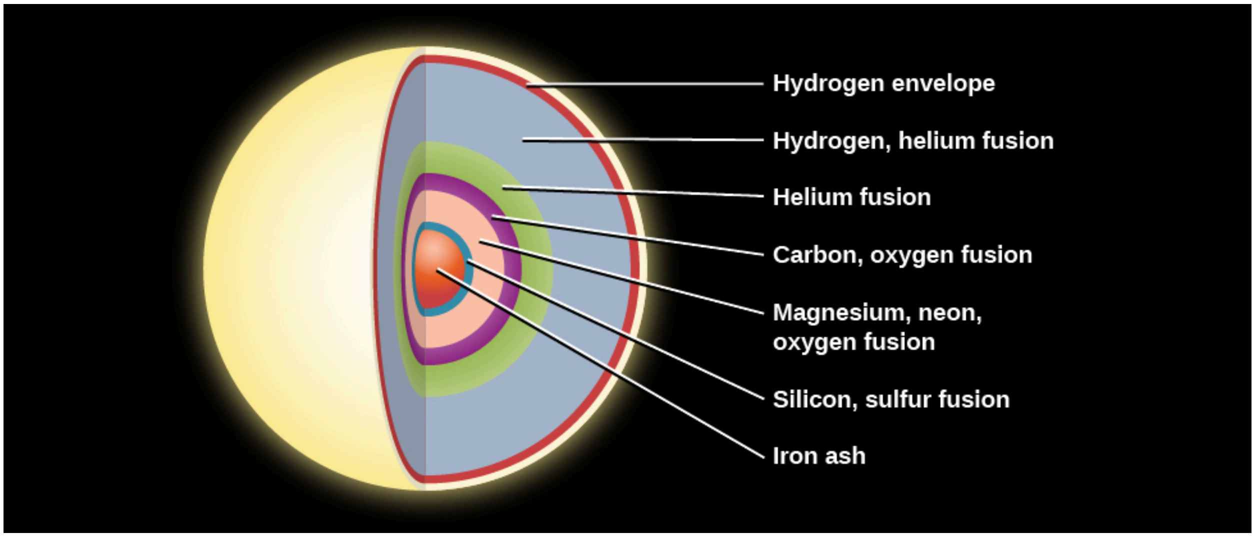


Figure 24.33 High-mass stars can fuse elements heavier than carbon. As a massive star nears the end of its evolution, its interior resembles an onion. Hydrogen fusion is taking place in an outer shell, and progressively heavier elements are undergoing fusion in the higher-temperature layers closer to the center. All of these fusion reactions generate energy and enable the star to continue shining. Iron is different. The fusion of iron requires energy, and when iron is finally created in the core, the star has only minutes to live.

Physicists have now found nuclear pathways whereby virtually all chemical elements of atomic weights up to that of iron can be built up by this **nucleosynthesis** (the making of new atomic nuclei) in the centers of the more massive red giant stars. This still leaves the question of where elements *heavier* than iron come from. We will see in the next chapter that when massive stars finally exhaust their nuclear fuel, they most often die in a spectacular explosion—a supernova. Heavier elements can be synthesized in the stunning violence of such explosions.

Not only can we explain in this way where the elements that make up our world and others come from, but our theories of nucleosynthesis inside stars are even able to predict the *relative abundances* with which the elements occur in nature. The way stars build up elements during various nuclear reactions really can explain why some elements (oxygen, carbon, and iron) are common and others are quite rare (gold, silver, and uranium).

Elements in Globular Clusters and Open Clusters Are Not the Same

The fact that the elements are made in stars over time explains an important difference between globular and open clusters. Hydrogen and helium, which are the most abundant elements in stars in the solar neighborhood, are also the most abundant constituents of stars in both kinds of clusters. However, the abundances of the elements *heavier* than helium are very different.

In the Sun and most of its neighboring stars, the combined abundance (by mass) of the elements heavier than hydrogen and helium is 1–4% of the star’s mass. Spectra show that most open-cluster stars also have 1–4% of their matter in the form of heavy elements. Globular clusters, however, are a different story. The heavy-element abundance of stars in typical globular clusters is found to be only 1/10 to 1/100 that of the Sun. A few very old stars not in clusters have been discovered with even lower abundances of heavy elements.

The differences in chemical composition are a direct consequence of the formation of a cluster of stars. The very first generation of stars initially contained only hydrogen and helium. We have seen that these stars, in order to generate energy, created heavier elements in their interiors. In the last stages of their lives, they ejected matter, now enriched in heavy elements, into the reservoirs of raw material between the stars. Such matter was then incorporated into a new generation of stars.

This means that the relative abundance of the heavy elements must be less and less as we look further into the past. We saw that the globular clusters are much older than the open clusters. Since globular-cluster stars formed much earlier (that is, they are an earlier generation of stars) than those in open clusters, they have only a relatively small abundance of elements heavier than hydrogen and helium.

As time passes, the proportion of heavier elements in the “raw material” that makes new stars and planets increases. This means that the first generation of stars that formed in our Galaxy would not have been accompanied by a planet like Earth, full of silicon, iron, and many other heavy elements. Earth (and the astronomy students who live on it) was possible only after generations of stars had a chance to make and recycle their heavier elements.

Now the search is on for true *first*-generation stars, made only of hydrogen and helium. Theories predict that such stars should be very massive, live fast, and die quickly. They should have lived and died long ago. The place to look for them is in very distant galaxies that formed when the universe was only a few hundred million years old, but whose light is only arriving at Earth now.

Approaching Death

Compared with the main-sequence lifetimes of stars, the events that characterize the last stages of stellar evolution pass very quickly (especially for massive stars). As the star's luminosity increases, its rate of nuclear fuel consumption goes up rapidly—just at that point in its life when its fuel supply is beginning to run down.

After the prime fuel—hydrogen—is exhausted in a star's core, we saw that other sources of nuclear energy are available to the star in the fusion of, first, helium, and then of other more complex elements. But the energy yield of these reactions is much less than that of the fusion of hydrogen to helium. And to trigger these reactions, the central temperature must be higher than that required for the fusion of hydrogen to helium, leading to even more rapid consumption of fuel. Clearly this is a losing game, and very quickly the star reaches its end. As it does so, however, some remarkable things can happen, as we will see in **The Death of Stars**.

For Further Exploration

Websites

Formation of Stars: **https://www.spacetelescope.org/science/formation_of_stars/ (https://www.spacetelescope.org/science/formation_of_stars/)** . Star Formation page from the Hubble Space Telescope, with links to images and information.

BBC Page on Giant Stars: **http://www.bbc.co.uk/science/space/universe/sights/giant_stars (http://www.bbc.co.uk/science/space/universe/sights/giant_stars)** . Includes basic information and links to brief video excerpts.

Encylopedia Brittanica Article on Star Clusters: **http://www.britannica.com/topic/star-cluster (http://www.britannica.com/topic/star-cluster)** . Written by astronomer Helen Sawyer Hogg-Priestley.

Hubble Image Gallery: Planetary Nebulae: **http://hubblesite.org/gallery/album/nebula/planetary/ (http://hubblesite.org/gallery/album/nebula/planetary/)** . Click on each image to go to a page with more information available. (See also a similar gallery at the National Optical Astronomy Observatories: **https://www.noao.edu/image_gallery/planetary_nebulae.html (https://www.noao.edu/image_gallery/planetary_nebulae.html)**).

Hubble Image Gallery: Star Clusters: **http://hubblesite.org/gallery/album/star/star_cluster/ (http://hubblesite.org/gallery/album/star/star_cluster/)** . Each image comes with an explanatory caption when you click on it. (See also a similar European Southern Observatory Gallery at: **https://www.eso.org/public/images/archive/category/starclusters/ (https://www.eso.org/public/images/archive/category/starclusters/)**).

Measuring the Age of a Star Cluster: **https://www.e-education.psu.edu/astro801/content/l7_p6.html (https://www.e-education.psu.edu/astro801/content/l7_p6.html)** . From Penn State.

Videos

A Star Is Born: **http://www.discovery.com/tv-shows/other-shows/videos/how-the-universe-works-a-star-is-born/ (http://www.discovery.com/tv-shows/other-shows/videos/how-the-universe-works-a-star-is-born/)** . Discovery Channel video with astronomer Michelle Thaller (2:25).

Life Cycle of Stars: **https://www.youtube.com/watch?v=PM9CQDlQI0A (https://www.youtube.com/watch?v=PM9CQDlQI0A)** . Short summary of stellar evolution from the Institute of Physics in Great Britain, with astronomer Tim O'Brien (4:58).

Missions Take an Unparalleled Look into Superstar Eta Carinae: **https://www.youtube.com/watch?v=0rJQi6oaZf0 (https://www.youtube.com/watch?v=0rJQi6oaZf0)** . NASA Goddard video about observations in 2014 and what we know about the pair of stars in this complicated system (4:00).

Star Clusters: Open and Globular Clusters: **https://www.youtube.com/watch?v=rGPRLxrYbYA (https://www.youtube.com/watch?v=rGPRLxrYbYA)** . Three Short Hubblecast Videos from 2007–2008 on discoveries involving star clusters (12:24).

Tour of Planetary Nebula NGC 5189: **https://www.youtube.com/watch?v=1D2cwiZld0o (https://www.youtube.com/watch?v=1D2cwiZld0o)** . Brief Hubblecast episode with Joe Liske, explaining planetary nebulae in general and one example in particular (5:22).

CHAPTER 24 REVIEW

KEY TERMS

association a loose group of young stars whose spectral types, motions, and positions in the sky indicate a common origin

distance modulus the numerical difference between the apparent visual magnitude of an object and its absolute visual magnitude, i.e. m-M

giant molecular clouds large, cold interstellar clouds with diameters of dozens of light-years and typical masses of 10^5 solar masses; found in the spiral arms of galaxies, these clouds are where stars form

globular cluster one of about 150 large, spherical star clusters (each with hundreds of thousands of stars) that form a system of clusters in the center of our Galaxy

helium flash a nearly explosive ignition of helium in the triple-alpha process in the dense core of a red giant star

Herbig-Haro (HH) object luminous knots of gas in an area of star formation that are set to glow by jets of material from a protostar

main-sequence turnoff location in the H–R diagram where stars begin to leave the main sequence

nucleosynthesis the building up of heavy elements from lighter ones by nuclear fusion

open cluster a comparatively loose cluster of stars, containing from a few dozen to a few thousand members, located in the spiral arms or disk of our Galaxy; sometimes referred to as a galactic cluster

planetary nebula a shell of gas ejected by and expanding away from an extremely hot low-mass star that is nearing the end of its life (the nebulae glow because of the ultra-violet energy of the central star)

protostar a very young star still in the process of formation, before nuclear fusion begins

stellar wind the outflow of gas, sometimes at speeds as high as hundreds of kilometers per second, from a star

triple-alpha process a nuclear reaction by which three helium nuclei are built up (fused) into one carbon nucleus

zero-age main sequence a line denoting the main sequence on the H–R diagram for a system of stars that have completed their contraction from interstellar matter and are now deriving all their energy from nuclear reactions, but whose chemical composition has not yet been altered substantially by nuclear reactions

KEY EQUATIONS

Jeans mass $$M_{\text{Jeans}} = 18M_{\text{Sun}}\sqrt{\frac{T^3}{n}}$$

Spectroscopic parallax $$d = 10\times10^{0.2(m-M)}$$

SUMMARY

24.1 Star Formation

- Most stars form in giant molecular clouds with masses as large as 3×10^6 solar masses.
- The most well-studied molecular cloud is Orion, where star formation is currently taking place.
- Molecular clouds typically contain regions of higher density called clumps, which in turn contain several even-denser cores of gas and dust, each of which may become a star.
- A star can form inside a core if its density is high enough that gravity can overwhelm the internal pressure and cause the gas and dust to collapse.
- The accumulation of material halts when a protostar develops a strong stellar wind, leading to jets of material being observed coming from the star.
- These jets of material can collide with the material around the star and produce regions that emit light that are known

as Herbig-Haro objects.

24.2 The H–R Diagram and the Study of Stellar Evolution

- The evolution of a star can be described in terms of changes in its temperature and luminosity, which can best be followed by plotting them on an H-R diagram.
- Protostars generate energy (and internal heat) through gravitational contraction that typically continues for millions of years, until the star reaches the main sequence.

24.3 Evolution from the Main Sequence to Red Giants

- When stars first begin to fuse hydrogen to helium, they lie on the zero-age main sequence.
- The amount of time a star spends in the main-sequence stage depends on its mass.
- More massive stars complete each stage of evolution more quickly than lower-mass stars.
- The fusion of hydrogen to form helium changes the interior composition of a star, which in turn results in changes in its temperature, luminosity, and radius.
- Eventually, as stars age, they evolve away from the main sequence to become red giants or supergiants.
- The core of a red giant is contracting, but the outer layers are expanding as a result of hydrogen fusion in a shell outside the core.
- The star gets larger, redder, and more luminous as it expands and cools.

24.4 Star Clusters

- Star clusters provide one of the best tests of our calculations of what happens as stars age.
- The stars in a given cluster were formed at about the same time and have the same composition, so they differ mainly in mass, and thus, in their life stage.
- There are three types of star clusters: globular, open, and associations.
- Globular clusters have diameters of 50–450 light-years, contain hundreds of thousands of stars, and are distributed in a halo around the Galaxy.
- Open clusters typically contain hundreds of stars, are located in the plane of the Galaxy, and have diameters less than 30 light-years.
- Associations are found in regions of gas and dust and contain extremely young stars.

24.5 Checking Out the Theory

- The H–R diagram of stars in a cluster changes systematically as the cluster grows older.
- The most massive stars evolve most rapidly.
- In the youngest clusters and associations, highly luminous blue stars are on the main sequence; the stars with the lowest masses lie to the right of the main sequence and are still contracting toward it.
- With passing time, stars of progressively lower masses evolve away from (or turn off) the main sequence.
- In globular clusters, which are all at least 11 billion years old, there are no luminous blue stars at all.
- Astronomers can use the turnoff point from the main sequence to determine the age of a cluster.
- Spectroscopic parallax can be used to determine the distance of a star cluster from Earth.

24.6 Further Evolution of Stars

- After stars become red giants, their cores eventually become hot enough to produce energy by fusing helium to form carbon (and sometimes a bit of oxygen.)
- The fusion of three helium nuclei produces carbon through the triple-alpha process.
- The rapid onset of helium fusion in the core of a low-mass star is called the helium flash.

- After this, the star becomes stable and reduces its luminosity and size briefly.
- In stars with masses about twice the mass of the Sun or less, fusion stops after the helium in the core has been exhausted.
- Fusion of hydrogen and helium in shells around the contracting core makes the star a bright red giant again, but only temporarily.
- When the star is a red giant, it can shed its outer layers and thereby expose hot inner layers.
- Planetary nebulae (which have nothing to do with planets) are shells of gas ejected by such stars, set glowing by the ultraviolet radiation of the dying central star.

24.7 The Evolution of More Massive Stars

- In stars with masses higher than about 8 solar masses, nuclear reactions involving carbon, oxygen, and still heavier elements can build up nuclei as heavy as iron.
- The creation of new chemical elements is called nucleosynthesis.
- The late stages of evolution occur very quickly.
- Ultimately, all stars must use up all of their available energy supplies.
- In the process of dying, most stars eject some matter, enriched in heavy elements, into interstellar space where it can be used to form new stars.
- Each succeeding generation of stars therefore contains a larger proportion of elements heavier than hydrogen and helium.
- This progressive enrichment explains why the stars in open clusters (which formed more recently) contain more heavy elements than do those in ancient globular clusters, and it tells us where most of the atoms on Earth and in our bodies come from.

CONCEPTUAL QUESTIONS

24.1 Star Formation

1. Give several reasons the Orion molecular cloud is such a useful "laboratory" for studying the stages of star formation.

2. Why is star formation more likely to occur in cold molecular clouds than in regions where the temperature of the interstellar medium is several hundred thousand degrees?

3. Why have we learned a lot about star formation since the invention of detectors sensitive to infrared radiation?

4. Describe what happens when a star forms. Begin with a dense core of material in a molecular cloud and trace the evolution up to the time the newly formed star reaches the main sequence.

5. Describe how the T Tauri star stage in the life of a low-mass star can lead to the formation of a Herbig-Haro (H-H) object.

6. A friend of yours who did not do well in her astronomy class tells you that she believes all stars are old and none could possibly be born today. What arguments would you use to persuade her that stars are being born somewhere in the Galaxy during your lifetime?

24.2 The H–R Diagram and the Study of Stellar Evolution

7. Look at the four stages shown in **Figure 24.8**. In which stage(s) can we see the star in visible light? In infrared radiation?

8. The evolutionary track for a star of 1 solar mass remains nearly vertical in the H–R diagram for a while (see **Figure 24.12**). How is its luminosity changing during this time? Its temperature? Its radius?

9. Two protostars, one 10 times the mass of the Sun and one half the mass of the Sun are born at the same time in a molecular cloud. Which one will be first to reach the main sequence stage, where it is stable and getting energy from fusion?

24.3 Evolution from the Main Sequence to Red Giants

10. What is the first event that happens to a star with roughly the mass of our Sun that exhausts the hydrogen in its core and stops the generation of energy by the nuclear fusion of hydrogen to helium? Describe the sequence of events that the star undergoes.

11. Astronomers find that 90% of the stars observed in the sky are on the main sequence of an H–R diagram; why does this make sense? Why are there far fewer stars in the giant and supergiant region?

12. Describe the evolution of a star with a mass similar to that of the Sun, from the protostar stage to the time it first becomes a red giant. Give the description in words and then sketch the evolution on an H–R diagram.

13. On which edge of the main sequence band on an H–R diagram would the zero-age main sequence be?

14. Certain stars, like Betelgeuse, have a lower surface temperature than the Sun and yet are more luminous. How do these stars produce so much more energy than the Sun?

15. Is the Sun on the zero-age main sequence? Explain your answer.

16. Which of the planets in our solar system have orbits that are smaller than the photospheric radius of Betelgeuse listed in in **Table 24.2**?

24.4 Star Clusters

17. Why are star clusters so useful for astronomers who want to study the evolution of stars?

18. Would the Sun more likely have been a member of a globular cluster or open cluster in the past?

19. Suppose a star cluster were at such a large distance that it appeared as an unresolved spot of light through the telescope. What would you expect the overall color of the spot to be if it were the image of the cluster immediately after it was formed? How would the color differ after 10^{10} years? Why?

24.5 Checking Out the Theory

20. Explain how an H–R diagram of the stars in a cluster can be used to determine the age of the cluster.

21. In the H–R diagrams for some young clusters, stars of both very low and very high luminosity are off to the right of the main sequence, whereas those of intermediate luminosity are on the main sequence. Can you offer an explanation for that? Sketch an H–R diagram for such a cluster.

22. Stars that have masses approximately 0.8 times the mass of the Sun take about 18 billion years to turn into red giants. How does this compare to the current age of the universe? Would you expect to find a globular cluster with a main-sequence turnoff for stars of 0.8 solar mass or less? Why or why not?

24.6 Further Evolution of Stars

23. Describe the evolution of a star with a mass similar to that of the Sun, from just after it first becomes a red giant to the time it exhausts the last type of fuel its core is capable of fusing.

24. A star is often described as “moving” on an H–R diagram; why is this description used and what is actually happening with the star?

25. The nuclear process for fusing helium into carbon is often called the “triple-alpha process.” Why is it called as such, and why must it occur at a much higher temperature than the nuclear process for fusing hydrogen into helium?

26. Pictures of various planetary nebulae show a variety of shapes, but astronomers believe a majority of planetary nebulae have the same basic shape. How can this paradox be explained?

27. Where did the carbon atoms in the trunk of a tree on your college campus come from originally? Where did the neon in the fabled “neon lights of Broadway” come from originally?

28. What is a planetary nebula? Will we have one around the Sun?

29. How are planetary nebulae comparable to a fluorescent light bulb in your classroom?

24.7 The Evolution of More Massive Stars

30. Give several reasons the Orion molecular cloud is such a useful “laboratory” for studying the stages of star formation.

31. Why is star formation more likely to occur in cold molecular clouds than in regions where the temperature of the interstellar medium is several hundred thousand degrees?

32. Why have we learned a lot about star formation since the invention of detectors sensitive to infrared radiation?

33. Describe what happens when a star forms. Begin with a dense core of material in a molecular cloud and trace the evolution up to the time the newly formed star reaches the main sequence.

34. Describe how the T Tauri star stage in the life of a low-mass star can lead to the formation of a Herbig-Haro (H-H) object.

35. Look at the four stages shown in **Figure 24.8**. In which stage(s) can we see the star in visible light? In infrared radiation?

36. The evolutionary track for a star of 1 solar mass remains nearly vertical in the H–R diagram for a while (see **Figure 24.12**). How is its luminosity changing during this time? Its temperature? Its radius?

37. Two protostars, one 10 times the mass of the Sun and one half the mass of the Sun are born at the same time in a molecular cloud. Which one will be first to reach the main sequence stage, where it is stable and getting energy from fusion?

38. A friend of yours who did not do well in her astronomy class tells you that she believes all stars are old and none could possibly be born today. What arguments would you use to persuade her that stars are being born somewhere in the Galaxy during your lifetime?

39. Compare the following stages in the lives of a human being and a star: prenatal, birth, adolescence/adulthood, middle age, old age, and death. What does a star with the mass of our Sun do in each of these stages?

40. How do stars typically “move” through the main sequence band on an H–R diagram? Why?

41. Gravity always tries to collapse the mass of a star toward its center. What mechanism can oppose this gravitational collapse for a star? During what stages of a star’s life would there be a “balance” between them?

42. Suppose you were handed two H–R diagrams for two different clusters: diagram A has a majority of its stars plotted on the upper left part of the main sequence with the rest of the stars off the main sequence; and diagram B has a majority of its stars plotted on the lower right part of the main sequence with the rest of the stars off the main sequence. Which diagram would be for the older cluster? Why?

43. Referring to the H–R diagrams in **Exercise 24.42**, which diagram would more likely be the H–R diagram for an association?

44. Describe the two “recycling” mechanisms that are associated with stars (one during each star’s life and the other connecting generations of stars).

45. In which of these star groups would you mostly likely find the least heavy-element abundance for the stars within them: open clusters, globular clusters, or associations?

46. Would you expect to find an earthlike planet (with a solid surface) around a very low-mass star that formed right at the beginning of a globular cluster’s life? Explain.

47. If the Sun were a member of the cluster NGC 2264, would it be on the main sequence yet? Why or why not?

48. If all the stars in a cluster have nearly the same age, why are clusters useful in studying evolutionary effects (different stages in the lives of stars)?

49. Suppose an astronomer known for joking around told you she had found a type-O main-sequence star in our Milky Way Galaxy that contained no elements heavier than helium. Would you believe her? Why?

50. Automobiles are often used as an analogy to help people better understand how more massive stars have much shorter main-sequence lifetimes compared to less massive stars. Can you explain such an analogy using automobiles?

PROBLEMS

24.2 The H–R Diagram and the Study of Stellar Evolution

51. The computer models of the earliest stars in the universe show temperatures of about 200 K and molecular densities of about 300,000 particles per cm^3. How massive would such a cloud need to be in order to produce a star?

24.3 Evolution from the Main Sequence to Red Giants

52. The text says a star does not change its mass very much during the course of its main-sequence lifetime. While it is on the main sequence, a star converts about 10% of the hydrogen initially present into helium (remember it’s only the core of the star that is hot enough for fusion). Look in earlier chapters to find out what percentage of the hydrogen mass involved in fusion is lost because it is converted to energy. By how much does the mass of the whole star change as a result of fusion? Were we correct to say that the mass of a star does not change significantly while it is on the main sequence?

53. The text explains that massive stars have shorter lifetimes than low-mass stars. Even though massive stars have more fuel to burn, they use it up faster than low-mass stars. You can check and see whether this statement is true. The lifetime of a star is directly proportional to the amount of mass (fuel) it contains and inversely proportional to the rate at which it uses up that fuel (i.e., to its luminosity). Since the lifetime of the Sun is about 10^{10} y, we have the following relationship:

$$T = 10^{10}\frac{M}{L}\text{ y}$$

where T is the lifetime of a main-sequence star, M is its mass measured in terms of the mass of the Sun, and L is its luminosity measured in terms of the Sun's luminosity.

A. Explain in words why this equation works.
B. Use the data in **Table 22.5** to calculate the ages of the main-sequence stars listed.
C. Do low-mass stars have longer main-sequence lifetimes?
D. Do you get the same answers as those in **Table 24.1**?

54. If star A has a core temperature T, and star B has a core temperature $3T$, how does the rate of fusion of star A compare to the rate of fusion of star B?

24.5 Checking Out the Theory

55. You can use the equation in **Exercise 24.53** to estimate the approximate ages of the clusters in **Figure 24.21**, **Figure 24.23**, and **Figure 24.24**. Use the information in the figures to determine the luminosity of the most massive star still on the main sequence. Now use the data in **Table 22.5** to estimate the mass of this star. Then calculate the age of the cluster. This method is similar to the procedure used by astronomers to obtain the ages of clusters, except that they use actual data and model calculations rather than simply making estimates from a drawing. How do your ages compare with the ages in the text?

56. An exercise in plotting a proxy H-R diagram for the open cluster Pleiades yields an estimated distance modulus of $m - M = 5.67$.

How far away is the Pleiades cluster from Earth?

24.6 Further Evolution of Stars

57. You can estimate the age of the planetary nebula in image (c) in **Figure 24.30**. The diameter of the nebula is 600 times the diameter of our own solar system, or about 0.8 light-year. The gas is expanding away from the star at a rate of about 25 mi/s. Considering that distance = velocity $\times$ time, calculate how long ago the gas left the star if its speed has been constant the whole time. Make sure you use consistent units for time, speed, and distance.

25 | THE DEATHS OF STARS

Chapter Outline

Introduction

Figure 25.1 This remarkable picture of NGC 3603, a nebula in the Milky Way Galaxy, was taken with the Hubble Space Telescope. This image illustrates the life cycle of stars. In the bottom half of the image, we see clouds of dust and gas, where it is likely that star formation will take place in the near future. Near the center, there is a cluster of massive, hot young stars that are only a few million years old. Above and to the right of the cluster, there is an isolated star surrounded by a ring of gas. Perpendicular to the ring and on either side of it, there are two bluish blobs of gas. The ring and the blobs were ejected by the star, which is nearing the end of its life. (credit: modification of work by NASA, Wolfgang Brandner (JPL/IPAC), Eva K. Grebel (University of Washington), You-Hua Chu (University of Illinois Urbana-Champaign))

Do stars die with a bang or a whimper? In the preceding two chapters, we followed the life story of stars, from the process of birth to the brink of death. Now we are ready to explore the ways that stars end their lives. Sooner or later, each star exhausts its store of nuclear energy. Without a source of internal pressure to balance the weight of the overlying layers, every star eventually gives way to the inexorable pull of gravity and collapses under its own weight.

Following the rough distinction made in the last chapter, we will discuss the end-of-life evolution of stars of lower and higher mass separately. What determines the outcome—bang or whimper—is the mass of the star *when it is ready to die*, not the mass it was born with. As we noted in the last chapter, stars can lose a significant amount of mass in their middle and old age.

25.1 | The Death of Low-Mass Stars

Learning Objectives

By the end of this section, you will be able to:

- Describe the physical characteristics of degenerate matter and explain how the mass and radius of degenerate stars are related
- Plot the future evolution of a white dwarf and show how its observable features will change over time
- Distinguish which stars will become white dwarfs

Let's begin with those stars whose final mass just before death is less than about 1.4 times the mass of the Sun (M_{Sun}). (We will explain why this mass is the crucial dividing line in a moment.) Note that most stars in the universe fall into this category. The number of stars decreases as mass increases; really massive stars are rare (see **Stellar Properties**). This is similar to the music business where only a few musicians ever become superstars. Furthermore, many stars with an initial mass much greater than 1.4 M_{Sun} will be reduced to that level by the time they die. For example, we now know that stars that start out with masses of at least 8.0 M_{Sun} (and possibly as much as 10 M_{Sun}) manage to lose enough mass during their lives to fit into this category (an accomplishment anyone who has ever attempted to lose weight would surely envy).

A Star in Crisis

In the last chapter, we left the life story of a star with a mass like the Sun's just after it had climbed up to the red-giant region of the H–R diagram for a second time and had shed some of its outer layers to form a planetary nebula. Recall that during this time, the *core* of the star was undergoing an "energy crisis." Earlier in its life, during a brief stable period, helium in the core had gotten hot enough to fuse into carbon (and oxygen). But after this helium was exhausted, the star's core had once more found itself without a source of pressure to balance gravity and so had begun to contract.

This collapse is the final event in the life of the core. Because the star's mass is relatively low, it cannot push its core temperature high enough to begin another round of fusion (in the same way larger-mass stars can). The core continues to shrink until it reaches a density equal to nearly a million times the density of water! That is 200,000 times greater than the average density of Earth. At this extreme density, a new and different way for matter to behave kicks in and helps the star achieve a final state of equilibrium. In the process, what remains of the star becomes one of the strange *white dwarfs* that we met in **Stellar Properties**.

Degenerate Stars

Because white dwarfs are far denser than any substance on Earth, the matter inside them behaves in a very unusual way—unlike anything we know from everyday experience. At this high density, gravity is incredibly strong and tries to shrink the star still further, but all the *electrons* resist being pushed closer together and set up a powerful pressure inside the core. This pressure is the result of the fundamental rules that govern the behavior of electrons (the quantum physics you were introduced to in **Source of Sunshine: Nuclear Fusion!**). According to these rules (known to physicists as the Pauli exclusion principle), which have been verified in studies of atoms in the laboratory, no two electrons can be in the same place at the same time doing the same thing. We specify the *place* of an electron by its position in space, and we specify what it is doing by its motion and the way it is spinning.

The temperature in the interior of a star is always so high that the atoms are stripped of virtually all their electrons. For most of a star's life, the density of matter is also relatively low, and the electrons in the star are moving rapidly. This means that no two of them will be in the same place moving in exactly the same way at the same time. But this all changes when a star exhausts its store of nuclear energy and begins its final collapse.

As the star's core contracts, electrons are squeezed closer and closer together. Eventually, a star like the Sun becomes so dense that further contraction would in fact require two or more electrons to violate the rule against occupying the same place and moving in the same way. Such a dense gas is said to be degenerate (a term coined by physicists and not related to the electron's moral character). The electrons in a **degenerate gas** resist further crowding with tremendous pressure. (It's as if the electrons said, "You can press inward all you want, but there is simply no room for any other electrons to squeeze in here without violating the rules of our existence.")

The degenerate electrons do not require an input of heat to maintain the pressure they exert, and so a star with this kind of structure, if nothing disturbs it, can last essentially forever. (Note that the repulsive force between degenerate electrons is different from, and much stronger than, the normal electrical repulsion between charges that have the same sign.)

The electrons in a degenerate gas do move about, as do particles in any gas, but not with a lot of freedom. A particular electron cannot change position or momentum until another electron in an adjacent stage gets out of the way. The situation is much like that in the parking lot after a big football game. Vehicles are closely packed, and a given car cannot move until the one in front of it moves, leaving an empty space to be filled.

Of course, the dying star also has atomic nuclei in it, not just electrons, but it turns out that the nuclei must be squeezed to much higher densities before their quantum nature becomes apparent. As a result, in white dwarfs, the nuclei do not exhibit degeneracy pressure. Hence, in the white dwarf stage of stellar evolution, it is the degeneracy pressure of the electrons, and not of the nuclei, that halts the collapse of the core.

White Dwarfs

White dwarfs, then, are stable, compact objects with electron-degenerate cores that cannot contract any further. Calculations showing that white dwarfs are the likely end state of low-mass stars were first carried out by the Indian-American astrophysicist Subrahmanyan Chandrasekhar. He was able to show how much a star will shrink before the degenerate electrons halt its further contraction and hence what its final diameter will be (**Figure 25.2**).

When Chandrasekhar made his calculation about white dwarfs, he found something very surprising: the radius of a white dwarf shrinks as the mass in the star increases (the larger the mass, the more tightly packed the electrons can become, resulting in a smaller radius). According to the best theoretical models, a white dwarf with a mass of about 1.4 M_{Sun} or larger would have a radius of zero. What the calculations are telling us is that even the force of degenerate electrons cannot stop the collapse of a star with more mass than this. The maximum mass that a star can end its life with and still become a white dwarf—1.4 M_{Sun}—is called the **Chandrasekhar limit**. Stars with end-of-life masses that exceed this limit have a different kind of end in store—one that we will explore in the next section.

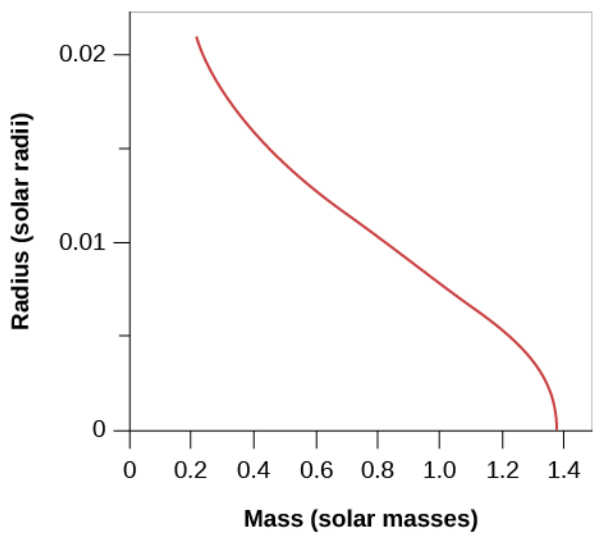


Figure 25.2 Models of white-dwarf structure predict that as the mass of the star increases (toward the right), its radius gets smaller and smaller.

Subrahmanyan Chandrasekhar

Born in 1910 in Lahore, India, Subrahmanyan Chandrasekhar (known as Chandra to his friends and colleagues) grew up in a home that encouraged scholarship and an interest in science (**Figure 25.3**). His uncle, C. V. Raman, was a physicist who won the 1930 Nobel Prize. A precocious student, Chandra tried to read as much as he could about the latest ideas in physics and astronomy, although obtaining technical books was not easy in India at the time. He finished college at age 19 and won a scholarship to study in England. It was during the long boat voyage to get to graduate school that he first began doing calculations about the structure of white dwarf stars.

Chandra developed his ideas during and after his studies as a graduate student, showing—as we have discussed—that white dwarfs with masses greater than 1.4 times the mass of the Sun cannot exist and that the theory predicts the existence of other kinds of stellar corpses. He wrote later that he felt very shy and lonely during this period, isolated from students, afraid to assert himself, and sometimes waiting for hours to speak with some of the famous professors he had read about in India. His calculations soon brought him into conflict with certain distinguished astronomers, including Sir Arthur Eddington, who publicly ridiculed Chandra's ideas. At a number of meetings of astronomers, such leaders in the field as Henry Norris Russell refused to give Chandra the opportunity to defend his ideas, while allowing his more senior critics lots of time to criticize them.

Yet Chandra persevered, writing books and articles elucidating his theories, which turned out not only to be correct, but to lay the foundation for much of our modern understanding of the death of stars. In 1983, he received the Nobel Prize in physics for this early work.

In 1937, Chandra came to the United States and joined the faculty at the University of Chicago, where he remained for the rest of his life. There he devoted himself to research and teaching, making major contributions to many fields of astronomy, from our understanding of the motions of stars through the Galaxy to the behavior of the bizarre objects called black holes. In 1999, NASA named its sophisticated orbiting X-ray telescope (designed in part to explore such stellar corpses) the Chandra X-ray Observatory.

Figure 25.3 Chandra's research provided the basis for much of what we now know about stellar corpses. (credit: modification of work by American Institute of Physics)

Chandra spent a great deal of time with his graduate students, supervising the research of more than 50 PhDs during his life. He took his teaching responsibilities very seriously: during the 1940s, while based at the Yerkes Observatory, he willingly drove the more than 100-mile trip to the university each week to teach a class of only a few students.

Chandra also had a deep devotion to music, art, and philosophy, writing articles and books about the relationship between the humanities and science. He once wrote that "one can learn science the way one enjoys music or art. . . . Heisenberg had a marvelous phrase 'shuddering before the beautiful'. . . that is the kind of feeling I have."

Using the Hubble Space Telescope, astronomers were able to detect **images (https://openstaxcollege.org/l/30hubimgwhidwa)** of faint white dwarf stars and other "stellar corpses" in the M4 star cluster, located about 7200 light-years away.

The Ultimate Fate of White Dwarfs

If the birth of a main-sequence star is defined by the onset of fusion reactions, then we must consider the end of all fusion reactions to be the time of a star's death. As the core is stabilized by degeneracy pressure, a last shudder of fusion passes through the outside of the star, consuming the little hydrogen still remaining. Now the star is a true white dwarf: nuclear fusion in its interior has ceased. **Figure 25.4** shows the path of a star like the Sun on the H–R diagram during its final stages.

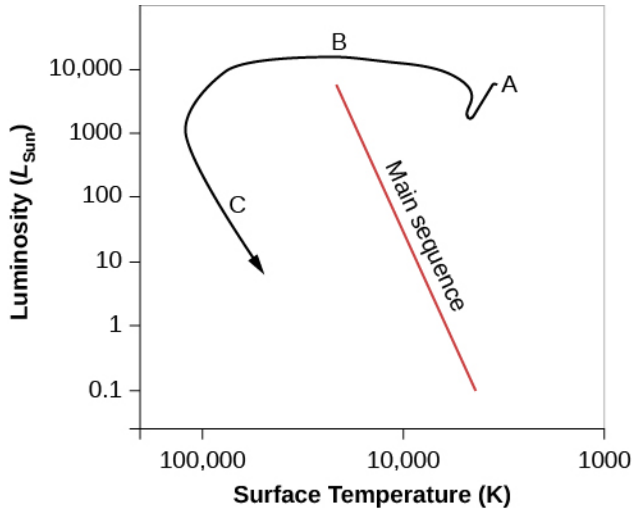


Figure 25.4 This diagram shows the changes in luminosity and surface temperature for a star with a mass like the Sun's as it nears the end of its life. After the star becomes a giant again (point A on the diagram), it will lose more and more mass as its core begins to collapse. The mass loss will expose the hot inner core, which will appear at the center of a planetary nebula. In this stage, the star moves across the diagram to the left as it becomes hotter and hotter during its collapse (point B). At first, the luminosity remains nearly constant, but as the star begins to cool off, it becomes less and less bright (point C). It is now a white dwarf and will continue to cool slowly for billions of years until all of its remaining store of energy is radiated away. (This assumes the Sun will lose between 46–50% of its mass during the giant stages, based upon various theoretical models).

Since a stable white dwarf can no longer contract or produce energy through fusion, its only energy source is the heat represented by the motions of the atomic nuclei in its interior. The light it emits comes from this internal stored heat, which is substantial. Gradually, however, the white dwarf radiates away all its heat into space. After many billions of years, the nuclei will be moving much more slowly, and the white dwarf will no longer shine (**Figure 25.5**). It will then be a black dwarf—a cold stellar corpse with the mass of a star and the size of a planet. It will be composed mostly of carbon, oxygen, and neon, the products of the most advanced fusion reactions of which the star was capable.

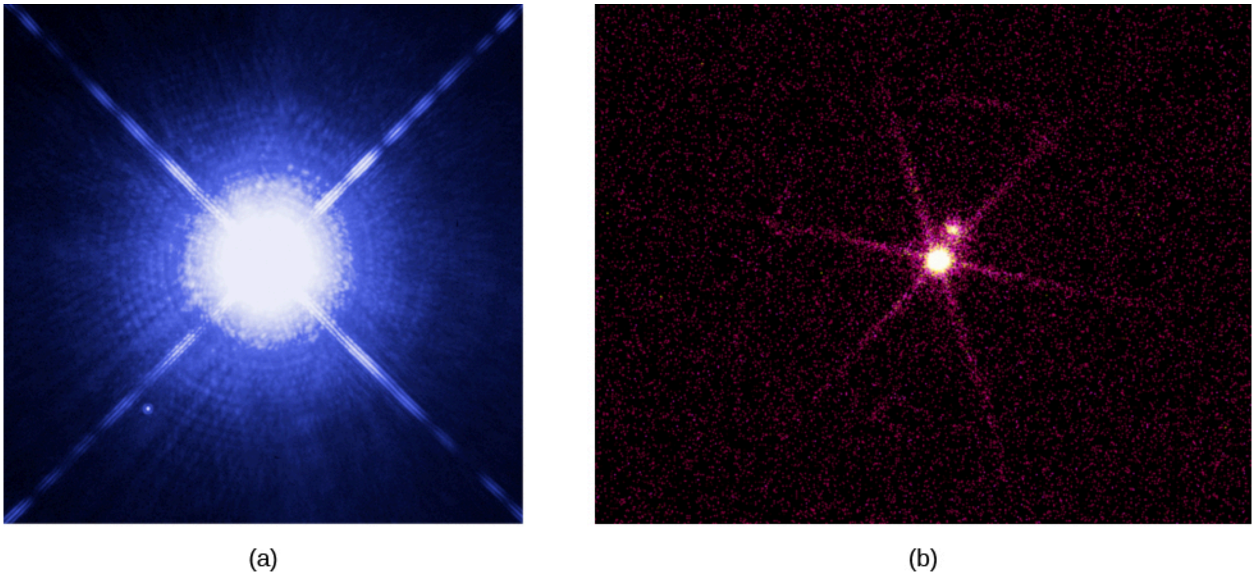


Figure 25.5 (a) This image taken by the Hubble Space Telescope shows Sirius A (the large bright star), and its companion star, the white dwarf known as Sirius B (the tiny, faint star at the lower left). Sirius A and B are 8.6 light-years from Earth and are our fifth-closest star system. Note that the image has intentionally been overexposed to allow us to see Sirius B. (b) The same system is shown in X-ray taken with the Chandra Space Telescope. Note that Sirius A is fainter in X-rays than the hot white dwarf that is Sirius B. (credit a: modification of work by NASA, ESA, H. Bond, M. Barstow(University of Leicester); credit b: modification of work by NASA/SAO/CXC)

We have one final surprise as we leave our low-mass star in the stellar graveyard. Calculations show that as a degenerate star cools, the atoms inside it in essence "solidify" into a giant, highly compact lattice (organized rows of atoms, just like in a crystal). When carbon is compressed and crystallized in this way, it becomes a giant *diamond-like* star. A white dwarf star might make the most impressive engagement present you could ever see, although any attempt to mine the diamond-like material inside would crush an ardent lover instantly!

Learn about a recent "diamond star" find, a **cold, white dwarf star (https://openstaxcollege.org/l/30diamondstar)** detected in 2014, which is considered the coldest and dimmest found to date, at the website of the National Radio Astronomy Observatory.

Evidence That Stars Can Shed a Lot of Mass as They Evolve

Whether or not a star will become a white dwarf depends on how much mass is lost in the red-giant and earlier phases of evolution. All stars that have masses below the Chandrasekhar limit when they run out of fuel will become white dwarfs, no matter what mass they were born with. But which stars shed enough mass to reach this limit?

One strategy for answering this question is to look in young, open clusters (which were discussed in **Star Clusters**). The basic idea is to search for young clusters that contain one or more white dwarf stars. Remember that more massive stars go through all stages of their evolution more rapidly than less massive ones. Suppose we find a cluster that has a white dwarf member and also contains stars on the main sequence that have 6 times the mass of the Sun. This means that only those stars with masses greater than 6 M_{Sun} have had time to exhaust their supply of nuclear energy and complete their evolution to the white dwarf stage. The star that turned into the white dwarf must therefore have had a main-sequence mass of more than 6 M_{Sun}, since stars with lower masses have not yet had time to use up their stores of nuclear energy. The star that became the white dwarf must, therefore, have gotten rid of at least 4.6 M_{Sun} so that its mass at the time nuclear energy generation ceased could be less than 1.4 M_{Sun}.

Astronomers continue to search for suitable clusters to make this test, and the evidence so far suggests that stars with masses up to about 8 M_{Sun} can shed enough mass to end their lives as white dwarfs. Stars like the Sun will probably lose about 45% of their initial mass and become white dwarfs with masses less than 1.4 M_{Sun}.

25.2 | Evolution of Massive Stars: An Explosive Finish

Learning Objectives

By the end of this section, you will be able to:

- Describe the interior of a massive star before a supernova
- Explain the steps of a core collapse and explosion
- List the hazards associated with nearby supernovae

Thanks to mass loss, then, stars with starting masses up to at least 8 M_{Sun} (and perhaps even more) probably end their lives as white dwarfs. But we know stars can have masses as large as 150 (or more) M_{Sun}. They have a different kind of death in store for them. As we will see, these stars die with a bang.

Nuclear Fusion of Heavy Elements

After the helium in its core is exhausted (see **The Evolution of More Massive Stars**), the evolution of a massive star takes a significantly different course from that of lower-mass stars. In a massive star, the weight of the outer layers is sufficient to force the carbon core to contract until it becomes hot enough to fuse carbon into oxygen, neon, and magnesium. This cycle of contraction, heating, and the ignition of another nuclear fuel repeats several more times. After each of the possible nuclear fuels is exhausted, the core contracts again until it reaches a new temperature high enough to fuse still-heavier nuclei. The products of carbon fusion can be further converted into silicon, sulfur, calcium, and argon. And these elements, when heated to a still-higher temperature, can combine to produce iron. Massive stars go through these stages very, very quickly. In really massive stars, some fusion stages toward the very end can take only months or even days! This is a far cry from the millions of years they spend in the main-sequence stage.

At this stage of its evolution, a massive star resembles an onion with an iron core. As we get farther from the center, we find shells of decreasing temperature in which nuclear reactions involve nuclei of progressively lower mass—silicon and sulfur, oxygen, neon, carbon, helium, and finally, hydrogen (**Figure 25.6**).

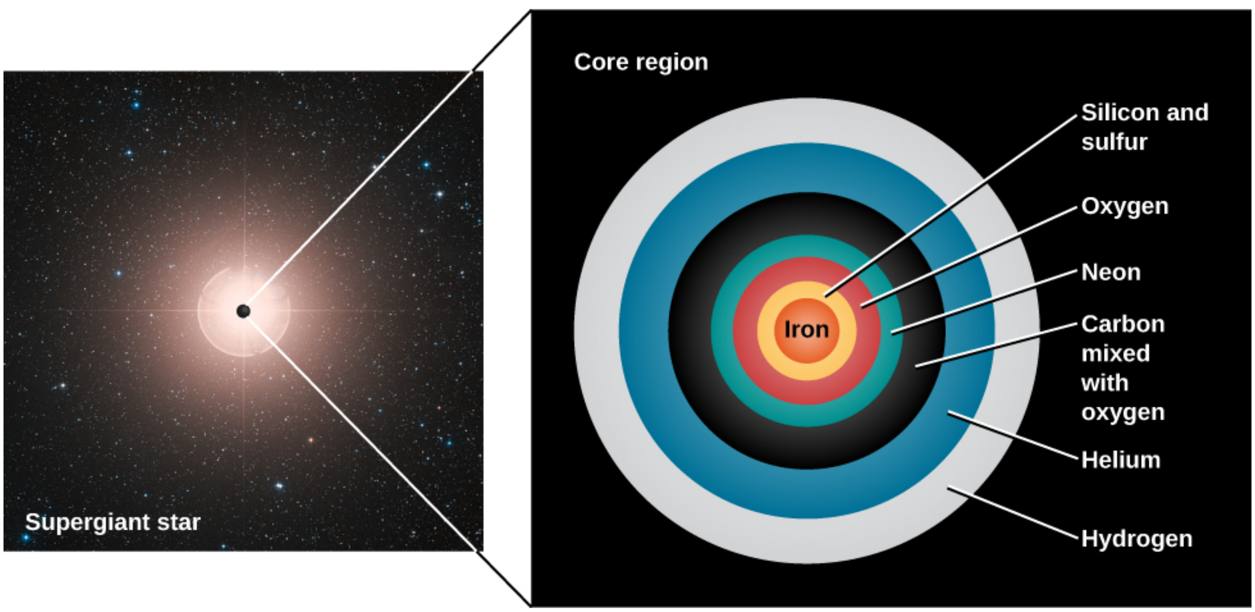


Figure 25.6 Just before its final gravitational collapse, the core of a massive star resembles an onion. The iron core is surrounded by layers of silicon and sulfur, oxygen, neon, carbon mixed with some oxygen, helium, and finally hydrogen. Outside the core, the composition is mainly hydrogen and helium. (Note that this diagram is not precisely to scale but is just meant to convey the general idea of what such a star would be like.) (credit: modification of work by ESO, Digitized Sky Survey)

But there is a limit to how long this process of building up elements by fusion can go on. The fusion of silicon into iron turns out to be the last step in the sequence of nonexplosive element production. Up to this point, each fusion reaction has *produced* energy because the nucleus of each fusion product has been a bit more stable than the nuclei that formed it. As discussed in **Source of Sunshine: Nuclear Fusion!**, light nuclei give up some of their binding energy in the process of fusing into more tightly bound, heavier nuclei. It is this released energy that maintains the outward pressure in the core so that the star does not collapse. But of all the nuclei known, iron is the most tightly bound and thus the most stable.

You might think of the situation like this: all smaller nuclei want to “grow up” to be like iron, and they are willing to pay (*produce* energy) to move toward that goal. But iron is a mature nucleus with good self-esteem, perfectly content being iron; it requires payment (must *absorb* energy) to change its stable nuclear structure. This is the exact opposite of what has happened in each nuclear reaction so far: instead of *providing* energy to balance the inward pull of gravity, any nuclear reactions involving iron would *remove* some energy from the core of the star.

Unable to generate energy, the star now faces catastrophe.

Collapse into a Ball of Neutrons

When nuclear reactions stop, the core of a massive star is supported by degenerate electrons, just as a white dwarf is. For stars that begin their evolution with masses of at least 10 M_{Sun}, this core is likely made mainly of iron. (For stars with initial masses in the range 8 to 10 M_{Sun}, the core is likely made of oxygen, neon, and magnesium, because the star never gets hot enough to form elements as heavy as iron. The exact composition of the cores of stars in this mass range is very difficult to determine because of the complex physical characteristics in the cores, particularly at the very high densities and temperatures involved.) We will focus on the more massive iron cores in our discussion.

While no energy is being generated within the white dwarf core of the star, fusion still occurs in the shells that surround the core. As the shells finish their fusion reactions and stop producing energy, the ashes of the last reaction fall onto the white dwarf core, increasing its mass. As **Figure 25.2** shows, a higher mass means a smaller core. The core can contract because even a degenerate gas is still mostly empty space. Electrons and atomic nuclei are, after all, extremely small. The electrons and nuclei in a stellar core may be crowded compared to the air in your room, but there is still lots of space between them.

The electrons at first resist being crowded closer together, and so the core shrinks only a small amount. Ultimately, however, the iron core reaches a mass so large that even degenerate electrons can no longer support it. When the density reaches 4×10^{11} g/cm^3 (400 billion times the density of water), some electrons are actually squeezed into the atomic nuclei, where they combine with protons to form neutrons and neutrinos. This transformation is not something that is familiar from everyday life, but becomes very important as such a massive star core collapses.

Some of the electrons are now gone, so the core can no longer resist the crushing mass of the star’s overlying layers. The core begins to shrink rapidly. More and more electrons are now pushed into the atomic nuclei, which ultimately become so saturated with neutrons that they cannot hold onto them.

At this point, the neutrons are squeezed out of the nuclei and can exert a new force. As is true for electrons, it turns out that the neutrons strongly resist being in the same place and moving in the same way. The force that can be exerted by such *degenerate neutrons* is much greater than that produced by degenerate electrons, so unless the core is too massive, they can ultimately stop the collapse.

This means the collapsing core can reach a stable state as a crushed ball made mainly of neutrons, which astronomers call a **neutron star**. We don’t have an exact number (a “Chandrasekhar limit”) for the maximum mass of a neutron star, but calculations tell us that the upper mass limit of a body made of neutrons might only be about 3 M_{Sun}. So if the mass of the core were greater than this, then even neutron degeneracy would not be able to stop the core from collapsing further. The dying star must end up as something even more extremely compressed, which until recently was believed to be only one possible type of object—the state of ultimate compaction known as a black hole (which is the subject of our next chapter). This is because no force was believed to exist that could stop a collapse beyond the neutron star stage.

Collapse and Explosion

When the collapse of a high-mass star’s core is stopped by degenerate neutrons, the core is saved from further destruction, but it turns out that the rest of the star is literally blown apart. Here’s how it happens.

The collapse that takes place when electrons are absorbed into the nuclei is very rapid. In less than a second, a core with a mass of about 1 M_{Sun}, which originally was approximately the size of Earth, collapses to a diameter of less than 20 kilometers. The speed with which material falls inward reaches one-fourth the speed of light. The collapse halts only when the density of the core exceeds the density of an atomic nucleus (which is the densest form of matter we know). A typical neutron star is so compressed that to duplicate its density, we would have to squeeze all the people in the world into a single sugar cube! This would give us one sugar cube’s worth (one cubic centimeter’s worth) of a neutron star.

The neutron degenerate core strongly resists further compression, abruptly halting the collapse. The shock of the sudden jolt initiates a shock wave that starts to propagate outward. However, this shock alone is not enough to create a star explosion.

The energy produced by the outflowing matter is quickly absorbed by atomic nuclei in the dense, overlying layers of gas, where it breaks up the nuclei into individual neutrons and protons.

Our understanding of nuclear processes indicates (as we mentioned above) that each time an electron and a proton in the star's core merge to make a neutron, the merger releases a *neutrino*. These ghostly subatomic particles, introduced in **Source of Sunshine: Nuclear Fusion!**, carry away some of the nuclear energy. It is their presence that launches the final disastrous explosion of the star. The total energy contained in the neutrinos is huge. In the initial second of the star's explosion, the power carried by the neutrinos (10^{46} watts) is greater than the power put out by all the stars in over a billion galaxies.

While neutrinos ordinarily do not interact very much with ordinary matter (we earlier accused them of being downright antisocial), matter near the center of a collapsing star is so dense that the neutrinos do interact with it to some degree. They deposit some of this energy in the layers of the star just outside the core. This huge, sudden input of energy reverses the infall of these layers and drives them explosively outward. Most of the mass of the star (apart from that which went into the neutron star in the core) is then ejected outward into space. As we saw earlier, such an explosion requires a star of at least 8 M_{Sun}, and the neutron star can have a mass of at most 3 $M_{Sun.}$ Consequently, at least five times the mass of our Sun is ejected into space in each such explosive event!

The resulting explosion is called a supernova (**Figure 25.7**). When these explosions happen close by, they can be among the most spectacular celestial events, as we will discuss in the next section. (Actually, there are at least two different types of supernova explosions: the kind we have been describing, which is the collapse of a massive star, is called, for historical reasons, a **type II supernova**. We will describe how the types differ later in this chapter).

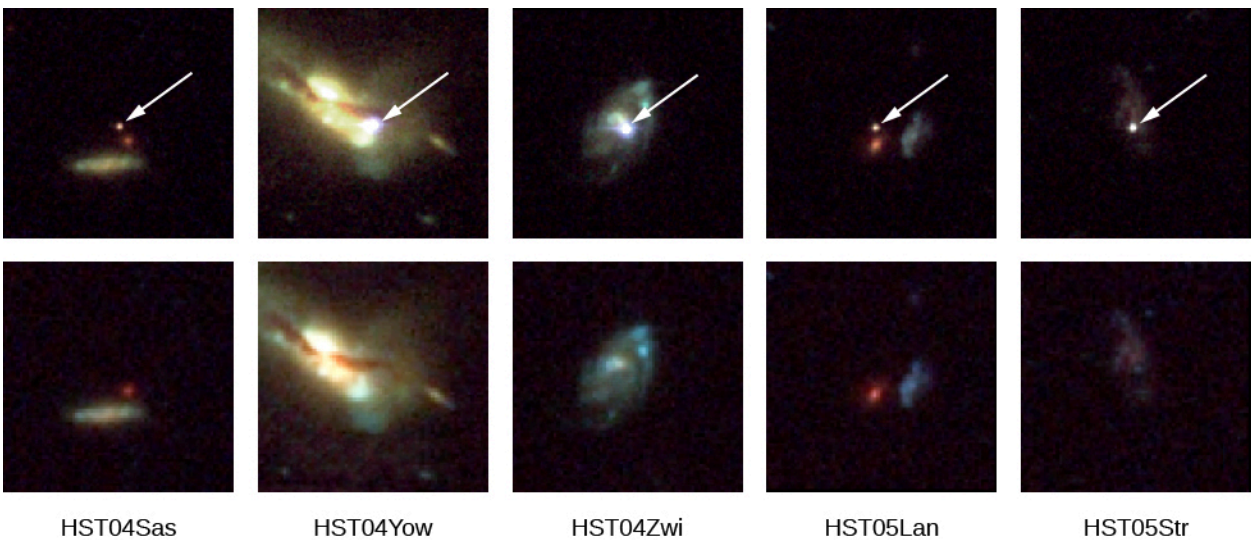


Figure 25.7 The arrows in the top row of images point to the supernovae. The bottom row shows the host galaxies before or after the stars exploded. Each of these supernovae exploded between 3.5 and 10 billion years ago. Note that the supernovae when they first explode can be as bright as an entire galaxy. (credit: modification of work by NASA, ESA, and A. Riess (STScI))

Table 25.1 summarizes the discussion so far about what happens to stars and substellar objects of different initial masses at the ends of their lives. Like so much of our scientific understanding, this list represents a progress report: it is the best we can do with our present models and observations. The mass limits corresponding to various outcomes may change somewhat as models are improved. There is much we do not yet understand about the details of what happens when stars die.

The Ultimate Fate of Stars and Substellar Objects with Different Masses

Initial Mass (Mass of Sun = 1)[1]	Final State at the End of Its Life
< 0.01	Planet

Table 25.1

1. Stars in the mass ranges 0.25–8 and 8–10 may later produce a type of supernova different from the one we have discussed so far. These are discussed in The Evolution of Binary Star Systems (https://legacy.cnx.org/content/m59936/latest/) .

The Ultimate Fate of Stars and Substellar Objects with Different Masses

Initial Mass (Mass of Sun = 1)	Final State at the End of Its Life
0.01 to 0.08	Brown dwarf
0.08 to 0.25	White dwarf made mostly of helium
0.25 to 8	White dwarf made mostly of carbon and oxygen
8 to 10	White dwarf made of oxygen, neon, and magnesium
10 to 40	Supernova explosion that leaves a neutron star
> 40	Supernova explosion that leaves a black hole

Table 25.1

The Supernova Giveth and the Supernova Taketh Away

After the supernova explosion, the life of a massive star comes to an end. But the death of each massive star is an important event in the history of its galaxy. The elements built up by fusion during the star's life are now "recycled" into space by the explosion, making them available to enrich the gas and dust that form new stars and planets. Because these heavy elements ejected by supernovae are critical for the formation of planets and the origin of life, it's fair to say that without mass loss from supernovae and planetary nebulae, neither the authors nor the readers of this book would exist.

But the supernova explosion has one more creative contribution to make, one we alluded to in **Stellar Life Cycles** when we asked where the atoms in your jewelry came from. The supernova explosion produces a flood of energetic neutrons that barrel through the expanding material. These neutrons can be absorbed by iron and other nuclei where they can turn into protons. Thus, they can build up elements that are more massive than iron, possibly including such terrestrial favorites as gold, silver and uranium. Supernovae (and, as we will shortly see, the explosive mergers of neutron stars) are the only candidates we have for places where such heavier atoms can be made. Next time you wear some gold jewelry (or give some to your sweetheart), bear in mind that those gold atoms were forged long ago in these kinds of celestial explosions!

When supernovae explode, these elements (as well as the ones the star made during more stable times) are ejected into the existing gas between the stars and mixed with it. Thus, supernovae play a crucial role in enriching their galaxy with heavier elements, allowing, among other things, the chemical elements that make up earthlike planets and the building blocks of life to become more common as time goes on (**Figure 25.8**).

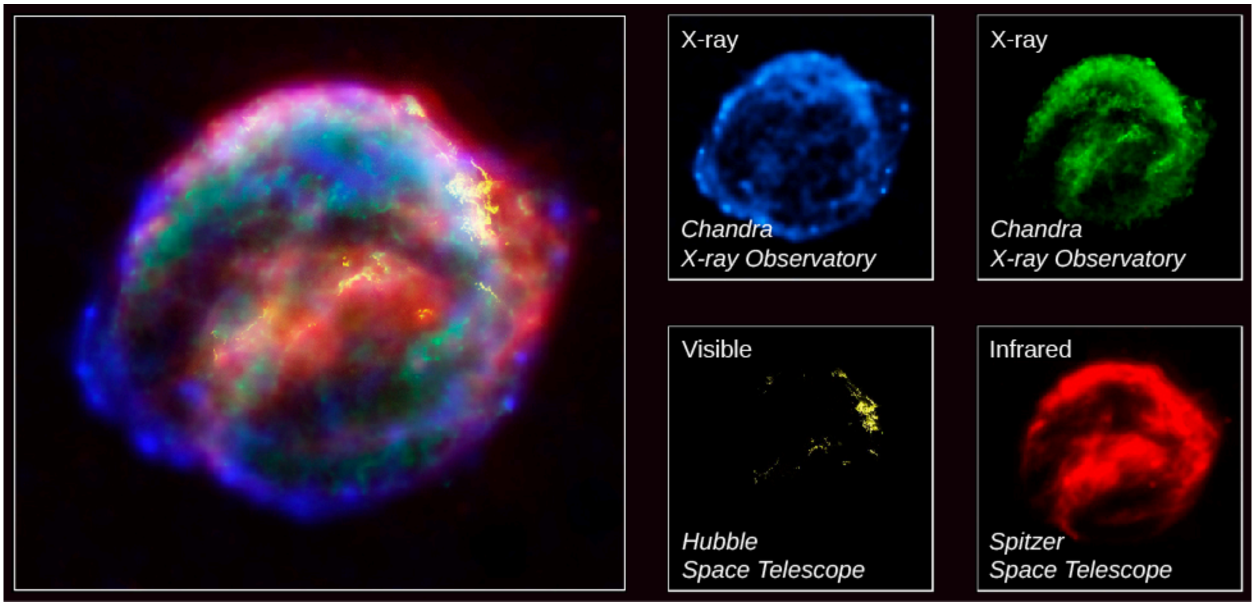


Figure 25.8 This image shows the expanding remains of a supernova explosion, which was first seen about 400 years ago by sky watchers, including the famous astronomer Johannes Kepler. The bubble-shaped shroud of gas and dust is now 14 light-years wide and is expanding at 2,000 kilometers per second (4 million miles per hour). The remnant emits energy at wavelengths from X-rays (shown in blue and green) to visible light (yellow) and into the infrared (red). The expanding shell is rich in iron, which was produced in the star that exploded. The main image combines the individual single-color images seen at the bottom into one multi-wavelength picture. (credit: modification of work by NASA, ESA, R. Sankrit and W. Blair (Johns Hopkins University))

Supernovae are also thought to be the source of many of the high-energy *cosmic ray* particles. Trapped by the magnetic field of the Galaxy, the particles from exploded stars continue to circulate around the vast spiral of the Milky Way. Scientists speculate that high-speed cosmic rays hitting the genetic material of Earth organisms over billions of years may have contributed to the steady *mutations*—subtle changes in the genetic code—that drive the evolution of life on our planet. In all the ways we have mentioned, supernovae have played a part in the development of new generations of stars, planets, and life.

But supernovae also have a dark side. Suppose a life form has the misfortune to develop around a star that happens to lie near a massive star destined to become a supernova. Such life forms may find themselves snuffed out when the harsh radiation and high-energy particles from the neighboring star's explosion reach their world. If, as some astronomers speculate, life can develop on many planets around long-lived (lower-mass) stars, then the suitability of that life's *own star* and planet may not be all that matters for its long-term evolution and survival. Life may well have formed around a number of pleasantly stable stars only to be wiped out because a massive nearby star suddenly went supernova. Just as children born in a war zone may find themselves the unjust victims of their violent neighborhood, life too close to a star that goes supernova may fall prey to having been born in the wrong place at the wrong time.

What is a safe distance to be from a supernova explosion? A lot depends on the violence of the particular explosion, what type of supernova it is (see **The Evolution of Binary Star Systems**), and what level of destruction we are willing to accept. Calculations suggest that a supernova less than 50 light-years away from us would certainly end all life on Earth, and that even one 100 light-years away would have drastic consequences for the radiation levels here. One minor extinction of sea creatures about 2 million years ago on Earth may actually have been caused by a supernova at a distance of about 120 light-years.

The good news is that there are at present no massive stars that promise to become supernovae within 50 light-years of the Sun. (This is in part because the kinds of massive stars that become supernovae are overall quite rare.) The massive star closest to us, Spica (in the constellation of Virgo), is about 260 light-years away, probably a safe distance, even if it were to explode as a supernova in the near future.

Example 25.1

Extreme Gravity

In this section, you were introduced to some very dense objects. How would those objects' gravity affect you? Recall that the force of gravity, F, between two bodies is calculated as

$$F = \frac{GM_1 M_2}{R^2}$$

where G is the gravitational constant, 6.67×10^{-11} Nm^2/kg^2, M_1 and M_2 are the masses of the two bodies, and R is their separation. Also, from Newton's second law,

$$F = M \times a$$

where a is the acceleration of a body with mass M.

So let's consider the situation of a mass—say, you—standing on a body, such as Earth or a white dwarf (where we assume you will be wearing a heat-proof space suit). You are M_1 and the body you are standing on is M_2. The distance between you and the center of gravity of the body on which you stand is its radius, R. The force exerted on you is

$$F = M_1 \times a = \text{GM}_1 M_2 \Big/ R^2$$

Solving for a, the acceleration of gravity on that world, we get

$$g = \frac{(G \times M)}{R^2}$$

Note that we have replaced the general symbol for acceleration, a, with the symbol scientists use for the acceleration of gravity, g.

Say that a particular white dwarf has the mass of the Sun (2×10^{30} kg) but the radius of Earth (6.4×10^6 m). What is the acceleration of gravity at the surface of the white dwarf?

Solution

The acceleration of gravity at the surface of the white dwarf is

$$g(\text{white dwarf}) = \frac{(G \times M_{\text{Sun}})}{{R_{\text{Earth}}}^2} = \frac{\left(6.67 \times 10^{-11}\ \text{m}^2/\text{kg s}^2 \times 2 \times 10^{30}\ \text{kg}\right)}{\left(6.4 \times 10^6\ \text{m}\right)^2} = 3.26 \times 10^6\ \text{m/s}^2$$

Compare this to g on the surface of Earth, which is 9.8 m/s^2.

25.1 Check Your Learning

What is the acceleration of gravity at the surface if the white dwarf has the twice the mass of the Sun and is only half the radius of Earth?

25.3 | Supernova Observations

Learning Objectives

By the end of this section, you will be able to:

- Describe the observed features of SN 1987A both before and after the supernova
- Explain how observations of various parts of the SN 1987A event helped confirm theories about supernovae

Supernovae were discovered long before astronomers realized that these spectacular cataclysms mark the death of stars. The word *nova* means "new" in Latin; before telescopes, when a star too dim to be seen with the unaided eye suddenly flared up in a brilliant explosion, observers concluded it must be a brand-new star. Twentieth-century astronomers reclassified the explosions with the greatest luminosity as *super*novae.

From historical records of such explosions, from studies of the remnants of supernovae in our Galaxy, and from analyses of supernovae in other galaxies, we estimate that, on average, one supernova explosion occurs somewhere in the Milky Way Galaxy every 25 to 100 years. Unfortunately, however, no supernova explosion has been observable in our Galaxy since the invention of the telescope. Either we have been exceptionally unlucky or, more likely, recent explosions have taken place in parts of the Galaxy where interstellar dust blocks light from reaching us.

Supernovae in History

Although many supernova explosions in our own Galaxy have gone unnoticed, a few were so spectacular that they were clearly seen and recorded by sky watchers and historians at the time. We can use these records, going back two millennia, to help us pinpoint where the exploding stars were and thus where to look for their remnants today.

The most dramatic supernova was observed in the year 1006. It appeared in May as a brilliant point of light visible during the daytime, perhaps 100 times brighter than the planet Venus. It was bright enough to cast shadows on the ground during the night and was recorded with awe and fear by observers all over Europe and Asia. No one had seen anything like it before; Chinese astronomers, noting that it was a temporary spectacle, called it a "guest star."

Astronomers David Clark and Richard Stephenson have scoured records from around the world to find more than 20 reports of the 1006 supernova (SN 1006) (**Figure 25.9**). This has allowed them to determine with some accuracy where in the sky the explosion occurred. They place it in the modern constellation of Lupus; at roughly the position they have determined, we find a supernova remnant, now quite faint. From the way its filaments are expanding, it indeed appears to be about 1000 years old.

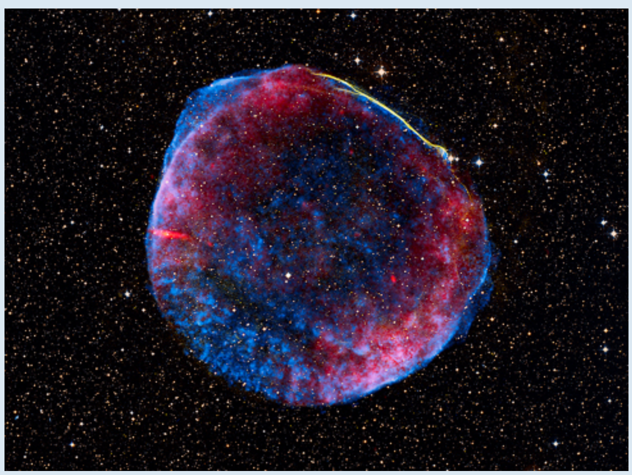


Figure 25.9 This composite view of SN 1006 from the Chandra X-Ray Observatory shows the X-rays coming from the remnant in blue, visible light in white-yellow, and radio emission in red. (credit: modification of work by NASA, ESA, Zolt Levay(STScI))

Another guest star, now known as SN 1054, was clearly recorded in Chinese records in July 1054. The remnant of that star is one of the most famous and best-studied objects in the sky, called the Crab Nebula (**Figure 25.14**). It is a marvelously complex object, which has been key to understanding the death of massive stars. When its explosion was first seen, we estimate that it was about as bright as the planet Jupiter: nowhere near as dazzling as the 1006 event but still quite dramatic to anyone who kept track of objects in the sky. Another fainter supernova was seen in 1181.

The next supernova became visible in November 1572 and, being brighter than the planet Venus, was quickly spotted by a number of observers, including the young Tycho Brahe (see **Kepler's Laws of Planetary Motion**). His careful measurements of the star over a year and a half showed that it was not a comet or something in Earth's atmosphere since it did not move relative to the stars. He correctly deduced that it must be a phenomenon belonging to the realm of the stars, not of the solar system. The remnant of Tycho's Supernova (as it is now called) can still be detected in many different bands of the electromagnetic spectrum.

Not to be outdone, Johannes Kepler, Tycho Brahe's scientific heir, found his own supernova in 1604, now known as Kepler's Supernova (**Figure 25.8**). Fainter than Tycho's, it nevertheless remained visible for about a year. Kepler wrote a book about his observations that was read by many with an interest in the heavens, including Galileo.

No supernova has been spotted in our Galaxy for the past 300 years. Since the explosion of a visible supernova is a chance event, there is no way to say when the next one might occur. Around the world, dozens of professional and amateur astronomers keep a sharp lookout for "new" stars that appear overnight, hoping to be the first to spot the next guest star in our sky and make a little history themselves.

At their maximum brightness, the most luminous supernovae have about 10 billion times the luminosity of the Sun. For a brief time, a supernova may outshine the entire galaxy in which it appears. After maximum brightness, the star's light fades and disappears from telescopic visibility within a few months or years. At the time of their outbursts, supernovae eject material at typical velocities of 10,000 kilometers per second (and speeds twice that have been observed). A speed of 20,000 kilometers per second corresponds to about 45 million miles per hour, truly an indication of great cosmic violence.

Supernovae are classified according to the appearance of their spectra, but in this chapter, we will focus on the two main causes of supernovae. Type Ia supernovae are ignited when a lot of material is dumped on degenerate white dwarfs (**Figure 25.10**); these supernovae will be discussed later in this chapter. For now, we will continue our story about the death of massive stars and focus on type II supernovae, which are produced when the core of a massive star collapses.

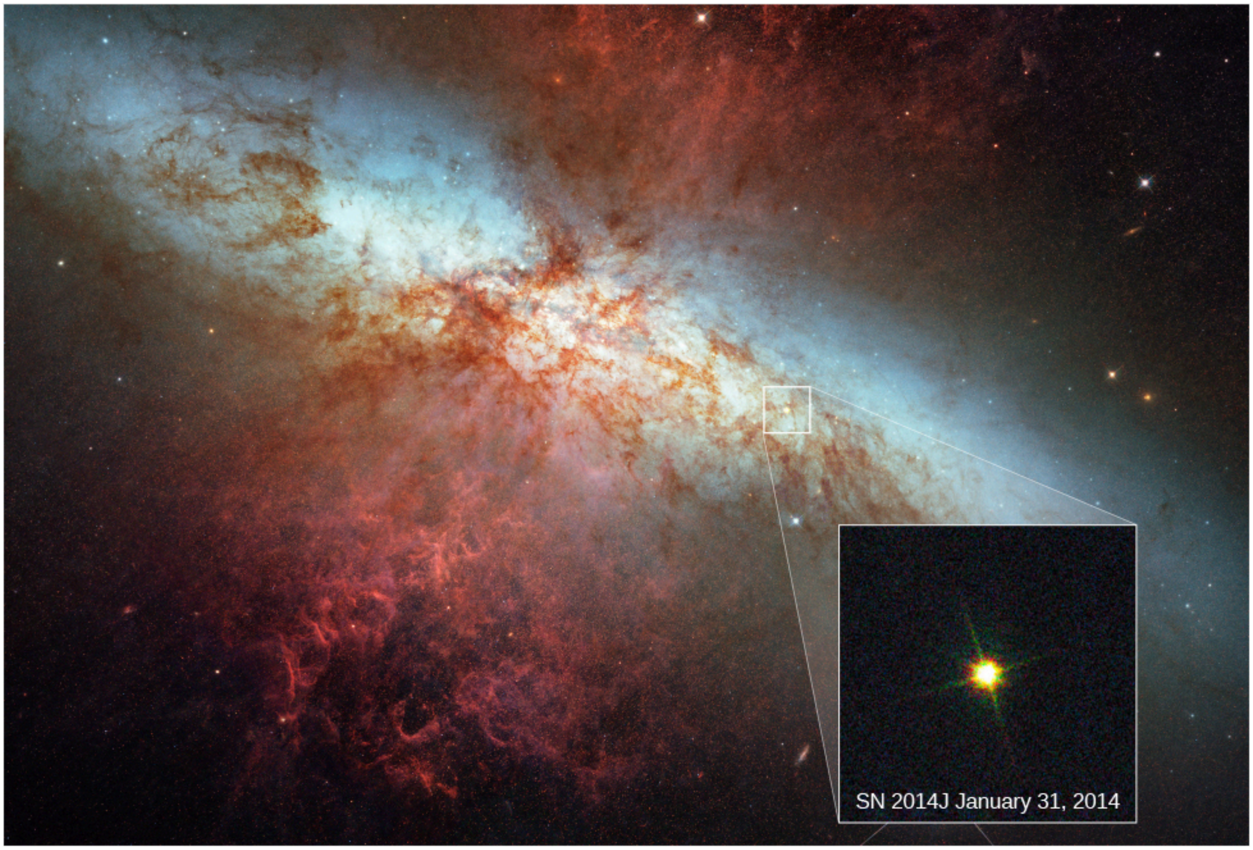


Figure 25.10 This image of supernova 2014J, located in Messier 82 (M82), which is also known as the Cigar galaxy, was taken by the Hubble Space Telescope and is superposed on a mosaic image of the galaxy also taken with Hubble. The supernova event is indicated by the box and the inset. This explosion was produced by a type Ia supernova, which is theorized to be triggered in binary systems consisting of a white dwarf and another star—and could be a second white dwarf, a star like our Sun, or a giant star. This type of supernova will be discussed later in this chapter. At a distance of approximately 11.5 million light-years from Earth, this is the closest supernova of type Ia discovered in the past few decades. In the image, you can see reddish plumes of hydrogen coming from the central region of the galaxy, where a considerable number of young stars are being born. (credit: modification of work by NASA, ESA, A. Goobar (Stockholm University), and the Hubble Heritage Team (STScI/AURA))

Supernova 1987A

Our most detailed information about what happens when a type II supernova occurs comes from an event that was observed in 1987. Before dawn on February 24, Ian Shelton, a Canadian astronomer working at an observatory in Chile, pulled a photographic plate from the developer. Two nights earlier, he had begun a survey of the Large Magellanic Cloud, a small galaxy that is one of the Milky Way's nearest neighbors in space. Where he expected to see only faint stars, he saw a large bright spot. Concerned that his photograph was flawed, Shelton went outside to look at the Large Magellanic Cloud . . . and saw that a new object had indeed appeared in the sky (see **Figure 25.11**). He soon realized that he had discovered a supernova, one that could be seen with the unaided eye even though it was about 160,000 light-years away.

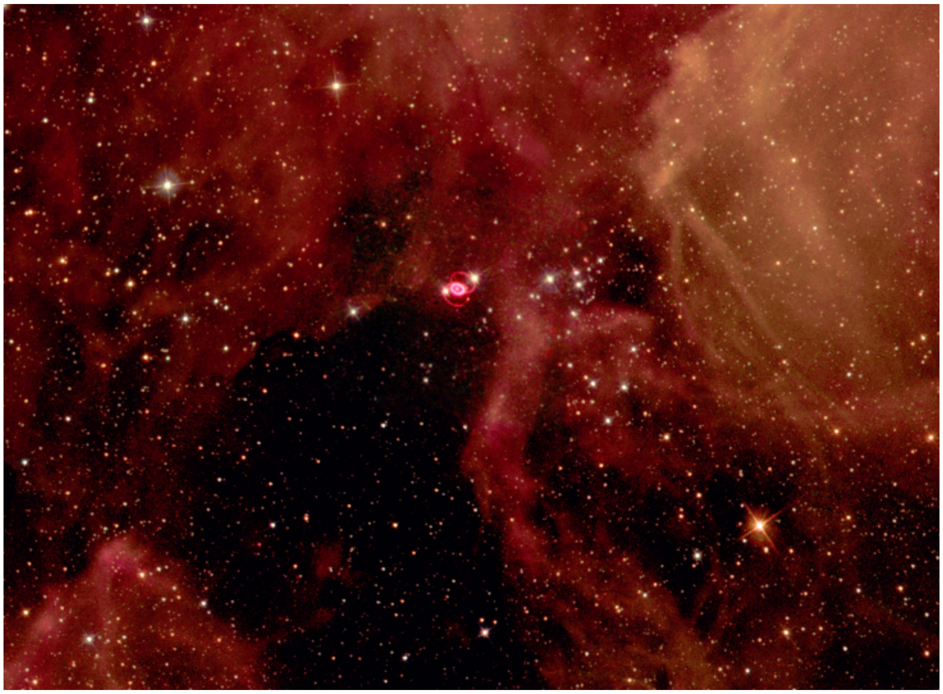


Figure 25.11 The supernova remnant with its inner and outer red rings of material is located in the Large Magellanic Cloud. This image is a composite of several images taken in 1994, 1996, and 1997—about a decade after supernova 1987A was first observed. (credit: modification of work by the Hubble Heritage Team (AURA/STScI/NASA/ESA))

Now known as SN 1987A, since it was the first supernova discovered in 1987, this brilliant newcomer to the southern sky gave astronomers their first opportunity to study the death of a relatively nearby star with modern instruments. It was also the first time astronomers had observed a star *before* it became a supernova. The star that blew up had been included in earlier surveys of the Large Magellanic Cloud, and as a result, we know the star was a blue supergiant just before the explosion.

By combining theory and observations at many different wavelengths, astronomers have reconstructed the life story of the star that became SN 1987A. Formed about 10 million years ago, it originally had a mass of about 20 M_{Sun}. For 90% of its life, it lived quietly on the main sequence, converting hydrogen into helium. At this time, its luminosity was about 60,000 times that of the Sun (L_{Sun}), and its spectral type was O. When the hydrogen in the center of the star was exhausted, the core contracted and ultimately became hot enough to fuse helium. By this time, the star was a red supergiant, emitting about 100,000 times more energy than the Sun. While in this stage, the star lost some of its mass.

This lost material has actually been detected by observations with the Hubble Space Telescope (**Figure 25.12**). The gas driven out into space by the subsequent supernova explosion is currently colliding with the material the star left behind when it was a red giant. As the two collide, we see a glowing ring.

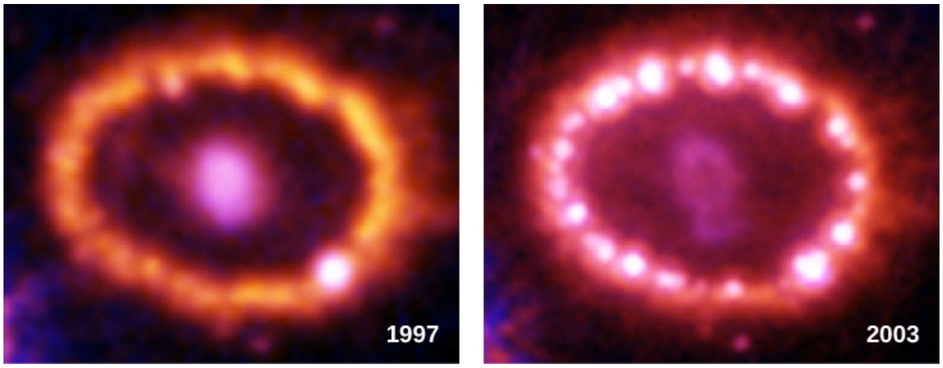


Figure 25.12 These two images show a ring of gas expelled by a red giant star about 30,000 years before the star exploded and was observed as Supernova 1987A. The supernova, which has been artificially dimmed, is located at the center of the ring. The left-hand image was taken in 1997 and the right-hand image in 2003. Note that the number of bright spots has increased from 1 to more than 15 over this time interval. These spots occur where high-speed gas ejected by the supernova and moving at millions of miles per hour has reached the ring and blasted into it. The collision has heated the gas in the ring and caused it to glow more brightly. The fact that we see individual spots suggests that material ejected by the supernova is first hitting narrow, inward-projecting columns of gas in the clumpy ring. The hot spots are the first signs of a dramatic and violent collision between the new and old material that will continue over the next few years. By studying these bright spots, astronomers can determine the composition of the ring and hence learn about the nuclear processes that build heavy elements inside massive stars. (credit: modification of work by NASA, P. Challis, R. Kirshner (Harvard-Smithsonian Center for Astrophysics) and B. Sugerman (STScI))

Helium fusion lasted only about 1 million years. When the helium was exhausted at the center of the star, the core contracted again, the radius of the surface also decreased, and the star became a blue supergiant with a luminosity still about equal to 100,000 L_{Sun}. This is what it still looked like on the outside when, after brief periods of further fusion, it reached the iron crisis we discussed earlier and exploded.

Some key stages of evolution of the star that became SN 1987A, including the ones following helium exhaustion, are listed in **Table 25.2**. While we don’t expect you to remember these numbers, note the patterns in the table: each stage of evolution happens more quickly than the preceding one, the temperature and pressure in the core increase, and progressively heavier elements are the source of fusion energy. Once iron was created, the collapse began. It was a catastrophic collapse, lasting only a few tenths of a second; the speed of infall in the outer portion of the iron core reached 70,000 kilometers per second, about one-fourth the speed of light.

Evolution of the Star That Exploded as SN 1987A

Phase	Central Temperature (K)	Central Density (g/cm^3)	Time Spent in This Phase
Hydrogen fusion	40×10^6	5	8×10^6 years
Helium fusion	190×10^6	970	10^6 years
Carbon fusion	870×10^6	170,000	2000 years
Neon fusion	1.6×10^9	3.0×10^6	6 months
Oxygen fusion	2.0×10^9	5.6×10^6	1 year
Silicon fusion	3.3×10^9	4.3×10^7	Days
Core collapse	200×10^9	2×10^{14}	Tenths of a second

Table 25.2

In the meantime, as the core was experiencing its last catastrophe, the outer shells of neon, oxygen, carbon, helium, and hydrogen in the star did not yet know about the collapse. Information about the physical movement of different layers travels through a star at the speed of sound and cannot reach the surface in the few tenths of a second required for the core collapse to occur. Thus, the surface layers of our star hung briefly suspended, much like a cartoon character who dashes off the edge of a cliff and hangs momentarily in space before realizing that he is no longer held up by anything.

The collapse of the core continued until the densities rose to several times that of an atomic nucleus. The resistance to further collapse then became so great that the core rebounded. Infalling material ran into the “brick wall” of the rebounding core and was thrown outward with a great shock wave. Neutrinos poured out of the core, helping the shock wave blow the star apart. The shock reached the surface of the star a few hours later, and the star began to brighten into the supernova Ian Shelton observed in 1987.

The Synthesis of Heavy Elements

The variations in the brightness of SN 1987A in the days and months after its discovery, which are shown in **Figure 25.13**, helped confirm our ideas about heavy element production. In a single day, the star soared in brightness by a factor of about 1000 and became just visible without a telescope. The star then continued to increase slowly in brightness until it was about the same apparent magnitude as the stars in the Little Dipper. Up until about day 40 after the outburst, the energy being radiated away was produced by the explosion itself. But then SN 1987A did not continue to fade away, as we might have expected the light from the explosion to do. Instead, SN 1987A remained bright as energy from newly created radioactive elements came into play.

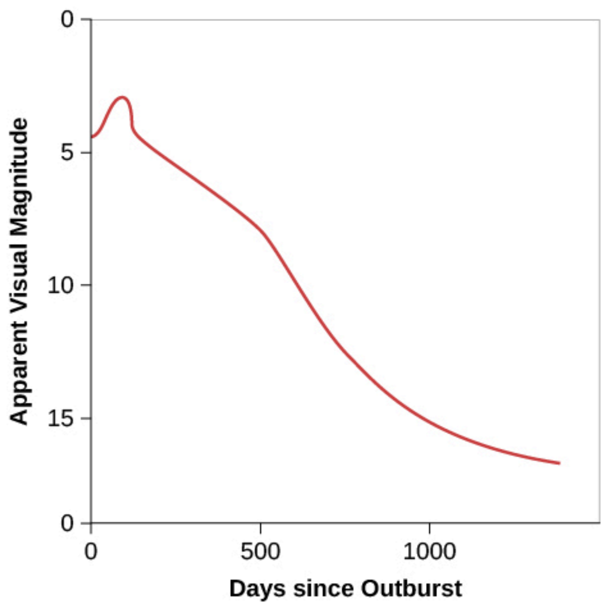


Figure 25.13 Note how the rate of decline of the supernova’s light slowed between days 40 and 500. During this time, the brightness was mainly due to the energy emitted by newly formed (and quickly decaying) radioactive elements. Remember that magnitudes are a backward measure of brightness: the larger the magnitude, the dimmer the object looks.

One of the elements formed in a supernova explosion is radioactive nickel, with an atomic mass of 56 (that is, the total number of protons plus neutrons in its nucleus is 56). Nickel-56 is unstable and changes spontaneously (with a half-life of about 6 days) to cobalt-56. (Recall that a half-life is the time it takes for half the nuclei in a sample to undergo radioactive decay.) Cobalt-56 in turn decays with a half-life of about 77 days to iron-56, which is stable. Energetic gamma rays are emitted when these radioactive nuclei decay. Those gamma rays then serve as a new source of energy for the expanding layers of the supernova. The gamma rays are absorbed in the overlying gas and re-emitted at visible wavelengths, keeping the remains of the star bright.

As you can see in **Figure 25.13**, astronomers did observe brightening due to radioactive nuclei in the first few months following the supernova’s outburst and then saw the extra light die away as more and more of the radioactive nuclei decayed to stable iron. The gamma-ray heating was responsible for virtually all of the radiation detected from SN 1987A after day 40. Some gamma rays also escaped directly without being absorbed. These were detected by Earth-orbiting telescopes at the wavelengths expected for the decay of radioactive nickel and cobalt, clearly confirming our understanding that new elements were indeed formed in the crucible of the supernova.

Neutrinos from SN 1987A

If there had been any human observers in the Large Magellanic Cloud about 160,000 years ago, the explosion we call SN 1987A would have been a brilliant spectacle in their skies. Yet we know that less than 1/10 of 1% of the energy of the explosion appeared as visible light. About 1% of the energy was required to destroy the star, and the rest was carried away by neutrinos. The overall energy in these neutrinos was truly astounding. In the initial second of the event, as we noted earlier in our general discussion of supernovae, their total luminosity exceeded the luminosity of all the stars in over a billion galaxies. And the supernova generated this energy in a volume less than 50 kilometers in diameter! Supernovae are one of the most violent events in the universe, and their *light* turns out to be only the tip of the iceberg in revealing how much energy they produce.

In 1987, the neutrinos from SN 1987A were detected by two instruments—which might be called "neutrino telescopes"—almost a full day before Shelton's observations. (This is because the neutrinos get out of the exploding star more easily than light does, and also because you don't need to wait until nightfall to catch a "glimpse" of them.) Both neutrino telescopes, one in a deep mine in Japan and the other under Lake Erie, consist of several thousand tons of purified water surrounded by several hundred light-sensitive detectors. Incoming neutrinos interact with the water to produce positrons and electrons, which move rapidly through the water and emit deep blue light.

Altogether, 19 neutrinos were detected. Since the neutrino telescopes were in the Northern Hemisphere and the supernova occurred in the Southern Hemisphere, the detected neutrinos had already passed through Earth and were on their way back out into space when they were captured.

Only a few neutrinos were detected because the probability that they will interact with ordinary matter is very, very low. It is estimated that the supernova actually released 10^{58} neutrinos. A tiny fraction of these, about 30 billion, eventually passed through each square centimeter of Earth's surface. About a million people actually experienced a neutrino interaction within their bodies as a result of the supernova. This interaction happened to only a single nucleus in each person and thus had absolutely no biological effect; it went completely unnoticed by everyone concerned.

Since the neutrinos come directly from the heart of the supernova, their energies provided a measure of the temperature of the core as the star was exploding. The central temperature was about 200 billion K, a stunning figure to which no earthly analog can bring much meaning. With neutrino telescopes, we are peering into the final moment in the life stories of massive stars and observing conditions beyond all human experience. Yet we are also seeing the unmistakable hints of our own origins.

25.4 | Pulsars and the Discovery of Neutron Stars

Learning Objectives

By the end of this section, you will be able to:

- Explain the research method that led to the discovery of neutron stars, located hundreds or thousands of light-years away
- Describe the features of a neutron star that allow it to be detected as a pulsar
- List the observational evidence that links pulsars and neutron stars to supernovae

After a type II supernova explosion fades away, all that is left behind is either a neutron star or something even more strange, a black hole. We will describe the properties of black holes in **Black Holes**, but for now, we want to examine how the neutron stars we discussed earlier might become observable.

Neutron stars are the densest objects in the universe; the force of gravity at their surface is 10^{11} times greater than what we experience at Earth's surface. The interior of a neutron star is composed of about 95% neutrons, with a small number of protons and electrons mixed in. In effect, a neutron star is a giant atomic nucleus, with a mass about 10^{57} times the mass of a proton. Its diameter is more like the size of a small town or an asteroid than a star. (**Table 25.3** compares the properties of neutron stars and white dwarfs.) Because it is so small, a neutron star probably strikes you as the object least likely to be observed from thousands of light-years away. Yet neutron stars do manage to signal their presence across vast gulfs of space.

Properties of a Typical White Dwarf and a Neutron Star

Property	White Dwarf	Neutron Star
Mass (Sun = 1)	0.6 (always <1.4)	Always >1.4 and <3
Radius	7000 km	10 km
Density	8×10^5 g/cm^3	10^{14} g/cm^3

Table 25.3

The Discovery of Neutron Stars

In 1967, Jocelyn Bell, a research student at Cambridge University, was studying distant radio sources with a special detector that had been designed and built by her advisor Antony Hewish to find rapid variations in radio signals. The project computers spewed out reams of paper showing where the telescope had surveyed the sky, and it was the job of Hewish's graduate students to go through it all, searching for interesting phenomena. In September 1967, Bell discovered what she called "a bit of scruff"—a strange radio signal unlike anything seen before.

What Bell had found, in the constellation of Vulpecula, was a source of rapid, sharp, intense, and extremely regular pulses of radio radiation. Like the regular ticking of a clock, the pulses arrived precisely every 1.33728 seconds. Such exactness first led the scientists to speculate that perhaps they had found signals from an intelligent civilization. Radio astronomers even half-jokingly dubbed the source "LGM" for "little green men." Soon, however, three similar sources were discovered in widely separated directions in the sky.

When it became apparent that this type of radio source was fairly common, astronomers concluded that they were highly unlikely to be signals from other civilizations. By today, more than 2500 such sources have been discovered; they are now called **pulsars**, short for "pulsating radio sources."

The pulse periods of different pulsars range from a little longer than 1/1000 of a second to nearly 10 seconds. At first, the pulsars seemed particularly mysterious because nothing could be seen at their location on visible-light photographs. But then a pulsar was discovered right in the center of the Crab Nebula, a cloud of gas produced by SN 1054, a supernova that was recorded by the Chinese in 1054 (Figure 25.14). The energy from the Crab Nebula pulsar arrives in sharp bursts that occur 30 times each second—with a regularity that would be the envy of a Swiss watchmaker. In addition to pulses of radio energy, we can observe pulses of visible light and X-rays from the Crab Nebula. The fact that the pulsar was just in the region of the supernova remnant where we expect the leftover neutron star to be immediately alerted astronomers that pulsars might be connected with these elusive "corpses" of massive stars.

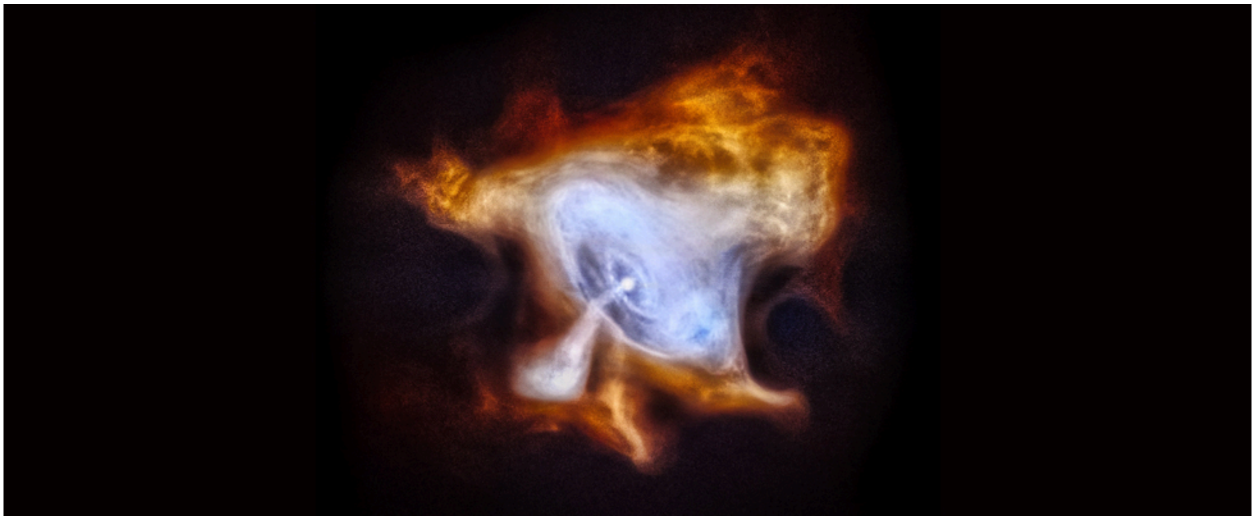


Figure 25.14 This image shows X-ray emmisions from the Crab Nebula, which is about 6500 light-years away. The pulsar is the bright spot at the center of the concentric rings. Data taken over about a year show that particles stream away from the inner ring at about half the speed of light. The jet that is perpendicular to this ring is a stream of matter and antimatter electrons also moving at half the speed of light. (credit: modification of work by NASA/CXC/SAO)

The Crab Nebula is a fascinating object. The whole nebula glows with radiation at many wavelengths, and its overall energy output is more than 100,000 times that of the Sun—not a bad trick for the remnant of a supernova that exploded almost a thousand years ago. Astronomers soon began to look for a connection between the pulsar and the large energy output of the surrounding nebula.

View an **interesting interview (https://openstaxcollege.org/l/30jocbellint)** with Jocelyn Bell (Burnell) to learn about her life and work (this is part of a project at the American Institute of Physics to record interviews with pathbreaking scientists while they are still alive).

A Spinning Lighthouse Model

By applying a combination of theory and observation, astronomers eventually concluded that pulsars must be *spinning neutron stars*. According to this model, a neutron star is something like a lighthouse on a rocky coast (**Figure 25.15**). To warn ships in all directions and yet not cost too much to operate, the light in a modern lighthouse turns, sweeping its beam across the dark sea. From the vantage point of a ship, you see a pulse of light each time the beam points in your direction. In the same way, radiation from a small region on a neutron star sweeps across the oceans of space, giving us a pulse of radiation each time the beam points toward Earth.

Figure 25.15 A lighthouse in California warns ships on the ocean not to approach too close to the dangerous shoreline. The lighted section at the top rotates so that its beam can cover all directions. (credit: Anita Ritenour)

Neutron stars are ideal candidates for such a job because the collapse has made them so small that they can turn very rapidly. Recall the principle of **Conservation of Angular Momentum**: if an object gets smaller, it can spin more rapidly. Even if the parent star was rotating very slowly when it was on the main sequence, its rotation had to speed up as it collapsed to form a neutron star. With a diameter of only 10 to 20 kilometers, a neutron star can complete one full spin in only a fraction of a second. This is just the sort of time period we observe between pulsar pulses.

Any magnetic field that existed in the original star will be highly compressed when the core collapses to a neutron star. At the surface of the neutron star, in the outer layer consisting of ordinary matter (and not just pure neutrons), protons and electrons are caught up in this spinning field and accelerated nearly to the speed of light. In only two places—the north and south magnetic poles—can the trapped particles escape the strong hold of the magnetic field (**Figure 25.16**). The same effect can be seen (in reverse) on Earth, where charged particles from space are *kept out* by our planet's magnetic field everywhere except near the poles. As a result, Earth's auroras (caused when charged particles hit the atmosphere at high speed) are seen mainly near the poles.

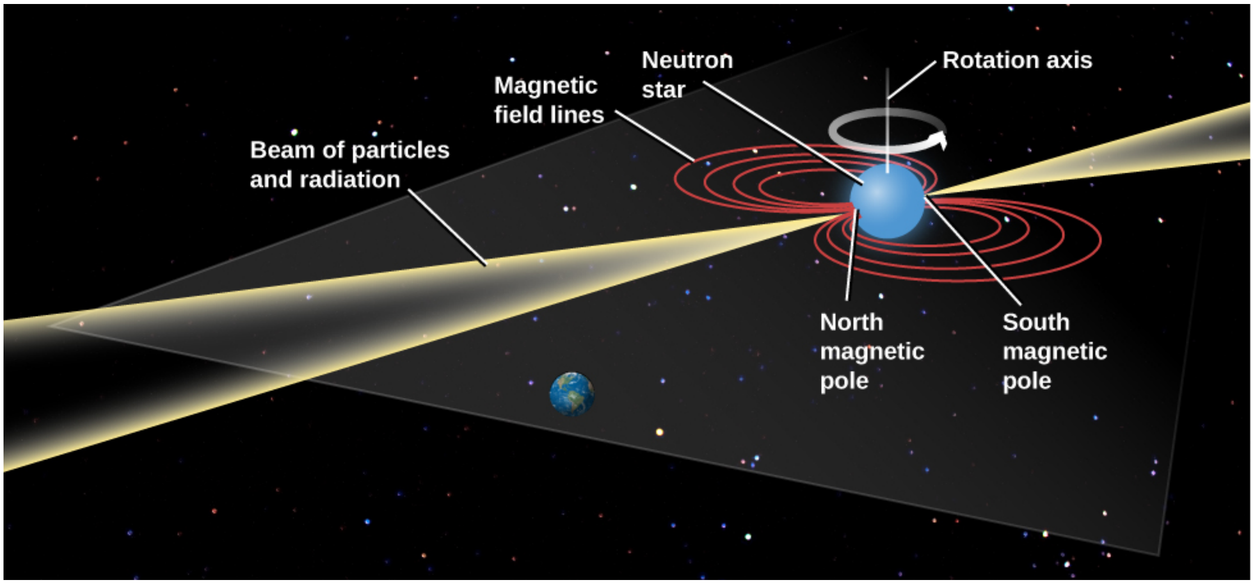


Figure 25.16 A diagram showing how beams of radiation at the magnetic poles of a neutron star can give rise to pulses of emission as the star rotates. As each beam sweeps over Earth, like a lighthouse beam sweeping over a distant ship, we see a short pulse of radiation. This model requires that the magnetic poles be located in different places from the rotation poles. (credit "stars": modification of work by Tony Hisgett)

Note that in a neutron star, the magnetic north and south poles do not have to be anywhere close to the north and south poles defined by the star's rotation.

In fact, the misalignment of the rotational axis with the magnetic axis plays a crucial role in the generation of the observed pulses in this model. At the two magnetic poles, the particles from the neutron star are focused into a narrow beam and come streaming out of the whirling magnetic region at enormous speeds. They emit energy over a broad range of the electromagnetic spectrum. The radiation itself is also confined to a narrow beam, which explains why the pulsar acts like a lighthouse. As the rotation carries first one and then the other magnetic pole of the star into our view, we see a pulse of radiation each time.

Tests of the Model

This explanation of pulsars in terms of beams of radiation from highly magnetic and rapidly spinning neutron stars is a very clever idea. But what evidence do we have that it is the correct model? First, we can measure the masses of some pulsars, and they do turn out be in the range of 1.4 to 1.8 times that of the Sun—just what theorists predict for neutron stars. The masses are found using Kepler's law for those few pulsars that are members of binary star systems.

But there is an even-better confirming argument, which brings us back to the Crab Nebula and its vast energy output. When the high-energy charged particles from the neutron star pulsar hit the slower-moving material from the supernova, they energize this material and cause it to "glow" at many different wavelengths—just what we observe from the Crab Nebula. The pulsar beams are a power source that "light up" the nebula long after the initial explosion of the star that made it.

Who "pays the bills" for all the energy we see coming out of a remnant like the Crab Nebula? After all, when energy emerges from one place, it must be depleted in another. The ultimate energy source in our model is the rotation of the neutron star, which propels charged particles outward and spins its magnetic field at enormous speeds. As its rotational energy is used to excite the Crab Nebula year after year, the pulsar inside the nebula slows down. As it slows, the pulses come a little less often; more time elapses before the slower neutron star brings its beam back around.

Several decades of careful observations have now shown that the Crab Nebula pulsar is not a perfectly regular clock as we originally thought: instead, it is gradually slowing down. Having measured how much the pulsar is slowing down, we can calculate how much rotation energy the neutron star is losing. Remember that it is very densely packed and spins amazingly quickly. Even a tiny slowing down can mean an immense loss of energy.

To the satisfaction of astronomers, the rotational energy lost by the pulsar turns out to be the same as the amount of energy emerging from the nebula surrounding it. In other words, the slowing down of a rotating neutron star can explain precisely why the Crab Nebula is glowing with the amount of energy we observe.

The Evolution of Pulsars

From observations of the pulsars discovered so far, astronomers have concluded that one new pulsar is born somewhere in the Galaxy every 25 to 100 years, the same rate at which supernovae are estimated to occur. Calculations suggest that the typical lifetime of a pulsar is about 10 million years; after that, the neutron star no longer rotates fast enough to produce significant beams of particles and energy, and is no longer observable. We estimate that there are about 100 million neutron stars in our Galaxy, most of them rotating too slowly to come to our notice.

The Crab pulsar is rather young (only about 960 years old) and has a short period, whereas other, older pulsars have already slowed to longer periods. Pulsars thousands of years old have lost too much energy to emit appreciably in the visible and X-ray wavelengths, and they are observed only as radio pulsars; their periods are a second or longer.

There is one other reason we can see only a fraction of the pulsars in the Galaxy. Consider our lighthouse model again. On Earth, all ships approach on the same plane—the surface of the ocean—so the lighthouse can be built to sweep its beam over that surface. But in space, objects can be anywhere in three dimensions. As a given pulsar's beam sweeps over a circle in space, there is absolutely no guarantee that this circle will include the direction of Earth. In fact, if you think about it, many more circles in space will *not* include Earth than will include it. Thus, we estimate that we are unable to observe a large number of neutron stars because their pulsar beams miss us entirely.

At the same time, it turns out that only a few of the pulsars discovered so far are embedded in the visible clouds of gas that mark the remnant of a supernova. This might at first seem mysterious, since we know that supernovae give rise to neutron stars and we should expect each pulsar to have begun its life in a supernova explosion. But the lifetime of a pulsar turns out to be about 100 times longer than the length of time required for the expanding gas of a supernova remnant to disperse into interstellar space. Thus, most pulsars are found with no other trace left of the explosion that produced them.

In addition, some pulsars are ejected by a supernova explosion that is not the same in all directions. If the supernova explosion is stronger on one side, it can kick the pulsar entirely out of the supernova remnant (some astronomers call this "getting a birth kick"). We know such kicks happen because we see a number of young supernova remnants in nearby galaxies where the pulsar is to one side of the remnant and racing away at several hundred miles per second (**Figure 25.17**).

Figure 25.17 This intriguing image (which combines X-ray, visible, and radio observations) shows the jet trailing behind a pulsar (at bottom right, lined up between the two bright stars). With a length of 37 light-years, the jet trail (seen in purple) is the longest ever observed from an object in the Milky Way. (There is also a mysterious shorter, comet-like tail that is almost perpendicular to the purple jet.) Moving at a speed between 2.5 and 5 million miles per hour, the pulsar is traveling away from the core of the supernova remnant where it originated. (credit: X-ray: NASA/CXC/ISDC/L.Pavan et al, Radio: CSIRO/ATNF/ATCA Optical: 2MASS/UMass/IPAC-Caltech/NASA/NSF)

Touched by a Neutron Star

On December 27, 2004, Earth was bathed with a stream of X-ray and gamma-ray radiation from a neutron star known as SGR 1806-20. What made this event so remarkable was that, despite the distance of the source, its tidal wave of radiation had measurable effects on Earth's atmosphere. The apparent brightness of this gamma-ray flare was greater than any historical star explosion.

The primary effect of the radiation was on a layer high in Earth's atmosphere called the *ionosphere*. At night, the ionosphere is normally at a height of about 85 kilometers, but during the day, energy from the Sun ionizes more molecules and lowers the boundary of the ionosphere to a height of about 60 kilometers. The pulse of X-ray and gamma-ray radiation produced about the same level of ionization as the daytime Sun. It also caused some sensitive satellites above the atmosphere to shut down their electronics.

Measurements by telescopes in space indicate that SGR 1806-20 was a special type of fast-spinning neutron star called a *magnetar*. Astronomers Robert Duncan and Christopher Thomson gave them this name because their magnetic fields are stronger than that of any other type of astronomical source—in this case, about 800 trillion times stronger than the magnetic field of Earth.

A magnetar is thought to consist of a superdense core of neutrons surrounded by a rigid crust of atoms about a mile deep with a surface made of iron. The magnetar's field is so strong that it creates huge stresses inside that can sometimes crack open the hard crust, causing a starquarke. The vibrating crust produces an enormous blast of radiation. An astronaut 0.1 light-year from this particular magnetar would have received a fatal does from the blast in less than a second.

Fortunately, we were far enough away from magnetar SGR 1806-20 to be safe. Could a magnetar ever present a real danger to Earth? To produce enough energy to disrupt the ozone layer, a magnetar would have to be located within the cloud of comets that surround the solar system, and we know no magnetars are that close. Nevertheless, it is a fascinating discovery that events on distant star corpses can have measurable effects on Earth.

25.5 | The Evolution of Binary Star Systems

Learning Objectives

By the end of this section, you will be able to:

- Describe the kind of binary star system that leads to a nova event
- Describe the type of binary star system that leads to a type Ia supernovae event
- Indicate how type Ia supernovae differ from type II supernovae

The discussion of the life stories of stars presented so far has suffered from a bias—what we might call "single-star chauvinism." Because the human race developed around a star that goes through life alone, we tend to think of most stars in isolation. But as we saw in **Stellar Properties**, it now appears that as many as half of all stars may develop in *binary* systems—those in which two stars are born in each other's gravitational embrace and go through life orbiting a common center of mass.

For these stars, the presence of a close-by companion can have a profound influence on their evolution. Under the right circumstances, stars can exchange material, especially during the stages when one of them swells up into a giant or supergiant, or has a strong wind. When this happens and the companion stars are sufficiently close, material can flow from one star to another, decreasing the mass of the donor and increasing the mass of the recipient. Such *mass transfer* can be especially dramatic when the recipient is a stellar remnant such as a white dwarf or a neutron star. While the detailed story of how such binary stars evolve is beyond the scope of our book, we do want to mention a few examples of how the stages of evolution described in this chapter may change when there are two stars in a system.

White Dwarf Explosions: The Mild Kind

Let's consider the following system of two stars: one has become a white dwarf and the other is gradually transferring material onto it. As fresh hydrogen from the outer layers of its companion accumulates on the surface of the hot white dwarf, it begins to build up a layer of hydrogen. As more and more hydrogen accumulates and heats up on the surface of the degenerate star, the new layer eventually reaches a temperature that causes fusion to begin in a sudden, explosive way, blasting much of the new material away.

In this way, the white dwarf quickly (but only briefly) becomes quite bright, hundreds or thousands of times its previous luminosity. To observers before the invention of the telescope, it seemed that a new star suddenly appeared, and they called it a **nova**.[2] Novae fade away in a few months to a few years.

Hundreds of novae have been observed, each occurring in a binary star system and each later showing a shell of expelled material. A number of stars have more than one nova episode, as more material from its neighboring star accumulates on the white dwarf and the whole process repeats. As long as the episodes do not increase the mass of the white dwarf beyond the Chandrasekhar limit (by transferring too much mass too quickly), the dense white dwarf itself remains pretty much unaffected by the explosions on its surface.

White Dwarf Explosions: The Violent Kind

If a white dwarf accumulates matter from a companion star at a much faster rate, it can be pushed over the Chandrasekhar limit. The evolution of such a binary system is shown in **Figure 25.18**. When its mass approaches the Chandrasekhar mass limit (exceeds 1.4 M_{Sun}), such an object can no longer support itself as a white dwarf, and it begins to contract. As it does so, it heats up, and new nuclear reactions can begin in the degenerate core. The star "simmers" for the next century or so, building up internal temperature. This simmering phase ends in less than a second, when an enormous amount of fusion (especially of carbon) takes place all at once, resulting in an explosion. The fusion energy produced during the final explosion is so great that it completely destroys the white dwarf. Gases are blown out into space at velocities of about 10,000 kilometers per second, and afterward, no trace of the white dwarf remains.

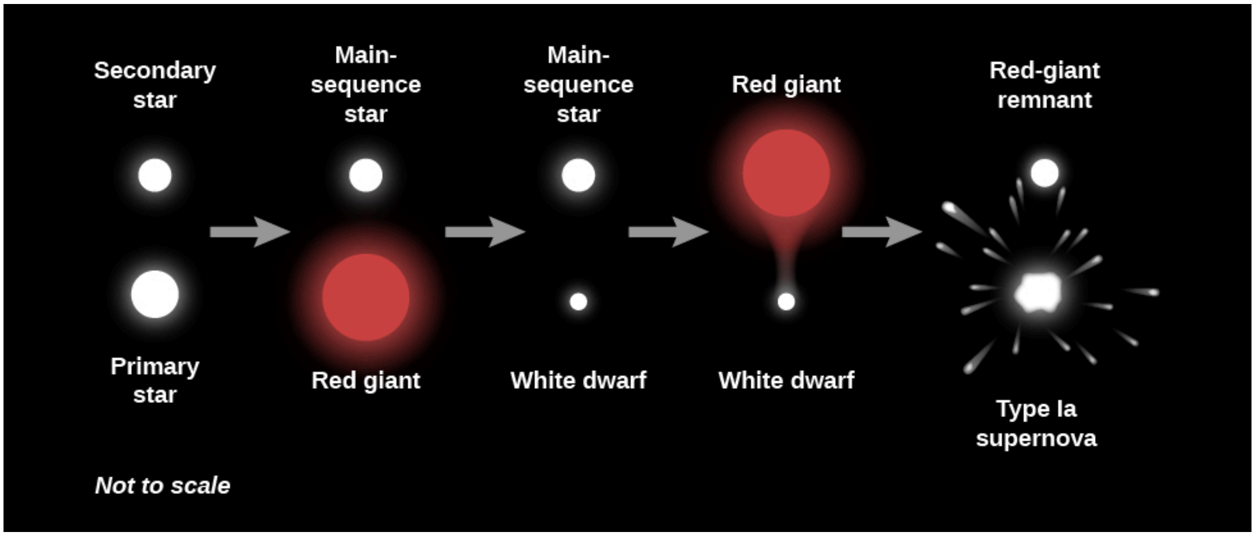


Figure 25.18 The more massive star evolves first to become a red giant and then a white dwarf. The white dwarf then begins to attract material from its companion, which in turn evolves to become a red giant. Eventually, the white dwarf acquires so much mass that it is pushed over the Chandrasekhar limit and becomes a type Ia supernova.

Such an explosion is also called a supernova, since, like the destruction of a high-mass star, it produces a huge amount of energy in a very short time. However, unlike the explosion of a high-mass star, which can leave behind a neutron star or black hole remnant, the white dwarf is completely destroyed in the process, leaving behind no remnant. We call these white dwarf explosions type Ia supernovae.

We distinguish type I supernovae from those of supernovae of type II originating from the death of massive stars discussed earlier by the absence of hydrogen in their observed spectra. Hydrogen is the most common element in the universe and is a major component of massive, evolved stars. However, as we learned earlier, hydrogen is absent from the white dwarf remnant, which is primarily composed of carbon and oxygen for masses comparable to the Chandrasekhar mass limit.

The "a" subdesignation of type Ia supernovae further refers to the presence of strong silicon absorption lines, which are absent from supernovae originating from the collapse of massive stars. Silicon is one of the products that results from the fusion of carbon and oxygen, which bears out the scenario we described above—that there is a sudden onset of the fusion of the carbon (and oxygen) of which the white dwarf was made.

2. We now know that this historical terminology is quite misleading since novae do not originate from new stars. In fact, quite to the contrary, novae originate from white dwarfs, which are actually the endpoint of stellar evolution for low-mass stars. But since the system of two stars was too faint to be visible to the naked eye, it did seem to people, before telescopes were invented, that a star had appeared where nothing had been visible.

Observational evidence now strongly indicates that SN 1006, Tycho's Supernova, and Kepler's Supernova (see **Supernovae in History**) were all type Ia supernovae. For instance, in contrast to the case of SN 1054, which yielded the spinning pulsar in the Crab Nebula, none of these historical supernovae shows any evidence of stellar remnants that have survived their explosions. Perhaps even more puzzling is that, so far, astronomers have not been able to identify the companion star feeding the white dwarf in any of these historical supernovae.

Consequently, in order to address the mystery of the absent companion stars and other outstanding puzzles, astronomers have recently begun to investigate alternative mechanisms of generating type Ia supernovae. All proposed mechanisms rely upon white dwarfs composed of carbon and oxygen, which are needed to meet the observed absence of hydrogen in the type Ia spectrum. And because any isolated white dwarf below the Chandrasekhar mass is stable, all proposed mechanisms invoke a binary companion to explode the white dwarf. The leading alternative mechanism scientists believe creates a type Ia supernova is the merger of two white dwarf stars in a binary system. The two white dwarfs may have unstable orbits, such that over time, they would slowly move closer together until they merge. If their combined mass is greater than the Chandrasekhar limit, the result could also be a type Ia supernova explosion.

You can watch a **short video (https://openstaxcollege.org/l/30supernovavid)** about Supernova SN 2014J, a type Ia supernova discovered in the Messier 82 (M82) galaxy on January 21, 2014, as well as see brief animations of the two mechanisms by which such a supernova could form.

Type Ia supernovae are of great interest to astronomers in other areas of research. This type of supernova is brighter than supernovae produced by the collapse of a massive star. Thus, type Ia supernovae can be seen at very large distances, and they are found in all types of galaxies. The energy output from most type Ia supernovae is consistent, with little variation in their maximum luminosities, or in how their light output initially increases and then slowly decreases over time. These properties make type Ia supernovae extremely valuable "standard bulbs" for astronomers looking out at great distances—well beyond the limits of our own Galaxy. You'll learn more about their use in measuring distances to other galaxies in **The Extragalactic Distance Scale**.

In contrast, type II supernovae are about 5 times less luminous than type Ia supernovae and are only seen in galaxies that have recent, massive star formation. Type II supernovae are also less consistent in their energy output during the explosion and can have a range a peak luminosity values.

Neutron Stars with Companions

Now let's look at an even-more mismatched pair of stars in action. It is possible that, under the right circumstances, a binary system can even survive the explosion of one of its members as a type II supernova. In that case, an ordinary star can eventually share a system with a neutron star. If material is then transferred from the "living" star to its "dead" (and highly compressed) companion, this material will be pulled in by the strong gravity of the neutron star. Such infalling gas will be compressed and heated to incredible temperatures. It will quickly become so hot that it will experience an explosive burst of fusion. The energies involved are so great that we would expect much of the radiation from the burst to emerge as X-rays. And indeed, high-energy observatories above Earth's atmosphere have recorded many objects that undergo just these types of X-ray *bursts*.

If the neutron star and its companion are positioned the right way, a significant amount of material can be transferred to the neutron star and can set it spinning faster (as spin energy is also transferred). The radius of the neutron star would also decrease as more mass was added. Astronomers have found pulsars in binary systems that are spinning at a rate of more than 500 times per second! (These are sometimes called **millisecond pulsars** since the pulses are separated by a few thousandths of a second.)

Such a rapid spin could not have come from the birth of the neutron star; it must have been externally caused. (Recall that the Crab Nebula pulsar, one of the youngest pulsars known, was spinning "only" 30 times per second.) Indeed, some of the fast pulsars are observed to be part of binary systems, while others may be alone only because they have "fully consumed" their former partner stars through the mass transfer process. (These have sometimes been called " black widow pulsars.")

View this **short video (https://openstaxcollege.org/l/30scotronvid)** to see Dr. Scott Ransom, of the National Radio Astronomy Observatory, explain how millisecond pulsars come about, with some nice animations.

And if you thought that a neutron star interacting with a "normal" star was unusual, there are also binary systems that consist of two neutron stars. One such system has the stars in very close orbits to one another, so much that they continually alter each other's orbit. Another binary neutron star system includes two pulsars that are orbiting each other every 2 hours and 25 minutes. As we discussed earlier, pulsars radiate away their energy, and these two pulsars are slowly moving toward one another, such that in about 85 million years, they will actually merge (see Gravitational Wave Astronomy for our first observations of such a merger).

We have now reached the end of our description of the final stages of stars, yet one piece of the story remains to be filled in. We saw that stars whose core masses are less than 1.4 M_{Sun} at the time they run out of fuel end their lives as white dwarfs. Dying stars with core masses between 1.4 and about 3 M_{Sun} become neutron stars. But there are stars whose core masses are greater than 3 M_{Sun} when they exhaust their fuel supplies. What becomes of them? The truly bizarre result of the death of such massive stellar cores (called a black hole) is the subject of our next chapter. But first, we will look at an astronomical mystery that turned out to be related to the deaths of stars and was solved through clever sleuthing and a combination of observation and theory.

25.6 | Introducing General Relativity

Learning Objectives

By the end of this section, you will be able to:

- Discuss some of the key ideas of the theory of general relativity
- Recognize that one's experiences of gravity and acceleration are interchangeable and indistinguishable
- Distinguish between Newtonian ideas of gravity and Einsteinian ideas of gravity
- Recognize why the theory of general relativity is necessary for understanding the nature of black holes
- Describe Einstein's view of gravity as the warping of spacetime in the presence of massive objects
- Understand that Newton's concept of the gravitational force between two massive objects and Einstein's concept of warped spacetime are different explanations for the same observed accelerations of one massive object in the presence of another massive object

Most stars end their lives as white dwarfs or neutron stars. When a *very* massive star collapses at the end of its life, however, not even the mutual repulsion between densely packed neutrons can support the core against its own weight. If the remaining mass of the star's core is more than about three times that of the Sun (M_{Sun}), our theories predict that *no known force can stop it from collapsing forever!* Gravity simply overwhelms all other forces and crushes the core until it occupies an infinitely small volume. A star in which this occurs may become one of the strangest objects ever predicted by theory—a black hole.

To understand what a black hole is like and how it influences its surroundings, we need a theory that can describe the action of gravity under such extreme circumstances. To date, our best theory of gravity is the **general theory of relativity**, which was put forward in 1916 by Albert Einstein.

General relativity was one of the major intellectual achievements of the twentieth century; if it were music, we would compare it to the great symphonies of Beethoven or Mahler. Until recently, however, scientists had little need for a better theory of gravity; Isaac Newton's ideas that led to his **Law of Universal Gravitation** are perfectly sufficient for most of the objects we deal with in everyday life. In the past half century, however, general relativity has become more than just a beautiful idea; it is now essential in understanding pulsars, quasars (which will be discussed in **The Evolution and Distribution of Galaxies**), and many other astronomical objects and events, including the black holes we will discuss here.

We should perhaps mention that this is the point in an astronomy course when many students start to feel a little nervous (and perhaps wish they had taken botany or some other earthbound course to satisfy the science requirement). This is because in popular culture, Einstein has become a symbol for mathematical brilliance that is simply beyond the reach of most people (**Figure 25.19**).

Figure 25.19 This famous scientist, seen here younger than in the usual photos, has become a symbol for high intellect in popular culture. (credit: NASA)

So, when we wrote that the theory of general relativity was Einstein's work, you may have worried just a bit, convinced that anything Einstein did must be beyond your understanding. This popular view is unfortunate and mistaken. Although the detailed calculations of general relativity do involve a good deal of higher mathematics, the basic ideas are not difficult to understand (and are, in fact, almost poetic in the way they give us a new perspective on the world). Moreover, general relativity goes beyond Newton's famous "inverse-square" law of gravity; it helps *explain* how matter interacts with other matter in space and time. This explanatory power is one of the requirements that any successful scientific theory must meet.

The Principle of Equivalence

The fundamental insight that led to the formulation of the general theory of relativity starts with a very simple thought: if you were able to jump off a high building and fall freely, you would not feel your own weight. In this chapter, we will describe how Einstein built on this idea to reach sweeping conclusions about the very fabric of space and time itself. He called it the "happiest thought of my life."

Einstein himself pointed out an everyday example that illustrates this effect (see **Figure 25.20**). Notice how your weight seems to be reduced in a high-speed elevator when it accelerates from a stop to a rapid descent. Similarly, your weight seems to increase in an elevator that starts to move quickly upward. This effect is not just a feeling you have: if you stood on a scale in such an elevator, you could measure your weight changing (you can actually perform this experiment in some science museums).

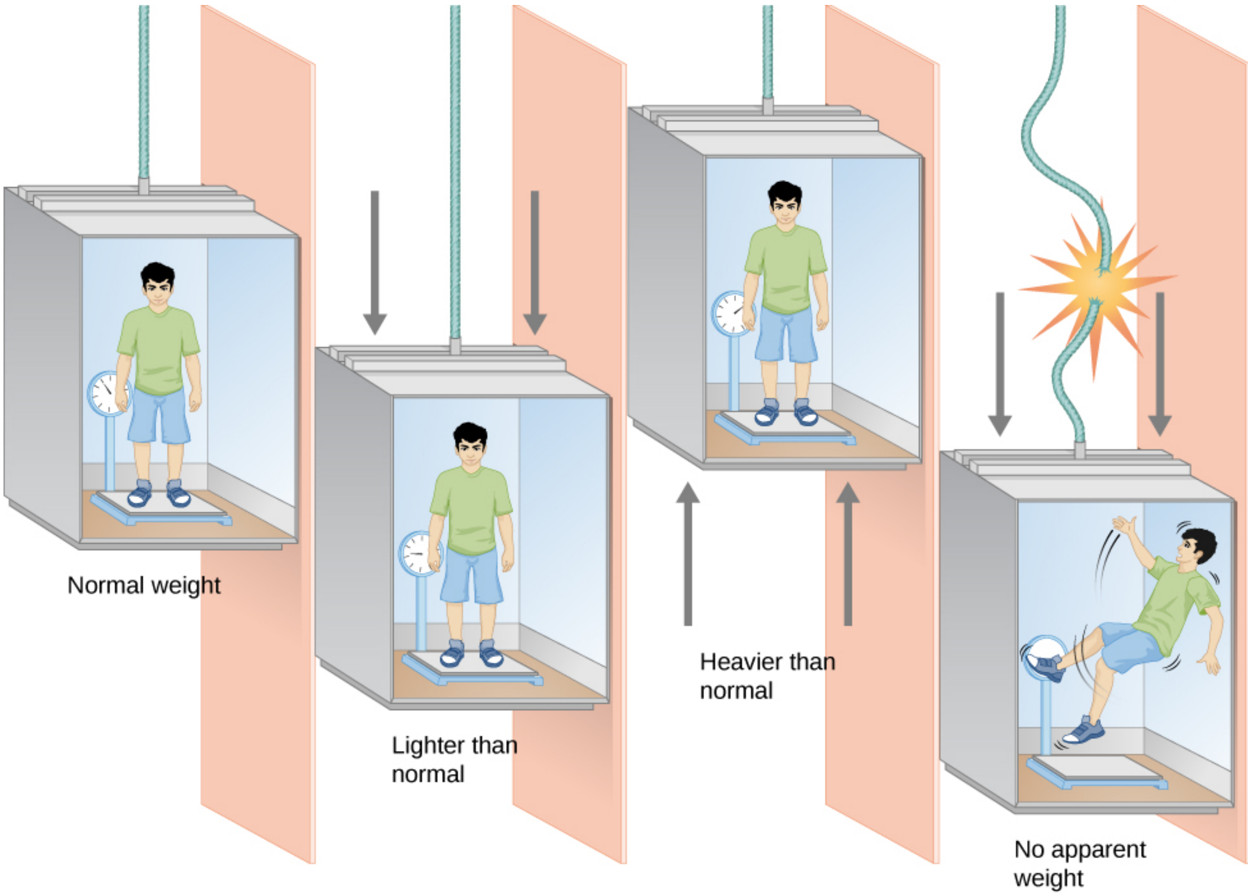


Figure 25.20 In an elevator at rest, you feel your normal weight. In an elevator that accelerates as it descends, you would feel lighter than normal. In an elevator that accelerates as it ascends, you would feel heavier than normal. If an evil villain cut the elevator cable, you would feel weightless as you fell to your doom.

In a *freely falling* elevator, with no air friction, you would lose your weight altogether. We generally don't like to cut the cables holding elevators to try this experiment, but near-weightlessness can be achieved by taking an airplane to high altitude and then dropping rapidly for a while. This is how NASA trains its astronauts for the experience of free fall in space; the scenes of weightlessness in the 1995 movie *Apollo 13* were filmed in the same way. (Moviemakers have since devised other methods using underwater filming, wire stunts, and computer graphics to create the appearance of weightlessness seen in such movies as *Gravity* and *The Martian*.)

Watch how NASA uses a **"weightless" environment (https://openstax.org/l/30NASAweightra)** to help train astronauts.

Another way to state Einstein's idea is this: suppose we have a spaceship that contains a windowless laboratory equipped with all the tools needed to perform scientific experiments. Now, imagine that an astronomer wakes up after a long night celebrating some scientific breakthrough and finds herself sealed into this laboratory. She has no idea how it happened but notices that she is weightless. This could be because she and the laboratory are far away from any source of gravity, and both are either at rest or moving at some steady speed through space (in which case she has plenty of time to wake up). But it could also be because she and the laboratory are falling freely toward a planet like Earth (in which case she might first want to check her distance from the surface before making coffee).

What Einstein postulated is that there is no experiment she can perform inside the sealed laboratory to determine whether she is floating in space or falling freely in a gravitational field.[3] As far as she is concerned, the two situations are

3. Strictly speaking, this is true only if the laboratory is infinitesimally small. Different locations in a real laboratory that is falling freely due to gravity cannot all be at identical distances from the object(s) responsible for producing the gravitational force. In this case, objects in different locations will experience slightly different accelerations. But this point does not invalidate the principle of equivalence that Einstein derived from this line of thinking.

completely *equivalent*. This idea that free fall is indistinguishable from, and hence equivalent to, zero gravity is called the **equivalence principle**.

Gravity or Acceleration?

Einstein's simple idea has big consequences. Let's begin by considering what happens if two foolhardy people jump from opposite banks into a bottomless chasm (**Figure 25.21**). If we ignore air friction, then we can say that while they freely fall, they both accelerate downward at the same rate and feel no external force acting on them. They can throw a ball back and forth, always aiming it straight at each other, as if there were no gravity. The ball falls at the same rate that they do, so it always remains in a line between them.

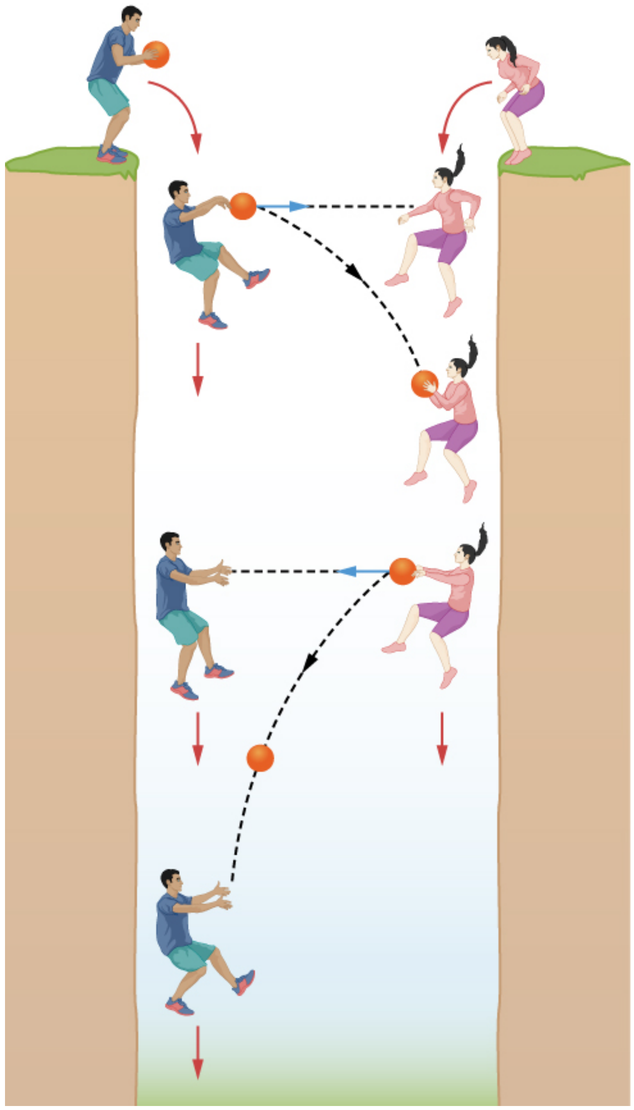


Figure 25.21 Two people play catch as they descend into a bottomless abyss. Since the people and ball all fall at the same speed, it appears to them that they can play catch by throwing the ball in a straight line between them. Within their frame of reference, there appears to be no gravity.

Such a game of catch is very different on the surface of Earth. Everyone who grows up feeling gravity knows that a ball, once thrown, falls to the ground. Thus, in order to play catch with someone, you must aim the ball upward so that it follows an arc—rising and then falling as it moves forward—until it is caught at the other end.

Now suppose we isolate our falling people and ball inside a large box that is falling with them. No one inside the box is aware of any gravitational force. If they let go of the ball, it doesn't fall to the bottom of the box or anywhere else but merely stays there or moves in a straight line, depending on whether it is given any motion.

Astronauts in the International Space Station (ISS) that is orbiting Earth live in an environment just like that of the people sealed in a freely falling box (**Figure 25.22**). The orbiting ISS is actually “falling” freely around Earth. While in free fall, the astronauts live in a strange world where there seems to be no gravitational force. One can give a wrench a shove, and it moves at constant speed across the orbiting laboratory. A pencil set in midair remains there as if no force were acting on it.

Figure 25.22 Shane Kimbrough and Sandra Magnus are shown aboard the Endeavour in 2008 with various fruit floating freely. Because the shuttle is in free fall as it orbits Earth, everything—including astronauts—stays put or moves uniformly relative to the walls of the spacecraft. This free-falling state produces a lack of apparent gravity inside the spacecraft. (credit: NASA)

In the “weightless” environment of the International Space Station, moving takes very little effort. Watch **astronaut Karen Nyberg (https://openstax.org/l/30ISSzerogravid)** demonstrate how she can propel herself with the force of a single human hair.

Appearances are misleading, however. There is a force in this situation. Both the ISS and the astronauts continually fall around Earth, pulled by its gravity. But since all fall together—shuttle, astronauts, wrench, and pencil—inside the ISS all gravitational forces appear to be absent.

Thus, the orbiting ISS provides an excellent example of the principle of equivalence—how local effects of gravity can be completely compensated by the right acceleration. To the astronauts, falling around Earth creates the same effects as being far off in space, remote from all gravitational influences.

The Paths of Light and Matter

Einstein postulated that the equivalence principle is a fundamental fact of nature, and that there is *no* experiment inside any spacecraft by which an astronaut can ever distinguish between being weightless in remote space and being in free fall near a planet like Earth. This would apply to experiments done with beams of light as well. But the minute we use light in our experiments, we are led to some very disturbing conclusions—and it is these conclusions that lead us to general relativity and a new view of gravity.

It seems apparent to us, from everyday observations, that beams of light travel in straight lines. Imagine that a spaceship is moving through empty space far from any gravity. Send a laser beam from the back of the ship to the front, and it will travel in a nice straight line and land on the front wall exactly opposite the point from which it left the rear wall. If the equivalence principle really applies universally, then this same experiment performed in free fall around Earth should give us the same result.

Now imagine that the astronauts again shine a beam of light along the length of their ship. But, as shown in **Figure 25.23**, this time the orbiting space station falls a bit between the time the light leaves the back wall and the time it hits the front wall. (The amount of the fall is grossly exaggerated in **Figure 25.23** to illustrate the effect.) Therefore, if the beam of light follows a straight line but the ship’s path curves downward, then the light should strike the front wall at a point higher than the point from which it left.

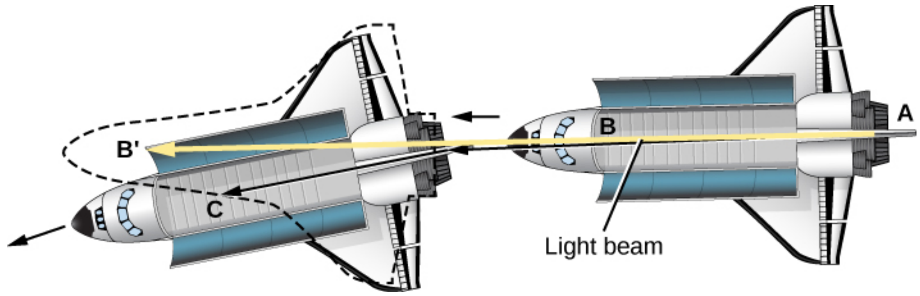


Figure 25.23 In a spaceship moving to the left (in this figure) in its orbit about a planet, light is beamed from the rear, A, toward the front, B. Meanwhile, the ship is falling out of its straight path (exaggerated here). We might therefore expect the light to strike at B′, above the target in the ship. Instead, the light follows a curved path and strikes at C. In order for the principle of equivalence to be correct, gravity must be able to curve the path of a light beam just as it curves the path of the spaceship.

However, this would violate the principle of equivalence—the two experiments would give different results. We are thus faced with giving up one of our two assumptions. Either the principle of equivalence is not correct, or light does not always travel in straight lines. Instead of dropping what probably seemed at the time like a ridiculous idea, Einstein worked out what happens if light sometimes does *not* follow a straight path.

Let's suppose the principle of equivalence is right. Then the light beam must arrive directly opposite the point from which it started in the ship. The light, like the ball thrown back and forth, *must fall with the ship* that is in orbit around Earth (see **Figure 25.23**). This would make its path curve downward, like the path of the ball, and thus the light would hit the front wall exactly opposite the spot from which it came.

Thinking this over, you might well conclude that it doesn't seem like such a big problem: why *can't* light fall the way balls do? But light is profoundly different from balls. Balls have mass, while light does not.

Here is where Einstein's intuition and genius allowed him to make a profound leap. He gave physical meaning to the strange result of our thought experiment. Einstein suggested that the light curves down to meet the front of the shuttle because Earth's gravity actually bends the *fabric of space and time*. This radical idea—which we will explain next—keeps the behavior of light the same in both empty space and free fall, but it changes some of our most basic and cherished ideas about space and time. The reason we take Einstein's suggestion seriously is that, as we will see, experiments now clearly show his intuitive leap was correct.

Is light actually bent from its straight-line path by the mass of Earth? How can light, which has no mass, be affected by gravity? Einstein preferred to think that it is *space and time* that are affected by the presence of a large mass; light beams, and everything else that travels through space and time, then find their paths affected. Light always follows the shortest path—but that path may not always be straight. This idea is true for human travel on the curved surface of planet Earth, as well. Say you want to fly from Chicago to Rome. Since an airplane can't go through the solid body of the Earth, the shortest distance is not a straight line but the arc of a *great circle*.

Linkages: Mass, Space, and Time

To show what Einstein's insight really means, let's first consider how we locate an event in space and time. For example, imagine you have to describe to worried school officials the fire that broke out in your room when your roommate tried cooking shish kebabs in the fireplace. You explain that your dorm is at 6400 College Avenue, a street that runs in the left-right direction on a map of your town; you are on the fifth floor, which tells where you are in the up-down direction; and you are the sixth room back from the elevator, which tells where you are in the forward-backward direction. Then you explain that the fire broke out at 6:23 p.m. (but was soon brought under control), which specifies the event in time. *Any* event in the universe, whether nearby or far away, can be pinpointed using the three dimensions of space and the one dimension of time.

Newton considered space and time to be completely independent, and that continued to be the accepted view until the beginning of the twentieth century. But Einstein showed that there is an intimate connection between space and time, and that only by considering the two together—in what we call **spacetime**—can we build up a correct picture of the physical world. We examine spacetime a bit more closely in the next subsection.

The gist of Einstein's general theory is that the presence of matter curves or warps the fabric of spacetime. This curving of spacetime is identified with gravity. When something else—a beam of light, an electron, or the starship *Enterprise*—enters such a region of distorted spacetime, its path will be different from what it would have been in the absence of the matter. As American physicist John Wheeler summarized it: "Matter tells spacetime how to curve; spacetime tells matter how to move."

The amount of distortion in spacetime depends on the mass of material that is involved and on how concentrated and compact it is. Terrestrial objects, such as the book you are reading, have far too little mass to introduce any significant distortion. Newton's view of gravity is just fine for building bridges, skyscrapers, or amusement park rides. General relativity does, however, have some practical applications. The GPS (Global Positioning System) in every smartphone can tell you where you are within 5 to 10 meters only because the effects of general and special relativity on the GPS satellites in orbit around the Earth are taken into account.

Unlike a book or your roommate, stars produce measurable distortions in spacetime. A white dwarf, with its stronger surface gravity, produces more distortion just above its surface than does a red giant with the same mass. So, you see, we *are* eventually going to talk about collapsing stars again, but not before discussing Einstein's ideas (and the evidence for them) in more detail.

Spacetime Examples

How can we understand the distortion of spacetime by the presence of some (significant) amount of mass? Let's try the following analogy. You may have seen maps of New York City that squeeze the full three dimensions of this towering metropolis onto a flat sheet of paper and still have enough information so tourists will not get lost. Let's do something similar with diagrams of spacetime.

Figure 25.24, for example, shows the progress of a motorist driving east on a stretch of road in Kansas where the countryside is absolutely flat. Since our motorist is traveling only in the east-west direction and the terrain is flat, we can ignore the other two dimensions of space. The amount of time elapsed since he left home is shown on the *y*-axis, and the distance traveled eastward is shown on the *x*-axis. From A to B he drove at a uniform speed; unfortunately, it was too fast a uniform speed and a police car spotted him. From B to C he stopped to receive his ticket and made no progress through space, only through time. From C to D he drove more slowly because the police car was behind him.

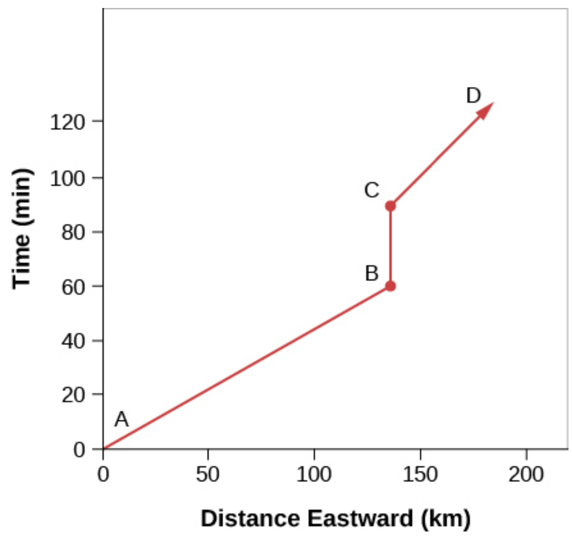


Figure 25.24 This diagram shows the progress of a motorist traveling east across the flat Kansas landscape. Distance traveled is plotted along the horizontal axis. The time elapsed since the motorist left the starting point is plotted along the vertical axis.

Now let's try illustrating the distortions of spacetime in two dimensions. In this case, we will (in our imaginations) use a rubber sheet that can stretch or warp if we put objects on it.

Let's imagine stretching our rubber sheet taut on four posts. To complete the analogy, we need something that normally travels in a straight line (as light does). Suppose we have an extremely intelligent ant—a friend of the comic book superhero Ant-Man, perhaps—that has been trained to walk in a straight line.

We begin with just the rubber sheet and the ant, simulating empty space with no mass in it. We put the ant on one side of the sheet and it walks in a beautiful straight line over to the other side (**Figure 25.25**). We next put a small grain of sand on the rubber sheet. The sand does distort the sheet a tiny bit, but this is not a distortion that we or the ant can measure. If we send the ant so it goes close to, but not on top of, the sand grain, it has little trouble continuing to walk in a straight line.

Now we grab something with a little more mass—say, a small pebble. It bends or distorts the sheet just a bit around its position. If we send the ant into this region, it finds its path slightly altered by the distortion of the sheet. The distortion is not large, but if we follow the ant's path carefully, we notice it deviating slightly from a straight line.

The effect gets more noticeable as we increase the mass of the object that we put on the sheet. Let's say we now use a massive paperweight. Such a heavy object distorts or warps the rubber sheet very effectively, putting a good sag in it. From our point of view, we can see that the sheet near the paperweight is no longer straight.

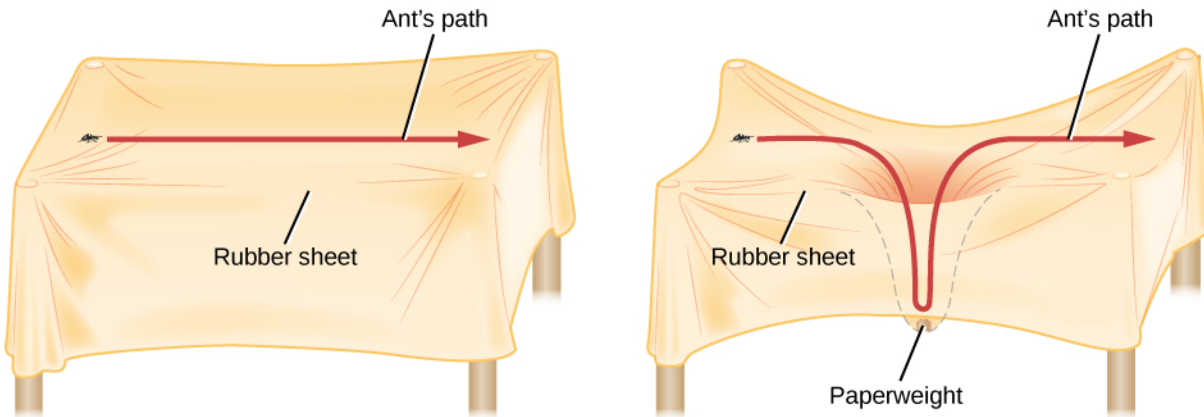


Figure 25.25 On a flat rubber sheet, a trained ant has no trouble walking in a straight line. When a massive object creates a big depression in the sheet, the ant, which must walk where the sheet takes it, finds its path changed (warped) dramatically.

Now let's again send the ant on a journey that takes it close to, but not on top of, the paperweight. Far away from the paperweight, the ant has no trouble doing its walk, which looks straight to us. As it nears the paperweight, however, the ant is forced down into the sag. It must then climb up the other side before it can return to walking on an undistorted part of the sheet. All this while, the ant is following the shortest path it can, but through no fault of its own (after all, ants can't fly, so it has to stay on the sheet) this path is curved by the distortion of the sheet itself.

In the same way, according to Einstein's theory, light always follows the shortest path through spacetime. But the mass associated with large concentrations of matter distorts spacetime, and the shortest, most direct paths are no longer straight lines, but curves.

How large does a mass have to be before we can measure a change in the path followed by light? In 1916, when Einstein first proposed his theory, no distortion had been detected at the surface of Earth (so Earth might have played the role of the grain of sand in our analogy). Something with a mass like our Sun's was necessary to detect the effect Einstein was describing (we will discuss how this effect was measured using the Sun in the next section).

The paperweight in our analogy might be a white dwarf or a neutron star. The distortion of spacetime is greater near the surfaces of these compact, massive objects than near the surface of the Sun. And when, to return to the situation described at the beginning of the chapter, a star core with more than three times the mass of the Sun collapses forever, the distortions of spacetime very close to it can become truly mind-boggling.

25.7 | Black Holes

Learning Objectives

By the end of this section, you will be able to:

- Explain the event horizon surrounding a black hole.
- Discuss why the popular notion of black holes as great sucking monsters that can ingest material at great distances from them is erroneous
- Use the concept of warped spacetime near a black hole to track what happens to any object that might fall into a black hole
- Recognize why the concept of a singularity—with its infinite density and zero volume—presents major challenges to our understanding of matter

Let's consider the collapsing core in a very massive star. We saw that if the core's mass is greater than about 3 M_{Sun}, theory says that nothing can stop the core from collapsing forever. We will examine this situation from two perspectives: first from a pre-Einstein point of view, and then with the aid of general relativity.

Classical Collapse

Let's begin with a thought experiment. We want to know what speeds are required to escape from the gravitational pull of different objects. A rocket must be launched from the surface of Earth at a very high speed if it is to escape the pull of Earth's gravity. In fact, any object—rocket, ball, astronomy book—that is thrown into the air with a velocity less than 11 kilometers per second will soon fall back to Earth's surface. Only those objects launched with a speed greater than this *escape velocity* can get away from Earth.

The escape velocity from the surface of the Sun is higher yet—618 kilometers per second. Now imagine that we begin to compress the Sun, forcing it to shrink in diameter. Recall that the pull of gravity depends on both the mass that is pulling you and your distance from the center of gravity of that mass. If the Sun is compressed, its *mass* will remain the same, but the *distance* between a point on the Sun's surface and the center will get smaller and smaller. Thus, as we compress the star, the pull of gravity for an object on the shrinking surface will get stronger and stronger (**Figure 25.26**).

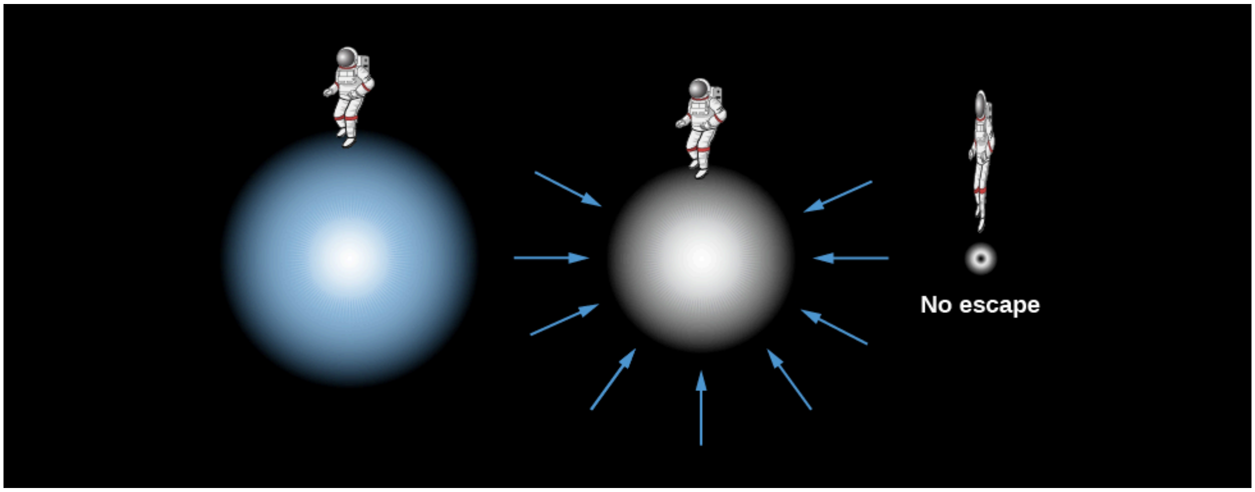


Figure 25.26 At left, an imaginary astronaut floats near the surface of a massive star-core about to collapse. As the same mass falls into a smaller sphere, the gravity at its surface goes up, making it harder for anything to escape from the stellar surface. Eventually the mass collapses into so small a sphere that the escape velocity exceeds the speed of light and nothing can get away. Note that the size of the astronaut has been exaggerated. In the last picture, the astronaut is just outside the sphere we will call the event horizon and is stretched and squeezed by the strong gravity.

When the shrinking Sun reaches the diameter of a neutron star (about 20 kilometers), the velocity required to escape its gravitational pull will be about half the speed of light. Suppose we continue to compress the Sun to a smaller and smaller diameter. (We saw this can't happen to a star like our Sun in the real world because of electron degeneracy, i.e., the mutual repulsion between tightly packed electrons; this is just a quick "thought experiment" to get our bearings).

Ultimately, as the Sun shrinks, the escape velocity near the surface would exceed the speed of light. If the speed you need to get away is faster than the fastest possible speed in the universe, then nothing, not even light, is able to escape. An object with such large escape velocity emits no light, and anything that falls into it can never return.

In modern terminology, we call an object from which light cannot escape a **black hole**, a name popularized by the America scientist John Wheeler starting in the late 1960s (**Figure 25.27**). The idea that such objects might exist is, however, not a new one. Cambridge professor and amateur astronomer John Michell wrote a paper in 1783 about the possibility that stars with escape velocities exceeding that of light might exist. And in 1796, the French mathematician Pierre-Simon, marquis de Laplace, made similar calculations using Newton's theory of gravity; he called the resulting objects "dark bodies."

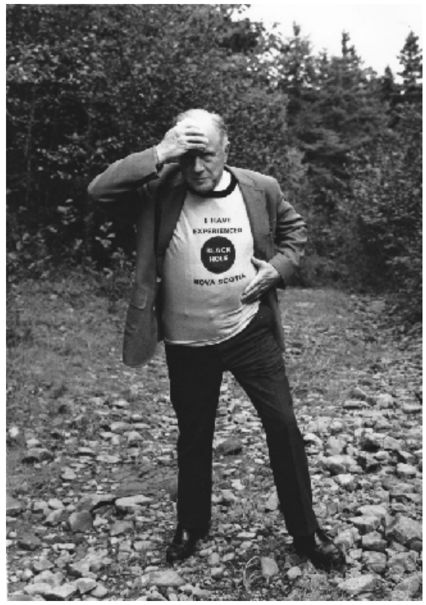


Figure 25.27 This brilliant physicist did much pioneering work in general relativity theory and popularized the term *black hole* starting in the late 1960s. (credit: modification of work by Roy Bishop)

While these early calculations provided strong hints that something strange should be expected if very massive objects collapse under their own gravity, we really need general relativity theory to give an adequate description of what happens in such a situation.

Collapse with Relativity

General relativity tells us that gravity is really a curvature of spacetime. As gravity increases (as in the collapsing Sun of our thought experiment), the curvature gets larger and larger. Eventually, if the Sun could shrink down to a diameter of about 6 kilometers, only light beams sent out perpendicular to the surface would escape. All others would fall back onto the star (**Figure 25.28**). If the Sun could then shrink just a little more, even that one remaining light beam would no longer be able to escape.

(a) (b)

Figure 25.28 Suppose a person could stand on the surface of a normal star with a flashlight. The light leaving the flashlight travels in a straight line no matter where the flashlight is pointed. Now consider what happens if the star collapses so that it is just a little larger than a black hole. All the light paths, except the one straight up, curve back to the surface. When the star shrinks inside the event horizon and becomes a black hole, even a beam directed straight up returns.

Keep in mind that gravity is not pulling on the light. The concentration of matter has curved spacetime, and light (like the trained ant of our earlier example) is "doing its best" to go in a straight line, yet is now confronted with a world in which straight lines that used to go outward have become curved paths that lead back in. The collapsing star is a *black hole* in this view, because the very concept of "out" has no geometrical meaning. The star has become trapped in its own little pocket of spacetime, from which there is no escape.

The star's geometry cuts off communication with the rest of the universe at precisely the moment when, in our earlier picture, the escape velocity becomes equal to the speed of light. The size of the star at this moment defines a surface that we call the **event horizon**. It's a wonderfully descriptive name: just as objects that sink below our horizon cannot be seen on Earth, so anything happening inside the event horizon can no longer interact with the rest of the universe.

Imagine a future spacecraft foolish enough to land on the surface of a massive star just as it begins to collapse in the way we have been describing. Perhaps the captain is asleep at the gravity meter, and before the crew can say "Albert Einstein," they have collapsed with the star inside the event horizon. Frantically, they send an escape pod straight outward. But paths outward twist around to become paths inward, and the pod turns around and falls toward the center of the black hole. They send a radio message to their loved ones, bidding good-bye. But radio waves, like light, must travel through spacetime, and curved spacetime allows nothing to get out. Their final message remains unheard. Events inside the event horizon can never again affect events outside it.

The characteristics of an event horizon were first worked out by astronomer and mathematician Karl Schwarzschild (**Figure 25.29**). A member of the German army in World War I, he died in 1916 of an illness he contracted while doing artillery shell calculations on the Russian front. His paper on the theory of event horizons was among the last things he finished as he was dying; it was the first exact solution to Einstein's equations of general relativity. The radius of the event horizon is called the *Schwarzschild radius* in his memory.

Figure 25.29 This German scientist was the first to demonstrate mathematically that a black hole is possible and to determine the size of a nonrotating black hole's event horizon.

The event horizon is the boundary of the black hole; calculations show that it does not get smaller once the whole star has collapsed inside it. It is the region that separates the things trapped inside it from the rest of the universe. Anything coming from the outside is also trapped once it comes inside the event horizon. The horizon's size turns out to depend only on the mass inside it. If the Sun, with its mass of 1 M_{Sun}, were to become a black hole (fortunately, it can't—this is just a thought experiment), the Schwarzschild radius would be about 3 kilometers; thus, the entire black hole would be about one-third the size of a neutron star of that same mass. Feed the black hole some mass, and the horizon will grow—but not very much. Doubling the mass will make the black hole 6 kilometers in radius, still very tiny on the cosmic scale.

The event horizons of more massive black holes have larger radii. For example, if a globular cluster of 100,000 stars (solar masses) could collapse to a black hole, it would be 300,000 kilometers in radius, a little less than half the radius of the Sun. If the entire Galaxy could collapse to a black hole, it would be only about 10^{12} kilometers in radius—about a tenth of a light year. Smaller masses have correspondingly smaller horizons: for Earth to become a black hole, it would have to be compressed to a radius of only 1 centimeter—less than the size of a grape. A typical asteroid, if crushed to a small enough size to be a black hole, would have the dimensions of an atomic nucleus.

Example 25.2

The Milky Way's Black Hole

The size of the event horizon of a black hole depends on the mass of the black hole. The greater the mass, the larger the radius of the event horizon. General relativity calculations show that the formula for the Schwarzschild radius (R_S) of the event horizon is

$$R_S = \frac{2GM}{c^2}$$

where c is the speed of light, G is the gravitational constant, and M is the mass of the black hole. Note that in this formula, 2, G, and c are all constant; only the mass changes from black hole to black hole. Note the consistency of this result with the **Equation 10.21** from the chapter on **Section 10.6**. That is, this equation gives the radius of a body of mass M from which the escape velocity would equal the speed of light.

As we will see in the chapter on **The Milky Way Galaxy**, astronomers have traced the paths of several stars near the center of our Galaxy and found that they seem to be orbiting an unseen object—dubbed Sgr A* (pronounced "Sagittarius A-star")—with a mass of about 4 million solar masses. What is the size of its Schwarzschild radius?

Solution

We can substitute data for G, M, and c (from **Appendix C**) directly into the equation:

$$\begin{aligned} R_S &= \frac{2GM}{c^2} = \frac{2(6.67 \times 10^{-11}\,\text{N} \cdot \text{m}^2/\text{kg}^2)(4 \times 10^6)(1.99 \times 10^{30}\ \text{kg})}{(3.00 \times 10^8\ \text{m/s})^2} \\ &= 1.18 \times 10^{10}\ \text{m} \end{aligned}$$

This distance is about one-fifth of the radius of Mercury's orbit around the Sun, yet the object contains 4 million solar masses and cannot be seen with our largest telescopes. You can see why astronomers are convinced this object is a black hole.

25.2 What would be the size of a black hole that contained only as much mass as a typical pickup truck (about 3000 kg)? (Note that something with so little mass could never actually form a black hole, but it's interesting to think about the result.)

A Black Hole Myth

Much of the modern folklore about black holes is misleading. One idea you may have heard is that black holes go about sucking things up with their gravity. Actually, it is only very close to a black hole that the strange effects we have been discussing come into play. The gravitational attraction far away from a black hole is the same as that of the star that collapsed to form it.

Remember that the gravity of any star some distance away acts as if all its mass were concentrated at a point in the center, which we call the center of gravity. For real stars, we merely *imagine* that all mass is concentrated there; for black holes, all the mass *really is* concentrated at a point in the center.

So, if you are a star or distant planet orbiting around a star that becomes a black hole, your orbit may not be significantly affected by the collapse of the star (although it may be affected by any mass loss that precedes the collapse). If, on the other hand, you venture close to the event horizon, it would be very hard for you to resist the "pull" of the warped spacetime near the black hole. You have to get really close to the black hole to experience any significant effect.

If another star or a spaceship were to pass one or two solar radii from a black hole, Newton's laws would be adequate to describe what would happen to it. Only very near the event horizon of a black hole is the gravitation so strong that Newton's laws break down. The black hole remnant of a massive star coming into our neighborhood would be far, far safer to us than its earlier incarnation as a brilliant, hot star.

Gravity and Time Machines

Time machines are one of the favorite devices of science fiction. Such a device would allow you to move through time at a different pace or in a different direction from everyone else. General relativity suggests that it is possible, in theory, to construct a time machine using gravity that could take you into the future.

Let's imagine a place where gravity is terribly strong, such as near a black hole. General relativity predicts that the stronger the gravity, the slower the pace of time (as seen by a distant observer). So, imagine a future astronaut, with a fast and strongly built spaceship, who volunteers to go on a mission to such a high-gravity environment. The astronaut leaves in the year 2222, just after graduating from college at age 22. She takes, let's say, exactly 10 years to get to the black hole. Once there, she orbits some distance from it, taking care not to get pulled in.

She is now in a high-gravity realm where time passes much more slowly than it does on Earth. This isn't just an effect on the mechanism of her clocks—*time itself* is running slowly. That means that every way she has of measuring time will give the same slowed-down reading when compared to time passing on Earth. Her heart will beat more slowly, her hair will grow more slowly, her antique wristwatch will tick more slowly, and so on. She is not aware of this slowing down because all her readings of time, whether made by her own bodily functions or with mechanical equipment, are measuring the same—slower—time. Meanwhile, back on Earth, time passes as it always does.

Our astronaut now emerges from the region of the black hole, her mission of exploration finished, and returns to Earth. Before leaving, she carefully notes that (according to her timepieces) she spent about 2 weeks around the black hole. She then takes exactly 10 years to return to Earth. Her calculations tell her that since she was 22 when she left the Earth, she will be 42 plus 2 weeks when she returns. So, the year on Earth, she figures, should be 2242, and her classmates should now be approaching their midlife crises.

But our astronaut should have paid more attention in her astronomy class! Because time slowed down near the black hole, much less time passed for her than for the people on Earth. While her clocks measured 2 weeks spent near the black hole, more than 2000 weeks (depending on how close she got) could well have passed on Earth. That's equal to 40 years, meaning her classmates will be senior citizens in their 80s when she (a mere 42-year-old) returns. On Earth it will be not 2242, but 2282—and she will say that she has arrived *in the future*.

Is this scenario real? Well, it has a few practical challenges: we don't think any black holes are close enough for us to reach in 10 years, and we don't think any spaceship or human can survive near a black hole. But the key point about the slowing down of time is a natural consequence of Einstein's general theory of relativity, and we saw that its predictions have been confirmed by experiment after experiment.

Such developments in the understanding of science also become inspiration for science fiction writers. Recently, the film *Interstellar* featured the protagonist traveling close to a massive black hole; the resulting delay in his aging relative to his earthbound family is a key part of the plot.

Science fiction novels, such as *Gateway* by Frederik Pohl and *A World out of Time* by Larry Niven, also make use of the slowing down of time near black holes as major turning points in the story. For a list of science fiction stories based on good astronomy, you can go to www.astrosociety.org/scifi.

A Trip into a Black Hole

The fact that scientists cannot see inside black holes has not kept them from trying to calculate what they are like. One of the first things these calculations showed was that the formation of a black hole obliterates nearly all information about the star that collapsed to form it. Physicists like to say "black holes have no hair," meaning that nothing sticks out of a black hole to give us clues about what kind of star produced it or what material has fallen inside. The only information a black hole can reveal about itself is its mass, its spin (rotation), and whether it has any electrical charge.

What happens to the collapsing star-core that made the black hole? Our best calculations predict that the material will continue to collapse under its own weight, forming an infinitely *squozen* point—a place of zero volume and infinite density—to which we give the name **singularity**. At the singularity, spacetime ceases to exist. The laws of physics as we know them break down. We do not yet have the physical understanding or the mathematical tools to describe the singularity itself, or even if singularities actually occur. From the outside, however, the entire structure of a basic black hole (one that is not rotating) can be described as a singularity surrounded by an event horizon. Compared to humans, black holes are really very simple objects.

Scientists have also calculated what would happen if an astronaut were to fall into a black hole. Let's take up an observing position a long, safe distance away from the event horizon and watch this astronaut fall toward it. At first he falls away from us, moving ever faster, just as though he were approaching any massive star. However, as he nears the event horizon of the black hole, things change. The strong gravitational field around the black hole will make his clocks run more slowly, when seen from our outside perspective.

If, as he approaches the event horizon, he sends out a signal once per second according to his clock, we will see the spacing between his signals grow longer and longer until it becomes infinitely long when he reaches the event horizon. (Recalling our discussion of gravitational redshift, we could say that if the infalling astronaut uses a blue light to send his signals every second, we will see the light get redder and redder until its wavelength is nearly infinite.) As the spacing between clock ticks approaches infinity, it will appear to us that the astronaut is slowly coming to a stop, frozen in time at the event horizon.

In the same way, all matter falling into a black hole will also appear to an outside observer to stop at the event horizon, frozen in place and taking an infinite time to fall through it. But don't think that matter falling into a black hole will therefore be easily visible at the event horizon. The tremendous redshift will make it very difficult to observe any radiation from the "frozen" victims of the black hole.

This, however, is only how we, located far away from the black hole, see things. To the astronaut, his time goes at its normal rate and he falls right on through the event horizon into the black hole. (Remember, this horizon is not a physical barrier, but only a region in space where the curvature of spacetime makes escape impossible.)

You may have trouble with the idea that you (watching from far away) and the astronaut (falling in) have such different ideas about what has happened. This is the reason Einstein's ideas about space and time are called theories of *relativity*. What each observer measures about the world depends on (is relative to) his or her frame of reference. The observer in strong gravity measures time and space differently from the one sitting in weaker gravity. When Einstein proposed these ideas, many scientists also had difficulty with the idea that two such different views of the same event could be correct, each in its own "world," and they tried to find a mistake in the calculations. There were no mistakes: we and the astronaut really would see him fall into a black hole very differently.

For the astronaut, there is no turning back. Once inside the event horizon, the astronaut, along with any signals from his radio transmitter, will remain hidden forever from the universe outside. He will, however, not have a long time (from his perspective) to feel sorry for himself as he approaches the black hole. Suppose he is falling feet first. The force of gravity that the singularity exerts on his feet is greater than on his head, so he will be stretched slightly. Because the singularity is a point, the left side of his body will be pulled slightly toward the right, and the right slightly toward the left, bringing each side closer to the singularity. The astronaut will therefore be slightly squeezed in one direction and stretched in the other. Some scientists like to call this process of stretching and narrowing *spaghettification*. The point at which the astronaut becomes so stretched that he perishes depends on the size of the black hole. For black holes with masses billions of times the mass of the Sun, such as those found at the centers of galaxies, the spaghettification becomes significant only after the astronaut passes through the event horizon. For black holes with masses of a few solar masses, the astronaut will be stretched and ripped apart even before he reaches the event horizon.

Earth exerts similar *tidal forces* on an astronaut performing a spacewalk. In the case of Earth, the tidal forces are so small that they pose no threat to the health and safety of the astronaut. Not so in the case of a black hole. Sooner or later, as the astronaut approaches the black hole, the tidal forces will become so great that the astronaut will be ripped apart, eventually reduced to a collection of individual atoms that will continue their inexorable fall into the singularity.

From the previous discussion, you will probably agree that jumping into a black hole is definitely a once-in-a-lifetime experience! You can see an **engaging explanation (https://openstax.org/l/30ndegtystidfor)** of death by black hole by Neil deGrasse Tyson, where he explains the effect of tidal forces on the human body until it dies by spaghettification.

An overview of black holes is given in this **Discovery Channel video (https://openstax.org/l/30dischatidfor)** excerpt.

25.8 | Evidence for Black Holes

Learning Objectives

By the end of this section, you will be able to:

- Describe what to look for when seeking and confirming the presence of a stellar black hole
- Explain how a black hole is inherently black yet can be associated with luminous matter
- Differentiate between stellar black holes and the black holes in the centers of galaxies

Theory tells us what black holes are like. But do they actually exist? And how do we go about looking for something that is many light years away, only about a few dozen kilometers across (if a stellar black hole), and completely black? It turns out that the trick is not to look for the black hole itself but instead to look for what it does to a nearby companion star.

As we saw, when very massive stars collapse, they leave behind their gravitational influence. What if a member of a double-star system becomes a black hole, and its companion manages to survive the death of the massive star? While the black hole disappears from our view, we may be able to deduce its presence from the things it does to its companion.

Requirements for a Black Hole

So, here is a prescription for finding a black hole: start by looking for a star whose motion (determined from the Doppler shift of its spectral lines) shows it to be a member of a binary star system. If both stars are visible, neither can be a black hole, so focus your attention on just those systems where only one star of the pair is visible, even with our most sensitive telescopes.

Being invisible is not enough, however, because a relatively faint star might be hard to see next to the glare of a brilliant companion or if it is shrouded by dust. And even if the star really is invisible, it could be a neutron star. Therefore, we must also have evidence that the unseen star has a mass too high to be a neutron star and that it is a collapsed object—an extremely small stellar remnant.

We can use Kepler’s Third law (see **Kepler's Laws of Planetary Motion**) and our knowledge of the visible star to measure the mass of the invisible member of the pair. If the mass is greater than about 3 M_{Sun}, then we are likely seeing (or, more precisely, not seeing) a black hole—as long as we can make sure the object really is a collapsed star.

If matter falls toward a compact object of high gravity, the material is accelerated to high speed. Near the event horizon of a black hole, matter is moving at velocities that approach the speed of light. As the atoms whirl chaotically toward the event horizon, they rub against each other; internal friction can heat them to temperatures of 100 million K or more. Such hot matter emits radiation in the form of flickering X-rays. The last part of our prescription, then, is to look for a source of X-rays associated with the binary system. Since X-rays do not penetrate Earth’s atmosphere, such sources must be found using X-ray telescopes in space.

In our example, the infalling gas that produces the X-ray emission comes from the black hole’s companion star. As we saw in **The Death of Stars**, stars in close binary systems can exchange mass, especially as one of the members expands into a red giant. Suppose that one star in a double-star system has evolved to a black hole and that the second star begins to expand. If the two stars are not too far apart, the outer layers of the expanding star may reach the point where the black hole exerts more gravitational force on them than do the inner layers of the red giant to which the atmosphere belongs. The outer atmosphere then passes through the point of no return between the stars and falls toward the black hole.

The mutual revolution of the giant star and the black hole causes the material falling toward the black hole to spiral around it rather than flow directly into it. The infalling gas whirls around the black hole in a pancake of matter called an **accretion disk**. It is within the inner part of this disk that matter is revolving about the black hole so fast that internal friction heats it up to X-ray–emitting temperatures.

Another way to form an accretion disk in a binary star system is to have a powerful stellar wind come from the black hole’s companion. Such winds are a characteristic of several stages in a star’s life. Some of the ejected gas in the wind will then flow close enough to the black hole to be captured by it into the disk (**Figure 25.30**).

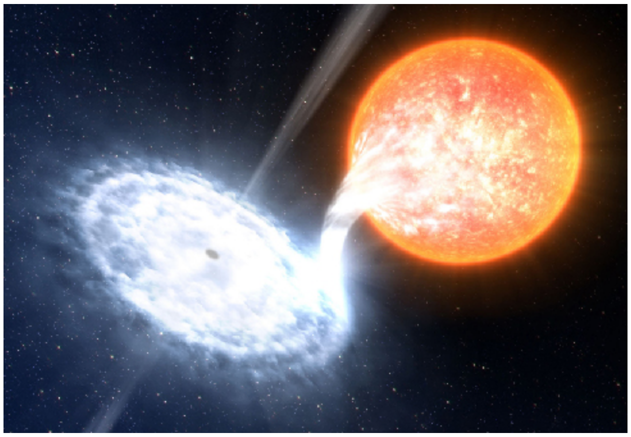


Figure 25.30 This artist’s rendition shows a black hole and star (red). As matter streams from the star, it forms a disk around the black hole. Some of the swirling material close to the black hole is pushed outward perpendicular to the disk in two narrow jets. (credit: modification of work by ESO/L. Calçada)

We should point out that, as often happens, the measurements we have been discussing are not quite as simple as they are described in introductory textbooks. In real life, Kepler's law allows us to calculate only the combined mass of the two stars in the binary system. We must learn more about the visible star of the pair and its history to ascertain the distance to the binary pair, the true size of the visible star's orbit, and how the orbit of the two stars is tilted toward Earth, something we can rarely measure. And neutron stars can also have accretion disks that produce X-rays, so astronomers must study the properties of these X-rays carefully when trying to determine what kind of object is at the center of the disk. Nevertheless, a number of systems that clearly contain black holes have now been found.

The Discovery of Stellar-Mass Black Holes

Because X-rays are such important tracers of black holes that are having some of their stellar companions for lunch, the search for black holes had to await the launch of sophisticated X-ray telescopes into space. These instruments must have the resolution to locate the X-ray sources accurately and thereby enable us to match them to the positions of binary star systems.

The first black hole binary system to be discovered is called Cygnus X-1. The visible star in this binary system is spectral type O. Measurements of the Doppler shifts of the O star's spectral lines show that it has an unseen companion. The X-rays flickering from it strongly indicate that the companion is a small collapsed object. The mass of the invisible collapsed companion is about 15 times that of the Sun. The companion is therefore too massive to be either a white dwarf or a neutron star.

A number of other binary systems also meet all the conditions for containing a black hole. **Table 25.4** lists the characteristics of some of the best examples.

Some Black Hole Candidates in Binary Star Systems

Name/Catalog Designation[4]	Companion Star Spectral Type	Orbital Period (days)	Black Hole Mass Estimates (M_{Sun})
LMC X-1	O giant	3.9	10.9
Cygnus X-1	O supergiant	5.6	15
XTE J1819.3-254 (V4641 Sgr)	B giant	2.8	6–7
LMC X-3	B main sequence	1.7	7
4U1543-475 (IL Lup)	A main sequence	1.1	9
GRO J1655-40 (V1033 Sco)	F subgiant	2.6	7
GRS 1915+105	K giant	33.5	14
GS202+1338 (V404 Cyg)	K giant	6.5	12
XTE J1550-564	K giant	1.5	11
A0620-00 (V616 Mon)	K main sequence	0.33	9–13
H1705-250 (Nova Oph 1977)	K main sequence	0.52	5–7
GRS1124-683 (Nova Mus 1991)	K main sequence	0.43	7
GS2000+25 (QZ Vul)	K main sequence	0.35	5–10
GRS1009-45 (Nova Vel 1993)	K dwarf	0.29	8–9

Table 25.4

4. As you can tell, there is no standard way of naming these candidates. The chain of numbers is the location of the source in right ascension and declination (the longitude and latitude system of the sky); some of the letters preceding the numbers refer to objects (e.g., LMC) and constellations (e.g., Cygnus), while other letters refer to the satellite that discovered the candidate—A for Ariel, G for Ginga, and so on. The notations in parentheses are those used by astronomers who study binary star system or novae.

Some Black Hole Candidates in Binary Star Systems

Name/Catalog Designation	Companion Star Spectral Type	Orbital Period (days)	Black Hole Mass Estimates (M_{Sun})
XTE J1118+480	K dwarf	0.17	7
XTE J1859+226	K dwarf	0.38	5.4
GRO J0422+32	M dwarf	0.21	4

Table 25.4

Feeding a Black Hole

After an isolated star, or even one in a binary star system, becomes a black hole, it probably won't be able to grow much larger. Out in the suburban regions of the Milky Way Galaxy where we live (see **The Milky Way Galaxy**), stars and star systems are much too far apart for other stars to provide "food" to a hungry black hole. After all, material must approach very close to the event horizon before the gravity is any different from that of the star before it became the black hole.

But, as will see, the central regions of galaxies are quite different from their outer parts. Here, stars and raw material can be quite crowded together, and they can interact much more frequently with each other. Therefore, black holes in the centers of galaxies may have a much better opportunity to find mass close enough to their event horizons to pull in. Black holes are not particular about what they "eat": they are happy to consume other stars, asteroids, gas, dust, and even other black holes. (If two black holes merge, you just get a black hole with more mass and a larger event horizon.)

As a result, black holes in crowded regions can grow, eventually swallowing thousands or even millions of times the mass of the Sun. Ground-based observations have provided compelling evidence that there is a black hole in the center of our own Galaxy with a mass of about 4 million times the mass of the Sun (we'll discuss this further in the chapter on **The Milky Way Galaxy**). Observations with the Hubble Space Telescope have shown dramatic evidence for the existence of black holes in the centers of many other galaxies. These black holes can contain more than a billion solar masses. The feeding frenzy of such supermassive black holes may be responsible for some of the most energetic phenomena in the universe (see **The Evolution and Distribution of Galaxies**). And evidence from more recent X-ray observations is also starting to indicate the existence of "middle-weight" black holes, whose masses are dozens to thousands of times the mass of the Sun. The crowded inner regions of the globular clusters we described in **Star Clusters** may be just the right breeding grounds for such intermediate-mass black holes.

Over the past decades, many observations, especially with the Hubble Space Telescope and with X-ray satellites, have been made that can be explained only if black holes really do exist. Furthermore, the observational tests of Einstein's general theory of relativity have convinced even the most skeptical scientists that his picture of warped or curved spacetime is indeed our best description of the effects of gravity near these black holes.

25.9 | Gravitational Wave Astronomy

Learning Objectives

By the end of this section, you will be able to:

- Describe what a gravitational wave is, what can produce it, and how fast it propagates
- Understand the basic mechanisms used to detect gravitational waves

Another part of Einstein's ideas about gravity can be tested as a way of checking the theory that underlies black holes. According to general relativity, the geometry of spacetime depends on where matter is located. Any rearrangement of matter—say, from a sphere to a sausage shape—creates a disturbance in spacetime. This disturbance is called a **gravitational wave**, and relativity predicts that it should spread outward at the speed of light. The big problem with trying to study such waves is that they are tremendously weaker than electromagnetic waves and correspondingly difficult to detect.

Proof from a Pulsar

We've had indirect evidence for some time that gravitational waves exist. In 1974, astronomers Joseph Taylor and Russell Hulse discovered a pulsar (with the designation PSR1913+16) orbiting another neutron star. Pulled by the powerful gravity of its companion, the pulsar is moving at about one-tenth the speed of light in its orbit.

According to general relativity, this system of stellar corpses should be radiating energy in the form of gravitational waves at a high enough rate to cause the pulsar and its companion to spiral closer together. If this is correct, then the orbital period should decrease (according to Kepler's third law) by one ten-millionth of a second per orbit. Continuing observations showed that the period is decreasing by precisely this amount. Such a loss of energy in the system can be due only to the radiation of gravitational waves, thus confirming their existence. Taylor and Hulse shared the 1993 Nobel Prize in physics for this work.

Direct Observations

Although such an indirect proof convinced physicists that gravitational waves exist, it is even more satisfying to detect the waves directly. What we need are phenomena that are powerful enough to produce gravitational waves with amplitudes large enough that we can measure them. Theoretical calculations suggest some of the most likely events that would give a burst of gravitational waves strong enough that our equipment on Earth could measure it:

- the coalescence of two neutron stars in a binary system that spiral together until they merge
- the swallowing of a neutron star by a black hole
- the coalescence (merger) of two black holes
- the implosion of a really massive star to form a neutron star or a black hole
- the first "shudder" when space and time came into existence and the universe began

For the last four decades, scientists have been developing an audacious experiment to try to detect gravitational waves from a source on this list. The US experiment, which was built with collaborators from the UK, Germany, Australia and other countries, is named LIGO (Laser Interferometer Gravitational-Wave Observatory). LIGO currently has two observing stations, one in Louisiana and the other in the state of Washington. The effects of gravitational waves are so small that confirmation of their detection will require simultaneous measurements by two widely separated facilities. Local events that might cause small motions within the observing stations and mimic gravitational waves—such as small earthquakes, ocean tides, and even traffic—should affect the two sites differently.

Each of the LIGO stations consists of two 4-kilometer-long, 1.2-meter-diameter vacuum pipes arranged in an L-shape. A test mass with a mirror on it is suspended by wire at each of the four ends of the pipes. Ultra-stable laser light is reflected from the mirrors and travels back and forth along the vacuum pipes (Figure 25.31). If gravitational waves pass through the LIGO instrument, then, according to Einstein's theory, the waves will affect local spacetime—they will alternately stretch and shrink the distance the laser light must travel between the mirrors ever so slightly. When one arm of the instrument gets longer, the other will get shorter, and vice versa.

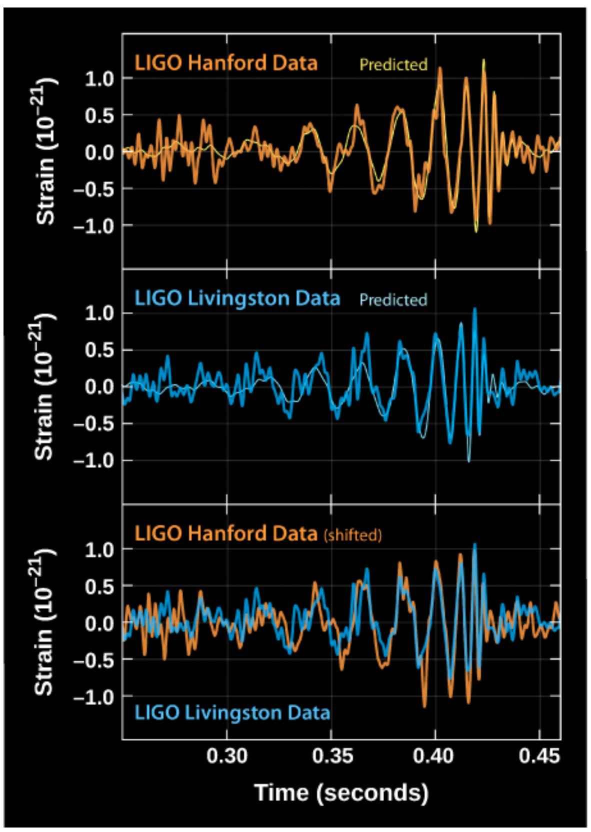
Figure 25.31 An aerial view of the LIGO facility at Livingston, Louisiana. Extending to the upper left and far right of the image are the 4-kilometer-long detectors. (credit: modification of work by Caltech/MIT/LIGO Laboratory)

The challenge of this experiment lies in that phrase "ever so slightly." In fact, to detect a gravitational wave, the change in the distance to the mirror must be measured with an accuracy of *one ten-thousandth the diameter of a proton*. In 1972, Rainer Weiss of MIT wrote a paper suggesting how this seemingly impossible task might be accomplished.

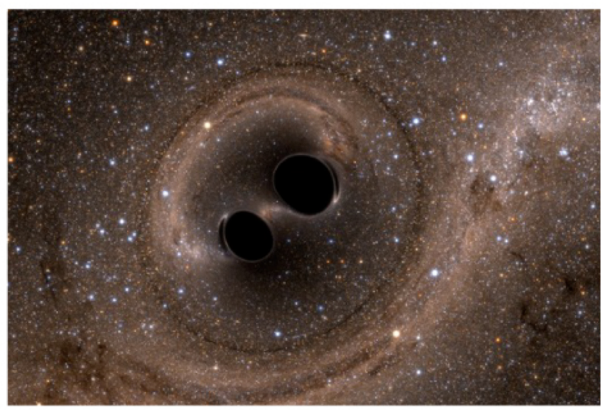
A great deal of new technology had to be developed, and work on the laboratory, with funding from the National Science Foundation, began in 1979. A full-scale prototype to demonstrate the technology was built and operated from 2002 to 2010, but the prototype was not expected to have the sensitivity required to actually detect gravitational waves from an astronomical source. Advanced LIGO, built to be more precise with the improved technology developed in the prototype, went into operation in 2015—and almost immediately detected gravitational waves.

What LIGO found was gravitational waves produced in the final fraction of a second of the merger of two black holes (**Figure 25.32**). The black holes had masses of 20 and 36 times the mass of the Sun, and the merger took place 1.3 billion years ago—the gravitational waves occurred so far away that it has taken that long for them, traveling at the speed of light, to reach us.

In the cataclysm of the merger, about three times the mass of the Sun was converted to energy (recall $E = mc^2$). During the tiny fraction of a second for the merger to take place, this event produced power about 10 times the power produced by all the stars in the entire visible universe—but the power was all in the form of gravitational waves and hence was invisible to our instruments, except to LIGO. The event was recorded in Louisiana about 7 milliseconds before the detection in Washington—just the right distance given the speed at which gravitational waves travel—and indicates that the source was located somewhere in the southern hemisphere sky. Unfortunately, the merger of two black holes is not expected to produce any light, so this is the only observation we have of the event.



(a)



(b)

Figure 25.32 (a) The top panel shows the signal measured at Hanford, Washington; the middle panel shows the signal measured at Livingston, Louisiana. The smoother thin curve in each panel shows the predicted signal, based on Einstein's general theory of relativity, produced by the merger of two black holes. The bottom panel shows a superposition of the waves detected at the two LIGO observatories. Note the remarkable agreement of the two independent observations and of the observations with theory. (b) The painting shows an artist's impression of two massive black holes spiraling inward toward an eventual merger. (credit a, b: modification of work by SXS)

This detection by LIGO (and another one of a different black hole merger a few months later) opened a whole new window on the universe. One of the experimenters compared the beginning of gravitational wave astronomy to the era when silent films were replaced by movies with sound (comparing the vibration of spacetime during the passing of a gravitational wave to the vibrations that sound makes).

By the end of 2018, LIGO had detected eight more mergers of black holes. Six of these, like the initial discovery, involved mergers of black holes with a range of masses that have been observed only by gravitational waves. In one merger, black holes with masses of 31 and 25 times the mass of the Sun merged to form a spinning black hole with a mass of about 53 times the mass the Sun. Some of these events were detected not only by the two LIGO detectors, but also by a newly operational European gravitational wave observatory, Virgo. Another event was caused by the merger of 40- and 29-solar-mass black holes, and resulted in a 66-mass black hole. Astronomers are not yet sure just how black holes in this mass range form.

Two other mergers detected by LIGO involved black holes with stellar masses comparable to those of black holes in X-ray binary systems. In one case, the merging black holes had masses of 14 and 8 times the mass of the Sun. The other event, again detected by both LIGO and Virgo, was produced by a merger of black holes with masses of 7 and 12 times the mass of the Sun. None of the mergers of black holes was detected in any other way besides gravitational waves. It is quite likely that the merger of black holes does not produce any electromagnetic radiation.

In late 2017, data from all three gravitational wave observatories was used to locate the position in the sky of a fifth event, which was produced by the merger of objects with masses of 1.1 to 1.6 times the mass of the Sun. This is the mass range for neutron stars (see **The Milky Way Galaxy**), so in this case, what was observed was the spiraling together of two neutron stars. Data obtained from all three observatories enabled scientists to narrow down the area in the sky where the event occurred. The Fermi satellite offered a fourth set of observational data, detecting a flash of gamma rays at the same time, which confirms the long-standing hypothesis that mergers of neutron stars are progenitors of short gamma-ray bursts. The *Swift* satellite also detected a flash of ultraviolet light at the same time, and in the same part of the sky. This was the first time that a gravitational wave event had been detected with any kind of electromagnetic wave.

The combined observations from LIGO, Virgo, Fermi, and *Swift* showed that this source was located in NGC 4993, a galaxy at a distance of about 130 million light-years in the direction of the constellation Hydra. With a well-defined position, ground-based observatories could point their telescopes directly at the source and obtain its spectrum. These observations showed that the merger ejected material with a mass of about 6 percent of the mass of the Sun, and a speed of one-tenth the speed of light. This material is rich in heavy elements. First estimates suggest that the merger produced about 200 Earth masses of gold, and around 500 Earth masses of platinum. This makes clear that neutron star mergers are a significant source of heavy elements. As additional detections of such events improve theoretical estimates of the frequency at which neutron star mergers occur, it may well turn out that the vast majority of heavy elements have been created in such cataclysms.

Observing the merger of black holes via gravitational waves also means that we can now test Einstein’s general theory of relativity where its effects are very strong—close to black holes—and not weak, as they are near Earth. One remarkable result from these detections is that the signals measured so closely match the theoretical predictions made using Einstein’s theory. Once again, Einstein’s revolutionary idea is found to be the correct description of nature.

Because of the scientific significance of the observations of gravitational waves, three of the LIGO project leaders—Rainer Weiss of MIT, and Kip Thorne and Barry Barish of Caltech—were awarded the Nobel Prize in 2017.

Several facilities similar to LIGO and Virgo are under construction in other countries to contribute to gravitational wave astronomy and help us pinpoint more precisely pinpoint the location of signals we detect in the sky. The European Space Agency (ESA) is exploring the possibility of building an even larger detector for gravitational waves in space. The goal is to launch a facility called eLISA sometime in the mid 2030s. The design calls for three detector arms, each a million kilometers in length, for the laser light to travel in space. This facility could detect the merger of distant supermassive black holes, which might have occurred when the first generation of stars formed only a few hundred million years after the Big Bang.

In December 2015, ESA launched LISA Pathfinder and successfully tested the technology required to hold two gold-platinum cubes in a state of weightless, perfect rest, relative to one another. While LISA Pathfinder cannot detect gravitational waves, such stability is required if eLISA is to be able to detect the small changes in path length produced by passing gravitational waves.

We should end by acknowledging that the ideas discussed in this chapter may seem strange and overwhelming, especially the first time you read them. The consequences of the general theory of relatively take some getting used to. But they make the universe more bizarre—and interesting—than you probably thought before you took this course.

For Further Exploration

Websites

Death of Stars

Crab Nebula: **http://chandra.harvard.edu/xray_sources/crab/crab.html (http://chandra.harvard.edu/xray_sources/crab/crab.html)** . A short, colorfully written introduction to the history and science involving the best-known supernova remant.

Introduction to Neutron Stars: **https://www.astro.umd.edu/~miller/nstar.html (https://www.astro.umd.edu/~miller/nstar.html)** . Coleman Miller of the University of Maryland maintains this site, which goes from easy to hard as you get into it, but it has lots of good information about corpses of massive stars.

Introduction to Pulsars (by Maryam Hobbs at the Australia National Telescope Facility): **http://www.atnf.csiro.au/outreach/education/everyone/pulsars/index.html (http://www.atnf.csiro.au/outreach/education/everyone/pulsars/index.html)** .

Magnetars, Soft Gamma Repeaters, and Very Strong Magnetic Fields: **http://solomon.as.utexas.edu/magnetar.html (http://solomon.as.utexas.edu/magnetar.html)** . Robert Duncan, one of the originators of the idea of magnetars, assembled this site some years ago.

Black Holes

Black Hole Encyclopedia: **http://blackholes.stardate.org (http://blackholes.stardate.org)** . From StarDate at the University of Texas McDonald Observatory.

Black Holes: **http://science.nasa.gov/astrophysics/focus-areas/black-holes (http://science.nasa.gov/astrophysics/focus-areas/black-holes)** . NASA overview of black holes, along with links to the most recent news and discoveries.

Black Holes FAQ: **http://cfpa.berkeley.edu/Education/BHfaq.html (http://cfpa.berkeley.edu/Education/BHfaq.html)** . Frequently asked questions about black holes, answered by Ted Bunn of UC–Berkeley's Center for Particle Astrophysics.

Black Holes: Gravity's Relentless Pull: **http://hubblesite.org/explore_astronomy/black_holes/home.html (http://hubblesite.org/explore_astronomy/black_holes/home.html)** . The Hubble Space Telescope's Journey to a Black Hole and Black Hole Encyclopedia (a good introduction for beginners).

Introduction to Black Holes: **http://www.damtp.cam.ac.uk/research/gr/public/bh_intro.html (http://www.damtp.cam.ac.uk/research/gr/public/bh_intro.html)** . The Cambridge University Relativity Group's pages on black holes and related calculations.

March 1918: Testing Einstein: **http://www.nature.com/nature/podcast/index-pastcast-2014-03-20.html (http://www.nature.com/nature/podcast/index-pastcast-2014-03-20.html)** . Nature Podcast about the 1919 eclipse expedition that proved Einstein's General Theory of Relativity.

Movies from the Edge of Spacetime: **http://archive.ncsa.illinois.edu/Cyberia/NumRel/MoviesEdge.html (http://archive.ncsa.illinois.edu/Cyberia/NumRel/MoviesEdge.html)** . Physicists simulate the behavior of various black holes.

Virtual Trips into Black Holes and Neutron Stars: **http://antwrp.gsfc.nasa.gov/htmltest/rjn_bht.html (http://antwrp.gsfc.nasa.gov/htmltest/rjn_bht.html)** . By Robert Nemiroff at Michigan Technological University.

Gravitational Waves

Advanced LIGO: **https://www.advancedligo.mit.edu (https://www.advancedligo.mit.edu)** . The full story on this gravitational wave observatory.

eLISA: **https://www.elisascience.org (https://www.elisascience.org)** .

Gravitational Waves Detected, Confirming Einstein's Theory: **http://www.nytimes.com/2016/02/12/science/ligo-gravitational-waves-black-holes-einstein.html (http://www.nytimes.com/2016/02/12/science/ligo-gravitational-waves-black-holes-einstein.html)** . *New York Times* article and videos on the discovery of gravitational waves.

Gravitational Waves Discovered from Colliding Black Holes: **http://www.scientificamerican.com/article/gravitational-waves-discovered-from-colliding-black-holes1 (http://www.scientificamerican.com/article/gravitational-waves-discovered-from-colliding-black-holes1)** . *Scientific American* coverage of the discovery of gravitational waves (note the additional materials available in the menu at the right).

LIGO Caltech: **https://www.ligo.caltech.edu (https://www.ligo.caltech.edu)** .

Videos

Death of Stars

BBC interview with Antony Hewish: **http://www.bbc.co.uk/archive/scientists/10608.shtml (http://www.bbc.co.uk/archive/scientists/10608.shtml)** . (40:54).

Black Widow Pulsars: The Vengeful Corpses of Stars: **https://www.youtube.com/watch?v=Fn-3G_N0hy4 (https://www.youtube.com/watch?v=Fn-3G_N0hy4)** . A public talk in the Silicon Valley Astronomy Lecture Series by Dr. Roger Romani (Stanford University) (1:01:47).

Hubblecast 64: It all ends with a bang!: **http://www.spacetelescope.org/videos/hubblecast64a/ (http://www.spacetelescope.org/videos/hubblecast64a/)** . HubbleCast Program introducing Supernovae with Dr. Joe Liske (9:48).

Space Movie Reveals Shocking Secrets of the Crab Pulsar: **http://hubblesite.org/newscenter/archive/releases/2002/24/video/c/ (http://hubblesite.org/newscenter/archive/releases/2002/24/video/c/)** . A sequence of Hubble and Chandra Space Telescope images of the central regions of the Crab Nebula have been assembled into a very brief movie accompanied by animation showing how the pulsar affects its environment; it comes with some useful background material (40:06).

Black Holes

Black Holes: The End of Time or a New Beginning?: **https://www.youtube.com/watch?v=mgtJRsdKe6Q (https://www.youtube.com/watch?v=mgtJRsdKe6Q)** . 2012 Silicon Valley Astronomy Lecture by Roger Blandford (1:29:52).

Death by Black Hole: **http://www.openculture.com/2009/02/death_by_black_hole_and_its_kind_of_funny.htm (http://www.openculture.com/2009/02/death_by_black_hole_and_its_kind_of_funny.htm)** . Neil deGrasse Tyson explains spaghettification with only his hands (5:34).

Hearts of Darkness: Black Holes in Space: **https://www.youtube.com/watch?v=4tiAOldypLk (https://www.youtube.com/watch?v=4tiAOldypLk)** . 2010 Silicon Valley Astronomy Lecture by Alex Filippenko (1:56:11).

Gravitational Waves

Journey of a Gravitational Wave: **https://www.youtube.com/watch?v=FlDtXIBrAYE (https://www.youtube.com/watch?v=FlDtXIBrAYE)** . Introduction from LIGO Caltech (2:55).

LIGO's First Detection of Gravitational Waves: **https://www.youtube.com/watch?v=gw-i_VKd6Wo (https://www.youtube.com/watch?v=gw-i_VKd6Wo)** . Explanation and animations from PBS Digital Studio (9:31).

Two Black Holes Merge into One: **https://www.youtube.com/watch?v=I_88S8DWbcU (https://www.youtube.com/watch?v=I_88S8DWbcU)** . Simulation from LIGO Caltech (0:35).

What the Discovery of Gravitational Waves Means: **https://www.youtube.com/watch?v=jMVAgCPYYHY (https://www.youtube.com/watch?v=jMVAgCPYYHY)** . TED Talk by Allan Adams (10:58).

CHAPTER 25 REVIEW

KEY TERMS

accretion disk the disk of gas and dust found orbiting newborn stars, as well as compact stellar remnants such as white dwarfs, neutron stars, and black holes when they are in binary systems and are sufficiently close to their binary companions to draw off material

black hole a region in spacetime where gravity is so strong that nothing—not even light—can escape

Chandrasekhar limit the upper limit to the mass of a white dwarf (equals 1.4 times the mass of the Sun)

degenerate gas a gas that resists further compression because no two electrons can be in the same place at the same time doing the same thing (Pauli exclusion principle)

equivalence principle concept that a gravitational force and a suitable acceleration are indistinguishable within a sufficiently local environment

event horizon a boundary in spacetime such that events inside the boundary can have no effect on the world outside it—that is, the boundary of the region around a black hole where the curvature of spacetime no longer provides any way out

general theory of relativity Einstein's theory relating gravity and the structure (geometry) of space and time

gravitational wave a disturbance in the curvature of spacetime caused by changes in how matter is distributed; gravitational waves propagate at (or near) the speed of light.

millisecond pulsar a pulsar that rotates so quickly that it can give off hundreds of pulses per second (and its period is therefore measured in milliseconds)

neutron star a compact object of extremely high density composed almost entirely of neutrons

nova the cataclysmic explosion produced in a binary system, temporarily increasing its luminosity by hundreds to thousands of times

pulsar a variable radio source of small physical size that emits very rapid radio pulses in very regular periods that range from fractions of a second to several seconds; now understood to be a rotating, magnetic neutron star that is energetic enough to produce a detectable beam of radiation and particles

singularity the point of zero volume and infinite density to which any object that becomes a black hole must collapse, according to the theory of general relativity

spacetime system of one time and three space coordinates, with respect to which the time and place of an event can be specified

type II supernova a stellar explosion produced at the endpoint of the evolution of stars whose mass exceeds roughly 10 times the mass of the Sun

KEY EQUATIONS

Schwarzschild radius of a black hole	$R_S = \frac{2GM}{c^2}$

SUMMARY

25.1 The Death of Low-Mass Stars

- During the course of their evolution, stars shed their outer layers and lose a significant fraction of their initial mass.
- Stars with masses of 8 M_{Sun} or less can lose enough mass to become white dwarfs, which have masses less than the Chandrasekhar limit (about 1.4 M_{Sun}).
- The pressure exerted by degenerate electrons keeps white dwarfs from contracting to still-smaller diameters.
- Eventually, white dwarfs cool off to become black dwarfs, stellar remnants made mainly of carbon, oxygen, and

neon.

25.2 Evolution of Massive Stars: An Explosive Finish

- In a massive star, hydrogen fusion in the core is followed by several other fusion reactions involving heavier elements.
- Just before it exhausts all sources of energy, a massive star has an iron core surrounded by shells of silicon, sulfur, oxygen, neon, carbon, helium, and hydrogen.
- The fusion of iron requires energy (rather than releasing it).
- If the mass of a star's iron core exceeds the Chandrasekhar limit (but is less than 3 M_{Sun}), the core collapses until its density exceeds that of an atomic nucleus, forming a neutron star with a typical diameter of 20 kilometers.
- The core rebounds and transfers energy outward, blowing off the outer layers of the star in a type II supernova explosion.

25.3 Supernova Observations

- A supernova occurs on average once every 25 to 100 years in the Milky Way Galaxy.
- Despite the odds, no supernova in our Galaxy has been observed from Earth since the invention of the telescope.
- One nearby supernova (SN 1987A) has been observed in a neighboring galaxy, the Large Magellanic Cloud.
- The star that evolved to become SN 1987A began its life as a blue supergiant, evolved to become a red supergiant, and returned to being a blue supergiant at the time it exploded.
- Studies of SN 1987A have detected neutrinos from the core collapse and confirmed theoretical calculations of what happens during such explosions, including the formation of elements beyond iron.
- Supernovae are a main source of high-energy cosmic rays and can be dangerous for any living organisms in nearby star systems.

25.4 Pulsars and the Discovery of Neutron Stars

- At least some supernovae leave behind a highly magnetic, rapidly rotating neutron star, which can be observed as a pulsar if its beam of escaping particles and focused radiation is pointing toward us.
- Pulsars emit rapid pulses of radiation at regular intervals; their periods are in the range of 0.001 to 10 seconds.
- The rotating neutron star acts like a lighthouse, sweeping its beam in a circle and giving us a pulse of radiation when the beam sweeps over Earth.
- As pulsars age, they lose energy, their rotations slow, and their periods increase.

25.5 The Evolution of Binary Star Systems

- When a white dwarf or neutron star is a member of a close binary star system, its companion star can transfer mass to it.
- Material falling *gradually* onto a white dwarf can explode in a sudden burst of fusion and make a nova.
- If material falls *rapidly* onto a white dwarf, it can push it over the Chandrasekhar limit and cause it to explode completely as a type Ia supernova.
- Another possible mechanism for a type Ia supernova is the merger of two white dwarfs.
- Material falling onto a neutron star can cause powerful bursts of X-ray radiation.
- Transfer of material and angular momentum can speed up the rotation of pulsars until their periods are just a few thousandths of a second.

25.6 Introducing General Relativity

- Einstein proposed the equivalence principle as the foundation of the theory of general relativity. According to this principle, there is no way that anyone or any experiment in a sealed environment can distinguish between free fall and the absence of gravity.

- By considering the consequences of the equivalence principle, Einstein concluded that we live in a curved spacetime.
- The distribution of matter determines the curvature of spacetime; other objects (and even light) entering a region of spacetime must follow its curvature.
- Light must change its path near a massive object not because light is bent by gravity, but because spacetime is.

25.7 Black Holes

- Theory suggests that stars with stellar cores more massive than three times the mass of the Sun at the time they exhaust their nuclear fuel will collapse to become black holes.
- The surface surrounding a black hole, where the escape velocity equals the speed of light, is called the event horizon, and the radius of the surface is called the Schwarzschild radius.
- Nothing, not even light, can escape through the event horizon from the black hole.
- At its center, each black hole is thought to have a singularity, a point of infinite density and zero volume.
- Matter falling into a black hole appears, as viewed by an outside observer, to freeze in position at the event horizon.
- However, if we were riding on the infalling matter, we would pass through the event horizon.
- As we approach the singularity, the tidal forces would tear our bodies apart even before we reach the singularity.

25.8 Evidence for Black Holes

- The best evidence of stellar-mass black holes comes from binary star systems in which (1) one star of the pair is not visible, (2) the flickering X-ray emission is characteristic of an accretion disk around a compact object, and (3) the orbit and characteristics of the visible star indicate that the mass of its invisible companion is greater than 3 M_{Sun}.
- A number of systems with these characteristics have been found.
- Black holes with masses of millions to billions of solar masses are found in the centers of large galaxies.

25.9 Gravitational Wave Astronomy

- General relativity predicts that the rearrangement of matter in space should produce gravitational waves.
- The existence of such waves was first confirmed in observations of a pulsar in orbit around another neutron star whose orbits were spiraling closer and losing energy in the form of gravitational waves.
- In 2015, LIGO found gravitational waves directly by detecting the signal produced by the merger of two stellar-mass black holes, opening a new window on the universe.

CONCEPTUAL QUESTIONS

25.1 The Death of Low-Mass Stars

1. Describe the evolution of a star with a mass like that of the Sun, from the main-sequence phase of its evolution until it becomes a white dwarf.

2. Describe the evolution of a white dwarf over time, in particular how the luminosity, temperature, and radius change.

3. How would a white dwarf that formed from a star that had an initial mass of 1 M_{Sun} be different from a white dwarf that formed from a star that had an initial mass of 9 M_{Sun}?

4. If most stars become white dwarfs at the ends of their lives and the formation of white dwarfs is accompanied by the production of a planetary nebula, why are there more white dwarfs than planetary nebulae in the Galaxy?

25.2 Evolution of Massive Stars: An Explosive Finish

5. Describe the evolution of a massive star (say, 20 times the mass of the Sun) up to the point at which it becomes a supernova. How does the evolution of a massive star differ from that of the Sun? Why?

6. Arrange the following stars in order of their evolution:

A. A star with no nuclear reactions going on in the core, which is made primarily of carbon and oxygen.
B. A star of uniform composition from center to surface; it contains hydrogen but has no nuclear reactions going on in the core.
C. A star that is fusing hydrogen to form helium in its core.
D. A star that is fusing helium to carbon in the core and hydrogen to helium in a shell around the core.
E. A star that has no nuclear reactions going on in the core but is fusing hydrogen to form helium in a shell around the core.

25.3 Supernova Observations

7. What observations from SN 1987A helped confirm theories about supernovae?

8. The Large Magellanic Cloud has about one-tenth the number of stars found in our own Galaxy. Suppose the mix of high- and low-mass stars is exactly the same in both galaxies. Approximately how often does a supernova occur in the Large Magellanic Cloud?

9. Look at the list of the nearest stars in **Appendix D**. Would you expect any of these to become supernovae? Why or why not?

25.4 Pulsars and the Discovery of Neutron Stars

10. How does a white dwarf differ from a neutron star? How does each form? What keeps each from collapsing under its own weight?

11. If the formation of a neutron star leads to a supernova explosion, explain why only three of the hundreds of known pulsars are found in supernova remnants.

12. Describe the evolution of a pulsar over time, in particular how the rotation and pulse signal changes over time.

13. Astronomers believe there are something like 100 million neutron stars in the Galaxy, yet we have only found about 2000 pulsars in the Milky Way. Give several reasons these numbers are so different. Explain each reason.

25.5 The Evolution of Binary Star Systems

14. Would you be more likely to observe a type II supernova (the explosion of a massive star) in a globular cluster or in an open cluster? Why?

15. Would you expect to observe every supernova in our own Galaxy? Why or why not?

16. If a 3 and 8 M_{Sun} star formed together in a binary system, which star would:

A. Evolve off the main sequence first?
B. Form a carbon- and oxygen-rich white dwarf?
C. Be the location for a nova explosion?

17. What observations or types of telescopes would you use to distinguish a binary system that includes a main-sequence star and a white dwarf star from one containing a main-sequence star and a neutron star?

18. How would the spectra of a type II supernova be different from a type Ia supernova? Hint: Consider the characteristics of the objects that are their source.

19. How do the two types of supernovae discussed in this chapter differ? What kind of star gives rise to each type?

20. How is a nova different from a type Ia supernova? How does it differ from a type II supernova?

21. Apart from the masses, how are binary systems with a neutron star different from binary systems with a white dwarf?

25.7 Black Holes

22. A student becomes so excited by the whole idea of black holes that he decides to jump into one. It has a mass 10 times the mass of our Sun. What is the trip like for him? What is it like for the rest of the class, watching from afar?

23. What is an event horizon? Does our Sun have an event horizon around it?

25.8 Evidence for Black Holes

24. If a black hole itself emits no radiation, what evidence do astronomers and physicists today have that the theory of black holes is correct?

25. What characteristics must a binary star have to be a good candidate for a black hole? Why is each of these characteristics important?

26. Why would we not expect to detect X-rays from a disk of matter about an ordinary star?

25.9 Gravitational Wave Astronomy

27. A star begins its life with a mass of 5 M_{Sun} but ends its life as a white dwarf with a mass of 0.8 M_{Sun}. List the stages in the star's life during which it most likely lost some of the mass it started with. How did mass loss occur in each stage?

28. How can the Crab Nebula shine with the energy of something like 100,000 Suns when the star that formed the nebula exploded almost 1000 years ago? Who "pays the bills" for much of the radiation we see coming from the nebula?

29. Suppose no stars more massive than about 2 M_{Sun} had ever formed. Would life as we know it have been able to develop? Why or why not?

30. You have discovered two star clusters. The first cluster contains mainly main-sequence stars, along with some red giant stars and a few white dwarfs. The second cluster also contains mainly main-sequence stars, along with some red giant stars, and a few neutron stars—but no white dwarf stars. What are the relative ages of the clusters? How did you determine your answer?

31. A supernova remnant was recently discovered and found to be approximately 150 years old. Provide possible reasons that this supernova explosion escaped detection.

32. Based upon the evolution of stars, place the following elements in order of least to most common in the Galaxy: gold, carbon, neon. What aspects of stellar evolution formed the basis for how you ordered the elements?

33. What is a gravitational wave and why was it so hard to detect?

34. What are some strong sources of gravitational waves that astronomers hope to detect in the future?

35. Suppose the amount of mass in a black hole doubles. Does the event horizon change? If so, how does it change?

36. Look elsewhere in this book for necessary data, and indicate what the final stage of evolution—white dwarf, neutron star, or black hole—will be for each of these kinds of stars.

A. Spectral type-O main-sequence star
B. Spectral type-B main-sequence star
C. Spectral type-A main-sequence star
D. Spectral type-G main-sequence star
E. Spectral type-M main-sequence star

37. Which is likely to be more common in our Galaxy: white dwarfs or black holes? Why?

38. If the Sun could suddenly collapse to a black hole, how would the period of Earth's revolution about it differ from what it is now?

PROBLEMS

25.1 The Death of Low-Mass Stars

39. What is the average density of the Sun? How does it compare to the average density of Earth?

40. Say that a particular white dwarf has the mass of the Sun but the radius of Earth. What is the acceleration of gravity at the surface of the white dwarf? How much greater is this than g at the surface of Earth? What would you weigh at the surface of the white dwarf (again granting us the dubious notion that you could survive there)?

41. What is the escape velocity from the white dwarf in **Exercise 25.40**? How much greater is it than the escape velocity from Earth?

42. What is the average density of the white dwarf in **Exercise 25.40**? How does it compare to the average density of Earth?

43. If the Sun were replaced by a white dwarf with a surface temperature of 10,000 K and a radius equal to Earth's, how would its luminosity compare to that of the Sun?

25.2 Evolution of Massive Stars: An Explosive Finish

44. The ring around SN 1987A (**Figure 25.12**) initially became illuminated when energetic photons from the supernova interacted with the material in the ring. The radius of the ring is approximately 0.75 light-year from the supernova location. How long after the supernova did the ring become illuminated?

25.3 Supernova Observations

45. A supernova can eject material at a velocity of 10,000 km/s. How long would it take a supernova remnant to expand to a radius of 1 AU? How long would it take to expand to a radius of 1 light-years? Assume that the expansion velocity remains constant and use the relationship: $\text{expansion time} = \frac{\text{distance}}{\text{expansion velocity}}$.

46. A supernova remnant was observed in 2007 to be expanding at a velocity of 14,000 km/s and had a radius of 6.5 light-years. Assuming a constant expansion velocity, in what year did this supernova occur?

47. The ring around SN 1987A (**Figure 25.12**) started interacting with material propelled by the shockwave from the supernova beginning in 1997 (10 years after the explosion). The radius of the ring is approximately 0.75 light-year from the supernova location. How fast is the supernova material moving, assume a constant rate of motion in km/s?

48. Before the star that became SN 1987A exploded, it evolved from a red supergiant to a blue supergiant while remaining at the same luminosity. As a red supergiant, its surface temperature would have been approximately 4000 K, while as a blue supergiant, its surface temperature was 16,000 K. How much did the radius change as it evolved from a red to a blue supergiant?

49. What is the radius of the progenitor star that became SN 1987A? Its luminosity was 100,000 times that of the Sun, and it had a surface temperature of 16,000 K.

50. What is the acceleration of gravity at the surface of the star that became SN 1987A? How does this g compare to that at the surface of Earth? The mass was 20 times that of the Sun and the radius was 41 times that of the Sun.

51. What was the escape velocity from the surface of the SN 1987A progenitor star? How much greater is it than the escape velocity from Earth? The mass was 20 times that of the Sun and the radius was 41 times that of the Sun.

52. What was the average density of the star that became SN 1987A? How does it compare to the average density of Earth? The mass was 20 times that of the Sun and the radius was 41 times that of the Sun.

25.4 Pulsars and the Discovery of Neutron Stars

53. What is the acceleration of gravity (g) at the surface of the Sun? (See **Appendix D** for the Sun's key characteristics.) How much greater is this than g at the surface of Earth? Calculate what you would weigh on the surface of the Sun. Your weight would be your Earth weight multiplied by the ratio of the acceleration of gravity on the Sun to the acceleration of gravity on Earth. (Okay, we know that the Sun does not have a solid surface to stand on and that you would be vaporized if you were at the Sun's photosphere. Humor us for the sake of doing these calculations.)

54. What is the escape velocity from the Sun? How much greater is it than the escape velocity from Earth?

55. Now take a neutron star that has twice the mass of the Sun but a radius of 10 km. What is the acceleration of gravity at the surface of the neutron star? How much greater is this than g at the surface of Earth? What would you weigh at the surface of the neutron star (provided you could somehow not become a puddle of protoplasm)?

56. What is the escape velocity from the neutron star in **Exercise 25.55**? How much greater is it than the escape velocity from Earth?

57. What is the average density of the neutron star in **Exercise 25.55**? How does it compare to the average density of Earth?

58. According to a model described in the text, a neutron star has a radius of about 10 km. Assume that the pulses occur once per rotation. According to Einstein's theory of relatively, nothing can move faster than the speed of light. Check to make sure that this pulsar model does not violate relativity. Calculate the rotation speed of the Crab Nebula pulsar at its equator, given its period of 0.033 s. (Remember that distance equals velocity × time and that the circumference of a circle is given by $2\pi R$).

59. Do the same calculations as in **Exercise 25.58** but for a pulsar that rotates 1000 times per second.

60. If the pulsar shown in **Figure 25.16** is rotating 100 times per second, how many pulses would be detected in one minute? The two beams are located along the pulsar's equator, which is aligned with Earth.

25.7 Black Holes

61. Look up G and c in **Appendix C**, as well as the mass of the Sun in **Appendix D**, and calculate the radius of a black hole that has the same mass as the Sun. (Note that this is only a theoretical calculation. The Sun does not have enough mass to become a black hole.)

62. Suppose you wanted to know the size of black holes with masses that are larger or smaller than the Sun. You could go through all the steps in **Exercise 25.61**, wrestling with a lot of large numbers with large exponents. You could be clever, however, and evaluate all the constants in the equation once and then simply vary the mass. You could even express the mass in terms of the Sun's mass and make future calculations really easy. Show that the event horizon equation is equivalent to saying that the radius of the event horizon is equal to 3 km times the mass of the black hole in units of the Sun's mass.

63. Use the result from **Exercise 25.62** to calculate the radius of a black hole with a mass equal to: the Earth, a B0-type main-sequence star, a globular cluster, and the Milky Way Galaxy. Look elsewhere in this text and the appendixes for tables that provide data on the mass of these four objects.

64. Since the force of gravity a significant distance away from the event horizon of a black hole is the same as that of an ordinary object of the same mass, Kepler's third law is valid. Suppose that Earth collapsed to the size of a golf ball. What would be the period of revolution of the Moon, orbiting at its current distance of 400,000 km? Use Kepler's third law to calculate the period of revolution of a spacecraft orbiting at a distance of 6000 km.

ADDITIONAL PROBLEMS

65. One way to calculate the radius of a star is to use its luminosity and temperature and assume that the star radiates approximately like a blackbody. Astronomers have measured the characteristics of central stars of planetary nebulae and have found that a typical central star is 16 times as luminous and 20 times as hot (about 110,000 K) as the Sun. Find the radius in terms of the Sun's. How does this radius compare with that of a typical white dwarf?

26 | THE MILKY WAY GALAXY

Chapter Outline

Introduction

Figure 26.1 The Milky Way rises over Square Tower, an ancestral pueblo building at Hovenweep National Monument in Utah. Many stars and dark clouds of dust combine to make a spectacular celestial sight of our home Galaxy. The location has been designated an International Dark Sky Park by the International Dark Sky Association.

Today, we know that our Sun is just one of the many billions of stars that make up the huge cosmic island we call the Milky Way Galaxy. How can we "weigh" such an enormous system of stars and measure its total mass?

One of the most striking features you can see in a truly dark sky—one without light pollution—is the band of faint white light called the Milky Way, which stretches from one horizon to the other. The name comes from an ancient Greek legend that compared its faint white splash of light to a stream of spilled milk. But folktales differ from culture to culture: one East African tribe thought of the hazy band as the smoke of ancient campfires, several Native American stories tell of a path across the sky traveled by sacred animals, and in Siberia, the diffuse arc was known as the seam of the tent of the sky.

In 1610, Galileo made the first telescopic survey of the Milky Way and discovered that it is composed of a multitude of individual stars. Today, we know that the Milky Way comprises our view inward of the huge cosmic pinwheel that we call the Milky Way Galaxy and that is our home. Moreover, our Galaxy is now recognized as just one galaxy among many billions of other galaxies in the cosmos.

26.1 | The Architecture of the Galaxy

Learning Objectives

By the end of this section, you will be able to:

- Explain why William and Caroline Herschel concluded that the Milky Way has a flattened structure centered on the Sun and solar system
- Describe the challenges of determining the Galaxy's structure from our vantage point within it
- Identify the main components of the Galaxy

The **Milky Way Galaxy** surrounds us, and you might think it is easy to study because it is so close. However, the very fact that we are embedded within it presents a difficult challenge. Suppose you were given the task of mapping New York City. You could do a much better job from a helicopter flying over the city than you could if you were standing in Times Square. Similarly, it would be easier to map our Galaxy if we could only get a little way outside it, but instead we are trapped inside and way out in its suburbs—far from the galactic equivalent of Times Square.

Herschel Measures the Galaxy

In 1785, William Herschel (**Figure 26.2**) made the first important discovery about the architecture of the Milky Way Galaxy. Using a large reflecting telescope that he had built, William and his sister Caroline counted stars in different directions of the sky. They found that most of the stars they could see lay in a flattened structure encircling the sky, and that the numbers of stars were about the same in any direction around this structure. Herschel therefore concluded that the stellar system to which the Sun belongs has the shape of a disk or wheel (he might have called it a Frisbee except Frisbees hadn't been invented yet), and that the Sun must be near the hub of the wheel (**Figure 26.3**).

Figure 26.2 William Herschel was a German musician who emigrated to England and took up astronomy in his spare time. He discovered the planet Uranus, built several large telescopes, and made measurements of the Sun's place in the Galaxy, the Sun's motion through space, and the comparative brightnesses of stars. This painting shows William and his sister Caroline polishing a telescope lens. (credit: modification of work by the Wellcome Library)

To understand why Herschel reached this conclusion, imagine that you are a member of a band standing in formation during halftime at a football game. If you count the band members you see in different directions and get about the same number each time, you can conclude that the band has arranged itself in a circular pattern with you at the center. Since you see no band members above you or underground, you know that the circle made by the band is much flatter than it is wide.

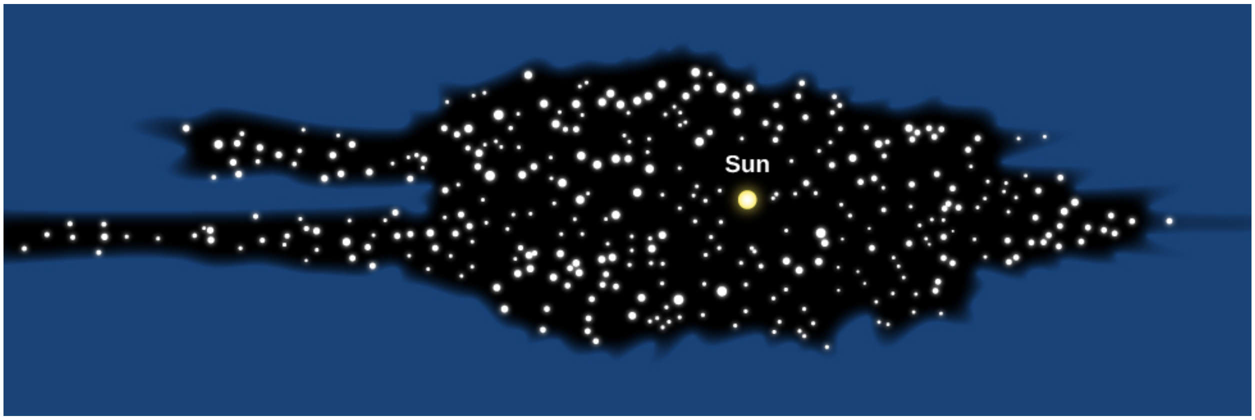


Figure 26.3 Herschel constructed this cross section of the Galaxy by counting stars in various directions.

We now know that Herschel was right about the shape of our system, but wrong about where the Sun lies within the disk. We live in a dusty Galaxy. Because interstellar dust absorbs the light from stars, Herschel could see only those stars within about 6000 light-years of the Sun. Today we know that this is a very small section of the entire 100,000-light-year-diameter disk of stars that makes up the Galaxy.

Harlow Shapley: Mapmaker to the Stars

Until the early 1900s, astronomers generally accepted Herschel's conclusion that the Sun is near the center of the Galaxy. The discovery of the Galaxy's true size and our actual location came about largely through the efforts of Harlow Shapley. In 1917, he was studying RR Lyrae variable stars in globular clusters. By comparing the known intrinsic luminosity of these stars to how bright they appeared, Shapley could calculate how far away they are. (Recall that it is distance that makes the stars look dimmer than they would be "up close," and that the brightness fades as the distance squared.) Knowing the distance to any star in a cluster then tells us the distance to the cluster itself.

Globular clusters can be found in regions that are free of interstellar dust and so can be seen at very large distances. When Shapley used the distances and directions of 93 globular clusters to map out their positions in space, he found that the clusters are distributed in a spherical volume, which has its center not at the Sun but at a distant point along the Milky Way in the direction of Sagittarius. Shapley then made the bold assumption, verified by many other observations since then, that the point on which the system of globular clusters is centered is also the center of the entire Galaxy (**Figure 26.4**).

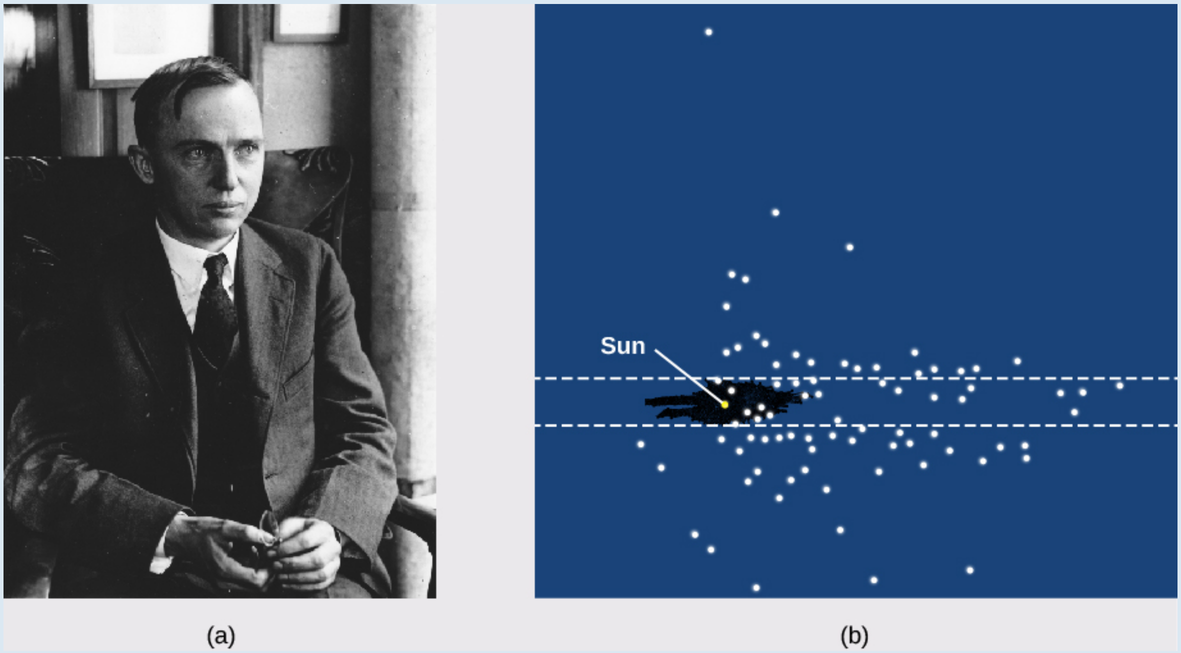


Figure 26.4 (a) Shapley poses for a formal portrait. (b) His diagram shows the location of globular clusters, with the position of the Sun also marked. The black area shows Herschel's old diagram, centered on the Sun, approximately to scale.

Shapley's work showed once and for all that our star has no special place in the Galaxy. We are in a nondescript region of the Milky Way, only one of 200 to 400 billion stars that circle the distant center of our Galaxy.

Born in 1885 on a farm in Missouri, Harlow Shapley at first dropped out of school with the equivalent of only a fifth-grade education. He studied at home and at age 16 got a job as a newspaper reporter covering crime stories. Frustrated by the lack of opportunities for someone who had not finished high school, Shapley went back and completed a six-year high-school program in only two years, graduating as class valedictorian.

In 1907, at age 22, he went to the University of Missouri, intent on studying journalism, but found that the school of journalism would not open for a year. Leafing through the college catalog (or so he told the story later), he chanced to see "Astronomy" among the subjects beginning with "A." Recalling his boyhood interest in the stars, he decided to study astronomy for the next year (and the rest, as the saying goes, is history).

Upon graduation Shapley received a fellowship for graduate study at Princeton and began to work with the brilliant Henry Norris Russell (see the **Henry Norris Russell** feature box). For his PhD thesis, Shapley made major contributions to the methods of analyzing the behavior of eclipsing binary stars. He was also able to show that cepheid variable stars are not binary systems, as some people thought at the time, but individual stars that pulsate with striking regularity.

Impressed with Shapley's work, George Ellery Hale offered him a position at the Mount Wilson Observatory, where the young man took advantage of the clear mountain air and the 60-inch reflector to do his pioneering study of variable stars in globular clusters.

Shapley subsequently accepted the directorship of the Harvard College Observatory, and over the next 30 years, he and his collaborators made contributions to many fields of astronomy, including the study of neighboring galaxies, the discovery of dwarf galaxies, a survey of the distribution of galaxies in the universe, and much more. He wrote a series of nontechnical books and articles and became known as one of the most effective popularizers of astronomy. Shapley enjoyed giving lectures around the country, including at many smaller colleges where students and faculty rarely got to interact with scientists of his caliber.

During World War II, Shapley helped rescue many scientists and their families from Eastern Europe; later, he helped found UNESCO, the United Nations Educational, Scientific, and Cultural Organization. He wrote a pamphlet called *Science from Shipboard* for men and women in the armed services who had to spend many weeks on board transport ships to Europe. And during the difficult period of the 1950s, when congressional committees began their "witch hunts" for communist sympathizers (including such liberal leaders as Shapley), he spoke out forcefully and fearlessly in defense of the freedom of thought and expression. A man of many interests, he was fascinated by the behavior of ants, and wrote scientific papers about them as well as about galaxies.

By the time he died in 1972, Shapley was acknowledged as one of the pivotal figures of modern astronomy, a "twentieth-century Copernicus" who mapped the Milky Way and showed us our place in the Galaxy.

To find more information about **Shapley's life and work (https://openstaxcollege.org/l/30shapbrumed)**, see the entry for him on the Bruce Medalists website. (This site features the winners of the Bruce Medal of the Astronomical Society of the Pacific, one of the highest honors in astronomy; the list is a who's who of some of the greatest astronomers of the last twelve decades.)

Disks and Haloes

With modern instruments, astronomers can now penetrate the "smog" of the Milky Way by studying radio and infrared emissions from distant parts of the Galaxy. Measurements at these wavelengths (as well as observations of other galaxies like ours) have given us a good idea of what the Milky Way would look like if we *could* observe it from a distance.

Figure 26.5 sketches what we would see if we could view the Galaxy face-on and edge-on. The brightest part of the Galaxy consists of a thin, circular, rotating disk of stars distributed across a region about 100,000 light-years in diameter and about 2000 light-years thick. (Given how thin the disk is, perhaps a CD is a more appropriate analogy than a wheel.) The very youngest stars, and the dust and gas from which stars form, are found typically within 100 light-years of the plane of the Milky Way Galaxy. The mass of the interstellar matter is about 15% of the mass of the stars in this disk.

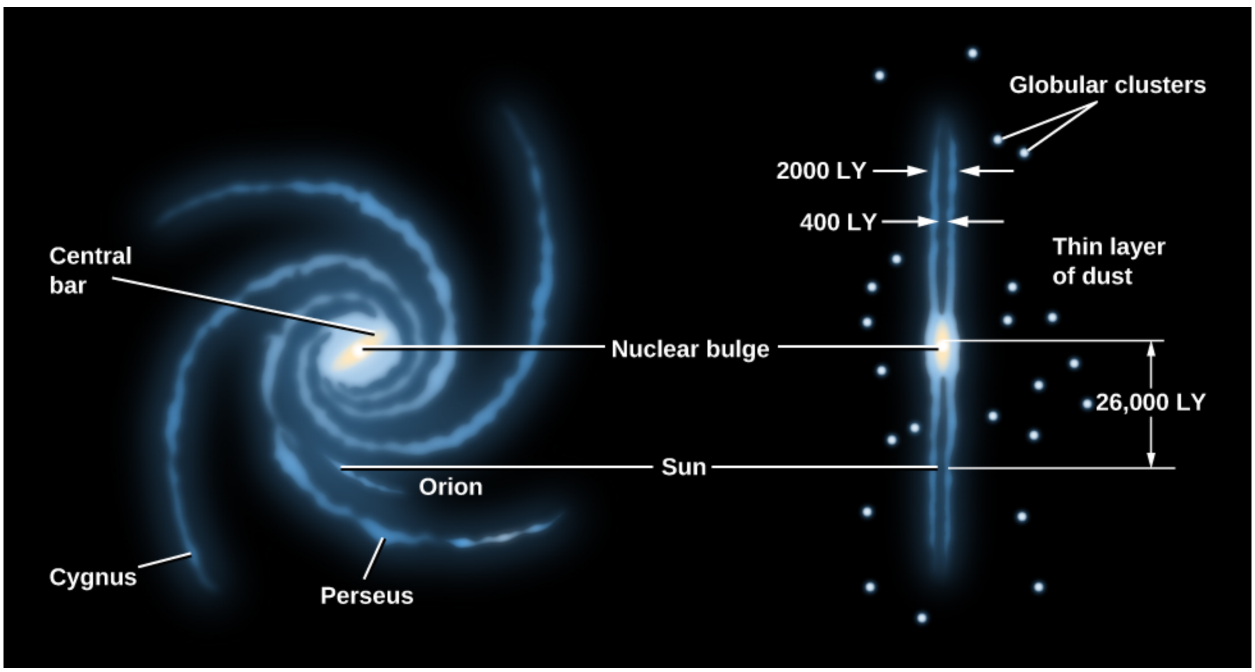


Figure 26.5 The left image shows the face-on view of the spiral disk; the right image shows the view looking edge-on along the disk. The major spiral arms are labeled. The Sun is located on the inside edge of the short Orion spur.

As the diagram in **Figure 26.5** shows, the stars, gas, and dust are not spread evenly throughout the disk but are concentrated into a central bar and a series of spiral arms. Recent infrared observations have confirmed that the central bar is composed mostly of old yellow-red stars. The two main spiral arms appear to connect with the ends of the bar. They are highlighted by the blue light from young hot stars. We know many other spiral galaxies that also have bar-shaped concentrations of stars in their central regions; for that reason they are called *barred spirals*. **Figure 26.6** shows two other galaxies—one without a bar and one with a strong bar—to give you a basis for comparison to our own. We will describe our spiral structure in more detail shortly. The Sun is located about halfway between the center of the Galaxy and the edge of the disk and only about 70 light-years above its central plane.

Figure 26.6 (a) This image shows the unbarred spiral galaxy M74. It contains a small central bulge of mostly old yellow-red stars, along with spiral arms that are highlighted with the blue light from young hot stars. (b) This image shows the strongly barred spiral galaxy NGC 1365. The bulge and the fainter bar both appear yellowish because the brightest stars in them are mostly old yellow and red giants. Two main spiral arms project from the ends of the bar. As in M74, these spiral arms are populated with blue stars and red patches of glowing gas—hallmarks of recent star formation. The Milky Way Galaxy is thought to have a barred spiral structure that is intermediate between these two examples. (credit a: modification of work by ESO/PESSTO/S. Smartt; credit b: modification of work by ESO)

Our thin disk of young stars, gas, and dust is embedded in a thicker but more diffuse disk of older stars; this thicker disk extends about 1000 light-years above and 1000 light-years below the midplane of the thin disk and contains only about 5% as much mass as the thin disk. The stars thin out with distance from the galactic plane and don't have a sharp edge. Approximately 2/3 of the stars in the thick disk are within 1000 light-years of midplane.

Close in to the galactic center (within about 10,000 light-years), the stars are no longer confined to the disk but form a **central bulge** (or nuclear bulge). When we observe with visible light, we can glimpse the stars in the bulge only in those rare directions where there happens to be relatively little interstellar dust. The first picture that actually succeeded in showing the bulge as a whole was taken at infrared wavelengths (**Figure 26.7**).

Figure 26.7 This beautiful infrared map, showing half a billion stars, was obtained as part of the Two Micron All Sky Survey (2MASS). Because interstellar dust does not absorb infrared as strongly as visible light, this view reveals the previously hidden bulge of old stars that surrounds the center of our Galaxy, along with the Galaxy's thin disk component. (credit: modification of work by 2MASS/J. Carpenter, T. H. Jarrett, and R. Hurt)

The fact that much of the bulge is obscured by dust makes its shape difficult to determine. For a long time, astronomers assumed it was spherical. However, infrared images and other data indicate that the bulge is about two times longer than it is wide, and shaped rather like a peanut. The relationship between this elongated inner bulge and the larger bar of stars remains uncertain. At the very center of the nuclear bulge is a tremendous concentration of matter, which we will discuss later in this chapter.

In our Galaxy, the thin and thick disks and the nuclear bulge are embedded in a spherical **halo** of very old, faint stars that extends to a distance of at least 150,000 light-years from the galactic center. Most of the globular clusters are also found in this halo.

The mass in the Milky Way extends even farther out, well beyond the boundary of the luminous stars to a distance of at least 200,000 light-years from the center of the Galaxy. This invisible mass has been give the name *dark matter* because it emits no light and cannot be seen with any telescope. Its composition is unknown, and it can be detected only because of its gravitational effects on the motions of luminous matter that we can see. We know that this extensive **dark matter halo** exists because of its effects on the orbits of distant star clusters and other dwarf galaxies that are associated with the Galaxy. This mysterious halo will be a subject of the section on **The Mass of the Galaxy**, and the properties of dark matter will be discussed more in the chapter on **Big Bang Cosmology**.

Some vital statistics of the thin and thick disks and the stellar halo are given in **Table 26.1**, with an illustration in **Figure 26.8**. Note particularly how the ages of stars correlate with where they are found. As we shall see, this information holds important clues to how the Milky Way Galaxy formed.

Characteristics of the Milky Way Galaxy

Property	Thin Disk	Thick Disk	Stellar Halo (Excludes Dark Matter)
Stellar mass	4×10^{10} M_{Sun}	A few percent of the thin disk mass	10^{10} M_{Sun}
Luminosity	3×10^{10} L_{Sun}	A few percent of the thin disk luminosity	8×10^{8} L_{Sun}
Typical age of stars	1 million to 10 billion years	11 billion years	13 billion years
Heavier-element abundance	High	Intermediate	Very low
Rotation	High	Intermediate	Very low

Table 26.1

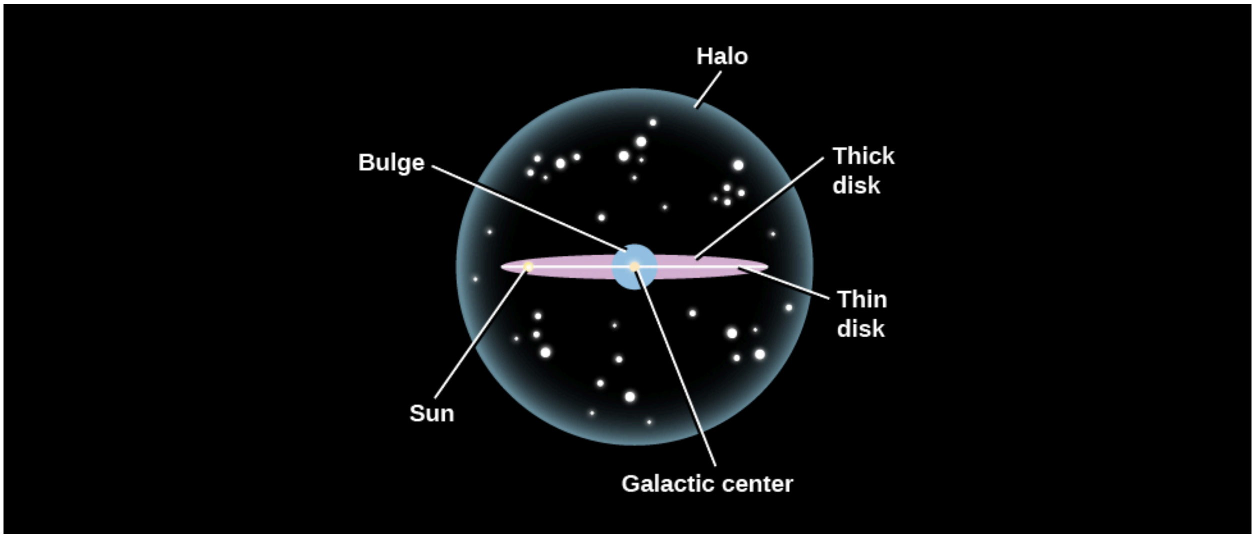


Figure 26.8 This schematic shows the major components of our Galaxy.

Establishing this overall picture of the Galaxy from our dust-shrouded viewpoint inside the thin disk has been one of the great achievements of modern astronomy (and one that took decades of effort by astronomers working with a wide range of telescopes). One thing that helped enormously was the discovery that our Galaxy is not unique in its characteristics. There are many other flat, spiral-shaped islands of stars, gas, and dust in the universe. For example, the Milky Way somewhat resembles the Andromeda galaxy, which, at a distance of about 2.3 million light-years, is our nearest neighboring giant spiral galaxy. Just as you can get a much better picture of yourself if someone else takes the photo from a distance away, pictures and other diagnostic observations of nearby galaxies that resemble ours have been vital to our understanding of the properties of the Milky Way.

The Milky Way Galaxy in Myth and Legend

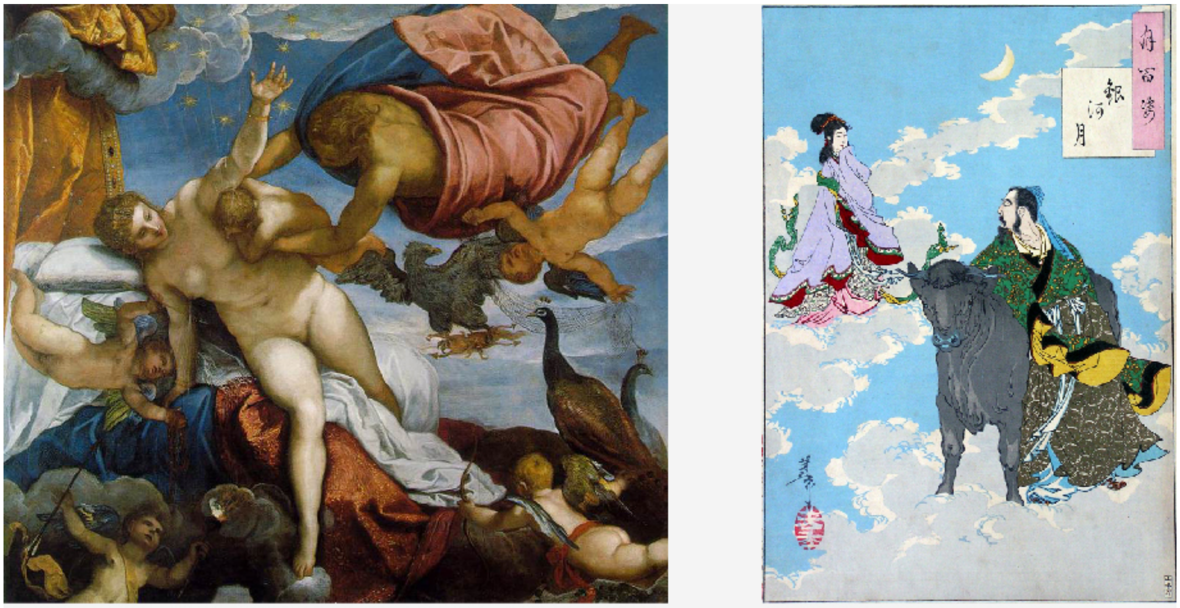
To most of us living in the twenty-first century, the Milky Way Galaxy is an elusive sight. We must make an effort to leave our well-lit homes and streets and venture beyond our cities and suburbs into less populated environments. Once the light pollution subsides to negligible levels, the Milky Way can be readily spotted arching over the sky on clear, moonless nights. The Milky Way is especially bright in late summer and early fall in the Northern Hemisphere. Some of the best places to view the Milky Way are in our national and state parks, where residential and industrial developments have been kept to a minimum. Some of these parks host special sky-gazing events that are definitely worth checking out—especially during the two weeks surrounding the new moon, when the faint stars and Milky Way don’t have to compete with the Moon’s brilliance.

Go back a few centuries, and these starlit sights would have been the norm rather than the exception. Before the advent of electric or even gas lighting, people relied on short-lived fires to illuminate their homes and byways. Consequently, their night skies were typically much darker. Confronted by myriad stellar patterns and the Milky Way’s gauzy band of diffuse light, people of all cultures developed myths to make sense of it all.

Some of the oldest myths relating to the Milky Way are maintained by the aboriginal Australians through their rock painting and storytelling. These legacies are thought to go back tens of thousands of years, to when the aboriginal people were being “dreamed” along with the rest of the cosmos. The Milky Way played a central role as an arbiter of the Creation. Taking the form of a great serpent, it joined with the Earth serpent to dream and thus create all the creatures on Earth.

The ancient Greeks viewed the Milky Way as a spray of milk that spilled from the breast of the goddess Hera. In this legend, Zeus had secretly placed his infant son Heracles at Hera’s breast while she was asleep in order to give his half-human son immortal powers. When Hera awoke and found Heracles suckling, she pushed him away, causing her milk to spray forth into the cosmos (**Figure 26.9**).

The dynastic Chinese regarded the Milky Way as a “silver river” that was made to separate two star-crossed lovers. To the east of the Milky Way, Zhi Nu, the weaving maiden, was identified with the bright star Vega in the constellation of Lyra the Harp. To the west of the Milky Way, her lover Niu Lang, the cowherd, was associated with the star Altair in the constellation of Aquila the Eagle. They had been exiled on opposite sides of the Milky Way by Zhi Nu’s mother, the Queen of Heaven, after she heard of their secret marriage and the birth of their two children. However, once a year, they are permitted to reunite. On the seventh day of the seventh lunar month (which typically occurs in our month of August), they would meet on a bridge over the Milky Way that thousands of magpies had made (**Figure 26.9**). This romantic time continues to be celebrated today as Qi Xi, meaning “Double Seventh,” with couples reenacting the cosmic reunion of Zhi Nu and Niu Lang.



(a) (b)

Figure 26.9 (a) *Origin of the Milky Way* by Jacopo Tintoretto (circa 1575) illustrates the Greek myth that explains the formation of the Milky Way. (b) *The Moon of the Milky Way* by Japanese painter Tsukioka Yoshitoshi depicts the Chinese legend of Zhi Nu and Niu Lang.

To the Quechua Indians of Andean Peru, the Milky Way was seen as the celestial abode for all sorts of cosmic creatures. Arrayed along the Milky Way are myriad dark patches that they identified with partridges, llamas, a toad, a snake, a fox, and other animals. The Quechua's orientation toward the dark regions rather than the glowing band of starlight appears to be unique among all the myth makers. Likely, their access to the richly structured southern Milky Way had something to do with it.

Among Finns, Estonians, and related northern European cultures, the Milky Way is regarded as the "pathway of birds" across the night sky. Having noted that birds seasonally migrate along a north-south route, they identified this byway with the Milky Way. Recent scientific studies have shown that this myth is rooted in fact: the birds of this region use the Milky Way as a guide for their annual migrations.

Today, we regard the Milky Way as our galactic abode, where the foment of star birth and star death plays out on a grand stage, and where sundry planets have been found to be orbiting all sorts of stars. Although our perspective on the Milky Way is based on scientific investigations, we share with our forebears an affinity for telling stories of origin and transformation. In these regards, the Milky Way continues to fascinate and inspire us.

26.2 | Spiral Structure

Learning Objectives

By the end of this section, you will be able to:

- Describe the structure of the Milky Way Galaxy and how astronomers discovered it
- Compare theoretical models for the formation of spiral arms in disk galaxies

Astronomers were able to make tremendous progress in mapping the spiral structure of the Milky Way after the discovery of the 21-cm line that comes from cool hydrogen. The obscuring effect of interstellar dust prevents us from seeing stars at large distances in the disk at visible wavelengths. However, radio waves of 21-cm wavelength pass right through the dust, enabling astronomers to detect hydrogen atoms throughout the Galaxy. More recent surveys of the infrared emission from stars in the disk have provided a similar dust-free perspective of our Galaxy's stellar distribution. Despite all this progress over the past fifty years, we are still just beginning to pin down the precise structure of our Galaxy.

The Arms of the Milky Way

Our radio observations of the disk's gaseous component indicate that the Galaxy has two major spiral arms that emerge from the bar and several fainter arms and shorter spurs. You can see a recently assembled map of our Galaxy's arm structure—derived from studies in the infrared—in **Figure 26.10**.

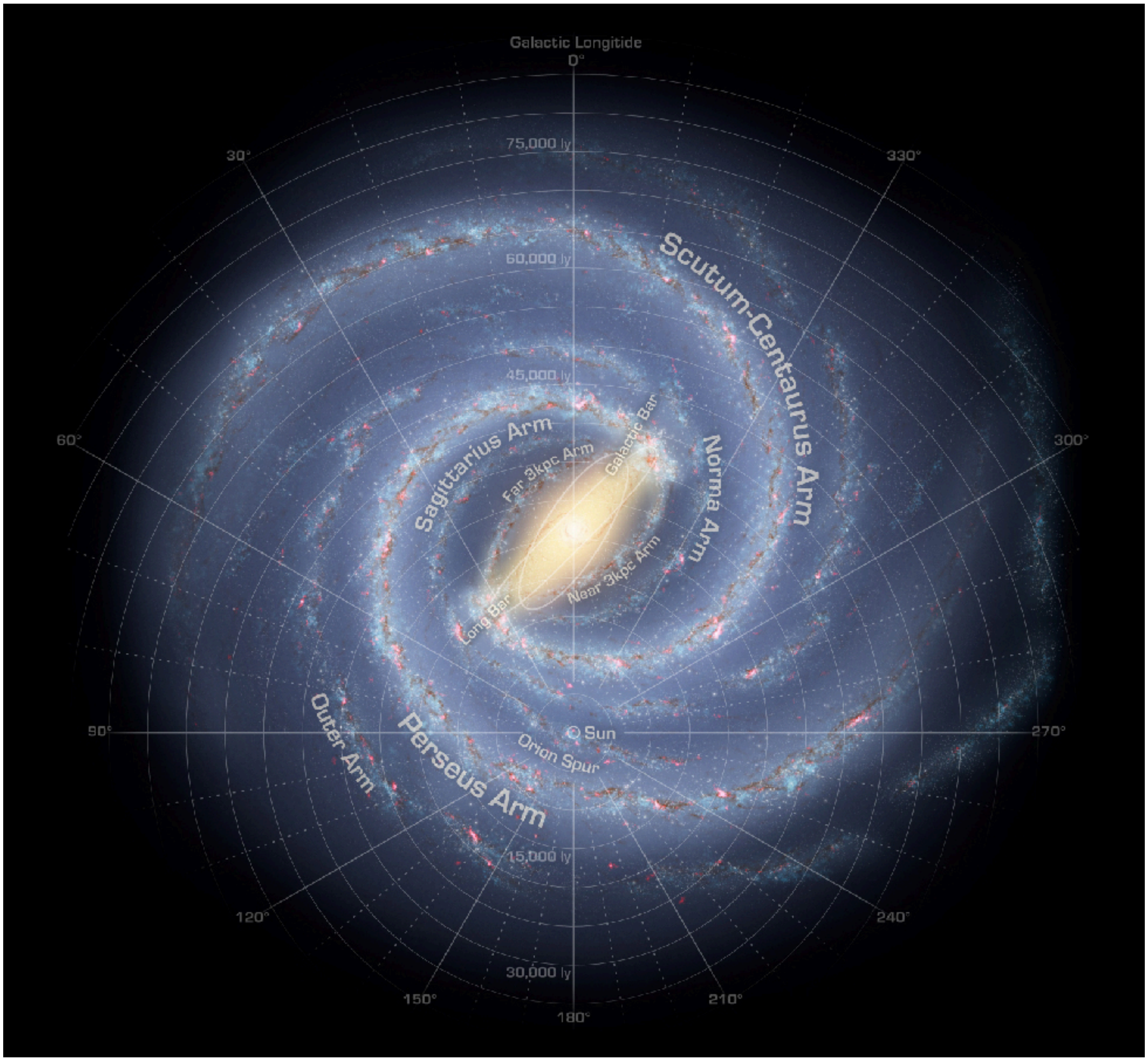


Figure 26.10 Here, we see the Milky Way Galaxy as it would look from above. This image, assembled from data from NASA's WISE mission, shows that the Milky Way Galaxy has a modest bar in its central regions. Two spiral arms, Scutum-Centaurus and Perseus, emerge from the ends of the bar and wrap around the bulge. The Sagittarius and Outer arms have fewer stars than the other two arms. (credit: modification of work by NASA/JPL-Caltech/R. Hurt (SSC/Caltech))

The Sun is near the inner edge of a short arm called the Orion Spur, which is about 10,000 light-years long and contains such conspicuous features as the Cygnus Rift (the great dark nebula in the summer Milky Way) and the bright Orion Nebula. **Figure 26.11** shows a few other objects that share this small section of the Galaxy with us and are easy to see. Remember, the farther away we try to look from our own arm, the more the dust in the Galaxy builds up and makes it hard to see with visible light.

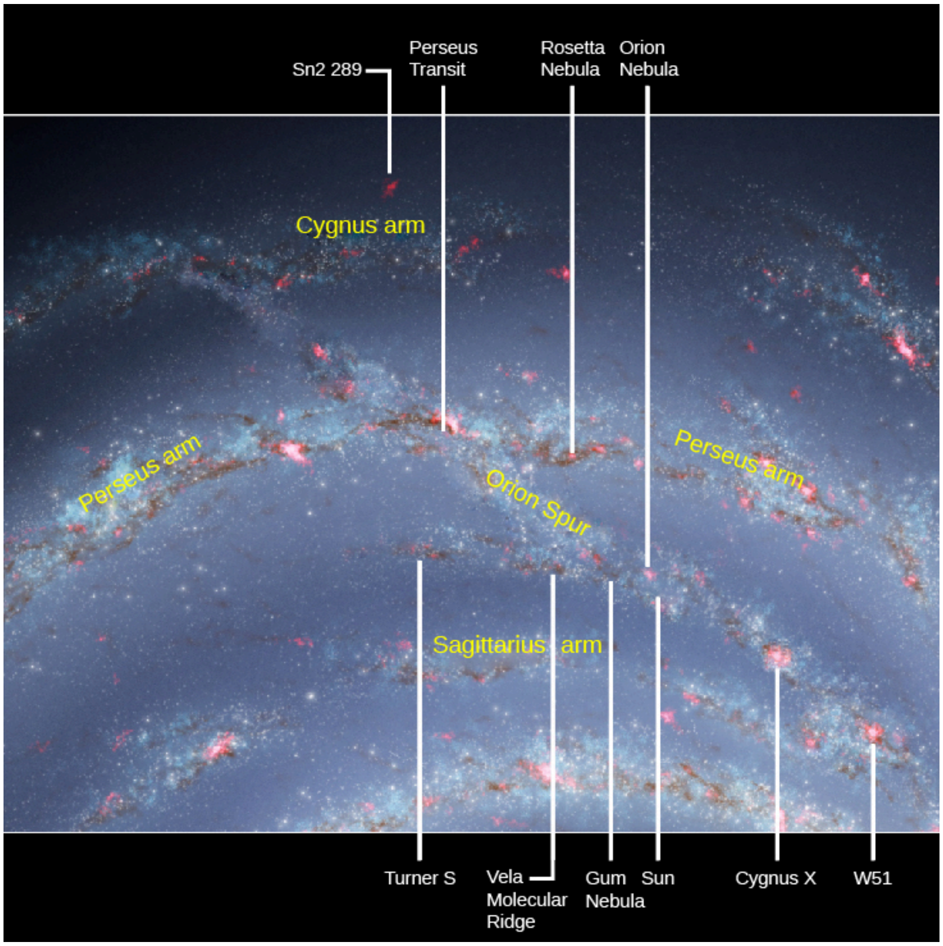


Figure 26.11 The Sun is located in the Orion Spur, which is a minor spiral arm located between two other arms. In this diagram, the white lines point to some other noteworthy objects that share this feature of the Milky Way Galaxy with the Sun. (credit: modification of work by NASA/JPL-Caltech)

Formation of Spiral Structure

At the Sun's distance from its center, the Galaxy does not rotate like a solid wheel or a CD inside your player. Instead, the way individual objects turn around the center of the Galaxy is more like the solar system. Stars, as well as the clouds of gas and dust, obey Kepler's third law. Objects farther from the center take longer to complete an orbit around the Galaxy than do those closer to the center. In other words, stars (and interstellar matter) in larger orbits in the Galaxy trail behind those in smaller ones. This effect is called **differential galactic rotation**.

Differential rotation would appear to explain why so much of the material in the disk of the Milky Way is concentrated into elongated features that resemble **spiral arms**. No matter what the original distribution of the material might be, the differential rotation of the Galaxy can stretch it out into spiral features. **Figure 26.12** shows the development of spiral arms from two irregular blobs of interstellar matter. Notice that as the portions of the blobs closest to the galactic center move faster, those farther out trail behind.

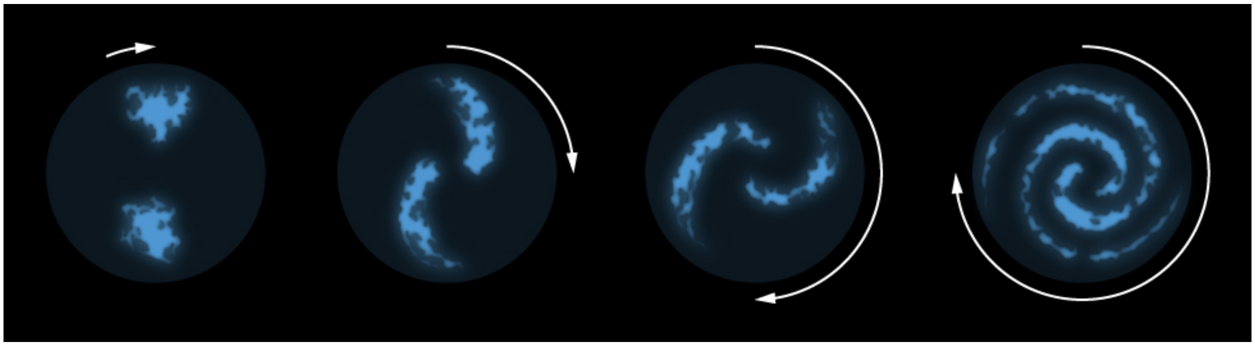


Figure 26.12 This sketch shows how spiral arms might form from irregular clouds of interstellar material stretched out by the different rotation rates throughout the Galaxy. The regions farthest from the galactic center take longer to complete their orbits and thus lag behind the inner regions. If this were the only mechanism for creating spiral arms, then over time the spiral arms would completely wind up and disappear. Since many galaxies have spiral arms, they must be long-lived, and there must be other processes at work to maintain them.

But this picture of spiral arms presents astronomers with an immediate problem. If that's all there were to the story, differential rotation—over the roughly 13-billion-year history of the Galaxy—would have wound the Galaxy's arms tighter and tighter until all semblance of spiral structure had disappeared. But did the Milky Way actually have spiral arms when it formed 13 billion years ago? And do spiral arms, once formed, last for that long a time?

With the advent of the Hubble Space Telescope, it has become possible to observe the structure of very distant galaxies and to see what they were like shortly after they began to form more than 13 billion years ago. What the observations show is that galaxies in their infancy had bright, clumpy star-forming regions, but no regular spiral structure.

Over the next few billion years, the galaxies began to "settle down." The galaxies that were to become spirals lost their massive clumps and developed a central bulge. The turbulence in these galaxies decreased, rotation began to dominate the motions of the stars and gas, and stars began to form in a much quieter disk. Smaller star-forming clumps began to form fuzzy, not-very-distinct spiral arms. Bright, well-defined spiral arms began to appear only when the galaxies were about 3.6 billion years old. Initially, there were two well-defined arms. Multi-armed structures in galaxies like we see in the Milky Way appeared only when the universe was about 8 billion years old.

We will discuss the history of galaxies in more detail in **The Evolution and Distribution of Galaxies**. But, even from our brief discussion, you can get the sense that the spiral structures we now observe in mature galaxies have come along later in the full story of how things develop in the universe.

Scientists have used supercomputer calculations to model the formation and evolution of the arms. These calculations follow the motions of up to 100 million "star particles" to see whether gravitational forces can cause them to form spiral structure. What these calculations show is that giant molecular clouds have enough gravitational influence over their surroundings to initiate the formation of structures that look like spiral arms. These arms then become self-perpetuating and can survive for at least several billion years. The arms may change their brightness over time as star formation comes and goes, but they are not temporary features. The concentration of matter in the arms exerts sufficient gravitational force to keep the arms together over long periods of time.

26.3 | The Mass of the Galaxy

Learning Objectives

By the end of this section, you will be able to:

- Describe historical attempts to determine the mass of the Galaxy,
- Use the orbital velocity law to calculate the total amount of mass that resides inside a circular orbit of s given radius.
- Interpret the observed rotation curve of our Galaxy to suggest the presence of dark matter whose distribution extends well beyond the Sun's orbit.

When we described the sections of the Milky Way, we said that the stars are now known to be surrounded by a much larger halo of invisible matter. Let's see how this surprising discovery was made.

Kepler Helps Weigh the Galaxy

The Sun, like all the other stars in the Galaxy, orbits the center of the Milky Way. Our star's orbit is nearly circular and lies in the Galaxy's disk. The speed of the Sun in its orbit is about 200 kilometers per second, which means it takes us approximately 225 million years to go once around the center of the Galaxy. We call the period of the Sun's revolution the *galactic year*. It is a long time compared to human time scales; during the entire lifetime of Earth, only about 20 galactic years have passed. This means that we have gone only a tiny fraction of the way around the Galaxy in all the time that humans have gazed into the sky.

We can use the information about the Sun's orbit to estimate the mass of the Galaxy (just as we could "weigh" the Sun by monitoring the orbit of a planet around it—see **Kepler's laws**). Let's assume that the Sun's orbit is circular and that the Galaxy is roughly spherical, (we know the Galaxy is shaped more like a disk, but to simplify the calculation we will make this assumption, which illustrates the basic approach). Long ago, Newton showed that if you have matter distributed in the shape of a sphere, then it is simple to calculate the pull of gravity on some object just outside that sphere: you can assume that gravity acts as if all the matter were concentrated at a point in the center of the sphere. For our calculation, then, we can assume that all the mass that lies inward of the Sun's position is concentrated at the center of the Galaxy, and that the Sun orbits that point from a distance of about 26,000 light-years.

The Orbital Velocity Law

We start with Newton's version of Kepler's Third Law:

$$T^2 = \frac{4\pi^2}{GM}a^3 \quad (26.1)$$

where T is the orbital period and a is the average orbital radius.

Newton showed that, for continuous distribution of mass in three dimensions, the mass in the equation can be replaced by the total mass, M_r, contained within a sphere of radius a.

Since the orbital velocity is merely the circumference divided by the period:

$$v = \frac{2\pi a}{T} \quad (26.2)$$

we can combine **Equation 26.1** and **Equation 26.2** to generate:

$$M_r = \frac{av^2}{G} \quad (26.3)$$

Determining the total mass of our galaxy is just the sort of situation to which the orbital velocity law can be applied. Plugging numbers into **Equation 26.3**, we can calculate the total galactic mass that lies inside the orbit of the Sun. The result (about 100 billion times the mass of the Sun) is an estimate of the mass of the Milky Way. More sophisticated calculations based on more sophisticated models give a similar result.

Our estimate tells us how much mass is contained in the volume inside the Sun's orbit. This is a good estimate for the total mass of the Galaxy only if hardly any mass lies outside the Sun's orbit. For many years astronomers thought this assumption was reasonable. The number of bright stars and the amount of *luminous matter* (meaning any material from which we can detect electromagnetic radiation) both drop off dramatically at distances of more than about 30,000 light-years from the galactic center. Little did we suspect how wrong our assumption was.

A Galaxy of Mostly Invisible Matter

In science, what seems to be a reasonable assumption can later turn out to be wrong (which is why we continue to do observations and experiments every chance we get). There is a lot more to the Milky Way than meets the eye (or our instruments). While there is relatively little luminous matter beyond 30,000 light-years, we now know that a lot of *invisible matter* exists at great distances from the galactic center.

We can understand how astronomers detected this invisible matter by remembering that according to Kepler's third law, objects orbiting at large distances from a massive object will move more slowly than objects that are closer to that central mass. In the case of the solar system, for example, the outer planets move more slowly in their orbits than the planets close to the Sun.

There are a few objects, including globular clusters and some nearby small satellite galaxies, that lie well outside the luminous boundary of the Milky Way. If most of the mass of our Galaxy were concentrated within the luminous region, then these very distant objects should travel around their galactic orbits at lower speeds than, for example, the Sun does.

It turns out, however, that the few objects seen at large distances from the luminous boundary of the Milky Way Galaxy are *not* moving more slowly than the Sun. There are some globular clusters and RR Lyrae stars between 30,000 and 150,000 light-years from the center of the Galaxy, and their orbital velocities are even greater than the Sun's (**Figure 26.13**).

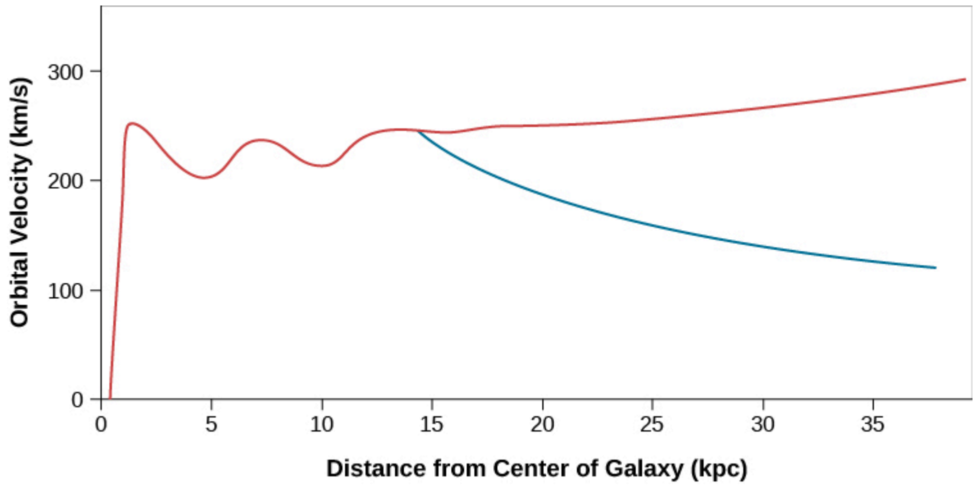


Figure 26.13 The orbital speed of carbon monoxide (CO) and hydrogen (H) gas at different distances from the center of the Milky Way Galaxy is shown in red. The blue curve shows what the rotation curve would look like if all the matter in the Galaxy were located inside a radius of 50,000 light-years. Instead of going down, the speed of gas clouds farther out remains high, indicating a great deal of mass beyond the Sun's orbit. The horizontal axis shows the distance from the galactic center in kiloparsecs (where a kiloparsec equals 3,260 light-years).

What do these higher speeds mean? Kepler's third law tells us how fast objects must orbit a source of gravity if they are neither to fall in (because they move too slowly) nor to escape (because they move too fast). If the Galaxy had only the mass calculated by Kepler, then the high-speed outer objects should long ago have escaped the grip of the Milky Way. The fact that they have not done so means that our Galaxy must have more gravity than can be supplied by the luminous matter—in fact, a *lot* more gravity. The high speed of these outer objects tells us that the source of this extra gravity must extend outward from the center far beyond the Sun's orbit.

If the gravity were supplied by stars or by something else that gives off radiation, we should have spotted this additional outer material long ago. We are therefore forced to the reluctant conclusion that this matter is invisible and has, except for its gravitational pull, gone entirely undetected.

Studies of the motions of the most remote globular clusters and the small galaxies that orbit our own show that the total mass of the Galaxy is at least $2 \times 10^{12}\ M_{\text{Sun}}$, which is about twenty times greater than the amount of luminous matter. Moreover, the **dark matter** (as astronomers have come to call the invisible material) extends to a distance of at least 200,000 light-years from the center of the Galaxy. Observations indicate that this dark matter halo is almost but not quite spherical.

The obvious question is: what is the dark matter made of? Let's look at a list of "suspects" taken from our study of astronomy so far. Since this matter is invisible, it clearly cannot be in the form of ordinary stars. And it cannot be gas in any form (remember that there has to be a lot of it). If it were neutral hydrogen gas, its 21-cm wavelength spectral-line emission would have been detected as radio waves. If it were ionized hydrogen, it should be hot enough to emit visible radiation. If a lot of hydrogen atoms out there had combined into hydrogen molecules, these should produce dark features in the ultraviolet spectra of objects lying beyond the Galaxy, but such features have not been seen. Nor can the dark matter consist of interstellar dust, since in the required quantities, the dust would significantly obscure the light from distant galaxies.

What are our other possibilities? The dark matter cannot be a huge number of black holes (of stellar mass) or old neutron stars, since interstellar matter falling onto such objects would produce more X-rays than are observed. Also, recall that the formation of black holes and neutron stars is preceded by a substantial amount of mass loss, which scatters heavy elements into space to be incorporated into subsequent generations of stars. If the dark matter consisted of an enormous number of any of those objects, they would have blown off and recycled a lot of heavier elements over the history of the Galaxy. In that case, the young stars we observe in our Galaxy today would contain much greater abundances of heavy elements than they actually do.

Brown dwarfs and lone Jupiter-like planets have also been ruled out. First of all, there would have to be an awful lot of them to make up so much dark matter. But we have a more direct test of whether so many low-mass objects could actually be lurking out there. As we learned in **Introducing General Relativity**, the general theory of relativity predicts that the path traveled by light is changed when it passes near a concentration of mass. It turns out that when the two objects appear close enough together in the sky, the mass closer to us can bend the light from farther away. With just the right alignment, the image of the more distant object also becomes significantly brighter. By looking for the temporary brightening that occurs when a dark matter object in our own Galaxy moves across the path traveled by light from stars in the Magellanic Clouds, astronomers have now shown that the dark matter cannot be made up of a lot of small objects with masses between one-millionth and one-tenth the mass of the Sun.

What's left? One possibility is that the dark matter is composed of exotic subatomic particles of a type not yet detected on Earth. Very sophisticated (and difficult) experiments are now under way to look for such particles. Stay tuned to see whether anything like that turns up.

We should add that the problem of dark matter is by no means confined to the Milky Way. Observations show that dark matter must also be present in other galaxies (whose outer regions also orbit too fast "for their own good"—they also have flat rotation curves). As we will see, dark matter even exists in great clusters of galaxies whose members are now known to move around under the influence of far more gravity than can be accounted for by luminous matter alone.

Stop a moment and consider how astounding the conclusion we have reached really is. Perhaps as much as 95% of the mass in our Galaxy (and many other galaxies) is not only invisible, but we do not even know what it is made of. The stars and raw material we can observe may be merely the tip of the cosmic iceberg; underlying it all may be other matter, perhaps familiar, perhaps startlingly new. Understanding the nature of this dark matter is one of the great challenges of astronomy today; you will learn more about this in **The Challenge of Dark Matter**.

26.4 | The Center of the Galaxy

Learning Objectives

By the end of this section, you will be able to:

- Describe the radio and X-ray observations that indicate energetic phenomena are occurring at the galactic center
- Explain what has been revealed by high-resolution near-infrared imaging of the galactic center
- Discuss how these near-infrared images, when combined with Kepler's third law of motion, can be used to derive the mass of the central gravitating object

At the beginning of this chapter, we hinted that the core of our Galaxy contains a large concentration of mass. In fact, we now have evidence that the very center contains a black hole with a mass equivalent to 4.6 million Suns and that all this mass fits within a sphere that has less than the diameter of Mercury's orbit. Such monster black holes are called **supermassive black holes** by astronomers, to indicate that the mass they contain is far greater than that of the typical black hole created by the death of a single star. It is amazing that we have very convincing evidence that this black hole really does exist. After all, recall from the chapter on **Black Holes** that we cannot see a black hole directly because by definition it radiates no energy. And we cannot even see into the center of the Galaxy in visible light because of absorption by the interstellar dust that lies between us and the galactic center. Light from the central region of the Galaxy is dimmed by a factor of a trillion (10^{12}) by all this dust.

Fortunately, we are not so blind at other wavelengths. Infrared and radio radiation, which have long wavelengths compared to the sizes of the interstellar dust grains, flow unimpeded past the dust particles and so reach our telescopes with hardly any dimming. In fact, the very bright radio source in the nucleus of the Galaxy, now known as Sagittarius A* (pronounced "Sagittarius A-star" and abbreviated Sgr A*), was the first cosmic radio source astronomers discovered.

A Journey toward the Center

Let's take a voyage to the mysterious heart of our Galaxy and see what's there. **Figure 26.14** is a radio image of a region about 1500 light-years across, centered on Sagittarius A, a bright radio source that contains the smaller Sagittarius A*. Much of the radio emission comes from hot gas heated either by clusters of hot stars (the stars themselves do not produce radio emission and can't be seen in the image) or by supernova blast waves. Most of the hollow circles visible on the radio image are supernova remnants. The other main source of radio emission is from electrons moving at high speed in regions with strong magnetic fields. The bright thin arcs and "threads" on the figure show us where this type of emission is produced.

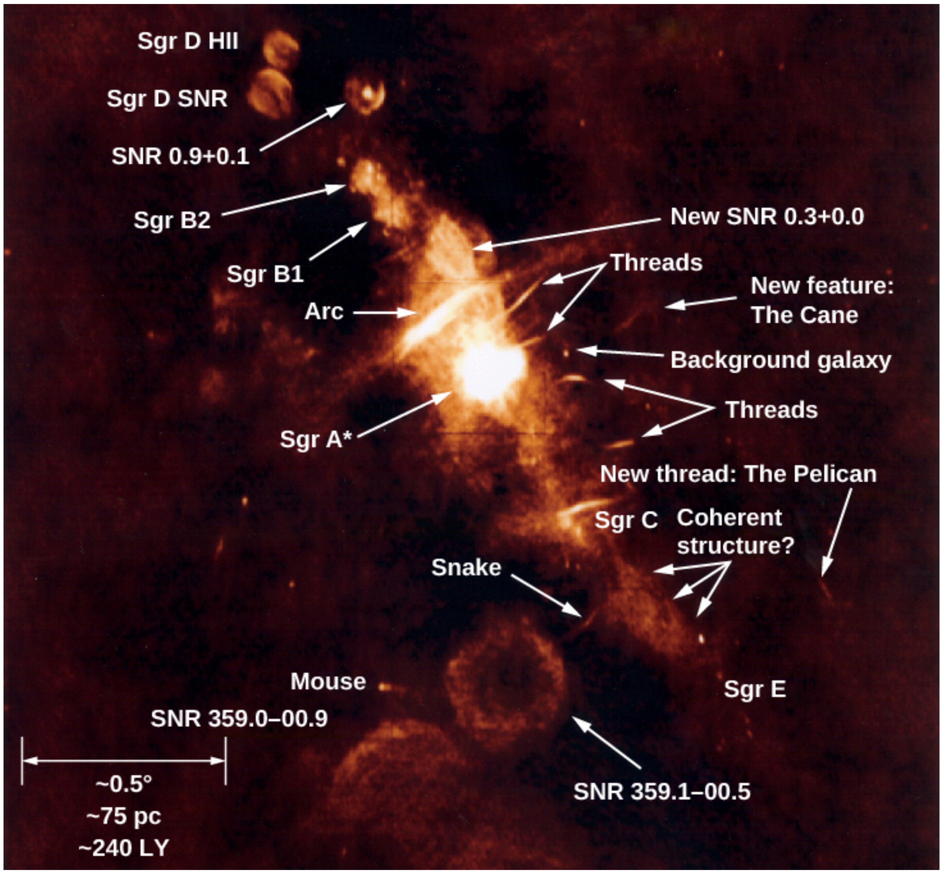


Figure 26.14 This radio map of the center of the Galaxy (at a wavelength of 90 centimeters) was constructed from data obtained with the Very Large Array (VLA) of radio telescopes in Socorro, New Mexico. Brighter regions are more intense in radio waves. The galactic center is inside the region labeled Sagittarius A. Sagittarius B1 and B2 are regions of active star formation. Many filaments or threadlike features are seen, as well as a number of shells (labeled SNR), which are supernova remnants. The scale bar at the bottom left is about 240 light-years long. Notice that radio astronomers also give fanciful animal names to some of the structures, much as visible-light nebulae are sometimes given the names of animals they resemble. (credit: modification of work by N. E. Kassim, D. S. Briggs, T. J. W. Lazio, T. N. LaRosa, and J. Imamura (NRL/RSD))

Now let's focus in on the central region using a more energetic form of electromagnetic radiation. **Figure 26.15** shows the X-ray emission from a smaller region 400 light-years wide and 900 light-years across centered in Sagittarius A*. Seen in this picture are hundreds of hot white dwarfs, neutron stars, and stellar black holes with accretion disks glowing with X-rays. The diffuse haze in the picture is emission from gas that lies among the stars and is at a temperature of 10 million K.

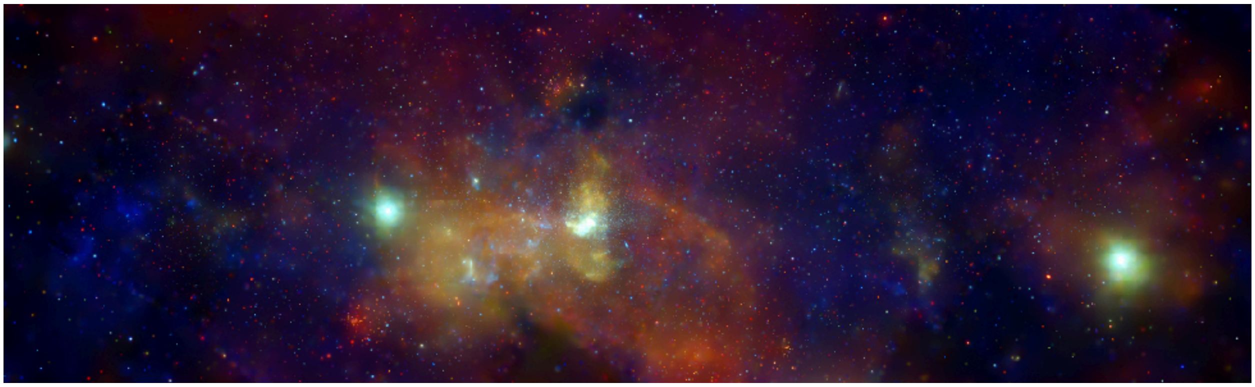


Figure 26.15 This artificial-color mosaic of 30 images taken with the Chandra X-ray satellite shows a region 400 × 900 light-years in extent and centered on Sagittarius A*, the bright white source in the center of the picture. The X-ray-emitting point sources are white dwarfs, neutron stars, and stellar black holes. The diffuse “haze” is emission from gas at a temperature of 10 million K. This hot gas is flowing away from the center out into the rest of the Galaxy. The colors indicate X-ray energy bands: red (low energy), green (medium energy), and blue (high energy). (credit: modification of work by NASA/CXC/ UMass/D. Wang et al.)

As we approach the center of the Galaxy, we find the supermassive black hole Sagittarius A*. There are also thousands of stars within a parsec of Sagittarius A*. Most of these are old, reddish main-sequence stars. But there are also about a hundred hot OB stars that must have formed within the last few million years. There is as yet no good explanation for how stars could have formed recently so close to a supermassive black hole. Perhaps they formed in a dense cluster of stars that was originally at a larger distance from the black hole and subsequently migrated closer.

There is currently no star formation at the galactic center, but there is lots of dust and molecular gas that is revolving around the black hole, along with some ionized gas streamers that are heated by the hot stars. **Figure 26.16** is a radio map that shows these gas streamers.

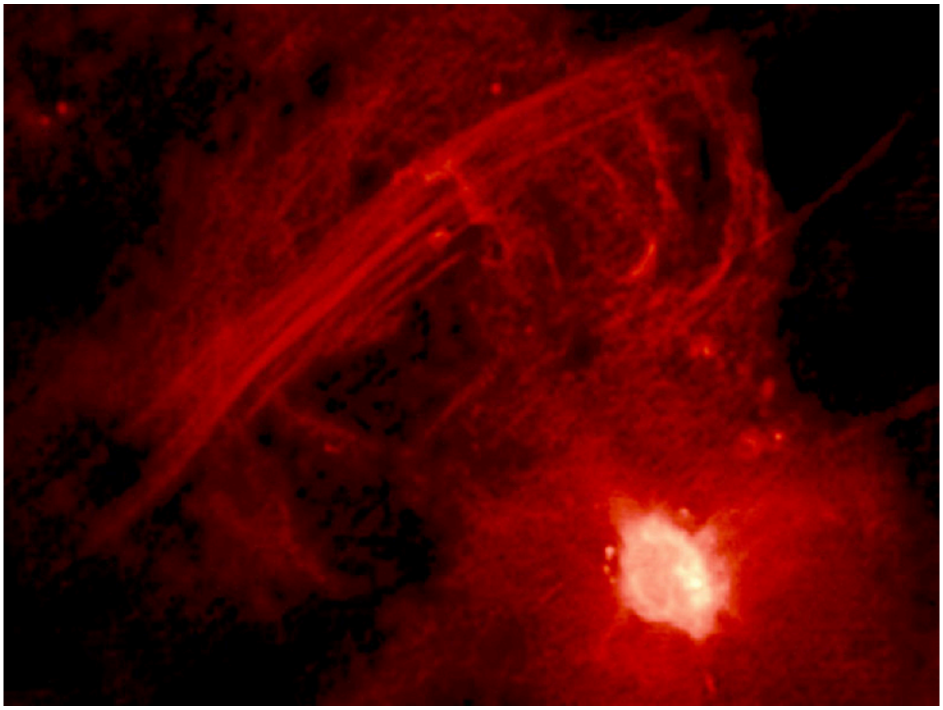


Figure 26.16 This image, taken with the Very Large Array of radio telescopes, shows the radio emission from hot, ionized gas in the center of the Milky Way. The lines slanting across the top of the image are gas streamers. Sagittarius A* is the bright spot in the lower right. (credit: modification of work by Farhad Zadeh et al. (Northwestern), VLA, NRAO)

Finding the Heart of the Galaxy

Just what is Sagittarius A*, which lies right at the center our Galaxy? To establish that there really is a black hole there, we must show that there is a very large amount of mass crammed into a very tiny volume. As we saw in **Black Holes**, proving that a black hole exists is a challenge because the black hole itself emits no radiation. What astronomers must do is prove that a black hole is the only possible explanation for our observations—that a small region contains far more mass than could be accounted for by a very dense cluster of stars or something else made of ordinary matter.

To put some numbers with this discussion, the radius of the event horizon of a *galactic black hole* with a mass of about 4 million M_{Sun} would be only about 17 times the size of the Sun—the equivalent of a single red giant star. The corresponding density within this region of space would be much higher than that of any star cluster or any other ordinary astronomical object. Therefore, we must measure both the diameter of Sagittarius A* and its mass. Both radio and infrared observations are required to give us the necessary evidence.

First, let's look at how the mass can be measured. If we zero in on the inner few light-days of the Galaxy with an infrared telescope equipped with adaptive optics, we see a region crowded with individual stars (**Figure 26.17**). These stars have now been observed for almost two decades, and astronomers have detected their rapid orbital motions around the very center of the Galaxy.

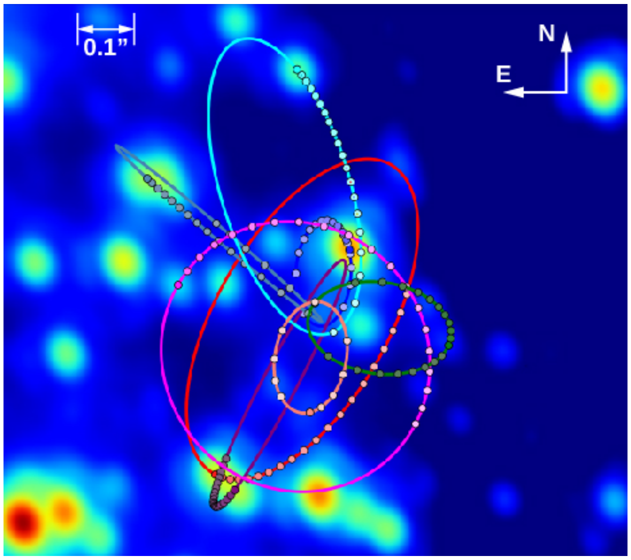


Figure 26.17 This image shows the inner 1 arcsecond, or 0.13 light-year, at the center of the Galaxy, as observed with the giant Keck Telescope. Tracks of the orbiting stars measured from 1995 to 2014 have been added to this "snapshot." The stars are moving around the center very fast, and their tracks are all consistent with a single massive "gravitator" that resides in the very center of this image. (credit: modification of work by Andrea Ghez, UCLA Galactic Center Group, W.M. Keck Observatory Laser Team)

Check out an **animated version (https://openstaxcollege.org/l/30anifiginfgal)** of **Figure 26.17**, showing the motion of the stars over the years.

If we combine observations of their periods and the size of their orbits with Kepler's third law, we can estimate the mass of the object that keeps them in their orbits. One of the stars has been observed for its full orbit of 15.6 years. Its closest approach takes it to a distance of only 124 AU or about 17 light-hours from the black hole. This orbit, when combined with observations of other stars close to the galactic center, indicates that a mass of 4.6 million M_{Sun} must be concentrated inside the orbit—that is, within 17 light-hours of the center of the Galaxy.

Even tighter limits on the size of the concentration of mass at the center of the Galaxy come from radio astronomy, which provided the first clue that a black hole might lie at the center of the Galaxy. As matter spirals inward toward the event horizon of a black hole, it is heated in a whirling *accretion disk* and produces radio radiation. (Such accretion disks were explained in **Black Holes**.) Measurements of the size of the accretion disk with the Very Long Baseline Array, which provides very high spatial resolution, show that the diameter of the radio source Sagittarius A* is no larger than about 0.3 AU, or about the size of Mercury's orbit. (In light units, that's only 2.5 light-*minutes*!)

The observations thus show that 4.6 million solar masses are crammed into a volume that has a diameter that is no larger than the orbit of Mercury. If this were anything other than a supermassive black hole—low-mass stars that emit very little light or neutron stars or a very large number of small black holes— calculations show that these objects would be so densely packed that they would collapse to a single black hole within a hundred thousand years. That is a very short time compared with the age of the Galaxy, which probably began forming more than 13 billion years ago. Since it seems very unlikely that we would have caught such a complex cluster of objects just before it collapsed, the evidence for a supermassive black hole at the center of the Galaxy is convincing indeed.

Finding the Source

Where did our galactic black hole come from? The origin of supermassive black holes in galaxies like ours is currently an active field of research. One possibility is that a large cloud of gas near the center of the Milky Way collapsed directly to form a black hole. Since we find large black holes at the centers of most other large galaxies (see **Supermassive Black Holes**)—even ones that are very young—this collapse probably would have taken place when the Milky Way was just beginning to take shape. The initial mass of this black hole might have been only a few tens of solar masses. Another way it could have started is that a massive star might have exploded to leave behind a seed black hole, or a dense cluster of stars might have collapsed into a black hole.

Once a black hole exists at the center of a galaxy, it can grow over the next several billion years by devouring nearby stars and gas clouds in the crowded central regions. It can also grow by merging with other black holes.

It appears that the monster black hole at the center of our Galaxy is not finished "eating." At the present time, we observe clouds of gas and dust falling into the galactic center at the rate of about 1 M_{Sun} per thousand years. Stars are also on the black hole's menu. The density of stars near the galactic center is high enough that we would expect a star to pass near the black hole and be swallowed by it every ten thousand years or so. As this happens, some of the energy of infall is released as radiation. As a result, the center of the Galaxy might flare up and even briefly outshine all the stars in the Milky Way. Other objects might also venture too close to the black hole and be pulled in. How great a flare we observe would depend on the mass of the object falling in.

In 2013, the Chandra X-ray satellite detected a flare from the center of our Galaxy that was 400 times brighter than the usual output from Sagittarius A*. A year later, a second flare, only half as bright, was also detected. This is much less energy than swallowing a whole star would produce. There are two theories to account for the flares. First, an asteroid might have ventured too close to the black hole and been heated to a very high temperature before being swallowed up. Alternatively, the flares might have involved interactions of the magnetic fields near the galactic center in a process similar to the one described for solar flares (see **The Structure and Composition of the Sun**). Astronomers continue to monitor the galactic center area for flares or other activity. Although the monster in the center of the Galaxy is not close enough to us to represent any danger, we still want to keep our eyes on it.

Andrea Ghez

A lover of puzzles, Andrea Ghez has been pursuing one of the greatest mysteries in astronomy: what strange entity lurks within the center of our Milky Way Galaxy?

Figure 26.18 Research by Ghez and her team has helped shape our understanding of supermassive black holes. (credit: modification of work by John D. and Catherine T. MacArthur Foundation)

As a child living in Chicago during the late 1960s, Andrea Ghez (**Figure 26.18**) was fascinated by the Apollo Moon landings. But she was also drawn to ballet and to solving all sorts of puzzles. By high school, she had lost the ballet bug in favor of competing in field hockey, playing the flute, and digging deeper into academics. Her undergraduate years at MIT were punctuated by a number of changes in her major—from mathematics to chemistry, mechanical engineering, aerospace engineering, and finally physics—where she felt her options were most open. As a physics major, she became involved in astronomical research under the guidance of one of her instructors. Once she got to do some actual observing at Kitt Peak National Observatory in Arizona, and later at Cerro Tololo Inter-American Observatory in Chile, Ghez had found her calling.

Pursuing her graduate studies at Caltech, she stuck with physics but oriented her efforts toward observational astrophysics, an area where Caltech had access to cutting-edge facilities. Though initially attracted to studying the black holes that were suspected of dwelling inside most massive galaxies, Ghez ended up spending most of her graduate study and later postdoctoral research at the University of Arizona studying stars in formation. By taking very high-resolution (detailed) imaging of regions where new stars are born, she discovered that most stars form as members of binary systems. As technologies advanced, she was able to track the orbits danced by these stellar pairings and thereby could ascertain their respective masses.

Now an astronomy professor at UCLA, Ghez has since used similar high-resolution imaging techniques to study the orbits of stars in the innermost core of the Milky Way. These orbits take years to delineate, so Ghez and her science team have logged more than 20 years of taking super-resolution infrared images with the giant Keck telescopes in Hawaii. Based on the resulting stellar orbits, the UCLA Galactic Center Group has settled (as we saw) on a gravitational solution that requires the presence of a supermassive black hole with a mass equivalent to 4.6 million Suns—all nestled within a space smaller than that occupied by our solar system. Ghez's achievements have been recognized with one of the "genius" awards given by the MacArthur Foundation. More recently, her team discovered glowing clouds of warm ionized gas that co-orbit with the stars but may be more vulnerable to the disruptive effects of the central black hole. By monitoring these clouds, the team hopes to better understand the evolution of supermassive black holes and their immediate environs. They also hope to test Einstein's theory of general relativity by carefully scrutinizing the orbits of stars that careen closest to the intensely gravitating black hole.

Besides her pioneering work as an astronomer, Ghez competes as a master swimmer, enjoys family life as a mother of two children, and actively encourages other women to pursue scientific careers.

26.5 | Stellar Populations in the Galaxy

Learning Objectives

By the end of this section, you will be able to:

- Distinguish between population I and population II stars according to their locations, motions, heavy-element abundances, and ages
- Explain why the oldest stars in the Galaxy are poor in elements heavier than hydrogen and helium, while stars like the Sun and even younger stars are typically richer in these heavy elements

In the first section of his chapter, we described the thin disk, thick disk, and stellar halo. Look back at **Table 26.1** and note some of the patterns. Young stars lie in the thin disk, are rich in metals, and orbit the Galaxy's center at high speed. The stars in the halo are old, have low abundances of elements heavier than hydrogen and helium, and have highly elliptical orbits randomly oriented in direction (see **Figure 26.19**). Halo stars can plunge through the disk and central bulge, but they spend most of their time far above or below the plane of the Galaxy. The stars in the thick disk are intermediate between these two extremes. Let's first see why age and heavier-element abundance are correlated and then see what these correlations tell us about the origin of our Galaxy.

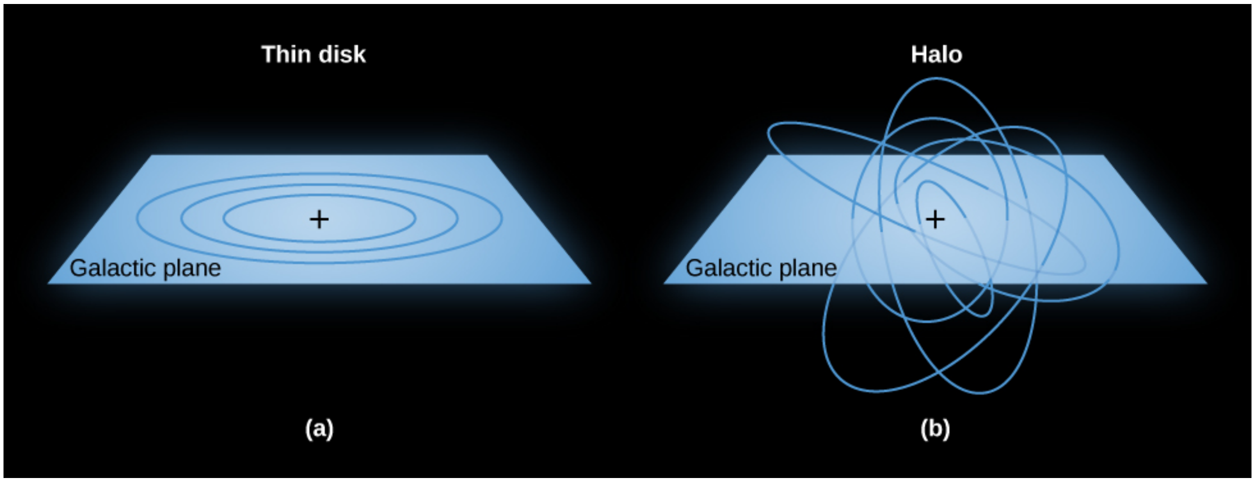


Figure 26.19 (a) In this image, you see stars in the thin disk of our Galaxy in nearly circular orbits. (b) In this image, you see the motion of stars in the Galaxy's halo in randomly oriented and elliptical orbits.

Two Kinds of Stars

The discovery that there are two different kinds of stars was first made by Walter Baade during World War II. As a German national, Baade was not allowed to do war research as many other U.S.-based scientists were doing, so he was able to make regular use of the Mount Wilson telescopes in southern California. His observations were aided by the darker skies that resulted from the wartime blackout of Los Angeles.

Among the things a large telescope and dark skies enabled Baade to examine carefully were *other* galaxies—neighbors of our Milky Way Galaxy. We will discuss other galaxies in the next chapter (**Galaxies**), but for now we will just mention that the nearest Galaxy that resembles our own (with a similar disk and spiral structure) is often called the Andromeda galaxy, after the constellation in which we find it.

Baade was impressed by the similarity of the mainly reddish stars in the Andromeda galaxy's nuclear bulge to those in our Galaxy's globular clusters and the halo. He also noted the difference in color between all these and the bluer stars found in the spiral arms near the Sun (**Figure 26.20**). On this basis, he called the bright blue stars in the spiral arms **population I** and all the stars in the halo and globular clusters **population II**.

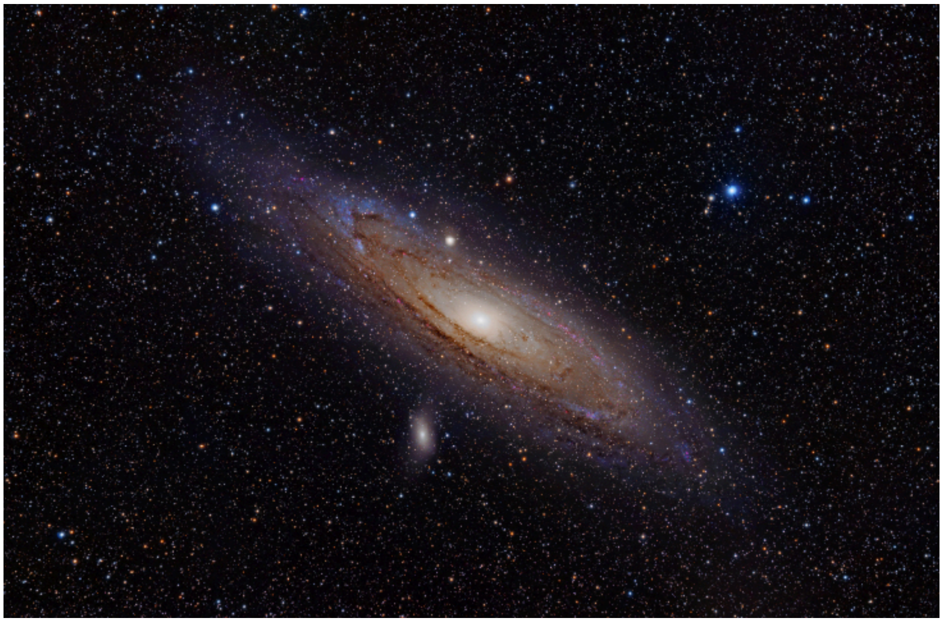


Figure 26.20 This neighboring spiral looks similar to our own Galaxy in that it is a disk galaxy with a central bulge. Note the bulge of older, yellowish stars in the center, the bluer and younger stars in the outer regions, and the dust in the disk that blocks some of the light from the bulge. (credit: Adam Evans)

We now know that the populations differ not only in their locations in the Galaxy, but also in their chemical composition, age, and orbital motions around the center of the Galaxy. Population I stars are found only in the disk and follow nearly circular orbits around the galactic center. Examples are bright supergiant stars, main-sequence stars of high luminosity (spectral classes O and B), which are concentrated in the spiral arms, and members of young open star clusters. Interstellar matter and molecular clouds are found in the same places as population I stars.

Population II stars show no correlation with the location of the spiral arms. These objects are found throughout the Galaxy. Some are in the disk, but many others follow eccentric elliptical orbits that carry them high above the galactic disk into the halo. Examples include stars surrounded by planetary nebulae and RR Lyrae variable stars. The stars in globular clusters, found almost entirely in the Galaxy's halo, are also classified as population II.

Today, we know much more about stellar evolution than astronomers did in the 1940s, and we can determine the ages of stars. Population I includes stars with a wide range of ages. While some are as old as 10 billion years, others are still forming today. For example, the Sun, which is about 5 billion years old, is a population I star. But so are the massive young stars in the Orion Nebula that have formed in the last few million years. Population II, on the other hand, consists entirely of old stars that formed very early in the history of the Galaxy; typical ages are 11 to 13 billion years.

We also now have good determinations of the compositions of stars. These are based on analyses of the stars' detailed spectra. Nearly all stars appear to be composed mostly of hydrogen and helium, but their abundances of the heavier elements differ. In the Sun and other population I stars, the heavy elements (those heavier than hydrogen and helium) account for 1–4% of the total stellar mass. Population II stars in the outer galactic halo and in globular clusters have much lower abundances of the heavy elements—often less than one-hundredth the concentrations found in the Sun and in rare cases even lower. The oldest population II star discovered to date has less than one ten-millionth as much iron as the Sun, for example.

As we discussed in earlier chapters, heavy elements are created deep within the interiors of stars. They are added to the Galaxy's reserves of raw material when stars die, and their material is recycled into new generations of stars. Thus, as time goes on, stars are born with larger and larger supplies of heavy elements. Population II stars formed when the abundance of elements heavier than hydrogen and helium was low. Population I stars formed later, after mass lost by dying members of the first generations of stars had seeded the interstellar medium with elements heavier than hydrogen and helium. Some are still forming now, when further generations have added to the supply of heavier elements available to new stars.

The Real World

With rare exceptions, we should never trust any theory that divides the world into just two categories. While they can provide a starting point for hypotheses and experiments, they are often oversimplifications that need refinement a research continue. The idea of two populations helped organize our initial thoughts about the Galaxy, but we now know it cannot explain everything we observe. Even the different structures of the Galaxy—disk, halo, central bulge—are not so cleanly separated in terms of their locations, ages, and the heavy element content of the stars within them.

The exact definition of the Galaxy's disk depends on what objects we use to define it, and, as we saw earlier, it has no sharp boundary. The hottest young stars and their associated gas and dust clouds are mostly in a region about 200 light-years thick. Older stars define a thicker disk that is about 2000 light-years thick. Halo stars spend most of their time high above or below the disk but pass through it on their highly elliptical orbits and so are sometimes found relatively near the Sun.

The highest density of stars is found in the central bulge, that bar-shaped inner region of the Galaxy. There are a few hot, young stars in the bulge, but most of the bulge stars are more than 10 billion years old. Yet unlike the halo stars of similar age, the abundance of heavy elements in the bulge stars is about the same as in the Sun. Why would that be?

Astronomers think that star formation in the crowded nuclear bulge occurred very rapidly just after the Milky Way Galaxy formed. After a few million years, the first generation of massive and short-lived stars then expelled heavy elements in supernova explosions and thereby enriched subsequent generations of stars. Thus, even stars that formed in the bulge more than 10 billion years ago started with a good supply of heavy elements.

Exactly the opposite occurred in the Small Magellanic Cloud, a small galaxy near the Milky Way, visible from Earth's Southern Hemisphere. Even the youngest stars in this galaxy are deficient in heavy elements. We think this is because the little galaxy is not especially crowded, and star formation has occurred quite slowly. As a result there have been, so far, relatively few supernova explosions. Smaller galaxies also have more trouble holding onto the gas expelled by supernova explosions in order to recycle it. Low-mass galaxies exert only a modest gravitational force, and the high-speed gas ejected by supernovae can easily escape from them.

Which elements a star is endowed with thus depends not only on when the star formed in the history of its galaxy, but also on how many stars in its part of the galaxy had already completed their lives by the time the star is ready to form.

26.6 | The Formation of the Galaxy

Learning Objectives

By the end of this section, you will be able to:

- Describe the roles played by the collapse of a single cloud and mergers with other galaxies in building the Milky Way Galaxy we see today
- Provide examples of globular clusters and satellite galaxies affected by the Milky Way's strong gravity.

Information about stellar populations holds vital clues to how our Galaxy was built up over time. The flattened disk shape of the Galaxy suggests that it formed through a process similar to the one that leads to the formation of a protostar (see **Star Formation**). Building on this idea, astronomers first developed models that assumed the Galaxy formed from a single rotating cloud. But, as we shall see, this turns out to be only part of the story.

The Protogalactic Cloud and the Monolithic Collapse Model

Because the oldest stars—those in the halo and in globular clusters—are distributed in a sphere centered on the nucleus of the Galaxy, it makes sense to assume that the *protogalactic* cloud that gave birth to our Galaxy was roughly spherical. The oldest stars in the halo have ages of 12 to 13 billion years, so we estimate that the formation of the Galaxy began about that long ago. (See the chapter on **Big Bang Cosmology** for other evidence that galaxies in general began forming a little more than 13 billion years ago.) Then, just as in the case of star formation, the protogalactic cloud collapsed and formed a thin rotating disk. Stars born before the cloud collapsed did not participate in the collapse, but have continued to orbit in the halo to the present day (**Figure 26.21**).

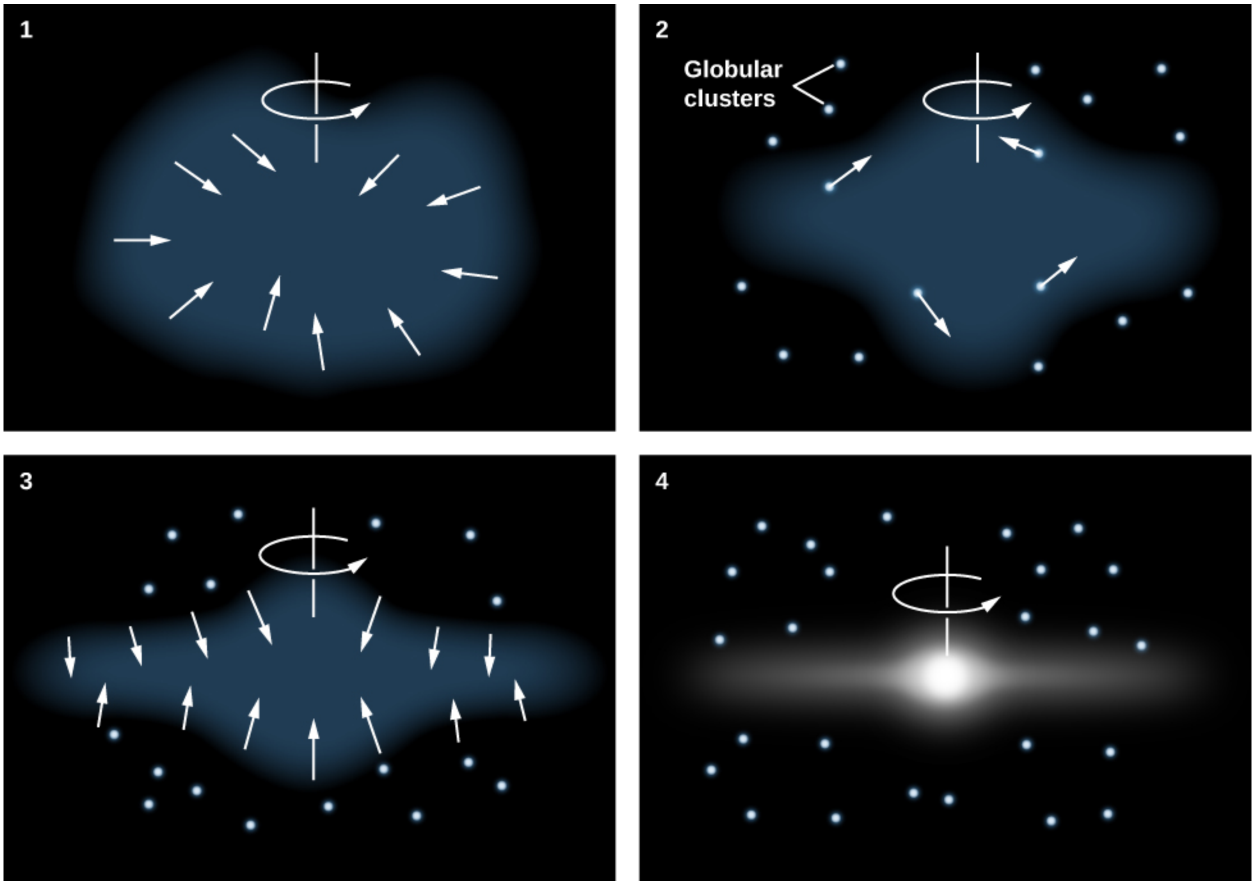


Figure 26.21 According to this model, the Milky Way Galaxy initially formed from a rotating cloud of gas that collapsed due to gravity. Halo stars and globular clusters either formed prior to the collapse or were formed elsewhere. Stars in the disk formed later, when the gas from which they were made was already "contaminated" with heavy elements produced in earlier generations of stars.

Gravitational forces caused the gas in the thin disk to fragment into clouds or clumps with masses like those of star clusters. These individual clouds then fragmented further to form stars. Since the oldest stars in the disk are nearly as old as the youngest stars in the halo, the collapse must have been rapid (astronomically speaking), requiring perhaps no more than a few hundred million years.

Collision Victims and the Multiple Merger Model

In past decades, astronomers have learned that the evolution of the Galaxy has not been quite as peaceful as this monolithic collapse model suggests. In 1994, astronomers discovered a small new galaxy in the direction of the constellation of Sagittarius. The Sagittarius dwarf galaxy is currently about 70,000 light-years away from Earth and 50,000 light-years from the center of the Galaxy. It is the closest galaxy known (**Figure 26.22**). It is very elongated, and its shape indicates that it is being torn apart by our Galaxy's gravitational tides—just as Comet Shoemaker-Levy 9 was torn apart when it passed too close to Jupiter in 1992.

The Sagittarius galaxy is much smaller than the Milky Way and is about 10,000 times less massive than our Galaxy. All of the stars in the Sagittarius dwarf galaxy seem destined to end up in the bulge and halo of the Milky Way. But don't sound the funeral bells for the little galaxy quite yet; the ingestion of the Sagittarius dwarf will take another 100 million years or so, and the stars themselves will survive.

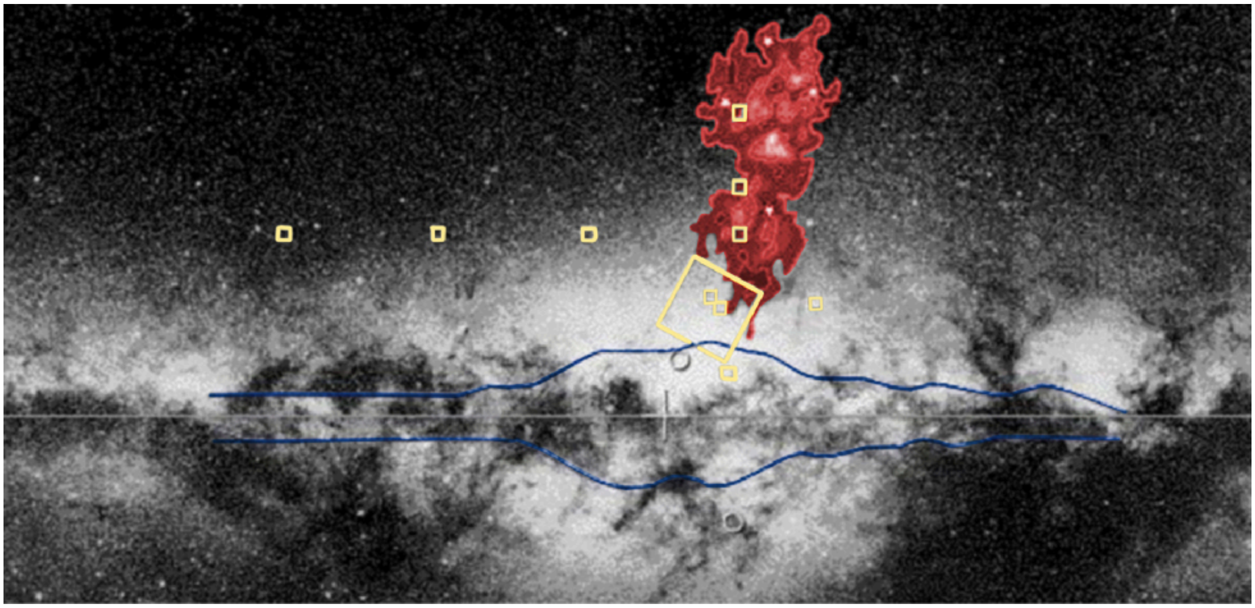


Figure 26.22 In 1994, British astronomers discovered a galaxy in the constellation of Sagittarius, located only about 50,000 light-years from the center of the Milky Way and falling into our Galaxy. This image covers a region approximately 70° × 50° and combines a black-and-white view of the disk of our Galaxy with a red contour map showing the brightness of the dwarf galaxy. The dwarf galaxy lies on the other side of the galactic center from us. The white stars in the red region mark the locations of several globular clusters contained within the Sagittarius dwarf galaxy. The cross marks the galactic center. The horizontal line corresponds to the galactic plane. The blue outline on either side of the galactic plane corresponds to the infrared image in **Figure 26.7**. The boxes mark regions where detailed studies of individual stars led to the discovery of this galaxy. (credit: modification of work by R. Ibata (UBC), R. Wyse (JHU), R. Sword (IoA))

Since that discovery, evidence has been found for many more close encounters between our Galaxy and other neighbor galaxies. When a small galaxy ventures too close, the force of gravity exerted by our Galaxy tugs harder on the near side than on the far side. The net effect is that the stars that originally belonged to the small galaxy are spread out into a long stream that orbits through the halo of the Milky Way (**Figure 26.23**).

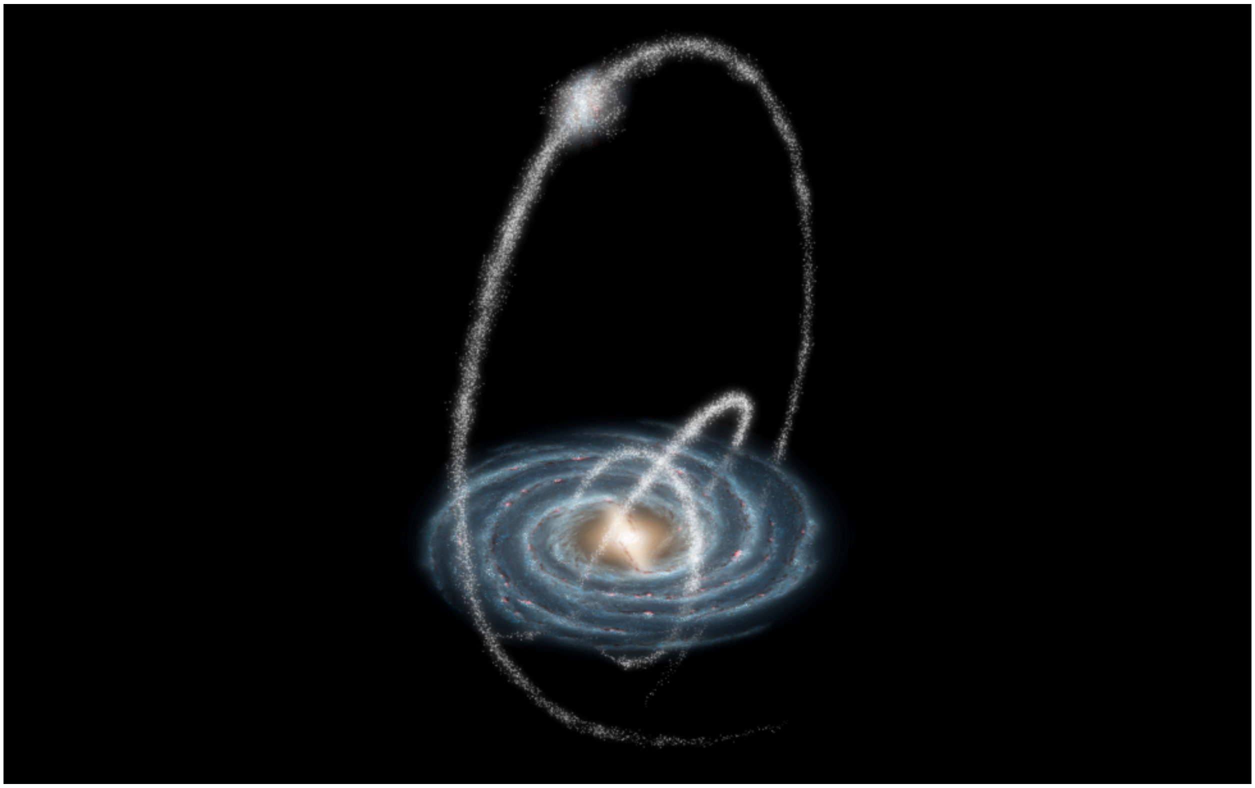


Figure 26.23 When a small galaxy is swallowed by the Milky Way, its member stars are stripped away and form streams of stars in the galactic halo. This image is based on calculations of what some of these tidal streams might look like if the Milky Way swallowed 50 dwarf galaxies over the past 10 billion years. (credit: modification of work by NASA/JPL-Caltech/R. Hurt (SSC/Caltech))

Such a tidal stream can maintain its identity for billions of years. To date, astronomers have now identified streams originating from 12 small galaxies that ventured too close to the much larger Milky Way. Six more streams are associated with globular clusters. It has been suggested that large globular clusters, like Omega Centauri, are actually dense nuclei of cannibalized dwarf galaxies. The globular cluster M54 is now thought to be the nucleus of the Sagittarius dwarf we discussed earlier, which is currently merging with the Milky Way (**Figure 26.24**). The stars in the outer regions of such galaxies are stripped off by the gravitational pull of the Milky Way, but the central dense regions may survive.

Figure 26.24 This beautiful Hubble Space Telescope image shows the globular cluster that is now believed to be the nucleus of the Sagittarius Dwarf Galaxy. (credit: ESA/Hubble & NASA)

Calculations indicate that the Galaxy's thick disk may be a product of one or more such collisions with other galaxies. Accretion of a satellite galaxy would stir up the orbits of the stars and gas clouds originally in the thin disk and cause them to move higher above and below the mid-plane of the Galaxy. Meanwhile, the Galaxy's stars would add to the fluffed-up mix. If such a collision happened about 10 billion years ago, then any gas in the two galaxies that had not yet formed into stars would have had plenty of time to settle back down into the thin disk. The gas could then have begun forming subsequent generations of population I stars. This timing is also consistent with the typical ages of stars in the thick disk.

The Milky Way has more collisions in store. An example is the Canis Major dwarf galaxy, which has a mass of about 1% of the mass of the Milky Way. Already long tidal tails have been stripped from this galaxy, which have wrapped themselves around the Milky Way three times. Several of the globular clusters found in the Milky Way may also have come from the Canis Major dwarf, which is expected to merge gradually with the Milky Way over about the next billion years.

In about 3 billion years, the Milky Way itself will be swallowed up, since it and the Andromeda galaxy are on a collision course. Our computer models show that after a complex interaction, the two will merge to form a larger, more rounded galaxy (**Figure 26.25**).

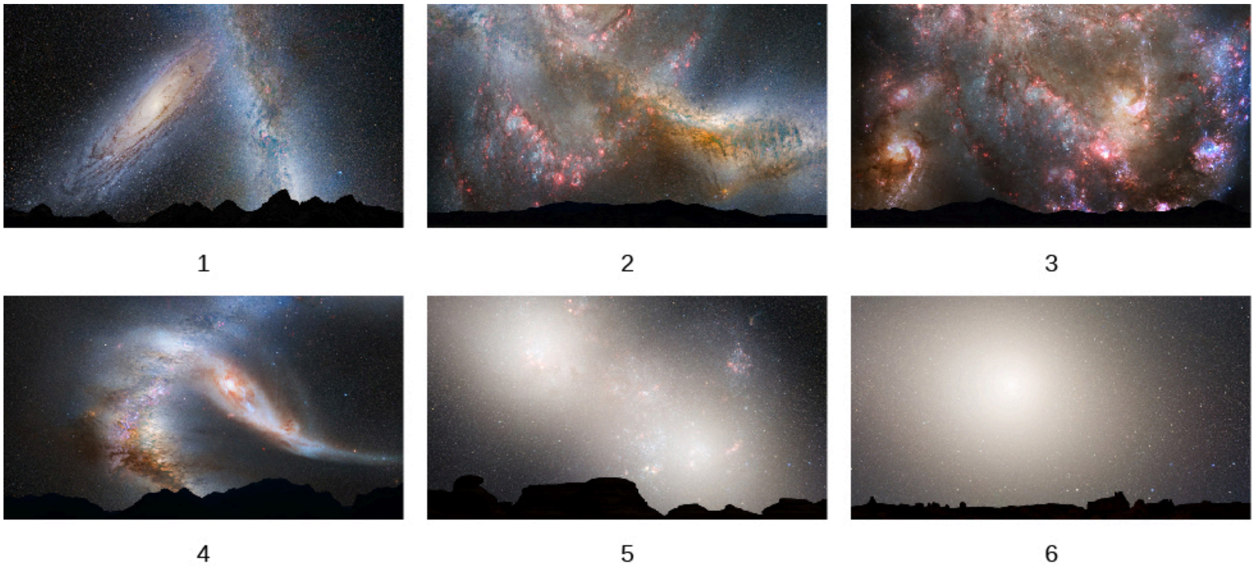


Figure 26.25 In about 3 billion years, the Milky Way Galaxy and Andromeda Galaxy will begin a long process of colliding, separating, and then coming back together to form an elliptical galaxy. The whole interaction will take 3 to 4 billion years. These computer-simulated images show the following sequence: (1) In 3.75 billion years, Andromeda has approached the Milky Way. (2) New star formation fills the sky 3.85 billion years from now. (3) Star formation continues at 3.9 billion years. (4) The galaxy shapes change as they interact, with Andromeda being stretched and our Galaxy becoming warped, about 4 billion years from now. (5) In 5.1 billion years, the cores of the two galaxies are bright lobes. (6) In 7 billion years, the merged galaxies form a huge elliptical galaxy whose brightness fills the night sky. This artist's illustrations show events from a vantage point 25,000 light-years from the center of the Milky Way. However, we should mention that the Sun may not be at that distance throughout the sequence of events, as the collision readjusts the orbits of many stars within each galaxy. (credit: NASA; ESA; Z. Levay, R. van der Marel, STScl; T. Hallas, and A. Mellinger)

We are thus coming to realize that "environmental influences" (and not just a galaxy's original characteristics) play an important role in determining the properties and development of our Galaxy. In future chapters we will see that collisions and mergers are a major factor in the evolution of many other galaxies as well.

For Further Exploration

Websites

International Dark Sky Sanctuaries: **http://darksky.org/idsp/sanctuaries/ (http://darksky.org/idsp/sanctuaries/)** . A listing of dark-sky sanctuaries, parks, and reserves.

Multiwavelength Milky Way: **http://mwmw.gsfc.nasa.gov/mmw_sci.html (http://mwmw.gsfc.nasa.gov/mmw_sci.html)** . This NASA site shows the plane of our Galaxy in a variety of wavelength bands, and includes background material and other resources.

Shapley-Curtis Debate in 1920: **http://apod.nasa.gov/diamond_jubilee/debate_1920.html (http://apod.nasa.gov/diamond_jubilee/debate_1920.html)** . In 1920, astronomers Harlow Shapley and Heber Curtis engaged in a historic debate about how large our Galaxy was and whether other galaxies existed. Here you can find historical and educational material about the debate.

UCLA Galactic Center Group: **http://www.galacticcenter.astro.ucla.edu/ (http://www.galacticcenter.astro.ucla.edu/)** . Learn more about the work of Andrea Ghez and colleagues on the central region of the Milky Way Galaxy.

Videos

Crash of the Titans: **http://www.spacetelescope.org/videos/hubblecast55a/ (http://www.spacetelescope.org/videos/hubblecast55a/)** . This Hubblecast from 2012 features Jay Anderson and Roeland van der Marel explaining how Andromeda will collide with the Milky Way in the distant future (5:07).

Diner at the Center of the Galaxy: **https://www.youtube.com/watch?v=UP7ig8Gxftw (https://www.youtube.com/watch?v=UP7ig8Gxftw)** . A short discussion from NASA ScienceCast of NuSTAR observations of flares from our Galaxy’s central black hole (3:23).

Hunt for a Supermassive Black Hole: **https://www.ted.com/talks/andrea_ghez_the_hunt_for_a_supermassive_black_hole (https://www.ted.com/talks/andrea_ghez_the_hunt_for_a_supermassive_black_hole)** . 2009 TED talk by Andrea Ghez on searching for supermassive black holes, particularly the one at the center of the Milky Way (16:19).

Journey to the Galactic Center: **https://www.youtube.com/watch?v=36xZsgZ0oSo (https://www.youtube.com/watch?v=36xZsgZ0oSo)** . A brief silent trip into the cluster of stars near the galactic center showing their motions around the center (3:00).

CHAPTER 26 REVIEW

KEY TERMS

central bulge (or nuclear bulge) the central (round) part of the Milky Way or a similar galaxy

dark matter nonluminous mass, whose presence can be inferred only because of its gravitational influence on luminous matter; the composition of the dark matter is not known

dark matter halo the mass in the Milky Way that extends well beyond the boundary of the luminous stars to a distance of at least 200,000 light-years from the center of the Galaxy; although we deduce its existence from its gravity, the composition of this matter remains a mystery

differential galactic rotation the idea that different parts of the Galaxy turn at different rates, since the parts of the Galaxy follow Kepler's third law: more distant objects take longer to complete one full orbit around the center of the Galaxy

halo the outermost extent of our Galaxy (or another galaxy), containing a sparse distribution of stars and globular clusters in a more or less spherical distribution

Milky Way Galaxy the band of light encircling the sky, which is due to the many stars and diffuse nebulae lying near the plane of the Milky Way Galaxy

population I star a star containing heavy elements; typically young and found in the disk

population II star a star with very low abundance of heavy elements; found throughout the Galaxy

spiral arm a spiral-shaped region, characterized by relatively dense interstellar material and young stars, that is observed in the disks of spiral galaxies

supermassive black hole the object in the center of most large galaxies that is so massive and compact that light cannot escape from it; the Milky Way's supermassive black hole contains 4.6 millions of Suns' worth of mass

KEY EQUATIONS

Orbital Velocity Law $M_r = \frac{av^2}{G}$

SUMMARY

26.1 The Architecture of the Galaxy

- The Milky Way Galaxy consists of a thin disk containing dust, gas, and young and old stars; a spherical halo containing populations of very old stars, including RR Lyrae variable stars and globular star clusters; a thick, more diffuse disk with stars that have properties intermediate between those in the thin disk and the halo; a peanut-shaped nuclear bulge of mostly old stars around the center; and a supermassive black hole at the very center.
- The Sun is located roughly halfway out of the Milky Way, about 26,000 light-years from the center.

26.2 Spiral Structure

- The gaseous distribution in the Galaxy's disk has two main spiral arms that emerge from the ends of the central bar, along with several fainter arms and short spurs; the Sun is located in one of those spurs.
- Measurements show that the Galaxy does not rotate as a solid body, but instead its stars and gas follow differential rotation, such that the material closer to the galactic center completes its orbit more quickly.
- Observations show that galaxies like the Milky Way take several billion years after they began to form to develop spiral structure.

26.3 The Mass of the Galaxy

- The Sun revolves completely around the galactic center in about 225 million years (a galactic year).

- The mass of the Galaxy can be determined by measuring the orbital velocities of stars and interstellar matter.
- The total mass of the Galaxy is about 2×10^{12} $M_{\text{Sun.}}$
- As much as 95% of this mass consists of dark matter that emits no electromagnetic radiation and can be detected only because of the gravitational force it exerts on visible stars and interstellar matter.
- This dark matter is located mostly in the Galaxy's halo; its nature is not well understood at present.

26.4 The Center of the Galaxy

- A supermassive black hole is located at the center of the Galaxy.
- Measurements of the velocities of stars located within a few light-days of the center show that the mass inside their orbits around the center is about 4.6 million M_{Sun}.
- Radio observations show that this mass is concentrated in a volume with a diameter similar to that of Mercury's orbit.
- The density of this matter concentration exceeds that of the densest known star clusters by a factor of nearly a million.
- The only known object with such a high density and total mass is a black hole.

26.5 Stellar Populations in the Galaxy

- We can roughly divide the stars in the Galaxy into two categories.
- Old stars with few heavy elements are referred to as population II stars and are found in the halo and in globular clusters.Population I stars contain more heavy elements than globular cluster and halo stars, are typically younger and found in the disk, and are especially concentrated in the spiral arms.
- The Sun is a member of population I.
- Population I stars formed after previous generations of stars had produced heavy elements and ejected them into the interstellar medium.
- The bulge stars, most of which are more than 10 billion years old, have unusually high amounts of heavy elements, presumably because there were many massive first-generation stars in this dense region, and these quickly seeded the next generations of stars with heavier elements.

26.6 The Formation of the Galaxy

- The Galaxy began forming a little more than 13 billion years ago.
- Models suggest that the stars in the halo and globular clusters formed first, while the Galaxy was spherical.
- The gas, somewhat enriched in heavy elements by the first generation of stars, then collapsed from a spherical distribution to a rotating disk-shaped distribution.
- Stars are still forming today from the gas and dust that remain in the disk. Star formation occurs most rapidly in the spiral arms, where the density of interstellar matter is highest.
- The Galaxy captured (and still is capturing) additional stars and globular clusters from small galaxies that ventured too close to the Milky Way.
- In 3 to 4 billion years, the Galaxy will begin to collide with the Andromeda galaxy, and after about 7 billion years, the two galaxies will merge to form a giant elliptical galaxy.

CONCEPTUAL QUESTIONS

26.1 The Architecture of the Galaxy

1. Explain why we see the Milky Way as a faint band of light stretching across the sky.

2. Briefly describe the main parts of our Galaxy.

3. Suppose the Milky Way was a band of light extending only halfway around the sky (that is, in a semicircle). What, then, would you conclude about the Sun's location in the Galaxy? Give your reasoning.

26.2 Spiral Structure

4. Suppose three stars lie in the disk of the Galaxy at distances of 20,000 light-years, 25,000 light-years, and 30,000 light-years from the galactic center, and suppose that right now all three are lined up in such a way that it is possible to draw a straight line through them and on to the center of the Galaxy. How will the relative positions of these three stars change with time? Assume that their orbits are all circular and lie in the plane of the disk.

26.4 The Center of the Galaxy

5. Describe the evidence indicating that a black hole may be at the center of our Galaxy.

6. Suppose somebody proposed that rather than invoking dark matter to explain the increased orbital velocities of stars beyond the Sun's orbit, the problem could be solved by assuming that the Milky Way's central black hole was much more massive. Does simply increasing the assumed mass of the Milky Way's central supermassive black hole correctly resolve the issue of unexpectedly high orbital velocities in the Galaxy? Why or why not?

26.5 Stellar Populations in the Galaxy

7. Explain where in a spiral galaxy you would expect to find globular clusters, molecular clouds, and atomic hydrogen.

8. Describe several characteristics that distinguish population I stars from population II stars.

9. Explain why the abundances of heavy elements in stars correlate with their positions in the Galaxy.

10. The globular clusters revolve around the Galaxy in highly elliptical orbits. Where would you expect the clusters to spend most of their time? (Think of Kepler's laws.) At any given time, would you expect most globular clusters to be moving at high or low speeds with respect to the center of the Galaxy? Why?

11. Shapley used the positions of globular clusters to determine the location of the galactic center. Could he have used open clusters? Why or why not?

12. Consider the following five kinds of objects: open cluster, giant molecular cloud, globular cluster, group of O and B stars, and planetary nebulae.

A. Which occur only in spiral arms?
B. Which occur only in the parts of the Galaxy other than the spiral arms?
C. Which are thought to be very young?
D. Which are thought to be very old?
E. Which have the hottest stars?

13. The dwarf galaxy in Sagittarius is the one closest to the Milky Way, yet it was discovered only in 1994. Can you think of a reason it was not discovered earlier? (Hint: Think about what else is in its constellation.)

14. Why does star formation occur primarily in the disk of the Galaxy?

15. Where in the Galaxy would you expect to find Type II supernovae, which are the explosions of massive stars that go through their lives very quickly? Where would you expect to find Type I supernovae, which involve the explosions of white dwarfs?

26.6 The Formation of the Galaxy

16. What will be the long-term future of our Galaxy?

17. Suppose that stars evolved without losing mass—that once matter was incorporated into a star, it remained there forever. How would the appearance of the Galaxy be different from what it is now? Would there be population I and population II stars? What other differences would there be?

PROBLEMS

26.3 The Mass of the Galaxy

18. Assume that the Sun orbits the center of the Galaxy at a speed of 220 km/s and a distance of 26,000 light-years from the center.

A. Calculate the circumference of the Sun's orbit, assuming it to be approximately circular. (Remember that the circumference of a circle is given by $2\pi R$, where R is the radius of the circle. Be sure to use consistent units. The conversion from light-years to km/s can be found in an online calculator or appendix, or you can calculate it for yourself: the speed of light is 300,000 km/s, and you can determine the number of seconds in a year.)
B. Calculate the Sun's period, the "galactic year." Again, be careful with the units. Does it agree with the number we gave above?

19. The Sun orbits the center of the Galaxy in 225 million years at a distance of 26,000 light-years. use the orbital velocity law to calculate the mass of the Galaxy within the Sun's orbit?

20. Suppose the Sun orbited a little farther out, but the mass of the Galaxy inside its orbit remained the same as we calculated in **Exercise 26.19**. What would be its period at a distance of 30,000 light-years?

21. We have said that the Galaxy rotates differentially; that is, stars in the inner parts complete a full 360° orbit around the center of the Galaxy more rapidly than stars farther out. Use Kepler's third law and the mass we derived in **Exercise 26.19** to calculate the period of a star that is only 5000 light-years from the center. Now do the same calculation for a globular cluster at a distance of 50,000 light-years. Suppose the Sun, this star, and the globular cluster all fall on a straight line through the center of the Galaxy. Where will they be relative to each other after the Sun completes one full journey around the center of the Galaxy? (Assume that all the mass in the Galaxy is concentrated at its center.)

22. If our solar system is 4.6 billion years old, how many galactic years has planet Earth been around?

23. Suppose the average mass of a star in the Galaxy is one-third of a solar mass. Use the value for the mass of the Galaxy that we calculated in **Exercise 26.19**, and estimate how many stars are in the Milky Way. Give some reasons it is reasonable to assume that the mass of an average star is less than the mass of the Sun.

24. The first clue that the Galaxy contains a lot of dark matter was the observation that the orbital velocities of stars did not decreases with increasing distance from the center of the Galaxy. Construct a rotation curve for the solar system by using the orbital velocities of the planets, which can be found in **Appendix D**. How does this curve differ from the rotation curve for the Galaxy? What does it tell you about where most of the mass in the solar system is concentrated?

25. A certain globular cluster has a radius of about 50 light-years. Stars near the outskirts of the cluster have roughly circular orbits and travel at a speed of 12 km/s. Use the orbital velocity law to find the mass of the cluster.

26.4 The Center of the Galaxy

26. The best evidence for a black hole at the center of the Galaxy also comes from the application of the orbital velocity law. Suppose a star at a distance of 20 light-hours from the center of the Galaxy has an orbital speed of 6200 km/s. How much mass must be located inside its orbit?

27. The next step in deciding whether the object in **Exercise 26.26** is a black hole is to estimate the density of this mass. Assume that all of the mass is spread uniformly throughout a sphere with a radius of 20 light-hours. What is the density in kg/km^3? (Remember that the volume of a sphere is given by $V = \frac{4}{3}\pi R^3$.) Explain why the density might be even higher than the value you have calculated. How does this density compare with that of the Sun or other objects we have talked about in this book?

26.6 The Formation of the Galaxy

28. Suppose the Sagittarius dwarf galaxy merges completely with the Milky Way and adds 150,000 stars to it. Estimate the percentage change in the mass of the Milky Way. Will this be enough mass to affect the orbit of the Sun around the galactic center? Assume that all of the Sagittarius galaxy's stars end up in the nuclear bulge of the Milky Way Galaxy and explain your answer.

27 | GALAXIES

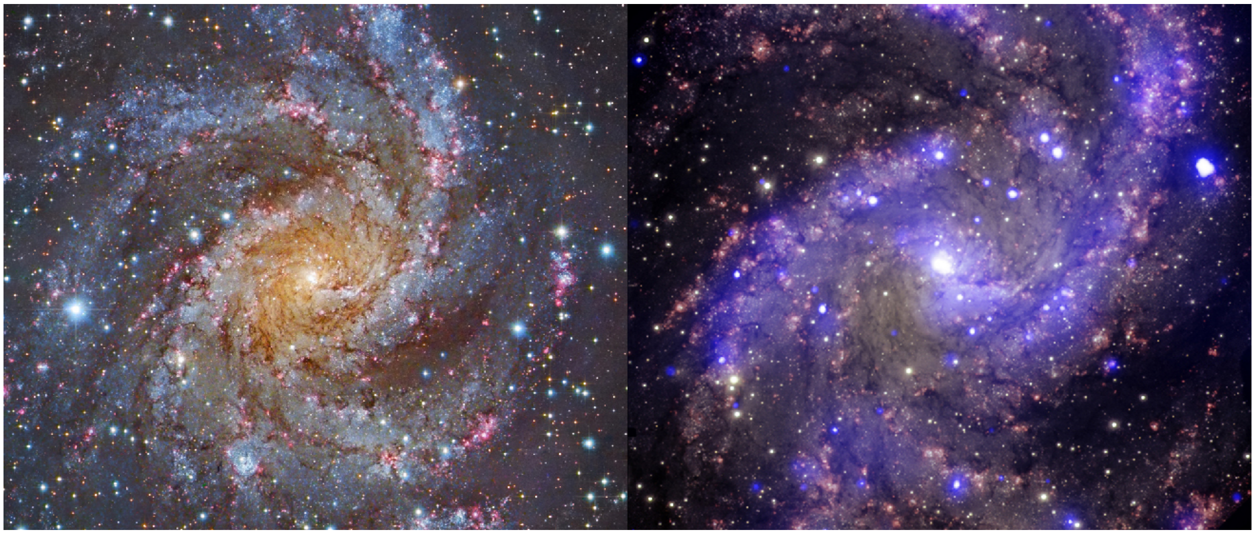


Figure 27.1 NGC 6946 is a spiral galaxy also known as the "Fireworks galaxy." It is at a distance of about 18 million light-years, in the direction of the constellations Cepheus and Cygnus. It was discovered by William Herschel in 1798. This galaxy is about one-third the size of the Milky Way. Note on the left how the colors of the galaxy change from the yellowish light of old stars in the center to the blue color of hot, young stars and the reddish glow of hydrogen clouds in the spiral arms. As the image shows, this galaxy is rich in dust and gas, and new stars are still being born here. In the right-hand image, the x-rays coming from this galaxy are shown in purple, which has been added to other colors showing visible light. (Credit left: modification of work by NASA, ESA, STScI, R. Gendler, and the Subaru Telescope (NAOJ); credit right: modification of work by X-ray: NASA/CXC/MSSL/R.Soria et al, Optical: AURA/Gemini OBs)

Chapter Outline

Introduction

In the last chapter, we explored our own Galaxy. But is it the only one? If there are others, are they like the Milky Way? How far away are they? Can we see them? As we shall learn, some galaxies turn out to be so far away that it has taken billions of years for their light to reach us. These remote galaxies can tell us what the universe was like when it was young.

We begin our voyage with a look at our own galaxy, the Milky Way. (After all, a tourist's basis for understanding new places is what she knows from her own home.) Then, we'll examine a guide to the properties of galaxies, much as a tourist begins with a guidebook to the main features of the cities on the itinerary. In the end, we will look more carefully at the past history of galaxies, how they have changed over time, and how they acquired their many different forms.

27.1 | The Discovery of Galaxies

Learning Objectives

By the end of this section, you will be able to:

- Describe the discoveries that confirmed the existence of galaxies that lie far beyond the Milky Way Galaxy
- Explain why galaxies used to be called nebulae and why we don't include them in that category any more

Growing up at a time when the Hubble Space Telescope orbits above our heads and giant telescopes are springing up on the great mountaintops of the world, you may be surprised to learn that we were not sure about the existence of other galaxies for a very long time. The very idea that other galaxies exist used to be controversial. Even into the 1920s, many astronomers thought the Milky Way encompassed *all* that exists in the universe. The evidence found in 1924 that meant our Galaxy is not alone was one of the great scientific discoveries of the twentieth century.

It was not that scientists weren't asking questions. They questioned the composition and structure of the universe as early as the eighteenth century. However, with the telescopes available in earlier centuries, galaxies looked like small fuzzy patches of light that were difficult to distinguish from the star clusters and gas-and-dust clouds that are part of our own Galaxy. All objects that were not sharp points of light were given the same name, *nebulae*, the Latin word for "clouds." Because their precise shapes were often hard to make out and no techniques had yet been devised for measuring their distances, the nature of the nebulae was the subject of much debate.

As early as the eighteenth century, the philosopher Immanuel Kant (1724–1804) suggested that some of the nebulae might be distant systems of stars (other Milky Ways), but the evidence to support this suggestion was beyond the capabilities of the telescopes of that time.

Other Galaxies

By the early twentieth century, some nebulae had been correctly identified as star clusters, and others (such as the Orion Nebula) as gaseous nebulae. Most nebulae, however, looked faint and indistinct, even with the best telescopes, and their distances remained unknown. If these nebulae were nearby, with distances comparable to those of observable stars, they were most likely clouds of gas or groups of stars within our Galaxy. If, on the other hand, they were remote, far beyond the edge of the Galaxy, they could be other star systems containing billions of stars.

To determine what the nebulae are, astronomers had to find a way of measuring the distances to at least some of them. When the 2.5-meter (100-inch) telescope on Mount Wilson in Southern California went into operation, astronomers finally had the large telescope they needed to settle the controversy.

Working with the 2.5-meter telescope, Edwin Hubble was able to resolve individual stars in several of the brighter spiral-shaped nebulae, including M31, the great spiral in Andromeda (**Figure 27.2**). Among these stars, he discovered some faint variable stars that—when he analyzed their light curves—turned out to be cepheids. Here were reliable indicators that Hubble could use to measure the distances to the nebulae using the technique pioneered by Henrietta Leavitt (see the chapter on **Celestial Distances**). After painstaking work, he estimated that the Andromeda galaxy was about 900,000 light-years away from us. At that enormous distance, it had to be a separate galaxy of stars located well outside the boundaries of the Milky Way. Today, we know the Andromeda galaxy is actually slightly more than twice as distant as Hubble's first estimate, but his conclusion about its true nature remains unchanged.

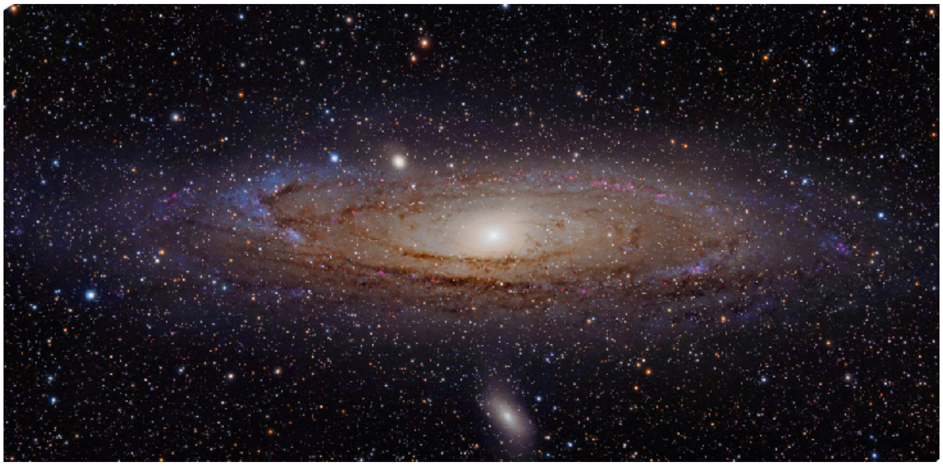


Figure 27.2 Also known by its catalog number M31, the Andromeda galaxy is a large spiral galaxy very similar in appearance to, and slightly larger than, our own Galaxy. At a distance of about 2.5 million light-years, Andromeda is the spiral galaxy that is nearest to our own in space. Here, it is seen with two of its satellite galaxies, M32 (top) and M110 (bottom). (credit: Adam Evans)

No one in human history had ever measured a distance so great. When Hubble's paper on the distances to nebulae was read before a meeting of the American Astronomical Society on the first day of 1925, the entire room erupted in a standing ovation. A new era had begun in the study of the universe, and a new scientific field—extragalactic astronomy—had just been born.

Edwin Hubble: Expanding the Universe

The son of a Missouri insurance agent, Edwin Hubble (**Figure 27.3**) graduated from high school at age 16. He excelled in sports, winning letters in track and basketball at the University of Chicago, where he studied both science and languages. Both his father and grandfather wanted him to study law, however, and he gave in to family pressure. He received a prestigious Rhodes scholarship to Oxford University in England, where he studied law with only middling enthusiasm. Returning to be the United States, he spent a year teaching high school physics and Spanish as well as coaching basketball, while trying to determine his life's direction.

Figure 27.3 Edwin Hubble established some of the most important ideas in the study of galaxies.

The pull of astronomy eventually proved too strong to resist, and so Hubble went back to the University of Chicago for graduate work. Just as he was about to finish his degree and accept an offer to work at the soon-to be completed 5-meter telescope, the United States entered World War I, and Hubble enlisted as an officer. Although the war had ended by the time he arrived in Europe, he received more officer's training abroad and enjoyed a brief time of further astronomical study at Cambridge before being sent home.

In 1919, at age 30, he joined the staff at Mount Wilson and began working with the world's largest telescope. Ripened by experience, energetic, disciplined, and a skillful observer, Hubble soon established some of the most important ideas in modern astronomy. He showed that other galaxies existed, classified them on the basis of their shapes, found a pattern to their motion (and thus put the notion of an expanding universe on a firm observational footing), and began a lifelong program to study the distribution of galaxies in the universe. Although a few others had glimpsed pieces of the puzzle, it was Hubble who put it all together and showed that an understanding of the large-scale structure of the universe was feasible.

His work brought Hubble much renown and many medals, awards, and honorary degrees. As he became better known (he was the first astronomer to appear on the cover of *Time* magazine), he and his wife enjoyed and cultivated friendships with movie stars and writers in Southern California. Hubble was instrumental (if you'll pardon the pun) in the planning and building of the 5-meter telescope on Palomar Mountain, and he had begun to use it for studying galaxies when he passed away from a stroke in 1953.

When astronomers built a space telescope that would allow them to extend Hubble's work to distances he could only dream about, it seemed natural to name it in his honor. It was fitting that observations with the Hubble Space Telescope (and his foundational work on expansion of the universe) contributed to the 2011 Nobel Prize in Physics, given for the discovery that the expansion of the universe is accelerating (a topic we will expand upon in the chapter on **Big Bang Cosmology**).

27.2 | Types of Galaxies

Learning Objectives

By the end of this section, you will be able to:

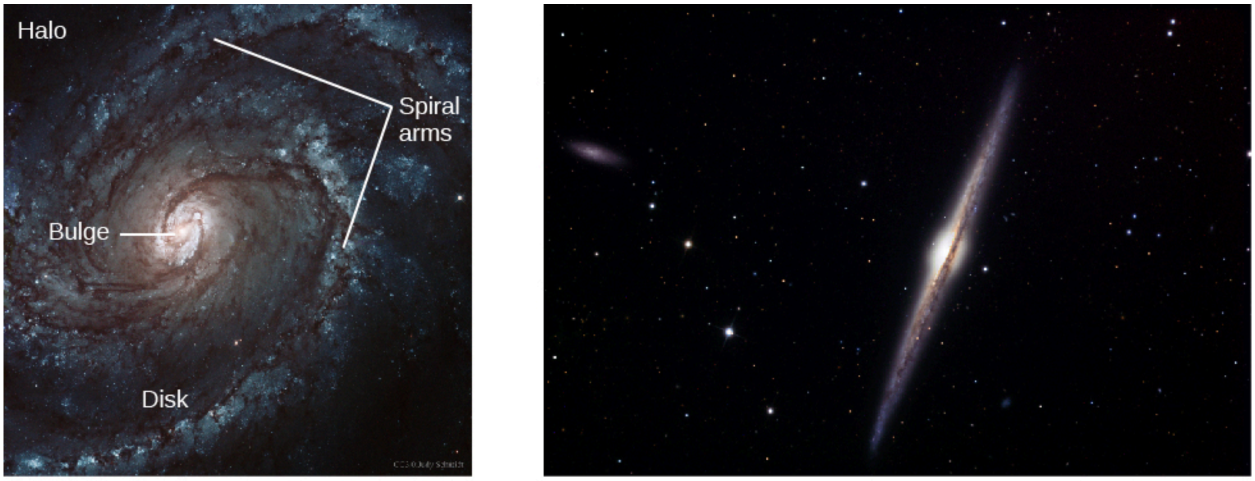
- Describe the properties and features of elliptical, spiral, and irregular galaxies
- Explain what may cause a galaxy's appearance to change over time

Having established the existence of other galaxies, Hubble and others began to observe them more closely—noting their shapes, their contents, and as many other properties as they could measure. This was a daunting task in the 1920s when obtaining a single photograph or spectrum of a galaxy could take a full night of tireless observing. Today, larger telescopes and electronic detectors have made this task less difficult, although observing the most distant galaxies (those that show us the universe in its earliest phases) still requires enormous effort.

The first step in trying to understand a new type of object is often simply to describe it. Remember, the first step in understanding stellar spectra was simply to sort them according to appearance (see **Using Spectra to Measure Stellar Composition and Motion**). As it turns out, the biggest and most luminous galaxies come in one of two basic shapes: either they are flatter and have spiral arms, like our own Galaxy, or they appear to be elliptical (blimp- or cigar-shaped). Many smaller galaxies, in contrast, have an irregular shape.

Spiral Galaxies

Our own Galaxy and the Andromeda galaxy are typical, large **spiral galaxies** (see **Figure 27.2**). They consist of a central bulge, a halo, a disk, and spiral arms. Interstellar material is usually spread throughout the disks of spiral galaxies. Bright emission nebulae and hot, young stars are present, especially in the spiral arms, showing that new star formation is still occurring. The disks are often dusty, which is especially noticeable in those systems that we view almost edge on (**Figure 27.4**).



(a) (b)

Figure 27.4 (a) The spiral arms of M100, shown here, are bluer than the rest of the galaxy, indicating young, high-mass stars and star-forming regions. (b) We view this spiral galaxy, NGC 4565, almost exactly edge on, and from this angle, we can see the dust in the plane of the galaxy; it appears dark because it absorbs the light from the stars in the galaxy. (credit a: modification of work by Hubble Legacy Archive, NASA, ESA, and Judy Schmidt; credit b: modification of work by “Jschulman555”/ Wikimedia)

In galaxies that we see face on, the bright stars and emission nebulae make the arms of spirals stand out like those of a pinwheel on the fourth of July. Open star clusters can be seen in the arms of nearer spirals, and globular clusters are often visible in their halos. Spiral galaxies contain a mixture of young and old stars, just as the Milky Way does. All spirals rotate, and the direction of their spin is such that the arms appear to trail much like the wake of a boat.

About two-thirds of the nearby spiral galaxies have boxy or peanut-shaped bars of stars running through their centers (**Figure 27.5**). Showing great originality, astronomers call these galaxies barred spirals.

Figure 27.5 NGC 1300, shown here, is a barred spiral galaxy. Note that the spiral arms begin at the ends of the bar. (credit: NASA, ESA, and the Hubble Heritage Team(STScI/AURA))

As we noted in **The Milky Way Galaxy** chapter, our Galaxy has a modest bar too (see **Figure 26.10**). The spiral arms usually begin from the ends of the bar. The fact that bars are so common suggests that they are long lived; it may be that most spiral galaxies form a bar at some point during their evolution.

In both barred and unbarred spiral galaxies, we observe a range of different shapes. At one extreme, the central bulge is large and luminous, the arms are faint and tightly coiled, and bright emission nebulae and supergiant stars are inconspicuous. Hubble, who developed a system of classifying galaxies by shape, gave these galaxies the designation Sa. Galaxies at this extreme may have no clear spiral arm structure, resulting in a lens-like appearance (they are sometimes referred to as lenticular galaxies). These galaxies seem to share as many properties with elliptical galaxies as they do with spiral galaxies

At the other extreme, the central bulge is small and the arms are loosely wound. In these Sc galaxies, luminous stars and emission nebulae are very prominent. Our Galaxy and the Andromeda galaxy are both intermediate between the two extremes. Photographs of spiral galaxies, illustrating the different types, are shown in **Figure 27.6**, along with elliptical galaxies for comparison.

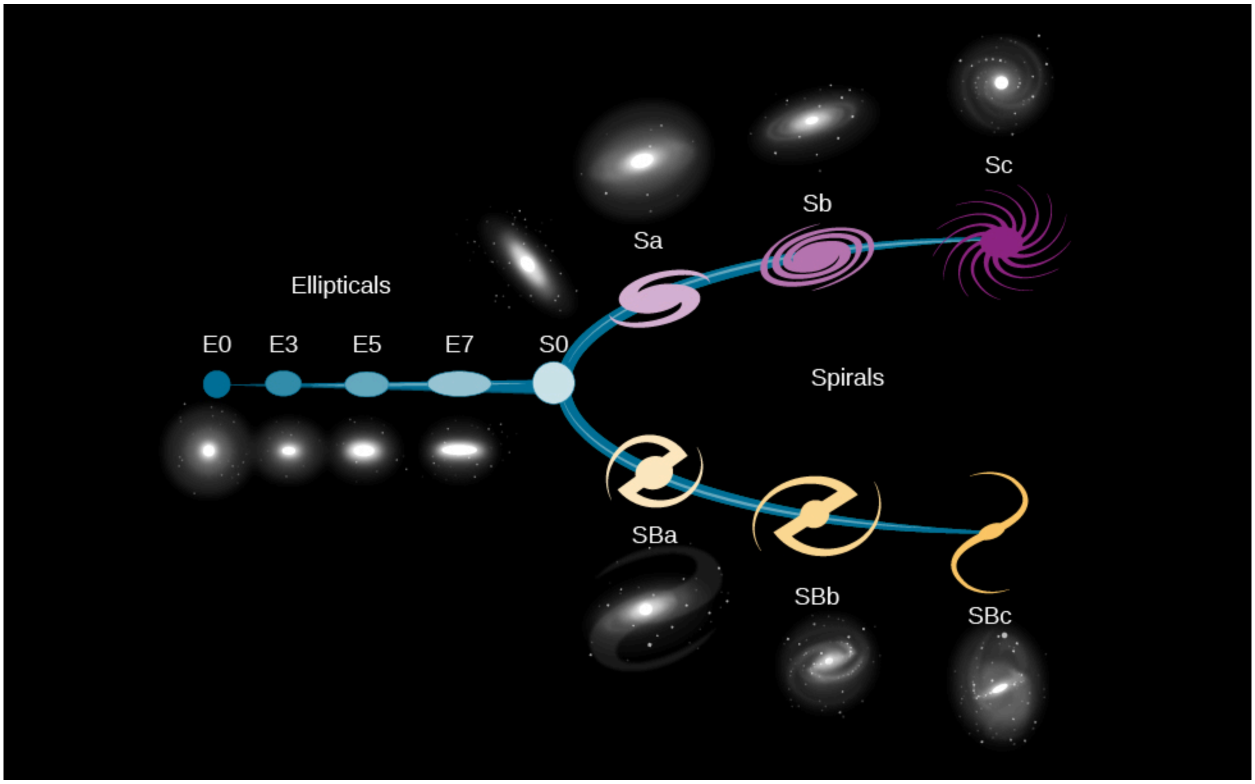


Figure 27.6 This figure shows Edwin Hubble's original classification of galaxies. Elliptical galaxies are on the left. On the right, you can see the basic spiral shapes illustrated, alongside images of actual barred and unbarred spirals. (credit: modification of work by NASA, ESA)

The luminous parts of spiral galaxies appear to range in diameter from about 20,000 to more than 100,000 light-years. Recent studies have found that there is probably a large amount of galactic material that extends well beyond the apparent edge of galaxies. This material appears to be thin, cold gas that is difficult to detect in most observations.

From the observational data available, the masses of the visible portions of spiral galaxies are estimated to range from 1 billion to 1 trillion Suns (10^9 to 10^{12} M_{Sun}). The total luminosities of most spirals fall in the range of 100 million to 100 billion times the luminosity of our Sun (10^8 to 10^{11} L_{Sun}). Our Galaxy and M31 are relatively large and massive, as spirals go. There is also considerable dark matter in and around the galaxies, just as there is in the Milky Way; we deduce its presence from how fast stars in the outer parts of the Galaxy are moving in their orbits.

Elliptical Galaxies

Elliptical galaxies consist almost entirely of old stars and have shapes that are spheres or ellipsoids (somewhat squashed spheres) (**Figure 27.7**). They contain no trace of spiral arms. Their light is dominated by older reddish stars (the population II stars discussed in **The Milky Way Galaxy**). In the larger nearby ellipticals, many globular clusters can be identified. Dust and emission nebulae are not conspicuous in elliptical galaxies, but many do contain a small amount of interstellar matter.

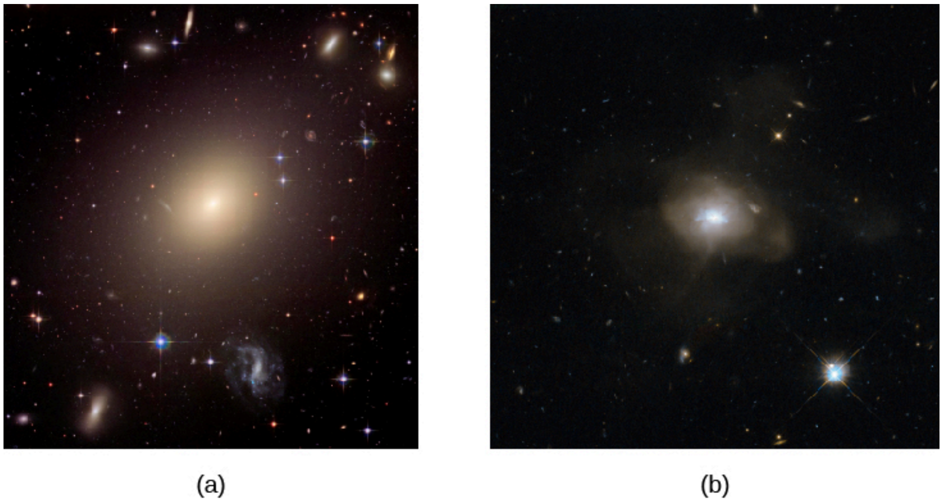


Figure 27.7 (a) ESO 325-G004 is a giant elliptical galaxy. Other elliptical galaxies can be seen around the edges of this image. (b) This elliptical galaxy probably originated from the collision of two spiral galaxies. (credit a: modification of work by NASA, ESA, and The Hubble Heritage Team (STScI/AURA); credit b: modification of work by ESA/Hubble, NASA)

Elliptical galaxies show various degrees of flattening, ranging from systems that are approximately spherical to those that approach the flatness of spirals. The rare giant ellipticals (for example, ESO 325-G004 in **Figure 27.7**) reach luminosities of 10^{11} L_{Sun}. The mass in a giant elliptical can be as large as 10^{13} M_{Sun}. The diameters of these large galaxies extend over several hundred thousand light-years and are considerably larger than the largest spirals. Although individual stars orbit the center of an elliptical galaxy, the orbits are not all in the same direction, as occurs in spirals. Therefore, ellipticals don't appear to rotate in a systematic way, making it difficult to estimate how much dark matter they contain.

We find that elliptical galaxies range all the way from the giants, just described, to dwarfs, which may be the most common kind of galaxy. *Dwarf ellipticals* (sometimes called dwarf spheroidals) escaped our notice for a long time because they are very faint and difficult to see. An example of a dwarf elliptical is the Leo I Dwarf Spheroidal galaxy shown in **Figure 27.8**. The luminosity of this typical dwarf is about equal to that of the brightest globular clusters.

Intermediate between the giant and dwarf elliptical galaxies are systems such as M32 and M110, the two companions of the Andromeda galaxy. While they are often referred to as dwarf ellipticals, these galaxies are significantly larger than galaxies such as Leo I.

Figure 27.8 M32, a dwarf elliptical galaxy and one of the companions to the giant Andromeda galaxy M31. M32 is a dwarf by galactic standards, as it is only 2400 light-years across. (credit: NOAO/AURA/NSF)

Irregular Galaxies

Hubble classified galaxies that do not have the regular shapes associated with the categories we just described into the catchall bin of an **irregular galaxy**, and we continue to use his term. Typically, irregular galaxies have lower masses and luminosities than spiral galaxies. Irregular galaxies often appear disorganized, and many are undergoing relatively intense star formation activity. They contain both young population I stars and old population II stars.

The two best-known irregular galaxies are the Large Magellanic Cloud and Small Magellanic Cloud (**Figure 27.9**), which are at a distance of a little more than 160,000 light-years away and are among our nearest extragalactic neighbors. Their names reflect the fact that Ferdinand Magellan and his crew, making their round-the-world journey, were the first European travelers to notice them. Although not visible from the United States and Europe, these two systems are prominent from the Southern Hemisphere, where they look like wispy clouds in the night sky. Since they are only about one-tenth as distant as the Andromeda galaxy, they present an excellent opportunity for astronomers to study nebulae, star clusters, variable stars, and other key objects in the setting of another galaxy. For example, the Large Magellanic Cloud contains the 30 Doradus complex (also known as the Tarantula Nebula), one of the largest and most luminous groups of supergiant stars known in any galaxy.

Figure 27.9 The Milky Way is seen to the right of the dome, and the Large and Small Magellanic Clouds are seen to the left. (credit: Roger Smith/NOAO/AURA/NSF)

The Small Magellanic Cloud is considerably less massive than the Large Magellanic Cloud, and it is six times longer than it is wide. This narrow wisp of material points directly toward our Galaxy like an arrow. The Small Magellanic Cloud was most likely contorted into its current shape through gravitational interactions with the Milky Way. A large trail of debris from this interaction between the Milky Way and the Small Magellanic Cloud has been strewn across the sky and is seen as a series of gas clouds moving at abnormally high velocity, known as the Magellanic Stream. We will see that this kind of interaction between galaxies will help explain the irregular shapes of this whole category of small galaxies,

View this **beautiful album showcasing the different types of galaxies (https://openstaxcollege.org/l/30galaxphohubb)** that have been photographed by the Hubble Space Telescope.

Galaxy Evolution

Encouraged by the success of **the H-R diagram** for stars, astronomers studying galaxies hoped to find some sort of comparable scheme, where differences in appearance could be tied to different evolutionary stages in the life of galaxies. Wouldn't it be nice if every elliptical galaxy evolved into a spiral, for example, just as every main-sequence star evolves into a red giant? Several simple ideas of this kind were tried, some by Hubble himself, but none stood the test of time (and observation).

Because no simple scheme for evolving one type of galaxy into another could be found, astronomers then tended to the opposite point of view. For a while, most astronomers thought that all galaxies formed very early in the history of the universe and that the differences between them had to do with the rate of star formation. Ellipticals were those galaxies in which all the interstellar matter was converted rapidly into stars. Spirals were galaxies in which star formation occurred slowly over the entire lifetime of the galaxy. This idea turned out to be too simple as well.

Today, we understand that at least some galaxies have changed types over the billions of years since the universe began. As we shall see in later chapters, collisions and mergers between galaxies may dramatically change spiral galaxies into elliptical galaxies. Even isolated spirals (with no neighbor galaxies in sight) can change their appearance over time. As they consume their gas, the rate of star formation will slow down, and the spiral arms will gradually become less conspicuous. Over long periods, spirals therefore begin to look more like the galaxies at the middle of **Figure 27.6** (which astronomers refer to as S0 types).

Over the past several decades, the study of how galaxies evolve over the lifetime of the universe has become one of the most active fields of astronomical research. We will discuss the evolution of galaxies in more detail in **The Evolution and Distribution of Galaxies**, but let's first see in a little more detail just what different galaxies are like.

27.3 | Properties of Galaxies

Learning Objectives

By the end of this section, you will be able to:

- Describe the methods through which astronomers can estimate the mass of a galaxy
- Characterize each type of galaxy by its mass-to-light ratio

The technique for deriving the masses of galaxies is basically the same as that used to estimate the mass of the Sun, the stars, and our own Galaxy. We measure how fast objects in the outer regions of the galaxy are orbiting the center, and then we use this information along with Kepler's third law to calculate how much mass is inside that orbit.

Masses of Galaxies

Astronomers can measure the rotation speed in spiral galaxies by obtaining spectra of either stars or gas, and looking for wavelength shifts produced by the Doppler effect. Remember that the faster something is moving toward or away from us, the greater the shift of the lines in its spectrum. Kepler's law, together with such observations of the part of the Andromeda galaxy that is bright in visible light, for example, show it to have a galactic mass of about 4×10^{11} M_{Sun} (enough material to make 400 billion stars like the Sun).

The total mass of the Andromeda galaxy is greater than this, however, because we have not included the mass of the material that lies beyond its visible edge. Fortunately, there is a handful of objects—such as isolated stars, star clusters, and satellite galaxies—beyond the visible edge that allows astronomers to estimate how much additional matter is hidden out there. Recent studies show that the amount of dark matter beyond the visible edge of Andromeda may be as large as the mass of the bright portion of the galaxy. Indeed, using Kepler's third law and the velocities of its satellite galaxies, the Andromeda galaxy is estimated to have a mass closer to 1.4×10^{12} M_{Sun}. The mass of the Milky Way Galaxy is estimated to be 8.5×10^{11} M_{Sun}, and so our Milky Way is turning out to be somewhat smaller than Andromeda.

Elliptical galaxies do not rotate in a systematic way, so we cannot determine a rotational velocity; therefore, we must use a slightly different technique to measure their mass. Their stars are still orbiting the galactic center, but not in the organized way that characterizes spirals. Since elliptical galaxies contain stars that are billions of years old, we can assume that the galaxies themselves are not flying apart. Therefore, if we can measure the various speeds with which the stars are moving in their orbits around the center of the galaxy, we can calculate how much mass the galaxy must contain in order to hold the stars within it.

In practice, the spectrum of a galaxy is a composite of the spectra of its many stars, whose different motions produce different Doppler shifts (some red, some blue). The result is that the lines we observe from the entire galaxy contain the combination of many Doppler shifts. When some stars provide blueshifts and others provide redshifts, they create a wider or broader absorption or emission feature than would the same lines in a hypothetical galaxy in which the stars had no orbital motion. Astronomers call this phenomenon line broadening. The amount by which each line broadens indicates the range of speeds at which the stars are moving with respect to the center of the galaxy. The range of speeds depends, in turn, on the force of gravity that holds the stars within the galaxies. With information about the speeds, it is possible to calculate the mass of an elliptical galaxy.

Table 27.1 summarizes the range of masses (and other properties) of the various types of galaxies. Interestingly enough, the most and least massive galaxies are ellipticals. On average, irregular galaxies have less mass than spirals.

Characteristics of the Different Types of Galaxies

Characteristic	Spirals	Ellipticals	Irregulars
Mass (M_{Sun})	10^9 to 10^{12}	10^5 to 10^{13}	10^8 to 10^{11}
Diameter (thousands of light-years)	15 to 150	3 to >700	3 to 30
Luminosity (L_{Sun})	10^8 to 10^{11}	10^6 to 10^{11}	10^7 to 2×10^9
Populations of stars	Old and young	Old	Old and young
Interstellar matter	Gas and dust	Almost no dust; little gas	Much gas; some have little dust, some much dust
Mass-to-light ratio in the visible part	2 to 10	10 to 20	1 to 10
Mass-to-light ratio for total galaxy	100	100	?

Table 27.1

Mass-to-Light Ratio

A useful way of characterizing a galaxy is by noting the ratio of its mass (in units of the Sun's mass) to its light output (in units of the Sun's luminosity). This single number tells us roughly what kind of stars make up most of the luminous population of the galaxy, and it also tells us whether a lot of dark matter is present. For stars like the Sun, the **mass-to-light ratio** is 1 by our definition.

Galaxies are not, of course, composed entirely of stars that are identical to the Sun. The overwhelming majority of stars are less massive and less luminous than the Sun, and usually these stars contribute most of the mass of a system without accounting for very much light. The mass-to-light ratio for low-mass stars is greater than 1 (you can verify this using the data in **Table 22.5**). Therefore, a galaxy's mass-to-light ratio is also generally greater than 1, with the exact value depending on the ratio of high-mass stars to low-mass stars.

Galaxies in which star formation is still occurring have many massive stars, and their mass-to-light ratios are usually in the range of 1 to 10. Galaxies consisting mostly of an older stellar population, such as ellipticals, in which the massive stars have already completed their evolution and have ceased to shine, have mass-to-light ratios of 10 to 20.

But these figures refer only to the inner, conspicuous parts of galaxies (**Figure 27.10**). In **The Milky Way Galaxy** and above, we discussed the evidence for dark matter in the outer regions of our own Galaxy, extending much farther from the galactic center than do the bright stars and gas. Recent measurements of the rotation speeds of the outer parts of nearby galaxies, such as the Andromeda galaxy we discussed earlier, suggest that they too have extended distributions of dark matter around the visible disk of stars and dust. This largely invisible matter adds to the mass of the galaxy while contributing nothing to its luminosity, thus increasing the mass-to-light ratio. If dark invisible matter is present in a galaxy, its mass-to-light ratio can be as high as 100. The two different mass-to-light ratios measured for various types of galaxies are given in **Table 27.1**.

Figure 27.10 This galaxy is a face-on spiral at a distance of 21 million light-years. M101 is almost twice the diameter of the Milky Way, and it contains at least 1 trillion stars. (credit: NASA, ESA, K. Kuntz (Johns Hopkins University), F. Bresolin (University of Hawaii), J. Trauger (Jet Propulsion Lab), J. Mould (NOAO), Y.-H. Chu (University of Illinois, Urbana), and STScI)

These measurements of other galaxies support the conclusion already reached from studies of the rotation of our own Galaxy—namely, that most of the material in the universe cannot at present be observed directly in any part of the electromagnetic spectrum. An understanding of the properties and distribution of this invisible matter is crucial to our understanding of galaxies. It's becoming clearer and clearer that, through the gravitational force it exerts, dark matter plays a dominant role in galaxy formation and early evolution. There is an interesting parallel here between our time and the time during which Edwin Hubble was receiving his training in astronomy. By 1920, many scientists were aware that astronomy stood on the brink of important breakthroughs—if only the nature and behavior of the nebulae could be settled with better observations. In the same way, many astronomers today feel we may be closing in on a far more sophisticated understanding of the large-scale structure of the universe—if only we can learn more about the nature and properties of dark matter. If you follow astronomy articles in the news (as we hope you will), you should be hearing more about dark matter in the years to come.

27.4 | The Extragalactic Distance Scale

Learning Objectives

By the end of this section, you will be able to:

- Describe the use of variable stars to estimate distances to galaxies
- Explain how standard candles and the Tully-Fisher relation can be used to estimate distances to galaxies

To determine many of the properties of a galaxy, such as its luminosity or size, we must first know how far away it is. If we know the distance to a galaxy, we can convert how bright the galaxy appears to us in the sky into its true luminosity because we know the precise way light is dimmed by distance. (The same galaxy 10 times farther away, for example, would look 100 times dimmer.) But the measurement of galaxy distances is one of the most difficult problems in modern astronomy: all galaxies are far away, and most are so distant that we cannot even make out individual stars in them.

For decades after Hubble's initial work, the techniques used to measure galaxy distances were relatively inaccurate, and different astronomers derived distances that differed by as much as a factor of two. (Imagine if the distance between your home or dorm and your astronomy class were this uncertain; it would be difficult to make sure you got to class on time.) In the past few decades, however, astronomers have devised new techniques for measuring distances to galaxies; most importantly, all of them give the same answer to within an accuracy of about 10%. As we will see, this means we may finally be able to make reliable estimates of the size of the universe.

Variable Stars

Before astronomers could measure distances to other galaxies, they first had to establish the scale of cosmic distances using objects in our own Galaxy. We described the chain of these distance methods in **Celestial Distances** (and we recommend that you review that chapter if it has been a while since you've read it). Astronomers were especially delighted when they discovered that they could measure distances using certain kinds of intrinsically luminous *variable stars*, such as cepheids, which can be seen at very large distances (**Figure 27.11**).

After the variables in nearby galaxies had been used to make distance measurements for a few decades, Walter Baade showed that there were actually two kinds of cepheids and that astronomers had been unwittingly mixing them up. As a result, in the early 1950s, the distances to all of the galaxies had to be increased by about a factor of two. We mention this because we want you to bear in mind, as you read on, that science is always a study in progress. Our first tentative steps in such difficult investigations are always subject to future revision as our techniques become more reliable.

The amount of work involved in finding cepheids and measuring their periods can be enormous. Hubble, for example, obtained 350 long-exposure photographs of the Andromeda galaxy over a period of 18 years and was able to identify only 40 cepheids. Even though cepheids are fairly luminous stars, they can be detected in only about 30 of the nearest galaxies with the world's largest ground-based telescopes.

As mentioned in **Celestial Distances**, one of the main projects carried out during the first years of operation of the Hubble Space Telescope was the measurement of cepheids in more distant galaxies to improve the accuracy of the extragalactic distance scale. Recently, astronomers working with the Hubble Space Telescope have extended such measurements out to 108 million light-years—a triumph of technology and determination.

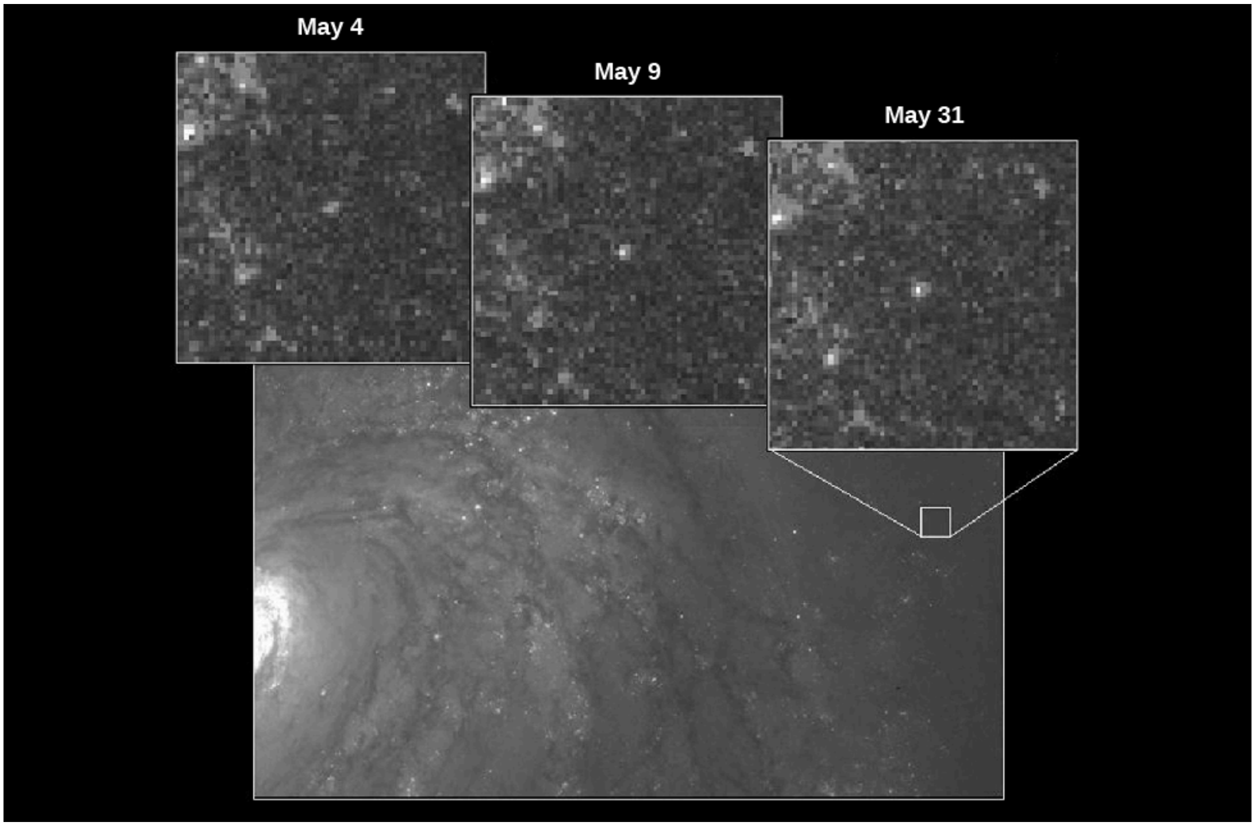


Figure 27.11 In 1994, using the Hubble Space Telescope, astronomers were able to make out an individual cepheid variable star in the galaxy M100 and measure its distance to be 56 million light-years. The insets show the star on three different nights; you can see that its brightness is indeed variable. (credit: modification of work by Wendy L. Freedman, Observatories of the Carnegie Institution of Washington, and NASA/ESA)

Nevertheless, we can only use cepheids to measure distances within a small fraction of the universe of galaxies. After all, to use this method, we must be able to resolve single stars and follow their subtle variations. Beyond a certain distance, even our finest space telescopes cannot help us do this. Fortunately, there are other ways to measure the distances to galaxies.

Standard Candles

We discussed in **Celestial Distances** the great frustration that astronomers felt when they realized that the stars in general were not standard *candles*. If every light bulb in a huge auditorium is a standard 100-watt bulb, then bulbs that look brighter to us must be closer, whereas those that look dimmer must be farther away. If every star were a standard luminosity (or wattage), then we could similarly “read off” their distances based on how bright they appear to us. Alas, as we have learned, neither stars nor galaxies come in one standard-issue luminosity. Nonetheless, astronomers have been searching for objects out there that do act in some way like a standard candle—that have the same intrinsic (built-in) brightness wherever they are. (The use of the term "standard candle" instead of "standard bulb" is historical, although either term would be appropriate.)

As you will recall from the section on **The Brightness of Stars**, if we know the luminosity, L, of a light source (in Watts), then its apparent brightness, b (in Watts/m^2) falls off as the square of the distance, d, from the source. Mathematically:

$$b = \frac{L}{4\pi d^2} \qquad (27.1)$$

Standard Spiral Galaxies

It turns out that many standard, bright spiral galaxies are found which have about the same composition and number of stars, and therefore almost the same luminosity. These can, then, be used as standard candles. Generally, they are known to have a luminosity that corresponds to an absolute magnitude of about M = -21.5. All that is needed, in order to use the **spectroscopic parallax** formula to find the distance to such a galaxy, is a photometric measurement of its apparent magnitude, m. Once again, the distance in parsecs to the galaxy is just $d = 10 \times 10^{0.2(m - M)}$.

Other Standard Candles

A number of suggestions have been made for what sorts of objects might be effective standard candles, including the brightest supergiant stars, planetary nebulae (which give off a lot of ultraviolet radiation), and the average globular cluster in a galaxy. One object turns out to be particularly useful: the **type Ia supernova**. These supernovae involve the explosion of a white dwarf in a binary system (see **The Evolution of Binary Star Systems**) Observations show that supernovae of this type all reach nearly the same luminosity (about 4.5×10^9 L_{Sun}) at maximum light. With such tremendous luminosities, these supernovae have been detected out to a distance of more than 8 billion light-years and are therefore especially attractive to astronomers as a way of determining distances on a large scale (**Figure 27.12**).

Figure 27.12 The bright object at the bottom left of center is a type Ia supernova near its peak intensity. The supernova easily outshines its host galaxy. This extreme increase and luminosity help astronomers use Ia supernova as standard candles. (credit: NASA, ESA, A. Riess (STScI))

Several other kinds of standard candles visible over great distances have also been suggested, including the overall brightness of, for example, giant ellipticals and the brightest member of a galaxy cluster. Type Ia supernovae, however, have proved to be the most accurate standard candles, and they can be seen in more distant galaxies than the other types of calibrators. As we will see in the chapter on **Big Bang Cosmology**, observations of this type of supernova have profoundly changed our understanding of the evolution of the universe.

Other Measuring Techniques

Another technique for measuring galactic distances makes use of an interesting relationship noticed in the late 1970s by Brent Tully of the University of Hawaii and Richard Fisher of the National Radio Astronomy Observatory. They discovered that the luminosity of a spiral galaxy is related to its rotational velocity (how fast it spins). Why would this be true?

The more mass a galaxy has, the faster the objects in its outer regions must orbit. A more massive galaxy has more stars in it and is thus more luminous (ignoring dark matter for a moment). Thinking back to our discussion from the previous section, we can say that if the mass-to-light ratios for various spiral galaxies are pretty similar, then we can estimate the luminosity of a spiral galaxy by measuring its mass, and we can estimate its mass by measuring its rotational velocity.

Tully and Fisher used the 21-cm line of cold hydrogen gas to determine how rapidly material in spiral galaxies is orbiting their centers. Since 21-cm radiation from stationary atoms comes in a nice narrow line, the width of the 21-cm line produced by a whole rotating galaxy tells us the range of orbital velocities of the galaxy’s hydrogen gas. The broader the line, the faster the gas is orbiting in the galaxy, and the more massive and luminous the galaxy turns out to be.

It is somewhat surprising that this technique works, since much of the mass associated with galaxies is dark matter, which does not contribute at all to the luminosity but does affect the rotation speed. There is also no obvious reason why the mass-to-light ratio should be similar for all spiral galaxies. Nevertheless, observations of nearer galaxies (where we have other ways of measuring distance) show that measuring the rotational velocity of a galaxy provides an accurate estimate of its intrinsic luminosity. Once we know how luminous the galaxy really is, we can compare the luminosity to the apparent brightness and use the difference to calculate its distance.

While the Tully-Fisher relation works well, it is limited—we can only use it to determine the distance to a spiral galaxy. There are other methods that can be used to estimate the distance to an elliptical galaxy; however, those methods are beyond the scope of our introductory astronomy course.

Table 27.2 lists the type of galaxy for which each of the distance techniques is useful, and the range of distances over which the technique can be applied.

Some Methods for Estimating Distance to Galaxies

Method	Galaxy Type	Approximate Distance Range (millions of light-years)
Planetary nebulae	All	0–70
Cepheid variables	Spiral, irregulars	0–110
Tully-Fisher relation	Spiral	0–300
Type Ia supernovae	All	0–11,000
Redshifts (Hubble's law)	All	300–13,000

Table 27.2

27.5 | The Expanding Universe

Learning Objectives

By the end of this section, you will be able to:

- Describe the discovery that galaxies getting farther apart as the universe evolves
- Explain how to use Hubble's law to determine distances to remote galaxies
- Describe models for the nature of an expanding universe
- Explain the variation in Hubble's constant

We now come to one of the most important discoveries ever made in astronomy—the fact that the universe is expanding. Before we describe how the discovery was made, we should point out that the first steps in the study of galaxies came at a time when the techniques of spectroscopy were also making great strides. Astronomers using large telescopes could record the spectrum of a faint star or galaxy on photographic plates, guiding their telescopes so they remained pointed to the same object for many hours and collected more light. The resulting spectra of galaxies contained a wealth of information about the composition of the galaxy and the velocities of these great star systems.

Slipher's Pioneering Observations

Curiously, the discovery of the expansion of the universe began with the search for Martians and other solar systems. In 1894, the controversial (and wealthy) astronomer Percival Lowell established an observatory in Flagstaff, Arizona, to study the planets and search for life in the universe. Lowell thought that the spiral nebulae might be solar systems in the process of formation. He therefore asked one of the observatory's young astronomers, Vesto M. Slipher (Figure 27.13), to photograph the spectra of some of the spiral nebulae to see if their spectral lines might show chemical compositions like those expected for newly forming planets.

Figure 27.13 Slipher spent his entire career at the Lowell Observatory, where he discovered the large radial velocities of galaxies. (credit: Lowell Observatory)

The Lowell Observatory's major instrument was a 24-inch refracting telescope, which was not at all well suited to observations of faint spiral nebulae. With the technology available in those days, photographic plates had to be exposed for 20 to 40 hours to produce a good spectrum (in which the positions of the lines could reveal a galaxy's motion). This often meant continuing to expose the same photograph over several nights. Beginning in 1912, and making heroic efforts over a period of about 20 years, Slipher managed to photograph the spectra of more than 40 of the spiral nebulae (which would all turn out to be galaxies).

To his surprise, the spectral lines of most galaxies showed an astounding **redshift**. By "redshift" we mean that the lines in the spectra are displaced toward longer wavelengths (toward the red end of the visible spectrum). Recall from the chapter on **Using Spectra to Measure Stellar Composition and Motion** that a redshift is seen when the source of the waves is moving away from us. Slipher's observations showed that most spirals are racing away at huge speeds; the highest velocity he measured was 1800 kilometers per second.

Only a few spirals—such as the Andromeda and Triangulum Galaxies and M81—all of which are now known to be our close neighbors, turned out to be approaching us. All the other galaxies were moving away. Slipher first announced this discovery in 1914, years before Hubble showed that these objects were other galaxies and before anyone knew how far away they were. No one at the time quite knew what to make of this discovery.

Hubble's Law

The profound implications of Slipher's work became apparent only during the 1920s. Georges Lemaître was a Belgian priest and a trained astronomer. In 1927, he published a paper in French in an obscure Belgian journal in which he suggested that we live in an expanding universe. The title of the paper (translated into English) is "A Homogenous Universe of Constant Mass and Growing Radius Accounting for the Radial Velocity of Extragalactic Nebulae." Lemaître had discovered that Einstein's equations of relativity were consistent with an expanding universe (as had the Russian scientist Alexander Friedmann independently in 1922). Lemaître then went on to use Slipher's data to support the hypothesis that the universe actually is expanding and to estimate the rate of expansion. Initially, scientists paid little attention to this paper, perhaps because the Belgian journal was not widely available.

In the meantime, Hubble was making observations of galaxies with the 2.5-meter telescope on Mt. Wilson, which was then the world's largest. Hubble carried out the key observations in collaboration with a remarkable man, Milton Humason, who dropped out of school in the eighth grade and began his astronomical career by driving a mule train up the trail on Mount Wilson to the observatory (**Figure 27.14**). In those early days, supplies had to be brought up that way; even astronomers hiked up to the mountaintop for their turns at the telescope. Humason became interested in the work of the astronomers and, after marrying the daughter of the observatory's electrician, took a job as janitor there. After a time, he became a night assistant, helping the astronomers run the telescope and record data. Eventually, he made such a mark that he became a full astronomer at the observatory.

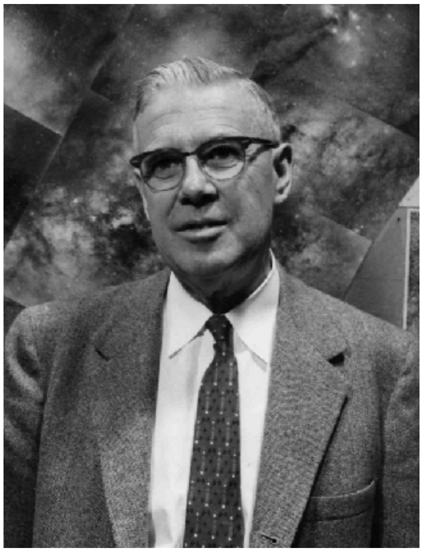


Figure 27.14 Humason was Hubble's collaborator on the great task of observing, measuring, and classifying the characteristics of many galaxies. (credit: Caltech Archives)

By the late 1920s, Humason was collaborating with Hubble by photographing the spectra of faint galaxies with the 2.5-meter telescope. (By then, there was no question that the spiral nebulae were in fact galaxies.) Hubble had found ways to improve the accuracy of the estimates of distances to spiral galaxies, and he was able to measure much fainter and more distant galaxies than Slipher could observe with his much-smaller telescope. When Hubble laid his own distance estimates next to measurements of the recession velocities (the speed with which the galaxies were moving away), he found something stunning: there was a relationship between distance and velocity for galaxies. *The more distant the galaxy, the faster it was receding from us*.

In 1931, Hubble and Humason jointly published the seminal paper where they compared distances and velocities of remote galaxies moving away from us at speeds as high as 20,000 kilometers per second and were able to show that the recession velocities of galaxies are directly proportional to their distances from us (**Figure 27.15**), just as Lemaître had suggested.

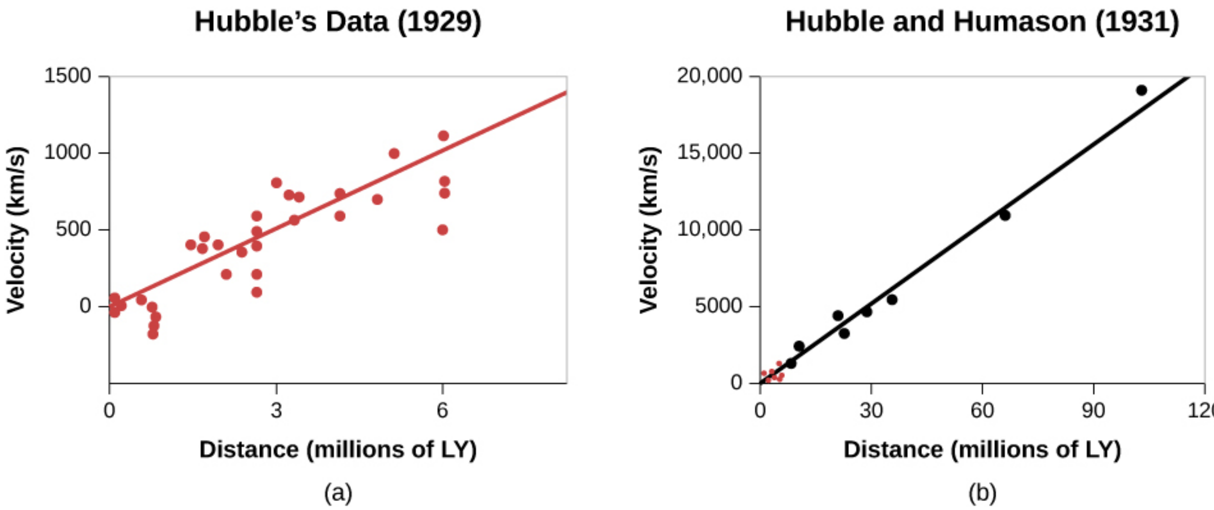


Figure 27.15 (a) These data show Hubble's original velocity-distance relation, adapted from his 1929 paper in the *Proceedings of the National Academy of Sciences*. (b) These data show Hubble and Humason's velocity-distance relation, adapted from their 1931 paper in *The Astrophysical Journal*. The red dots at the lower left are the points in the diagram in the 1929 paper. Comparison of the two graphs shows how rapidly the determination of galactic distances and redshifts progressed in the 2 years between these publications.

We now know that this relationship holds for every galaxy except a few of the nearest ones. Nearly all of the galaxies that are approaching us turn out to be part of the Milky Way's own group of galaxies, which have their own individual motions, just as birds flying in a group may fly in slightly different directions at slightly different speeds even though the entire flock travels through space together.

Written as a formula, the relationship between velocity and distance is

$$V = H_0 \times d \tag{27.2}$$

where v is the recession speed, d is the distance, and H_0 is a number called the **Hubble constant**. This equation is now known as **Hubble's law**.

Astronomers express the value of Hubble's constant in units that relate to how they measure speed and velocity for galaxies. In this book, we will use kilometers per second per million light-years as that unit. For many years, estimates of the value of the Hubble constant have been in the range of 15 to 30 kilometers per second per million light-years The most recent work appears to be converging on a value near 22 kilometers per second per million light-years If H_0 is 22 kilometers per second per million light-years, a galaxy moves away from us at a speed of 22 kilometers per second for every million light-years of its distance. As an example, a galaxy 100 million light-years away is moving away from us at a speed of 2200 kilometers per second.

Hubble's law tells us something fundamental about the universe. Since all but the nearest galaxies appear to be in motion away from us, with the most distant ones moving the fastest, we must be living in an expanding universe. We will explore the implications of this idea shortly, as well as in the final chapters of this text. For now, we will just say that Hubble's observation underlies all our theories about the origin and evolution of the universe.

Hubble's Law and Distances

The regularity expressed in Hubble's law has a built-in bonus: it gives us a new way to determine the distances to remote galaxies. First, we must reliably establish Hubble's constant by measuring both the distance and the velocity of many galaxies in many directions to be sure Hubble's law is truly a universal property of galaxies. But once we have calculated the value of this constant and are satisfied that it applies everywhere, much more of the universe opens up for distance determination. Basically, if we can obtain a spectrum of a galaxy, we can immediately tell how far away it is.

The procedure works like this. We use the spectrum to measure the speed with which the galaxy is moving away from us. If we then put this speed and the Hubble constant into Hubble's law equation, we can solve for the distance.

Example 27.1

Hubble's Law

Hubble's law ($v = H_0 \times d$) allows us to calculate the distance to any galaxy. Here is how we use it in practice.

We have measured Hubble's constant to be 22 km/s per million light-years. This means that if a galaxy is 1 million light-years farther away, it will move away 22 km/s faster. So, if we find a galaxy that is moving away at 18,000 km/s, what does Hubble's law tells us about the distance to the galaxy?

Solution

$$d = \frac{v}{H_0} = \frac{18{,}000 \text{ km/s}}{\frac{22 \text{ km/s}}{1 \text{ million light-years}}} = \frac{18{,}000}{22} \times \frac{1 \text{ million light-years}}{1} = 818 \text{ million light-years}$$

Note how we handled the units here: the km/s in the numerator and denominator cancel, and the factor of million light-years in the denominator of the constant must be divided correctly before we get our distance of 818 million light-years.

Exercise 27.1

Check Your Learning

Using 22 km/s/million light-years for Hubble's constant, what recessional velocity do we expect to find if we observe a galaxy at 500 million light-years?

Solution

$$v = d \times H = 500 \text{ million light-years} \times \frac{22 \text{ km/s}}{1 \text{ million light-years}} = 11{,}000 \text{ km/s}$$

Variation of Hubble's Constant

The use of redshift is potentially a very important technique for determining distances because as we have seen, most of our methods for determining galaxy distances are limited to approximately the nearest few hundred million light-years (and they have large uncertainties at these distances). The use of Hubble's law as a distance indicator requires only a spectrum of a galaxy and a measurement of the Doppler shift, and with large telescopes and modern spectrographs, spectra can be taken of extremely faint galaxies.

But, as is often the case in science, things are not so simple. This technique works if, and only if, the Hubble constant has been truly constant throughout the entire life of the universe. We have used the symbol H_0, with the subscript zero, to mean, specifically, the value of the Hubble constant **today**. When we observe galaxies billions of light-years away, we are seeing them as they were billions of years ago. But what if the Hubble "constant", H, was different billions of years ago? Before 1998, astronomers thought that, although the universe is expanding, the expansion should be slowing down, or decelerating, because the overall gravitational pull of all matter in the universe would have a dominant, measurable effect. If the expansion is decelerating, then the Hubble constant should be decreasing over time.

The discovery that type Ia supernovae are standard bulbs gave astronomers the tool they needed to observe extremely distant galaxies and measure the rate of expansion billions of years ago. The results were completely unexpected. It turns out that the expansion of the universe is *accelerating* over time! What makes this result so astounding is that there is no way that existing physical theories can account for this observation. While a decelerating universe could easily be explained by gravity, there was no force or property in the universe known to astronomers that could account for the acceleration. In **Big Bang Cosmology** chapter, we will look in more detail at the observations that led to this totally unexpected result and explore its implications for the ultimate fate of the universe.

In any case, if the Hubble constant is not really a constant when we look over large spans of space and time, then the calculation of galaxy distances using the Hubble constant won't be accurate. As we shall see in the chapter on **Big Bang Cosmology**, the accurate calculation of distances requires a model for how the Hubble constant has changed over time. The farther away a galaxy is (and the longer ago we are seeing it), the more important it is to include the effects of the change in the Hubble constant. For galaxies within a few billion light-years, however, the assumption that the Hubble constant is indeed constant gives good estimates of distance.

Models for an Expanding Universe

At first, thinking about Hubble's law and being a fan of the work of Copernicus and Harlow Shapley, you might be shocked. Are all the galaxies really moving *away from us*? Is there, after all, something special about our position in the universe? Worry not; the fact that galaxies are receding from us and that more distant galaxies are moving away more rapidly than nearby ones shows only that the universe is expanding uniformly.

A uniformly expanding universe is one that is expanding at the same rate everywhere. In such a universe, we and all other observers, no matter where they are located, must observe a proportionality between the velocities and distances of equivalently remote galaxies. (Here, we are ignoring the fact that the Hubble constant is not constant over all time, but if at any given time in the evolution of the universe the Hubble constant has the same value everywhere, this argument still works.)

To see why, first imagine a ruler made of stretchable rubber, with the usual lines marked off at each centimeter. Now suppose someone with strong arms grabs each end of the ruler and slowly stretches it so that, say, it doubles in length in 1 minute (**Figure 27.16**). Consider an intelligent ant sitting on the mark at 2 centimeters—a point that is not at either end nor in the middle of the ruler. He measures how fast other ants, sitting at the 4-, 7-, and 12-centimeter marks, move away from him as the ruler stretches.

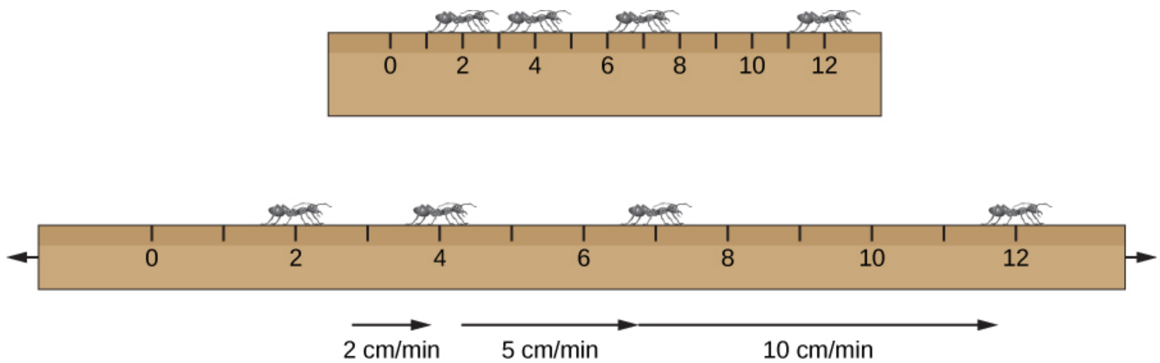


Figure 27.16 Ants on a stretching ruler see other ants move away from them. The speed with which another ant moves away is proportional to its distance.

The ant at 4 centimeters, originally 2 centimeters away from our ant, has doubled its distance in 1 minute; it therefore moved away at a speed of 2 centimeters per minute. The ant at the 7-centimeters mark, which was originally 5 centimeters away from our ant, is now 10 centimeters away; it thus had to move at 5 centimeters per minute. The one that started at the 12-centimeters mark, which was 10 centimeters away from the ant doing the counting, is now 20 centimeters away, meaning it must have raced away at a speed of 10 centimeters per minute. Ants at different distances move away at different speeds, and their speeds are proportional to their distances (just as Hubble's law indicates for galaxies). Yet, notice in our example that all the ruler was doing was stretching uniformly. Also, notice that none of the ants were actually moving of their own accord, it was the stretching of the ruler that moved them apart.

Now let's repeat the analysis, but put the intelligent ant on some other mark—say, on 7 or 12 centimeters. We discover that, as long as the ruler stretches uniformly, this ant also finds every other ant moving away at a speed proportional to its distance. In other words, the kind of relationship expressed by Hubble's law can be explained by a uniform stretching of the "world" of the ants. And all the ants in our simple diagram will see the other ants moving away from them as the ruler stretches.

For a three-dimensional analogy, let's look at the loaf of raisin bread in **Figure 27.17**. The chef has accidentally put too much yeast in the dough, and when she sets the bread out to rise, it doubles in size during the next hour, causing all the raisins to move farther apart. On the figure, we again pick a representative raisin (that is not at the edge or the center of the loaf) and show the distances from it to several others in the figure (before and after the loaf expands).

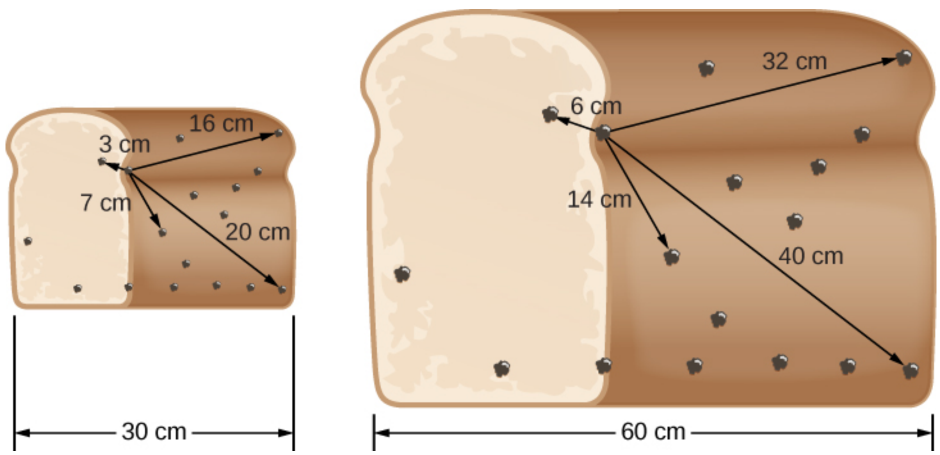


Figure 27.17 As the raisin bread rises, the raisins "see" other raisins moving away. More distant raisins move away faster in a uniformly expanding bread.

Measure the increases in distance and calculate the speeds for yourself on the raisin bread, just like we did for the ruler. You will see that, since each distance doubles during the hour, each raisin moves away from our selected raisin at a speed proportional to its distance. The same is true no matter which raisin you start with.

Our two analogies are useful for clarifying our thinking, but you must not take them literally. On both the ruler and the raisin bread, there are points that are at the end or edge. You can use these to pinpoint the middle of the ruler and the loaf. While our models of the universe have some resemblance to the properties of the ruler and the loaf, the universe has no boundaries, no edges, and no center (all mind-boggling ideas that we will discuss in a later chapter).

What is useful to notice about both the ants and the raisins is that they themselves did not "cause" their motion. It isn't as if the raisins decided to take a trip away from each other and then hopped on a hoverboard to get away. No, in both our analogies, it was the stretching of the medium (the ruler or the bread) that moved the ants or the raisins farther apart. In the same way, we will see in **Big Bang Cosmology** chapter that the galaxies don't have rocket motors propelling them away from each other. Instead, they are passive participants in the *expansion of space*. As space stretches, the galaxies are carried farther and farther apart much as the ants and the raisins were. (If this notion of the "stretching" of space surprises or bothers you, now would be a good time to review the information about spacetime in **Introducing General Relativity**. We will discuss these ideas further as our discussion broadens from galaxies to the whole universe.)

The expansion of the universe, by the way, does not imply that the individual galaxies and clusters of galaxies themselves are expanding. Neither raisins nor the ants in our analogy grow in size as the loaf expands. Similarly, gravity holds galaxies and clusters of galaxies together, and they get farther away from each other—without themselves changing in size—as the universe expands.

For Further Exploration

Websites

ABC's of Distance: **http://www.astro.ucla.edu/~wright/distance.htm (http://www.astro.ucla.edu/~wright/distance.htm)** . A concise summary by astronomer Ned Wright of all the different methods we use to get distances in astronomy.

Cosmic Times 1929: **http://cosmictimes.gsfc.nasa.gov/online_edition/1929Cosmic/index.html (http://cosmictimes.gsfc.nasa.gov/online_edition/1929Cosmic/index.html)** . NASA project explaining Hubble's work and surrounding discoveries as if you were reading newspaper articles.

Edwin Hubble: The Man Behind the Name: **https://www.spacetelescope.org/about/history/the_man_behind_the_name/ (https://www.spacetelescope.org/about/history/the_man_behind_the_name/)** . Concise biography from the people at the Hubble Space Telescope.

Edwin Hubble: **http://apod.nasa.gov/diamond_jubilee/d_1996/sandage_hubble.html (http://apod.nasa.gov/diamond_jubilee/d_1996/sandage_hubble.html)** . An article on the life and work of Hubble by his student and successor, Allan Sandage. A bit technical in places, but giving a real picture of the man and the science.

NASA Science: Introduction to Galaxies: **http://science.nasa.gov/astrophysics/focus-areas/what-are-galaxies/ (http://science.nasa.gov/astrophysics/focus-areas/what-are-galaxies/)** . A brief overview with links to other pages, and recent Hubble Space Telescope discoveries.

National Optical Astronomy Observatories Gallery of Galaxies: **https://www.noao.edu/image_gallery/galaxies.html (https://www.noao.edu/image_gallery/galaxies.html)** . A collection of images and information about galaxies and galaxy groups of different types. Another impressive archive can be found at the European Southern Observatory site: **https://www.eso.org/public/images/archive/category/galaxies/ (https://www.eso.org/public/images/archive/category/galaxies/)** .

Sloan Digital Sky Survey: Introduction to Galaxies: **http://skyserver.sdss.org/dr1/en/astro/galaxies/galaxies.asp (http://skyserver.sdss.org/dr1/en/astro/galaxies/galaxies.asp)** . Another brief overview.

Universe Expansion: **http://hubblesite.org/newscenter/archive/releases/1999/19 (http://hubblesite.org/newscenter/archive/releases/1999/19)** . The background material here provides a nice chronology of how we discovered and measured the expansion of the universe.

Videos

Edwin Hubble (Hubblecast Episode 89): **http://www.spacetelescope.org/videos/hubblecast89a/ (http://www.spacetelescope.org/videos/hubblecast89a/)** . (5:59).

Galaxies: **https://www.youtube.com/watch?v=I82ADyJC7wE (https://www.youtube.com/watch?v=I82ADyJC7wE)** . An introduction.

Hubble's Views of the Deep Universe: **https://www.youtube.com/watch?v=argR2U15w-M (https://www.youtube.com/watch?v=argR2U15w-M)** . A 2015 public talk by Brandon Lawton of the Space Telescope Science Institute about galaxies and beyond (1:26:20).

CHAPTER 27 REVIEW

KEY TERMS

elliptical galaxy a galaxy whose shape is an ellipse and that contains no conspicuous interstellar material

Hubble constant a constant of proportionality in the law relating the velocities of remote galaxies to their distances

Hubble's law a rule that the radial velocities of remote galaxies are proportional to their distances from us

irregular galaxy a galaxy without any clear symmetry or pattern; neither a spiral nor an elliptical galaxy

mass-to-light ratio the ratio of the total mass of a galaxy to its total luminosity, usually expressed in units of solar mass and solar luminosity; the mass-to-light ratio gives a rough indication of the types of stars contained within a galaxy and whether or not substantial quantities of dark matter are present

redshift when lines in the spectra are displaced toward longer wavelengths (toward the red end of the visible spectrum)

spiral galaxy a flattened, rotating galaxy with pinwheel-like arms of interstellar material and young stars, winding out from its central bulge

type Ia supernova a supernova formed by the explosion of a white dwarf in a binary system and reach a luminosity of about $4.5 \times 10^9\ L_{\text{Sun}}$; can be used to determine distances to galaxies on a large scale

KEY EQUATIONS

Hubble's law $V = H_0 \times d$

SUMMARY

27.1 The Discovery of Galaxies

- Faint star clusters, clouds of glowing gas, and galaxies all appeared as faint patches of light (or nebulae) in the telescopes available at the beginning of the twentieth century.
- It was only when Hubble measured the distance to the Andromeda galaxy using cepheid variables with the giant 2.5-meter reflector on Mount Wilson in 1924 that the existence of other galaxies similar to the Milky Way in size and content was established.

27.2 Types of Galaxies

- The majority of bright galaxies are either spirals or ellipticals.
- Spiral galaxies contain both old and young stars, as well as interstellar matter, and have typical masses in the range of 10^9 to $10^{12}\ M_{\text{Sun}}$.
- Our own Galaxy is a large spiral.
- Ellipticals are spheroidal or slightly elongated systems that consist almost entirely of old stars, with very little interstellar matter.
- Elliptical galaxies range in size from giants, more massive than any spiral, down to dwarfs, with masses of only about $10^6\ M_{\text{Sun}}$.
- Dwarf ellipticals are probably the most common type of galaxy in the nearby universe.
- A small percentage of galaxies with more disorganized shapes are classified as irregulars.
- Galaxies may change their appearance over time due to collisions with other galaxies or by a change in the rate of star formation.

27.3 Properties of Galaxies

- The masses of spiral galaxies are determined from measurements of their rates of rotation.

- The masses of elliptical galaxies are estimated from analyses of the motions of the stars within them.
- Galaxies can be characterized by their mass-to-light ratios.
- The luminous parts of galaxies with active star formation typically have mass-to-light ratios in the range of 1 to 10; the luminous parts of elliptical galaxies, which contain only old stars, typically have mass-to-light ratios of 10 to 20.
- The mass-to-light ratios of whole galaxies, including their outer regions, are as high as 100, indicating the presence of a great deal of dark matter.

27.4 The Extragalactic Distance Scale

- Astronomers determine the distances to galaxies using a variety of methods, including the period-luminosity relationship for cepheid variables; objects such as type Ia supernovae, which appear to be standard bulbs; and the Tully-Fisher relation, which connects the line broadening of 21-cm radiation to the luminosity of spiral galaxies.
- Each method has limitations in terms of its precision, the kinds of galaxies with which it can be used, and the range of distances over which it can be applied.

27.5 The Expanding Universe

- The universe is expanding.
- Observations show that the spectral lines of distant galaxies are redshifted, and that their recession velocities are proportional to their distances from us, a relationship known as Hubble's law.
- The rate of recession, called the Hubble constant, is approximately 22 kilometers per second per million light-years.
- We are not at the center of this expansion: an observer in any other galaxy would see the same pattern of expansion that we do.
- The expansion described by Hubble's law is best understood as a stretching of space.

CONCEPTUAL QUESTIONS

27.1 The Discovery of Galaxies

2. Why did it take so long for the existence of other galaxies to be established?

3. Why can we not determine distances to galaxies by the same method used to measure the parallaxes of stars?

27.2 Types of Galaxies

4. Describe the main distinguishing features of spiral, elliptical, and irregular galaxies.

5. If we now realize dwarf ellipticals are the most common type of galaxy, why did they escape our notice for so long?

6. Where might the gas and dust (if any) in an elliptical galaxy come from?

7. Which is redder—a spiral galaxy or an elliptical galaxy?

27.3 Properties of Galaxies

8. Explain what the mass-to-light ratio is and why it is smaller in spiral galaxies with regions of star formation than in elliptical galaxies.

9. Does an elliptical galaxy rotate like a spiral galaxy? Explain.

10. Why does the disk of a spiral galaxy appear dark when viewed edge on?

11. What causes the largest mass-to-light ratio: gas and dust, dark matter, or stars that have burnt out?

12. When comparing two isolated spiral galaxies that have the same apparent brightness, but rotate at different rates, what can you say about their relative luminosity?

13. What does it mean if one elliptical galaxy has broader spectrum lines than another elliptical galaxy?

14. Based on your analysis of galaxies in Table 27.1, is there a correlation between the population of stars and the quantity of gas or dust? Explain why this might be.

15. Can a higher mass-to-light ratio mean that there is gas and dust present in the system that is being analyzed?

27.4 The Extragalactic Distance Scale

16. What are the two best ways to measure the distance to a nearby spiral galaxy, and how would it be measured?

17. What are the two best ways to measure the distance to a distant, isolated spiral galaxy, and how would it be measured?

18. What is the most useful standard bulb method for determining distances to galaxies?

27.5 The Expanding Universe

19. Why is Hubble's law considered one of the most important discoveries in the history of astronomy?

20. What does it mean to say that the universe is expanding? What is expanding? For example, is your astronomy classroom expanding? Is the solar system? Why or why not?

21. Was Hubble's original estimate of the distance to the Andromeda galaxy correct? Explain.

22. If all distant galaxies are expanding away from us, does this mean we're at the center of the universe?

23. Is the Hubble constant actually constant?

24. Suppose the stars in an elliptical galaxy all formed within a few million years shortly after the universe began. Suppose these stars have a range of masses, just as the stars in our own galaxy do. How would the color of the elliptical change over the next several billion years? How would its luminosity change? Why?

25. Starting with the determination of the size of Earth, outline a sequence of steps necessary to obtain the distance to a remote cluster of galaxies. (Hint: Review the chapter on **Celestial Distances**.)

26. Suppose the Milky Way Galaxy were truly isolated and that no other galaxies existed within 100 million light-years. Suppose that galaxies were observed in larger numbers at distances greater than 100 million light-years. Why would it be more difficult to determine accurate distances to those galaxies than if there were also galaxies relatively close by?

27. Suppose you were Hubble and Humason, working on the distances and Doppler shifts of the galaxies. What sorts of things would you have to do to convince yourself (and others) that the relationship you were seeing between the two quantities was a real feature of the behavior of the universe? (For example, would data from two galaxies be enough to demonstrate Hubble's law? Would data from just the nearest galaxies—in what astronomers call "the Local Group"—suffice?)

PROBLEMS

27.3 Properties of Galaxies

28. Calculate the mass-to-light ratio for a globular cluster with a luminosity of 10^6 L_{Sun} and 10^5 stars. (Assume that the average mass of a star in such a cluster is 1 M_{Sun}.)

29. Calculate the mass-to-light ratio for a luminous star of 100 M_{Sun} having the luminosity of 10^6 L_{Sun}.

27.4 The Extragalactic Distance Scale

30. A standard, bright spiral galaxy is discovered in a distant cluster. Its apparent magnitude is 15.6. How far away is this galaxy from Earth?

27.5 The Expanding Universe

31. According to Hubble's law, what is the recessional velocity of a galaxy that is 10^8 light-years away from us? (Assume a Hubble constant of 22 km/s per million light-years.)

32. A cluster of galaxies is observed to have a recessional velocity of 60,000 km/s. Find the distance to the cluster. (Assume a Hubble constant of 22 km/s per million light-years.)

33. Suppose we could measure the distance to a galaxy using one of the distance techniques listed in **Table 27.2** and it turns out to be 200 million light-years. The galaxy's redshift tells us its recessional velocity is 5000 km/s. What is the Hubble constant?

28 | THE EVOLUTION AND DISTRIBUTION OF GALAXIES

Chapter Outline

Introduction

Figure 28.1 The deepest picture of the sky in visible light (left) shows huge numbers of galaxies in a tiny patch of sky, only 1/100 the area of the full Moon. In contrast, the deepest picture of the sky taken in X-rays (right) shows large numbers of point-like quasars, which astronomers have shown are supermassive black holes at the very centers of galaxies. (credit left: modification of work by NASA, ESA, H. Teplitz and M. Rafelski (IPAC/Caltech), A. Koekemoer (STScI), R. Windhorst (Arizona State University), and Z. Levay (STScI); credit right: modification of work by ESO/Mario Nonino, Piero Rosati, ESO GOODS Team)

During the first half of the twentieth century, astronomers viewed the universe of galaxies as a mostly peaceful place. They assumed that galaxies formed billions of years ago and then evolved slowly as the populations of stars within them formed, aged, and died. That placid picture completely changed in the last few decades of the twentieth century.

Today, astronomers can see that the universe is often shaped by violent events, including cataclysmic explosions of supernovae, collisions of whole galaxies, and the tremendous outpouring of energy as matter interacts in the environment surrounding very massive black holes. The key event that began to change our view of the universe was the discovery of a new class of objects: quasars.

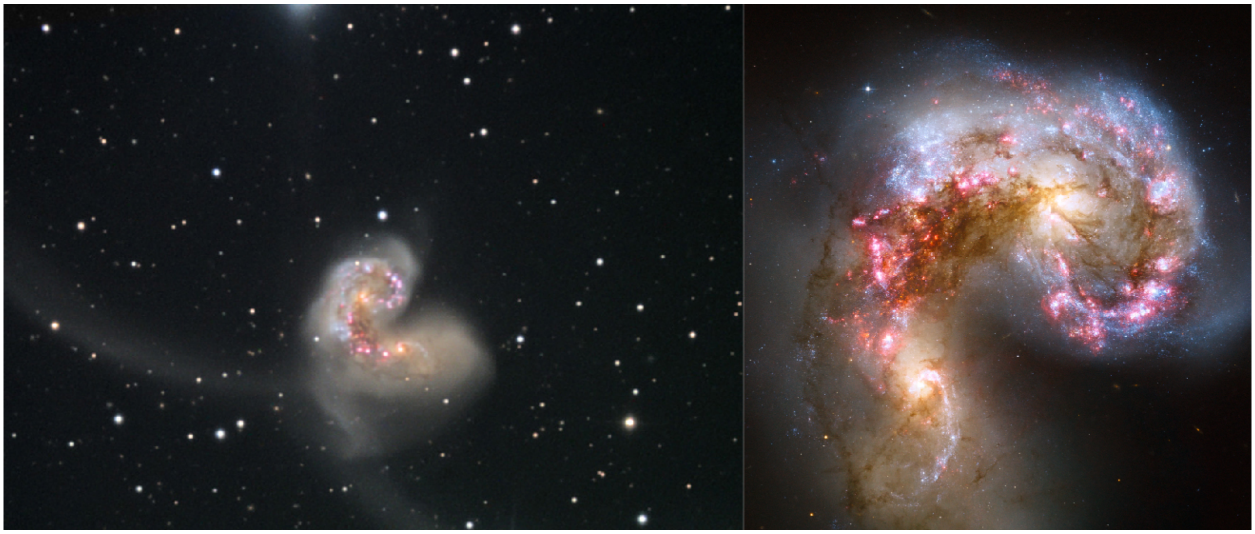


Figure 28.2 Collisions and mergers of galaxies strongly influence their evolution. On the left is a ground-based image of two colliding galaxies (NCG 4038 and 4039), sometimes nicknamed the Antennae galaxies. The long, luminous tails are material torn out of the galaxies by tidal forces during the collision. The right image shows the inner regions of these two galaxies, as taken by the Hubble Space Telescope. The cores of the twin galaxies are the orange blobs to the lower left and upper right of the center of the image. Note the dark lanes of dust crossing in front of the bright regions. The bright pink and blue star clusters are the result of a burst of star formation stimulated by the collision. (credit left: modification of work by Bob and Bill Twardy/Adam Block/ NOAO/AURA/NSF; credit right: modification of work by NASA, ESA, and the Hubble Heritage Team (STScI/AURA)-ESA/ Hubble Collaboration)

How and when did galaxies like our Milky Way form? Which formed first: stars or galaxies? Can we see direct evidence of the changes galaxies undergo over their lifetimes? If so, what determines whether a galaxy will "grow up" to be spiral or elliptical? And what is the role of "nature versus nurture"? That is to say, how much of a galaxy's development is determined by what it looks like when it is born and how much is influenced by its environment?

Astronomers today have the tools needed to explore the universe almost back to the time it began. The huge new telescopes and sensitive detectors built in the last decades make it possible to obtain both images and spectra of galaxies so distant that their light has traveled to reach us for more than 13 billion years—more than 90% of the way back to the Big Bang: we can use the finite speed of light and the vast size of the universe as a cosmic time machine to peer back and observe how galaxies formed and evolved over time. Studying galaxies so far away in any detail is always a major challenge, largely because their distance makes them appear very faint. However, today's large telescopes on the ground and in space are finally making such a task possible.

28.1 | Quasars

Learning Objectives

By the end of this section, you will be able to:

- Describe how quasars were discovered
- Explain how astronomers determined that quasars are at the distances implied by their redshifts
- Justify the statement that the enormous amount of energy produced by quasars is generated in a very small volume of space

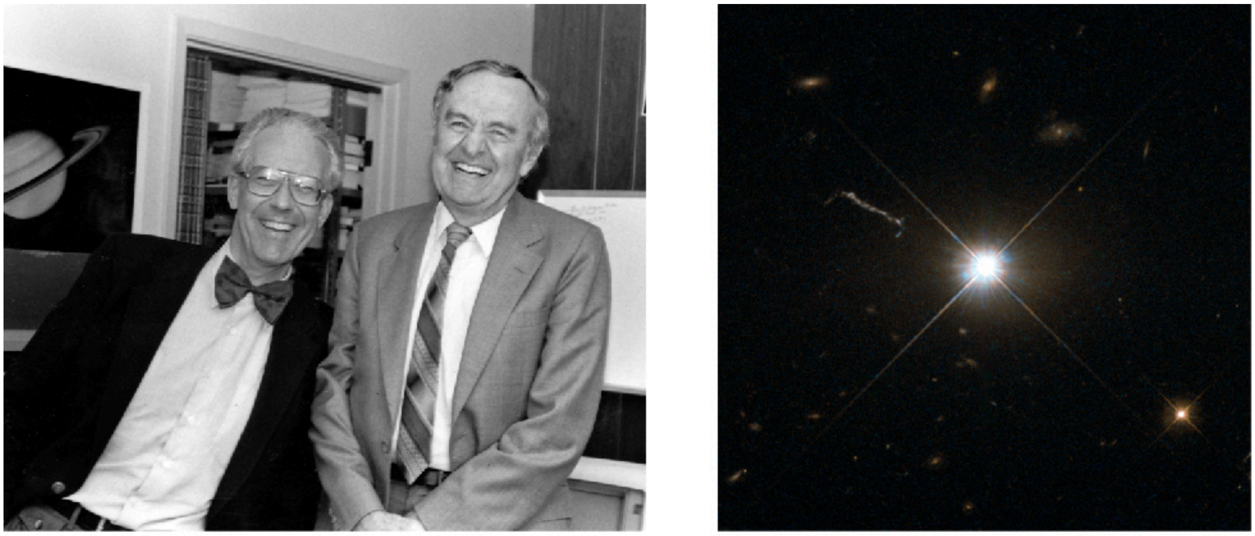
The name " quasars" started out as short for "quasi-stellar radio sources" (here "quasi-stellar" means "sort of like stars"). The discovery of radio sources that appeared point-like, just like stars, came with the use of surplus World War II radar equipment in the 1950s. Although few astronomers would have predicted it, the sky turned out to be full of strong sources of radio waves. As they improved the images that their new radio telescopes could make, scientists discovered that some radio sources were in the same location as faint blue "stars." No known type of star in our Galaxy emits such powerful radio radiation. What then were these “quasi-stellar radio sources”?

Redshifts: The Key to Quasars

The answer came when astronomers obtained visible-light spectra of two of those faint “blue stars” that were strong sources of radio waves (**Figure 28.3**). Spectra of these radio “stars” only deepened the mystery: they had emission lines, but astronomers at first could not identify them with any known substance. By the 1960s, astronomers had a century of experience in identifying elements and compounds in the spectra of stars. Elaborate tables had been published showing the lines that each element would produce under a wide range of conditions. A “star” with unidentifiable lines in the ordinary visible light spectrum had to be something completely new.

Figure 28.3 The arrow in this image marks the quasar known by its catalog number, PKS 1117-248. Note that nothing in this image distinguishes the quasar from an ordinary star. Its spectrum, however, shows that it is moving away from us at a speed of 36% the speed of light, or 67,000 miles per second. In contrast, the maximum speed observed for any star is only a few hundred miles per second. (credit: modification of work by WIYN Telescope, Kitt Peak National Observatory, NOAO)

In 1963 at Caltech’s Palomar Observatory, Maarten Schmidt (**Figure 28.4**) was puzzling over the spectrum of one of the radio stars, which was named 3C 273 because it was the 273rd entry in the third Cambridge catalog of radio sources (part (b) of **Figure 28.4**). There were strong emission lines in the spectrum, and Schmidt recognized that they had the same spacing between them as the Balmer lines of hydrogen (see **Using Spectra to Measure Stellar Composition and Motion**). But the lines in 3C 273 were shifted far to the red of the wavelengths at which the Balmer lines are normally located. Indeed, these lines were at such long wavelengths that if the redshifts were attributed to the Doppler effect, 3C 273 was receding from us at a speed of 45,000 kilometers per second, or about 15% the speed of light! Since stars don’t show Doppler shifts this large, no one had thought of considering high redshifts to be the cause of the strange spectra.



(a) (b)

Figure 28.4 (a) Maarten Schmidt (left), who solved the puzzle of the quasar spectra in 1963, shares a joke in this 1987 photo with Allan Sandage, who took the first spectrum of a quasar. Sandage was also instrumental in measuring the value of Hubble's constant. (b) This is the first quasar for which a redshift was measured. The redshift showed that the light from it took about 2.5 billion years to reach us. Despite this great distance, it is still one of the quasars closest to the Milky Way Galaxy. Note also the faint streak going toward the upper left from the quasar. Some quasars, like 3C 273, eject super-fast jets of material. The jet from 3C 273 is about 200,000 light-years long. (credit a: modification of work by Andrew Fraknoi; credit b: modification of work by ESA/Hubble/NASA)

The puzzling emission lines in other star-like radio sources were then reexamined to see if they, too, might be well-known lines with large redshifts. This proved to be the case, but the other objects were found to be receding from us at even greater speeds. Their astounding speeds showed that the radio "stars" could not possibly be stars in our own Galaxy. Any true star moving at more than a few hundred kilometers per second would be able to overcome the gravitational pull of the Galaxy and completely escape from it. (As we shall see later in this chapter, astronomers eventually discovered that there was also more to these "stars" than just a point of light.)

It turns out that these high-velocity objects only look like stars because they are compact and very far away. Later, astronomers discovered objects with large redshifts that appear star-like but have no radio emission. Observations also showed that quasars were bright in the infrared and X-ray bands too, and not all these X-ray or infrared-bright quasars could be seen in either the radio or the visible-light bands of the spectrum. Today, all these objects are referred to as *quasi-stellar objects* (*QSOs*), or, as they are more popularly known, **quasars**. (The name was also soon appropriated by a manufacturer of home electronics.)

Read **an interview (https://openstax.org/l/30SchmidtIntv)** with Maarten Schmidt on the fiftieth anniversary of his insight about the spectrum of quasars and their redshifts.

Over a million quasars have now been discovered, and spectra are available for over a hundred thousand. All these spectra show redshifts, none show blueshifts, and their redshifts can be very large. Yet in a photo they look just like stars (**Figure 28.5**).

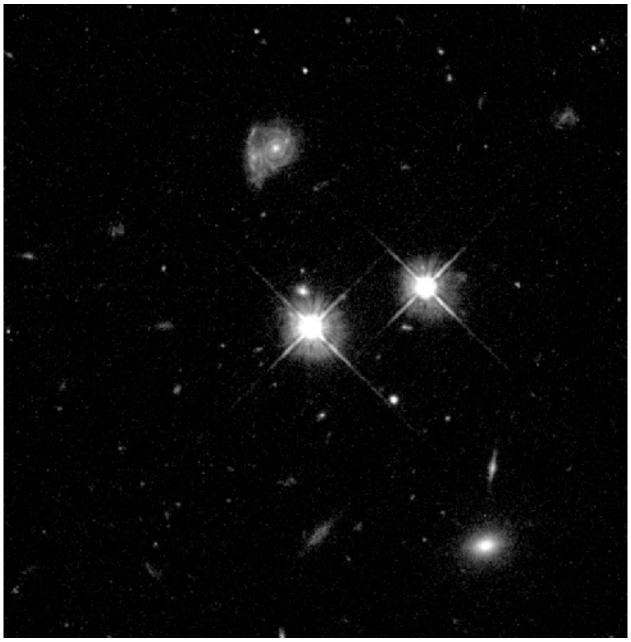


Figure 28.5 One of these two bright “stars” in the middle is in our Galaxy, while the other is a quasar 9 billion light-years away. From this picture alone, there’s no way to say which is which. (The quasar is the one in the center of the picture.) (credit: Charles Steidel (CIT)/NASA/ESA)

In the record-holding quasars, the first Lyman series line of hydrogen, with a laboratory wavelength of 121.5 nanometers in the ultraviolet portion of the spectrum, is shifted all the way through the visible region to the infrared. At such high redshifts, the simple formula for converting a **Doppler shift** to speed must be modified to take into account the effects of the theory of relativity. If we apply the relativistic form of the Doppler shift formula, we find that these redshifts correspond to velocities of about 96% of the speed of light.

Example 28.1

Recession Speed of a Quasar

The formula for the Doppler shift (or simply the "redshift"), which astronomers denote by the letter *z*, is

$$z = \frac{\Delta\lambda}{\lambda} = \frac{v}{c}$$

where λ is the wavelength emitted by a source of radiation that is not moving, Δλ is the difference between that wavelength and the wavelength we measure, *v* is the speed with which the source moves away, and *c* (as usual) is the speed of light.

A line in the spectrum of a galaxy is at a wavelength of 393 nanometers (nm, or 10^{-9} m) when the source is at rest. Let’s say the line is measured to be longer than this value (redshifted) by 7.86 nm. Then its redshift $z = \frac{7.86 \text{ nm}}{393 \text{ nm}} = 0.02$, so its speed away from us is 2% of the speed of light $\left(\frac{v}{c} = 0.02\right)$.

This formula is fine for galaxies that are relatively nearby and are moving away from us slowly in the expansion of the universe. But the quasars and distant galaxies we discuss in this chapter are moving away at speeds close to the speed of light. In that case, converting a Doppler shift (redshift) to a distance must include the effects of the special theory of relativity, which explains how measurements of space and time change when we see things moving at high speeds. The details of how this is done are beyond the level of this text, but we can share with you the relativistic formula for the Doppler shift:

$$\frac{v}{c} = \frac{(z+1)^2 - 1}{(z+1)^2 + 1}$$

Let's do an example. Suppose a distant quasar has a redshift of 5. At what fraction of the speed of light is the quasar moving away?

Solution

We calculate the following:

$$\frac{v}{c} = \frac{(5+1)^2 - 1}{(5+1)^2 + 1} = \frac{36-1}{36+1} = \frac{35}{37} = 0.946$$

The quasar is thus receding from us at about 95% the speed of light.

28.1 Several lines of hydrogen absorption in the visible spectrum have rest wavelengths of 410 nm, 434 nm, 486 nm, and 656 nm. In a spectrum of a distant galaxy, these same lines are observed to have wavelengths of 492 nm, 521 nm, 583 nm, and 787 nm respectively. What is the redshift of this galaxy? What is the recession speed of this galaxy?

Quasars Obey the Hubble Law

The first question astronomers asked was whether quasars obeyed the Hubble law and were really at the large distances implied by their redshifts. If they did not obey the rule that large redshift means large distance, then they could be much closer, and their luminosity could be a lot less. One straightforward way to show that quasars had to obey the Hubble law was to demonstrate that they were actually part of galaxies, and that their redshift was the same as the galaxy that hosted them. Since ordinary galaxies *do* obey the Hubble law, anything within them would be subject to the same rules.

Observations with the Hubble Space Telescope provided the strongest evidence showing that quasars are located at the centers of galaxies. Hints that this is true had been obtained with ground-based telescopes, but space observations were required to make a convincing case. The reason is that quasars can outshine their entire galaxies by factors of 10 to 100 or even more. When this light passes through Earth's atmosphere, it is blurred by turbulence and drowns out the faint light from the surrounding galaxy—much as the bright headlights from an oncoming car at night make it difficult to see anything close by.

The Hubble Space Telescope, however, is not affected by atmospheric turbulence and can detect the faint glow from some of the galaxies that host quasars (**Figure 28.6**). Quasars have been found in the cores of both spiral and elliptical galaxies, and each quasar has the same redshift as its host galaxy. A wide range of studies with the Hubble Space Telescope now clearly demonstrate that quasars are indeed far away. If so, they must be producing a truly impressive amount of energy to be detectable as points of light that are much brighter than their galaxy. Interestingly, many quasar host galaxies are found to be involved in a collision with a second galaxy, providing, as we shall see, an important clue to the source of their prodigious energy output.

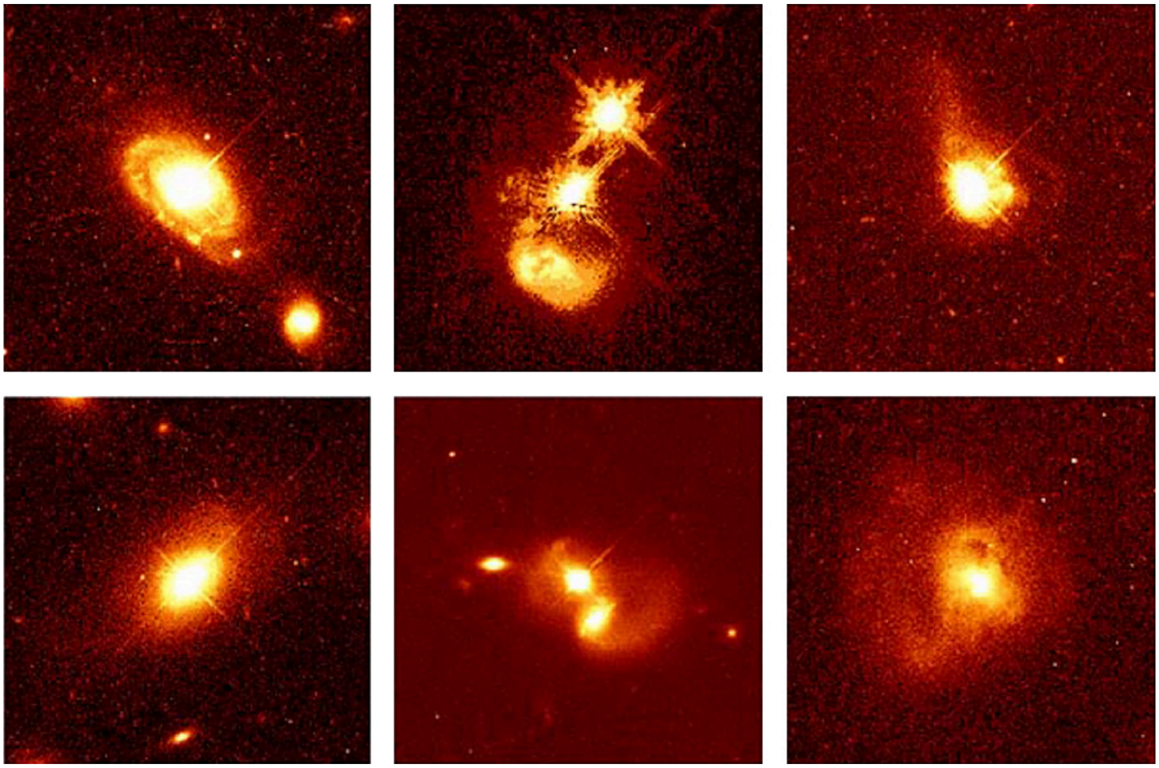


Figure 28.6 The Hubble Space Telescope reveals the much fainter "host" galaxies around quasars. The top left image shows a quasar that lies at the heart of a spiral galaxy 1.4 billion light-years from Earth. The bottom left image shows a quasar that lies at the center of an elliptical galaxy some 1.5 billion light-years from us. The middle images show remote pairs of interacting galaxies, one of which harbors a quasar. Each of the right images shows long tails of gas and dust streaming away from a galaxy that contains a quasar. Such tails are produced when one galaxy collides with another. (credit: modification of work by John Bahcall, Mike Disney, NASA)

The Size of the Energy Source

Given their large distances, quasars have to be extremely luminous to be visible to us at all—far brighter than any normal galaxy. In visible light alone, most are far more energetic than the brightest elliptical galaxies. But, as we saw, quasars also emit energy at X-ray and ultraviolet wavelengths, and some are radio sources as well. When all their radiation is added together, some QSOs have total luminosities as large as a hundred trillion Suns (10^{14} L_{Sun}), which is 10 to 100 times the brightness of luminous elliptical galaxies.

Finding a mechanism to produce the large amount of energy emitted by a quasar would be difficult under any circumstances. But there is an additional problem. When astronomers began monitoring quasars carefully, they found that some vary in luminosity on time scales of months, weeks, or even, in some cases, days. This variation is irregular and can change the brightness of a quasar by a few tens of percent in both its visible light and radio output.

Think about what such a change in luminosity means. A quasar at its dimmest is still more brilliant than any normal galaxy. Now imagine that the brightness increases by 30% in a few weeks. Whatever mechanism is responsible must be able to release new energy at rates that stagger our imaginations. The most dramatic changes in quasar brightness are equivalent to the energy released by 100,000 billion Suns. To produce this much energy we would have to convert the total mass of about ten Earths into energy every minute.

Moreover, because the fluctuations occur in such short times, the part of a quasar that is varying must be smaller than the distance light travels in the time it takes the variation to occur—typically a few months. To see why this must be so, let's consider a cluster of stars 10 light-years in diameter at a very large distance from Earth (see **Figure 28.7**, in which Earth is off to the right). Suppose every star in this cluster somehow brightens simultaneously and remains bright. When the light from this event arrives at Earth, we would first see the brighter light from stars on the near side; 5 years later we would see increased light from stars at the center. Ten years would pass before we detected more light from stars on the far side.

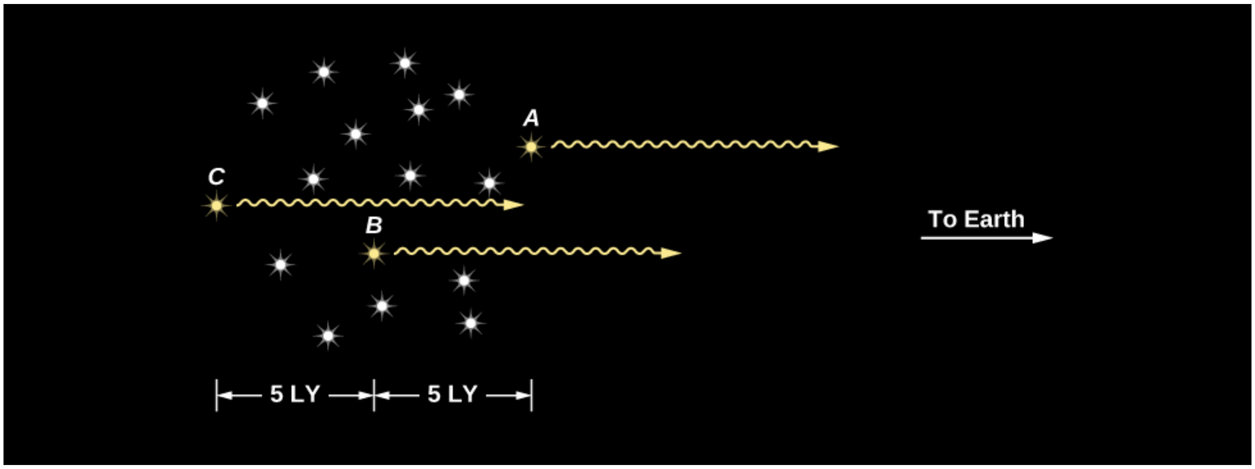


Figure 28.7 This diagram shows why light variations from a large region in space appear to last for an extended period of time as viewed from Earth. Suppose all the stars in this cluster, which is 10 light-years across, brighten simultaneously and instantaneously. From Earth, star *A* will appear to brighten 5 years before star *B*, which in turn will appear to brighten 5 years earlier than star *C*. It will take 10 years for an Earth observer to get the full effect of the brightening.

Even though all stars in the cluster brightened at the same time, the fact that the cluster is 10 light-years wide means that 10 years must elapse before the increased light from every part of the cluster reaches us. From Earth we would see the cluster get brighter and brighter, as light from more and more stars began to reach us. Not until 10 years after the brightening began would we see the cluster reach maximum brightness. In other words, if an extended object suddenly flares up, it will seem to brighten over a period of time equal to the time it takes light to travel across the object from its far side.

We can apply this idea to brightness changes in quasars to estimate their diameters. Because quasars typically vary (get brighter and dimmer) over periods of a few months, the region where the energy is generated can be no larger than a few light-months across. If it were larger, it would take longer than a few months for the light from the far side to reach us.

How large is a region of a few light-months? Pluto, usually the outermost (dwarf) planet in our solar system, is about 5.5 light-hours from us, while the nearest star is 4 light-years away. Clearly a region a few light months across is tiny relative to the size of the entire Galaxy. And some quasars vary even more rapidly, which means their energy is generated in an even smaller region. Whatever mechanism powers the quasars must be able to generate more energy than that produced by an entire galaxy in a volume of space that, in some cases, is not much larger than our solar system.

Earlier Evidence

Even before the discovery of quasars, there had been hints that something very strange was going on in the centers of at least some galaxies. Back in 1918, American astronomer Heber Curtis used the large Lick Observatory telescope to photograph the galaxy Messier 87 in the constellation Virgo. On that photograph, he saw what we now call a jet coming from the center, or nucleus, of the galaxy (**Figure 28.8**). This jet literally and figuratively pointed to some strange activity going on in that galaxy nucleus. But he had no idea what it was. No one else knew what to do with this space oddity either.

The random factoid that such a central jet existed lay around for a quarter century, until Carl Seyfert, a young astronomer at Mount Wilson Observatory, also in California, found half a dozen galaxies with extremely bright nuclei that were almost stellar, rather than fuzzy in appearance like most galaxy nuclei. Using spectroscopy, he found that these nuclei contain gas moving at up to two percent the speed of light. That may not sound like much, but it is 6 million miles per hour, and more than 10 times faster than the typical motions of stars in galaxies.

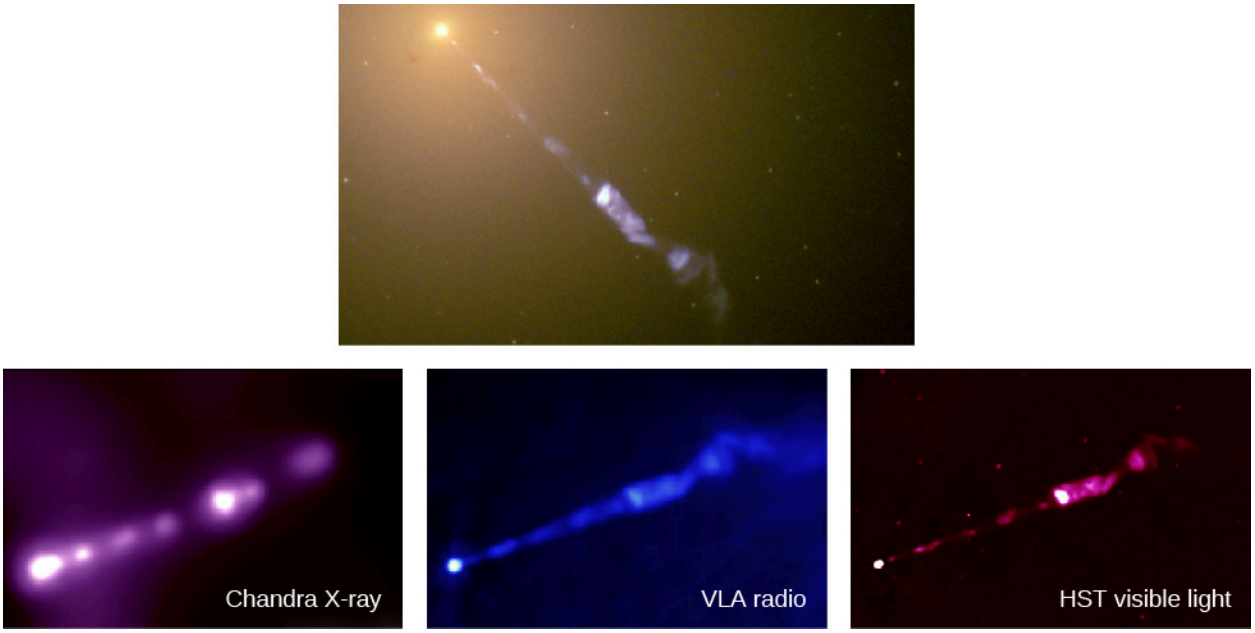


Figure 28.8 Streaming out like a cosmic searchlight from the center of the galaxy, M87 is one of nature’s most amazing phenomena, a huge jet of electrons and other particles traveling at nearly the speed of light. In this Hubble Space Telescope image, the blue of the jet contrasts with the yellow glow from the combined light of billions of unseen stars and yellow, point-like globular clusters that make up the galaxy (at the upper left). As we shall see later in this chapter, the jet, which is several thousand light-years long, originates in a disk of superheated gas swirling around a giant black hole at the center of M87. The light that we see is produced by electrons twisting along magnetic field lines in the jet, a process known as synchrotron radiation, which gives the jet its bluish tint. The jet in M87 can be observed in X-ray, radio, and visible light, as shown in the bottom three images. At the extreme left of each bottom image, we see the bright galactic nucleus harboring a supermassive black hole. (credit top: modification of work by NASA, The Hubble Heritage Team(STScI/AURA); credit bottom: modification of work by X-ray: H. Marshall (MIT), et al., CXC, NASA; Radio: F. Zhou, F. Owen (NRAO), J. Biretta (STScI); Optical: E. Perlman (UMBC), et al.)

After decades of study, astronomers identified many other strange objects beyond our Milky Way Galaxy; they populate a whole “zoo” of what are now called **active galaxies** or **active galactic nuclei (AGN)**. Astronomers first called them by many different names, depending on what sorts of observations discovered each category, but now we know that we are always looking at the same basic mechanism. What all these galaxies have in common is some activity in their nuclei that produces an enormous amount of energy in a very small volume of space. In the next section, we describe a model that explains all these galaxies with strong central activity—both the AGNs and the QSOs.

To see a jet for yourself, check out a **time-lapse video (https://openstax.org/l/30timelapsejet)** of the jet ejected from NGC 3862.

28.2 | Supermassive Black Holes

Learning Objectives

By the end of this section, you will be able to:

- Describe the characteristics common to all quasars
- Justify the claim that supermassive black holes are the source of the energy emitted by quasars (and AGNs)
- Explain how a quasar’s energy is produced

In order to find a common model for quasars (and their cousins, the AGNs), let’s first list the common characteristics we have been describing—and add some new ones:

- Quasars are hugely powerful, emitting more power in radiated light than all the stars in our Galaxy combined.
- Quasars are tiny, about the size of our solar system (to astronomers, that is really small!).
- Some quasars are observed to be shooting out pairs of straight jets at close to the speed of light, in a tight beam, to distances far beyond the galaxies they live in. These jets are themselves powerful sources of radio and gamma-ray radiation.
- Because quasars put out so much power from such a small region, they can't be powered by nuclear fusion the way stars are; they must use some process that is far more efficient.
- As we shall see later in this chapter, quasars were much more common when the universe was young than they are today. That means they must have been able to form in the first billion years or so after the universe began to expand.

The readers of this text are in a much better position than the astronomers who discovered quasars in the 1960s to guess what powers the quasars. That's because the key idea in solving the puzzle came from observations of the black holes. The discovery of the first stellar mass black hole in the binary system Cygnus X-1 was announced in 1971, several years after the discovery of quasars. Proof that there is a black hole at the center of our own Galaxy came even later. Back when astronomers first began trying to figure out what powered quasars, black holes were simply one of the more exotic predictions of the general theory of relativity that still waited to be connected to the real world.

It was only as proof of the existence of black holes accumulated over several decades that it became clearer that only supermassive black holes could account for all the observed properties of quasars and AGNs. As we saw in **The Milky Way Galaxy**, our own Galaxy has a black hole in its center, and the energy is emitted from a small central region. While our black hole doesn't have the mass or energy of the quasar black holes, the mechanism that powers them is similar. The evidence now shows that most—and probably all—elliptical galaxies and all spirals with nuclear bulges have black holes at their centers. The amount of energy emitted by material near the black hole depends on two things: the mass of the black hole and the amount of matter that is falling into it.

If a black hole with a billion Suns' worth of mass inside (10^9 M_{Sun}) accretes (gathers) even a relatively modest amount of additional material—say, about 10 M_{Sun} per year—then (as we shall see) it can, in the process, produce as much energy as a thousand normal galaxies. This is enough to account for the total energy of a quasar. If the mass of the black hole is smaller than a billion solar masses or the accretion rate is low, then the amount of energy emitted can be much smaller, as it is in the case of the Milky Way.

Watch a **video (https://openstax.org/l/30mataccrsupblh)** of an artist's impression of matter accreting around a supermassive black hole.

Observational Evidence for Black Holes

In order to prove that a black hole is present at the center of a galaxy, we must demonstrate that so much mass is crammed into so small a volume that no normal objects—massive stars or clusters of stars—could possibly account for it (just as we did for the black hole in the Milky Way). We already know from observations (discussed in **Black Holes**) that an accreting black hole is surrounded by a hot *accretion disk* with gas and dust that swirl around the black hole before it falls in.

If we assume that the energy emitted by quasars is also produced by a hot accretion disk, then, as we saw in the previous section, the size of the disk must be given by the time the quasar energy takes to vary. For quasars, the emission in visible light varies on typical time scales of 5 to 2000 days, limiting the size of the disk to that many light-days.

In the X-ray band, quasars vary even more rapidly, so the light travel time argument tells us that this more energetic radiation is generated in an even smaller region. Therefore, the mass around which the accretion disk is swirling must be confined to a space that is even smaller. If the quasar mechanism involves a great deal of mass, then the only astronomical object that can confine a lot of mass into a very small space is a black hole. In a few cases, it turns out that the X-rays are emitted from a region just a few times the size of the black hole event horizon.

The next challenge, then, is to "weigh" this central mass in a quasar. In the case of our own Galaxy, we used observations of the orbits of stars very close to the galactic center, along with Kepler's third law, to estimate the mass of the central black hole (**The Milky Way Galaxy**). In the case of distant galaxies, we cannot measure the orbits of individual stars, but we can measure the orbital speed of the gas in the rotating accretion disk. The Hubble Space Telescope is especially well suited to this task because it is above the blurring of Earth's atmosphere and can obtain spectra very close to the bright central regions of active galaxies. The Doppler effect is then used to measure radial velocities of the orbiting material and so derive the speed with which it moves around.

One of the first galaxies to be studied with the Hubble Space Telescope is our old favorite, the giant elliptical M87. Hubble

Space Telescope images showed that there is a disk of hot (10,000 K) gas swirling around the center of M87 (**Figure 28.9**). It was surprising to find hot gas in an elliptical galaxy because this type of galaxy is usually devoid of gas and dust. But the discovery was extremely useful for pinning down the existence of the black hole. Astronomers measured the Doppler shift of spectral lines emitted by this gas, found its speed of rotation, and then used the speed to derive the amount of mass inside the disk—applying Kepler’s third law.

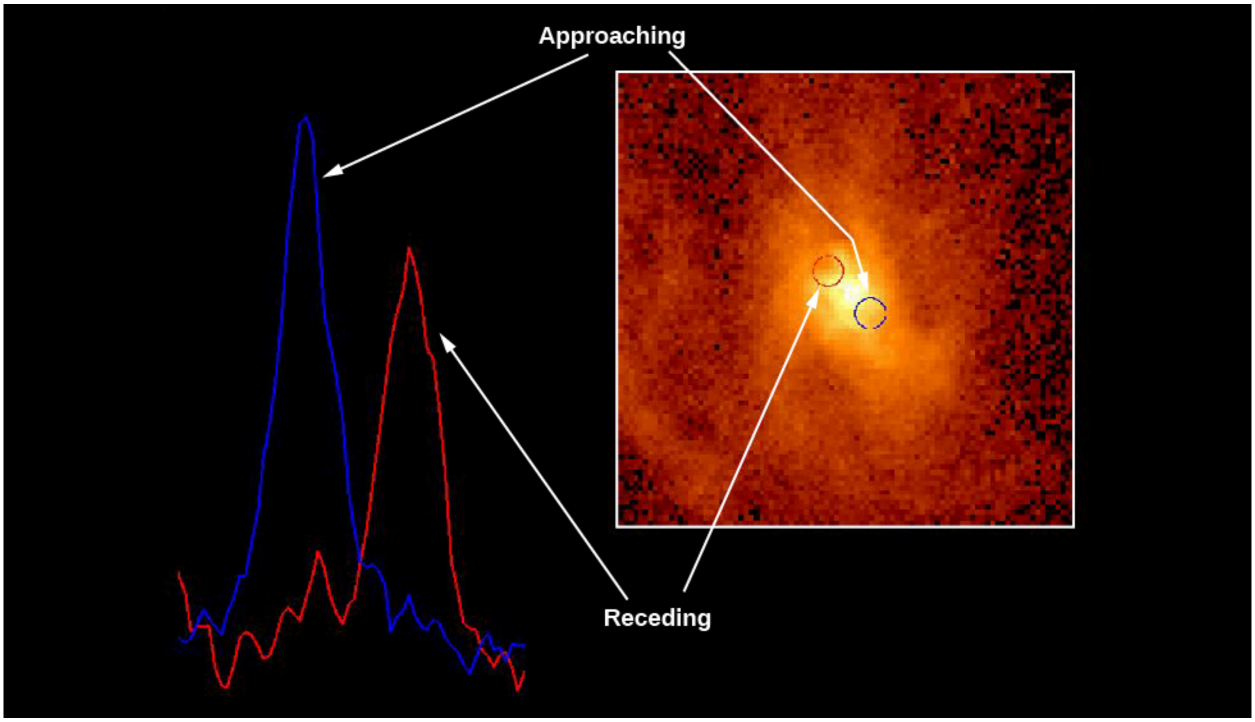


Figure 28.9 The disk of whirling gas at right was discovered at the center of the giant elliptical galaxy M87 with the Hubble Space Telescope. Observations made on opposite sides of the disk show that one side is approaching us (the spectral lines are blueshifted by the Doppler effect) while the other is receding (lines redshifted), a clear indication that the disk is rotating. The rotation speed is about 550 kilometers per second or 1.2 million miles per hour. Such a high rotation speed is evidence that there is a very massive black hole at the center of M87. (credit: modification of work by Holland Ford, STScI/JHU; Richard Harms, Linda Dressel, Ajay K. Kochhar, Applied Research Corp.; Zlatan Tsvetanov, Arthur Davidsen, Gerard Kriss, Johns Hopkins; Ralph Bohlin, George Hartig, STScI; Bruce Margon, University of Washington in Seattle; NASA)

Modern estimates show that there is a mass of at least 3.5 billion M_{Sun} concentrated in a tiny region at the very center of M87. So much mass in such a small volume of space must be a black hole. Let’s stop for a moment and take in this figure: a single black hole that has swallowed enough material to make 3.5 billion stars like the Sun. Few astronomical measurements have ever led to so mind-boggling a result. What a strange environment the neighborhood of such a supermassive black hole must be.

Another example is shown in **Figure 28.10**. Here, we see a disk of dust and gas that surrounds a 300-million-M_{Sun} black hole in the center of an elliptical galaxy. (The bright spot in the center is produced by the combined light of stars that have been pulled close together by the gravitational force of the black hole.) The mass of the black hole was again derived from measurements of the rotational speed of the disk. The gas in the disk is moving around at 155 kilometers per second at a distance of only 186 light-years from its center. Given the pull of the mass at the center, we expect that the whole dust disk should be swallowed by the black hole in several billion years.

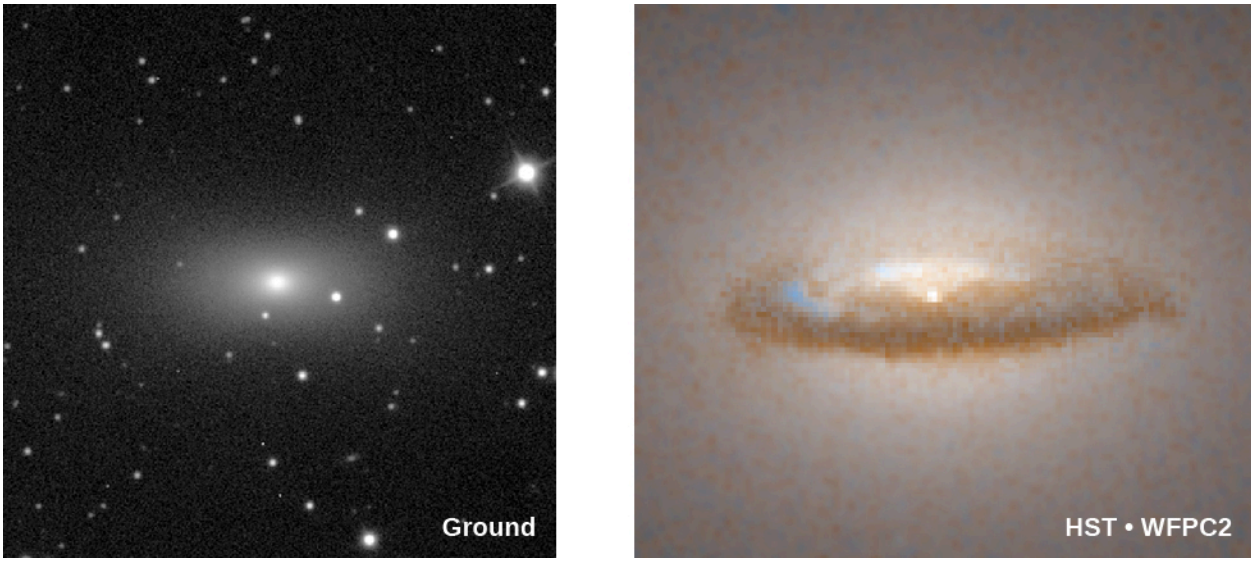


Figure 28.10 The ground-based image shows an elliptical galaxy called NGC 7052 located in the constellation of Vulpecula, almost 200 million light-years from Earth. At the galaxy's center (right) is a dust disk roughly 3700 light-years in diameter. The disk rotates like a giant merry-go-round: gas in the inner part (186 light-years from the center) whirls around at a speed of 155 kilometers per second (341,000 miles per hour). From these measurements and Kepler's third law, it is possible to estimate that the disk is orbiting around a central black hole with a mass of 300 million Suns. (credit: modification of work by Roeland P. van der Marel (STScI), Frank C. van den Bosch (University of Washington), NASA)

But do we *have* to accept black holes as the only explanation of what lies at the center of these galaxies? What else could we put in such a small space other than a giant black hole? The alternative is stars. But to explain the masses in the centers of galaxies without a black hole we need to put at least a million stars in a region the size of the solar system. To fit, they would have be only 2 star diameters apart. Collisions between stars would happen all the time. And these collisions would lead to mergers of stars, and very soon the one giant star that they form would collapse into a black hole. So there is really no escape: only a black hole can fit so much mass into so small a space.

As we saw earlier, observations now show that all the galaxies with a spherical concentration of stars—either elliptical galaxies or spiral galaxies with nuclear bulges (see the chapter on **Galaxies**)—harbor one of these giant black holes at their centers. Among them is our neighbor spiral galaxy, the Andromeda galaxy, M31. The masses of these central black holes range from a just under a million up to at least 30 billion times the mass of the Sun. Several black holes may be even more massive, but the mass estimates have large uncertainties and need verification. We call these black holes "supermassive" to distinguish them from the much smaller black holes that form when some stars die (see **The Deaths of Stars**). So far, the most massive black holes from stars—those detected through gravitational waves detected by LIGO—have masses only a little over 30 solar masses.

Energy Production around a Black Hole

By now, you may be willing to entertain the idea that huge black holes lurk at the centers of active galaxies. But we still need to answer the question of how such a black hole can account for one of the most powerful sources of energy in the universe. As we saw in **Black Holes**, a black hole itself can radiate no energy. Any energy we detect from it must come from material very close to the black hole, but not inside its event horizon.

In a galaxy, a central black hole (with its strong gravity) attracts matter—stars, dust, and gas—orbiting in the dense nuclear regions. This matter spirals in toward the spinning black hole and forms an accretion disk of material around it. As the material spirals ever closer to the black hole, it accelerates and becomes compressed, heating up to temperatures of millions of degrees. Such hot matter can radiate prodigious amounts of energy as it falls in toward the black hole.

To convince yourself that falling into a region with strong gravity can release a great deal of energy, imagine dropping a printed version of your astronomy textbook out the window of the ground floor of the library. It will land with a thud, and maybe give a surprised pigeon a nasty bump, but the energy released by its fall will not be very great. Now take the same book up to the fifteenth floor of a tall building and drop it from there. For anyone below, astronomy could suddenly become a deadly subject; when the book hits, it does so with a great deal of energy.

Dropping things from far away into the much stronger gravity of a black hole is much more effective in turning the energy released by infall into other forms of energy. Just as the falling book can heat up the air, shake the ground, or produce sound energy that can be heard some distance away, so the energy of material falling toward a black hole can be converted to significant amounts of electromagnetic radiation.

What a black hole has to work with is not textbooks but streams of infalling gas. If a dense blob of gas moves through a thin gas at high speed, it heats up as it slows by friction. As it slows down, kinetic (motion) energy is turned into heat energy. Just like a spaceship reentering the atmosphere (**Figure 28.11**), gas approaching a black hole heats up and glows where it meets other gas. But this gas, as it approaches the event horizon, reaches speeds of 10% the speed of light and more. It therefore gets far, far hotter than a spaceship, which reaches no more than about 1500 K. Indeed, gas near a supermassive black hole reaches a temperature of about 150,000 K, about 100 times hotter than a spaceship returning to Earth. It can even get so hot—millions of degrees—that it radiates X-rays.

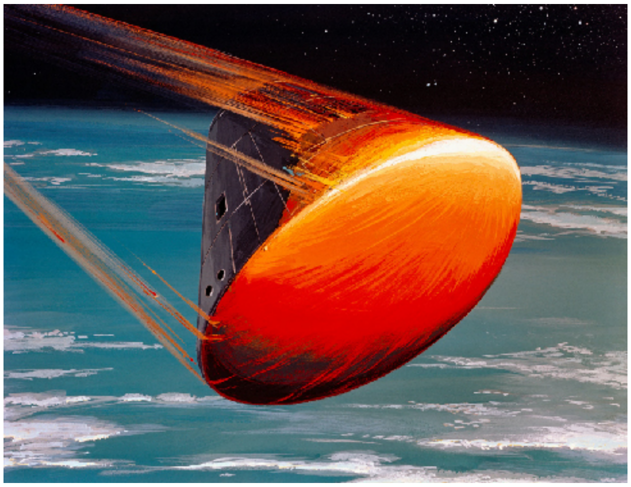


Figure 28.11 In this artist's impression, the rapid motion of a spacecraft (the Apollo mission reentry capsule) through the atmosphere compresses and heats the air ahead of it, which heats the spacecraft in turn until it glows red hot. Pushing on the air slows down the spacecraft, turning the kinetic energy of the spacecraft into heat. Fast-moving gas falling into a quasar heats up in a similar way. (credit: modification of work by NASA)

The amount of energy that can be liberated this way is enormous. Einstein showed that mass and energy are interchangeable with his famous formula $E = mc^2$ (see **Source of Sunshine: Nuclear Fusion!**). A hydrogen bomb releases just 1% of that energy, as does a star. Quasars are much more efficient than that. The energy released falling to the event horizon of a black hole can easily reach 10% or, in the extreme theoretical limit, 32%, of that energy. (Unlike the hydrogen atoms in a bomb or a star, the gas falling into the black hole is not actually losing mass from its atoms to free up the energy; the energy is produced just because the gas is falling closer and closer to the black hole.) This huge energy release explains how a tiny volume like the region around a black hole can release as much power as a whole galaxy. But to radiate all that energy, instead of just falling inside the event horizon with barely a peep, the hot gas must take the time to swirl around the star in the accretion disk and emit some of its energy.

Most black holes don't show any signs of quasar emission. We call them "quiescent." But, like sleeping dragons, they can be woken up by being roused with a fresh supply of gas. Our own Milky Way black hole is currently quiescent, but it may have been a quasar just a few million years ago (**Figure 28.12**). Two giant bubbles that extend 25,000 light-years above and below the galactic center are emitting gamma rays. Were these produced a few million years ago when a significant amount of matter fell into the black hole at the center of the galaxy? Astronomers are still working to understand what remarkable event might have formed these enormous bubbles.

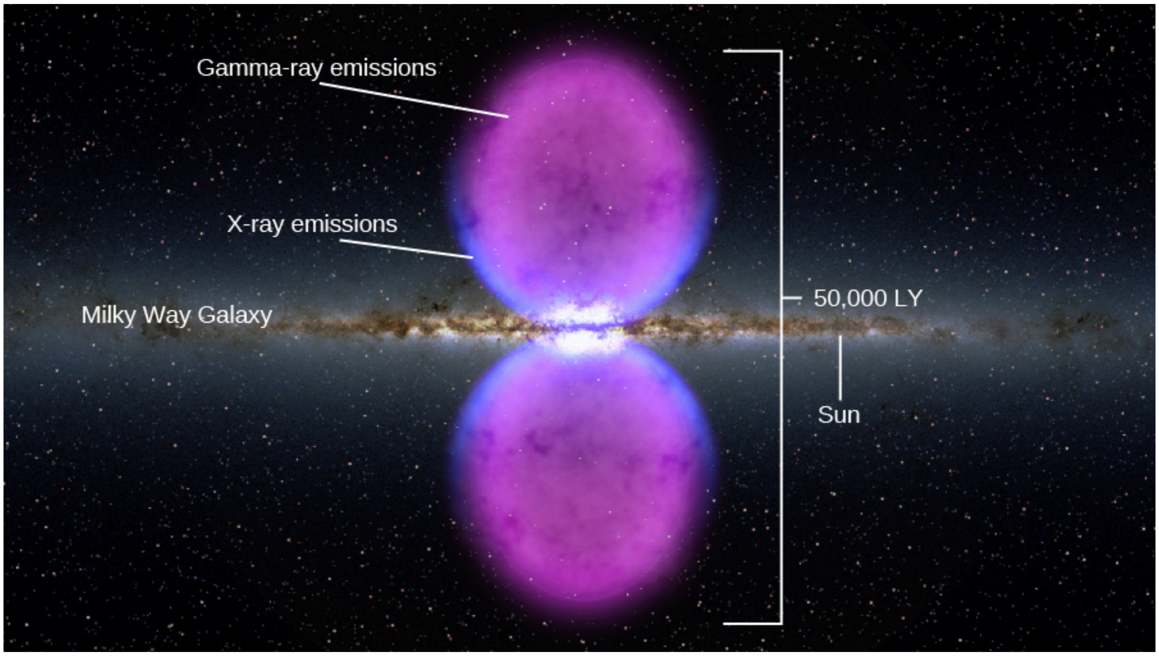


Figure 28.12 Giant bubbles shining in gamma-ray light lie above and below the center of the Milky Way Galaxy, as seen by the Fermi satellite. (The gamma-ray and X-ray image is superimposed on a visible-light image of the inner parts of our Galaxy.) The bubbles may be evidence that the supermassive black hole at the center of our Galaxy was a quasar a few million years ago. (credit: modification of work by NASA's Goddard Space Flight Center)

The physics required to account for the exact way in which the energy of infalling material is converted to radiation near a black hole is far more complicated than our simple discussion suggests. To understand what happens in the "rough and tumble" region around a massive black hole, astronomers and physicists must resort to computer simulations (and they require supercomputers, fast machines capable of awesome numbers of calculations per second). The details of these models are beyond the scope of our book, but they support the basic description presented here.

Radio Jets

So far, our model seems to explain the central energy source in quasars and active galaxies. But, as we have seen, there is more to quasars and other active galaxies than the point-like energy source. They can also have long jets that glow with radio waves, light, and sometimes even X-rays, and that extend far beyond the limits of the parent galaxy. Can we find a way for our black hole and its accretion disk to produce these jets of energetic particles as well?

Many different observations have now traced these jets to within 3 to 30 light-years of the parent quasar or galactic nucleus. While the black hole and accretion disk are typically smaller than 1 light-year, we nevertheless presume that if the jets come this close, they probably originate in the vicinity of the black hole. Another characteristic of the jets we need to explain is that they contain matter moving close to the speed of light.

Why are energetic electrons and other particles near a supermassive black hole ejected into jets, and often into two oppositely directed jets, rather than in all directions? Again, we must use theoretical models and supercomputer simulations of what happens when a lot of material whirls inward in a crowded black hole accretion disk. Although there is no agreement on exactly how jets form, it has become clear that any material escaping from the neighborhood of the black hole has an easier time doing so *perpendicular to* the disk.

In some ways, the inner regions of black hole accretion disks resemble a baby that is just learning to eat by herself. As much food as goes into the baby's mouth can sometimes wind up being spit out in various directions. In the same way, some of the material whirling inward toward a black hole finds itself under tremendous pressure and orbiting with tremendous speed. Under such conditions, simulations show that a significant amount of material can be flung outward—not back along the disk, where more material is crowding in, but above and below the disk. If the disk is thick (as it tends to be when a lot of material falls in quickly), it can channel the outrushing material into narrow beams perpendicular to the disk (**Figure 28.13**).

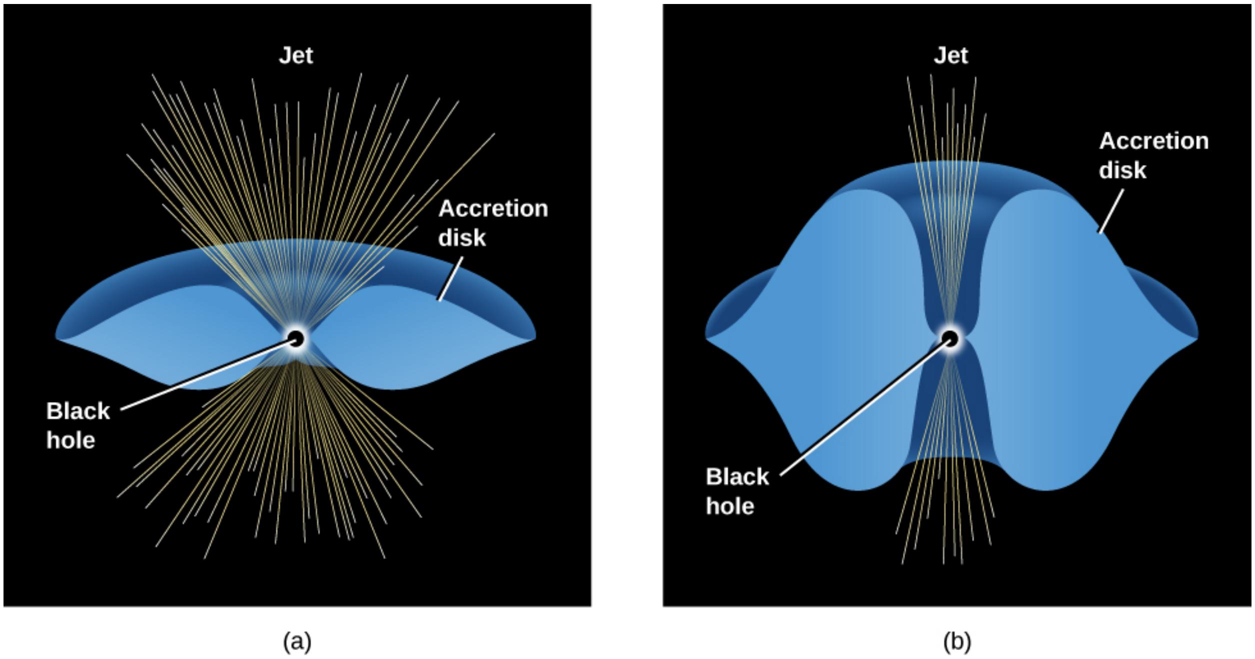


Figure 28.13 These schematic drawings show what accretion disks might look like around large black holes for (a) a thin accretion disk and (b) a "fat" disk—the type needed to account for channeling the outflow of hot material into narrow jets oriented perpendicular to the disk.

Figure 28.14 shows observations of an elliptical galaxy that behaves in exactly this way. At the center of this active galaxy, there is a ring of dust and gas about 400 light-years in diameter, surrounding a 1.2-billion-M_{Sun} black hole. Radio observations show that two jets emerge in a direction perpendicular to the ring, just as the model predicts.

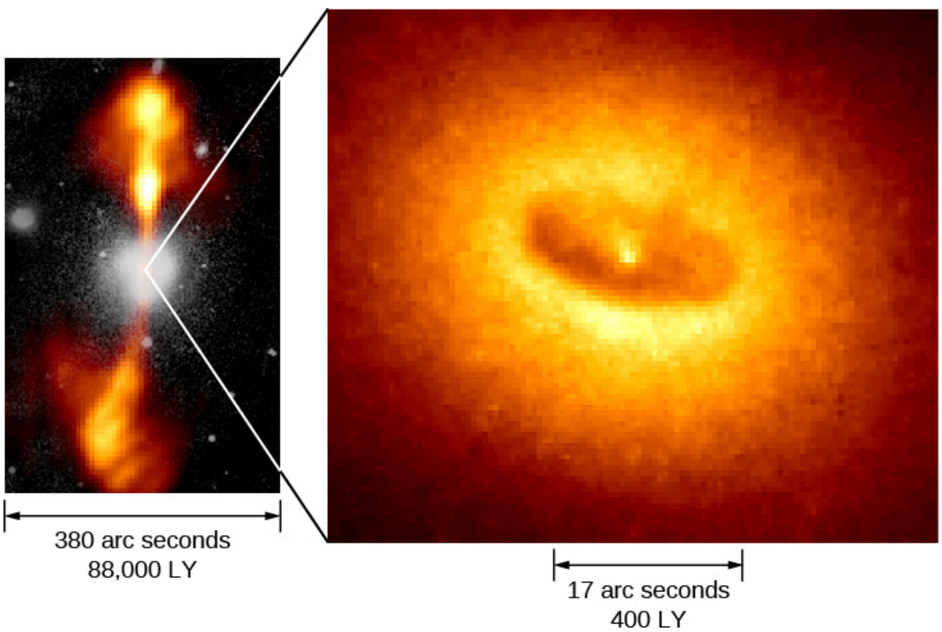


Figure 28.14 The picture on the left shows the active elliptical galaxy NGC 4261, which is located in the Virgo Cluster at a distance of about 100 million light-years. The galaxy itself—the white circular region in the center—is shown the way it looks in visible light, while the jets are seen at radio wavelengths. A Hubble Space Telescope image of the central portion of the galaxy is shown on the right. It contains a ring of dust and gas about 800 light-years in diameter, surrounding a supermassive black hole. Note that the jets emerge from the galaxy in a direction perpendicular to the plane of the ring. (credit: modification of work by ESA/HST)

Quasars and the Attitudes of Astronomers

The discovery of quasars in the early 1960s was the first in a series of surprises astronomers had in store. Within another decade they would find neutron stars (in the form of pulsars), the first hints of black holes (in binary X-ray sources), and even the radio echo of the Big Bang itself. Many more new discoveries lay ahead.

As Maarten Schmidt reminisced in 1988, "This had, I believe, a profound impact on the conduct of those practicing astronomy. Before the 1960s, there was much authoritarianism in the field. New ideas expressed at meetings would be instantly judged by senior astronomers and rejected if too far out." We saw a good example of this in the trouble Chandrasekhar had in finding acceptance for his ideas about the death of stars with cores greater than 1.4 M_{Sun} (see the feature box on **Subrahmanyan Chandrasekhar**).

"The discoveries of the 1960s," Schmidt continued, "were an embarrassment, in the sense that they were totally unexpected and could not be evaluated immediately. In reaction to these developments, an attitude has evolved where even outlandish ideas in astronomy are taken seriously. Given our lack of solid knowledge in extragalactic astronomy, this is probably to be preferred over authoritarianism."[1]

That is not to say that astronomers (being human) don't continue to have prejudices and preferences. For example, a small group of astronomers who thought that the redshifts of quasars were not connected with their distances (which was definitely a minority opinion) often felt excluded from meetings or from access to telescopes in the 1960s and 1970s. It's not so clear that they actually *were ex*cluded, as much as that they felt the very difficult pressure of knowing that most of their colleagues strongly disagreed with them. As it turned out, the evidence—which must ultimately decide all scientific questions—was not on their side either.

But today, as better instruments bring solutions to some problems and starkly illuminate our ignorance about others, the entire field of astronomy seems more open to discussing unusual ideas. Of course, before any hypotheses become accepted, they must be tested—again and again—against the evidence that nature itself reveals. Still, the many strange proposals published about what dark matter might be (see **The Evolution and Distribution of Galaxies**) attest to the new openness that Schmidt described.

With this black hole model, we have come a long way toward understanding the quasars and active galaxies that seemed very mysterious only a few decades ago. As often happens in astronomy, a combination of better instruments (making better observations) and improved theoretical models enabled us to make significant progress on a puzzling aspect of the cosmos.

28.3 | Quasars as Probes of Evolution in the Universe

Learning Objectives

By the end of this section, you will be able to:

- Trace the rise and fall of quasars over cosmic time
- Describe some of the ways in which galaxies and black holes influence each other's growth
- Describe some ways the first black holes may have formed
- Explain why some black holes are not producing quasar emission but rather are quiescent

The quasars' brilliance and large distance make them ideal probes of the far reaches of the universe and its remote past. Recall that when first introducing quasars, we mentioned that they generally tend to be far away. When we see extremely distant objects, we are seeing them as they were long ago. Radiation from a quasar 8 billion light-years away is telling us what that quasar and its environment were like 8 billion years ago, much closer to the time that the galaxy that surrounds it first formed. Astronomers have now detected light emitted from quasars that were already formed only a few hundred million years after the universe began its expansion 13.8 billion years ago. Thus, they give us a remarkable opportunity to learn about the time when large structures were first assembling in the cosmos.

The Evolution of Quasars

Quasars provide compelling evidence that we live in an evolving universe—one that changes with time. They tell us that astronomers living billions of years ago would have seen a universe that is very different from the universe of today. Counts

1. M. Schmidt, "The Discovery of Quasars," in *Modern Cosmology in Retrospect*, ed. B. Bertotti et al. (Cambridge University Press, 1990).

of the number of quasars at different redshifts (and thus at different times in the evolution of the universe) show us how dramatic these changes are (**Figure 28.15**). We now know that the number of quasars was greatest at the time when the universe was only 20% of its present age.

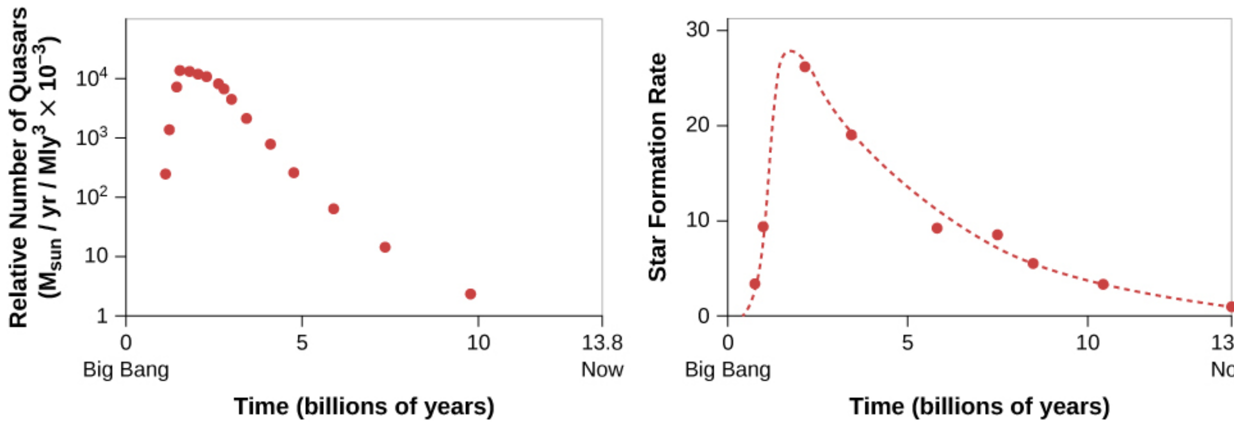


Figure 28.15 An age of 0 on the plots corresponds to the beginning of the universe; an age of 13.8 corresponds to the present time. Both the number of quasars and the rate of star formation were at a peak when the universe was about 20% as old as it is now.

As you can see, the drop-off in the numbers of quasars as time gets nearer to the present day is quite abrupt. Observations also show that the emission from the accretion disks around the most massive black holes peaks early and then fades. The most powerful quasars are seen only at early times. In order to explain this result, we make use of our model of the energy source of the quasars—namely that quasars are black holes with enough fuel to make a brilliant accretion disk right around them.

The fact that there were more quasars long ago (far away) than there are today (nearby) could be explained if there was more material available to be accreted by black holes early in the history of the universe. You might say that the quasars were more active when their black holes had fuel for their "energy-producing engines." If that fuel was mostly consumed in the first few billion years after the universe began its expansion, then later in its life, a "hungry" black hole would have very little left with which to light up the galaxy's central regions.

In other words, if matter in the accretion disk is continually being depleted by falling into the black hole or being blown out from the galaxy in the form of jets, then a quasar can continue to radiate only as long as new gas is available to replenish the accretion disk.

In fact, there *was* more gas around to be accreted early in the history of the universe. Back then, most gas had not yet collapsed to form stars, so there was more fuel available for both the feeding of black holes and the forming of new stars. Much of that fuel was subsequently consumed in the formation of stars during the first few billion years after the universe began its expansion. Later in its life, a galaxy would have little left to feed a hungry black hole or to form more new stars. As we see from **Figure 28.15**, both star formation and black hole growth peaked together when the universe was about 2 billion years old. Ever since, both have been in sharp decline. We are late to the party of the galaxies and have missed some of the early excitement.

Observations of nearer galaxies (seen later in time) indicate that there is another source of fuel for the central black holes—the collision of galaxies. If two galaxies of similar mass collide and merge, or if a smaller galaxy is pulled into a larger one, then gas and dust from one may come close enough to the black hole in the other to be devoured by it and so provide the necessary fuel. Astronomers have found that collisions were also much more common early in the history of the universe than they are today. There were more small galaxies in those early times because over time, as we shall see (in **Galaxy Mergers and Active Galactic Nuclei**), small galaxies tend to combine into larger ones. Again, this means that we would expect to see more quasars long ago (far away) than we do today (nearby)—as we in fact do.

Codependence of Black Holes and Galaxies

Once black hole masses began to be measured reliably in the late 1990s, they posed an enigma. It looked as though the mass of the central black hole depended on the mass of the galaxy. The black holes in galaxies always seem to be just 1/200 the mass of the galaxy they live in. This result is shown schematically in **Figure 28.16**, and some of the observations are plotted in **Figure 28.17**.

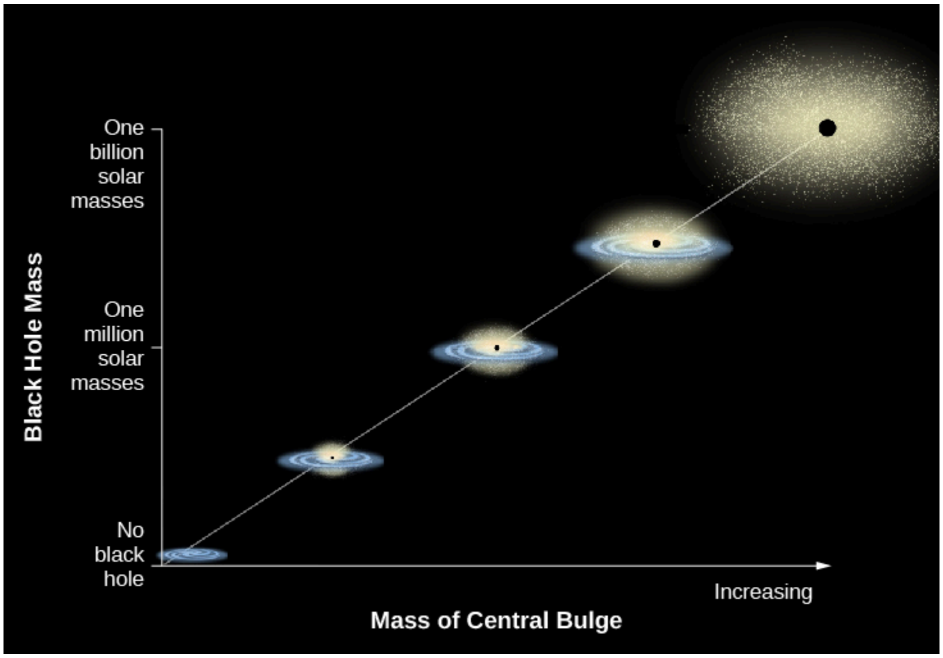


Figure 28.16 Observations show that there is a close correlation between the mass of the black hole at the center of a galaxy and the mass of the spherical distribution of stars that surrounds the black hole. That spherical distribution may be in the form of either an elliptical galaxy or the central bulge of a spiral galaxy. (credit: modification of work by K. Cordes, S. Brown (STScI))

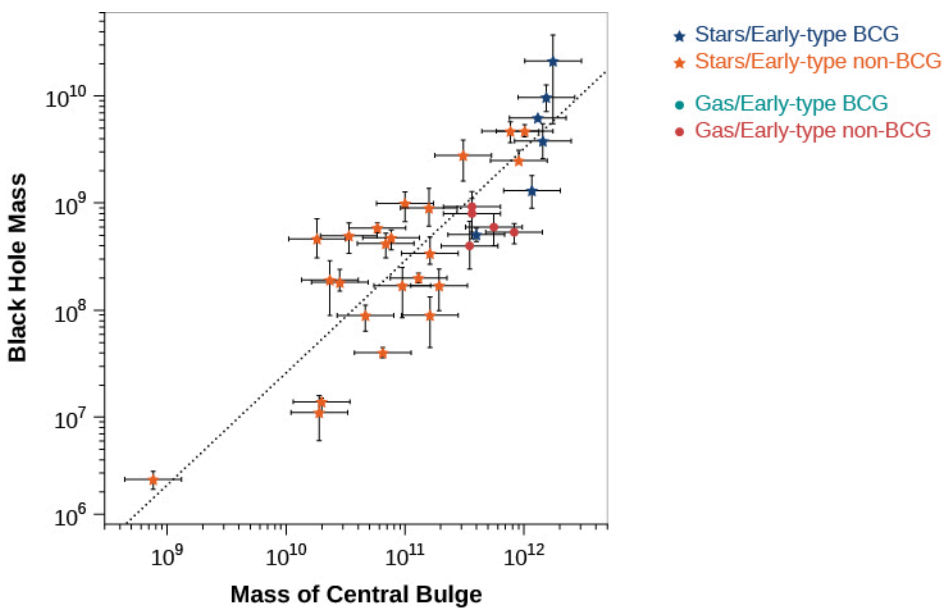


Figure 28.17 The black hole always turns out to be about 1/200 the mass of the stars surrounding it. The horizontal and vertical bars surrounding each point show the uncertainty of the measurement. (credit: modification of work by Nicholas J. McConnell, Chung-Pei Ma, "Revisiting the Scaling Relations of Black Hole Masses and Host Galaxy Properties," *The Astrophysical Journal*, 764:184 (14 pp.), February 20, 2013.)

Somehow black hole mass and the mass of the surrounding bulge of stars are connected. But why does this correlation exist? Unfortunately, astronomers do not yet know the answer to this question. We do know, however, that the black hole can influence the rate of star formation in the galaxy, and that the properties of the surrounding galaxy can influence how fast the black hole grows. Let's see how these processes work.

How a Galaxy Can Influence a Black Hole in Its Center

Let's look first at how the surrounding galaxy might influence the growth and size of the black hole. Without large quantities of fresh "food," the surroundings of black holes glow only weakly as bits of local material spiral inward toward the black hole. So somehow large amounts of gas have to find their way to the black hole from the galaxy in order to feed the quasar and make it grow and give off the energy to be noticed. Where does this "food" for the black hole come from originally and how might it be replenished? The jury is still out, but the options are pretty clear.

One obvious source of fuel for the black hole is matter from the host galaxy itself. Galaxies start out with large amounts of interstellar gas and dust, and at least some of this interstellar matter is gradually converted into stars as the galaxy evolves. On the other hand, as stars go through their lives and die, they lose mass all the time into the space between them, thereby returning some of the gas and dust to the interstellar medium. We expect to find more gas and dust in the central regions early in a galaxy's life than later on, when much of it has been converted into stars. Any of the interstellar matter that ventures too close to the black hole may be accreted by it. This means that we would expect that the number and luminosity of quasars powered in this way would decline with time. And as we have seen, that is just what we find.

Today both *elliptical galaxies* and the *nuclear bulges of spiral galaxies* have very little raw material left to serve as a source of fuel for the black hole. And most of the giant black holes in nearby galaxies, including the one in our own Milky Way, are now dark and relatively quiet—mere shadows of their former selves. So that fits with our observations.

We should note that even if you have a quiescent supermassive black hole, a star in the area could occasionally get close to it. Then the powerful tidal forces of the black hole can pull the whole star apart into a stream of gas. This stream quickly forms an accretion disk that gives off energy in the normal way and makes the black hole region into a temporary quasar. However, the material will fall into the black hole after only a few weeks or months. The black hole then goes back into its lurking, quiescent state, until another victim wanders by.

This sort of "cannibal" event happens only once every 100,000 years or so in a typical galaxy. But we can monitor millions of galaxies in the sky, so a few of these " tidal disruption events" are found each year (**Figure 28.18**). However, these individual events, dramatic as they are, are too rare to account for the huge masses of the central black holes.

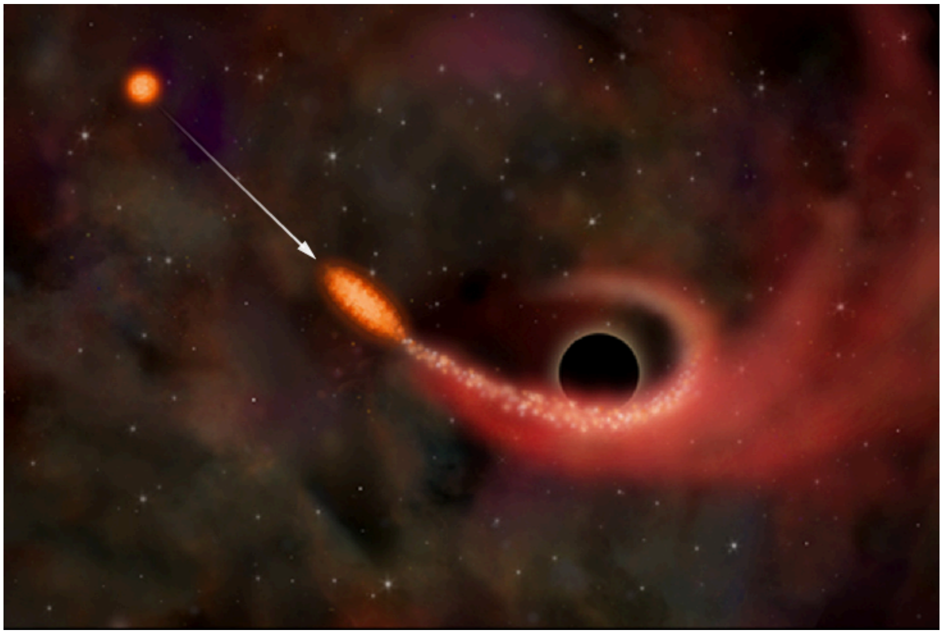


Figure 28.18 This artist's impression shows three stages of a star (red) swinging too close to a giant black hole (black circle). The star starts off (top left) in its normal spherical shape, then begins to be pulled into a long football shape by tides raised by the black hole (center). When the star gets closer still, the tides become stronger than the gravity holding the star together, and it breaks up into a streamer (right). Much of the star's matter forms a temporary accretion disk that lights up as a quasar for a few weeks or months. (credit: modification of work by NASA/CXC/M. Weiss)

Another source of fuel for the black hole is the collision of its host galaxy with another galaxy. Some of the brightest galaxies turn out, when a detailed picture is taken, to be pairs of colliding galaxies. And most of them have quasars inside them, not easily visible to us because they are buried by enormous amounts of dust and gas.

A collision between two cars creates quite a mess, pushing parts out of their regular place. In the same way, if two galaxies collide and merge, then gas and dust (though not so much the stars) can get pushed out of their regular orbits. Some may veer close enough to the black hole in one galaxy or the other to be devoured by it and so provide the necessary fuel to power a quasar. As we saw, galaxy collisions and mergers happened most frequently when the universe was young and probably help account for the fact that quasars were most common when the universe was only about 20% of its current age.

Collisions in today's universe are less frequent, but they do happen. Once a galaxy reaches the size of the Milky Way, most of the galaxies it merges with will be much smaller galaxies—*dwarf galaxies* (see the chapter on **Galaxies**). These don't disrupt the big galaxy much, but they can supply some additional gas to its black hole.

By the way, if two galaxies, each of which contains a black hole, collide, then the two black holes may merge and form an even larger black hole (**Figure 28.19**). In this process they will emit a burst of gravitational waves. One of the main goals of the European Space Agency's planned LISA (Laser Interferometer Space Antenna) mission is to detect the gravitational wave signals from the merging of supermassive black holes.

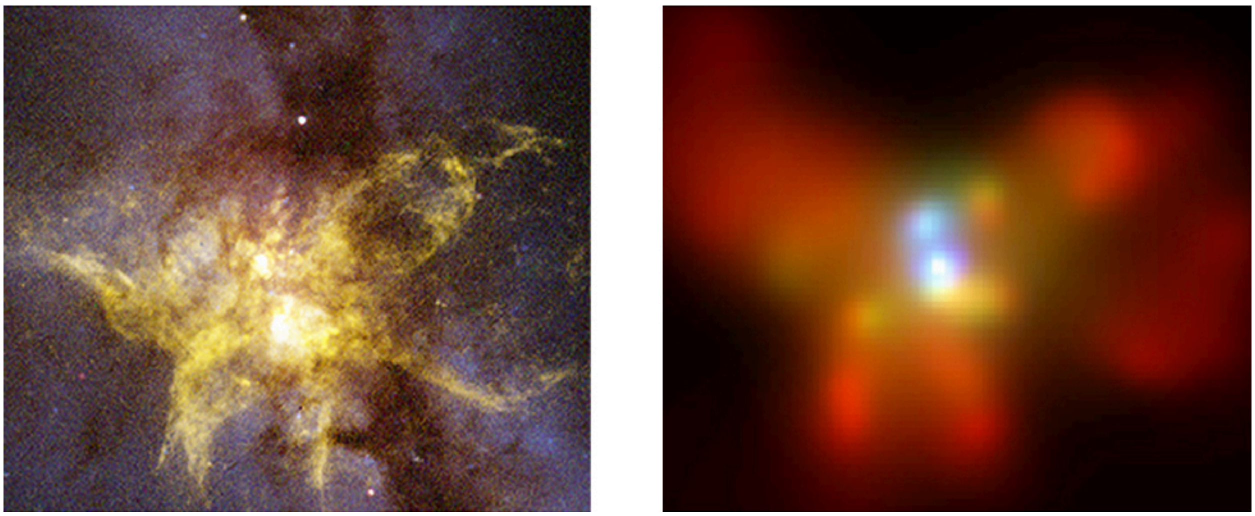


Figure 28.19 We compare Hubble Space Telescope visible-light (left) and Chandra X-ray (right) images of the central regions of NGC 6240, a galaxy about 400 million light-years away. It is a prime example of a galaxy in which stars are forming, evolving, and exploding at an exceptionally rapid rate due to a relatively recent merger (30 million years ago). The Chandra image shows two bright X-ray sources, each produced by hot gas surrounding a black hole. Over the course of the next few hundred million years, the two supermassive black holes, which are about 3000 light-years apart, will drift toward each other and merge to form an even larger black hole. This detection of a binary black hole supports the idea that black holes can grow to enormous masses in the centers of galaxies by merging with nearby galaxies. (credit left: modification of work by NASA/CXC/MPE/S.Komossa et al; credit right: NASA/STScI/R. P. van der Marel, J. Gerssen)

Watch **two galaxies collide (https://openstax.org/l/30galcolsuperma)** to form a supermassive black hole.

How Does the Black Hole Influence the Formation of Stars in the Galaxy?

We have seen that the material in galaxies can influence the growth of the black hole. The black hole in turn can also influence the galaxy in which it resides. It can do so in three ways: through its jets, through winds of particles that manage to stream away from the accretion disk, and through radiation from the accretion disk. As they stream away from the black hole, all three can either promote star formation by compressing the surrounding gas and dust—or instead suppress star formation by heating the surrounding gas and shredding molecular clouds, thereby inhibiting or preventing star formation. The outflowing energy can even be enough to halt the accretion of new material and starve the black hole of fuel. Astronomers are still trying to evaluate the relative importance of these effects in determining the overall evolution of galactic bulges and the rates of star formation.

In summary, we have seen how galaxies and supermassive black holes can each influence the evolution of the other: the galaxy supplies fuel to the black hole, and the quasar can either support or suppress star formation. The balance of these processes probably helps account for the correlation between black hole and bulge masses, but there are as yet no theories that explain quantitatively and in detail why the correlation between black hole and bulge masses is as tight as it is or why the black hole mass is always about 1/200 times the mass of the bulge.

The Birth of Black Holes and Galaxies

While the connection between quasars and galaxies is increasingly clear, the biggest puzzle of all—namely, how the supermassive black holes in galaxies got started—remains unsolved. Observations show that they existed when the universe was very young. One dramatic example is the discovery of a quasar that was already shining when the universe was only 700 million years old. What does it take to create a large black hole so quickly? A related problem is that in order to eventually build black holes containing more than 2 billion solar masses, it is necessary to have giant "seed" black holes with masses at least 2000 times the mass of the Sun—and they must somehow have been created shortly after the expansion of the universe began.

Astronomers are now working actively to develop models for how these seed black holes might have formed. Theories suggest that galaxies formed from collapsing clouds of dark matter and gas. Some of the gas formed stars, but perhaps some of the gas settled to the center where it became so concentrated that it formed a black hole. If this happened, the black hole could form right away—although this requires that the gas should not be rotating very much initially.

A more likely scenario is that the gas will have some angular momentum (rotation) that will prevent direct collapse to a black hole. In that case, the very first generation of stars will form, and some of them, according to calculations, will have masses hundreds of times that of the Sun. When these stars finish burning hydrogen, just a few million years later, the supernovae they end with will create black holes a hundred or so times the mass of the Sun. These can then merge with others or accrete the rich gas supply available at these early times.

The challenge is growing these smaller black holes quickly enough to make the much larger black holes we see a few hundred million years later. It turns out to be difficult because there are limits on how fast they can accrete matter. These should make sense to you from what we discussed earlier in the chapter. If the rate of accretion becomes too high, then the energy streaming outward from the black hole's accretion disk will become so strong as to blow away the infalling matter.

What if, instead, a collapsing gas cloud doesn't form a black hole directly or break up and form a group of regular stars, but stays together and makes one fairly massive star embedded within a dense cluster of thousands of lower mass stars and large quantities of dense gas? The massive star will have a short lifetime and will soon collapse to become a black hole. It can then begin to attract the dense gas surrounding it. But calculations show that the gravitational attraction of the many nearby stars will cause the black hole to zigzag randomly within the cluster and will prevent the formation of an accretion disk. If there is no accretion disk, then matter can fall freely into the black hole from all directions. Calculations suggest that under these conditions, a black hole even as small as 10 times the mass of the Sun could grow to more than 10 billion times the mass of the Sun by the time the universe is a billion years old.

Scientists are exploring other ideas for how to form the seeds of supermassive black holes, and this remains a very active field of research. Whatever mechanism caused the rapid formation of these supermassive black holes, they do give us a way to observe the youthful universe when it was only about five percent as old as it is now.

Take a look at some **new results (https://openstax.org/l/30chanxrayobser)** from the Chandra X-ray Observatory about the formation of supermassive black holes in the early universe.

28.4 | Observations of Distant Galaxies

Learning Objectives

By the end of this section, you will be able to:

- Explain how astronomers use light to learn about distant galaxies long ago
- Discuss the evidence showing that the first stars formed when the universe was less than 10% of its current age
- Describe the major differences observed between galaxies seen in the distant, early universe and galaxies seen in the nearby universe today

Let's now begin exploring some techniques astronomers use to study how galaxies are born and change over cosmic time. Suppose you wanted to understand how adult humans got to be the way they are. If you were very dedicated and patient, you could actually observe a sample of babies from birth, following them through childhood, adolescence, and into adulthood, and making basic measurements such as their heights, weights, and the proportional sizes of different parts of their bodies to understand how they change over time.

Unfortunately, we have no such possibility for understanding how galaxies grow and change over time: in a human lifetime—or even over the entire history of human civilization—individual galaxies change hardly at all. We need other tools than just patiently observing single galaxies in order to study and understand those long, slow changes.

We do, however, have one remarkable asset in studying galactic evolution. As we have seen, the universe itself is a kind of time machine that permits us to observe remote galaxies as they were long ago. For the closest galaxies, like the Andromeda galaxy, the time the light takes to reach us is on the order of a few hundred thousand to a few million years. Typically not much changes over times that short—individual stars in the galaxy may be born or die, but the overall structure and appearance of the galaxy will remain the same. But we have observed galaxies so far away that we are seeing them as they were when the light left them more than 10 billion years ago.

By observing more distant objects, we look further back toward a time when both galaxies and the universe were young (**Figure 28.20**). This is a bit like getting letters in the mail from several distant friends: the farther the friend was when she mailed the letter to you, the longer the letter must have been in transit, and so the older the news is when it arrives in your mailbox; you are learning something about her life at an earlier time than when you read the letter.

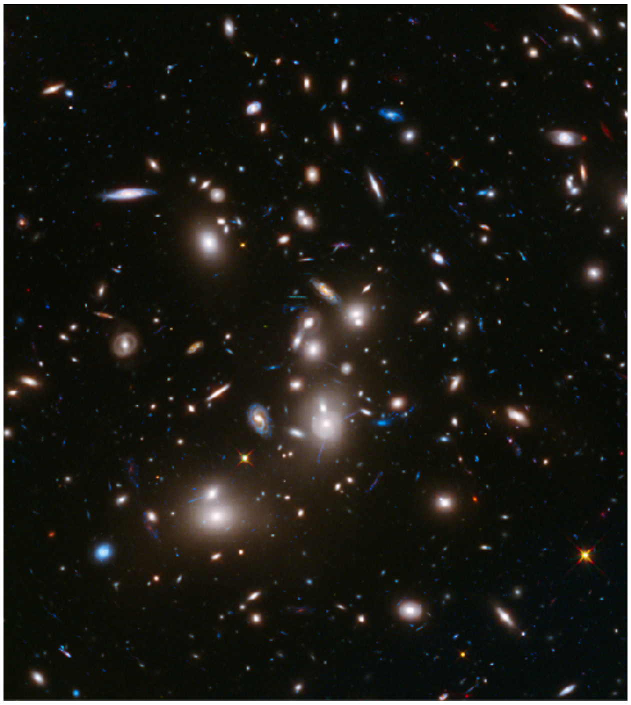


Figure 28.20 This true-color, long-exposure image, made during 70 orbits of Earth with the Hubble Space Telescope, shows a small area in the direction of the constellation Sculptor. The massive cluster of galaxies named Abell 2744 appears in the foreground of this image. It contains several hundred galaxies, and we are seeing them as they looked 3.5 billion years ago. The immense gravity in Abell 2744 acts as a gravitational lens (see the Astronomy Basics feature box on **Gravitational Lensing** later in this chapter) to warp space and brighten and magnify images of nearly 3000 distant background galaxies. The more distant galaxies (many of them quite blue) appear as they did more than 12 billion years ago, not long after the Big Bang. Blue galaxies were much more common in that earlier time than they are today. These galaxies appear blue because they are undergoing active star formation and making hot, bright blue stars. (credit: NASA, ESA, STScI)

If we can't directly detect the changes over time in individual galaxies because they happen too slowly, how then can we ever understand those changes and the origins of galaxies? The solution is to observe many galaxies at many different cosmic distances and, therefore, look-back times (how far back in time we are seeing the galaxy). If we can study a thousand very distant "baby" galaxies when the universe was 1 billion years old, and another thousand slightly closer "toddler" galaxies when it was 2 billion years old, and so on until the present 13.8-billion-year-old universe of mature "adult" galaxies near us today, then maybe we can piece together a coherent picture of how the whole ensemble of galaxies evolves over time. This allows us to reconstruct the "life story" of galaxies since the universe began, even though we can't follow a single galaxy from infancy to old age.

Fortunately, there is no shortage of galaxies to study. Hold up your pinky at arm's length: the part of the sky blocked by your fingernail contains about one million galaxies, layered farther and farther back in space and time. In fact, the sky is filled with galaxies, all of them, except for Andromeda and the Magellanic Clouds, too faint to see with the naked eye—more than 2 trillion (2000 billion) galaxies in the observable universe, each one with about 100 billion stars.

This cosmic time machine, then, lets us peer into the past to answer fundamental questions about where galaxies come from and how they got to be the way they are today. Astronomers call those galactic changes over cosmic time **evolution**, a word that recalls the work of Darwin and others on the development of life on Earth. But note that galaxy evolution refers to the changes in *individual* galaxies over time, while the kind of evolution biologists study is changes in *successive generations* of living organisms over time.

Spectra, Colors, and Shapes

Astronomy is one of the few sciences in which all measurements must be made at a distance. Geologists can take samples of the objects they are studying; chemists can conduct experiments in their laboratories to determine what a substance is made of; archeologists can use carbon dating to determine how old something is. But astronomers can't pick up and play with a star or galaxy. As we have seen throughout this book, if they want to know what galaxies are made of and how they have changed over the lifetime of the universe, they must decode the messages carried by the small number of photons that reach Earth.

Fortunately (as you have learned) electromagnetic radiation is a rich source of information. The distance to a galaxy is derived from its *redshift* (how much the lines in its spectrum are shifted to the red because of the expansion of the universe). The conversion of redshift to a distance depends on certain properties of the universe, including the value of the Hubble constant and how much mass it contains. We will describe the currently accepted model of the universe in **Big Bang Cosmology**. For the purposes of this chapter, it is enough to know that the current best estimate for the age of the universe is 13.8 billion years. In that case, if we see an object that emitted its light 6 billion light-years ago, we are seeing it as it was when the universe was almost 8 billion years old. If we see something that emitted its light 13 billion years ago, we are seeing it as it was when the universe was less than a billion years old. So astronomers measure a galaxy's redshift from its spectrum, use the Hubble constant plus a model of the universe to turn the redshift into a distance, and use the distance and the constant speed of light to infer how far back in time they are seeing the galaxy—the look-back time.

In addition to distance and look-back time, studies of the Doppler shifts of a galaxy's spectral lines can tell us how fast the galaxy is rotating and hence how massive it is (as explained in **Galaxies**). Detailed analysis of such lines can also indicate the types of stars that inhabit a galaxy and whether it contains large amounts of interstellar matter.

Unfortunately, many galaxies are so faint that collecting enough light to produce a detailed spectrum is currently impossible. Astronomers thus have to use a much rougher guide to estimate what kinds of stars inhabit the faintest galaxies—their overall colors. Look again at **Figure 28.20** and notice that some of the galaxies are very blue and others are reddish-orange. Now remember that hot, luminous blue stars are very massive and have lifetimes of only a few million years. If we see a galaxy where blue colors dominate, we know that it must have many hot, luminous blue stars, and that star formation must have taken place in the few million years before the light left the galaxy. In a yellow or red galaxy, on the other hand, the young, luminous blue stars that surely were made in the galaxy's early bursts of star formation must have died already; it must contain mostly old yellow and red stars that last a long time in their main-sequence stages and thus typically formed billions of years before the light that we now see was emitted.

Another important clue to the nature of a galaxy is its shape. Spiral galaxies can be distinguished from elliptical galaxies by shape. Observations show that spiral galaxies contain young stars and large amounts of interstellar matter, while elliptical galaxies have mostly old stars and very little or no star formation. Elliptical galaxies turned most of their interstellar matter into stars many billions of years ago, while star formation has continued until the present day in spiral galaxies.

If we can count the number of galaxies of each type during each epoch of the universe, it will help us understand how the pace of star formation changes with time. As we will see later in this chapter, galaxies in the distant universe—that is, young galaxies—look very different from the older galaxies that we see nearby in the present-day universe.

The First Generation of Stars

In addition to looking at the most distant galaxies we can find, astronomers look at the oldest stars (what we might call the fossil record) of our own Galaxy to probe what happened in the early universe. Since stars are the source of nearly all the light emitted by galaxies, we can learn a lot about the evolution of galaxies by studying the stars within them. What we find is that nearly all galaxies contain at least some very old stars. For example, our own Galaxy contains globular clusters with stars that are at least 13 billion years old, and some may be even older than that. Therefore, if we count the age of the Milky Way as the age of its oldest constituents, the Milky Way must have been born at least 13 billion years ago.

As we will discuss in **Big Bang Cosmology**, astronomers have discovered that the universe is expanding, and have traced the expansion backward in time. In this way, they have discovered that the universe itself is only about 13.8 billion years old. Thus, it appears that at least some of the globular-cluster stars in the Milky Way must have formed less than a billion years after the expansion began.

Several other observations also establish that star formation in the cosmos began very early. Astronomers have used spectra to determine the composition of some elliptical galaxies that are so far away that the light we see left them when the universe was only half as old as it is now. Yet these ellipticals contain old red stars, which must have formed billions of years earlier still.

When we make computer models of how such galaxies evolve with time, they tell us that star formation in elliptical galaxies began less than a billion years or so after the universe started its expansion, and new stars continued to form for a few billion years. But then star formation apparently stopped. When we compare distant elliptical galaxies with ones nearby, we find that ellipticals have not changed very much since the universe reached about half its current age. We'll return to this idea later in the chapter.

Observations of the most luminous galaxies take us even further back in time. Recently, as we have already noted, astronomers have discovered a few galaxies that are so far away that the light we see now left them less than a billion years or so after the beginning (**Figure 28.21**). Yet the spectra of some of these galaxies already contain lines of heavy elements, including carbon, silicon, aluminum, and sulfur. These elements were not present when the universe began but had to be manufactured in the interiors of stars. This means that when the light from these galaxies was emitted, an entire generation of stars had already been born, lived out their lives, and died—spewing out the new elements made in their interiors through supernova explosions—even before the universe was a billion years old. And it wasn't just a few stars in each galaxy that got started this way. Enough had to live and die to affect the overall composition of the galaxy, in a way that we can still measure in the spectrum from far away.

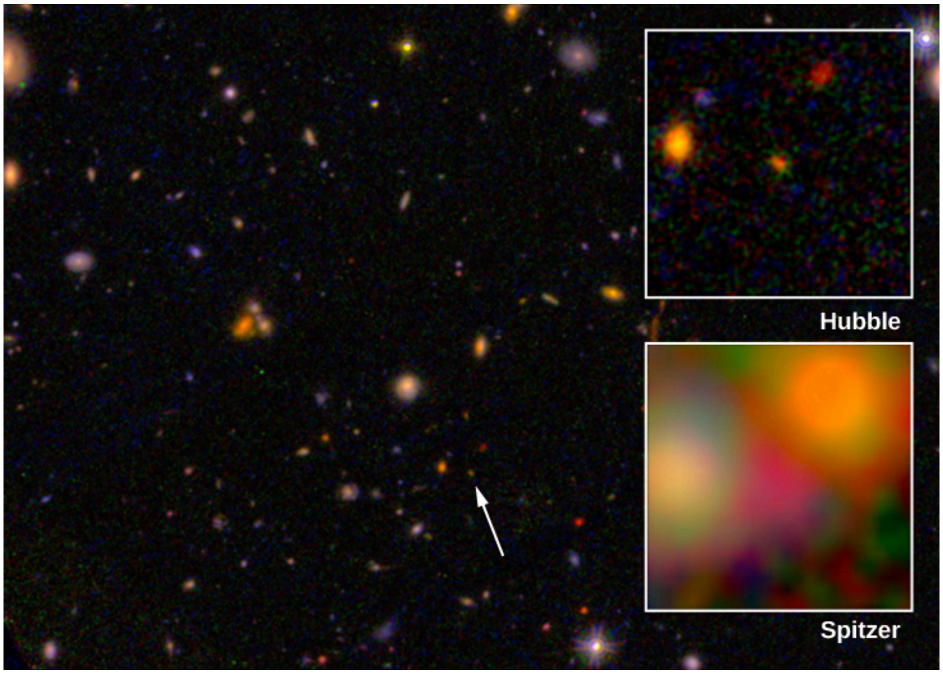


Figure 28.21 This image was made with the Hubble Space Telescope and shows the field around a luminous galaxy at a redshift $z = 8.68$, which corresponds to 13.2 billion light years. This means that we are seeing this galaxy as it appeared about 13.2 billion years ago. The galaxy itself is indicated by the arrow. Long exposures in the far-red and infrared wavelengths were combined to make the image, and additional infrared exposures with the Spitzer Space Telescope, which has lower spatial resolution than the Hubble (lower inset), show the redshifted light of normal stars. The very distant galaxy was detected because it has a strong emission line of hydrogen. This line is produced in regions where the formation of hot, young stars is taking place. (credit: modification of work by I. Labbé (Leiden University), NASA/ESA/JPL-Caltech)

Observations of *quasars* (galaxies whose centers contain a supermassive black hole) support this conclusion. We can measure the abundances of heavy elements in the gas near quasar black holes (explained in **Supermassive Black Holes**). The composition of this gas in quasars that emitted their light 12.5 billion light-years ago is very similar to that of the Sun. This means that a large portion of the gas surrounding the black holes must have already been cycled through stars during the first 1.3 billion years after the expansion of the universe began. If we allow time for this cycling, then their first stars must have formed when the universe was only a few hundred million years old.

A Changing Universe of Galaxies

Back in the middle decades of the twentieth century, the observation that all galaxies contain some old stars led astronomers to the hypothesis that galaxies were born fully formed near the time when the universe began its expansion. This hypothesis was similar to suggesting that human beings were born as adults and did not have to pass through the various stages of development from infancy through the teens. If this hypothesis were correct, the most distant galaxies should have shapes and sizes very much like the galaxies we see nearby. According to this old view, galaxies, after they formed, should then change only slowly, as successive generations of stars within them formed, evolved, and died. As the interstellar matter was slowly used up and fewer new stars formed, the galaxies would gradually become dominated by fainter, older stars and look dimmer and dimmer.

Thanks to the new generation of large ground- and space-based telescopes, we now know that this picture of galaxies evolving peacefully and in isolation from one another is completely wrong. As we will see later in this chapter, galaxies in the distant universe do not look like the Milky Way and nearby galaxies such as Andromeda, and the story of their development is more complex and involves far more interaction with their neighbors.

Why were astronomers so wrong? Up until the early 1990s, the most distant normal galaxy that had been observed emitted its light 8 billion years ago. Since that time, many galaxies—and particularly the giant ellipticals, which are the most luminous and therefore the easiest to see at large distances—did evolve peacefully and slowly. But the Hubble, Spitzer, Herschel, Keck, and other powerful new telescopes that have come on line since the 1990s make it possible to pierce the 8-billion-light-year barrier. We now have detailed views of many thousands of galaxies that emitted their light much earlier (some more than 13 billion years ago—see **Figure 28.21**).

Much of the recent work on the evolution of galaxies has progressed by studying a few specific small regions of the sky where the Hubble, Spitzer, and ground-based telescopes have taken extremely long exposure images. This allowed astronomers to detect very faint, very distant, and therefore very *young* galaxies (**Figure 28.22**). Our deep space telescope images show some galaxies that are 100 times fainter than the faintest objects that can be observed spectroscopically with today's giant ground-based telescopes. This turns out to mean that we can obtain the spectra needed to determine redshifts for only the very brightest five percent of the galaxies in these images.

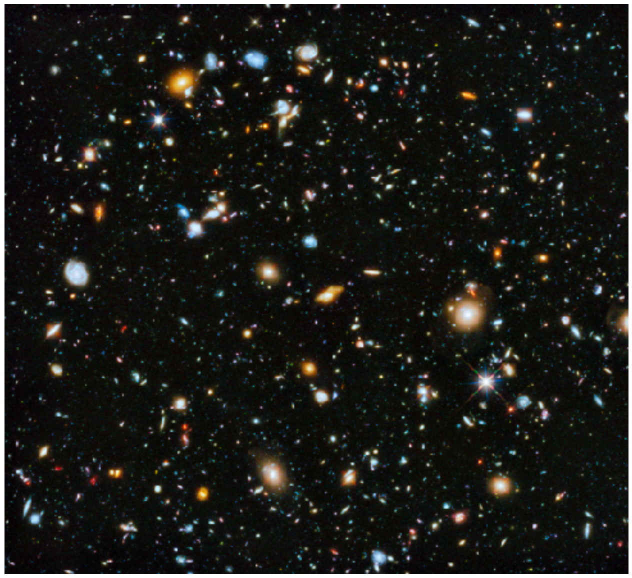


Figure 28.22 This image is the result of an 11-day-long observation with the Hubble Space Telescope of a tiny region of sky, located toward the constellation Fornax near the south celestial pole. This is an area that has only a handful of Milky Way stars. (Since the Hubble orbits Earth every 96 minutes, the telescope returned to view the same tiny piece of sky over and over again until enough light was collected and added together to make this very long exposure.) There are about 10,000 objects in this single image, nearly all of them galaxies, each with tens or hundreds of billions of stars. We can see some pinwheel-shaped spiral galaxies, which are like the Milky Way. But we also find a large variety of peculiar-shaped galaxies that are in collision with companion galaxies. Elliptical galaxies, which contain mostly old stars, appear as reddish blobs. (credit: modification of work by NASA, ESA, H. Teplitz and M. Rafelski (IPAC/Caltech), A. Koekemoer (STScI), R. Windhorst (Arizona State University), and Z. Levay (STScI))

Although we do not have spectra for most of the faint galaxies, the Hubble Space Telescope is especially well suited to studying their *shapes* because the images taken in space are not blurred by Earth's atmosphere. To the surprise of astronomers, the distant galaxies did not fit Hubble's classification scheme at all. Remember that Hubble found that nearly all nearby galaxies could be classified into a few categories, depending on whether they were ellipticals or spirals. The distant galaxies observed by the Hubble Space Telescope look very different from present-day galaxies, without identifiable spiral arms, disks, and bulges (**Figure 28.23**). They also tend to be much clumpier than most galaxies today. In other words, it's becoming clear that the shapes of galaxies have changed significantly over time. In fact, we now know that the Hubble scheme works well for only the last half of the age of the universe. Before then, galaxies were much more chaotic.

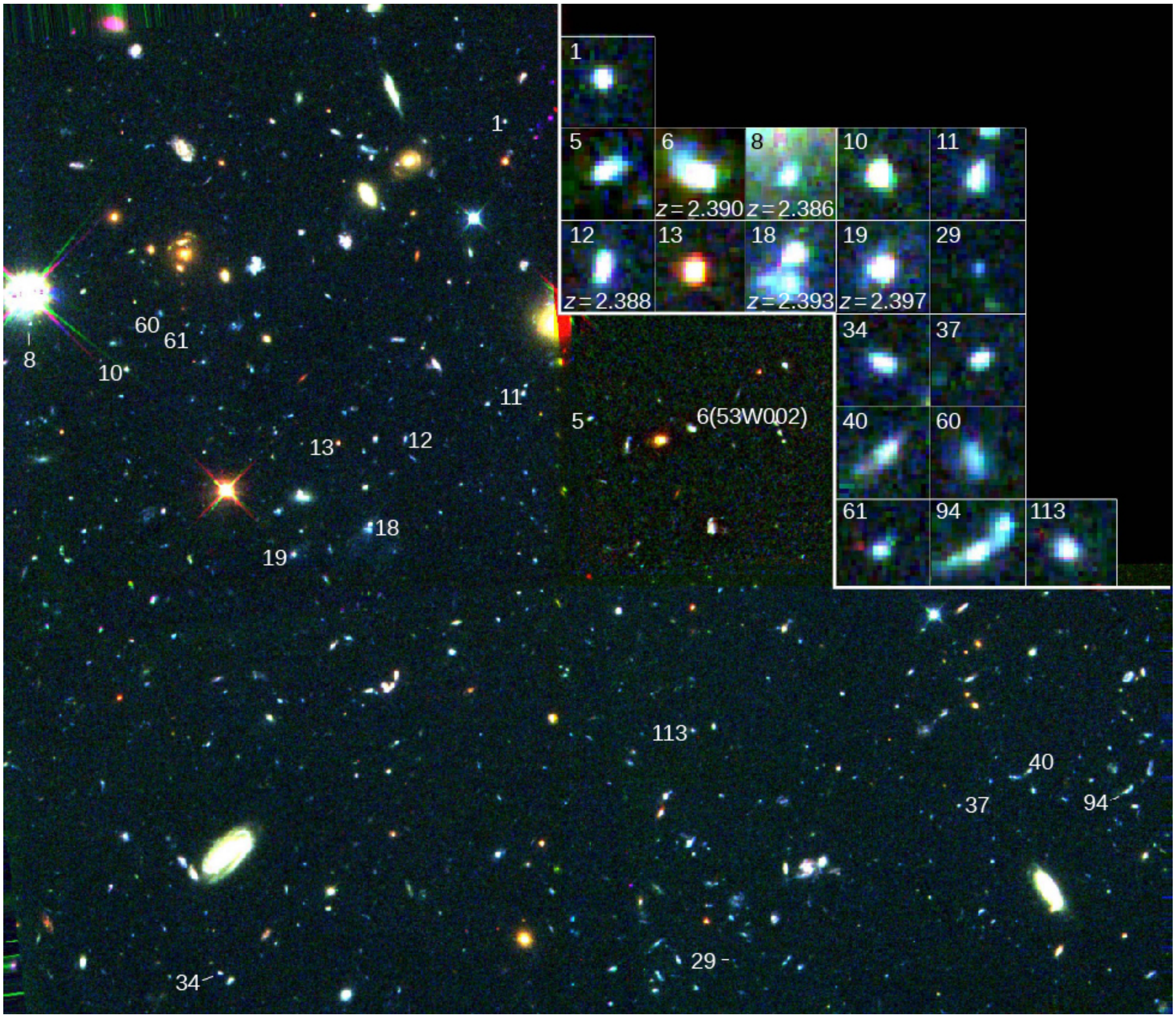


Figure 28.23 This Hubble Space Telescope image shows what are probably "galaxies under construction" in the early universe. The boxes in this color image show enlargements of 18 groups of stars smaller than galaxies as we know them. All these objects emitted their light about 11 billion years ago. They are typically only about 2,000 light-years across, which is much smaller than the Milky Way, with its diameter of 100,000 light-years. These 18 objects are found in a region only 2 million light-years across and are close enough together that they will probably collide and merge to build one or more normal galaxies. (credit: modification of work by Rogier Windhorst (Arizona State University) and NASA)

It's not just the shapes that are different. Nearly all the galaxies with red-shifts that correspond to 11 billion light-years or more—that is, galaxies that we are seeing when they were less than 3 billion years old—are extremely blue, indicating that they contain a lot of young stars and that star formation in them is occurring at a higher rate than in nearby galaxies. Observations also show that very distant galaxies are systematically smaller on average than nearby galaxies. Relatively few galaxies present before the universe was about 8 billion years old have masses greater than 10^{11} M_{Sun}. That's 1/20 the mass of the Milky Way if we include its dark matter halo. Eleven billion years ago, there were only a few galaxies with masses greater than 10^{10} M_{Sun}. What we see instead seem to be small pieces or fragments of galactic material (**Figure 28.24**). When we look at galaxies that emitted their light 11 to 12 billion years ago, we now believe we are seeing the *seeds* of elliptical galaxies and of the central bulges of spirals. Over time, these smaller galaxies collided and merged to build up today's large galaxies.

Bear in mind that stars that formed more than 11 billion years ago will be very old stars today. Indeed when we look nearby (at galaxies we see closer to our time), we find mostly old stars in the nuclear bulges of nearby spirals and in elliptical galaxies.

Figure 28.24 The small white boxes, labeled *a*, *b*, and *c*, mark the positions of three images of the same galaxy. These multiple images were produced by the massive cluster of galaxies known as Abell 2744, which is located between us and the galaxy and acts as a gravitational lens. The arrows in the enlarged insets at right point to the galaxy. Each magnified image makes the galaxy appear as much as 10 times larger and brighter than it would look without the intervening lens. This galaxy emitted the light we observe today when the universe was only about 500 million years old. When the light was emitted the galaxy was tiny—only 850 light-years across, or 500 times smaller than the Milky, and its mass was only 40 million times the mass of the Sun. Star formation is going on in this galaxy, but it appears red in the image because of its large redshift. (credit: modification of work by NASA, ESA, A. Zitrin (California Institute of Technology), and J. Lotz, M. Mountain, A. Koekemoer, and the HFF Team (STScI))

What such observations are showing us is that galaxies have grown in size as the universe has aged. Not only were galaxies smaller several billion years ago, but there were more of them; gas-rich galaxies, particularly the less luminous ones, were much more numerous then than they are today.

Those are some of the basic observations we can make of individual galaxies (and their evolution) looking back in cosmic time. Now we want to turn to the larger context. If stars are grouped into galaxies, are the galaxies also grouped in some way? In the third section of this chapter, we'll explore the largest structures known in the universe.

28.5 | Galaxy Mergers and Active Galactic Nuclei

Learning Objectives

By the end of this section, you will be able to:

- Explain how galaxies grow by merging with other galaxies and by consuming smaller galaxies (for lunch)
- Describe the effects that supermassive black holes in the centers of most galaxies have on the fate of their host galaxies

One of the conclusions astronomers have reached from studying distant galaxies is that collisions and mergers of whole galaxies play a crucial role in determining how galaxies acquired the shapes and sizes we see today. Only a few of the nearby galaxies are currently involved in collisions, but detailed studies of those tell us what to look for when we seek evidence of mergers in very distant and very faint galaxies. These in turn give us important clues about the different evolutionary paths galaxies have taken over cosmic time. Let's examine in more detail what happens when two galaxies collide.

Mergers and Cannibalism

Figure 28.2 shows a dynamic view of two galaxies that are colliding. The stars themselves in this pair of galaxies will not be affected much by this cataclysmic event. (See the Astronomy Basics feature box **Why Galaxies Collide but Stars Rarely Do**.) Since there is a lot of space between the stars, a direct collision between two stars is very unlikely. However, the *orbits* of many of the stars will be changed as the two galaxies move through each other, and the change in orbits can totally alter the appearance of the interacting galaxies. A gallery of interesting colliding galaxies is shown in **Figure 28.25**. Great rings, huge tendrils of stars and gas, and other complex structures can form in such cosmic collisions. Indeed, these strange shapes are the signposts that astronomers use to identify colliding galaxies.

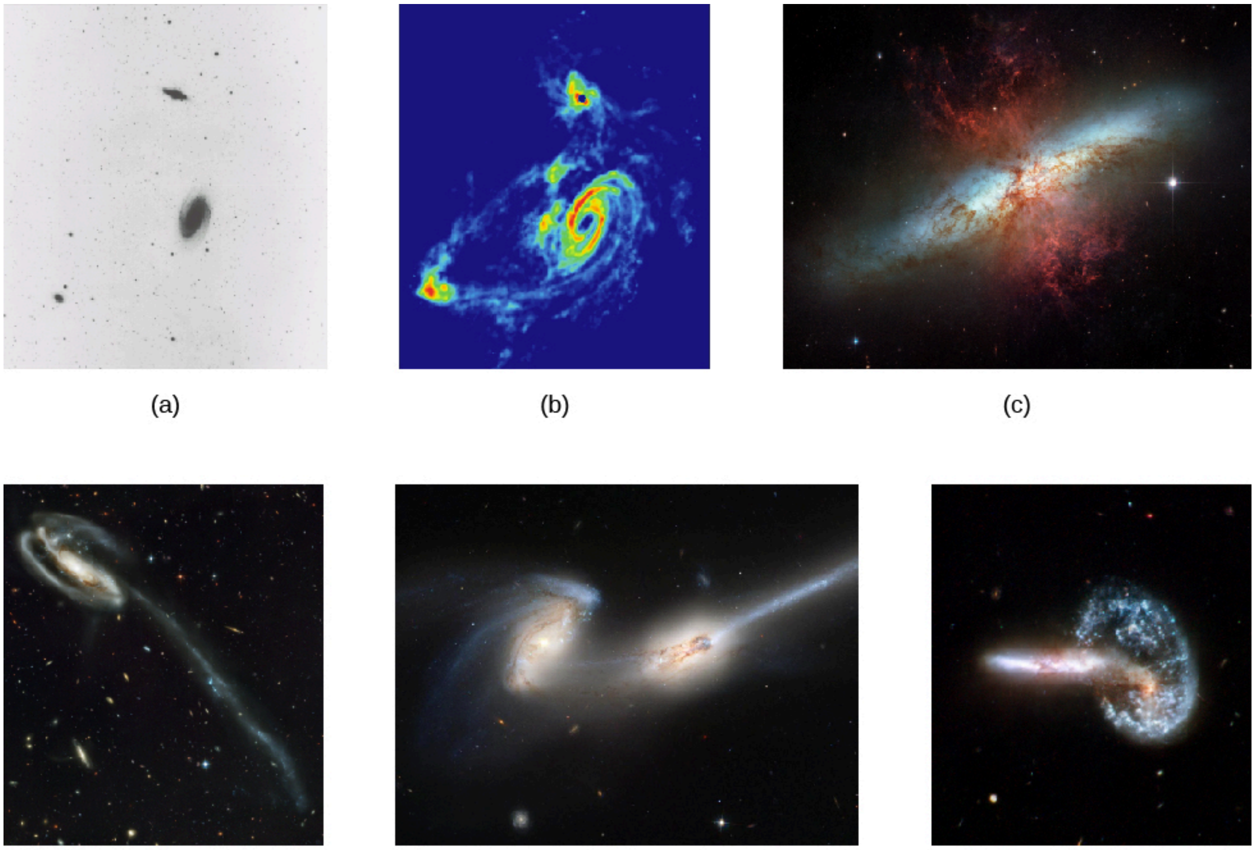


Figure 28.25 (a and b) M82 (smaller galaxy at top) and M83 (spiral) are seen (a) in a black-and-white visible light image and (b) in radio waves given off by cold hydrogen gas. The hydrogen image shows that the two galaxies are wrapped in a common shroud of gas that is being tugged and stretched by the gravity of the two galaxies. (c) This close-up view by the Hubble Space Telescope shows some of the effects of this interaction on galaxy M82, including gas streaming outward (red tendrils) powered by supernovae explosions of massive stars formed in the burst of star formation that was a result of the collision. (d) Galaxy UGC 10214 ("The Tadpole") is a barred spiral galaxy 420 million light-years from the Milky Way that has been disrupted by the passage of a smaller galaxy. The interloper's gravity pulled out the long tidal tail, which is about 280,000 light-years long, and triggered bursts of star formation seen as blue clumps along the tail. (e) Galaxies NGC 4676 A and B are nicknamed "The Mice." In this Hubble Space Telescope image, you can see the long, narrow tails of stars pulled away from the galaxies by the interactions of the two spirals. (f) Arp 148 is a pair of galaxies that are caught in the act of merging to become one new galaxy. The two appear to have already passed through each other once, causing a shockwave that reformed one into a bright blue ring of star formation, like the ripples from a stone tossed into a pond. (credit a, b: modification of work by NRAO/AUI; credit c: modification of work by NASA, ESA, and The Hubble Heritage Team (STScI/AURA); credit d, e: modification of work by NASA, H. Ford (JHU), G. Illingworth (UCSC/LO), M.Clampin (STScI), G. Hartig (STScI), the ACS Science Team, and ESA; credit f: modification of work by NASA, ESA, the Hubble Heritage (STScI/AURA)-ESA/Hubble Collaboration, and A. Evans (University of Virginia, Charlottesville/NRAO/Stony Brook University))

Why Galaxies Collide but Stars Rarely Do

Throughout this book we have emphasized the large distances between objects in space. You might therefore have been surprised to hear about collisions between galaxies. Yet (except at the very cores of galaxies) we have not worried at all about stars inside a galaxy colliding with each other. Let's see why there is a difference.

The reason is that stars are pitifully small compared to the distances between them. Let's use our Sun as an example. The Sun is about 1.4 million kilometers wide, but is separated from the closest other star by about 4 light-years, or about 38 trillion kilometers. In other words, the Sun is 27 million of its own diameters from its nearest neighbor. If the Sun were a grapefruit in New York City, the nearest star would be another grapefruit in San Francisco. This is typical of stars that are not in the nuclear bulge of a galaxy or inside star clusters. Let's contrast this with the separation of galaxies.

The visible disk of the Milky Way is about 100,000 light-years in diameter. We have three satellite galaxies that are just one or two Milky Way diameters away from us (and will probably someday collide with us). The closest major spiral is the Andromeda Galaxy (M31), about 2.4 million light-years away. If the Milky Way were a pancake at one end of a big breakfast table, M31 would be another pancake at the other end of the same table. Our nearest large galaxy neighbor is only 24 of our Galaxy's diameters from us, and it will begin to crash into the Milky Way in about 3 billion years.

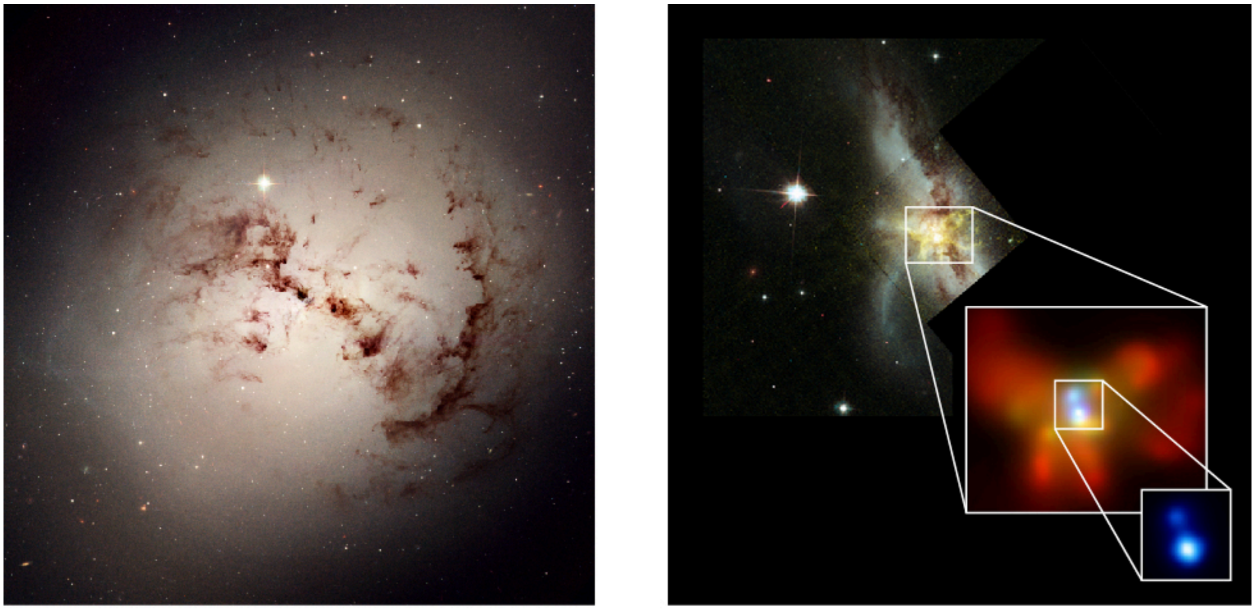
Galaxies in rich clusters are even closer together than those in our neighborhood (see **The Distribution of Galaxies in Space**). Thus, the chances of galaxies colliding are far greater than the chances of stars in the disk of a galaxy colliding. And we should note that the difference between the separation of galaxies and stars also means that when galaxies do collide, their stars almost always pass right by each other like smoke passing through a screen door.

The details of galaxy collisions are complex, and the process can take hundreds of millions of years. Thus, collisions are best simulated on a computer (**Figure 28.26**), where astronomers can calculate the slow interactions of stars, and clouds of gas and dust, via gravity. These calculations show that if the collision is slow, the colliding galaxies may coalesce to form a single galaxy.

Figure 28.26 This computer simulation starts with two spiral galaxies merging and ends with a single elliptical galaxy. The colors show the colors of stars in the system; note the bursts of blue color as copious star formation gets triggered by the interaction. The timescale from start to finish in this sequence is about a billion years. (credit: modification of work by P. Jonsson (Harvard-Smithsonian Center for Astrophysics), G. Novak (Princeton University), and T. J. Cox (Carnegie Observatories))

When two galaxies of equal size are involved in a collision, we call such an interaction a **merger** (the term applied in the business world to two equal companies that join forces). But small galaxies can also be swallowed by larger ones—a process astronomers have called, with some relish, **galactic cannibalism** (**Figure 28.27**).

Modern personal computers are more than powerful enough to compute what happens when galaxies collide. Here's a **website and Java applet (https://openstax.org/l/30whgalaxcoll)** that will let you try your own hand at crashing two spiral galaxies together from the comfort of your own home or dorm room. By changing a few basic controls such as the relative masses, their separation, and the orientation of each galaxy's disk, you can create a wide range of resulting merger results. (You can also download a similar **app (https://openstax.org/l/30iphoneapp)** for your iPhone or iPad.)



(a) (b)

Figure 28.27 (a) This Hubble image shows the eerie silhouette of dark dust clouds against the glowing nucleus of the elliptical galaxy NGC 1316. Elliptical galaxies normally contain very little dust. These clouds are probably the remnant of a small companion galaxy that was cannibalized (eaten) by NGC 1316 about 100 million years ago. (b) The highly disturbed galaxy NGC 6240, imaged by Hubble Space Telescope (background image) and Chandra X-ray Telescope (both insets) is apparently the product of a merger between two gas-rich spiral galaxies. The X-ray images show that there is not one but two nuclei, both glowing brightly in X-rays and separated by only 4000 light-years. These are likely the locations of two supermassive black holes that inhabited the cores of the two galaxies pre-merger; here they are participating in a kind of "death spiral," in which the two black holes themselves will merge to become one. (credit a: modification of work by NASA, ESA, and The Hubble Heritage Team (STScI/AURA); credit b: X-ray: NASA/CXC/MPE/S.Komossa et al.; Optical: NASA/STScI/R.P.van der Marel & J.Gerssen)

The very large elliptical galaxies we discussed in **Galaxies** probably form by cannibalizing a variety of smaller galaxies in their clusters. These "monster" galaxies frequently possess more than one nucleus and have probably acquired their unusually high luminosities by swallowing nearby galaxies. The multiple nuclei are the remnants of their victims (**Figure 28.27**). Many of the large, peculiar galaxies that we observe also owe their chaotic shapes to past interactions. Slow collisions and mergers can even transform two or more spiral galaxies into a single elliptical galaxy.

A change in shape is not all that happens when galaxies collide. If either galaxy contains interstellar matter, the collision can compress the gas and trigger an increase in the rate at which stars are being formed—by as much as a factor of 100. Astronomers call this abrupt increase in the number of stars being formed a **starburst**, and the galaxies in which the increase occurs are termed starburst galaxies (**Figure 28.28**). In some interacting galaxies, star formation is so intense that all the available gas is exhausted in only a few million years; the burst of star formation is clearly only a temporary phenomenon. While a starburst is going on, however, the galaxy where it is taking place becomes much brighter and much easier to detect at large distances.

(a) (b)

Figure 28.28 (a) Three of the galaxies in the small group known as Stephan's Quintet are interacting gravitationally with each other (the galaxy at upper left is actually much closer than the other three and is not part of this interaction), resulting in the distorted shapes seen here. Long strings of young, massive blue stars and hundreds of star formation regions glowing in the pink light of excited hydrogen gas are also results of the interaction. The ages of the star clusters range from 2 million to 1 billion years old, suggesting that there have been several different collisions within this group of galaxies, each leading to bursts of star formation. The three interacting members of Stephan's Quintet are located at a distance of 270 million light-years. (b) Most galaxies form new stars at a fairly slow rate, but members of a rare class known as starburst galaxies blaze with extremely active star formation. The galaxy II Zw 096 is one such starburst galaxy, and this combined image using both Hubble and Spitzer Space Telescope data shows that it is forming bright clusters of new stars at a prodigious rate. The blue colors show the merging galaxies in visible light, while the red colors show infrared radiation from the dusty region where star formation is happening. This galaxy is at a distance of 500 million light-years and has a diameter of about 50,000 light-years, about half the size of the Milky Way. (credit a: modification of work by NASA, ESA, and the Hubble SM4 ERO Team; credit b: modification of work by NASA/JPL-Caltech/STScI)

When astronomers finally had the tools to examine a significant number of galaxies that emitted their light 11 to 12 billion years ago, they found that these very young galaxies often resemble nearby starburst galaxies that are involved in mergers: they also have multiple nuclei and peculiar shapes, they are usually clumpier than normal galaxies today, with multiple intense knots and lumps of bright starlight, and they have higher rates of star formation than isolated galaxies. They also contain lots of blue, young, type O and B stars, as do nearby merging galaxies.

Galaxy mergers in today's universe are rare. Only about five percent of nearby galaxies are currently involved in interactions. Interactions were much more common billions of years ago (**Figure 28.29**) and helped build up the "more mature" galaxies we see in our time. Clearly, interactions of galaxies have played a crucial role in their evolution.

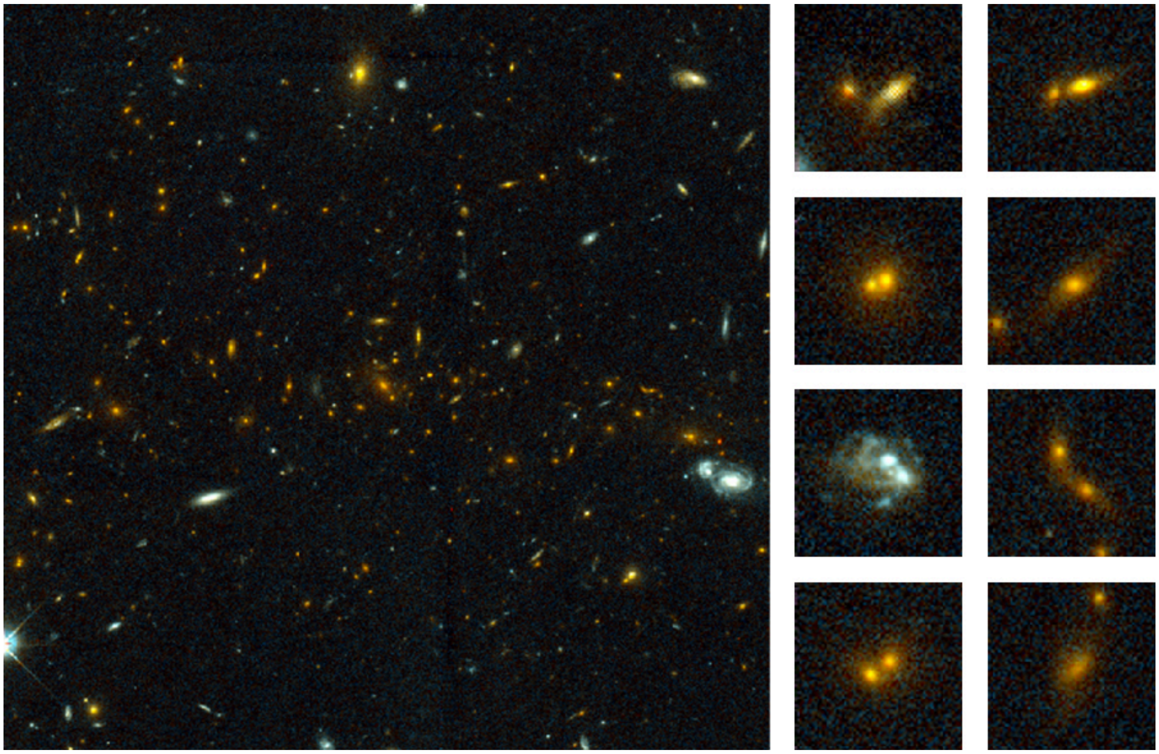


Figure 28.29 The large picture on the left shows the Hubble Space Telescope image of a cluster of galaxies at a distance of about 8 billion light-years. Among the 81 galaxies in the cluster that have been examined in some detail, 13 are the result of recent collisions of pairs of galaxies. The eight smaller images on the right are close-ups of some of the colliding galaxies. The merger process typically takes a billion years or so. (credit: modification of work by Pieter van Dokkum, Marijn Franx (University of Groningen/Leiden), ESA and NASA)

Active Galactic Nuclei and Galaxy Evolution

While galaxy mergers are huge, splashy events that completely reshape entire galaxies on scales of hundreds of thousands of light-years and can spark massive bursts of star formation, accreting black holes inside galaxies can also disturb and alter the evolution of their host galaxies. You learned in **Supermassive Black Holes** about a family of objects known as *active galactic nuclei* (AGN), all of them powered by supermassive black holes. If the black hole is surrounded by enough gas, some of the gas can fall into the black hole, getting swept up on the way into an accretion disk, a compact, swirling maelstrom perhaps only 100 AU across (the size of our solar system).

Within the disk the gas heats up until it shines brilliantly even in X-rays, often outshining the rest of the host galaxy with its billions of stars. Supermassive black holes and their accretion disks can be violent and powerful places, with some material getting sucked into the black hole but even more getting shot out along huge jets perpendicular to the disk. These powerful jets can extend far outside the starry edge of the galaxy.

AGN were much more common in the early universe, in part because frequent mergers provided a fresh gas supply for the black hole accretion disks. Examples of AGN in the nearby universe today include the one in galaxy M87 (see **Figure 28.8**), which sports a jet of material shooting out from its nucleus at speeds close to the speed of light, and the one in the bright galaxy NGC 5128, also known as Centaurus A (see **Figure 28.30**).

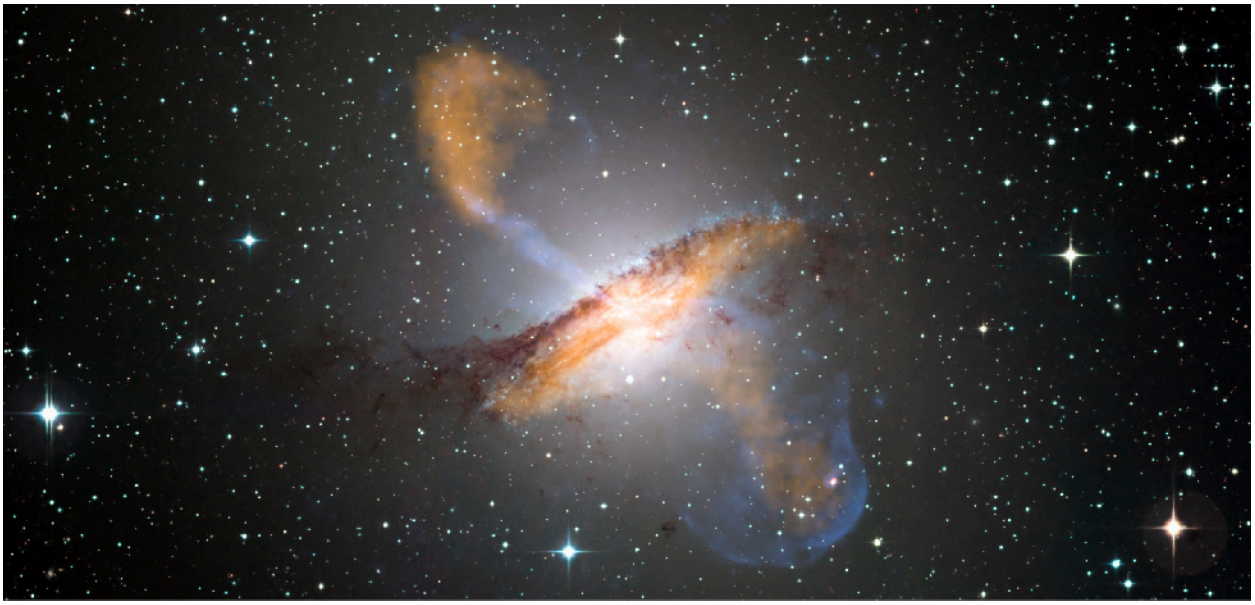


Figure 28.30 This artificially colored image was made using data from three different telescopes: submillimeter radiation with a wavelength of 870 microns is shown in orange; X-rays are seen in blue; and visible light from stars is shown in its natural color. Centaurus A has an active galactic nucleus that is powering two jets, seen in blue and orange, reaching in opposite directions far outside the galaxy's stellar disk, and inflating two huge lobes, or clouds, of hot X-ray-emitting gas. Centaurus is at a distance of 13 million light-years, making it one of the closest active galaxies we know. (credit: modification of work by ESO/WFI (Optical); MPIfR/ESO/APEX/A. Weiss et al. (Submillimeter); NASA/CXC/CfA/R.Kraft et al. (X-ray))

Many highly accelerated particles move with the jets in such galaxies. Along the way, the particles in the jets can plow into gas clouds in the interstellar medium, breaking them apart and scattering them. Since denser clouds of gas and dust are required for material to clump together to make stars, the disruption of the clouds can halt star formation in the host galaxy or cut it off before it even begins.

In this way, quasars and other kinds of AGN can play a crucial role in the evolution of their galaxies. For example, there is growing evidence that the merger of two gas-rich galaxies not only produces a huge burst of star formation, but also triggers AGN activity in the core of the new galaxy. That activity, in turn, could then slow down or shut off the burst of star formation—which could have significant implications for the apparent shape, brightness, chemical content, and stellar components of the entire galaxy. Astronomers refer to that process as *AGN feedback*, and it is apparently an important factor in the evolution of most galaxies.

28.6 | The Distribution of Galaxies in Space

Learning Objectives

By the end of this section, you will be able to:

- Explain the cosmological principle and summarize the evidence that it applies on the largest scales of the known universe
- Describe the contents of the Local Group of galaxies
- Distinguish among groups, clusters, and superclusters of galaxies
- Describe the largest structures seen in the universe, including voids

In the preceding section, we emphasized the role of mergers in shaping the evolution of galaxies. In order to collide, galaxies must be fairly close together. To estimate how often collisions occur and how they affect galaxy evolution, astronomers need to know how galaxies are distributed in space and over cosmic time. Are most of them isolated from one another or do they congregate in groups? If they congregate, how large are the groups and how and when did they form? And how, in general, are galaxies and their groups arranged in the cosmos? Are there as many in one direction of the sky as in any other, for example? How did galaxies get to be arranged the way we find them today?

Edwin Hubble found answers to some of these questions only a few years after he first showed that the spiral nebulae were galaxies and not part of our Milky Way. As he examined galaxies all over the sky, Hubble made two discoveries that turned out to be crucial for studies of the evolution of the universe.

The Cosmological Principle

Hubble made his observations with what were then the world's largest telescopes—the 100-inch and 60-inch reflectors on Mount Wilson. These telescopes have small fields of view: they can see only a small part of the heavens at a time. To photograph the entire sky with the 100-inch telescope, for example, would have taken longer than a human lifetime. So instead, Hubble sampled the sky in many regions, much as Herschel did with his star gauging (see **The Architecture of the Galaxy**). In the 1930s, Hubble photographed 1283 sample areas, and on each print, he carefully counted the numbers of galaxy images (**Figure 28.31**).

The first discovery Hubble made from his survey was that the number of galaxies visible in each area of the sky is about the same. (Strictly speaking, this is true only if the light from distant galaxies is not absorbed by dust in our own Galaxy, but Hubble made corrections for this absorption.) He also found that the numbers of galaxies increase with faintness, as we would expect if the density of galaxies is about the same at all distances from us.

To understand what we mean, imagine you are taking snapshots in a crowded stadium during a sold-out concert. The people sitting near you look big, so only a few of them will fit into a photo. But if you focus on the people sitting in seats way on the other side of the stadium, they look so small that many more will fit into your picture. If all parts of the stadium have the same seat arrangements, then as you look farther and farther away, your photo will get more and more crowded with people. In the same way, as Hubble looked at fainter and fainter galaxies, he saw more and more of them.

Figure 28.31 Edwin Hubble at the 100-inch telescope on Mount Wilson. (credit: NASA)

Hubble's findings are enormously important, for they indicate that the universe is both **isotropic** and **homogeneous**—it looks the same in all directions, and a large volume of space at any given redshift or distance is much like any other volume at that redshift. If that is so, it does not matter what section of the universe we observe (as long as it's a sizable portion): any section will look the same as any other.

Hubble's results—and many more that have followed in the nearly 100 years since then—imply not only that the universe is about the same everywhere (apart from changes with time) but also that aside from small-scale local differences, the part we can see around us is representative of the whole. The idea that the universe is the same everywhere is called the **cosmological principle** and is the starting assumption for nearly all theories that describe the entire universe (see **Big Bang Cosmology**).

Without the cosmological principle, we could make no progress at all in studying the universe. Suppose our own local neighborhood were unusual in some way. Then we could no more understand what the universe is like than if we were marooned on a warm south-sea island without outside communication and were trying to understand the geography of Earth. From our limited island vantage point, we could not know that some parts of the planet are covered with snow and ice, or that large continents exist with a much greater variety of terrain than that found on our island.

Hubble merely counted the numbers of galaxies in various directions without knowing how far away most of them were. With modern instruments, astronomers have measured the velocities and distances of hundreds of thousands of galaxies, and so built up a meaningful picture of the large-scale structure of the universe. In the rest of this section, we describe what we know about the distribution of galaxies, beginning with those that are nearby.

The Local Group

The region of the universe for which we have the most detailed information is, as you would expect, our own local neighborhood. It turns out that the Milky Way Galaxy is a member of a small group of galaxies called, not too imaginatively, the **Local Group**. It is spread over about 3 million light-years and contains 60 or so members. There are three large spiral galaxies (our own, the Andromeda galaxy, and M33), two intermediate ellipticals, and many dwarf ellipticals and irregular galaxies.

New members of the Local Group are still being discovered. We mentioned in **The Milky Way Galaxy** a dwarf galaxy only about 80,000 light-years from Earth and about 50,000 light-years from the center of the galaxy that was discovered in 1994 in the constellation of Sagittarius. (This dwarf is actually venturing too close to the much larger Milky Way and will eventually be consumed by it.)

Many of the recent discoveries have been made possible by the new generation of automated, sensitive, wide-field surveys, such as the Sloan Digital Sky Survey, that map the positions of millions of stars across most of the visible sky. By digging into the data with sophisticated computer programs, astronomers have turned up numerous tiny, faint dwarf galaxies that are all but invisible to the eye even in those deep telescopic images. These new findings may help solve a long-standing problem: the prevailing theories of how galaxies form predicted that there should be more dwarf galaxies around big galaxies like the Milky Way than had been observed—and only now do we have the tools to find these faint and tiny galaxies and begin to compare the numbers of them with theoretical predictions.

You can read more about the **Sloan survey (https://openstax.org/l/30sloansurvey)** and its dramatic results. And check out this **brief animation (https://openstax.org/l/30anifliarrgal)** of a flight through the arrangement of the galaxies as revealed by the survey.

Several new dwarf galaxies have also been found near the Andromeda galaxy. Such dwarf galaxies are difficult to find because they typically contain relatively few stars, and it is hard to distinguish them from the foreground stars in our own Milky Way.

Figure 28.32 is a rough sketch showing where the brighter members of the Local Group are located. The average of the motions of all the galaxies in the Local Group indicates that its total mass is about 4×10^{12} M_{Sun}, and at least half of this mass is contained in the two giant spirals—the Andromeda galaxy and the Milky Way Galaxy. And bear in mind that a substantial amount of the mass in the Local Group is in the form of dark matter.

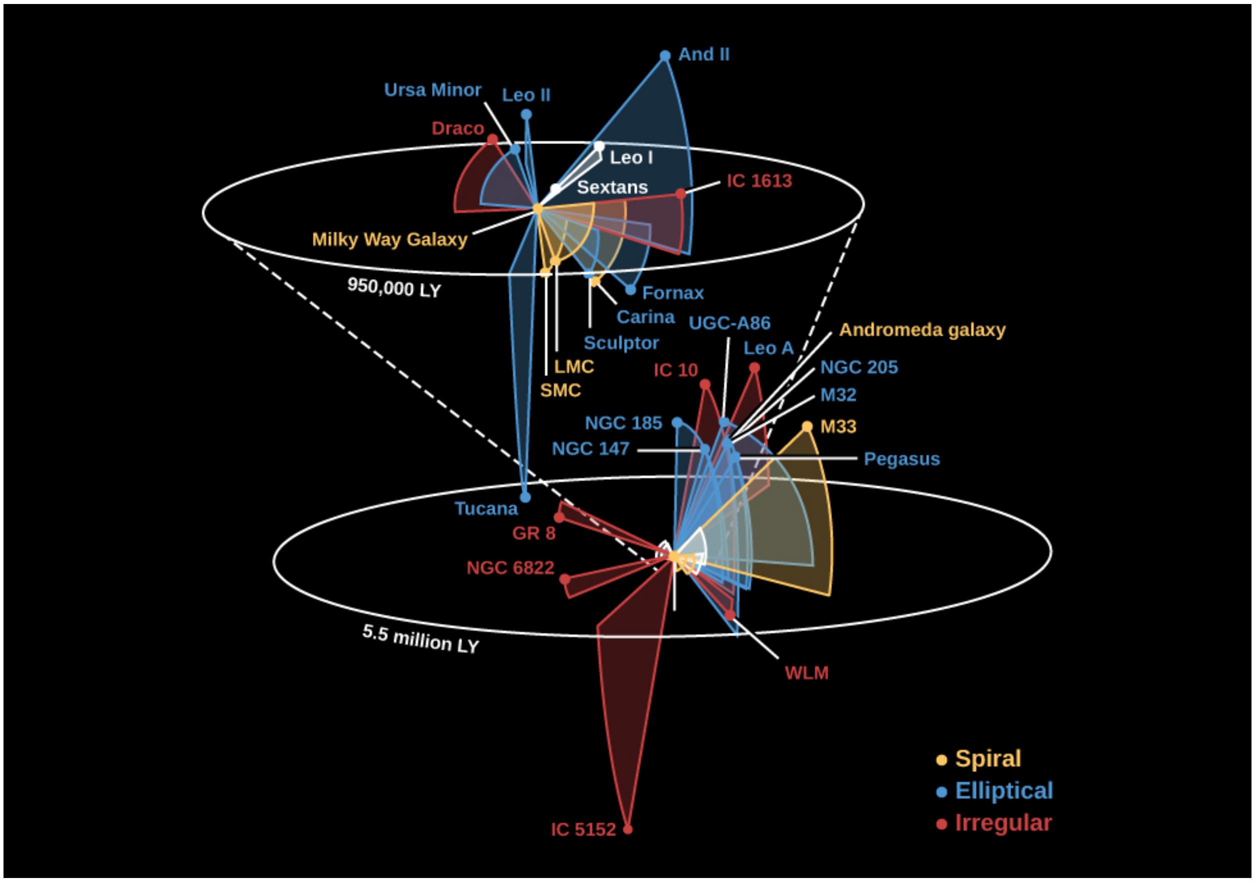


Figure 28.32 This illustration shows some members of the Local Group of galaxies, with our Milky Way at the center. The exploded view at the top shows the region closest to the Milky Way and fits into the bigger view at the bottom as shown by the dashed lines. The three largest galaxies among the three dozen or so members of the Local Group are all spirals; the others are small irregular galaxies and dwarf ellipticals. A number of new members of the group have been found since this map was made.

Neighboring Groups and Clusters

Small galaxy groups like ours are hard to notice at larger distances. However, there are much more substantial groups called galaxy clusters that are easier to spot even many millions of light-years away. Such clusters are described as *poor* or *rich* depending on how many galaxies they contain. Rich clusters have thousands or even tens of thousands of galaxies, although many of the galaxies are quite faint and hard to detect.

The nearest moderately rich galaxy cluster is called the Virgo Cluster, after the constellation in which it is seen. It is about 50 million light-years away and contains thousands of members, of which a few are shown in **Figure 28.33**. The giant elliptical (and very active) galaxy M87, which you came to know and love in **Supermassive Black Holes**, belongs to the Virgo Cluster.

Figure 28.33 Virgo is the nearest rich cluster and is at a distance of about 50 million light-years. It contains hundreds of bright galaxies. In this picture you can see only the central part of the cluster, including the giant elliptical galaxy M87, just below center. Other spirals and ellipticals are visible; the two galaxies to the top right are known as “The Eyes.” (credit: modification of work by Chris Mihos (Case Western Reserve University)/ESO)

A good example of a cluster that is much larger than the Virgo complex is the Coma cluster, with a diameter of at least 10 million light-years (**Figure 28.34**). A little over 300 million light-years distant, this cluster is centered on two giant ellipticals whose luminosities equal about 400 billion Suns each. Thousands of galaxies have been observed in Coma, but the galaxies we see are almost certainly only part of what is really there. Dwarf galaxies are too faint to be seen at the distance of Coma, but we expect they are part of this cluster just as they are part of nearer ones. If so, then Coma likely contains tens of thousands of galaxies. The total mass of this cluster is about 4×10^{15} M_{Sun} (enough mass to make 4 million billion stars like the Sun).

Let’s pause here for a moment of perspective. We are now discussing numbers by which even astronomers sometimes feel overwhelmed. The Coma cluster may have 10, 20, or 30 thousand galaxies, and each galaxy has billions and billions of stars. If you were traveling at the speed of light, it would still take you more than 10 million years (longer than the history of the human species) to cross this giant swarm of galaxies. And if you lived on a planet on the outskirts of one of these galaxies, many other members of the cluster would be close enough to be noteworthy sights in your nighttime sky.

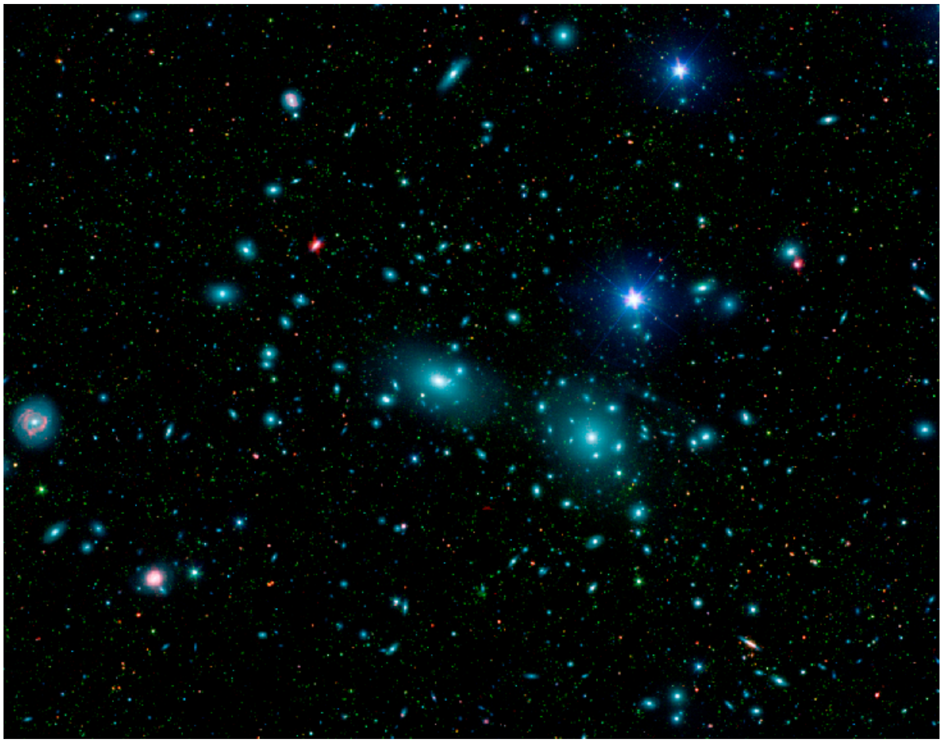


Figure 28.34 This combined visible-light (from the Sloan Digital Sky Survey) and infrared (from the Spitzer Space Telescope) image has been color coded so that faint dwarf galaxies are seen as green. Note the number of little green smudges on the image. The cluster is roughly 320 million light-years away from us. (credit: modification of work by NASA/JPL-Caltech/L. Jenkins (GSFC))

Really rich clusters such as Coma usually have a high concentration of galaxies near the center. We can see giant elliptical galaxies in these central regions but few, if any, spiral galaxies. The spirals that do exist generally occur on the outskirts of clusters.

We might say that ellipticals are highly "social": they are often found in groups and very much enjoy "hanging out" with other ellipticals in crowded situations. It is precisely in such crowds that collisions are most likely and, as we discussed earlier, we think that most large ellipticals are built through mergers of smaller galaxies.

Spirals, on the other hand, are more "shy": they are more likely to be found in poor clusters or on the edges of rich clusters where collisions are less likely to disrupt the spiral arms or strip out the gas needed for continued star formation.

Gravitational Lensing

As we saw in **Introducing General Relativity**, spacetime is more strongly curved in regions where the gravitational field is strong. Light passing very near a concentration of matter appears to follow a curved path. In the case of starlight passing close to the Sun, we measure the position of the distant star to be slightly different from its true position.

Now let's consider the case of light from a distant galaxy or quasar that passes near a concentration of matter such as a cluster of galaxies on its journey to our telescopes. According to general relativity, the light path may be bent in a variety of ways; as a result we can observe distorted and even multiple images (**Figure 28.35**).

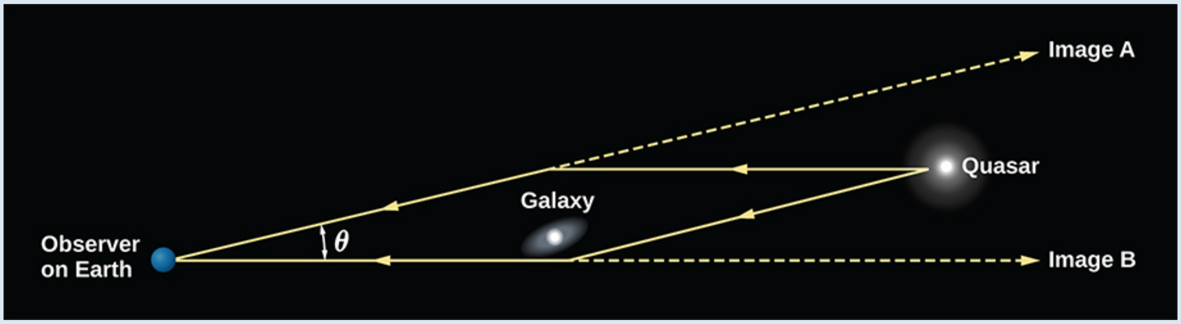


Figure 28.35 This drawing shows how a gravitational lens can make two images. Two light rays from a distant quasar are shown being bent while passing a foreground galaxy; they then arrive together at Earth. Although the two beams of light contain the same information, they now appear to come from two different points on the sky. This sketch is oversimplified and not to scale, but it gives a rough idea of the lensing phenomenon.

Gravitational lenses can produce not only double images, as shown in **Figure 28.35**, but also multiple images, arcs, or rings. The first gravitational lens discovered, in 1979, showed two images of the same distant object. Eventually, astronomers used the Hubble Space Telescope to capture remarkable images of the effects of gravitational lenses. One example is shown in **Figure 28.36**.

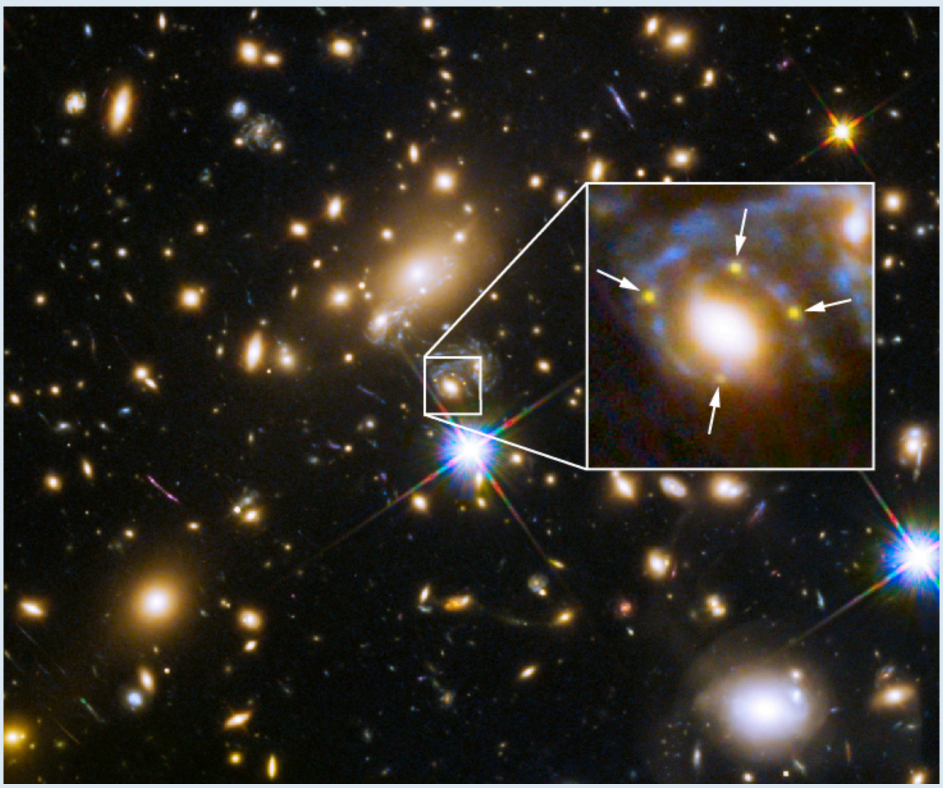


Figure 28.36 Light from a supernova at a distance of 9 billion light-years passed near a galaxy in a cluster at a distance of about 5 billion light-years. In the enlarged inset view of the galaxy, the arrows point to the multiple images of the exploding star. The images are arranged around the galaxy in a cross-shaped pattern called an Einstein Cross. The blue streaks wrapping around the galaxy are the stretched images of the supernova's host spiral galaxy, which has been distorted by the warping of space. (credit: modification of work by NASA, ESA, and S. Rodney (JHU) and the FrontierSN team; T. Treu (UCLA), P. Kelly (UC Berkeley), and the GLASS team; J. Lotz (STScI) and the Frontier Fields team; M. Postman (STScI) and the CLASH team; and Z. Levay (STScI))

General relativity predicts that the light from a distant object may also be amplified by the lensing effect, thereby making otherwise invisible objects bright enough to detect. This is particularly useful for probing the earliest stages of galaxy formation, when the universe was young. **Figure 28.37** shows an example of a very distant faint galaxy that we can study in detail only because its light path passes through a large concentration of massive galaxies and we now see a brighter image of it.

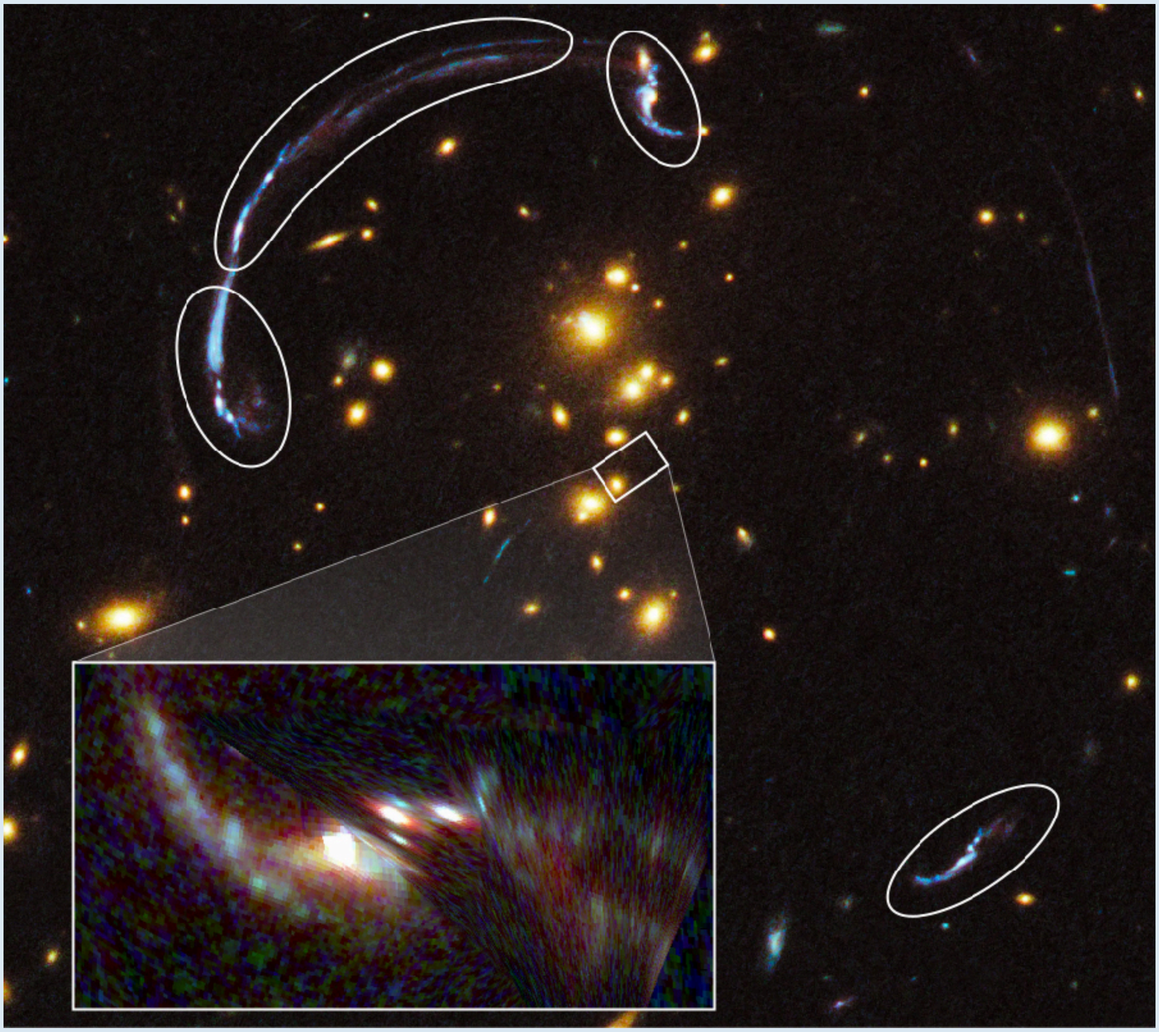


Figure 28.37 The rounded outlines show the location of distinct, distorted images of the background galaxy resulting from lensing by the mass in the cluster. The image in the box at lower left is a reconstruction of what the lensed galaxy would look like in the absence of the cluster, based on a model of the cluster's mass distribution, which can be derived from studying the distorted galaxy images. The reconstruction shows far more detail about the galaxy than could have been seen in the absence of lensing. As the image shows, this galaxy contains regions of star formation glowing like bright Christmas tree bulbs. These are much brighter than any star-formation regions in our Milky Way Galaxy. (credit: modification of work by NASA, ESA, and Z. Levay (STScI))

We should note that the visible mass in a galaxy is not the only possible gravitational lens. Dark matter can also reveal itself by producing this effect. Astronomers are using lensed images from all over the sky to learn more about where dark matter is located and how much of it exists.

Superclusters and Voids

After astronomers discovered clusters of galaxies, they naturally wondered whether there were still larger structures in the universe. Do clusters of galaxies gather together? To answer this question, we must be able to map large parts of the universe in three dimensions. We must know not only the position of each galaxy on the sky (that's two dimensions) but also its distance from us (the third dimension).

This means we must be able to measure the redshift of each galaxy in our map. Taking a spectrum of each individual galaxy to do this is a much more time-consuming task than simply counting galaxies seen in different directions on the sky, as Hubble did. Today, astronomers have found ways to get the spectra of many galaxies in the same field of view (sometimes hundreds or even thousands at a time) to cut down the time it takes to finish their three-dimensional maps. Larger telescopes are also able to measure the redshifts—and therefore the distances—of much more distant galaxies and (again) to do so much more quickly than previously possible.

Another challenge astronomers faced in deciding how to go about constructing a map of the universe is similar to that confronted by the first team of explorers in a huge, uncharted territory on Earth. Since there is only one band of explorers and an enormous amount of land, they have to make choices about where to go first. One strategy might be to strike out in a straight line in order to get a sense of the terrain. They might, for example, cross some mostly empty prairies and then hit a dense forest. As they make their way through the forest, they learn how thick it is in the direction they are traveling, but not its width to their left or right. Then a river crosses their path; as they wade across, they can measure its width but learn nothing about its length. Still, as they go on in their straight line, they begin to get some sense of what the landscape is like and can make at least part of a map. Other explorers, striking out in other directions, will someday help fill in the remaining parts of that map.

Astronomers have traditionally had to make the same sort of choices. We cannot explore the universe in every direction to infinite "depth" or sensitivity: there are far too many galaxies and far too few telescopes to do the job. But we can pick a single direction or a small slice of the sky and start mapping the galaxies. Margaret Geller, the late John Huchra, and their students at the Harvard-Smithsonian Center for Astrophysics pioneered this technique, and several other groups have extended their work to cover larger volumes of space.

Margaret Geller: Cosmic Surveyor

Born in 1947, Margaret Geller is the daughter of a chemist who encouraged her interest in science and helped her visualize the three-dimensional structure of molecules as a child. (It was a skill that would later come in very handy for visualizing the three-dimensional structure of the universe.) She remembers being bored in elementary school, but she was encouraged to read on her own by her parents. Her recollections also include subtle messages from teachers that mathematics (her strong early interest) was not a field for girls, but she did not allow herself to be deterred.

Geller obtained a BA in physics from the University of California at Berkeley and became the second woman to receive a PhD in physics from Princeton. There, while working with James Peebles, one of the world's leading cosmologists, she became interested in problems relating to the large-scale structure of the universe. In 1980, she accepted a research position at the Harvard-Smithsonian Center for Astrophysics, one of the nation's most dynamic institutions for astronomy research. She saw that to make progress in understanding how galaxies and clusters are organized, a far more intensive series of surveys was required. Although it would not bear fruit for many years, Geller and her collaborators began the long, arduous task of mapping the galaxies (**Figure 28.38**).

Figure 28.38 Geller's work mapping and researching galaxies has helped us to better understand the structure of the universe. (credit: modification of work by Massimo Ramella)

Her team was fortunate to be given access to a telescope that could be dedicated to their project, the 60-inch reflector on Mount Hopkins, near Tucson, Arizona, where they and their assistants took spectra to determine galaxy distances. To get a slice of the universe, they pointed their telescope at a predetermined position in the sky and then let the rotation of Earth bring new galaxies into their field of view. In this way, they measured the positions and redshifts of over 18,000 galaxies and made a wide range of interesting maps to display their data. Their surveys now include “slices” in both the Northern and Southern Hemispheres.

As news of her important work spread beyond the community of astronomers, Geller received a MacArthur Foundation Fellowship in 1990. These fellowships, popularly called “genius awards,” are designed to recognize truly creative work in a wide range of fields. Geller continues to have a strong interest in visualization and has (with filmmaker Boyd Estus) made several award-winning videos explaining her work to nonscientists (one is titled *So Many Galaxies . . . So Little Time*). She has appeared on a variety of national news and documentary programs, including the *MacNeil/Lehrer NewsHour*, *The Astronomers*, and *The Infinite Voyage*. Energetic and outspoken, she has given talks on her work to many audiences around the country, and works hard to find ways to explain the significance of her pioneering surveys to the public.

“It’s exciting to discover something that nobody’s seen before. [To be] one of the first three people to ever see that slice of the universe [was] sort of being like Columbus. . . . Nobody expected such a striking pattern!”—Margaret Geller

Gustavus Connections - She's a Gustie!

As interesting side notes, Dr. Geller was a speaker at the 1991 **Nobel Conference XXVII, The Evolving Cosmos (https://gustavus.edu/events/nobelconference/about/)** . Then, on June 1, 1997, she gave the Gustavus Commencement address, entitled "When You're Ten Feet Tall". On that day she was awarded an Honorary Doctor of Science degree from the College. She also served as the **Rydell Visiting Professor (https://gustavus.edu/events/rydell/MargaretGeller.php)** , in residence at Gustavus for Spring Semester 1999.

Find out more about Geller and Huchra’s work (including interviews with Geller) in this 4-minute **NOVA (https://openstax.org/l/30gellhucwork)** video. You can also learn more about their **conclusions (https://openstax.org/l/30gellhucconc)** and additional research it led to.

The largest universe mapping project to date is the Sloan Digital Sky Survey (see the Making Connections feature box **Astronomy and Technology: The Sloan Digital Sky Survey** at the end of this section). A plot of the distribution of galaxies mapped by the Sloan survey is shown in **Figure 28.39**. To the surprise of astronomers, maps like the one in the figure showed that clusters of galaxies are not arranged uniformly throughout the universe, but are found in huge filamentary **superclusters** that look like great arcs of inkblots splattered across a page. The superclusters resemble an irregularly torn sheet of paper or a pancake in shape—they can extend for hundreds of millions of light-years in two dimensions, but are only 10 to 20 million light-years thick in the third dimension. Detailed study of some of these structures shows that their masses are a few times 10^{16} M_{Sun}, which is 10,000 times more massive than the Milky Way Galaxy.

Check out this **animated visualization (https://openstax.org/l/30anivisslosur)** of large-scale structure from the Sloan survey.

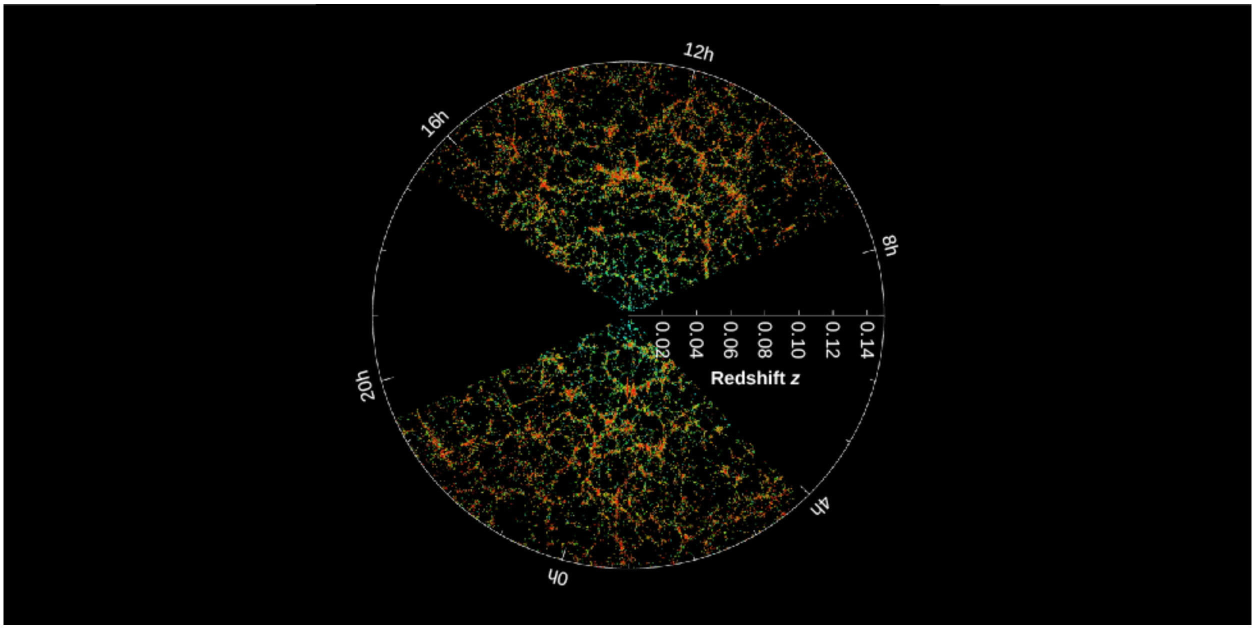


Figure 28.39 This image shows slices from the SDSS map. The point at the center corresponds to the Milky Way and might say “You Are Here!” Points on the map moving outward from the center are farther away. The distance to the galaxies is indicated by their redshifts (following Hubble’s law), shown on the horizontal line going right from the center. The redshift $z = \Delta\lambda/\lambda$, where $\Delta\lambda$ is the difference between the observed wavelength and the wavelength λ emitted by a nonmoving source in the laboratory. Hour angle on the sky is shown around the circumference of the circular graph. The colors of the galaxies indicate the ages of their stars, with the redder color showing galaxies that are made of older stars. The outer circle is at a distance of two billion light-years from us. Note that red (older stars) galaxies are more strongly clustered than blue galaxies (young stars). The unmapped areas are where our view of the universe is obstructed by dust in our own Galaxy. (credit: modification of work by M. Blanton and the Sloan Digital Sky Survey)

Separating the filaments and sheets in a supercluster are **voids**, which look like huge empty bubbles walled in by the great arcs of galaxies. They have typical diameters of 150 million light-years, with the clusters of galaxies concentrated along their walls. The whole arrangement of filaments and voids reminds us of a sponge, the inside of a honeycomb, or a hunk of Swiss cheese with very large holes. If you take a good slice or cross-section through any of these, you will see something that looks roughly like Figure 28.39.

Before these voids were discovered, most astronomers would probably have predicted that the regions between giant clusters of galaxies were filled with many small groups of galaxies, or even with isolated individual galaxies. Careful searches within these voids have found few galaxies of any kind. Apparently, 90 percent of the galaxies occupy less than 10 percent of the volume of space.

Example 28.2

Galaxy Distribution

To determine the distribution of galaxies in three-dimensional space, astronomers have to measure their positions and their redshifts. The larger the volume of space surveyed, the more likely the measurement is a fair sample of the universe as a whole. However, the work involved increases very rapidly as you increase the volume covered by the survey.

Let’s do a quick calculation to see why this is so.

Suppose that you have completed a survey of all the galaxies within 30 million light-years and you now want to survey to 60 million light-years. What volume of space is covered by your second survey? How much larger is this volume than the volume of your first survey? Remember that the volume of a sphere, V, is given by the formula $V = 4/3\pi R^3$, where R is the radius of the sphere.

Solution

Since the volume of a sphere depends on R^3 and the second survey reaches twice as far in distance, it will cover a volume that is $2^3 = 8$ times larger. The total volume covered by the second survey will be $(4/3)\pi \times (60 \text{ million light-years})^3 = 9 \times 10^{23}$ light-years^3.

28.2 Suppose you now want to expand your survey to 90 million light-years. What volume of space is covered, and how much larger is it than the volume of the second survey?

Even larger, more sensitive telescopes and surveys are currently being designed and built to peer farther and farther out in space and back in time. The new 50-meter Large Millimeter Telescope in Mexico and the Atacama Large Millimeter Array in Chile can detect far-infrared and millimeter-wave radiation from massive starbursting galaxies at redshifts and thus distances more than 90% of the way back to the Big Bang. These cannot be observed with visible light because their star formation regions are wrapped in clouds of thick dust. And in 2018, the 6.5-meter-diameter James Webb Space Telescope is scheduled to launch. It will be the first new major visible light and near-infrared telescope in space since Hubble was launched more than 25 years earlier. One of the major goals of this telescope is to observe directly the light of the first galaxies and even the first stars to shine, less than half a billion years after the Big Bang.

At this point, if you have been thinking about our discussions of the expanding universe in **Galaxies**, you may be wondering what exactly in **Figure 28.39** is expanding. We know that the galaxies and clusters of galaxies are held together by their gravity and do not expand as the universe does. However, the voids do grow larger and the filaments move farther apart as space stretches (see **Big Bang Cosmology**).

Astronomy and Technology: The Sloan Digital Sky Survey

In Edwin Hubble's day, spectra of galaxies had to be taken one at a time. The faint light of a distant galaxy gathered by a large telescope was put through a slit, and then a spectrometer (also called a spectrograph) was used to separate the colors and record the spectrum. This was a laborious process, ill suited to the demands of making large-scale maps that require the redshifts of many thousands of galaxies.

But new technology has come to the rescue of astronomers who seek three-dimensional maps of the universe of galaxies. One ambitious survey of the sky was produced using a special telescope, camera, and spectrograph atop the Sacramento Mountains of New Mexico. Called the Sloan Digital Sky Survey (SDSS), after the foundation that provided a large part of the funding, the program used a 2.5-meter telescope (about the same aperture as the Hubble) as a wide-angle astronomical camera. During a mapping program lasting more than ten years, astronomers used the SDSS's 30 charge-coupled devices (CCDs)—sensitive electronic light detectors similar to those used in many digital cameras and cell phones—to take images of over 500 million objects and spectra of over 3 million, covering more than one-quarter of the celestial sphere. Like many large projects in modern science, the Sloan Survey involved scientists and engineers from many different institutions, ranging from universities to national laboratories.

Every clear night for more than a decade, astronomers used the instrument to make images recording the position and brightness of celestial objects in long strips of the sky. The information in each strip was digitally recorded and preserved for future generations. When the seeing was only adequate, the telescope was used for taking spectra of galaxies and quasars—but it did so for up to *640 objects at a time*.

The key to the success of the project was a series of *optical fibers*, thin tubes of flexible glass that can transmit light from a source to the CCD that then records the spectrum. After taking images of a part of the sky and identifying which objects are galaxies, project scientists drilled an aluminum plate with holes for attaching fibers at the location of each galaxy. The telescope was then pointed at the right section of the sky, and the fibers led the light of each galaxy to the spectrometer for individual recording (**Figure 28.40**).

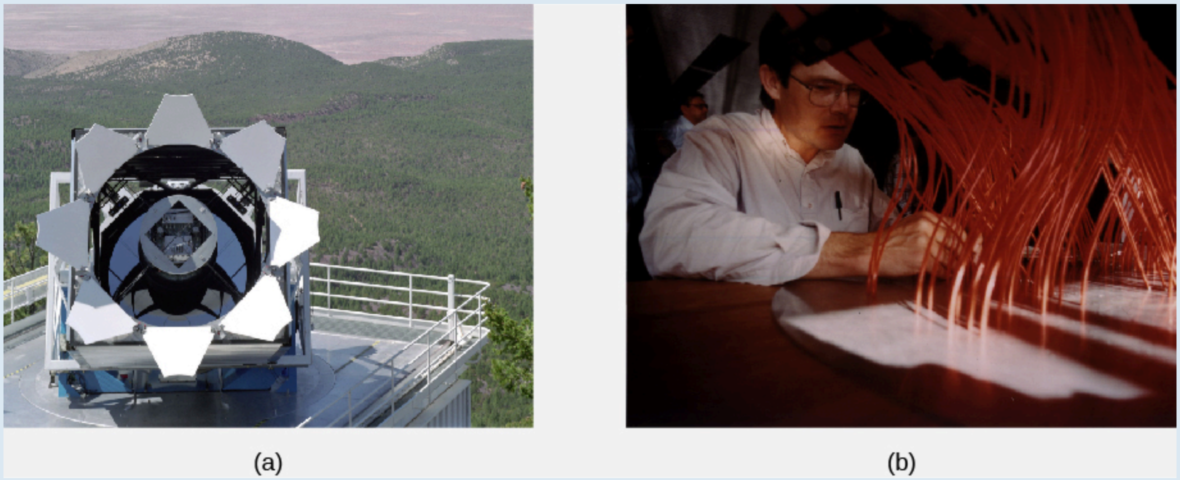


Figure 28.40 (a) The Sloan Digital Sky Survey telescope is seen here in front of the Sacramento Mountains in New Mexico. (b) Astronomer Richard Kron inserts some of the optical fibers into the pre-drilled plate to enable the instruments to make many spectra of galaxies at the same time. (credit a, b: modification of work by the Sloan Digital Sky Survey)

About an hour was sufficient for each set of spectra, and the pre-drilled aluminum plates could be switched quickly. Thus, it was possible to take as many as 5000 spectra in one night (provided the weather was good enough).

The galaxy survey led to a more comprehensive map of the sky than has ever before been possible, allowing astronomers to test their ideas about large-scale structure and the evolution of galaxies against an impressive array of real data.

The information recorded by the Sloan Survey staggers the imagination. The data came in at 8 megabytes per second (this means 8 million individual numbers or characters every second). Over the course of the project, scientists recorded over 15 terabytes, or 15 thousand billion bytes, which they estimate is comparable to the information contained in the Library of Congress. Organizing and sorting this volume of data and extracting the useful scientific results it contains is a formidable challenge, even in our information age. Like many other fields, astronomy has now entered an era of "Big Data," requiring supercomputers and advanced computer algorithms to sift through all those terabytes of data efficiently.

One very successful solution to the challenge of dealing with such large datasets is to turn to "citizen science," or crowd-sourcing, an approach the SDSS helped pioneer. The human eye is very good at recognizing subtle differences among shapes, such as between two different spiral galaxies, while computers often fail at such tasks. When Sloan project astronomers wanted to catalog the shapes of some of the millions of galaxies in their new images, they launched the "Galaxy Zoo" project: volunteers around the world were given a short training course online, then were provided with a few dozen galaxy images to classify by eye. The project was wildly successful, resulting in over 40 million galaxy classifications by more than 100,000 volunteers and the discovery of whole new types of galaxies.

Learn more about how you can be part of **the project of classifying galaxies (https://openstax.org/l/30classgalax)** in this citizen science effort. This program is part of a whole series of **"citizen science" projects (https://openstax.org/l/30citizscien)** that enable people in all walks of life to be part of the research that professional astronomers (and scholars in a growing number of fields) need help with.

28.7 | The Challenge of Dark Matter

Learning Objectives

By the end of this section, you will be able to:

- Explain how astronomers know that the solar system contains very little dark matter
- Summarize the evidence for dark matter in most galaxies
- Explain how we know that galaxy clusters are dominated by dark matter
- Relate the presence of dark matter to the average mass-to-light ratio of huge volumes of space containing many galaxies

So far this chapter has focused almost entirely on matter that radiates electromagnetic energy—stars, planets, gas, and dust. But, as we have pointed out in several earlier chapters (especially **The Milky Way Galaxy**), it is now clear that galaxies contain large amounts of dark matter as well. There is much more dark matter, in fact, than matter we can see—which means it would be foolish to ignore the effect of this unseen material in our theories about the structure of the universe. (As many a ship captain in the polar seas found out too late, the part of the iceberg visible above the ocean's surface was not necessarily the only part he needed to pay attention to.) Dark matter turns out to be extremely important in determining the evolution of galaxies and of the universe as a whole.

The idea that much of the universe is filled with dark matter may seem like a bizarre concept, but we can cite a historical example of "dark matter" much closer to home. In the mid-nineteenth century, measurements showed that the planet Uranus did not follow exactly the orbit predicted from Newton's laws if one added up the gravitational forces of all the known objects in the solar system. Some people worried that Newton's laws may simply not work so far out in our solar system. But the more straightforward interpretation was to attribute Uranus' orbital deviations to the gravitational effects of a new planet that had not yet been seen. Calculations showed where that planet had to be, and Neptune was discovered just about in the predicted location.

In the same way, astronomers now routinely determine the location and amount of dark matter in galaxies by measuring its gravitational effects on objects we can see. And, by measuring the way that galaxies move in clusters, scientists have discovered that dark matter is also distributed among the galaxies in the clusters. Since the environment surrounding a galaxy is important in its development, dark matter must play a central role in galaxy evolution as well. Indeed, it appears that dark matter makes up most of the matter in the universe. But what *is* dark matter? What is it made of? We'll look next at the search for dark matter and the quest to determine its nature.

Dark Matter in the Local Neighborhood

Is there dark matter in our own solar system? Astronomers have examined the orbits of the known planets and of spacecraft as they journey to the outer planets and beyond. No deviations have been found from the orbits predicted on the basis of the masses of objects already discovered in our solar system and the theory of gravity. We therefore conclude that there is no evidence that there are large amounts of dark matter nearby.

Astronomers have also looked for evidence of dark matter in the region of the Milky Way Galaxy that lies within a few hundred light-years of the Sun. In this vicinity, most of the stars are restricted to a thin disk. It is possible to calculate how much mass the disk must contain in order to keep the stars from wandering far above or below it. The total matter that must be in the disk is less than twice the amount of luminous matter. This means that no more than half of the mass in the region near the Sun can be dark matter.

Dark Matter in and around Galaxies

In contrast to our local neighborhood near the Sun and solar system, there is (as we saw in **The Milky Way Galaxy**) ample evidence strongly suggesting that about 90% of the mass in the entire galaxy is in the form of a halo of dark matter. In other words, there is apparently about nine times more dark matter than visible matter. Astronomers have found some stars in the outer regions of the Milky Way beyond its bright disk, and these stars are revolving very rapidly around its center. The mass contained in all the stars and all the interstellar matter we can detect in the galaxy does not exert enough gravitational force to explain how those fast-moving stars remain in their orbits and do not fly away. Only by having large amounts of unseen matter could the galaxy be holding on to those fast-moving outer stars. The same result is found for other spiral galaxies as well.

Figure 28.41 is an example of the kinds of observations astronomers are making, for the Andromeda galaxy, a member of our Local Group. The observed rotation of spiral galaxies like Andromeda is usually seen in plots, known as *rotation curves,* that show velocity versus distance from the galaxy center. Such plots suggest that the dark matter is found in a large halo surrounding the luminous parts of each galaxy. The radius of the halos around the Milky Way and Andromeda may be as large as 300,000 light-years, much larger than the visible size of these galaxies.

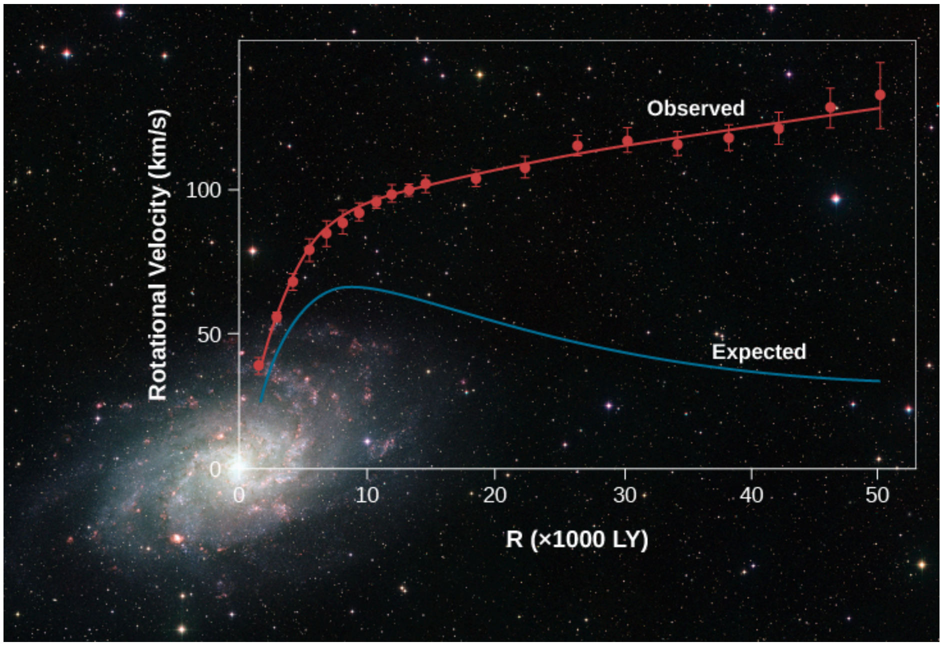


Figure 28.41 We see the Milky Way's sister, the spiral Andromeda galaxy, with a graph that shows the velocity at which stars and clouds of gas orbit the galaxy at different distances from the center (red line). As is true of the Milky Way, the rotational velocity (or orbital speed) does not decrease with distance from the center, which is what you would expect if an assembly of objects rotates around a common center. A calculation (blue line) based on the total mass visible as stars, gas, and dust predicts that the velocity should be much lower at larger distances from the center. The discrepancy between the two curves implies the presence of a halo of massive dark matter extending outside the boundary of the luminous matter. This dark matter causes everything in the galaxy to orbit faster than the observed matter alone could explain. (credit background: modification of work by ESO)

Dark Matter in Clusters of Galaxies

Galaxies in clusters also move around: they orbit the cluster's center of mass. It is not possible for us to follow a galaxy around its entire orbit because that typically takes about a billion years. It is possible, however, to measure the velocities with which galaxies in a cluster are moving, and then estimate what the total mass in the cluster must be to keep the individual galaxies from flying out of the cluster. The observations indicate that the mass of the galaxies alone cannot keep the cluster together—some other gravity must again be present. The total amount of dark matter in clusters exceeds by more than ten times the luminous mass contained within the galaxies themselves, indicating that dark matter exists between galaxies as well as inside them.

There is another approach we can take to measuring the amount of dark matter in clusters of galaxies. As we saw, the universe is expanding, but this expansion is not perfectly uniform, thanks to the interfering hand of gravity. Suppose, for example, that a galaxy lies outside but relatively close to a rich cluster of galaxies. The gravitational force of the cluster will tug on that neighboring galaxy and slow down the rate at which it moves away from the cluster due to the expansion of the universe.

Consider the Local Group of galaxies, lying on the outskirts of the Virgo Supercluster. The mass concentrated at the center of the Virgo Cluster exerts a gravitational force on the Local Group. As a result, the Local Group is moving away from the center of the Virgo Cluster at a velocity a few hundred kilometers per second slower than the Hubble law predicts. By measuring such deviations from a smooth expansion, astronomers can estimate the total amount of mass contained in large clusters.

There are two other very useful methods for measuring the amount of dark matter in galaxy clusters, and both of them have produced results in general agreement with the method of measuring galaxy velocities: gravitational lensing and X-ray emission. Let's take a look at both.

As Albert Einstein showed in his theory of general relativity, the presence of mass bends the surrounding fabric of spacetime. Light follows those bends, so very massive objects can bend light significantly. You saw examples of this in the Astronomy Basics feature box **Gravitational Lensing** in the previous section. Visible galaxies are not the only possible gravitational lenses. Dark matter can also reveal its presence by producing this effect. **Figure 28.42** shows a galaxy cluster that is acting like a gravitational lens; the streaks and arcs you see on the picture are lensed images of more distant galaxies. Gravitational lensing is well enough understood that astronomers can use the many ovals and arcs seen in this image to calculate detailed maps of how much matter there is in the cluster and how that mass is distributed. The result from studies of many such gravitational lens clusters shows that, like individual galaxies, galaxy clusters contain more than ten times as much dark matter as luminous matter.

Figure 28.42 This view from the Hubble Space Telescope shows the massive galaxy cluster Abell 2218 at a distance of about 2 billion light-years. Most of the yellowish objects are galaxies belonging to the cluster. But notice the numerous long, thin streaks, many of them blue; those are the distorted and magnified images of even more distant background galaxies, gravitationally lensed by the enormous mass of the intervening cluster. By carefully analyzing the lensed images, astronomers can construct a map of the dark matter that dominates the mass of the cluster. (credit: modification of work by NASA, ESA, and Johan Richard (Caltech))

The third method astronomers use to detect and measure dark matter in galaxy clusters is to image them in the light of X-rays. When the first sensitive X-ray telescopes were launched into orbit around Earth in the 1970s and trained on massive galaxy clusters, it was quickly discovered that the clusters emit copious X-ray radiation (see **Figure 28.43**). Most stars do not emit much X-ray radiation, and neither does most of the gas or dust between the stars inside galaxies. What could be emitting the X-rays seen from virtually all massive galaxy clusters?

It turns out that just as galaxies have gas distributed between their stars, clusters of galaxies have gas distributed between their galaxies. The particles in these huge reservoirs of gas are not just sitting still; rather, they are constantly moving, zooming around under the influence of the cluster's immense gravity like mini planets around a giant sun. As they move and bump against each other, the gas heats up hotter and hotter until, at temperatures as high as 100 million K, it shines brightly at X-ray wavelengths. The more mass the cluster has, the faster the motions, the hotter the gas, and the brighter the X-rays. Astronomers calculate that the mass present to induce those motions must be about ten times the mass they can see in the clusters, including all the galaxies and all the gas. Once again, this is evidence that the galaxy clusters are seen to be dominated by dark matter.

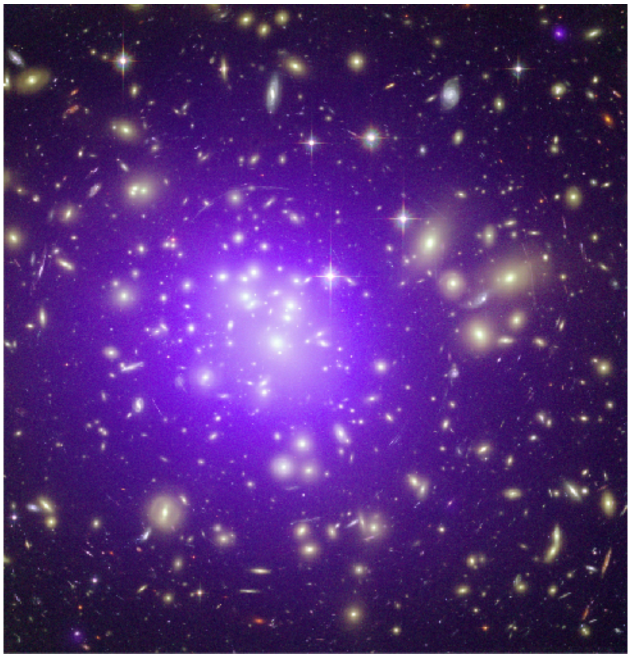


Figure 28.43 This composite image shows the galaxy cluster Abell 1689 at a distance of 2.3 billion light-years. The finely detailed views of the galaxies, most of them yellow, are in visible and near-infrared light from the Hubble Space Telescope, while the diffuse purple haze shows X-rays as seen by Chandra X-ray Observatory. The abundant X-rays, the gravitationally lensed images (thin curving arcs) of background galaxies, and the measured velocities of galaxies in the clusters all show that the total mass of Abell 1689—most of it dark matter—is about 10^{15} solar masses. (credit: modification of work by NASA/ESA/JPL-Caltech/Yale/CNRS)

Mass-to-Light Ratio

We described the use of the mass-to-light ratio to characterize the matter in galaxies or clusters of galaxies in **Properties of Galaxies**. For systems containing mostly old stars, the mass-to-light ratio is typically 10 to 20, where mass and light are measured in units of the Sun's mass and luminosity. A mass-to-light ratio of 100 or more is a signal that a substantial amount of dark matter is present. **Table 28.1** summarizes the results of measurements of mass-to-light ratios for various classes of objects. Very large mass-to-light ratios are found for all systems of galaxy size and larger, indicating that dark matter is present in all these types of objects. This is why we say that dark matter apparently makes up most of the total mass of the universe.

Mass-To-Light Ratios

Type of Object	Mass-to-Light Ratio
Sun	1
Matter in vicinity of Sun	2
Mass in Milky Way within 80,000 light-years of the center	10
Small groups of galaxies	50–150
Rich clusters of galaxies	250–300

Table 28.1

The clustering of galaxies can be used to derive the total amount of mass in a given region of space, while visible radiation is a good indicator of where the luminous mass is. Studies show that the dark matter and luminous matter are very closely associated. The dark matter halos do extend beyond the luminous boundaries of the galaxies that they surround. However, where there are large clusters of galaxies, you will also find large amounts of dark matter. Voids in the galaxy distribution are also voids in the distribution of dark matter.

What Is the Dark Matter?

How do we go about figuring out what the dark matter consists of? The technique we might use depends on its composition. Let's consider the possibility that some of the dark matter is made up of normal particles: protons, neutrons, and electrons. Suppose these particles were assembled into black holes, brown dwarfs, or white dwarfs. If the black holes had no accretion disks, they would be invisible to us. White and brown dwarfs do emit some radiation but have such low luminosities that they cannot be seen at distances greater than a few thousand light-years.

We can, however, look for such compact objects because they can act as gravitational lenses. (See the Astronomy Basics feature box **Gravitational Lensing**.) Suppose the dark matter in the halo of the Milky Way were made up of black holes, brown dwarfs, and white dwarfs. These objects have been whimsically dubbed MACHOs (MAssive Compact Halo Objects). If an invisible MACHO passes directly between a distant star and Earth, it acts as a gravitational lens, focusing the light from the distant star. This causes the star to appear to brighten over a time interval of a few hours to several days before returning to its normal brightness. Since we can't predict when any given star might brighten this way, we have to monitor huge numbers of stars to catch one in the act. There are not enough astronomers to keep monitoring so many stars, but today's automated telescopes and computer systems can do it for us.

Research teams making observations of millions of stars in the nearby galaxy called the Large Magellanic Cloud have reported several examples of the type of brightening expected if MACHOs are present in the halo of the Milky Way (**Figure 28.44**). However, there are not enough MACHOs in the halo of the Milky Way to account for the mass of the dark matter in the halo.

Figure 28.44 Here, the two small galaxies we call the Large Magellanic Cloud and Small Magellanic Cloud can be seen above the auxiliary telescopes for the Very Large Telescope Array on Cerro Paranal in Chile. You can see from the number of stars that are visible that this is a very dark site for doing astronomy. (credit: ESO/J. Colosimo)

This result, along with a variety of other experiments, leads us to conclude that the types of matter we are familiar with can make up only a tiny portion of the dark matter. Another possibility is that dark matter is composed of some new type of particle—one that researchers are now trying to detect in laboratories here on Earth (see **Big Bang Cosmology**).

The kinds of dark matter particles that astronomers and physicists have proposed generally fall into two main categories: hot and cold dark matter. The terms *hot* and *cold* don't refer to true temperatures, but rather to the average velocities of the particles, analogous to how we might think of particles of air moving in your room right now. In a cold room, the air particles move more slowly on average than in a warm room.

In the early universe, if dark matter particles easily moved fast and far compared to the lumps and bumps of ordinary matter that eventually became galaxies and larger structures, we call those particles **hot dark matter**. In that case, smaller lumps and bumps would be smeared out by the particle motions, meaning fewer small galaxies would get made.

On the other hand, if the dark matter particles moved slowly and covered only small distances compared to the sizes of the lumps in the early universe, we call that **cold dark matter**. Their slow speeds and energy would mean that even the smaller lumps of ordinary matter would survive to grow into small galaxies. By looking at when galaxies formed and how they evolve, we can use observations to distinguish between the two kinds of dark matter. So far, observations seem most consistent with models based on cold dark matter.

Solving the dark matter problem is one of the biggest challenges facing astronomers. After all, we can hardly understand the evolution of galaxies and the long-term history of the universe without understanding what its most massive component is made of. For example, we need to know just what role dark matter played in starting the higher-density "seeds" that led to the formation of galaxies. And since many galaxies have large halos made of dark matter, how does this affect their interactions with one another and the shapes and types of galaxies that their collisions create?

Astronomers armed with various theories are working hard to produce models of galaxy structure and evolution that take dark matter into account in just the right way. Even though we don't know what the dark matter is, we do have some clues about how it affected the formation of the very first galaxies. As we will see in **Big Bang Cosmology**, careful measurements of the microwave radiation left over after the Big Bang have allowed astronomers to set very tight limits on the actual sizes of those early seeds that led to the formation of the large galaxies that we see in today's universe. Astronomers have also measured the relative numbers and distances between galaxies and clusters of different sizes in the universe today. So far, most of the evidence seems to weigh heavily in favor of cold dark matter, and most current models of galaxy and large-scale structure formation use cold dark matter as their main ingredient.

As if the presence of dark matter—a mysterious substance that exerts gravity and outweighs all the known stars and galaxies in the universe but does not emit or absorb light—were not enough, there is an even more baffling and equally important constituent of the universe that has only recently been discovered: we have called it **dark energy** in parallel with dark matter. We will say more about it and explore its effects on the evolution of the universe in **Big Bang Cosmology**. For now, we can complete our inventory of the contents of the universe by noting that it appears that the entire universe contains some mysterious energy that pushes spacetime apart, taking galaxies and the larger structures made of galaxies along with it. Observations show that dark energy becomes more and more important relative to gravity as the universe ages. As a result, the expansion of the universe is accelerating, and this acceleration seems to be happening mostly since the universe was about half its current age.

What we see when we peer out into the universe—the light from trillions of stars in hundreds of billions of galaxies wrapped in intricate veils of gas and dust—is therefore actually only a sprinkling of icing on top of the cake: as we will see in **Big Bang Cosmology**, when we look outside galaxies and clusters of galaxies at the universe as a whole, astronomers find that for every gram of luminous normal matter, such as protons, neutrons, electrons, and atoms in the universe, there are about 4 grams of nonluminous normal matter, mainly intergalactic hydrogen and helium. There are about 27 grams of dark matter, and the energy equivalent (remember Einstein's famous $E = mc^2$) of about 68 grams of dark energy. Dark matter, and (as we will see) even more so dark energy, are dramatic demonstrations of what we have tried to emphasize throughout this book: science is always a "progress report," and we often encounter areas where we have more questions than answers.

Let's next put together all these clues to trace the life history of galaxies and large-scale structure in the universe. What follows is the current consensus, but research in this field is moving rapidly, and some of these ideas will probably be modified as new observations are made.

28.8 | The Formation and Evolution of Galaxies and Structure in the Universe

Learning Objectives

By the end of this section, you will be able to:

- Summarize the main theories attempting to explain how individual galaxies formed
- Explain how tiny "seeds" of dark matter in the early universe grew by gravitational attraction over billions of years into the largest structures observed in the universe: galaxy clusters and superclusters, filaments, and voids

As with most branches of natural science, astronomers and cosmologists always want to know the answer to the question, "How did it get that way?" What made galaxies and galaxy clusters, superclusters, voids, and filaments look the way they do? The existence of such large filaments of galaxies and voids is an interesting puzzle because we have evidence (to be discussed in **Big Bang Cosmology**) that the universe was extremely smooth even a few hundred thousand years after forming. The challenge for theoreticians is to understand how a nearly featureless universe changed into the complex and lumpy one that we see today. Armed with our observations and current understanding of galaxy evolution over cosmic time, dark matter, and large-scale structure, we are now prepared to try to answer that question on some of the largest possible scales in the universe. As we will see, the short answer to how the universe got this way is "dark matter + gravity + time."

How Galaxies Form and Grow

We've already seen that galaxies were more numerous, but smaller, bluer, and clumpier, in the distant past than they are today, and that galaxy mergers play a significant role in their evolution. At the same time, we have observed quasars and galaxies that emitted their light when the universe was less than a billion years old—so we know that large condensations of matter had begun to form at least that early. We also saw in **Supermassive Black Holes** that many quasars are found in the centers of elliptical galaxies. This means that some of the first large concentrations of matter must have evolved into the elliptical galaxies that we see in today's universe. It seems likely that the supermassive black holes in the centers of galaxies and the spherical distribution of ordinary matter around them formed at the same time and through related physical processes.

Dramatic confirmation of that picture arrived only in the last decade, when astronomers discovered a curious empirical relationship: as we saw in **Supermassive Black Holes**, the more massive a galaxy is, the more massive its central black hole is. Somehow, the black hole and the galaxy "know" enough about each other to match their growth rates.

There have been two main types of galaxy formation models to explain all those observations. The first asserts that massive elliptical galaxies formed in a single, rapid collapse of gas and dark matter, during which virtually all the gas was turned quickly into stars. Afterward the galaxies changed only slowly as the stars evolved. This is what astronomers call a "top-down" scenario.

The second model suggests that today's giant ellipticals were formed mostly through mergers of smaller galaxies that had already converted at least some of their gas into stars—a "bottom-up" scenario. In other words, astronomers have debated whether giant ellipticals formed most of their stars in the large galaxy that we see today or in separate small galaxies that subsequently merged.

Since we see some luminous quasars from when the universe was less than a billion years old, it is likely that at least some giant ellipticals began their evolution very early through the collapse of a single cloud. However, the best evidence also seems to show that mature *giant* elliptical galaxies like the ones we see nearby were rare before the universe was about 6 billion years old and that they are much more common today than they were when the universe was young. Observations also indicate that most of the gas in elliptical galaxies was converted to stars by the time the universe was about 3 billion years old, so it appears that elliptical galaxies have not formed many new stars since then. They are often said to be "red and dead"—that is, they mostly contain old, cool, red stars, and there is little or no new star formation going on.

These observations (when considered together) suggest that the giant elliptical galaxies that we see nearby formed from a combination of both top-down and bottom-up mechanisms, with the most massive galaxies forming in the densest clusters where both processes happened very early and quickly in the history of the universe.

The situation with spiral galaxies is apparently very different. The bulges of these galaxies formed early, like the elliptical galaxies (**Figure 28.45**). However, the disks formed later (remember that the stars in the disk of the Milky Way are younger than the stars in the bulge and the halo) and still contain gas and dust. However, the rate of star formation in spirals today is about ten times lower than it was 8 billion years ago. The number of stars being formed drops as the gas is used up. So spirals seem to form mostly "bottom up" but over a longer time than ellipticals and in a more complex way, with at least two distinct phases.

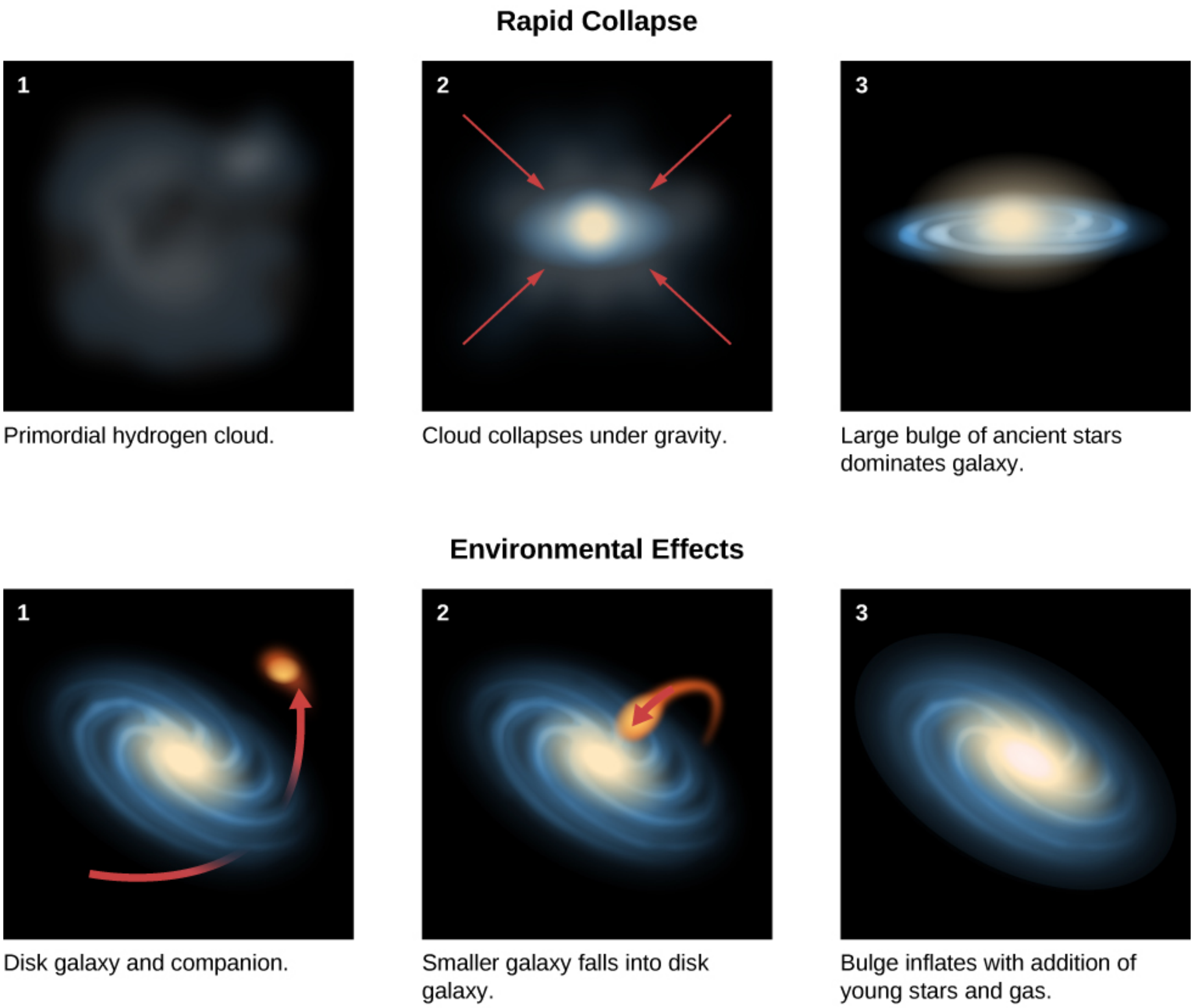


Figure 28.45 The nuclear bulges of some spiral galaxies formed through the collapse of a single protogalactic cloud (top row). Others grew over time through mergers with other smaller galaxies (bottom row).

Hubble originally thought that elliptical galaxies were young and would eventually turn into spirals, an idea we now know is not true. In fact, as we saw above, it's more likely the other way around: two spirals that crash together under their mutual gravity can turn into an elliptical.

Despite these advances in our understanding of how galaxies form and evolve, many questions remain. For example, it's even possible, given current evidence, that spiral galaxies could lose their spiral arms and disks in a merger event, making them look more like an elliptical or irregular galaxy, and then regain the disk and arms again later if enough gas remains available. The story of how galaxies assume their final shapes is still being written as we learn more about galaxies and their environment.

Forming Galaxy Clusters, Superclusters, Voids, and Filaments

If individual galaxies seem to grow mostly by assembling smaller pieces together gravitationally over cosmic time, what about the clusters of galaxies and larger structures such as those seen in **Figure 28.39**? How do we explain the large-scale maps that show galaxies distributed on the walls of huge sponge- or bubble-like structures spanning hundreds of millions of light-years?

As we saw, observations have found increasing evidence for concentrations, filaments, clusters, and superclusters of galaxies when the universe was less than 3 billion years old (**Figure 28.46**). This means that large concentrations of galaxies had already come together when the universe was less than a quarter as old as it is now.

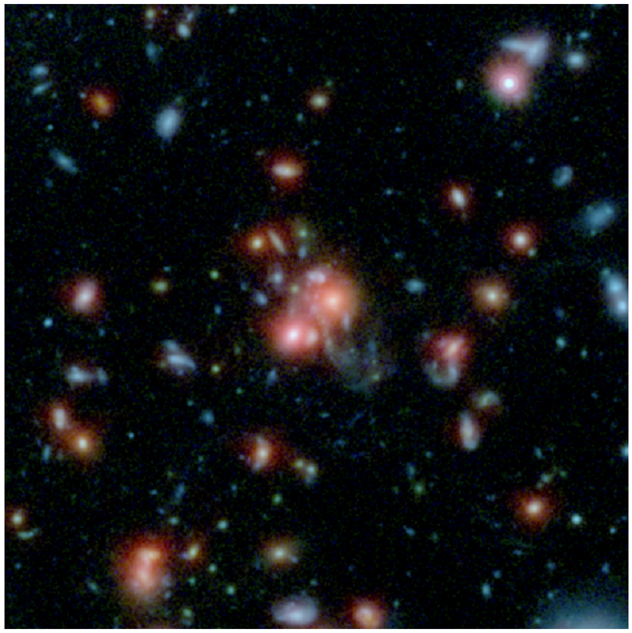


Figure 28.46 This Hubble image shows the core of one of the most distant galaxy clusters yet discovered, SpARCS 1049+56; we are seeing it as it was nearly 10 billion years ago. The surprise delivered by the image was the “train wreck” of chaotic galaxy shapes and blue tidal tails: apparently there are several galaxies right in the core that are merging together, the probable cause of a massive burst of star formation and bright infrared emission from the cluster. (credit: modification of work by NASA/STScI/ESA/JPL-Caltech/McGill)

Almost all the currently favored models of how large-scale structure formed in the universe tell a story similar to that for individual galaxies: tiny dark matter “seeds” in the hot cosmic soup after the Big Bang grew by gravity into larger and larger structures as cosmic time ticked on (**Figure 28.47**). The final models we construct will need to be able to explain the size, shape, age, number, and spatial distribution of galaxies, clusters, and filaments—not only today, but also far back in time. Therefore, astronomers are working hard to measure and then to model those features of large-scale structure as accurately as possible. So far, a mixture of 5% normal atoms, 27% cold dark matter, and 68% dark energy seems to be the best way to explain all the evidence currently available (see **Big Bang Cosmology**).

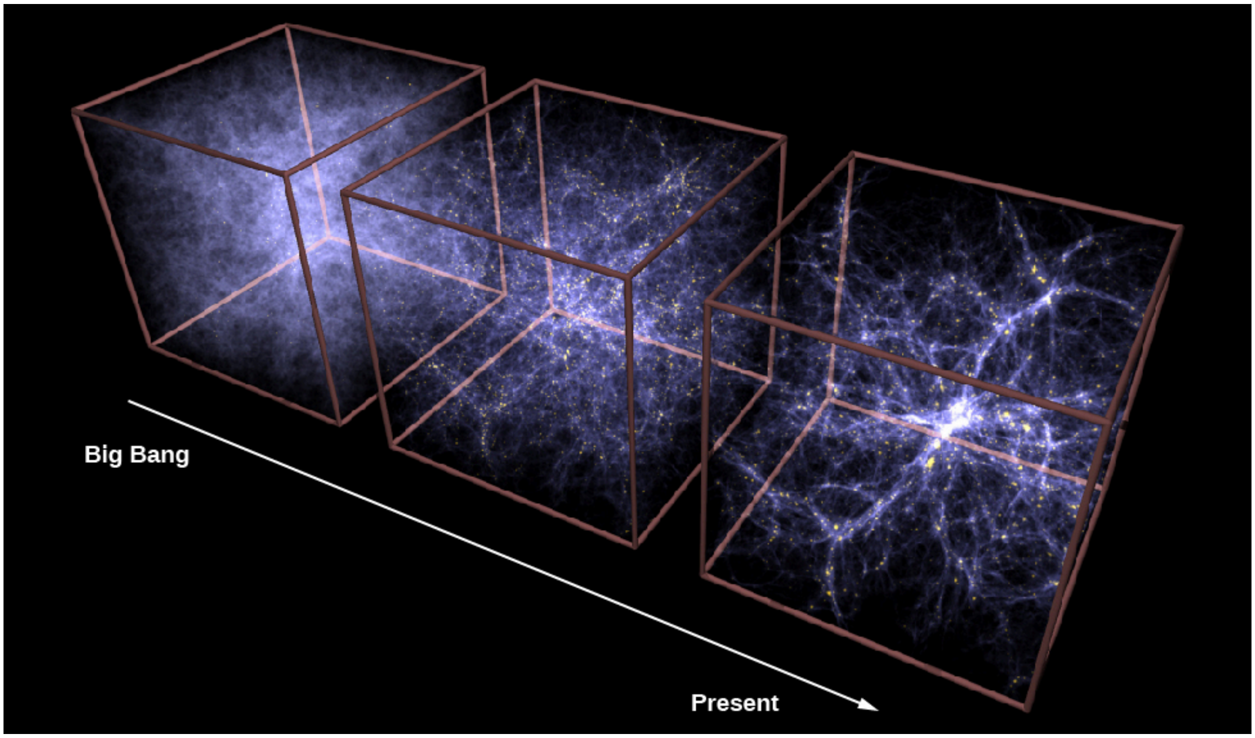


Figure 28.47 The boxes show how filaments and superclusters of galaxies grow over time, from a relatively smooth distribution of dark matter and gas, with few galaxies formed in the first 2 billion years after the Big Bang, to the very clumpy strings of galaxies with large voids today. Compare the last image in this sequence with the actual distribution of nearby galaxies shown in Figure 28.39. (credit: modification of work by CXC/MPE/V.Springel)

The box at left is labeled "Big Bang," the box at center is unlabeled and the box at right is labeled "Present". A white arrow points from left to right representing the direction of time.

Scientists even have a model to explain how a nearly uniform, hot "soup" of particles and energy at the beginning of time acquired the Swiss-cheese-like structure that we now see on the largest scales. As we will see in **Big Bang Cosmology**, when the universe was only a few hundred thousand years old, *everything* was at a temperature of a few thousand degrees. Theorists suggest that at that early time, all the hot gas was vibrating, much as sound waves vibrate the air of a nightclub with an especially loud band. This vibrating could have concentrated matter into high-density peaks and created emptier spaces between them. When the universe cooled, the concentrations of matter were "frozen in," and galaxies ultimately formed from the matter in these high-density regions.

The Big Picture

To finish this chapter, let's put all these ideas together to tell a coherent story of how the universe came to look the way it does. Initially, as we said, the distribution of matter (both luminous and dark) was nearly, but not quite exactly, smooth and uniform. That "not quite" is the key to everything. Here and there were lumps where the density of matter (both luminous and dark) was ever so slightly higher than average.

Initially, each individual lump expanded because the whole universe was expanding. However, as the universe continued to expand, the regions of higher density acquired still more mass because they exerted a slightly larger than average gravitational force on surrounding material. If the inward pull of gravity was high enough, the denser individual regions ultimately stopped expanding. They then began to collapse into irregularly shaped blobs (that's the technical term astronomers use!). In many regions the collapse was more rapid in one direction, so the concentrations of matter were not spherical but came to resemble giant clumps, pancakes, and rope-like filaments—each much larger than individual galaxies.

These elongated clumps existed throughout the early universe, oriented in different directions and collapsing at different rates. The clumps provided the framework for the large-scale filamentary and bubble-like structures that we see preserved in the universe today.

The universe then proceeded to "build itself" from the bottom up. Within the clumps, smaller structures formed first, then merged to build larger ones, like Lego pieces being put together one by one to create a giant Lego metropolis. The first dense concentrations of matter that collapsed were the size of small dwarf galaxies or globular clusters—which helps explain why globular clusters are the oldest things in the Milky Way and most other galaxies. These fragments then gradually assembled to build galaxies, galaxy clusters, and, ultimately, superclusters of galaxies.

According to this picture, small galaxies and large star clusters first formed in the highest density regions of all—the filaments and nodes where the pancakes intersect—when the universe was about two percent of its current age. Some stars may have formed even before the first star clusters and galaxies came into existence. Some galaxy-galaxy collisions triggered massive bursts of star formation, and some of these led to the formation of black holes. In that rich, crowded environment, black holes found constant food and grew in mass. The development of massive black holes then triggered quasars and other active galactic nuclei whose powerful outflows of energy and matter shut off the star formation in their host galaxies. The early universe must have been an exciting place!

Clusters of galaxies then formed as individual galaxies congregated, drawn together by their mutual gravitational attraction (**Figure 28.48**). First, a few galaxies came together to form groups, much like our own Local Group. Then the groups began combining to form clusters and, eventually, superclusters. This model predicts that clusters and superclusters should still be in the process of gathering together, and observations do in fact suggest that clusters are still gathering up their flocks of galaxies and collecting more gas as it flows in along filaments. In some instances we even see entire clusters of galaxies merging together.

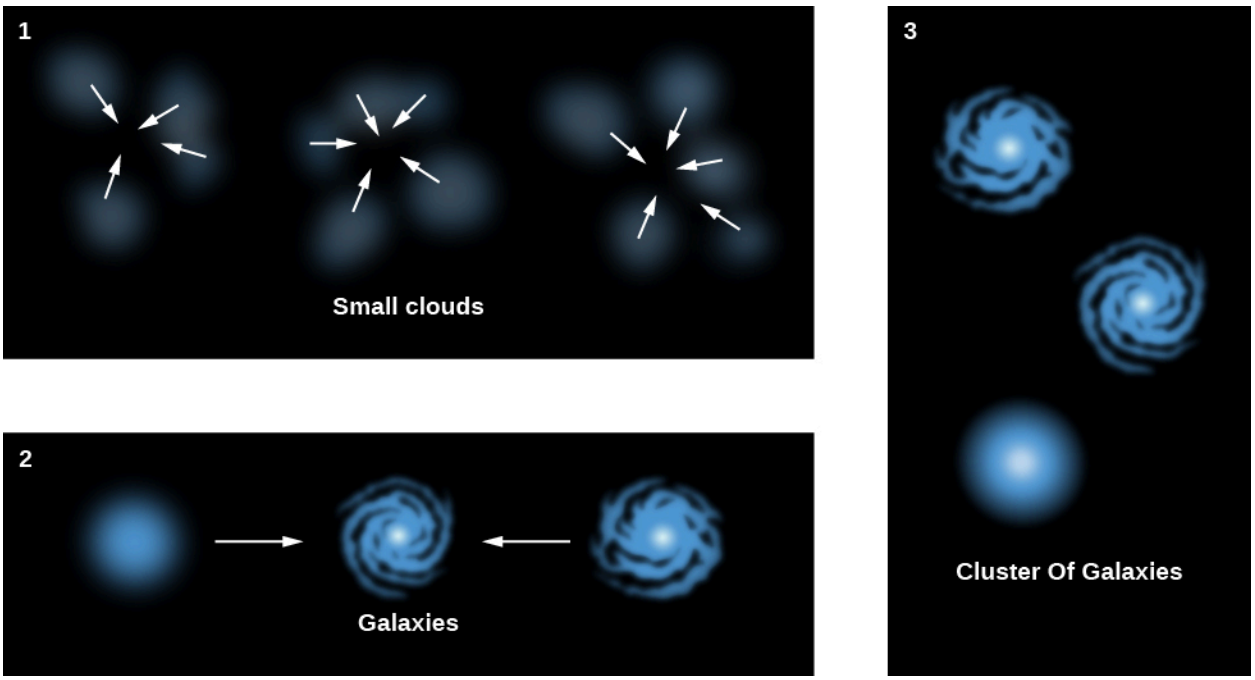


Figure 28.48 This schematic diagram shows how galaxies might have formed if small clouds formed first and then congregated to form galaxies and then clusters of galaxies.

Most giant elliptical galaxies formed through the collision and merger of many smaller fragments. Some spiral galaxies may have formed in relatively isolated regions from a single cloud of gas that collapsed to make a flattened disk, but others acquired additional stars, gas, and dark matter through collisions, and the stars acquired through these collisions now populate their halos and bulges. As we have seen, our Milky Way is still capturing small galaxies and adding them to its halo, and probably also pulling fresh gas from these galaxies into its disk.

For Further Exploration

Websites

Monsters in Galactic Nuclei: **http://chandra.as.utexas.edu/stardate.html (http://chandra.as.utexas.edu/stardate.html)** . An article on supermassive black holes by John Kormendy, from *StarDate* magazine.

Quasar Astronomy Forty Years On: **http://www.astr.ua.edu/keel/agn/quasar40.html (http://www.astr.ua.edu/keel/agn/quasar40.html)** . A 2003 popular article by William Keel.

Quasars and Active Galactic Nuclei: **www.astr.ua.edu/keel/agn/**. An annotated gallery of images showing the wide range of activity in galaxies. There is also an introduction, a glossary, and background information. Also by William Keel.

Quasars: "The Light Fantastic": **http://hubblesite.org/newscenter/archive/releases/1996/35/background/ (http://hubblesite.org/newscenter/archive/releases/1996/35/background/)** . This brief "backgrounder" from the public information office at the HubbleSite gives a bit of the history of the discovery and understanding of quasars.

Assembly of Galaxies: **http://jwst.nasa.gov/galaxies.html (http://jwst.nasa.gov/galaxies.html)** . Introductory background information about galaxies: what we know and what we want to learn.

Brief History of Gravitational Lensing: **http://www.einstein-online.info/spotlights/grav_lensing_history (http://www.einstein-online.info/spotlights/grav_lensing_history)** . From Einstein OnLine.

Cosmic Structures: **http://skyserver.sdss.org/dr1/en/astro/structures/structures.asp (http://skyserver.sdss.org/dr1/en/astro/structures/structures.asp)** . Brief review page on how galaxies are organized, from the Sloan Survey.

Discovery of the First Gravitational Lens: **http://astrosociety.org/wp-content/uploads/2013/02/ab2009-33.pdf (http://astrosociety.org/wp-content/uploads/2013/02/ab2009-33.pdf)** . By Ray Weymann, 2009.

Gravitational Lensing Discoveries from the Hubble Space Telescope: **http://hubblesite.org/newscenter/archive/releases/exotic/gravitational-lens/ (http://hubblesite.org/newscenter/archive/releases/exotic/gravitational-lens/)** . A chronological list of news releases and images.

Local Group of Galaxies: **http://www.atlasoftheuniverse.com/localgr.html (http://www.atlasoftheuniverse.com/localgr.html)** . Clickable map from the Atlas of the Universe project. See also their Virgo Cluster page: **http://www.atlasoftheuniverse.com/galgrps/vir.html (http://www.atlasoftheuniverse.com/galgrps/vir.html)** .

RotCurve: **http://burro.astr.cwru.edu/JavaLab/RotcurveWeb/main.html (http://burro.astr.cwru.edu/JavaLab/RotcurveWeb/main.html)** . Try your hand at using real galaxy rotation curve data to measure dark matter halos using this Java applet simulation.

Sloan Digital Sky Survey Website: **http://classic.sdss.org/ (http://classic.sdss.org/)** . Includes nontechnical and technical parts.

Spyglasses into the Universe: **http://www.spacetelescope.org/science/gravitational_lensing/ (http://www.spacetelescope.org/science/gravitational_lensing/)** . Hubble page on gravitational lensing; includes links to videos.

Virgo Cluster of Galaxies: **http://messier.seds.org/more/virgo.html (http://messier.seds.org/more/virgo.html)** . A page with brief information and links to maps, images, etc.

Videos

Active Galaxies: **https://vimeo.com/21079798 (https://vimeo.com/21079798)** . Part of the *Astronomy: Observations and Theories* series; half-hour introduction to quasars and related objects (27:28).

Black Hole Chaos: The Environments of the Most Supermassive Black Holes in the Universe: **https://www.youtube.com/watch?v=hzSgU-3d8QY (https://www.youtube.com/watch?v=hzSgU-3d8QY)** . May 2013 lecture by Dr. Belinda Wilkes and Dr. Francesca Civano of the Center for Astrophysics in the CfA Observatory Nights Lecture Series (50:14).

Hubble and Black Holes: **http://www.spacetelescope.org/videos/hubblecast43a/ (http://www.spacetelescope.org/videos/hubblecast43a/)** . Hubblecast on black holes and active galactic nuclei (9:10).

Monster Black Holes: **https://www.youtube.com/watch?v=LN9oYjNKBm8 (https://www.youtube.com/watch?v=LN9oYjNKBm8)** . May 2013 lecture by Professor Chung-Pei Ma of the University of California, Berkeley; part of the Silicon Valley Astronomy Lecture Series (1:18:03).

Cosmic Simulations: **http://www.tapir.caltech.edu/~phopkins/Site/Movies_cosmo.html (http://www.tapir.caltech.edu/~phopkins/Site/Movies_cosmo.html)** . Beautiful videos with computer simulations of how galaxies form, from the FIRE group.

Cosmology of the Local Universe: **http://irfu.cea.fr/cosmography (http://irfu.cea.fr/cosmography)** . Narrated flythrough of maps of galaxies showing the closer regions of the universe (17:35).

Gravitational Lensing: **https://www.youtube.com/watch?v=4Z71RtwoOas (https://www.youtube.com/watch?v=4Z71RtwoOas)** . Video from Fermilab, with Dr. Don Lincoln (7:14).

How Galaxies Were Cooked from the Primordial Soup: **https://www.youtube.com/watch?v=wqNNCm7SNyw (https://www.youtube.com/watch?v=wqNNCm7SNyw)** . A 2013 public talk by Dr. Sandra Faber of Lick Observatory about the evolution of galaxies; part of the Silicon Valley Astronomy Lecture Series (1:19:33).

Hubble Extreme Deep Field Pushes Back Frontiers of Time and Space: **https://www.youtube.com/watch?v=gu_VhzhlqGw (https://www.youtube.com/watch?v=gu_VhzhlqGw)** . Brief 2012 video (2:42).

Looking Deeply into the Universe in 3-D: **https://www.eso.org/public/videos/eso1507a/ (https://www.eso.org/public/videos/eso1507a/)** . 2015 ESOCast video on how the Very Large Telescopes are used to explore the Hubble Ultra-Deep Field and learn more about the faintest and most distant galaxies (5:12).

Millennium Simulation: **http://wwwmpa.mpa-garching.mpg.de/galform/virgo/millennium (http://wwwmpa.mpa-garching.mpg.de/galform/virgo/millennium)** . A supercomputer in Germany follows the evolution of a representative large box as the universe evolves.

Movies of flying through the large-scale local structure: **http://www.ifa.hawaii.edu/~tully/ (http://www.ifa.hawaii.edu/~tully/)** . By Brent Tully.

Shedding Light on Dark Matter: **https://www.youtube.com/watch?v=bZW_B9CC-gI (https://www.youtube.com/watch?v=bZW_B9CC-gI)** . 2008 TED talk on galaxies and dark matter by physicist Patricia Burchat (17:08).

Sloan Digital Sky Survey overview movies: **http://astro.uchicago.edu/cosmus/projects/sloanmovie/ (http://astro.uchicago.edu/cosmus/projects/sloanmovie/)** .

Virtual Universe: **https://www.youtube.com/watch?v=SY0bKE10ZDM (https://www.youtube.com/watch?v=SY0bKE10ZDM)** . An MIT model of a section of universe evolving, with dark matter included (4:11).

When Two Galaxies Collide: **http://www.openculture.com/2009/04/when_galaxies_collide.html (http://www.openculture.com/2009/04/when_galaxies_collide.html)** . Computer simulation, which stops at various points and shows a Hubble image of just such a system in nature (1:37).

CHAPTER 28 REVIEW

KEY TERMS

active galactic nuclei (AGN) galaxies that are almost as luminous as quasars and share many of their properties, although to a less spectacular degree; abnormal amounts of energy are produced in their centers

active galaxies galaxies that house active galactic nuclei

cold dark matter slow-moving massive particles, not yet identified, that don't absorb, emit, or reflect light or other electromagnetic radiation

cosmological principle the assumption that, on the large scale, the universe at any given time is the same everywhere—isotropic and homogeneous

dark energy an energy that is causing the expansion of the universe to accelerate; the source of this energy is not yet understood

evolution (of galaxies) changes in individual galaxies over cosmic time, inferred by observing snapshots of many different galaxies at different times in their lives

galactic cannibalism a process by which a larger galaxy strips material from or completely swallows a smaller one

homogeneous having a consistent and even distribution of matter that is the same everywhere

hot dark matter massive particles, not yet identified, that don't absorb, emit, or reflect light or other electromagnetic radiation; hot dark matter is faster-moving material than cold dark matter

isotropic the same in all directions

Local Group a small cluster of galaxies to which our Galaxy belongs

merger a collision between galaxies (of roughly comparable size) that combine to form a single new structure

quasar an object of very high redshift that looks like a star but is extragalactic and highly luminous; also called a quasi-stellar object, or QSO

starburst a galaxy or merger of multiple galaxies that turns gas into stars much faster than usual

supercluster a large region of space (more than 100 million light-years across) where groups and clusters of galaxies are more concentrated; a cluster of clusters of galaxies

void a region between clusters and superclusters of galaxies that appears relatively empty of galaxies

KEY EQUATIONS

Redshift $z = \frac{\Delta\lambda}{\lambda} = \frac{v}{c}$

Relativistic Doppler Formula $\frac{v}{c} = \frac{(z+1)^2 - 1}{(z+1)^2 + 1}$

Insert paragraph text here.

SUMMARY

28.1 Quasars

- The first quasars discovered looked like stars but had strong radio emission. Their visible-light spectra at first seemed confusing, but then astronomers realized that they had much larger redshifts than stars.
- The quasar spectra obtained so far show redshifts ranging from 15% to more than 96% the speed of light.
- Observations with the Hubble Space Telescope show that quasars lie at the centers of galaxies and that both spirals and ellipticals can harbor quasars.

- The redshifts of the underlying galaxies match the redshifts of the quasars embedded in their centers, thereby proving that quasars obey the Hubble law and are at the great distances implied by their redshifts.
- To be noticeable at such great distances, quasars must have 10 to 100 times the luminosity of the brighter normal galaxies.
- Their variations show that this tremendous energy output is generated in a small volume—in some cases, in a region not much larger than our own solar system.
- A number of galaxies closer to us also show strong activity at their centers—activity now known to be caused by the same mechanism as the quasars.

28.2 Supermassive Black Holes

- Both active galactic nuclei and quasars derive their energy from material falling toward, and forming a hot accretion disk around, a massive black hole.
- This model can account for the large amount of energy emitted and for the fact that the energy is produced in a relatively small volume of space.
- It can also explain why jets coming from these objects are seen in two directions: those directions are perpendicular to the accretion disk.

28.3 Quasars as Probes of Evolution in the Universe

- Quasars and galaxies affect each other: the galaxy supplies fuel to the black hole, and the quasar heats and disrupts the gas clouds in the galaxy.
- The balance between these two processes probably helps explain why the black hole seems always to be about 1/200 the mass of the spherical bulge of stars that surrounds the black hole.
- Quasars were much more common billions of years ago than they are now, and astronomers speculate that they mark an early stage in the formation of galaxies.
- Quasars were more likely to be active when the universe was young and fuel for their accretion disk was more available.
- Quasar activity can be re-triggered by a collision between two galaxies, which provides a new source of fuel to feed the black hole.

28.4 Observations of Distant Galaxies

- When we look at distant galaxies, we are looking back in time.
- We have now seen galaxies as they were when the universe was about 500 million years old—only about four percent as old as it is now.
- The universe now is 13.8 billion years old.
- The color of a galaxy is an indicator of the age of the stars that populate it.
- Blue galaxies must contain a lot of hot, massive, young stars.
- Galaxies that contain only old stars tend to be yellowish red.
- The first generation of stars formed when the universe was only a few hundred million years old.
- Galaxies observed when the universe was only a few billion years old tend to be smaller than today's galaxies, to have more irregular shapes, and to have more rapid star formation than the galaxies we see nearby in today's universe. This shows that the smaller galaxy fragments assembled themselves into the larger galaxies we see today.

28.5 Galaxy Mergers and Active Galactic Nuclei

- When galaxies of comparable size collide and coalesce we call it a merger, but when a small galaxy is swallowed by a much larger one, we use the term galactic cannibalism.
- Collisions play an important role in the evolution of galaxies.
- If the collision involves at least one galaxy rich in interstellar matter, the resulting compression of the gas will result

in a burst of star formation, leading to a starburst galaxy.

- Mergers were much more common when the universe was young, and many of the most distant galaxies that we see are starburst galaxies that are involved in collisions.
- Active galactic nuclei powered by supermassive black holes in the centers of most galaxies can have major effects on the host galaxy, including shutting off star formation.

28.6 The Distribution of Galaxies in Space

- Counts of galaxies in various directions establish that the universe on the large scale is homogeneous and isotropic (the same everywhere and the same in all directions, apart from evolutionary changes with time).
- The sameness of the universe everywhere is referred to as the cosmological principle.
- Galaxies are grouped together in clusters.
- The Milky Way Galaxy is a member of the Local Group, which contains at least 54 member galaxies.
- Rich clusters (such as Virgo and Coma) contain thousands or tens of thousands of galaxies.
- Galaxy clusters often group together with other clusters to form large-scale structures called superclusters, which can extend over distances of several hundred million light-years.
- Clusters and superclusters are found in filamentary structures that are huge but fill only a small fraction of space.
- Most of space consists of large voids between superclusters, with nearly all galaxies confined to less than 10% of the total volume.

28.7 The Challenge of Dark Matter

- Stars move much faster in their orbits around the centers of galaxies, and galaxies around centers of galaxy clusters, than they should according to the gravity of all the luminous matter (stars, gas, and dust) astronomers can detect.
- This discrepancy implies that galaxies and galaxy clusters are dominated by dark matter rather than normal luminous matter.
- Gravitational lensing and X-ray radiation from massive galaxy clusters confirm the presence of dark matter.
- Galaxies and clusters of galaxies contain about 10 times more dark matter than luminous matter.
- While some of the dark matter may be made up of ordinary matter (protons, neutrons, and electrons), perhaps in the form of very faint stars or black holes, most of it probably consists of some totally new type of particle not yet detected on Earth.
- Observations of gravitational lensing effects on distant objects have been used to look in the outer region of our Galaxy for any dark matter in the form of compact, dim stars or star remnants, but not enough such objects have been found to account for all the dark matter.

28.8 The Formation and Evolution of Galaxies and Structure in the Universe

- Initially, luminous and dark matter in the universe was distributed almost—but not quite—uniformly.
- The challenge for galaxy formation theories is to show how this “not quite” smooth distribution of matter developed the structures—galaxies and galaxy clusters—that we see today.
- It is likely that the filamentary distribution of galaxies and voids was built in near the beginning, before stars and galaxies began to form.
- The first condensations of matter were about the mass of a large star cluster or a small galaxy. These smaller structures then merged over cosmic time to form large galaxies, clusters of galaxies, and superclusters of galaxies.
- Superclusters today are still gathering up more galaxies, gas, and dark matter. And spiral galaxies like the Milky Way are still acquiring material by capturing small galaxies near them.

CONCEPTUAL QUESTIONS

28.1 Quasars

1. Describe some differences between quasars and normal galaxies.

2. Describe the arguments supporting the idea that quasars are at the distances indicated by their redshifts.

3. In what ways are active galaxies like quasars but different from normal galaxies?

4. Suppose you observe a star-like object in the sky. How can you determine whether it is actually a star or a quasar?

5. Why don't any of the methods for establishing distances to galaxies, described in **Galaxies** (other than Hubble's law itself), work for quasars?

6. One of the early hypotheses to explain the high redshifts of quasars was that these objects had been ejected at very high speeds from other galaxies. This idea was rejected, because no quasars with large blueshifts have been found. Explain why we would expect to see quasars with both blueshifted and redshifted lines if they were ejected from nearby galaxies.

7. Suppose we detect a powerful radio source with a radio telescope. How could we determine whether or not this was a newly discovered quasar and not some nearby radio transmission?

8. A friend tries to convince you that she can easily see a quasar in her backyard telescope. Would you believe her claim?

28.2 Supermassive Black Holes

9. Why could the concentration of matter at the center of an active galaxy like M87 not be made of stars?

10. Describe the process by which the action of a black hole can explain the energy radiated by quasars.

11. Describe the observations that convinced astronomers that M87 is an active galaxy.

12. A friend of yours who has watched many *Star Trek* episodes and movies says, "I thought that black holes pulled everything into them. Why then do astronomers think that black holes can explain the great *outpouring* of energy from quasars?" How would you respond?

13. Could the Milky Way ever become an active galaxy? Is it likely to ever be as luminous as a quasar?

28.3 Quasars as Probes of Evolution in the Universe

14. Why do astronomers believe that quasars represent an early stage in the evolution of galaxies?

15. Why were quasars and active galaxies not initially recognized as being "special" in some way?

16. What do we now understand to be the primary difference between normal galaxies and active galaxies?

17. What is the typical structure we observe in a quasar at radio frequencies?

18. What evidence do we have that the luminous central region of a quasar is small and compact?

19. Why are quasars generally so much more luminous (why do they put out so much more energy) than active galaxies?

28.4 Observations of Distant Galaxies

20. How are distant (young) galaxies different from the galaxies that we see in the universe today?

21. What is the evidence that star formation began when the universe was only a few hundred million years old?

22. Describe how you might use the color of a galaxy to determine something about what kinds of stars it contains.

28.5 Galaxy Mergers and Active Galactic Nuclei

23. Suppose a galaxy formed stars for a few million years and then stopped (and no other galaxy merged or collided with it). What would be the most massive stars on the main sequence after 500 million years? After 10 billion years? How would the color of the galaxy change over this time span? (Refer to **Evolution from the Main Sequence to Red Giants**.)

28.6 The Distribution of Galaxies in Space

24. If we see a double image of a quasar produced by a gravitational lens and can obtain a spectrum of the galaxy that is acting as the gravitational lens, we can then put limits on the distance to the quasar. Explain how.

25. The left panel of **Figure 28.1** shows a cluster of yellow galaxies that produces several images of blue galaxies through gravitational lensing. Which are more distant—the blue galaxies or the yellow galaxies? The light in the galaxies comes from stars. How do the temperatures of the stars that dominate the light of the cluster galaxies differ from the temperatures of the stars that dominate the light of the blue-lensed galaxy? Which galaxy's light is dominated by young stars?

26. Suppose you are standing in the center of a large, densely populated city that is exactly circular, surrounded by a ring of suburbs with lower-density population, surrounded in turn by a ring of farmland. From this specific location, would you say the population distribution is isotropic? Homogeneous?

27. Astronomers have been making maps by observing a slice of the universe and seeing where the galaxies lie within that slice. If the universe is isotropic and homogeneous, why do they need more than one slice? Suppose they now want to make each slice extend farther into the universe. What do they need to do?

28. Explain what we mean when we call the universe homogeneous and isotropic. Would you say that the distribution of elephants on Earth is homogeneous and isotropic? Why?

29. Describe the organization of galaxies into groupings, from the Local Group to superclusters.

28.7 The Challenge of Dark Matter

30. What is the evidence that a large fraction of the matter in the universe is invisible?

28.8 The Formation and Evolution of Galaxies and Structure in the Universe

31. Describe the evolution of an elliptical galaxy. How does the evolution of a spiral galaxy differ from that of an elliptical?

32. When astronomers make maps of the structure of the universe on the largest scales, how do they find the superclusters of galaxies to be arranged?

33. How does the presence of an active galactic nucleus in a starburst galaxy affect the starburst process?

34. Given the ideas presented here about how galaxies form, would you expect to find a giant elliptical galaxy in the Local Group? Why or why not? Is there in fact a giant elliptical in the Local Group?

35. Can an elliptical galaxy evolve into a spiral? Explain your answer. Can a spiral turn into an elliptical? How?

36. Human civilization is about 10,000 years old as measured by the development of agriculture. If your telescope collects starlight tonight that has been traveling for 10,000 years, is that star inside or outside our Milky Way Galaxy? Is it likely that the star has changed much during that time?

37. Given that only about 5% of the galaxies visible in the Hubble Deep Field are bright enough for astronomers to study spectroscopically, they need to make the most of the other 95%. One technique is to use their colors and apparent brightnesses to try to roughly estimate their redshift. How do you think the inaccuracy of this redshift estimation technique (compared to actually measuring the redshift from a spectrum) might affect our ability to make maps of large-scale structures such as the filaments and voids shown in **Figure 28.39**?

PROBLEMS

28.1 Quasars

38. Show that no matter how big a redshift (z) we measure, v/c will never be greater than 1. (In other words, no galaxy we observe can be moving away faster than the speed of light.)

39. If a quasar has a redshift of 3.3, at what fraction of the speed of light is it moving away from us?

40. If a quasar is moving away from us at $v/c = 0.8$, what is the measured redshift?

41. In the chapter, we discussed that the largest redshifts found so far are greater than 6. Suppose we find a quasar with a redshift of 6.1. With what fraction of the speed of light is it moving away from us?

42. Rapid variability in quasars indicates that the region in which the energy is generated must be small. You can show why this is true. Suppose, for example, that the region in which the energy is generated is a transparent sphere 1 light-year in diameter. Suppose that in 1 s this region brightens by a factor of 10 and remains bright for two years, after which it returns to its original luminosity. Draw its light curve (a graph of its brightness over time) as viewed from Earth.

43. Large redshifts move the positions of spectral lines to longer wavelengths and change what can be observed from the ground. For example, suppose a quasar has a redshift of $\frac{\Delta\lambda}{\lambda} = 4.1$. At what wavelength would you make observations in order to detect its Lyman line of hydrogen, which has a laboratory or rest wavelength of 121.6 nm? Would this line be observable with a ground-based telescope in a quasar with zero redshift? Would it be observable from the ground in a quasar with a redshift of $\frac{\Delta\lambda}{\lambda} = 4.1$?

44. The quasar that appears the brightest in our sky, 3C 273, is located at a distance of 2.4 billion light-years. The Sun would have to be viewed from a distance of 1300 light-years to have the same apparent magnitude as 3C 273. Using the inverse square law for light, estimate the luminosity of 3C 273 in solar units.

45. In the **Check Your Learning** section, you were told that several lines of hydrogen absorption in the visible spectrum have rest wavelengths of 410 nm, 434 nm, 486 nm, and 656 nm. In a spectrum of a distant galaxy, these same lines are observed to have wavelengths of 492 nm, 521 nm, 583 nm, and 787 nm, respectively. The example demonstrated that $z = 0.20$ for the 410 nm line. Show that you will obtain the same redshift regardless of which absorption line you measure.

46. In the **Check Your Learning** section, the author commented that even at $z = 0.2$, there is already an 11% deviation between the relativistic and the classical solution. What is the percentage difference between the classical and relativistic results at $z = 0.1$? What is it for $z = 0.5$? What is it for $z = 1$?

28.2 Supermassive Black Holes

47. Once again in this chapter, we see the use of Kepler's third law to estimate the mass of supermassive black holes. In the case of NGC 4261, this chapter supplied the result of the calculation of the mass of the black hole in NGC 4261. In order to get this answer, astronomers had to measure the velocity of particles in the ring of dust and gas that surrounds the black hole. How high were these velocities? Turn Kepler's third law around and use the information given in this chapter about the galaxy NGC 4261—the mass of the black hole at its center and the diameter of the surrounding ring of dust and gas—to calculate how long it would take a dust particle in the ring to complete a single orbit around the black hole. Assume that the only force acting on the dust particle is the gravitational force exerted by the black hole. Calculate the velocity of the dust particle in km/s.

28.6 The Distribution of Galaxies in Space

48. Using the information from **Example 28.2**, how much fainter an object will you have to be able to measure in order to include the same kinds of galaxies in your second survey? Remember that the brightness of an object varies as the inverse square of the distance.

49. Using the information from **Example 28.2**, if galaxies are distributed homogeneously, how many times more of them would you expect to count on your second survey?

50. Using the information from **Example 28.2**, how much longer will it take you to do your second survey?

51. Galaxies are found in the "walls" of huge voids; very few galaxies are found in the voids themselves. The text says that the structure of filaments and voids has been present in the universe since shortly after the expansion began 13.8 billion years ago. In science, we always have to check to see whether some conclusion is contradicted by any other information we have. In this case, we can ask whether the voids would have filled up with galaxies in roughly 14 billion years. Observations show that in addition to the motion associated with the expansion of the universe, the galaxies in the walls of the voids are moving in random directions at typical speeds of 300 km/s. At least some of them will be moving into the voids. How far into the void will a galaxy move in 14 billion years? Is it a reasonable hypothesis that the voids have existed for 14 billion years?

28.7 The Challenge of Dark Matter

52. Assume that dark matter is uniformly distributed throughout the Milky Way, not just in the outer halo but also throughout the bulge and in the disk, where the solar system lives. How much dark matter would you expect there to be inside the solar system? Would you expect that to be easily detectable? Hint: For the radius of the Milky Way's dark matter halo, use R = 300,000 light-years; for the solar system's radius, use 100 AU; and start by calculating the ratio of the two volumes.

28.8 The Formation and Evolution of Galaxies and Structure in the Universe

53. Calculate the velocity, the distance, and the look-back time of the most distant galaxies in **Figure 28.39** using the Hubble constant given in this text and the redshift given in the diagram. Remember the Doppler formula for velocity $\left(v = c \times \frac{\Delta\lambda}{\lambda}\right)$ and the Hubble law ($v = H \times d$, where d is the distance to a galaxy). For these low velocities, you can neglect relativistic effects.

54. The simulated box of galaxy filaments and superclusters shown in **Figure 28.47** stretches across 1 billion light-years. If you were to make a scale model where that box covered the core of a university campus, say 1 km, then how big would the Milky Way Galaxy be? How far away would the Andromeda galaxy be in the scale model?

55. The first objects to collapse gravitationally after the Big Bang might have been globular cluster-size galaxy pieces, with masses around 10^6 solar masses. Suppose you merge two of those together, then merge two larger pieces together, and so on, Lego-style, until you reach a Milky Way mass, about 10^{12} solar masses. How many merger generations would that take, and how many original pieces? (Hint: Think in powers of 2.)

29 | BIG BANG COSMOLOGY

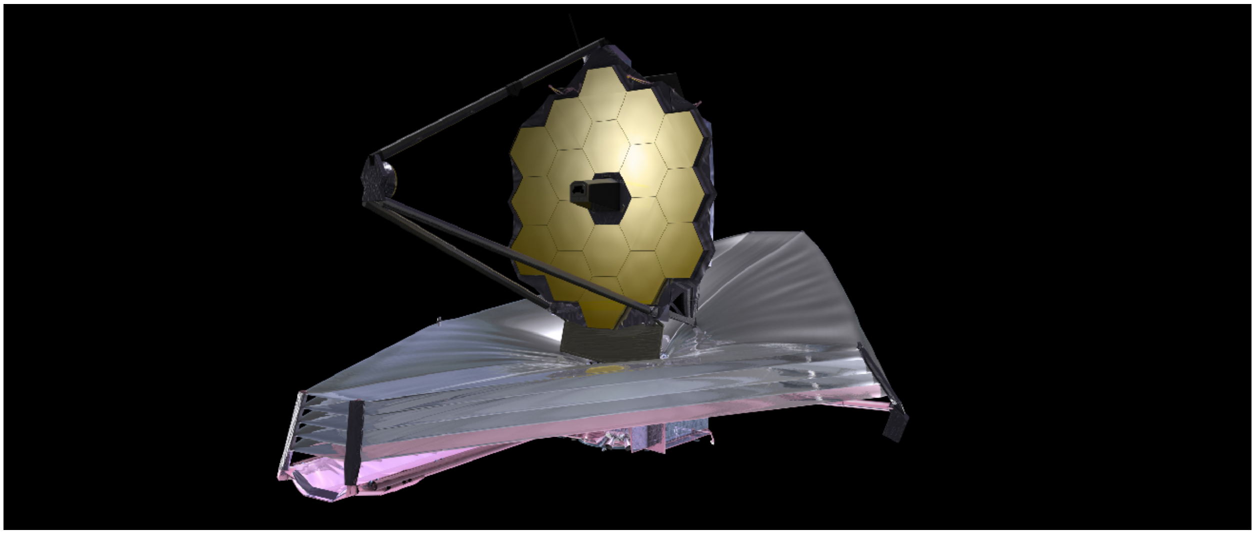


Figure 29.1 This drawing shows the James Webb Space Telescope, which is currently planned for launch in 2018. The silver sunshade shadows the primary mirror and science instruments. The primary mirror is 6.5 meters (21 feet) in diameter. Before and during launch, the mirror will be folded up. After the telescope is placed in its orbit, ground controllers will command it to unfold the mirror petals. To see distant galaxies whose light has been shifted to long wavelengths, the telescope will carry several instruments for taking infrared images and spectra. (credit: modification of work by NASA)

Chapter Outline

Introduction

In previous chapters, we explored the contents of the universe—planets, stars, and galaxies—and learned about how these objects change with time. But what about the universe as a whole? How old is it? What did it look like in the beginning? How has it changed since then? What will be its fate?

Cosmology is the study of the universe as a whole and is the subject of this chapter. The story of observational cosmology really begins in 1929 when Edwin Hubble published observations of redshifts and distances for a small sample of galaxies and showed the then-revolutionary result that we live in an expanding universe—one which in the past was denser, hotter, and smoother. From this early discovery, astronomers developed many predictions about the origin and evolution of the universe and then tested those predictions with observations. In this chapter, we will describe what we already know about the history of our dynamic universe and highlight some of the mysteries that remain.

29.1 | The Age of the Universe

Learning Objectives

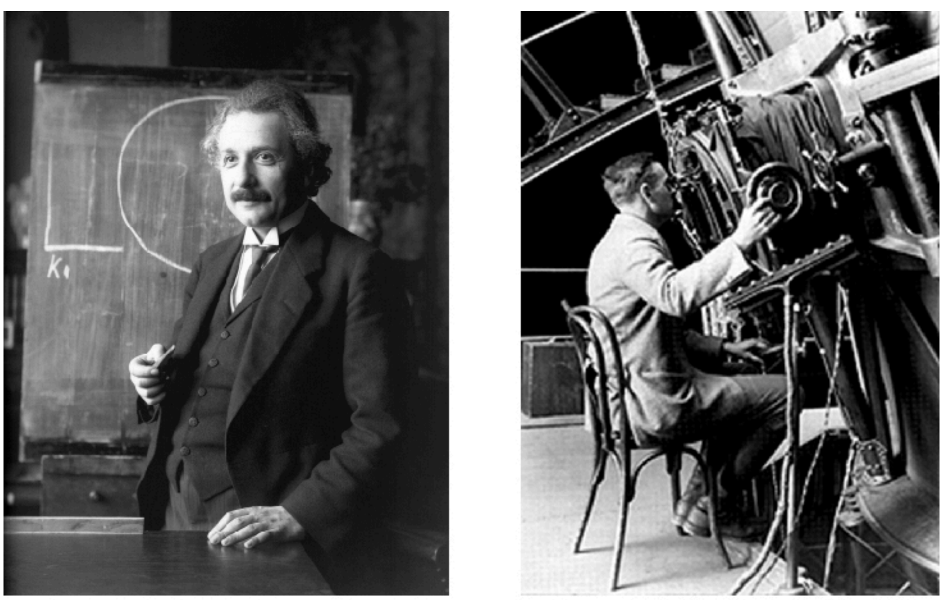
By the end of this section, you will be able to:

- Describe how we estimate the age of the universe
- Explain how changes in the rate of expansion over time affect estimates of the age of the universe
- Describe the evidence that dark energy exists and that the rate of expansion is currently accelerating
- Describe some independent evidence for the age of the universe that is consistent with the age estimate based on the rate of expansion

To explore the history of the universe, we will follow the same path that astronomers followed historically—beginning with studies of the nearby universe and then probing ever-more-distant objects and looking further back in time.

The realization that the universe changes with time came in the 1920s and 1930s when measurements of the redshifts of a large sample of galaxies became available. With hindsight, it is surprising that scientists were so shocked to discover that the universe is expanding. In fact, our theories of gravity demand that the universe must be either expanding or contracting. To show what we mean, let's begin with a universe of finite size—say a giant ball of a thousand galaxies. All these galaxies attract each other because of their gravity. If they were initially stationary, they would inevitably begin to move closer together and eventually collide. They could avoid this collapse only if for some reason they happened to be moving away from each other at high speeds. In just the same way, only if a rocket is launched at high enough speed can it avoid falling back to Earth.

The problem of what happens in an infinite universe is harder to solve, but Einstein (and others) used his theory of general relativity (which we described in **Introducing General Relativity**) to show that even infinite universes cannot be static. Since astronomers at that time did not yet know the universe was expanding (and Einstein himself was philosophically unwilling to accept a universe in motion), he changed his equations by introducing an arbitrary new term (we might call it a fudge factor) called the **cosmological constant**. This constant represented a hypothetical force of repulsion that could balance gravitational attraction on the largest scales and permit galaxies to remain at fixed distances from one another. That way, the universe could remain still.



(a) (b)

Figure 29.2 (a) Albert Einstein is shown in a 1921 photograph. (b) Edwin Hubble at work in the Mt. Wilson Observatory.

About a decade later, Hubble, and his coworkers reported that the universe is expanding, so that no mysterious balancing force is needed. (We discussed this in the chapter on **Galaxies**.) Einstein is reported to have said that the introduction of the cosmological constant was "the biggest blunder of my life." As we shall see later in this chapter, however, relatively recent observations indicate that the expansion is *accelerating*. Observations are now being carried out to determine whether this acceleration is consistent with a cosmological constant. In a way, it may turn out that Einstein was right after all.

View this **web exhibit (https://openstax.org/l/30exhcosmAIPCHP)** on the history of our thinking about cosmology, with images and biographies, from the American Institute of Physics Center for the History of Physics.

The Hubble Time

If we had a movie of the expanding universe and ran the film *backward*, what would we see? The galaxies, instead of moving apart, would move *together* in our movie—getting closer and closer all the time. Eventually, we would find that all the matter we can see today was once concentrated in an infinitesimally small volume. Astronomers identify this time with the *beginning of the universe*. The explosion of that concentrated universe at the beginning of time is called the **Big Bang** (not a bad term, since you can't have a bigger bang than one that creates the entire universe). But when did this bang occur?

We can make a reasonable estimate of the time since the universal expansion began. To see how astronomers do this, let's begin with an analogy. Suppose your astronomy class decides to have a party (a kind of "Big Bang") at someone's home to celebrate the end of the semester. Unfortunately, everyone is celebrating with so much enthusiasm that the neighbors call the police, who arrive and send everyone away at the same moment. You get home at 2 a.m., still somewhat upset about the way the party ended, and realize you forgot to look at your watch to see what time the police got there. But you use a map to measure that the distance between the party and your house is 40 kilometers. And you also remember that you drove the whole trip at a steady speed of 80 kilometers/hour (since you were worried about the police cars following you). Therefore, the trip must have taken:

$$\text{time} = \frac{\text{distance}}{\text{velocity}} = \frac{40\,\text{kilometers}}{80\,\text{kilometers/hour}} = 0.5\,\text{hours}$$

So the party must have broken up at 1:30 a.m.

No humans were around to look at their watches when the universe began, but we can use the same technique to estimate when the galaxies began moving away from each other. (Remember that, in reality, it is space that is expanding, not the galaxies that are moving through static space.) If we can measure how far apart the galaxies are now, and how fast they are moving, we can figure out how long a trip it's been.

Let's call the age of the universe measured in this way T_0. Let's first do a simple case by assuming that the expansion has been at a constant rate ever since the expansion of the universe began. In this case, the time it has taken a galaxy to move a distance, d, away from the Milky Way (remember that at the beginning the galaxies were all together in a very tiny volume) is (as in our example)

$$T_0 = d/v$$

where v is the velocity of the galaxy. If we can measure the speed with which galaxies are moving away, and also the distances between them, we can establish how long ago the expansion began.

Making such measurements should sound very familiar. This is just what Hubble and many astronomers after him needed to do in order to establish the Hubble law and the Hubble constant. We learned in **Galaxies** that a galaxy's distance and its velocity in the expanding universe are related by

$$V = H_0 \times d$$

where H_0 is the Hubble constant. Combining these two expressions gives us

$$T_0 = \frac{d}{v} = \frac{d}{(H_0 \times d)} = \frac{1}{H_0}$$

We see, then, that the work of calculating this time was already done for us when astronomers measured the Hubble constant. The age of the universe estimated in this way turns out to be just the *reciprocal of the Hubble constant* (that is, $1/H_0$). This age estimate is sometimes called the Hubble time. For a Hubble constant of 20 kilometers/second per million light-years, the Hubble time is about 15 billion years. The unit used by astronomers for the Hubble constant is kilometers/second per million parsecs. In these units, the Hubble constant is equal to about 70 kilometers/second per million parsecs, again with an uncertainty of about 5%.

To make numbers easier to remember, we have done some rounding here. Estimates for the Hubble constant are actually closer to 21 or 22 kilometers/second per million light-years, which would make the age closer to 14 billion years. But there is still about a 5% uncertainty in the Hubble constant, which means the age of the universe estimated in this way is also uncertain by about 5%.

To put these uncertainties in perspective, however, you should know that 50 years ago, the uncertainty was a factor of 2. Remarkable progress toward pinning down the Hubble constant has been made in the last couple of decades.

The Role of Deceleration

The Hubble time is the right age for the universe only if the expansion rate has been constant throughout the time since the expansion of the universe began. Continuing with our end-of-the-semester-party analogy, this is equivalent to assuming that you traveled home from the party at a constant rate, when in fact this may not have been the case. At first, mad about having to leave, you may have driven fast, but then as you calmed down—and thought about police cars on the highway—you may have begun to slow down until you were driving at a more socially acceptable speed (such as 80 kilometers/hour). In this case, given that you were driving faster at the beginning, the trip home would have taken less than a half-hour.

In the same way, in calculating the Hubble time, we have assumed that *H* has been constant throughout all of time. It turns out that this is not a good assumption. Earlier in their thinking about this, astronomers expected that the rate of expansion should be slowing down. We know that matter creates gravity, whereby all objects pull on all other objects. The mutual attraction between galaxies was expected to slow the expansion as time passed. This means that, if gravity were the only force acting (a big *if*, as we shall see in the next section), then the rate of expansion must have been faster in the past than it is today. In this case, we would say the universe has been *decelerating* since the beginning.

How much it has decelerated depends on the importance of gravity in slowing the expansion. If the universe were nearly empty, the role of gravity would be minor. Then the deceleration would be close to zero, and the universe would have been expanding at a constant rate. But in a universe with any significant density of matter, the pull of gravity means that the rate of expansion should be slower now than it used to be. If we use the current rate of expansion to estimate how long it took the galaxies to reach their current separations, we will overestimate the age of the universe—just as we may have overestimated the time it took for you to get home from the party.

A Universal Acceleration

Astronomers spent several decades looking for evidence that the expansion was decelerating, but they were not successful. What they needed were 1) larger telescopes so that they could measure the redshifts of more distant galaxies and 2) a very luminous *standard bulb* (or standard candle), that is, some astronomical object with known luminosity that produces an enormous amount of energy and can be observed at distances of a billion light-years or more.

Recall that we discussed standard bulbs in the chapter on **Galaxies**. If we compare how luminous a standard bulb is supposed to be and how dim it actually looks in our telescopes, the difference allows us to calculate its distance. The redshift of the galaxy such a bulb is in can tell us how fast it is moving in the universe. So we can measure its distance and motion independently.

These two requirements were finally met in the 1990s. Astronomers showed that supernovae of type Ia (see **The Deaths of Stars**), with some corrections based on the shapes of their light curves, are standard bulbs. This type of supernova occurs when a white dwarf accretes enough material from a companion star to exceed the Chandrasekhar limit and then collapses and explodes. At the time of maximum brightness, these dramatic supernovae can briefly outshine the galaxies that host them, and hence, they can be observed at very large distances. Large 8- to 10-meter telescopes can be used to obtain the spectra needed to measure the redshifts of the host galaxies (**Figure 29.3**).

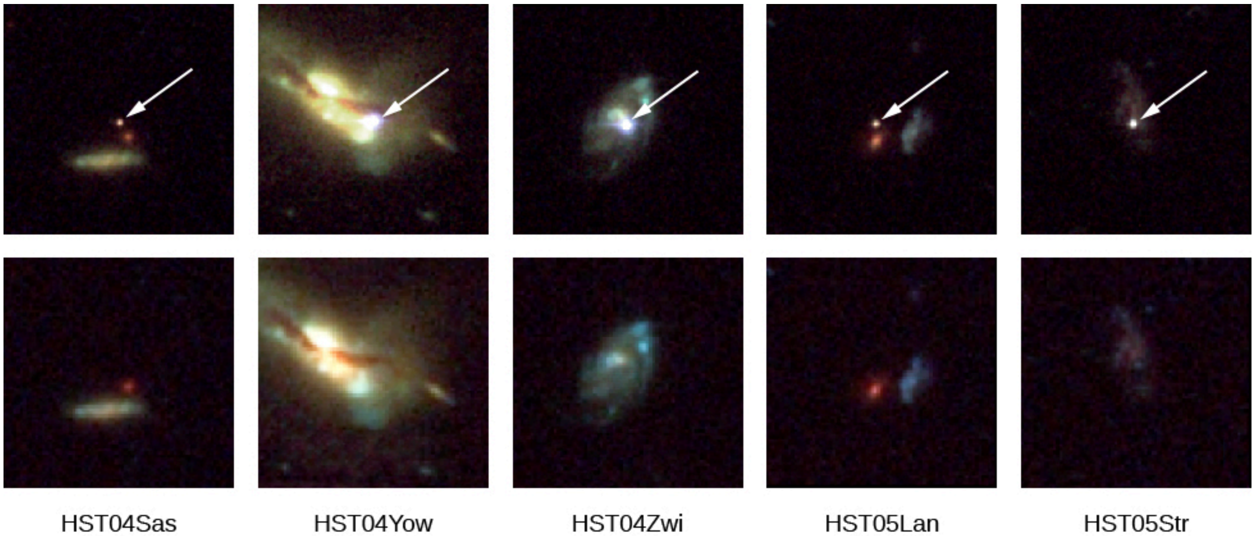


Figure 29.3 The top row shows each galaxy and its supernova (arrow). The bottom row shows the same galaxies either before or after the supernovae exploded. (credit: modification of work by NASA, ESA, and A. Riess (STScI))

The result of painstaking, careful study of these supernovae in a range of galaxies, carried out by two groups of researchers, was published in 1998. It was shocking—and so revolutionary that their discovery received the 2011 Nobel Prize in Physics. What the researchers found was that these type Ia supernovae in distant galaxies were fainter than expected from Hubble's law, given the measured redshifts of their host galaxies. In other words, distances estimated from the supernovae used as standard bulbs disagreed with the distances measured from the redshifts.

If the universe were decelerating, we would expect the far-away supernovae to be *brighter* than expected. The slowing down would have kept them closer to us. Instead, they were *fainter*, which at first seemed to make no sense.

Before accepting this shocking development, astronomers first explored the possibility that the supernovae might not really be as useful as standard bulbs as they thought. Perhaps the supernovae appeared too faint because dust along our line of sight to them absorbed some of their light. Or perhaps the supernovae at large distances were for some reason intrinsically less luminous than nearby supernovae of type Ia.

A host of more detailed observations ruled out these possibilities. Scientists then had to consider the alternative that the distance estimated from the redshift was incorrect. Distances derived from redshifts assume that the Hubble constant has been truly constant for all time. We saw that one way it might not be constant is that the expansion is slowing down. But suppose neither assumption is right (steady speed or slowing down.)

Suppose, instead, that the universe is *accelerating*. If the universe is expanding faster now than it was billions of years ago, our motion away from the distant supernovae has sped up since the explosion occurred, sweeping us farther away from them. The light of the explosion has to travel a greater distance to reach us than if the expansion rate were constant. The farther the light travels, the fainter it appears. This conclusion would explain the supernova observations in a natural way, and this has now been substantiated by many additional observations over the last couple of decades. It really seems that *the expansion of the universe is accelerating*, a notion so unexpected that astronomers at first resisted considering it.

How can the expansion of the universe be speeding up? If you want to accelerate your car, you must supply energy by stepping on the gas. Similarly, energy must be supplied to accelerate the expansion of the universe. The discovery of the acceleration was shocking because scientists still have no idea what the source of the energy is. Scientists call whatever it is **dark energy**, which is a clear sign of how little we understand it.

Note that this new component of the universe is not the dark matter we talked about in earlier chapters. Dark energy is something else that we have also not yet detected in our laboratories on Earth.

What is dark energy? One possibility is that it is the cosmological constant, which is an energy associated with the vacuum of "empty" space itself. Quantum mechanics (the intriguing theory of how things behave at the atomic and subatomic levels) tells us that the source of this vacuum energy might be tiny elementary particles that flicker in and out of existence everywhere throughout the universe. Various attempts have been made to calculate how big the effects of this vacuum energy should be, but so far these attempts have been unsuccessful. In fact, the order of magnitude of theoretical estimates of the vacuum energy based on the quantum mechanics of matter and the value required to account for the acceleration of the expansion of the universe differ by an incredible factor of at least 10^{120} (that is a 1 followed by 120 zeros)! Various other theories have been suggested, but the bottom line is that, although there is compelling evidence that dark energy exists, we do not yet know the source of that energy.

Whatever the dark energy turns out to be, we should note that the discovery that the rate of expansion has not been constant since the beginning of the universe complicates the calculation of the age of the universe. Interestingly, the acceleration seems not to have started with the Big Bang. During the first several billion years after the Big Bang, when galaxies were close together, gravity was strong enough to slow the expansion. As galaxies moved farther apart, the effect of gravity weakened. Several billion years after the Big Bang, dark energy took over, and the expansion began to accelerate (**Figure 29.4**).

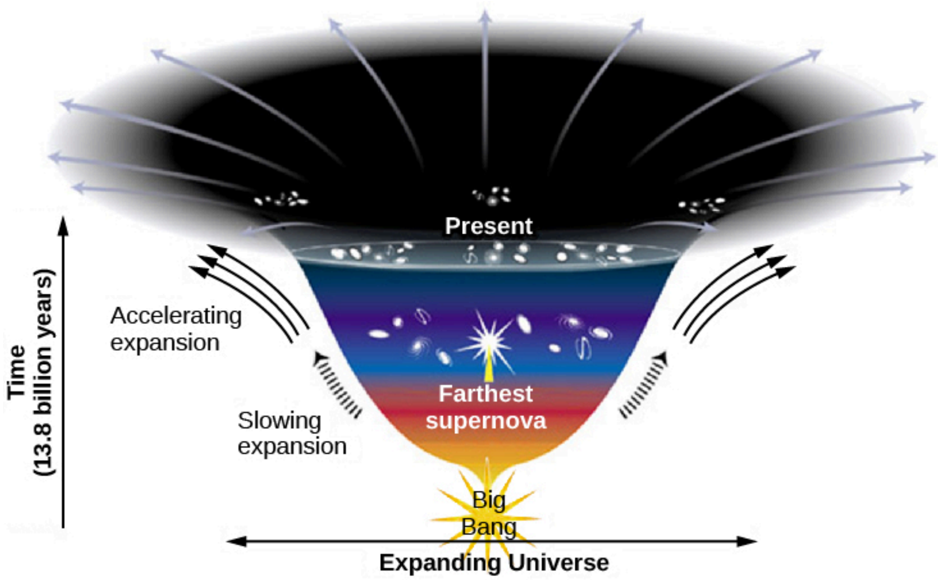


Figure 29.4 The more the diagram spreads out horizontally, the faster the change in the velocity of expansion. After a period of very rapid expansion at the beginning, which scientists call inflation and which we will discuss later in this chapter, the expansion began to decelerate. Galaxies were then close together, and their mutual gravitational attraction slowed the expansion. After a few billion years, when galaxies were farther apart, the influence of gravity began to weaken. Dark energy then took over and caused the expansion to accelerate. (credit: modification of work by Ann Feild (STScI))

Deceleration works to make the age of the universe estimated by the simple relation $T_0 = 1/H$ seem older than it really is, whereas acceleration works to make it seem younger. By happy coincidence, our best estimates of how much deceleration and acceleration occurred lead to an answer for the age very close to $T_0 = 1/H$. The best current estimate is that the universe is 13.8 billion years old with an uncertainty of only about 100 million years.

Throughout this chapter, we have referred to the Hubble *constant*. We now know that the Hubble constant does change with time. It is, however, constant everywhere in the universe at any given time. When we say the Hubble constant is about 70 kilometers/second/million parsecs, we mean that this is the value of the Hubble constant at the current time.

Comparing Ages

We now have one estimate for the age of the universe from its expansion. Is this estimate consistent with other observations? For example, are the oldest stars or other astronomical objects younger than 13.8 billion years? After all, the universe has to be at least as old as the oldest objects in it.

In our Galaxy and others, the oldest stars are found in the globular clusters (**Figure 29.5**), which can be dated using the models of stellar evolution described in the chapter **Stellar Life Cycles**.

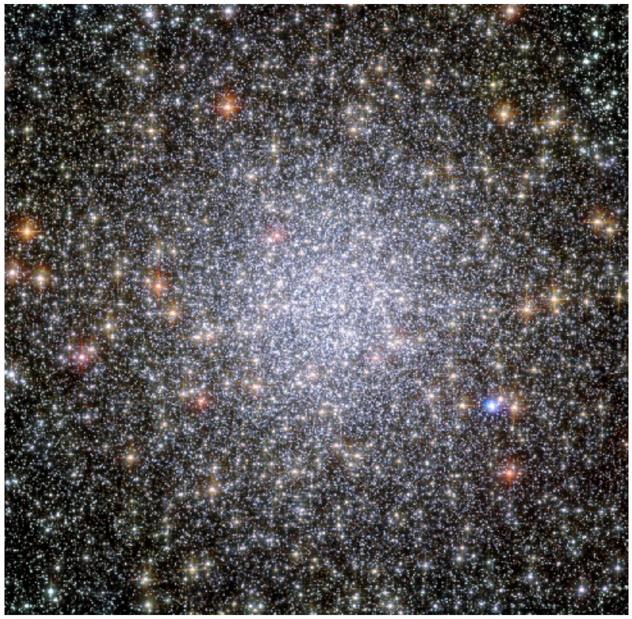


Figure 29.5 This NASA/ESA Hubble Space Telescope image shows a globular cluster known as 47 Tucanae, since it is in the constellation of Tucana (The Toucan) in the southern sky. The second-brightest globular cluster in the night sky, it includes hundreds of thousands of stars. Globular clusters are among the oldest objects in our Galaxy and can be used to estimate its age. (credit: NASA, ESA, and the Hubble Heritage (STScI/AURA)-ESA/Hubble Collaboration)

The accuracy of the age estimates of the globular clusters has improved markedly in recent years for two reasons. First, models of interiors of globular cluster stars have been improved, mainly through better information about how atoms absorb radiation as they make their way from the center of a star out into space. Second, observations from satellites have improved the accuracy of our measurements of the distances to these clusters. The conclusion is that the oldest stars formed about 12–13 billion years ago.

This age estimate has recently been confirmed by the study of the spectrum of uranium in the stars. The isotope uranium-238 is radioactive and decays (changes into another element) over time. (Uranium-238 gets its designation because it has 92 protons and 146 neutrons.) We know (from how stars and supernovae make elements) how much uranium-238 is generally made compared to other elements. Suppose we measure the amount of uranium relative to nonradioactive elements in a very old star and in our own Sun, and compare the abundances. With those pieces of information, we can estimate how much longer the uranium has been decaying in the very old star because we know from our own Sun how much uranium decays in 4.5 billion years.

The line of uranium is very weak and hard to make out even in the Sun, but it has now been measured in one extremely old star using the European Very Large Telescope (Figure 29.6). Comparing the abundance with that in the solar system, whose age we know, astronomers estimate the star is 12.5 billion years old, with an uncertainty of about 3 billion years. While the uncertainty is large, this work is important confirmation of the ages estimated by studies of the globular cluster stars. Note that the uranium age estimate is completely independent; it does not depend on either the measurement of distances or on models of the interiors of stars.

Figure 29.6 The European Extremely Large Telescope (E-ELT) is currently under construction in Chile. This image compares the size of the E-ELT (left) with the four 8-meter telescopes of the European Very Large Telescope (center) and with the Colosseum in Rome (right). The mirror of the E-ELT will be 39 meters in diameter. Astronomers are building a new generation of giant telescopes in order to observe very distant galaxies and understand what they were like when they were newly formed and the universe was young. (credit: modification of work by ESO)

As we shall see later in this chapter, the globular cluster stars probably did not form until the expansion of the universe had been underway for at least a few hundred million years. Accordingly, their ages are consistent with the 13.8 billion-year age estimated from the expansion rate.

29.2 | A Model of the Universe

Learning Objectives

By the end of this section, you will be able to:

- Explain how the rate of expansion of the universe affects its evolution
- Describe four possibilities for the evolution of the universe
- Explain what is expanding when we say that the universe is expanding
- Define critical density and the evidence that matter alone in the universe is much smaller than the critical density
- Describe what the observations say about the likely long-term future of the universe

Let's now use the results about the expansion of the universe to look at how these ideas might be applied to develop a model for the evolution of the universe as a whole. With this model, astronomers can make predictions about how the universe has evolved so far and what will happen to it in the future.

The Expanding Universe

Every model of the universe must include the expansion we observe. Another key element of the models is that the cosmological principle (which we discussed in **The Evolution and Distribution of Galaxies**) is valid: on the large scale, the universe at any given time is the same everywhere (homogeneous and isotropic). As a result, the expansion rate must be the same everywhere during any epoch of cosmic time. If so, we don't need to think about the entire universe when we think about the expansion, we can just look at any sufficiently large portion of it. (Some models for dark energy would allow the expansion rate to be different in different directions, and scientists are designing experiments to test this idea. However, until such evidence is found, we will assume that the cosmological principle applies throughout the universe.)

In **Galaxies**, we hinted that when we think of the expansion of the universe, it is more correct to think of space itself stretching rather than of galaxies moving through static space. Nevertheless, we have since been discussing the redshifts of galaxies as if they resulted from the motion of the galaxies themselves.

Now, however, it is time to finally put such simplistic notions behind us and take a more sophisticated look at the cosmic expansion. Recall from our discussion of Einstein's theory of general relativity (in the section **Introducing General Relativity**) that space—or, more precisely, spacetime—is not a mere backdrop to the action of the universe, as Newton thought. Rather, it is an active participant—affected by and in turn affecting the matter and energy in the universe.

Since the expansion of the universe is the stretching of all spacetime, all points in the universe are stretching together. Thus, the expansion began *everywhere at once*. Unfortunately for tourist agencies of the future, there is no location you can visit where the stretching of space began or where we can say that the Big Bang happened.

To describe just how space stretches, we say the cosmic expansion causes the universe to undergo a uniform change in *scale* over time. By scale we mean, for example, the distance between two clusters of galaxies. It is customary to represent the scale by the factor R; if R doubles, then the distance between the clusters has doubled. Since the universe is expanding at the same rate everywhere, the change in R tells us how much it has expanded (or contracted) at any given time. For a static universe, R would be constant as time passes. In an expanding universe, R increases with time.

If it is space that is stretching rather than galaxies moving through space, then why do the galaxies show redshifts in their spectra? When you were young and naïve—a few chapters ago—it was fine to discuss the redshifts of distant galaxies as if they resulted from their motion away from us. But now that you are an older and wiser student of cosmology, this view will simply not do.

A more accurate view of the redshifts of galaxies is that the light waves are stretched by the stretching of the space they travel through. Think about the light from a remote galaxy. As it moves away from its source, the light has to travel through space. If space is stretching during all the time the light is traveling, the light waves will be stretched as well. A redshift is a stretching of waves—the wavelength of each wave increases (**Figure 29.7**). Light from more distant galaxies travels for more time than light from closer ones. This means that the light has stretched more than light from closer ones and thus shows a greater redshift.

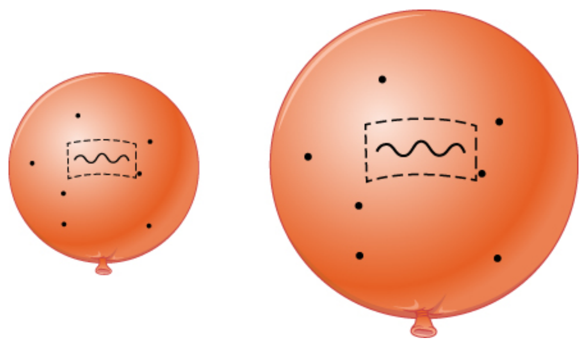


Figure 29.7 As an elastic surface expands, a wave on its surface stretches. For light waves, the increase in wavelength would be seen as a redshift.

Thus, what the measured redshift of light from an object is telling us is how much the universe has expanded since the light left the object. If the universe has expanded by a factor of 2, then the wavelength of the light (and all electromagnetic waves from the same source) will have doubled.

Models of the Expansion

Before astronomers knew about dark energy or had a good measurement of how much matter exists in the universe, they made speculative models about how the universe might evolve over time. The four possible scenarios are shown in **Figure 29.8**. In this diagram, time moves forward from the bottom upward, and the scale of space increases by the horizontal circles becoming wider.

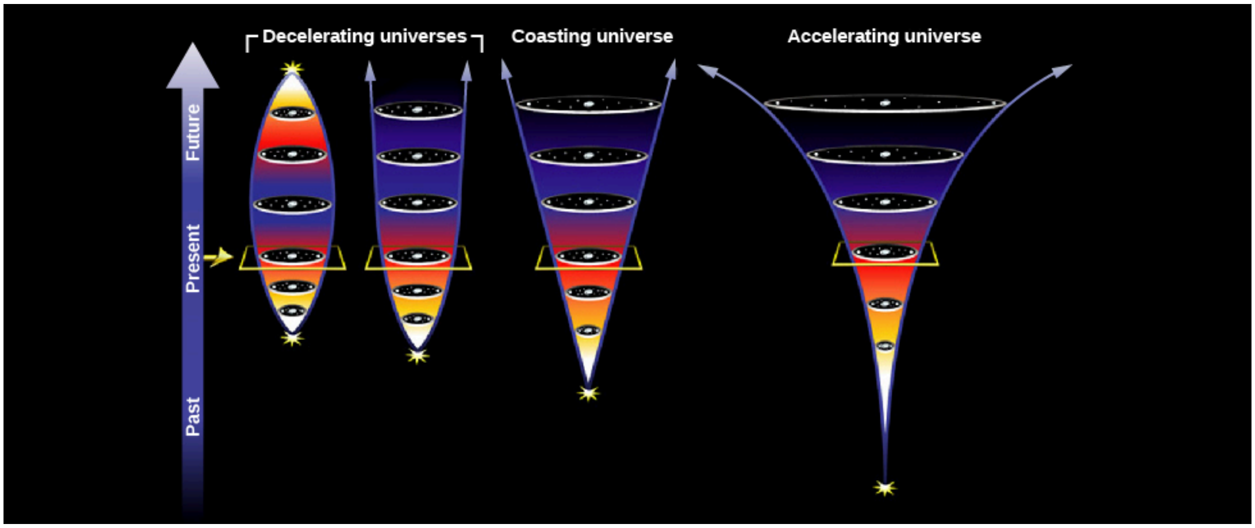


Figure 29.8 The yellow square marks the present in all four cases, and for all four, the Hubble constant is equal to the same value at the present time. Time is measured in the vertical direction. The first two universes on the left are ones in which the rate of expansion slows over time. The one on the left will eventually slow, come to a stop and reverse, ending up in a "big crunch," while the one next to it will continue to expand forever, but ever-more slowly as time passes. The "coasting" universe is one that expands at a constant rate given by the Hubble constant throughout all of cosmic time. The accelerating universe on the right will continue to expand faster and faster forever. (credit: modification of work by NASA/ESA)

The simplest scenario of an expanding universe would be one in which R increases with time at a constant rate. But you already know that life is not so simple. The universe contains a great deal of mass and its gravity decelerates the expansion—by a large amount if the universe contains a lot of matter, or by a negligible amount if the universe is nearly empty. Then there is the observed acceleration, which astronomers blame on a kind of dark energy.

Let's first explore the range of possibilities with models for different amounts of mass in the universe and for different contributions by dark energy. In some models—as we shall see—the universe expands forever. In others, it stops expanding and starts to contract. After looking at the extreme possibilities, we will look at recent observations that allow us to choose the most likely scenario.

We should perhaps pause for a minute to note how remarkable it is that we can do this at all. Our understanding of the principles that underlie how the universe works on the large scale and our observations of how the objects in the universe change with time allow us to model the evolution of the entire cosmos these days. It is one of the loftiest achievements of the human mind.

What astronomers look at in practice, to determine the kind of universe we live in, is the *average density* of the universe. This is the mass of matter (including the equivalent mass of energy)[1] that would be contained in each unit of volume (say, 1 cubic centimeter) if all the stars, galaxies, and other objects were taken apart, atom by atom, and if all those particles, along with the light and other energy, were distributed throughout all of space with absolute uniformity. If the average density is low, there is less mass and less gravity, and the universe will not decelerate very much. It can therefore expand forever. Higher average density, on the other hand, means there is more mass and more gravity and that the stretching of space might slow down enough that the expansion will eventually stop. An extremely high density might even cause the universe to collapse again.

For a given rate of expansion, there is a **critical density**—the mass per unit volume that will be just enough to slow the expansion to zero at some time infinitely far in the future. If the actual density is higher than this critical density, then the expansion will ultimately reverse and the universe will begin to contract. If the actual density is lower, then the universe will expand forever.

These various possibilities are illustrated in **Figure 29.9**. In this graph, one of the most comprehensive in all of science, we chart the development of the scale of space in the cosmos against the passage of time. Time increases to the right, and the scale of the universe, R, increases upward in the figure. Today, at the point marked "present" along the time axis, R is increasing in each model. We know that the galaxies are currently expanding away from each other, no matter which model is right. (The same situation holds for a baseball thrown high into the air. While it may eventually fall back down, near the

1. By equivalent mass we mean that which would result if the energy were turned into mass using Einstein's formula, $E = mc^2$.

beginning of the throw it moves upward most rapidly.)

The various lines moving across the graph correspond to different models of the universe. The straight dashed line corresponds to the empty universe with no deceleration; it intercepts the time axis at a time, T_0 (the Hubble time), in the past. This is not a realistic model but gives us a measure to compare other models to. The curves below the dashed line represent models with no dark energy and with varying amounts of deceleration, starting from the Big Bang at shorter times in the past. The curve above the dashed line shows what happens if the expansion is accelerating. Let's take a closer look at the future according to the different models.

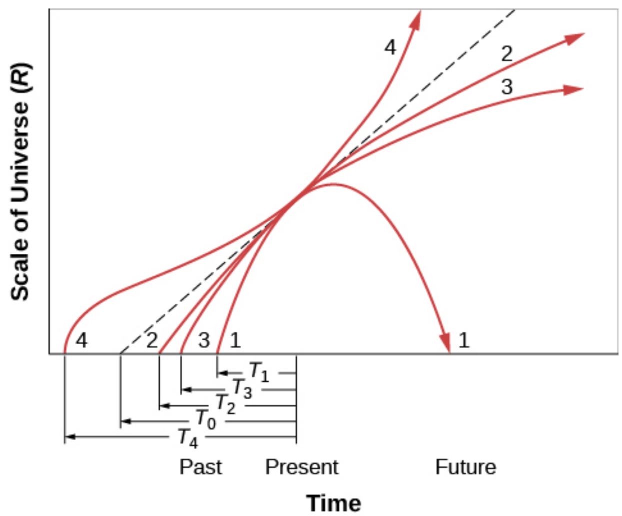


Figure 29.9 This graph plots *R*, the scale of the universe, against time for various cosmological models. Curve 1 represents a universe where the density is greater than the critical value; this model predicts that the universe will eventually collapse. Curve 2 represents a universe with a density lower than critical; the universe will continue to expand but at an ever-slower rate. Curve 3 is a critical-density universe; in this universe, the expansion will gradually slow to a stop infinitely far in the future. Curve 4 represents a universe that is accelerating because of the effects of dark energy. The dashed line is for an empty universe, one in which the expansion is not slowed by gravity or accelerated by dark energy. Time is very compressed on this graph.

Let's start with curve 1 in **Figure 29.9**. In this case, the actual density of the universe is higher than the critical density and there is no dark energy. This universe will stop expanding at some time in the future and begin contracting. This model is called a **closed universe** and corresponds to the universe on the left in **Figure 29.8**. Eventually, the scale drops to zero, which means that space will have shrunk to an infinitely small size. The noted physicist John Wheeler called this the " big crunch," because matter, energy, space, and time would all be crushed out of existence. Note that the "big crunch" is the opposite of the Big Bang—it is an *implosion*. The universe is not expanding but rather collapsing in upon itself.

Some scientists speculated that another Big Bang might follow the crunch, giving rise to a new expansion phase, and then another contraction—perhaps oscillating between successive Big Bangs and big crunches indefinitely in the past and future. Such speculation was sometimes referred to as the oscillating theory of the universe. The challenge for theorists was how to describe the transition from collapse (when space and time themselves disappear into the big crunch) to expansion. With the discovery of dark energy, however, it does not appear that the universe will experience a big crunch, so we can put worrying about it on the back burner.

If the density of the universe is less than the critical density (curve 2 in **Figure 29.9** and the universe second from the left in **Figure 29.8**), gravity is never important enough to stop the expansion, and so the universe expands forever. Such a universe is infinite and this model is called an **open universe**. Time and space begin with the Big Bang, but they have no end; the universe simply continues expanding, always a bit more slowly as time goes on. Groups of galaxies eventually get so far apart that it would be difficult for observers in any of them to see the others. (See the feature box on **What Might the Universe Be Like in the Distant Future?** for more about the distant future in the closed and open universe models.)

At the critical density (curve 3), the universe can just barely expand forever. The critical-density universe has an age of exactly two-thirds T_0, where T_0 is the age of the empty universe. Universes that will someday begin to contract have ages less than two-thirds T_0.

In an empty universe (the dashed line **Figure 29.9** and the coasting universe in **Figure 29.8**), neither gravity nor dark energy is important enough to affect the expansion rate, which is therefore constant throughout all time.

In a universe with dark energy, the rate of the expansion will increase with time, and the expansion will continue at an ever-faster rate. Curve 4 in **Figure 29.9**, which represents this universe, has a complicated shape. In the beginning, when the matter is all very close together, the rate of expansion is most influenced by gravity. Dark energy appears to act only over large scales and thus becomes more important as the universe grows larger and the matter begins to thin out. In this model, at first the universe slows down, but as space stretches, the acceleration plays a greater role and the expansion speeds up.

The Cosmic Tug of War

We might summarize our discussion so far by saying that a "tug of war" is going on in the universe between the forces that push everything apart and the gravitational attraction of matter, which pulls everything together. If we can determine who will win this tug of war, we will learn the ultimate fate of the universe.

The first thing we need to know is the density of the universe. Is it greater than, less than, or equal to the critical density? The critical density today depends on the value of the expansion rate today, H_0. If the Hubble constant is around 20 kilometers/second per million light-years, the critical density is about 10^{-26} kg/m^3. Let's see how this value compares with the actual density of the universe.

Example 29.1

Critical Density of the Universe

As we discussed, the critical density is that combination of matter and energy that brings the universe coasting to a stop at time infinity. Einstein's equations lead to the following expression for the critical density (ρ_{crit}):

$$\rho_{\text{crit}} = \frac{3H^2}{8\pi G} \quad \textbf{(29.1)}$$

where H is the Hubble constant and G is the universal constant of gravity (6.67×10^{-11} Nm^2/kg^2).

Solution

Let's substitute our values and see what we get. Take an H = 22 km/s per million light-years. We need to convert both km and light-years into meters for consistency. A million light-years = $10^6 \times 9.5 \times 10^{15}$ m = 9.5×10^{21} m. And 22 km/s = 2.2×10^4 m/s. That makes $H = 2.3 \times 10^{-18}$ /s and $H^2 = 5.36 \times 10^{-36}$ /s^2. So,

$$\rho_{\text{crit}} = \frac{3 \times 5.36 \times 10^{-36}}{8 \times 3.14 \times 6.67 \times 10^{-11}} = 9.6 \times 10^{-27}\ \text{kg/m}^3$$

which we can round off to the 10^{-26} kg/m^3. (To make the units work out, you have to know that N, the unit of force, is the same as kg × m/s^2.)

Now we can compare densities we measure in the universe to this critical value. Note that density is mass per unit volume, but energy has an equivalent mass of $m = E/c^2$ (from Einstein's equation $E = mc^2$).

29.1
a. A single grain of dust has a mass of about 1.1×10^{-13} kg. If the average mass-energy density of space is equal to the critical density on average, how much space would be required to produce a total mass-energy equal to a dust grain?

b. If the Hubble constant were twice what it actually is, how much would the critical density be?

We can start our survey of how dense the cosmos is by ignoring the dark energy and just estimating the density of all matter in the universe, including ordinary matter and dark matter. Here is where the cosmological principle really comes in handy. Since the universe is the same all over (at least on large scales), we only need to measure how much matter exists in a (large) representative sample of it. This is similar to the way a representative survey of a few thousand people can tell us whom the millions of residents of the US prefer for president.

There are several methods by which we can try to determine the average density of matter in space. One way is to count all the galaxies out to a given distance and use estimates of their masses, including dark matter, to calculate the average density. Such estimates indicate a density of about 1 to 2×10^{-27} kg/m^3 (10 to 20% of critical), which by itself is too small to stop the expansion.

A lot of the dark matter lies outside the boundaries of galaxies, so this inventory is not yet complete. But even if we add an estimate of the dark matter outside galaxies, our total won't rise beyond about 30% of the critical density. We'll pin these numbers down more precisely later in this chapter, where we will also include the effects of dark energy.

In any case, even if we ignore dark energy, the evidence is that the universe will continue to expand forever. The discovery of dark energy that is causing the rate of expansion to speed up only strengthens this conclusion. Things definitely do not look good for fans of the closed universe (big crunch) model.

What Might the Universe Be Like in the Distant Future?

" Some say the world will end in fire,
Some say in ice.
From what I've tasted of desire
I hold with those who favor fire."
—From the poem "Fire and Ice" by Robert Frost (1923)

Given the destructive power of impacting asteroids, expanding red giants, and nearby supernovae, our species may not be around in the remote future. Nevertheless, you might enjoy speculating about what it would be like to live in a much, much older universe.

The observed acceleration makes it likely that we will have continued expansion into the indefinite future. If the universe expands forever (R increases without limit), the clusters of galaxies will spread ever farther apart with time. As eons pass, the universe will get thinner, colder, and darker.

Within each galaxy, stars will continue to go through their lives, eventually becoming white dwarfs, neutron stars, and black holes. Low-mass stars might take a long time to finish their evolution, but in this model, we would literally have all the time in the world. Ultimately, even the white dwarfs will cool down to be black dwarfs, any neutron stars that reveal themselves as pulsars will slowly stop spinning, and black holes with accretion disks will one day complete their "meals." The remains of stars will all be dark and difficult to observe.

This means that the light that now reveals galaxies to us will eventually go out. Even if a small pocket of raw material were left in one unsung corner of a galaxy, ready to be turned into a fresh cluster of stars, we will only have to wait until the time that their evolution is also complete. And time is one thing this model of the universe has plenty of. There will surely come a time when all the stars are out, galaxies are as dark as space, and no source of heat remains to help living things survive. Then the lifeless galaxies will just continue to move apart in their lightless realm.

If this view of the future seems discouraging (from a human perspective), keep in mind that we fundamentally do not understand why the expansion rate is currently accelerating. Thus, our speculations about the future are just that: speculations. You might take heart in the knowledge that science is always a progress report. The most advanced ideas about the universe from a hundred years ago now strike us as rather primitive. It may well be that our best models of today will in a hundred or a thousand years also seem rather simplistic and that there are other factors determining the ultimate fate of the universe of which we are still completely unaware.

Ages of Distant Galaxies

In the chapter on **Galaxies**, we discussed how we can use Hubble's law to measure the distance to a galaxy. But that simple method only works with galaxies that are not too far away. Once we get to large distances, we are looking so far into the past that we must take into account changes in the rate of the expansion of the universe. Since we cannot measure these changes directly, we must assume one of the models of the universe to be able to convert large redshifts into distances.

This is why astronomers squirm when reporters and students ask them exactly how far away some newly discovered distant quasar or galaxy is. We really can't give an answer without first explaining the model of the universe we are assuming in calculating it (by which time a reporter or student is long gone or asleep). Specifically, we must use a model that includes the change in the expansion rate with time. The key ingredients of the model are the amounts of matter, including dark matter, and the equivalent mass (according to $E = mc^2$) of the dark energy along with the Hubble constant.

Elsewhere in this book, we have estimated the mass density of ordinary matter plus dark matter as roughly 0.3 times the critical density, and the mass equivalent of dark energy as roughly 0.7 times the critical density. We will refer to these values as the “standard model of the universe.” The latest (slightly improved) estimates for these values and the evidence for them will be given later in this chapter. Calculations also require the current value of the Hubble constant. For **Table 29.1**, we have adopted a Hubble constant of 67.3 kilometers/second/million parsecs (rather than rounding it to 70 kilometers/second/million parsecs), which is consistent with the 13.8 billion-year age of the universe estimated by the latest observations.

Once we assume a model, we can use it to calculate the age of the universe at the time an object emitted the light we see. As an example, **Table 29.1** lists the times that light was emitted by objects at different redshifts as fractions of the current age of the universe. The times are given for two very different models so you can get a feeling for the fact that the calculated ages are fairly similar. The first model assumes that the universe has a critical density of matter and no dark energy. The second model is the standard model described in the preceding paragraph. The first column in the table is the redshift, which is given by the equation $z = \Delta\lambda/\lambda_0$ and is a measure of how much the wavelength of light has been stretched by the expansion of the universe on its long journey to us.

Ages of the Universe at Different Redshifts

Redshift	Percent of Current Age of Universe When the Light Was Emitted (mass = critical density)	Percent of Current Age of Universe When the Light Was Emitted (mass = 0.3 critical density; dark energy = 0.7 critical density)
0	100 (now)	100 (now)
0.5	54	63
1.0	35	43
2.0	19	24
3.0	13	16
4.0	9	11
5.0	7	9
8.0	4	5
11.9	2.1	2.7
Infinite	0	0

Table 29.1

Notice that as we find objects with higher and higher redshifts, we are looking back to smaller and smaller fractions of the age of the universe. The highest observed redshifts as this book is being written are close to 12 (**Figure 29.10**). As **Table 29.1** shows, we are seeing these galaxies as they were when the universe was only about 3% as old as it is now. They were already formed only about 700 million years after the Big Bang.

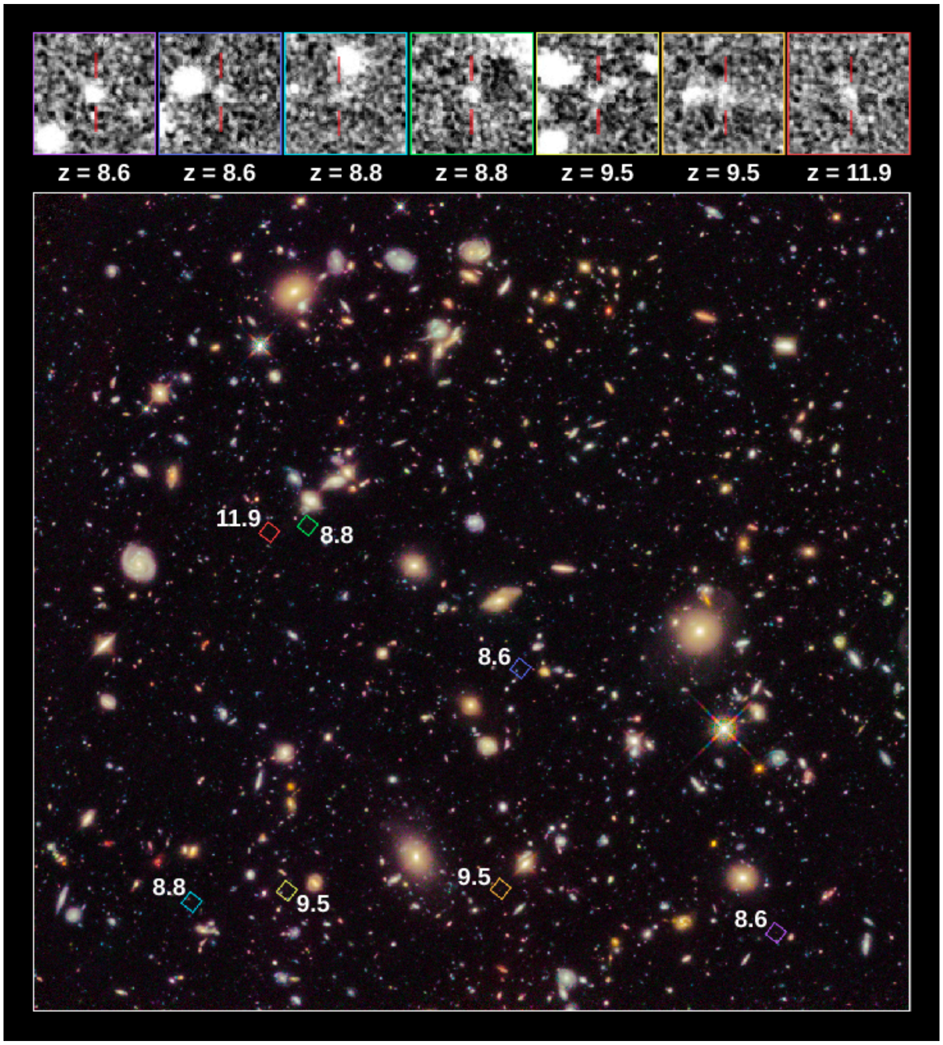


Figure 29.10 This image, called the Hubble Ultra Deep Field, shows faint galaxies, seen very far away and therefore very far back in time. The colored squares in the main image outline the locations of the galaxies. Enlarged views of each galaxy are shown in the black-and-white images. The red lines mark each galaxy's location. The "redshift" of each galaxy is indicated below each box, denoted by the symbol "z." The redshift measures how much a galaxy's ultraviolet and visible light has been stretched to infrared wavelengths by the universe's expansion. The larger the redshift, the more distant the galaxy, and therefore the further astronomers are seeing back in time. One of the seven galaxies may be a distance breaker, observed at a redshift of 11.9. If this redshift is confirmed by additional measurements, the galaxy is seen as it appeared only 380 million years after the Big Bang, when the universe was less than 3% of its present age. (credit: modification of work by NASA, ESA, R. Ellis (Caltech), and the UDF 2012 Team)

29.3 | The Beginning of the Universe

Learning Objectives

By the end of this section, you will be able to:

- Describe what the universe was like during the first few minutes after it began to expand
- Explain how the first new elements were formed during the first few minutes after the Big Bang
- Describe how the contents of the universe change as the temperature of the universe decreases

The best evidence we have today indicates that the first galaxies did not begin to form until a few hundred million years after the Big Bang. What were things like before there were galaxies and space had not yet stretched very significantly? Amazingly, scientists have been able to calculate in some detail what was happening in the universe in the first few minutes after the Big Bang.

The History of the Idea

It is one thing to say the universe had a beginning (as the equations of general relativity imply) and quite another to describe that beginning. The Belgian priest and cosmologist Georges Lemaître was probably the first to propose a specific model for the Big Bang itself (**Figure 29.11**). He envisioned all the matter of the universe starting in one great bulk he called the *primeval atom*, which then broke into tremendous numbers of pieces. Each of these pieces continued to fragment further until they became the present atoms of the universe, created in a vast nuclear fission. In a popular account of his theory, Lemaître wrote, “The evolution of the world could be compared to a display of fireworks just ended—some few red wisps, ashes, and smoke. Standing on a well-cooled cinder, we see the slow fading of the suns and we try to recall the vanished brilliance of the origin of the worlds.”

Figure 29.11 This Belgian cosmologist studied theology at Mechelen and mathematics and physics at the University of Leuven. It was there that he began to explore the expansion of the universe and postulated its explosive beginning. He actually predicted Hubble’s law 2 years before its verification, and he was the first to consider seriously the physical processes by which the universe began.

View a **short video (https://openstax.org/l/30Lemaitrevid)** about the work of Lemaître, considered by some to be the father of the Big Bang theory.

Physicists today know much more about nuclear physics than was known in the 1920s, and they have shown that the primeval fission model cannot be correct. Yet Lemaître’s vision was in some respects quite prophetic. We still believe that everything was together at the beginning; it was just not in the form of matter we now know. Basic physical principles tell us that when the universe was much denser, it was also much hotter, and that it cools as it expands, much as gas cools when sprayed from an aerosol can.

By the 1940s, scientists knew that fusion of hydrogen into helium was the source of the Sun’s energy. Fusion requires high temperatures, and the early universe must have been hot. Based on these ideas, American physicist George Gamow (**Figure 29.12**) suggested a universe with a different kind of beginning that involved nuclear **fusion** instead of fission. Ralph Alpher worked out the details for his PhD thesis, and the results were published in 1948. (Gamow, who had a quirky sense of humor, decided at the last minute to add the name of physicist Hans Bethe to their paper, so that the coauthors on this paper about the beginning of things would be Alpher, Bethe, and Gamow, a pun on the first three letters of the Greek alphabet: alpha, beta, and gamma.) Gamow’s universe started with fundamental particles that built up the heavy elements by fusion in the Big Bang.

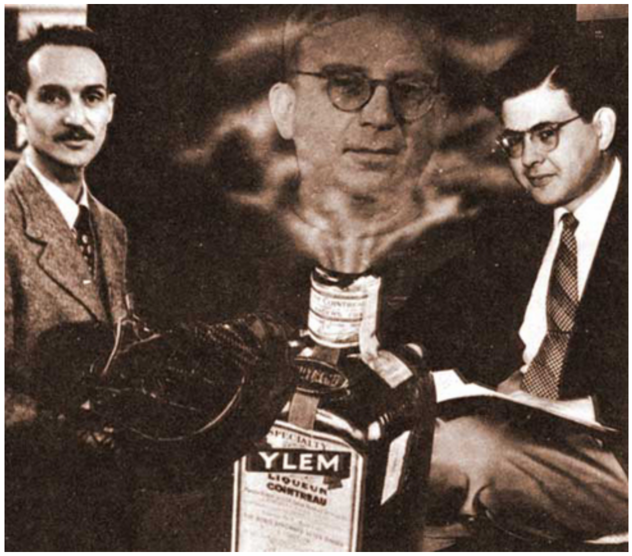


Figure 29.12 This composite image shows George Gamow emerging like a genie from a bottle of ylem, a Greek term for the original substance from which the world formed. Gamow revived the term to describe the material of the hot Big Bang. Flanking him are Robert Herman (left) and Ralph Alpher (right), with whom he collaborated in working out the physics of the Big Bang. (The modern composer Karlheinz Stockhausen was inspired by Gamow's ideas to write a piece of music called *Ylem*, in which the players actually move away from the stage as they perform, simulating the expansion of the universe.)

Gamow's ideas were close to our modern view, except we now know that the early universe remained hot enough for fusion for only a short while. Thus, only the three lightest elements—hydrogen, helium, and a small amount of lithium—were formed in appreciable abundances at the beginning. The heavier elements formed later in stars. Since the 1940s, many astronomers and physicists have worked on a detailed theory of what happened in the early stages of the universe.

The First Few Minutes

Let's start with the first few minutes following the Big Bang. Three basic ideas hold the key to tracing the changes that occurred during the time just after the universe began. The first, as we have already mentioned, is that the universe cools as it expands. **Figure 29.13** shows how the temperature changes with the passage of time. Note that a huge span of time, from a tiny fraction of a second to billions of years, is summarized in this diagram. In the first fraction of a second, the universe was unimaginably hot. By the time 0.01 second had elapsed, the temperature had dropped to 100 billion (10^{11}) K. After about 3 minutes, it had fallen to about 1 billion (10^9) K, still some 70 times hotter than the interior of the Sun. After a few hundred thousand years, the temperature was down to a mere 3000 K, and the universe has continued to cool since that time.

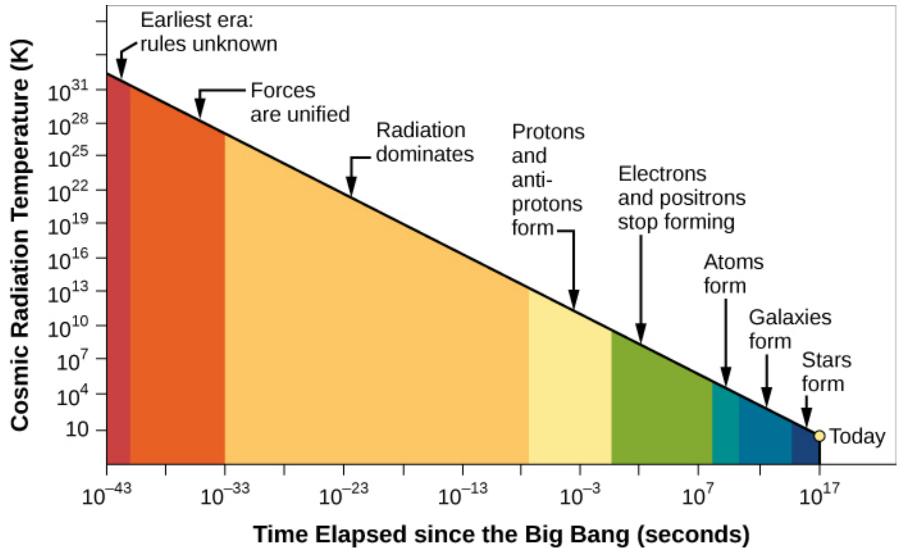


Figure 29.13 This graph shows how the temperature of the universe varies with time as predicted by the standard model of the Big Bang. Note that both the temperature (vertical axis) and the time in seconds (horizontal axis) change over vast scales on this compressed diagram.

All of these temperatures but the last are derived from theoretical calculations since (obviously) no one was there to measure them directly. As we shall see in the next section, however, we have actually detected the feeble glow of radiation emitted at a time when the universe was a few hundred thousand years old. We can measure the characteristics of that radiation to learn what things were like long ago. Indeed, the fact that we have found this ancient glow is one of the strongest arguments in favor of the Big Bang model.

The second step in understanding the evolution of the universe is to realize that at very early times, it was so hot that it contained mostly radiation (and not the matter that we see today). The photons that filled the universe could collide and produce material particles; that is, under the conditions just after the Big Bang, energy could turn into matter (and matter could turn into energy). We can calculate how much mass is produced from a given amount of energy by using Einstein's formula $E = mc^2$ (see the section on **Source of Sunshine: Nuclear Fusion!**).

The idea that energy could turn into matter in the universe at large is a new one for many students, since it is not part of our everyday experience. That's because, when we compare the universe today to what it was like right after the Big Bang, we live in cold, hard times. The photons in the universe today typically have far-less energy than the amount required to make new matter. In the discussion on the source of the Sun's energy in **Source of Sunshine: Nuclear Fusion!**, we briefly mentioned that when subatomic particles of matter and *antimatter* collide, they turn into pure energy. But the reverse, energy turning into matter and antimatter, is equally possible. This process has been observed in particle accelerators around the world. If we have enough energy, under the right circumstances, new particles of matter (and antimatter) are indeed created —and the conditions were right during the first few minutes after the expansion of the universe began.

Our third key point is that the hotter the universe was, the more energetic were the photons available to make matter and antimatter (see **Figure 29.13**). To take a specific example, at a temperature of 6 billion (6×10^9) K, the collision of two typical photons can create an electron and its antimatter counterpart, a positron. If the temperature exceeds 10^{14} K, much more massive protons and antiprotons can be created.

The Evolution of the Early Universe

Keeping these three ideas in mind, we can trace the evolution of the universe from the time it was about 0.01 second old and had a temperature of about 100 billion K. Why not begin at the very beginning? There are as yet no theories that allow us penetrate to a time before about $10^{–43}$ second (this number is a decimal point followed by 42 zeros and then a one). It is so small that we cannot relate it to anything in our everyday experience. When the universe was that young, its density was so high that the theory of general relativity is not adequate to describe it, and even the concept of time breaks down.

Scientists, by the way, have been somewhat more successful in describing the universe when it was older than $10^{–43}$ second but still less than about 0.01 second old. We will take a look at some of these ideas later in this chapter, but for now, we want to start with somewhat more familiar situations.

By the time the universe was 0.01 second old, it consisted of a soup of matter and radiation; the matter included protons and neutrons, leftovers from an even younger and hotter universe. Each particle collided rapidly with other particles. The temperature was no longer high enough to allow colliding photons to produce neutrons or protons, but it was sufficient for the production of electrons and positrons (Figure 29.14). There was probably also a sea of exotic subatomic particles that would later play a role as dark matter. All the particles jiggled about on their own; it was still much too hot for protons and neutrons to combine to form the nuclei of atoms.

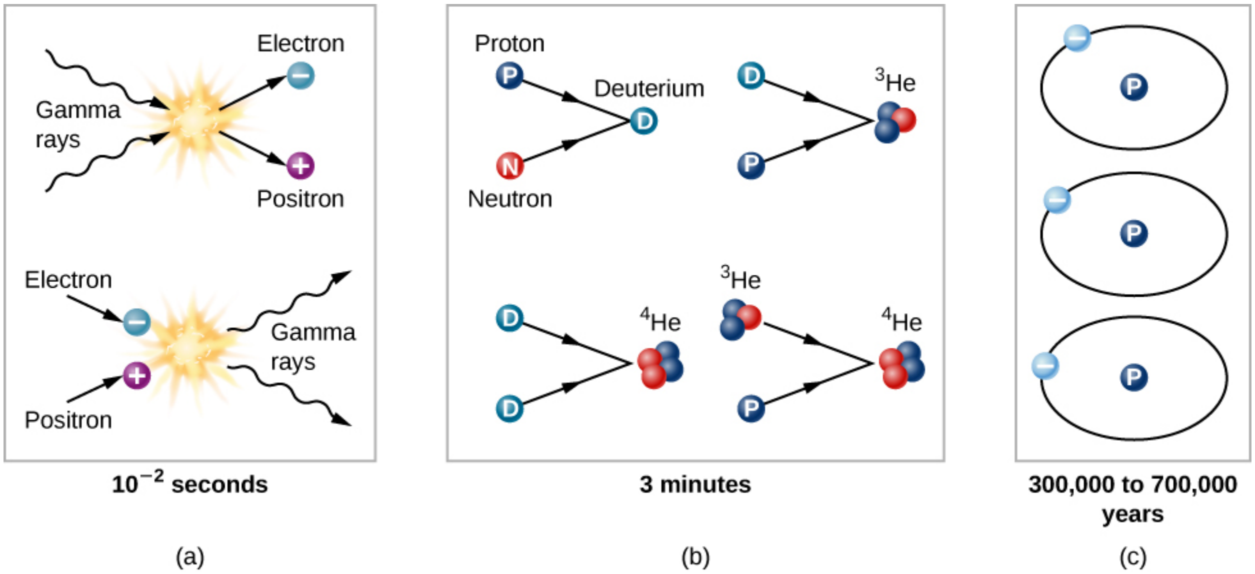


Figure 29.14 (a) In the first fractions of a second, when the universe was very hot, energy was converted into particles and antiparticles. The reverse reaction also happened: a particle and antiparticle could collide and produce energy. (b) As the temperature of the universe decreased, the energy of typical photons became too low to create matter. Instead, existing particles fused to create such nuclei as deuterium and helium. (c) Later, it became cool enough for electrons to settle down with nuclei and make neutral atoms. Most of the universe was still hydrogen.

Think of the universe at this time as a seething cauldron, with photons colliding and interchanging energy, and sometimes being destroyed to create a pair of particles. The particles also collided with one another. Frequently, a matter particle and an antimatter particle met and turned each other into a burst of gamma-ray radiation.

Among the particles created in the early phases of the universe was the ghostly neutrino (see Source of Sunshine: Nuclear Fusion!), which today interacts only very rarely with ordinary matter. In the crowded conditions of the very early universe, however, neutrinos ran into so many electrons and positrons that they experienced frequent interactions despite their “antisocial” natures.

By the time the universe was a little more than 1 second old, the density had dropped to the point where neutrinos no longer interacted with matter but simply traveled freely through space. In fact, these neutrinos should now be all around us. Since they have been traveling through space unimpeded (and hence unchanged) since the universe was 1 second old, measurements of their properties would offer one of the best tests of the Big Bang model. Unfortunately, the very characteristic that makes them so useful—the fact that they interact so weakly with matter that they have survived unaltered for all but the first second of time—also renders them unable to be measured, at least with present techniques. Perhaps someday someone will devise a way to capture these elusive messengers from the past.

Atomic Nuclei Form

When the universe was about 3 minutes old and its temperature was down to about 900 million K, protons and neutrons could combine. At higher temperatures, these atomic nuclei had immediately been blasted apart by interactions with high-energy photons and thus could not survive. But at the temperatures and densities reached between 3 and 4 minutes after the beginning, **deuterium** (a proton and neutron) lasted long enough that collisions could convert some of it into helium, (Figure 29.14). In essence, the entire universe was acting the way centers of stars do today—fusing new elements from simpler components. In addition, a little bit of element 3, **lithium**, could also form.

This burst of cosmic fusion was only a brief interlude, however. By 4 minutes after the Big Bang, more helium was having trouble forming. The universe was still expanding and cooling down. After the formation of helium and some lithium, the temperature had dropped so low that the fusion of helium nuclei into still-heavier elements could not occur. No elements beyond lithium could form in the first few minutes. That 4-minute period was the end of the time when the entire universe was a fusion factory. In the cool universe we know today, the fusion of new elements is limited to the centers of stars and the explosions of supernovae.

Still, the fact that the Big Bang model allows the creation of a good deal of helium is the answer to a long-standing mystery in astronomy. Put simply, there is just too much helium in the universe to be explained by what happens inside stars. All the generations of stars that have produced helium since the Big Bang cannot account for the quantity of helium we observe. Furthermore, even the oldest stars and the most distant galaxies show significant amounts of helium. These observations find a natural explanation in the synthesis of helium by the Big Bang itself during the first few minutes of time. We estimate that *10 times more helium* was manufactured in the first 4 minutes of the universe than in all the generations of stars during the succeeding 10 to 15 billion years.

These **nice animations (https://openstax.org/l/30origelemani)** that explain the way in which different elements formed in the history of the universe are from the University of Chicago's *Origins of the Elements* site.

Learning from Deuterium

We can learn many things from the way the early universe made atomic nuclei. It turns out that all of the deuterium (a hydrogen nucleus with a neutron in it) in the universe was formed during the first 4 minutes. In stars, any region hot enough to fuse two protons to form a deuterium nucleus is also hot enough to change it further—either by destroying it through a collision with an energetic photon or by converting it into helium through nuclear reactions.

The amount of deuterium that can be produced in the first 4 minutes of creation depends on the density of the universe at the time deuterium was formed. If the density were relatively high, nearly all the deuterium would have been converted into helium through interactions with protons, just as it is in stars. If the density were relatively low, then the universe would have expanded and thinned out rapidly enough that some deuterium would have survived. The amount of deuterium we see today thus gives us a clue to the density of the universe when it was about 4 minutes old. Theoretical models can relate the density then to the density now; thus, measurements of the abundance of deuterium today can give us an estimate of the current density of the universe.

The measurements of deuterium indicate that the present-day density of ordinary matter—protons and neutrons—is about 5×10^{-28} kg/m^3. Deuterium can only provide an estimate of the density of ordinary matter because the abundance of deuterium is determined by the particles that interact to form it, namely protons and neutrons alone. From the abundance of deuterium, we know that not enough protons and neutrons are present, by a factor of about 20, to produce a critical-density universe.

We do know, however, that there are dark matter particles that add to the overall matter density of the universe, which is then higher than what is calculated for ordinary matter alone. Because dark matter particles do not affect the production of deuterium, measurement of the deuterium abundance cannot tell us how much dark matter exists. Dark matter is made of some exotic kind of particle, not yet detected in any earthbound laboratory. It is definitely not made of protons and neutrons like the readers of this book.

29.4 | The Cosmic Microwave Background

Learning Objectives

By the end of this section, you will be able to:

- Explain why we can observe the afterglow of the hot, early universe
- Discuss the properties of this afterglow as we see it today, including its average temperature and the size of its temperature fluctuations
- Describe open, flat, and curved universes and explain which type of universe is supported by observations
- Summarize our current knowledge of the basic properties of the universe including its age and contents

The description of the first few minutes of the universe is based on theoretical calculations. It is crucial, however, that a scientific theory should be testable. What predictions does it make? And do observations show those predictions to be accurate? One success of the theory of the first few minutes of the universe is the correct prediction of the amount of helium in the universe.

Another prediction is that a significant milestone in the history of the universe occurred about 380,000 years after the Big Bang. Scientists have directly observed what the universe was like at this early stage, and these observations offer some of the strongest support for the Big Bang theory. To find out what this milestone was, let's look at what theory tells us about what happened during the first few hundred thousand years after the Big Bang.

The fusion of helium and lithium was completed when the universe was about 4 minutes old. The universe then continued to resemble the interior of a star in some ways for a few hundred thousand years more. It remained hot and opaque, with radiation being scattered from one particle to another. It was still too hot for electrons to "settle down" and become associated with a particular nucleus; such free electrons are especially effective at scattering photons, thus ensuring that no radiation ever got very far in the early universe without having its path changed. In a way, the universe was like an enormous crowd right after a popular concert; if you get separated from a friend, even if he is wearing a flashing button, it is impossible to see through the dense crowd to spot him. Only after the crowd clears is there a path for the light from his button to reach you.

The Universe Becomes Transparent

Not until a few hundred thousand years after the Big Bang, when the temperature had dropped to about 3000 K and the density of atomic nuclei to about 1000 per cubic centimeter, did the electrons and nuclei manage to combine to form stable atoms of hydrogen and helium (**Figure 29.14**). With no free electrons to scatter photons, the universe became transparent for the first time in cosmic history. From this point on, matter and radiation interacted much less frequently; we say that they *decoupled* from each other and evolved separately. Suddenly, electromagnetic radiation could really travel, and it has been traveling through the universe ever since.

Discovery of the Cosmic Background Radiation

If the model of the universe described in the previous section is correct, then—as we look far outward in the universe and thus far back in time—the first "afterglow" of the hot, early universe should still be detectable. Observations of it would be very strong evidence that our theoretical calculations about how the universe evolved are correct. As we shall see, we have indeed detected the radiation emitted at this **photon decoupling time**, when radiation began to stream freely through the universe without interacting with matter (**Figure 29.15**).

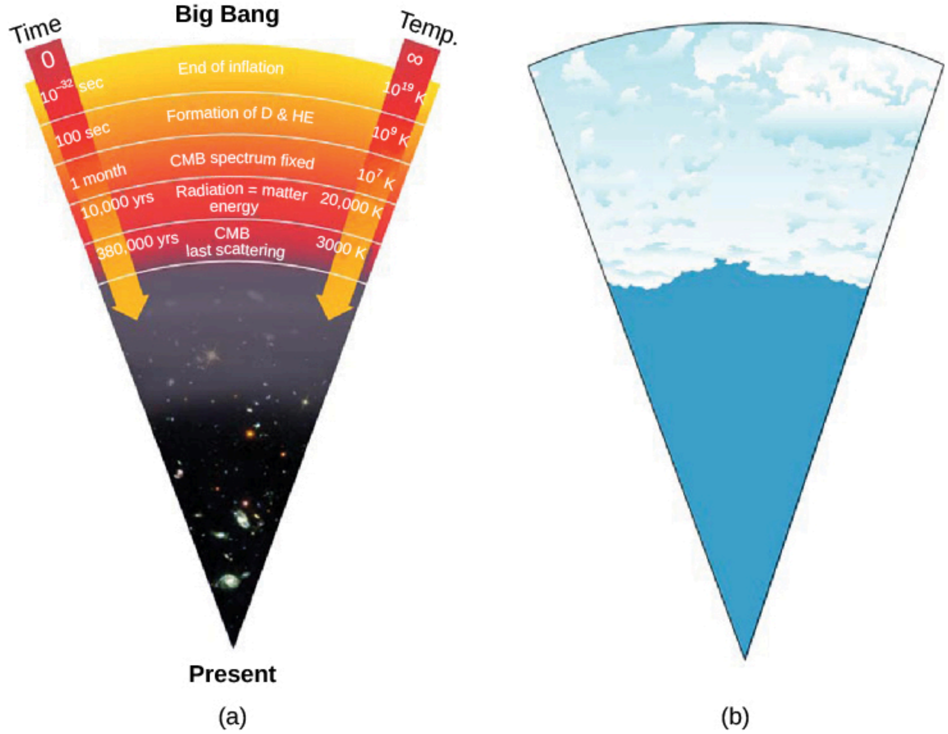


Figure 29.15 (a) Early in the universe, photons (electromagnetic energy) were scattering off the crowded, hot, charged particles and could not get very far without colliding with another particle. But after electrons and photons settled into neutral atoms, there was far less scattering, and photons could travel over vast distances. The universe became transparent. As we look out in space and back in time, we can't see back beyond this time. (b) This is similar to what happens when we see clouds in Earth's atmosphere. Water droplets in a cloud scatter light very efficiently, but clear air lets light travel over long distances. So as we look up into the atmosphere, our vision is blocked by the cloud layers and we can't see beyond them. (credit: modification of work by NASA)

The detection of this afterglow was initially an accident. In the late 1940s, Ralph Alpher and Robert Herman, working with George Gamow, realized that just before the universe became transparent, it must have been radiating like a blackbody at a temperature of about 3000 K—the temperature at which hydrogen atoms could begin to form. If we could have seen that radiation just after neutral atoms formed, it would have resembled radiation from a reddish star. It was as if a giant fireball filled the whole universe.

But that was nearly 14 billion years ago, and, in the meantime, the scale of the universe has increased a thousand fold. This expansion has increased the wavelength of the radiation by a factor of 1000 (see **Figure 29.7**). According to Wien's law, which relates wavelength and temperature, the expansion has correspondingly lowered the temperature by a factor of 1000 (see the section on **Blackbody Radiation**). The cosmic background behaves like a blackbody and should therefore have a spectrum that obeys Wien's Law.

Alpher and Herman predicted that the glow from the fireball should now be at radio wavelengths and should resemble the radiation from a blackbody at a temperature only a few degrees above absolute zero. Since the fireball was everywhere throughout the universe, the radiation left over from it should also be everywhere. If our eyes were sensitive to radio wavelengths, the whole sky would appear to glow very faintly. However, our eyes can't see at these wavelengths, and at the time Alpher and Herman made their prediction, there were no instruments that could detect the glow. Over the years, their prediction was forgotten.

In the mid-1960s, in Holmdel, New Jersey, Arno Penzias and Robert Wilson of AT&T's Bell Laboratories had built a delicate microwave antenna (**Figure 29.16**) to measure astronomical sources, including supernova remnants like Cassiopeia A (see the chapter on **The Deaths of Stars**). They were plagued with some unexpected background noise, just like faint static on a radio, which they could not get rid of. The puzzling thing about this radiation was that it seemed to be coming from all directions at once. This is very unusual in astronomy: after all, most radiation has a specific direction where it is strongest—the direction of the Sun, or a supernova remnant, or the disk of the Milky Way, for example.

Figure 29.16 These two scientists are standing in front of the horn-shaped antenna with which they discovered the cosmic background radiation. The photo was taken in 1978, just after they received the Nobel Prize in physics.

Penzias and Wilson at first thought that any radiation appearing to come from all directions must originate from inside their telescope, so they took everything apart to look for the source of the noise. They even found that some pigeons had roosted inside the big horn-shaped antenna and had left (as Penzias delicately put it) “a layer of white, sticky, dielectric substance coating the inside of the antenna.” However, nothing the scientists did could reduce the background radiation to zero, and they reluctantly came to accept that it must be real, and it must be coming from space.

Penzias and Wilson were not cosmologists, but as they began to discuss their puzzling discovery with other scientists, they were quickly put in touch with a group of astronomers and physicists at Princeton University (a short drive away). These astronomers had—as it happened—been redoing the calculations of Alpher and Herman from the 1940s and also realized that the radiation from the decoupling time should be detectable as a faint afterglow of radio waves. The different calculations of what the observed temperature would be for this **cosmic microwave background (CMB)**[2] were uncertain, but all predicted less than 40 K.

Penzias and Wilson found the distribution of intensity at different radio wavelengths to correspond to a temperature of 3.5 K. This is very cold—closer to absolute zero than most other astronomical measurements—and a testament to how much space (and the waves within it) has stretched. Their measurements have been repeated with better instruments, which give us a reading of 2.73 K. So Penzias and Wilson came very close. Rounding this value, scientists often refer to “the 3-degree microwave background.”

Many other experiments on Earth and in space soon confirmed the discovery by Penzias and Wilson: The radiation was indeed coming from all directions (it was isotropic) and matched the predictions of the Big Bang theory with remarkable precision. Penzias and Wilson had inadvertently observed the glow from the primeval fireball. They received the Nobel Prize for their work in 1978. And just before his death in 1966, Lemaître learned that his “vanished brilliance” had been discovered and confirmed.

You may enjoy watching *Three Degrees*, a **26-minute video (https://openstax.org/l/30threedegvid)** from Bell Labs about Penzias and Wilson’s discovery of the cosmic background radiation (with interesting historical footage).

Properties of the Cosmic Microwave Background

One issue that worried astronomers is that Penzias and Wilson were measuring the background radiation filling space through Earth’s atmosphere. What if that atmosphere is a source of radio waves or somehow affected their measurements? It would be better to measure something this important from space.

The first accurate measurements of the CMB were made with a satellite orbiting Earth. Named the Cosmic Background Explorer (COBE), it was launched by NASA in November 1989. The data it received quickly showed that the CMB closely matches that expected from a blackbody with a temperature of 2.73 K (**Figure 29.17**). This is exactly the result expected if the CMB was indeed redshifted radiation emitted by a hot gas that filled all of space shortly after the universe began.

2. Recall that microwaves are in the radio region of the electromagnetic spectrum.

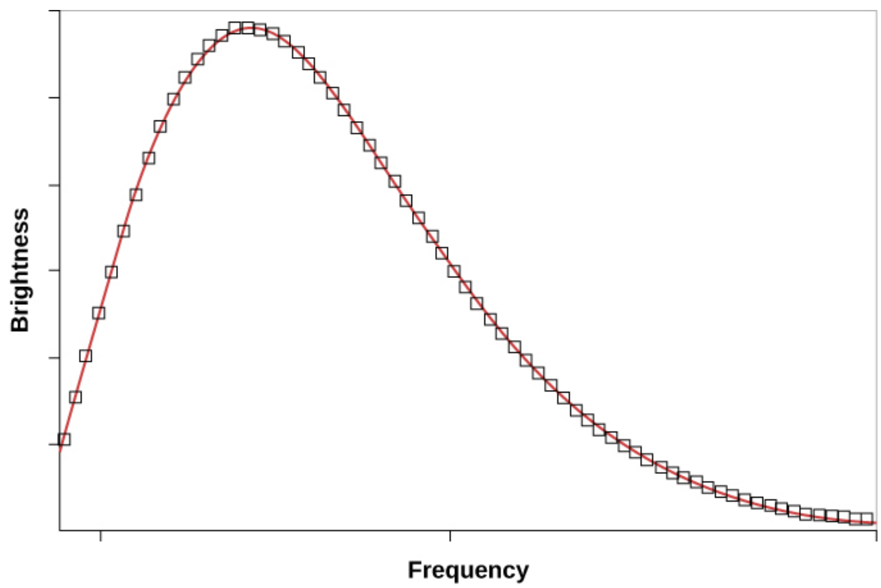


Figure 29.17 The solid line shows how the intensity of radiation should change with wavelength for a blackbody with a temperature of 2.73 K. The boxes show the intensity of the cosmic background radiation as measured at various wavelengths by COBE's instruments. The fit is perfect. When this graph was first shown at a meeting of astronomers, they gave it a standing ovation.

The first important conclusion from measurements of the CMB, therefore, is that the universe we have today has indeed evolved from a hot, uniform state. This observation also provides direct support for the general idea that we live in an evolving universe, since the universe is cooler today than it was in the beginning.

Small Differences in the CMB

It was known even before the launch of COBE that the CMB is extremely *isotropic*. In fact, its uniformity in every direction is one of the best confirmations of the cosmological principle— that the universe is homogenous and isotropic.

According to our theories, however, the temperature could not have been *perfectly* uniform when the CMB was emitted. After all, the CMB is radiation that was scattered from the particles in the universe at the time of decoupling. If the radiation were completely smooth, then all those particles must have been distributed through space absolutely evenly. Yet it is those particles that have become all the galaxies and stars (and astronomy students) that now inhabit the cosmos. Had the particles been completely smoothly distributed, they could not have formed all the large-scale structures now present in the universe—the clusters and superclusters of galaxies discussed in the last few chapters.

The early universe must have had tiny density fluctuations from which such structures could evolve. Regions of higher-than-average density would have attracted additional matter and eventually grown into the galaxies and clusters that we see today. It turned out that these denser regions would appear to us to be colder spots, that is, they would have lower-than-average temperatures.

The reason that temperature and density are related can be explained this way. At the time of decoupling, photons in a slightly denser portion of space had to expend some of their energy to escape the gravitational force exerted by the surrounding gas. In losing energy, the photons became slightly colder than the overall average temperature at the time of decoupling. Vice versa, photons that were located in a slightly less dense portion of space lost less energy upon leaving it than other photons, thus appearing slightly hotter than average. Therefore, if the seeds of present-day galaxies existed at the time that the CMB was emitted, we should see some slight variations in the CMB temperature as we look in different directions in the sky.

Scientists working with the data from the COBE satellite did indeed detect very subtle temperature differences—about 1 part in 100,000—in the CMB. The regions of lower-than-average temperature come in a variety of sizes, but even the smallest of the colder areas detected by COBE is far too large to be the precursor of an individual galaxy, or even a supercluster of galaxies. This is because the COBE instrument had "blurry vision" (poor resolution) and could only measure large patches of the sky. We needed instruments with "sharper vision."

The most detailed measurements of the CMB have been obtained by two satellites launched more recently than COBE. The results from the first of these satellites, the Wilkinson Microwave Anisotropy Probe (WMAP) spacecraft, were published in

2003. In 2015, measurements from the Planck satellite extended the WMAP measurements to even-higher spatial resolution and lower noise (**Figure 29.18**).

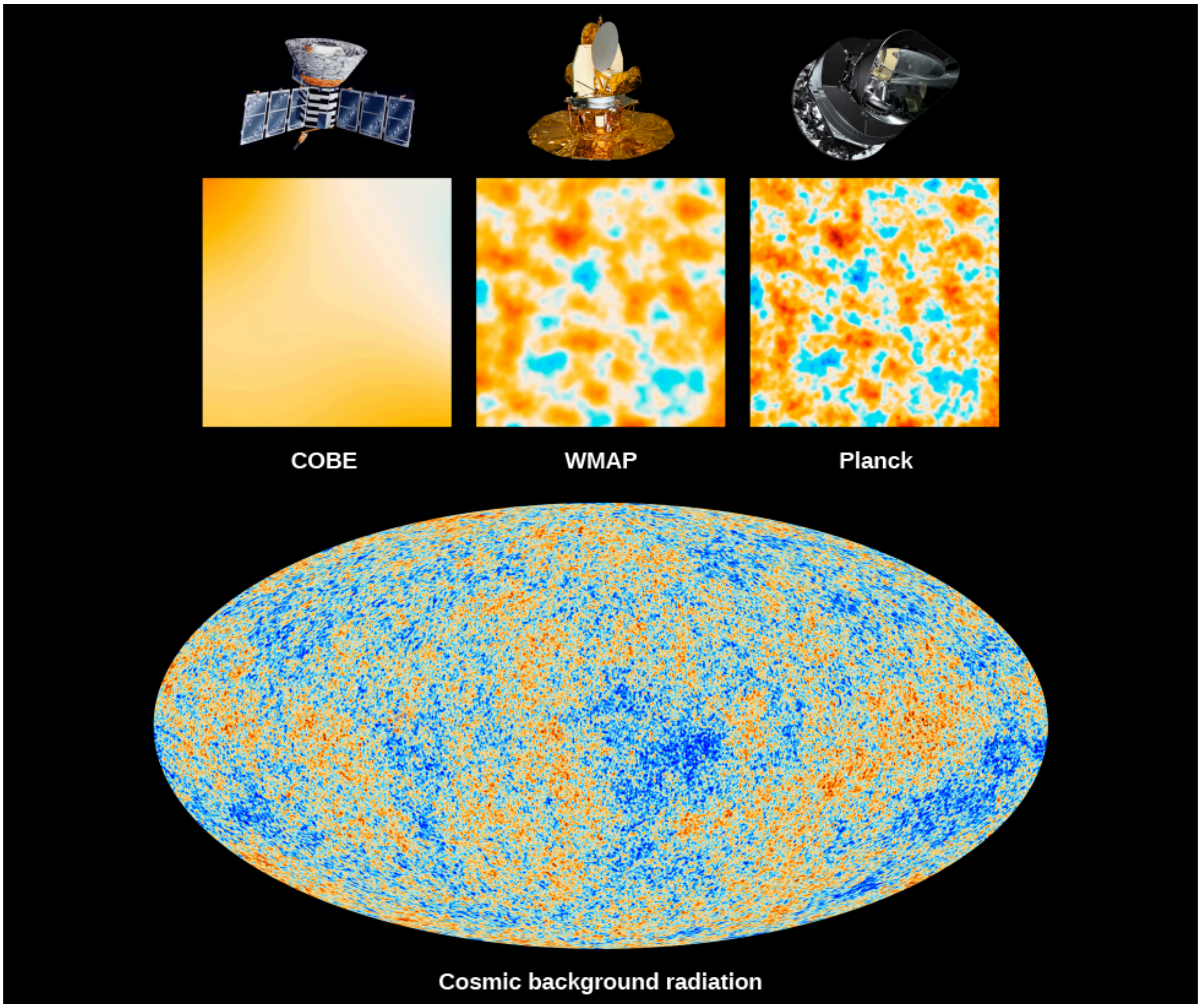


Figure 29.18 This comparison shows how much detail can be seen in the observations of three satellites used to measure the CMB. The CMB is a snapshot of the oldest light in our universe, imprinted on the sky when the universe was just about 380,000 years old. The first spacecraft, launched in 1989, is NASA's Cosmic Background Explorer, or COBE. WMAP was launched in 2001, and Planck was launched in 2009. The three panels show 10-square-degree patches of all-sky maps. This cosmic background radiation image (bottom) is an all-sky map of the CMB as observed by the Planck mission. The colors in the map represent different temperatures: red for warmer and blue for cooler. These tiny temperature fluctuations correspond to regions of slightly different densities, representing the seeds of all future structures: the stars, galaxies, and galaxy clusters of today. (credit top: modification of work by NASA/JPL-Caltech/ESA; credit bottom: modification of work by ESA and the Planck Collaboration)

Theoretical calculations show that the sizes of the hot and cold spots in the CMB depend on the geometry of the universe and hence on its total density. (It's not at all obvious that it should do so, and it takes some pretty fancy calculations—way beyond the level of our text—to make the connection, but having such a dependence is very useful.) The total density we are discussing here includes both the amount of mass in the universe and the mass equivalent of the dark energy. That is, we must add together mass and energy: ordinary matter, dark matter, and the dark energy that is speeding up the expansion.

To see why this works, remember from the section on **Introducing General Relativity** that Einstein showed that matter can curve space and that the amount of curvature depends on the amount of matter present. Therefore, the total amount of matter in the universe (including dark matter and the equivalent matter contribution by dark energy), determines the overall geometry of space. Just like the geometry of space around a black hole has a curvature to it, so the entire universe may have a curvature. Let's take a look at the possibilities (**Figure 29.19**).

If the density of matter is higher than the critical density, the universe will eventually collapse. In such a closed universe, two initially parallel rays of light will eventually meet. This kind of geometry is referred to as spherical geometry. If the density of matter is less than critical, the universe will expand forever. Two initially parallel rays of light will diverge, and this is referred to as hyperbolic geometry. In a critical-density universe, two parallel light rays never meet, and the expansion comes to a halt only at some time infinitely far in the future. We refer to this as a **flat universe**, and the kind of Euclidean geometry you learned in high school applies in this type of universe.

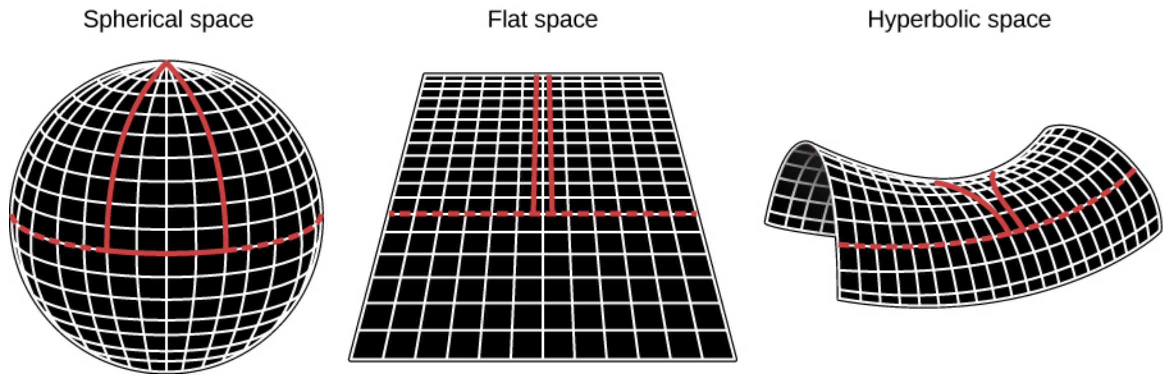


Figure 29.19 The density of matter and energy determines the overall geometry of space. If the density of the universe is greater than the critical density, then the universe will ultimately collapse and space is said to be *closed* like the surface of a sphere. If the density exactly equals the critical density, then space is *flat* like a sheet of paper; the universe will expand forever, with the rate of expansion coming to a halt infinitely far in the future. If the density is less than critical, then the expansion will continue forever and space is said to be *open* and negatively curved like the surface of a saddle (where more space than you expect opens up as you move farther away). Note that the red lines in each diagram show what happens in each kind of space—they are initially parallel but follow different paths depending on the curvature of space. Remember that these drawings are trying to show how space for the entire universe is "warped"—this can't be seen locally in the small amount of space that we humans occupy.

If the density of the universe is equal to the critical density, then the hot and cold spots in the CMB should typically be about a degree in size. If the density is greater than critical, then the typical sizes will be larger than one degree. If the universe has a density less than critical, then the structures will appear smaller. In **Figure 29.20**, you can see the differences easily. WMAP and Planck observations of the CMB confirmed earlier experiments that we do indeed live in a flat, critical-density universe.

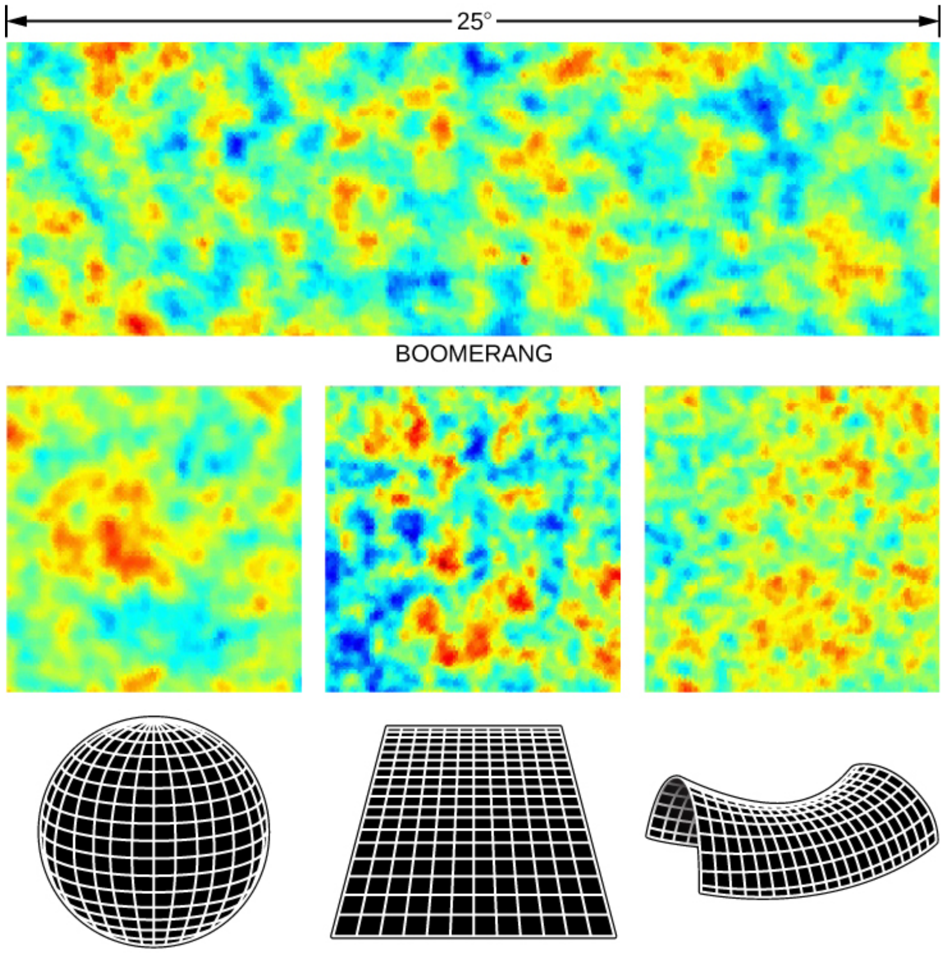


Figure 29.20 Cosmological simulations predict that if our universe has critical density, then the CMB images will be dominated by hot and cold spots of around one degree in size (bottom center). If, on the other hand, the density is higher than critical (and the universe will ultimately collapse), then the images' hot and cold spots will appear larger than one degree (bottom left). If the density of the universe is less than critical (and the expansion will continue forever), then the structures will appear smaller (bottom right). As the measurements show, the universe is at critical density. The measurements shown were made by a balloon-borne instrument called BOOMERanG (Balloon Observations of Millimetric Extragalactic Radiation and Geophysics), which was flown in Antarctica. Subsequent satellite observations by WMAP and Planck confirm the BOOMERanG result. (credit: modification of work by NASA)

Key numbers from an analysis of the Planck data give us the best values currently available for some of the basic properties of the universe:

- Age of universe: 13.799 ± 0.038 billion years (Note: That means we know the age of the universe to within 38 million years. Amazing!)
- Hubble constant: 67.31 ± 0.96 kilometers/second/million parsecs
- Fraction of universe's content that is "dark energy": 68.5% ± 1.3%
- Fraction of the universe's content that is matter: 31.5% ± 1.3%

Note that this value for the Hubble constant is slightly smaller than the value of 70 kilometers/second/million parsecs that we have adopted in this book. In fact, the value derived from measurements of redshifts is 73 kilometers/second/million parsecs. So precise is modern cosmology these days that scientists are working hard to resolve this discrepancy. The fact that the difference between these two independent measurements is so small is actually a remarkable achievement. Only a few decades ago, astronomers were arguing about whether the Hubble constant was around 50 kilometers/second/million parsecs or 100 kilometers/second/million parsecs.

Analysis of Planck data also shows that ordinary matter (mainly protons and neutrons) makes up 4.9% of the total density.

Dark matter plus normal matter add up to 31.5% of the total density. Dark energy contributes the remaining 68.5%. The age of the universe at decoupling—that is, when the CMB was emitted—was 380,000 years.

Perhaps the most surprising result from the high-precision measurements by WMAP and the even higher-precision measurements from Planck is that there were no surprises. The model of cosmology with ordinary matter at about 5%, dark matter at about 25%, and dark energy about 70% has survived since the late 1990s when cosmologists were forced in that direction by the supernovae data. In other words, the very strange universe that we have been describing, with only about 5% of its contents being made up of the kinds of matter we are familiar with here on Earth, really seems to be the universe we live in.

After the CMB was emitted, the universe continued to expand and cool off. By 400 to 500 million years after the Big Bang, the very first stars and galaxies had already formed. Deep in the interiors of stars, matter was reheated, nuclear reactions were ignited, and the more gradual synthesis of the heavier elements that we have discussed throughout this book began.

We conclude this quick tour of our model of the early universe with a reminder. You must not think of the Big Bang as a *localized* explosion *in space*, like an exploding superstar. There were no boundaries and there was no single site where the explosion happened. It was an explosion *of space* (and time and matter and energy) that happened everywhere in the universe. All matter and energy that exist today, including the particles of which you are made, came from the Big Bang. We were, and still are, in the midst of a Big Bang; it is all around us.

29.5 | What Is the Universe Really Made Of?

Learning Objectives

By the end of this section, you will be able to:

- Specify what fraction of the density of the universe is contributed by stars and galaxies and how much ordinary matter (such as hydrogen, helium, and other elements we are familiar with here on Earth) makes up the overall density
- Describe how ideas about the contents of the universe have changed over the last 50 years
- Explain why it is so difficult to determine what dark matter really is
- Explain why dark matter helped galaxies form quickly in the early universe
- Summarize the evolution of the universe from the time the CMB was emitted to the present day

The model of the universe we described in the previous section is the simplest model that explains the observations. It assumes that general relativity is the correct theory of gravity throughout the universe. With this assumption, the model then accounts for the existence and structure of the CMB; the abundances of the light elements deuterium, helium, and lithium; and the acceleration of the expansion of the universe. All of the observations to date support the validity of the model, which is referred to as the standard (or concordance) model of cosmology.

Figure 29.21 and **Table 29.2** summarize the current best estimates of the contents of the universe. Luminous matter in stars and galaxies and neutrinos contributes about 1% of the mass required to reach critical density. Another 4% is mainly in the form of hydrogen and helium in the space between stars and in intergalactic space. Dark matter accounts for about an additional 27% of the critical density. The mass equivalent of dark energy (according to $E = mc^2$) then supplies the remaining 68% of the critical density.

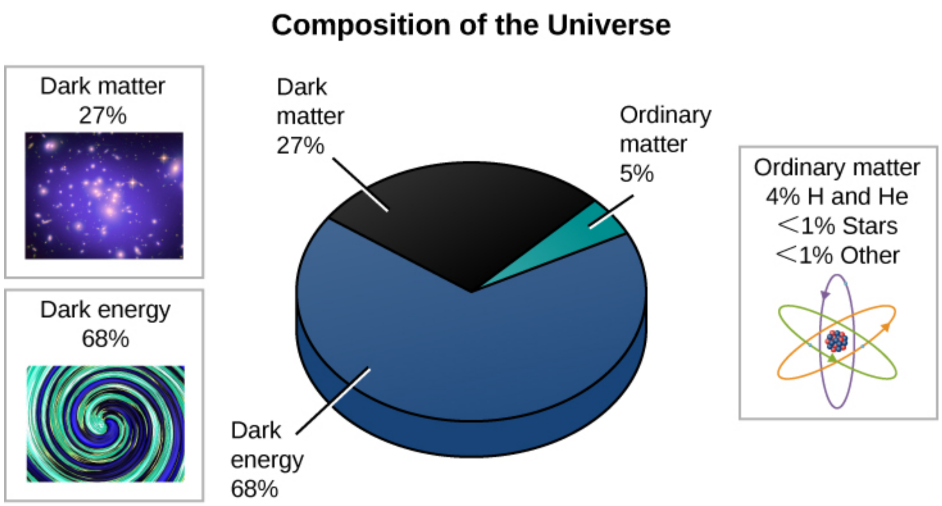


Figure 29.21 Only about 5% of all the mass and energy in the universe is matter with which we are familiar here on Earth. Most ordinary matter consists of hydrogen and helium located in interstellar and intergalactic space. Only about one-half of 1% of the critical density of the universe is found in stars. Dark matter and dark energy, which have not yet been detected in earthbound laboratories, account for 95% of the contents of the universe.

What Different Kinds of Objects Contribute to the Density of the Universe

Object	Density as a Percent of Critical Density
Luminous matter (stars, etc.)	<1
Hydrogen and helium in interstellar and intergalactic space	4
Dark matter	27
Equivalent mass density of the dark energy	68

Table 29.2

This table should shock you. What we are saying is that 95% of the stuff of the universe is either dark matter or dark energy—neither of which has ever been detected in a laboratory here on Earth. This whole textbook, which has focused on objects that emit electromagnetic radiation, has generally been ignoring 95% of what is out there. Who says there aren't big mysteries yet to solve in science!

Figure 29.22 shows how our ideas of the composition of the universe have changed over just the past three decades. The fraction of the universe that we think is made of the same particles as astronomy students has been decreasing steadily.

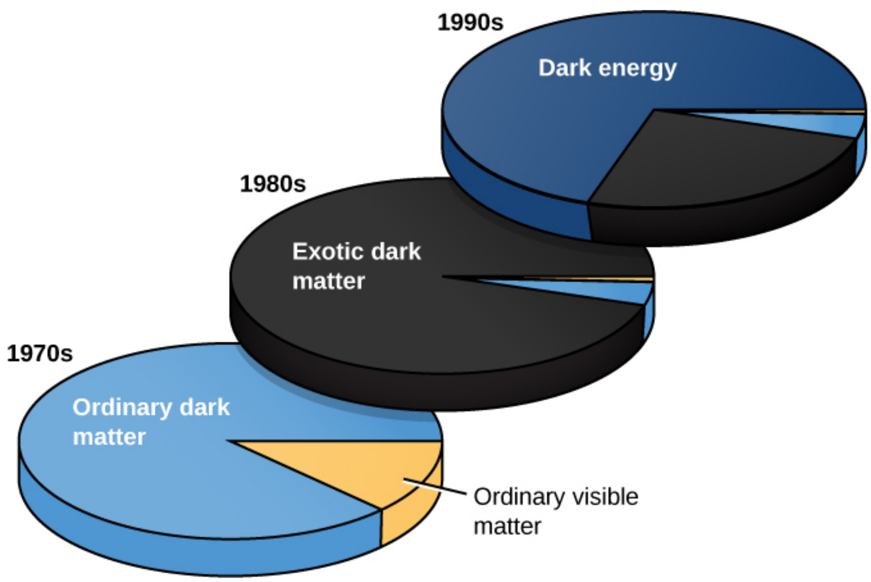


Figure 29.22 This diagram shows the changes in our understanding of the contents of the universe over the past three decades. In the 1970s, we suspected that most of the matter in the universe was invisible, but we thought that this matter might be ordinary matter (protons, neutrons, etc.) that was simply not producing electromagnetic radiation. By the 1980s, it was becoming likely that most of the dark matter was made of something we had not yet detected on Earth. By the late 1990s, a variety of experiments had shown that we live in a critical -density universe and that dark energy contributes about 70% of what is required to reach critical density. Note how the estimate of the relative importance of ordinary luminous matter (shown in yellow) has diminished over time.

What Is Dark Matter?

Many astronomers find the situation we have described very satisfying. Several independent experiments now agree on the type of universe we live in and on the inventory of what it contains. We seem to be very close to having a cosmological model that explains nearly everything. Others are not yet ready to jump on the bandwagon. They say, "show me the 96% of the universe we can't detect directly—for example, find me some dark matter!"

At first, astronomers thought that **dark matter** might be hidden in objects that appear dark because they emit no light (e.g., black holes) or that are too faint to be observed at large distances (e.g., planets or white dwarfs). However, these objects would be made of ordinary matter, and the deuterium abundance tells us that no more than 5% of the critical density consists of ordinary matter.

Another possible form that dark matter can take is some type of elementary particle that we have not yet detected here on Earth—a particle that has mass and exists in sufficient abundance to contribute 23% of the critical density. Some physics theories predict the existence of such particles. One class of these particles has been given the name WIMPs, which stands for **weakly interacting massive particles**. Since these particles do not participate in nuclear reactions leading to the production of deuterium, the deuterium abundance puts no limits on how many WIMPs might be in the universe. (A number of other exotic particles have also been suggested as prime constituents of dark matter, but we will confine our discussion to WIMPs as a useful example.)

If large numbers of WIMPs do exist, then some of them should be passing through our physics laboratories right now. The trick is to catch them. Since by definition they interact only weakly (infrequently) with other matter, the chances that they will have a measurable effect are small. We don't know the mass of these particles, but various theories suggest that it might be a few to a few hundred times the mass of a proton. If WIMPs are 60 times the mass of a proton, there would be about 10 million of them passing through your outstretched hand every second—with absolutely no effect on you. If that seems too mind-boggling, bear in mind that neutrinos interact weakly with ordinary matter, and yet we were able to "catch" them eventually.

Despite the challenges, more than 30 experiments designed to detect WIMPS are in operation or in the planning stages. Predictions of how many times WIMPs might actually collide with the nucleus of an atom in the instrument designed to detect them are in the range of 1 event per year to 1 event per 1000 years per kilogram of detector. The detector must therefore be large. It must be shielded from radioactivity or other types of particles, such as neutrons, passing through it, and hence these detectors are placed in deep mines. The energy imparted to an atomic nucleus in the detector by collision with a WIMP will be small, and so the detector must be cooled to a very low temperature.

The WIMP detectors are made out of crystals of germanium, silicon, or xenon. The detectors are cooled to a few thousandths of a degree—very close to absolute zero. That means that the atoms in the detector are so cold that they are scarcely vibrating at all. If a dark matter particle collides with one of the atoms, it will cause the whole crystal to vibrate and the temperature therefore to increase ever so slightly. Some other interactions may generate a detectable flash of light.

A different kind of search for WIMPs is being conducted at the Large Hadron Collider (LHC) at CERN, Europe's particle physics lab near Geneva, Switzerland. In this experiment, protons collide with enough energy potentially to produce WIMPs. The LHC detectors cannot detect the WIMPs directly, but if WIMPs are produced, they will pass through the detectors, carrying energy away with them. Experimenters will then add up all the energy that they detect as a result of the collisions of protons to determine if any energy is missing.

So far, none of these experiments has detected WIMPs. Will the newer experiments pay off? Or will scientists have to search for some other explanation for dark matter? Only time will tell (**Figure 29.23**).

Figure 29.23 This cartoon from NASA takes a humorous look at how little we yet understand about dark matter. (credit: NASA)

Dark Matter and the Formation of Galaxies

As elusive as dark matter may be in the current-day universe, galaxies could not have formed quickly without it. Galaxies grew from density fluctuations in the early universe, and some had already formed only about 400–500 million years after the Big Bang. The observations with WMAP, Planck, and other experiments give us information on the size of those density fluctuations. It turns out that the density variations we observe are too small to have formed galaxies so soon after the Big Bang. In the hot, early universe, energetic photons collided with hydrogen and helium, and kept them moving so rapidly that gravity was still not strong enough to cause the atoms to come together to form galaxies. How can we reconcile this with the fact that galaxies *did* form and are all around us?

Our instruments that measure the CMB give us information about density fluctuations only for *ordinary matter*, which interacts with radiation. Dark matter, as its name indicates, does not interact with photons at all. Dark matter could have had much greater variations in density and been able to come together to form gravitational "traps" that could then have begun to attract ordinary matter immediately after the universe became transparent. As ordinary matter became increasingly concentrated, it could have turned into galaxies quickly thanks to these dark matter traps.

For an analogy, imagine a boulevard with traffic lights every half mile or so. Suppose you are part of a motorcade of cars accompanied by police who lead you past each light, even if it is red. So, too, when the early universe was opaque, radiation interacted with ordinary matter, imparting energy to it and carrying it along, sweeping past the concentrations of dark matter. Now suppose the police leave the motorcade, which then encounters some red lights. The lights act as traffic traps; approaching cars now have to stop, and so they bunch up. Likewise, after the early universe became transparent, ordinary matter interacted with radiation only occasionally and so could fall into the dark matter traps.

The Universe in a Nutshell

In the previous sections of this chapter, we traced the evolution of the universe progressively further back in time. Astronomical discovery has followed this path historically, as new instruments and new techniques have allowed us to probe ever closer to the beginning of time. The rate of expansion of the universe was determined from measurements of nearby galaxies. Determinations of the abundances of deuterium, helium, and lithium based on nearby stars and galaxies were used to put limits on how much ordinary matter is in the universe. The motions of stars in galaxies and of galaxies within clusters of galaxies could only be explained if there were large quantities of dark matter. Measurements of supernovae that exploded when the universe was about half as old as it is now indicated that the rate of expansion of the universe has sped up since those explosions occurred. Observations of extremely faint galaxies show that galaxies had begun to form when the universe was only 400–500 million years old. And observations of the CMB confirmed early theories that the universe was initially very hot.

But all this moving further and further backward in time might have left you a bit dizzy. So now let’s instead show how the universe evolves as time moves forward.

Figure 29.24 summarizes the entire history of the observable universe from the beginning in a single diagram. The universe was very hot when it began to expand. We have fossil remnants of the very early universe in the form of neutrons, protons, electrons, and neutrinos, and the atomic nuclei that formed when the universe was 3–4 minutes old: deuterium, helium, and a small amount of lithium. Dark matter also remains, but we do not yet know what form it is in.

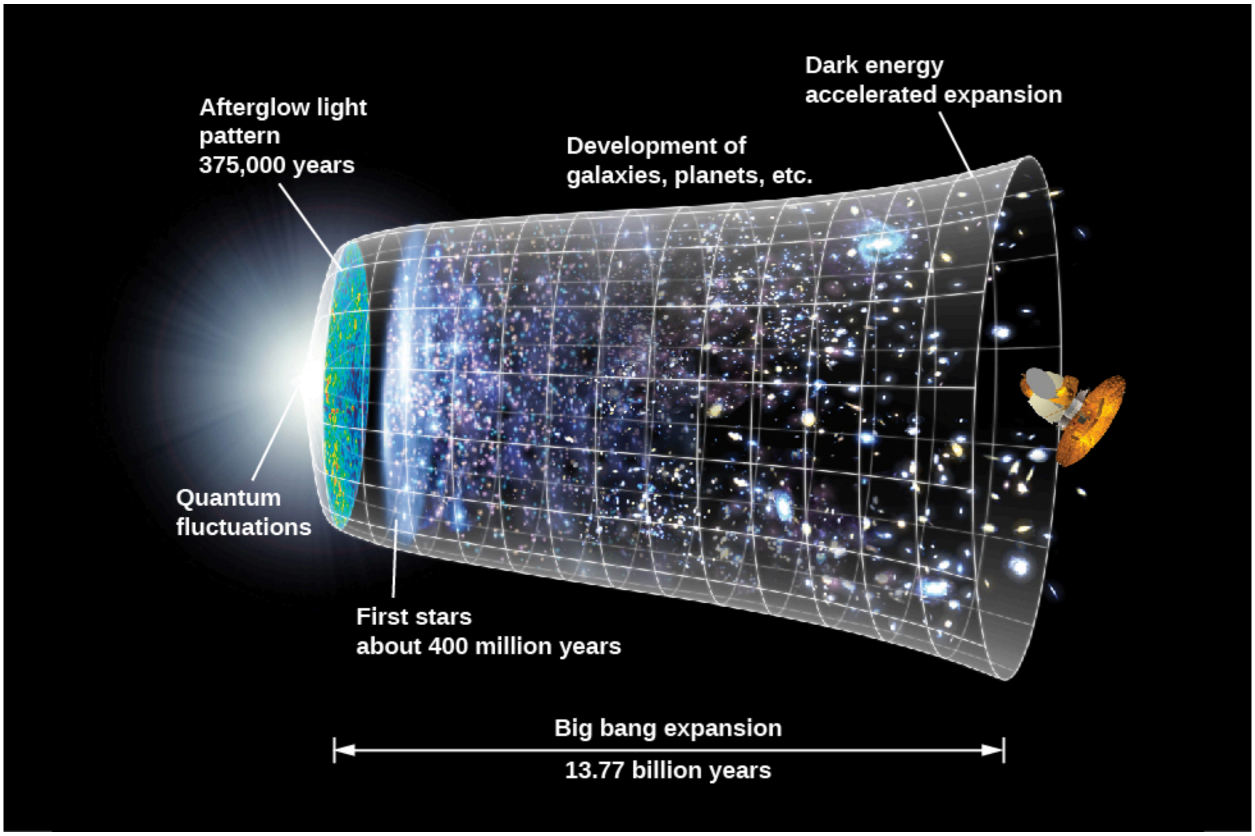


Figure 29.24 This image summarizes the changes that have occurred in the universe during the last 13.8 billion years. Protons, deuterium, helium, and some lithium were produced in the initial fireball. About 380,000 years after the Big Bang, the universe became transparent to electromagnetic radiation for the first time. COBE, WMAP, Planck, and other instruments have been used to study the radiation that was emitted at that time and that is still visible today (the CMB). The universe was then dark (except for this background radiation) until the first stars and galaxies began to form only a few hundred million years after the Big Bang. Existing space and ground-based telescopes have made substantial progress in studying the subsequent evolution of galaxies. (credit: modification of work by NASA/WMAP Science Team)

The universe gradually cooled; when it was about 380,000 years old, and at a temperature of about 3000 K, electrons combined with protons to form hydrogen atoms. At this point, as we saw, the universe became transparent to light, and astronomers have detected the CMB emitted at this time. The universe still contained no stars or galaxies, and so it entered what astronomers call "the dark ages" (since stars were not lighting up the darkness). During the next several hundred million years, small fluctuations in the density of the dark matter grew, forming gravitational traps that concentrated the ordinary matter, which began to form galaxies about 400–500 million years after the Big Bang.

By the time the universe was about a billion years old, it had entered its own renaissance: it was again blazing with radiation, but this time from newly formed stars, star clusters, and small galaxies. Over the next several billion years, small galaxies merged to form the giants we see today. Clusters and superclusters of galaxies began to grow, and the universe eventually began to resemble what we see nearby.

During the next 20 years, astronomers plan to build giant new telescopes both in space and on the ground to explore even further back in time. In 2018, the James Webb Space Telescope, a 6.5-meter telescope that is the successor to the Hubble Space Telescope, will be launched and assembled in space. The predictions are that with this powerful instrument (see **Figure 29.1**) we should be able to look back far enough to analyze in detail the formation of the first galaxies.

29.6 | The Inflationary Universe

Learning Objectives

By the end of this section, you will be able to:

- Describe two important properties of the universe that the simple Big Bang model cannot explain
- Explain why these two characteristics of the universe can be accounted for if there was a period of rapid expansion (inflation) of the universe just after the Big Bang
- Name the four forces that control all physical processes in the universe

The hot Big Bang model that we have been describing is remarkably successful. It accounts for the expansion of the universe, explains the observations of the CMB, and correctly predicts the abundances of the light elements. As it turns out, this model also predicts that there should be exactly three types of neutrinos in nature, and this prediction has been confirmed by experiments with high-energy accelerators. We can't relax just yet, however. This standard model of the universe doesn't explain *all* the observations we have made about the universe as a whole.

Problems with the Standard Big Bang Model

There are a number of characteristics of the universe that can only be explained by considering further what might have happened before the emission of the CMB. One problem with the standard Big Bang model is that it does not explain why the density of the universe is equal to the critical density. The mass density could have been, after all, so low and the effects of dark energy so high that the expansion would have been too rapid to form any galaxies at all. Alternatively, there could have been so much matter that the universe would have already begun to contract long before now. Why is the universe balanced so precisely on the knife edge of the critical density?

Another puzzle is the remarkable *uniformity* of the universe. The temperature of the CMB is the same to about 1 part in 100,000 everywhere we look. This sameness might be expected if all the parts of the visible universe were in contact at some point in time and had the time to come to the same temperature. In the same way, if we put some ice into a glass of lukewarm water and wait a while, the ice will melt and the water will cool down until they are the same temperature.

However, if we accept the standard Big Bang model, all parts of the visible universe were *not* in contact at any time. The fastest that information can go from one point to another is the speed of light. There is a maximum distance that light can have traveled from any point since the time the universe began—that's the distance light could have covered since then. This distance is called that point's *horizon distance* because anything farther away is "below its horizon"—unable to make contact with it. One region of space separated by more than the horizon distance from another has been completely isolated from it through the entire history of the universe.

If we measure the CMB in two opposite directions in the sky, we are observing regions that were significantly beyond each other's horizon distance at the time the CMB was emitted. We can see both regions, but *they* can never have seen each other. Why, then, are their temperatures so precisely the same? According to the standard Big Bang model, they have never been able to exchange information, and there is no reason they should have identical temperatures. (It's a little like seeing the clothes that all the students wear at two schools in different parts of the world become identical, without the students ever having been in contact.) The only explanation we could suggest was simply that the universe somehow *started out* being absolutely uniform (which is like saying all students were born liking the same clothes). Scientists are always uncomfortable when they must appeal to a special set of initial conditions to account for what they see.

The Inflationary Hypothesis

Some physicists suggested that these fundamental characteristics of the cosmos—its flatness and uniformity—can be explained if shortly after the Big Bang (and before the emission of the CMB), the universe experienced a sudden increase in size. A model universe in which this rapid, early expansion occurs is called an **inflationary universe**. The inflationary universe is identical to the Big Bang universe for all time after the first 10^{-30} second. Prior to that, the model suggests that there was a brief period of extraordinarily rapid expansion or inflation, during which the scale of the universe increased by a factor of about 10^{50} times more than predicted by standard Big Bang models (**Figure 29.25**).

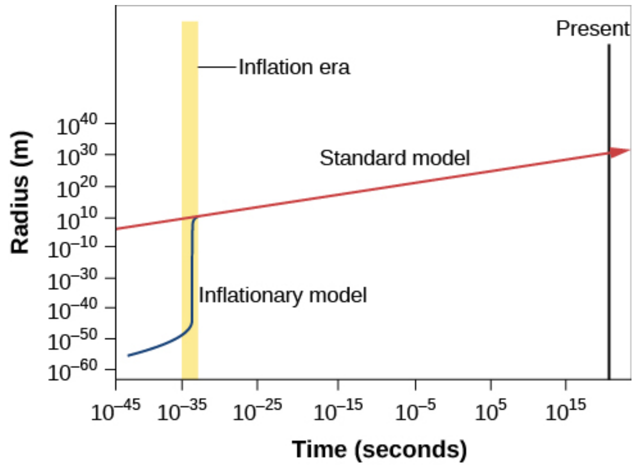


Figure 29.25 This graph shows how the scale factor of the observable universe changes with time for the standard Big Bang model (red line) and for the inflationary model (blue line). (Note that the time scale at the bottom is extremely compressed.) During inflation, regions that were very small and in contact with each other are suddenly blown up to be much larger and outside each other's horizon distance. The two models are the same for all times after 10–30 second.

Prior to (and during) inflation, all the parts of the universe that we can now see were so small and close to each other that they *could* exchange information, that is, the horizon distance included all of the universe that we can now observe. Before (and during) inflation, there was adequate time for the observable universe to homogenize itself and come to the same temperature. Then, inflation expanded those regions tremendously, so that many parts of the universe are now beyond each other's horizon.

Another appeal of the inflationary model is its prediction that the density of the universe should be exactly equal to the critical density. To see why this is so, remember that curvature of spacetime is intimately linked to the density of matter. If the universe began with some curvature of its spacetime, one analogy for it might be the skin of a balloon. The period of inflation was equivalent to blowing up the balloon to a tremendous size. The universe became so big that from our vantage point, no curvature should be visible (**Figure 29.26**). In the same way, Earth's surface is so big that it looks flat to us no matter where we are. Calculations show that a universe with no curvature is one that is at critical density. Universes with densities either higher or lower than the critical density would show marked curvature. But we saw that the observations of the CMB in **Figure 29.18**, which show that the universe has critical density, rule out the possibility that space is significantly curved.

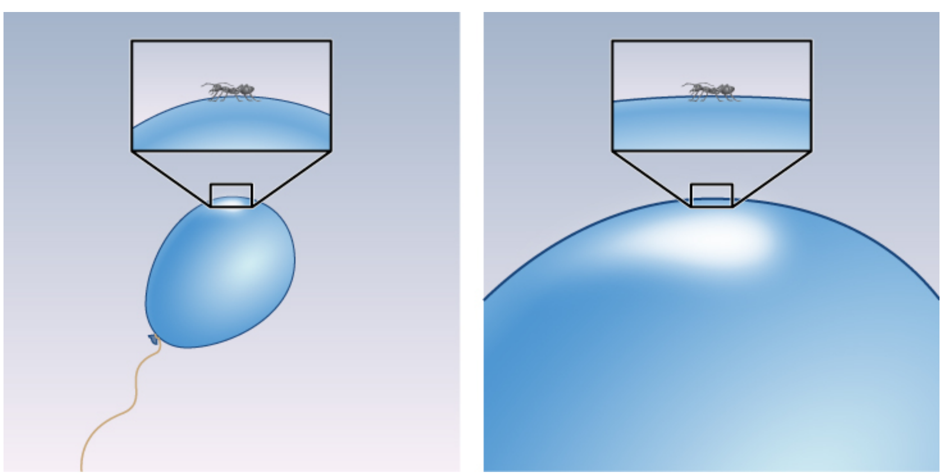


Figure 29.26 During a period of rapid inflation, a curved balloon grows so large that to any local observer it looks flat. The inset shows the geometry from the ant's point of view.

Grand Unified Theories

While inflation is an intriguing idea and widely accepted by researchers, we cannot directly observe events so early in the universe. The conditions at the time of inflation were so extreme that we cannot reproduce them in our laboratories or high-energy accelerators, but scientists have some ideas about what the universe might have been like. These ideas are called **grand unified theories** or GUTs.

In GUT models, the forces that we are familiar with here on Earth, including gravity and electromagnetism, behaved very differently in the extreme conditions of the early universe than they do today. In physical science, the term *force* is used to describe anything that can change the motion of a particle or body. One of the remarkable discoveries of modern science is that all known physical processes can be described through the action of just four forces: gravity, electromagnetism, the strong nuclear force, and the weak nuclear force (**Table 29.3**).

The Forces of Nature

Force	Relative Strength Today	Range of Action	Important Applications
Gravity	1	Whole universe	Motions of planets, stars, galaxies
Electromagnetism	10^{36}	Whole universe	Atoms, molecules, electricity, magnetic fields
Weak nuclear force	10^{33}	10^{-17} meters	Radioactive decay
Strong nuclear force	10^{38}	10^{-15} meters	The existence of atomic nuclei

Table 29.3

Gravity is perhaps the most familiar force, and certainly appears strong if you jump off a tall building. However, the force of gravity between two elementary particles—say two protons—is by far the weakest of the four forces. Electromagnetism—which includes both magnetic and electrical forces, holds atoms together, and produces the electromagnetic radiation that we use to study the universe—is much stronger, as you can see in **Table 29.3**. The weak nuclear force is only weak in comparison to its strong "cousin," but it is in fact much stronger than gravity.

Both the weak and strong nuclear forces differ from the first two forces in that they act only over very small distances—those comparable to the size of an atomic nucleus or less. The weak force is involved in radioactive decay and in reactions that result in the production of neutrinos. The strong force holds protons and neutrons together in an atomic nucleus.

Physicists have wondered why there are four forces in the universe—why not 300 or, preferably, just one? An important hint comes from the name *electromagnetic force*. For a long time, scientists thought that the forces of electricity and magnetism were separate, but James Clerk Maxwell was able to *unify* these forces—to show that they are aspects of the same phenomenon. In the same way, many scientists (including Einstein) have wondered if the four forces we now know could also be unified. Physicists have actually developed GUTs that unify three of the four forces (but not gravity).

In these theories, the strong, weak, and electromagnetic forces are not three independent forces but instead are different manifestations or aspects of what is, in fact, a single force. The theories predict that at high enough temperatures, there would be only one force. At lower temperatures (like the ones in the universe today), however, this single force has changed into three different forces (**Figure 29.27**). Just as different gases or liquids freeze at different temperatures, we can say that the different forces "froze out" of the unified force at different temperatures. Unfortunately, the temperatures at which the three forces acted as one force are so high that they cannot be reached in any laboratory on Earth. Only the early universe, at times prior to 10^{-35} second, was hot enough to unify these forces.

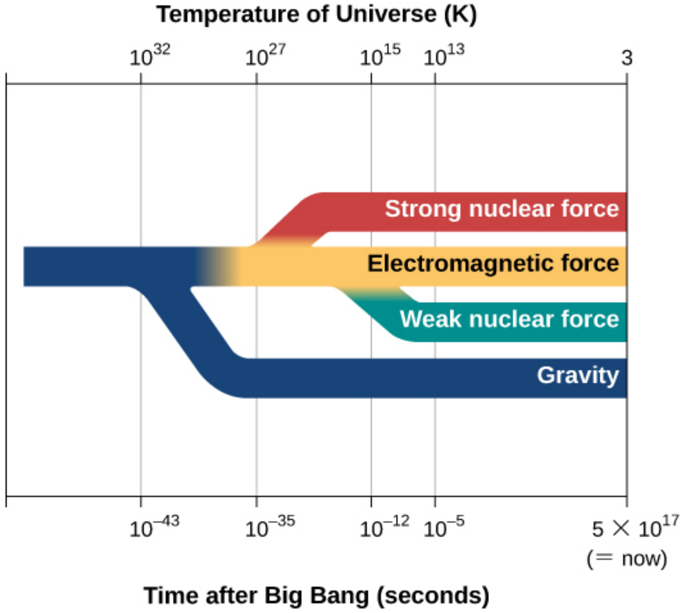


Figure 29.27 The behavior of the four forces depends on the temperature of the universe. This diagram (inspired by some grand unified theories) shows that at very early times when the temperature of the universe was very high, all four forces resembled one another and were indistinguishable. As the universe cooled, the forces took on separate and distinctive characteristics.

Many physicists think that gravity was also unified with the three other forces at still higher temperatures, and scientists have tried to develop a theory that combines all four forces. For example, in string theory, the point-like particles of matter that we have discussed in this book are replaced by one-dimensional objects called strings. In this theory, infinitesimal strings, which have length but not height or width, are the building blocks used to construct all the forms of matter and energy in the universe. These strings exist in 11-dimensional space (not the 4-dimensional spacetime with which we are familiar). The strings vibrate in the various dimensions, and depending on how they vibrate, they are seen in our world as matter or gravity or light. As you can imagine, the mathematics of string theory is very complex, and the theory remains untested by experiments. Even the largest particle accelerators on Earth do not achieve high enough energy to show whether string theory applies to the real world.

String theory is interesting to scientists because it is currently the only approach that seems to have the potential of combining all four forces to produce what physicists have termed the Theory of Everything.[3] Theories of the earliest phases of the universe must take both quantum mechanics and gravity into account, but at the simplest level, gravity and quantum mechanics are incompatible. General relativity, our best theory of gravity, says that the motions of objects can be predicted exactly. Quantum mechanics says you can only calculate the probability (chance) that an object will do something. String theory is an attempt to resolve this paradox. The mathematics that underpins string theory is elegant and beautiful, but it remains to be seen whether it will make predictions that can be tested by observations in yet-to-be-developed, high-energy accelerators on Earth or by observations of the early universe.

The earliest period in the history of the universe from time zero to 10^{-43} second is called the Planck time. The universe was unimaginably hot and dense, and theorists believe that at this time, quantum effects of gravity dominated physical interactions—and, as we have just discussed, we have no tested theory of quantum gravity. Inflation is hypothesized to have occurred somewhat later, when the universe was between perhaps 10^{-35} and 10^{-33} second old and the temperature was 10^{27} to 10^{28} K. This rapid expansion took place when three forces (electromagnetic, strong, and weak) are thought to have been unified, and this is when GUTs are applicable.

After inflation, the universe continued to expand (but more slowly) and to cool. An important milestone was reached when the temperature was down to 10^{15} K and the universe was 10^{-10} second old. Under these conditions, all four forces were separate and distinct. High-energy particle accelerators can achieve similar conditions, and so theories of the history of the universe from this point on have a sound basis in experiments.

As yet, we have no direct evidence of what the conditions were during the inflationary epoch, and the ideas presented here

3. This name became the title of a film about physicist Stephen Hawking in 2014.

are speculative. Researchers are trying to devise some experimental tests. For example, the quantum fluctuations in the very early universe would have caused variations in density and produced gravitational waves that may have left a detectable imprint on the CMB. Detection of such an imprint will require observations with equipment whose sensitivity is improved from what we have today. Ultimately, however, it may provide confirmation that we live in a universe that once experienced an epoch of rapid inflation.

If you are typical of the students who read this book, you may have found this brief discussion of dark matter, inflation, and cosmology a bit frustrating. We have offered glimpses of theories and observations, but have raised more questions than we have answered. What is dark matter? What is dark energy? Inflation explains the observations of flatness and uniformity of the university, but did it actually happen? These ideas are at the forefront of modern science, where progress almost always leads to new puzzles, and much more work is needed before we can see clearly. Bear in mind that less than a century has passed since Hubble demonstrated the existence of other galaxies. The quest to understand just how the universe of galaxies came to be will keep astronomers busy for a long time to come.

For Further Exploration

Websites

Cosmology Primer: **https://preposterousuniverse.com/cosmologyprimer/ (https://preposterousuniverse.com/cosmologyprimer/)** . Caltech Astrophysicist Sean Carroll offers a non-technical site with brief overviews of many key topics in modern cosmology.

Everyday Cosmology: **http://cosmology.carnegiescience.edu/ (http://cosmology.carnegiescience.edu/)** . An educational website from the Carnegie Observatories with a timeline of cosmological discovery, background materials, and activities.

How Big Is the Universe?: **http://www.pbs.org/wgbh/nova/space/how-big-universe.html (http://www.pbs.org/wgbh/nova/space/how-big-universe.html)** . A clear essay by a noted astronomer Brent Tully summarizes some key ideas in cosmology and introduces the notion of the acceleration of the universe.

Universe 101: WMAP Mission Introduction to the Universe: **http://map.gsfc.nasa.gov/universe/ (http://map.gsfc.nasa.gov/universe/)** . Concise NASA primer on cosmological ideas from the WMAP mission team.

Cosmic Times Project: **http://cosmictimes.gsfc.nasa.gov/ (http://cosmictimes.gsfc.nasa.gov/)** . James Lochner and Barbara Mattson have compiled a rich resource of twentieth-century cosmology history in the form of news reports on key events, from NASA's Goddard Space Flight Center.

Videos

The Day We Found the Universe: **http://www.cfa.harvard.edu/events/mon_video_archive09.html (http://www.cfa.harvard.edu/events/mon_video_archive09.html)** . Distinguished science writer Marcia Bartusiak discusses Hubble's work and the discovery of the expansion of the cosmos—one of the Observatory Night lectures at the Harvard-Smithsonian Center for Astrophysics (53:46).

Images of the Infant Universe: **https://www.youtube.com/watch?v=x0AqCwElyUk (https://www.youtube.com/watch?v=x0AqCwElyUk)** . Lloyd Knox's public talk on the latest discoveries about the CMB and what they mean for cosmology (1:16:00).

Runaway Universe: **https://www.youtube.com/watch?v=kNYVFrnmcOU (https://www.youtube.com/watch?v=kNYVFrnmcOU)** . Roger Blandford (Stanford Linear Accelerator Center) public lecture on the discovery and meaning of cosmic acceleration and dark energy (1:08:08).

From the Big Bang to the Nobel Prize and on to the James Webb Space Telescope and the Discovery of Alien Life: **http://svs.gsfc.nasa.gov/vis/a010000/a010300/a010370/index.html (http://svs.gsfc.nasa.gov/vis/a010000/a010300/a010370/index.html)** . John Mather, NASA Goddard (1:01:02). His Nobel Prize talk from Dec. 8, 2006 can be found at **http://www.nobelprize.org/mediaplayer/index.php?id=74&view=1 (http://www.nobelprize.org/mediaplayer/index.php?id=74&view=1)** .

Dark Energy and the Fate of the Universe: **https://webcast.stsci.edu/webcast/detail.xhtml?talkid=1961&parent=1 (https://webcast.stsci.edu/webcast/detail.xhtml?talkid=1961&parent=1)** . Adam Reiss (STScI), at the Space Telescope Science Institute (1:00:00).

CHAPTER 29 REVIEW

KEY TERMS

Big Bang the theory of cosmology in which the expansion of the universe began with a primeval explosion (of space, time, matter, and energy)

closed universe a model in which the universe expands from a Big Bang, stops, and then contracts to a big crunch

cosmic microwave background (CMB) microwave radiation coming from all directions that is the redshifted afterglow of the Big Bang

cosmological constant the term in the equations of general relativity that represents a repulsive force in the universe

cosmology the study of the organization and evolution of the universe

critical density in cosmology, the density that is just sufficient to bring the expansion of the universe to a stop after infinite time

dark energy the energy that is causing the expansion of the universe to accelerate; its existence is inferred from observations of distant supernovae

dark matter nonluminous material, whose nature we don't yet understand, but whose presence can be inferred because of its gravitational influence on luminous matter

deuterium a form of hydrogen in which the nucleus of each atom consists of one proton and one neutron

flat universe a model of the universe that has a critical density and in which the geometry of the universe is flat, like a sheet of paper

fusion the building of heavier atomic nuclei from lighter ones

grand unified theories (GUTs) physical theories that attempt to describe the four forces of nature as different manifestations of a single force

inflationary universe a theory of cosmology in which the universe is assumed to have undergone a phase of very rapid expansion when the universe was about 10^{-35} second old; after this period of rapid expansion, the standard Big Bang and inflationary models are identical

lithium the third element in the periodic table; lithium nuclei with three protons and four neutrons were manufactured during the first few minutes of the expansion of the universe

open universe a model in which the density of the universe is not high enough to bring the expansion of the universe to a halt

photon decoupling time when radiation began to stream freely through the universe without interacting with matter

weakly interacting massive particles (WIMPs) weakly interacting massive particles are one of the candidates for the composition of dark matter

KEY EQUATIONS

Critical density of the universe $\rho_{\text{crit}} = \frac{3H^2}{8\pi G}$

SUMMARY

29.1 The Age of the Universe

- Cosmology is the study of the organization and evolution of the universe.
- The universe is expanding, and this is one of the key observational starting points for modern cosmological theories.
- Modern observations show that the rate of expansion has not been constant throughout the life of the universe.
- Initially, when galaxies were close together, the effects of gravity were stronger than the effects of dark energy, and

the expansion rate gradually slowed.

- As galaxies moved farther apart, the influence of gravity on the expansion rate weakened.
- Measurements of distant supernovae show that when the universe was about half its current age, dark energy began to dominate the rate of expansion and caused it to speed up.
- In order to estimate the age of the universe, we must allow for changes in the rate of expansion. After allowing for these effects, astronomers estimate that all of the matter within the observable universe was concentrated in an extremely small volume 13.8 billion years ago, a time we call the Big Bang.

29.2 A Model of the Universe

- For describing the large-scale properties of the universe, a model that is isotropic and homogeneous (same everywhere) is a pretty good approximation of reality.
- The universe is expanding, which means that the universe undergoes a change in scale with time; space stretches and distances grow larger by the same factor everywhere at a given time.
- Observations show that the mass density of the universe is less than the critical density. In other words, there is not enough matter in the universe to stop the expansion.
- With the discovery of dark energy, which is accelerating the rate of expansion, the observational evidence is strong that the universe will expand forever.
- Observations tell us that the expansion started about 13.8 billion years ago.

29.3 The Beginning of the Universe

- Lemaître, Alpher, and Gamow first worked out the ideas that are today called the Big Bang theory.
- The universe cools as it expands.
- The energy of photons is determined by their temperature, and calculations show that in the hot, early universe, photons had so much energy that when they collided with one another, they could produce material particles.
- As the universe expanded and cooled, protons and neutrons formed first, then came electrons and positrons.
- Next, fusion reactions produced deuterium, helium, and lithium nuclei.
- Measurements of the deuterium abundance in today’s universe show that the total amount of ordinary matter in the universe is only about 5% of the critical density.

29.4 The Cosmic Microwave Background

- When the universe became cool enough to form neutral hydrogen atoms, the universe became transparent to radiation.
- Scientists have detected the cosmic microwave background (CMB) radiation from this time during the hot, early universe.
- Measurements with the COBE satellite show that the CMB acts like a blackbody with a temperature of 2.73 K.
- Tiny fluctuations in the CMB show us the seeds of large-scale structures in the universe. Detailed measurements of these fluctuations show that we live in a critical-density universe and that the critical density is composed of 31% matter, including dark matter, and 69% dark energy.
- Ordinary matter—the kinds of elementary particles we find on Earth—make up only about 5% of the critical density.
- CMB measurements also indicate that the universe is 13.8 billion years old.

29.5 What Is the Universe Really Made Of?

- Twenty-seven percent of the critical density of the universe is composed of dark matter.
- To explain so much dark matter, some physics theories predict that additional types of particles should exist.
- One type has been given the name of WIMPs (weakly interacting massive particles), and scientists are now conducting experiments to try to detect them in the laboratory.

- Dark matter plays an essential role in forming galaxies.
- Since, by definition, these particles interact only very weakly (if at all) with radiation, they could have congregated while the universe was still very hot and filled with radiation. They would thus have formed gravitational traps that quickly attracted and concentrated ordinary matter after the universe became transparent, and matter and radiation decoupled.
- This rapid concentration of matter enabled galaxies to form by the time the universe was only 400–500 million years old.

29.6 The Inflationary Universe

- The Big Bang model does not explain why the CMB has the same temperature in all directions.
- Neither does it explain why the density of the universe is so close to critical density.
- These observations can be explained if the universe experienced a period of rapid expansion, which scientists call inflation, about 10^{-35} second after the Big Bang.
- New grand unified theories (GUTs) are being developed to describe physical processes in the universe before and at the time that inflation occurred.

CONCEPTUAL QUESTIONS

29.1 The Age of the Universe

1. What are the basic observations about the universe that any theory of cosmology must explain?

2. What does the term Hubble time mean in cosmology, and what is the current best calculation for the Hubble time?

29.2 A Model of the Universe

3. Describe some possible futures for the universe that scientists have come up with. What property of the universe determines which of these possibilities is the correct one?

4. Some theorists expected that observations would show that the density of matter in the universe is just equal to the critical density. Do the current observations support this hypothesis?

5. What is dark energy and what evidence do astronomers have that it is an important component of the universe?

6. Would acceleration of the universe occur if it were composed entirely of matter (that is, if there were no dark energy)?

29.3 The Beginning of the Universe

7. Which formed first: hydrogen nuclei or hydrogen atoms? Explain the sequence of events that led to each.

29.4 The Cosmic Microwave Background

8. Penzias and Wilson's discovery of the Cosmic Microwave Background (CMB) is a nice example of scientific *serendipity*—something that is found by chance but turns out to have a positive outcome. What were they looking for and what did they discover?

29.5 What Is the Universe Really Made Of?

9. Why do astronomers believe there must be dark matter that is not in the form of atoms with protons and neutrons?

29.6 The Inflationary Universe

10. Describe at least two characteristics of the universe that are explained by the standard Big Bang model.

11. Describe two properties of the universe that are not explained by the standard Big Bang model (without inflation). How does inflation explain these two properties?

12. Describe the evidence that the expansion of the universe is accelerating.

PROBLEMS

29.1 The Age of the Universe

13. There is still some uncertainty in the Hubble constant. (a) Current estimates range from about 19.9 km/s per million light-years to 23 km/s per million light-years. Assume that the Hubble constant has been constant since the Big Bang. What is the possible range in the ages of the universe? Use the equation in the text, $T_0 = \frac{1}{H}$, and make sure you use consistent units. (b) Twenty years ago, estimates for the Hubble constant ranged from 50 to 100 km/s per Mps. What are the possible ages for the universe from those values? Can you rule out some of these possibilities on the basis of other evidence?

14. It is possible to derive the age of the universe given the value of the Hubble constant and the distance to a galaxy, again with the assumption that the value of the Hubble constant has not changed since the Big Bang. Consider a galaxy at a distance of 400 million light-years receding from us at a velocity, *v*. If the Hubble constant is 20 km/s per million light-years, what is its velocity? How long ago was that galaxy right next door to our own Galaxy if it has always been receding at its present rate? Express your answer in years. Since the universe began when all galaxies were very close together, this number is a rough estimate for the age of the universe.

29.2 A Model of the Universe

15. Suppose the Hubble constant were not 22 but 33 km/s per million light-years. Then what would the critical density be?

16. Assume that the average galaxy contains 10^{11} M_{Sun} and that the average distance between galaxies is 10 million light-years. Calculate the average density of matter (mass per unit volume) in galaxies. What fraction is this of the critical density we calculated in this section?

29.4 The Cosmic Microwave Background

17. The CMB contains roughly 400 million photons per m^3. The energy of each photon depends on its wavelength. Calculate the typical wavelength of a CMB photon. Hint: The CMB is blackbody radiation at a temperature of 2.73 K. According to Wien's law, the peak wave length in nanometers is given by $\lambda_{\text{max}} = \frac{3 \times 10^6}{T}$. Calculate the wavelength at which the CMB is a maximum and, to make the units consistent, convert this wavelength from nanometers to meters.

18. Following up on **Exercise 29.17** calculate the energy of a typical photon. Assume for this approximate calculation that each photon has the wavelength calculated in **Exercise 29.17**. The energy of a photon is given by $E = \frac{hc}{\lambda}$, where *h* is Planck's constant and is equal to 6.626×10^{-34} J × s, *c* is the speed of light in m/s, and λ is the wavelength in m.

19. Continuing the thinking in **Exercise 29.17** and , calculate the energy in a cubic meter of space, multiply the energy per photon calculated in **Exercise 29.18** by the number of photons per cubic meter given above.

20. Continuing the thinking in the last three exercises, convert this energy to an equivalent in mass, use Einstein's equation $E = mc^2$. Hint: Divide the energy per m^3 calculated in **Exercise 29.19** by the speed of light squared. Check your units; you should have an answer in kg/m^3. Now compare this answer with the critical density. Your answer should be several powers of 10 smaller than the critical density. In other words, you have found for yourself that the contribution of the CMB photons to the overall density of the universe is much, much smaller than the contribution made by stars and galaxies.

30 | GEOMETRIC OPTICS - LIGHT AS RAYS

Figure 30.1 Due to total internal reflection, an underwater swimmer's image is reflected back into the water where the camera is located. The circular ripple in the image center is actually on the water surface. Due to the viewing angle, total internal reflection is not occurring at the top edge of this image, and we can see a view of activities on the pool deck. (credit: modification of work by "jayhem"/Flickr)

Chapter Outline

Introduction

Our investigation of light revolves around two questions of fundamental importance: (1) What is the nature of light, and (2) how does light behave under various circumstances? Complete answers to these questions involve Maxwell's equations (from the study of electrodynamics), which predict the existence of electromagnetic waves and their behavior. Examples of light include radio and infrared waves, visible light, ultraviolet radiation, and X-rays. Interestingly, not all light phenomena can be explained by Maxwell's theory. Experiments performed early in the twentieth century showed that light has corpuscular, or particle-like, properties. The idea that light can display both wave and particle characteristics is called *wave-particle duality*, which is examined in **Section 13.**.

However, the basic study of light does not require most of the details from the study of electrodynamics. In this chapter, we study the basic properties of light. In the next few chapters, we investigate the behavior of light when it interacts with optical devices such as mirrors, lenses, and apertures.

30.1 | The Propagation of Light

Learning Objectives

By the end of this section, you will be able to:

- Determine the index of refraction, given the speed of light in a medium
- List the ways in which light travels from a source to another location

The speed of light in a vacuum *c* is one of the fundamental constants of physics. (This constancy is a central concept in Einstein's theory of relativity.) As the accuracy of the measurements of the speed of light improved, it was found that different observers, even those moving at large velocities with respect to each other, measure the same value for the speed of light. However, the speed of light does vary in a precise manner with the material it traverses. These facts have far-reaching implications, as we will see in later chapters.

The Speed of Light: Early Measurements

The first measurement of the speed of light was made by the Danish astronomer Ole Roemer (1644–1710) in 1675. He studied the orbit of Io, one of the four large moons of Jupiter, and found that it had a period of revolution of 42.5 h around Jupiter. He also discovered that this value fluctuated by a few seconds, depending on the position of Earth in its orbit around the Sun. Roemer realized that this fluctuation was due to the finite speed of light and could be used to determine *c*.

Roemer found the period of revolution of Io by measuring the time interval between successive eclipses by Jupiter. **Figure 30.2**(a) shows the planetary configurations when such a measurement is made from Earth in the part of its orbit where it is receding from Jupiter. When Earth is at point *A*, Earth, Jupiter, and Io are aligned. The next time this alignment occurs, Earth is at point *B*, and the light carrying that information to Earth must travel to that point. Since *B* is farther from Jupiter than *A*, light takes more time to reach Earth when Earth is at *B*. Now imagine it is about 6 months later, and the planets are arranged as in part (b) of the figure. The measurement of Io's period begins with Earth at point A' and Io eclipsed by Jupiter. The next eclipse then occurs when Earth is at point B', to which the light carrying the information of this eclipse must travel. Since B' is closer to Jupiter than A', light takes less time to reach Earth when it is at B'. This time interval between the successive eclipses of Io seen at A' and B' is therefore less than the time interval between the eclipses seen at *A* and *B*. By measuring the difference in these time intervals and with appropriate knowledge of the distance between Jupiter and Earth, Roemer calculated that the speed of light was 2.0×10^8 m/s, which is 33% below the value accepted today.

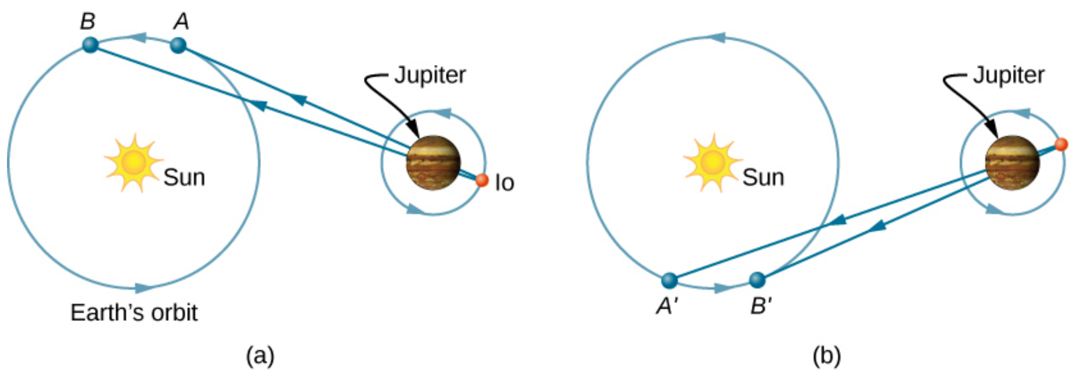


Figure 30.2 Roemer's astronomical method for determining the speed of light. Measurements of Io's period done with the configurations of parts (a) and (b) differ, because the light path length and associated travel time increase from *A* to *B* (a) but decrease from A' to B' (b).

The first successful terrestrial measurement of the speed of light was made by Armand Fizeau (1819–1896) in 1849. He placed a toothed wheel that could be rotated very rapidly on one hilltop and a mirror on a second hilltop 8 km away (**Figure 30.3**). An intense light source was placed behind the wheel, so that when the wheel rotated, it chopped the light beam into a succession of pulses. The speed of the wheel was then adjusted until no light returned to the observer located behind the wheel. This could only happen if the wheel rotated through an angle corresponding to a displacement of $(n + ½)$ teeth, while the pulses traveled down to the mirror and back. Knowing the rotational speed of the wheel, the number of teeth on the wheel, and the distance to the mirror, Fizeau determined the speed of light to be 3.15×10^8 m/s, which is only 5% too high.

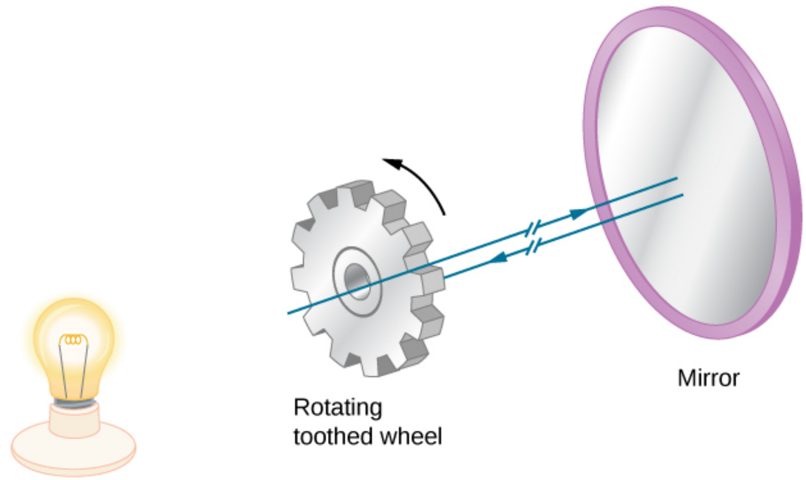


Figure 30.3 Fizeau's method for measuring the speed of light. The teeth of the wheel block the reflected light upon return when the wheel is rotated at a rate that matches the light travel time to and from the mirror.

The French physicist Jean Bernard Léon Foucault (1819–1868) modified Fizeau's apparatus by replacing the toothed wheel with a rotating mirror. In 1862, he measured the speed of light to be 2.98×10^8 m/s, which is within 0.6% of the presently accepted value. Albert Michelson (1852–1931) also used Foucault's method on several occasions to measure the speed of light. His first experiments were performed in 1878; by 1926, he had refined the technique so well that he found c to be $(2.99796 \pm 4) \times 10^8$ m/s.

Today, the speed of light is known to great precision. In fact, the speed of light in a vacuum c is so important that it is accepted as one of the basic physical quantities and has the value

$$c = 2.99792458 \times 10^8 \text{ m/s} \approx 3.00 \times 10^8 \text{ m/s} \tag{30.1}$$

where the approximate value of 3.00×10^8 m/s is used whenever three-digit accuracy is sufficient.

Speed of Light in Matter

The speed of light through matter is less than it is in a vacuum, because light interacts with atoms in a material. The speed of light depends strongly on the type of material, since its interaction varies with different atoms, crystal lattices, and other substructures. We can define a constant of a material that describes the speed of light in it, called the **index of refraction** n:

$$n = \frac{c}{v} \tag{30.2}$$

where v is the observed speed of light in the material.

Since the speed of light is always less than c in matter and equals c only in a vacuum, the index of refraction is always greater than or equal to one; that is, $n \geq 1$. **Table 30.1** gives the indices of refraction for some representative substances. The values are listed for a particular wavelength of light, because they vary slightly with wavelength. (This can have important effects, such as colors separated by a prism, as we will see in **Dispersion**.) Note that for gases, n is close to 1.0. This seems reasonable, since atoms in gases are widely separated, and light travels at c in the vacuum between atoms. It is common to take $n = 1$ for gases unless great precision is needed. Although the speed of light v in a medium varies considerably from its value c in a vacuum, it is still a large speed.

Medium	n
Gases at $0°C$, 1 atm	
Air	1.000293
Carbon dioxide	1.00045
Hydrogen	1.000139
Oxygen	1.000271
Liquids at $20°C$	
Benzene	1.501
Carbon disulfide	1.628
Carbon tetrachloride	1.461
Ethanol	1.361
Glycerine	1.473
Water, fresh	1.333
Solids at $20°C$	
Diamond	2.419
Fluorite	1.434
Glass, crown	1.52
Glass, flint	1.66
Ice (at $0°C$)	1.309
Polystyrene	1.49
Plexiglas	1.51
Quartz, crystalline	1.544
Quartz, fused	1.458
Sodium chloride	1.544
Zircon	1.923

Table 30.1 Index of Refraction in Various Media For light with a wavelength of 589 nm in a vacuum

Example 30.1

Speed of Light in Jewelry

Calculate the speed of light in zircon, a material used in jewelry to imitate diamond.

Strategy

We can calculate the speed of light in a material v from the index of refraction n of the material, using the equation $n = c/v$.

Solution

Rearranging the equation $n = c/v$ for v gives us

$$v = \frac{c}{n}.$$

The index of refraction for zircon is given as 1.923 in **Table 30.1**, and c is given in **Equation 30.1**. Entering these values in the equation gives

$$v = \frac{3.00 \times 10^8 \text{ m/s}}{1.923} = 1.56 \times 10^8 \text{ m/s}.$$

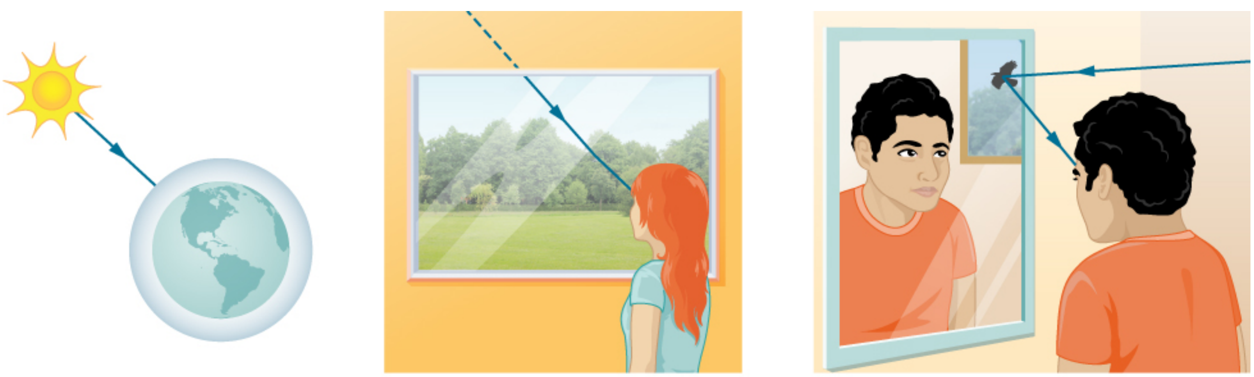
Significance

This speed is slightly larger than half the speed of light in a vacuum and is still high compared with speeds we normally experience. The only substance listed in **Table 30.1** that has a greater index of refraction than zircon is diamond. We shall see later that the large index of refraction for zircon makes it sparkle more than glass, but less than diamond.

30.1 Check Your Understanding **Table 30.1** shows that ethanol and fresh water have very similar indices of refraction. By what percentage do the speeds of light in these liquids differ?

The Ray Model of Light

You will eventually study **Physical Optics - Light as Waves**. In this chapter, however, we simplify things and start with the ray characteristics of light. There are three ways in which light can travel from a source to another location (**Figure 30.4**). It can come directly from the source through empty space, such as from the Sun to Earth. Or light can travel through various media, such as air and glass, to the observer. Light can also arrive after being reflected, such as by a mirror. In all of these cases, we can model the path of light as a straight line called a **ray**.



(a) (b) (c)

Figure 30.4 Three methods for light to travel from a source to another location. (a) Light reaches the upper atmosphere of Earth, traveling through empty space directly from the source. (b) Light can reach a person by traveling through media like air and glass. (c) Light can also reflect from an object like a mirror. In the situations shown here, light interacts with objects large enough that it travels in straight lines, like a ray.

Experiments show that when light interacts with an object several times larger than its wavelength, it travels in straight lines and acts like a ray. Its wave characteristics are not pronounced in such situations. Since the wavelength of visible light is less than a micron (a thousandth of a millimeter), it acts like a ray in the many common situations in which it encounters objects larger than a micron. For example, when visible light encounters anything large enough that we can observe it with unaided eyes, such as a coin, it acts like a ray, with generally negligible wave characteristics.

In all of these cases, we can model the path of light as straight lines. Light may change direction when it encounters objects (such as a mirror) or in passing from one material to another (such as in passing from air to glass), but it then continues in a straight line or as a ray. The word "ray" comes from mathematics and here means a straight line that originates at some point. It is acceptable to visualize light rays as laser rays. The *ray model of light* describes the path of light as straight lines.

Since light moves in straight lines, changing directions when it interacts with materials, its path is described by geometry and simple trigonometry. This part of optics, where the ray aspect of light dominates, is therefore called **geometric optics**. Two laws govern how light changes direction when it interacts with matter. These are the *law of reflection*, for situations in which light bounces off matter, and the *law of refraction*, for situations in which light passes through matter. We will examine more about each of these laws in upcoming sections of this chapter.

30.2 | The Law of Reflection

Learning Objectives

By the end of this section, you will be able to:

- Explain the reflection of light from polished and rough surfaces
- Describe the principle and applications of corner reflectors

Whenever we look into a mirror, or squint at sunlight glinting from a lake, we are seeing a reflection. When you look at a piece of white paper, you are seeing light scattered from it. Large telescopes use reflection to form an image of stars and other astronomical objects.

The **law of reflection** states that the angle of reflection equals the angle of incidence, or

$$\theta_r = \theta_i \tag{30.3}$$

The law of reflection is illustrated in **Figure 30.5**, which also shows how the angle of incidence and angle of reflection are measured relative to the perpendicular to the surface at the point where the light ray strikes.

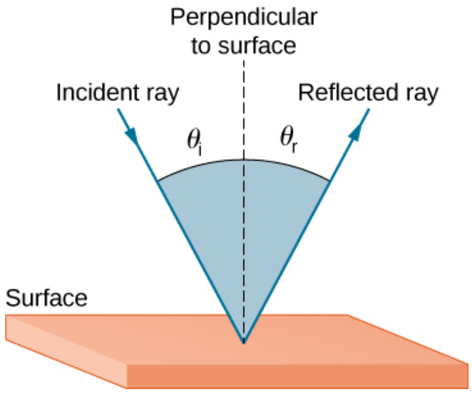


Figure 30.5 The law of reflection states that the angle of reflection equals the angle of incidence— $\theta_r = \theta_i$. The angles are measured relative to the perpendicular to the surface at the point where the ray strikes the surface.

We expect to see reflections from smooth surfaces, but **Figure 30.6** illustrates how a rough surface reflects light. Since the light strikes different parts of the surface at different angles, it is reflected in many different directions, or diffused. Diffused light is what allows us to see a sheet of paper from any angle, as shown in **Figure 30.7**(a). People, clothing, leaves, and walls all have rough surfaces and can be seen from all sides. A mirror, on the other hand, has a smooth surface (compared with the wavelength of light) and reflects light at specific angles, as illustrated in **Figure 30.7**(b). When the Moon reflects from a lake, as shown in **Figure 30.7**(c), a combination of these effects takes place.

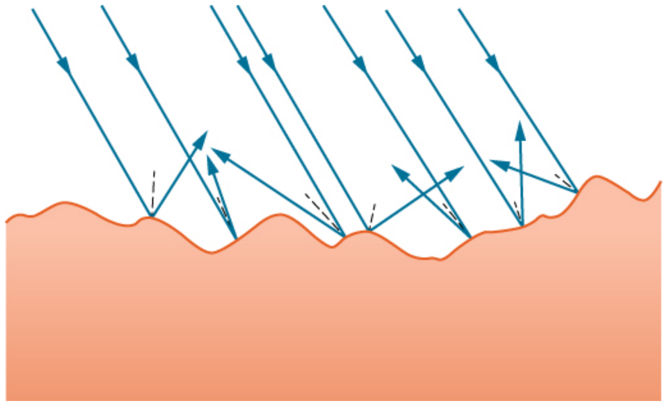


Figure 30.6 Light is diffused when it reflects from a rough surface. Here, many parallel rays are incident, but they are reflected at many different angles, because the surface is rough.

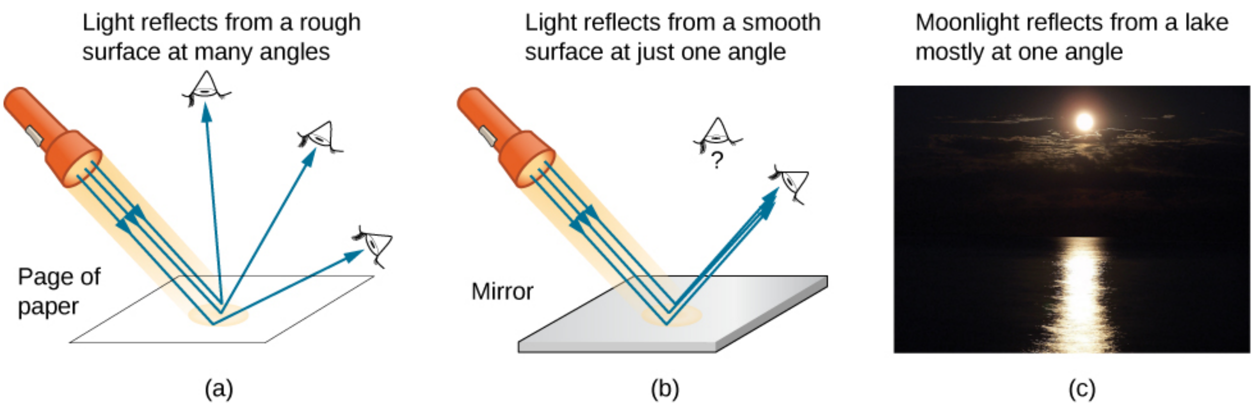


Figure 30.7 (a) When a sheet of paper is illuminated with many parallel incident rays, it can be seen at many different angles, because its surface is rough and diffuses the light. (b) A mirror illuminated by many parallel rays reflects them in only one direction, because its surface is very smooth. Only the observer at a particular angle sees the reflected light. (c) Moonlight is spread out when it is reflected by the lake, because the surface is shiny but uneven. (credit c: modification of work by Diego Torres Silvestre)

When you see yourself in a mirror, it appears that the image is actually behind the mirror (Figure 30.8). We see the light coming from a direction determined by the law of reflection. The angles are such that the image is exactly the same distance behind the mirror as you stand in front of the mirror. If the mirror is on the wall of a room, the images in it are all behind the mirror, which can make the room seem bigger. Although these mirror images make objects appear to be where they cannot be (like behind a solid wall), the images are not figments of your imagination. Mirror images can be photographed and videotaped by instruments and look just as they do with our eyes (which are optical instruments themselves). The precise manner in which images are formed by mirrors and lenses is discussed in an upcoming chapter on Image Formation.

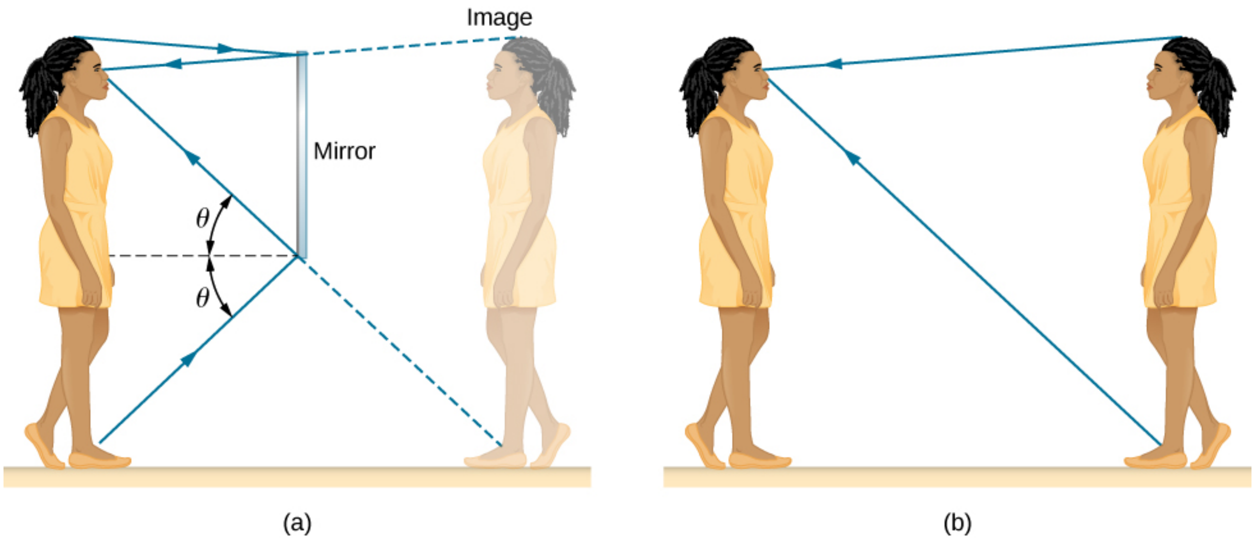


Figure 30.8 (a) Your image in a mirror is behind the mirror. The two rays shown are those that strike the mirror at just the correct angles to be reflected into the eyes of the person. The image appears to be behind the mirror at the same distance away as (b) if you were looking at your twin directly, with no mirror.

Corner Reflectors (Retroreflectors)

A light ray that strikes an object consisting of two mutually perpendicular reflecting surfaces is reflected back exactly parallel to the direction from which it came (**Figure 30.9**). This is true whenever the reflecting surfaces are perpendicular, and it is independent of the angle of incidence. (For proof, see at the end of this section.) Such an object is called a **corner reflector**, since the light bounces from its inside corner. Corner reflectors are a subclass of retroreflectors, which all reflect rays back in the directions from which they came. Although the geometry of the proof is much more complex, corner reflectors can also be built with three mutually perpendicular reflecting surfaces and are useful in three-dimensional applications.

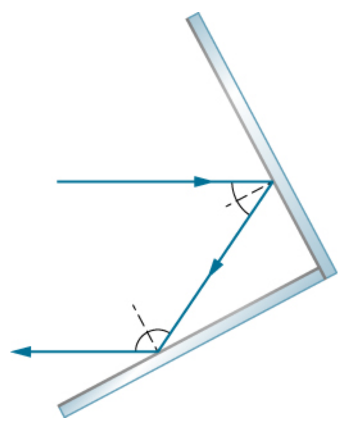
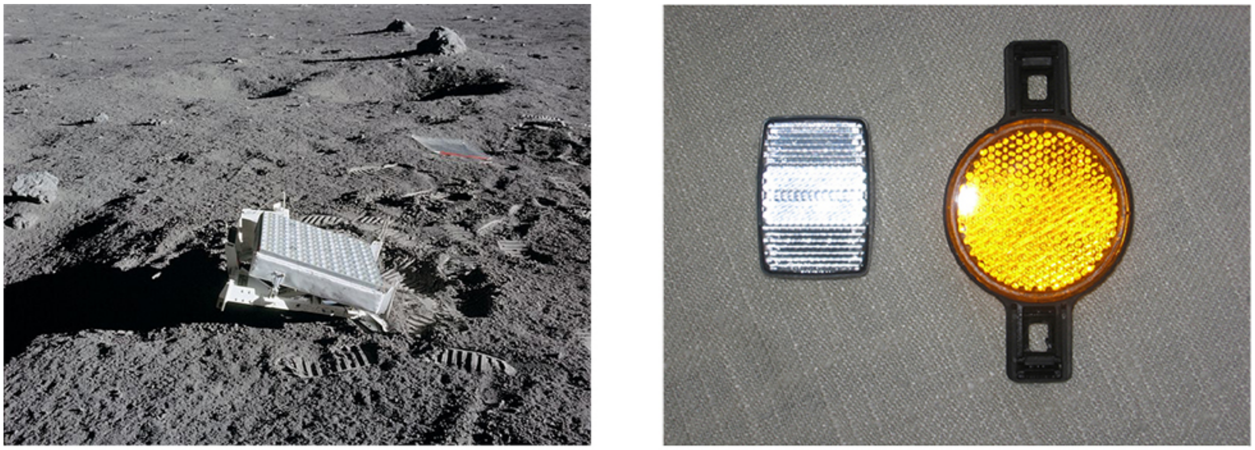


Figure 30.9 A light ray that strikes two mutually perpendicular reflecting surfaces is reflected back exactly parallel to the direction from which it came.

Many inexpensive reflector buttons on bicycles, cars, and warning signs have corner reflectors designed to return light in the direction from which it originated. Rather than simply reflecting light over a wide angle, retroreflection ensures high visibility if the observer and the light source are located together, such as a car's driver and headlights. The Apollo astronauts placed a true corner reflector on the Moon (**Figure 30.10**). Laser signals from Earth can be bounced from that corner reflector to measure the gradually increasing distance to the Moon of a few centimeters per year.



(a) (b)

Figure 30.10 (a) Astronauts placed a corner reflector on the Moon to measure its gradually increasing orbital distance. (b) The bright spots on these bicycle safety reflectors are reflections of the flash of the camera that took this picture on a dark night. (credit a: modification of work by NASA; credit b: modification of work by “Julo”/Wikimedia Commons)

Working on the same principle as these optical reflectors, corner reflectors are routinely used as radar reflectors (**Figure 30.11**) for radio-frequency applications. Under most circumstances, small boats made of fiberglass or wood do not strongly reflect radio waves emitted by radar systems. To make these boats visible to radar (to avoid collisions, for example), radar reflectors are attached to boats, usually in high places.

Figure 30.11 A radar reflector hoisted on a sailboat is a type of corner reflector. (credit: Tim Sheerman-Chase)

As a counterexample, if you are interested in building a stealth airplane, radar reflections should be minimized to evade detection. One of the design considerations would then be to avoid building $90°$ corners into the airframe.

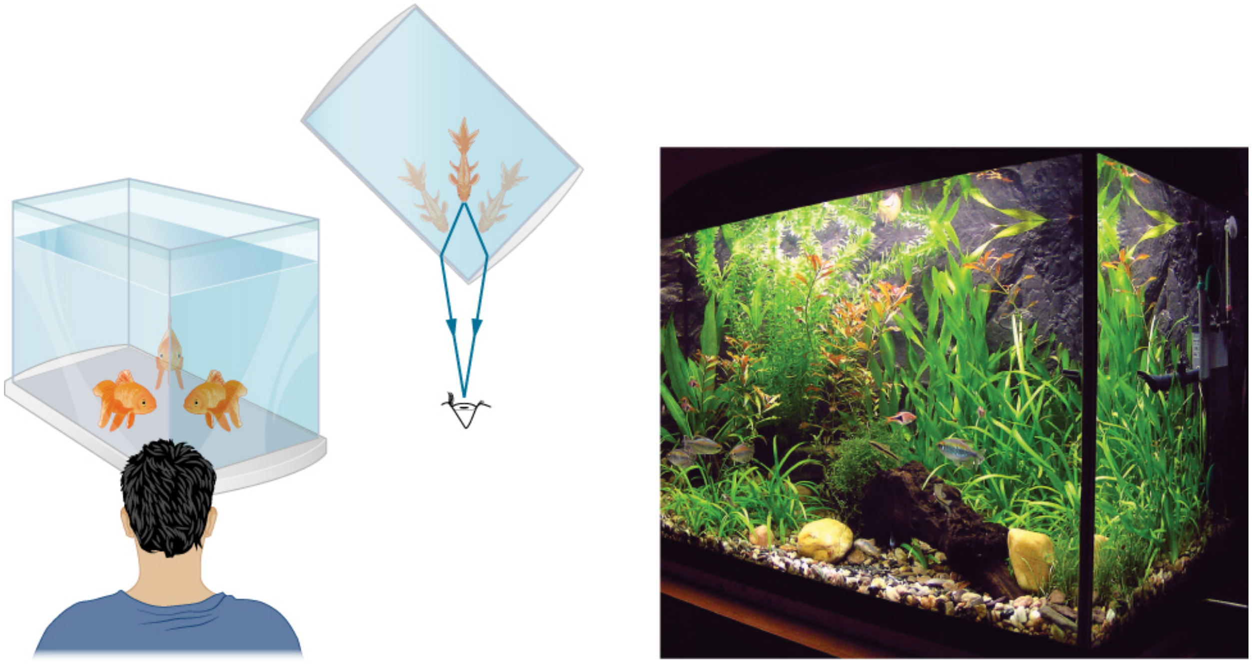
30.3 | Refraction

Learning Objectives

By the end of this section, you will be able to:

- Describe how rays change direction upon entering a medium
- Apply the law of refraction in problem solving

You may often notice some odd things when looking into a fish tank. For example, you may see the same fish appearing to be in two different places (**Figure 30.12**). This happens because light coming from the fish to you changes direction when it leaves the tank, and in this case, it can travel two different paths to get to your eyes. The changing of a light ray's direction (loosely called bending) when it passes through substances of different refractive indices is called **refraction** and is related to changes in the speed of light, $v = c/n$. Refraction is responsible for a tremendous range of optical phenomena, from the action of lenses to data transmission through optical fibers.



(a) (b)

Figure 30.12 (a) Looking at the fish tank as shown, we can see the same fish in two different locations, because light changes directions when it passes from water to air. In this case, the light can reach the observer by two different paths, so the fish seems to be in two different places. This bending of light is called refraction and is responsible for many optical phenomena. (b) This image shows refraction of light from a fish near the top of a fish tank.

Figure 30.13 shows how a ray of light changes direction when it passes from one medium to another. As before, the angles are measured relative to a perpendicular to the surface at the point where the light ray crosses it. (Some of the incident light is reflected from the surface, but for now we concentrate on the light that is transmitted.) The change in direction of the light ray depends on the relative values of the indices of refraction (**The Propagation of Light**) of the two media involved. In the situations shown, medium 2 has a greater index of refraction than medium 1. Note that as shown in **Figure 30.13**(a), the direction of the ray moves closer to the perpendicular when it progresses from a medium with a lower index of refraction to one with a higher index of refraction. Conversely, as shown in **Figure 30.13**(b), the direction of the ray moves away from the perpendicular when it progresses from a medium with a higher index of refraction to one with a lower index of refraction. The path is exactly reversible.

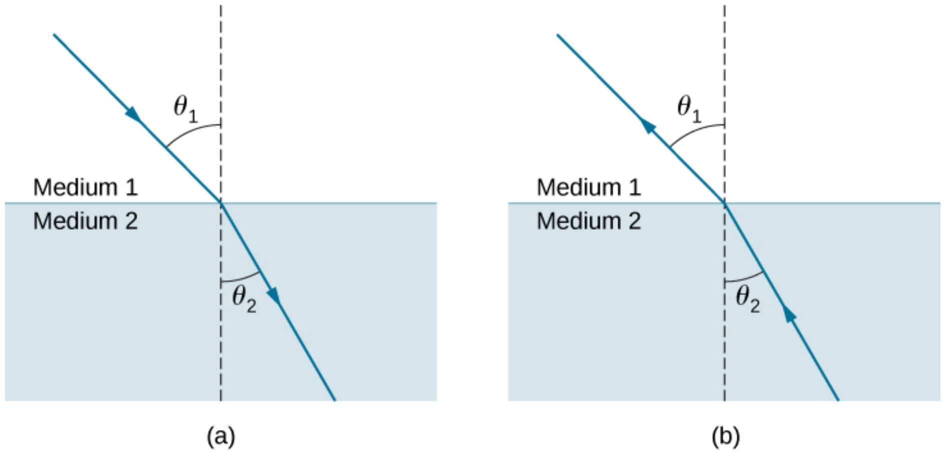


Figure 30.13 The change in direction of a light ray depends on how the index of refraction changes when it crosses from one medium to another. In the situations shown here, the index of refraction is greater in medium 2 than in medium 1. (a) A ray of light moves closer to the perpendicular when entering a medium with a higher index of refraction. (b) A ray of light moves away from the perpendicular when entering a medium with a lower index of refraction.

The amount that a light ray changes its direction depends both on the incident angle and the amount that the speed changes. For a ray at a given incident angle, a large change in speed causes a large change in direction and thus a large change in angle. The exact mathematical relationship is the **law of refraction**, or Snell's law, after the Dutch mathematician Willebrord Snell (1591–1626), who discovered it in 1621. The law of refraction is stated in equation form as

$$n_1 \sin\theta_1 = n_2 \sin\theta_2. \quad (30.4)$$

Here n_1 and n_2 are the indices of refraction for media 1 and 2, and θ_1 and θ_2 are the angles between the rays and the perpendicular in media 1 and 2. The incoming ray is called the incident ray, the outgoing ray is called the refracted ray, and the associated angles are the incident angle and the refracted angle, respectively.

Snell's experiments showed that the law of refraction is obeyed and that a characteristic index of refraction *n* could be assigned to a given medium and its value measured. (Snell was actually not aware that the speed of light varied in different media.)

Example 30.2

Determining the Index of Refraction

Find the index of refraction for medium 2 in **Figure 30.13**(a), assuming medium 1 is air and given that the incident angle is 30.0° and the angle of refraction is 22.0° .

Strategy

The index of refraction for air is taken to be 1 in most cases (and up to four significant figures, it is 1.000). Thus, $n_1 = 1.00$ here. From the given information, $\theta_1 = 30.0^\circ$ and $\theta_2 = 22.0^\circ$. With this information, the only unknown in Snell's law is n_2, so we can use Snell's law to find it.

Solution

From Snell's law we have

$$\begin{aligned} n_1 \sin\theta_1 &= n_2 \sin\theta_2 \\ n_2 &= n_1 \frac{\sin\theta_1}{\sin\theta_2}. \end{aligned}$$

Entering known values,

$$n_2 = 1.00\frac{\sin 30.0°}{\sin 22.0°} = \frac{0.500}{0.375} = 1.33.$$

Significance

This is the index of refraction for water, and Snell could have determined it by measuring the angles and performing this calculation. He would then have found 1.33 to be the appropriate index of refraction for water in all other situations, such as when a ray passes from water to glass. Today, we can verify that the index of refraction is related to the speed of light in a medium by measuring that speed directly.

Explore **bending of light (https://openstax.org/l/21bendoflight)** between two media with different indices of refraction. Use the "Intro" simulation and see how changing from air to water to glass changes the bending angle. Use the protractor tool to measure the angles and see if you can recreate the configuration in **Example 30.2**. Also by measurement, confirm that the angle of reflection equals the angle of incidence.

Example 30.3

A Larger Change in Direction

Suppose that in a situation like that in **Example 30.2**, light goes from air to diamond and that the incident angle is $30.0°$. Calculate the angle of refraction θ_2 in the diamond.

Strategy

Again, the index of refraction for air is taken to be $n_1 = 1.00$, and we are given $\theta_1 = 30.0°$. We can look up the index of refraction for diamond in **Table 30.1**, finding $n_2 = 2.419$. The only unknown in Snell's law is θ_2, which we wish to determine.

Solution

Solving Snell's law for $\sin\theta_2$ yields

$$\sin\theta_2 = \frac{n_1}{n_2}\sin\theta_1.$$

Entering known values,

$$\sin\theta_2 = \frac{1.00}{2.419}\sin 30.0° = (0.413)(0.500) = 0.207.$$

The angle is thus

$$\theta_2 = \sin^{-1}(0.207) = 11.9°.$$

Significance

For the same $30.0°$ angle of incidence, the angle of refraction in diamond is significantly smaller than in water ($11.9°$ rather than $22.0°$—see **Example 30.2**). This means there is a larger change in direction in diamond. The cause of a large change in direction is a large change in the index of refraction (or speed). In general, the larger the change in speed, the greater the effect on the direction of the ray.

30.2 Check Your Understanding In **Table 30.1**, the solid with the next highest index of refraction after diamond is zircon. If the diamond in **Example 30.3** were replaced with a piece of zircon, what would be the new angle of refraction?

30.4 | Total Internal Reflection

Learning Objectives

By the end of this section, you will be able to:

- Explain the phenomenon of total internal reflection
- Describe the workings and uses of optical fibers
- Analyze the reason for the sparkle of diamonds

A good-quality mirror may reflect more than 90% of the light that falls on it, absorbing the rest. But it would be useful to have a mirror that reflects all of the light that falls on it. Interestingly, we can produce total reflection using an aspect of refraction.

Consider what happens when a ray of light strikes the surface between two materials, as shown in Figure 30.14(a). Part of the light crosses the boundary and is refracted; the rest is reflected. If, as shown in the figure, the index of refraction for the second medium is less than for the first, the ray bends away from the perpendicular. (Since $n_1 > n_2,$ the angle of refraction is greater than the angle of incidence—that is, $\theta_2 > \theta_1$.) Now imagine what happens as the incident angle increases. This causes θ_2 to increase also. The largest the angle of refraction θ_2 can be is $90°$, as shown in part (b). The **critical angle** θ_c for a combination of materials is defined to be the incident angle θ_1 that produces an angle of refraction of $90°$. That is, θ_c is the incident angle for which $\theta_2 = 90°$. If the incident angle θ_1 is greater than the critical angle, as shown in Figure 30.14(c), then all of the light is reflected back into medium 1, a condition called **total internal reflection**. (As the figure shows, the reflected rays obey the law of reflection so that the angle of reflection is equal to the angle of incidence in all three cases.)

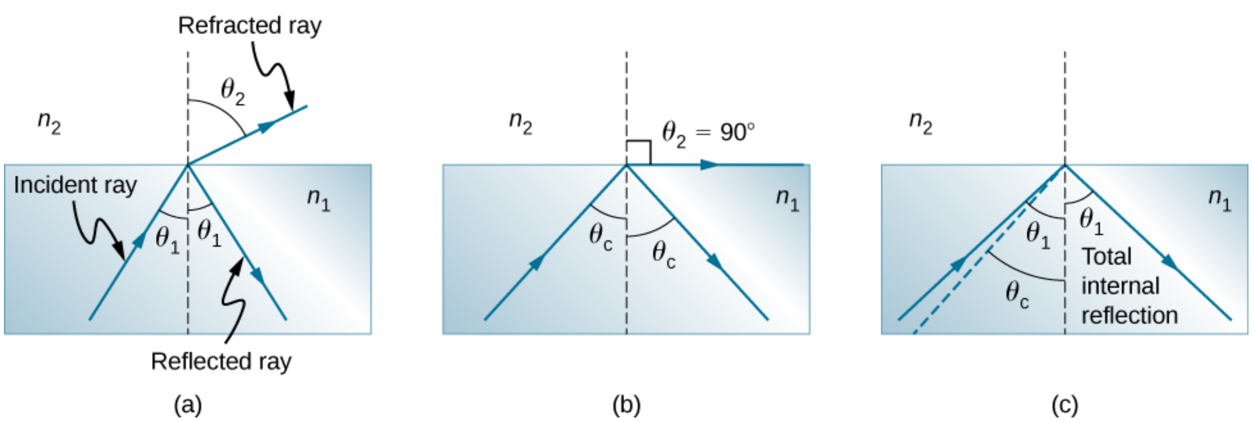


Figure 30.14 (a) A ray of light crosses a boundary where the index of refraction decreases. That is, $n_2 < n_1$. The ray bends away from the perpendicular. (b) The critical angle θ_c is the angle of incidence for which the angle of refraction is $90°$. (c) Total internal reflection occurs when the incident angle is greater than the critical angle.

Snell's law states the relationship between angles and indices of refraction. It is given by

$$n_1 \sin\theta_1 = n_2 \sin\theta_2.$$

When the incident angle equals the critical angle $(\theta_1 = \theta_c)$, the angle of refraction is $90°$ $(\theta_2 = 90°)$. Noting that $\sin 90° = 1,$ Snell's law in this case becomes

$$n_1 \sin\theta_1 = n_2.$$

The critical angle θ_c for a given combination of materials is thus

$$\theta_c = \sin^{-1}\left(\frac{n_2}{n_1}\right) \text{ for } n_1 > n_2. \quad (30.5)$$

Total internal reflection occurs for any incident angle greater than the critical angle θ_c, and it can only occur when the second medium has an index of refraction less than the first. Note that this equation is written for a light ray that travels in medium 1 and reflects from medium 2, as shown in **Figure 30.14**.

Example 30.4

Determining a Critical Angle

What is the critical angle for light traveling in a polystyrene (a type of plastic) pipe surrounded by air? The index of refraction for polystyrene is 1.49.

Strategy

The index of refraction of air can be taken to be 1.00, as before. Thus, the condition that the second medium (air) has an index of refraction less than the first (plastic) is satisfied, and we can use the equation

$$\theta_c = \sin^{-1}\left(\frac{n_2}{n_1}\right)$$

to find the critical angle θ_c, where $n_2 = 1.00$ and $n_1 = 1.49$.

Solution

Substituting the identified values gives

$$\theta_c = \sin^{-1}\left(\frac{1.00}{1.49}\right) = \sin^{-1}(0.671) = 42.2°.$$

Significance

This result means that any ray of light inside the plastic that strikes the surface at an angle greater than $42.2°$ is totally reflected. This makes the inside surface of the clear plastic a perfect mirror for such rays, without any need for the silvering used on common mirrors. Different combinations of materials have different critical angles, but any combination with $n_1 > n_2$ can produce total internal reflection. The same calculation as made here shows that the critical angle for a ray going from water to air is $48.6°$, whereas that from diamond to air is $24.4°$, and that from flint glass to crown glass is $66.3°$.

30.3 Check Your Understanding At the surface between air and water, light rays can go from air to water and from water to air. For which ray is there no possibility of total internal reflection?

In the photo that opens this chapter, the image of a swimmer underwater is captured by a camera that is also underwater. The swimmer in the upper half of the photograph, apparently facing upward, is, in fact, a reflected image of the swimmer below. The circular ripple near the photograph's center is actually on the water surface. The undisturbed water surrounding it makes a good reflecting surface when viewed from below, thanks to total internal reflection. However, at the very top edge of this photograph, rays from below strike the surface with incident angles less than the critical angle, allowing the camera to capture a view of activities on the pool deck above water.

Fiber Optics: Endoscopes to Telephones

Fiber optics is one application of total internal reflection that is in wide use. In communications, it is used to transmit telephone, internet, and cable TV signals. **Fiber optics** employs the transmission of light down fibers of plastic or glass. Because the fibers are thin, light entering one is likely to strike the inside surface at an angle greater than the critical angle and, thus, be totally reflected (**Figure 30.15**). The index of refraction outside the fiber must be smaller than inside. In fact, most fibers have a varying refractive index to allow more light to be guided along the fiber through total internal refraction. Rays are reflected around corners as shown, making the fibers into tiny light pipes.

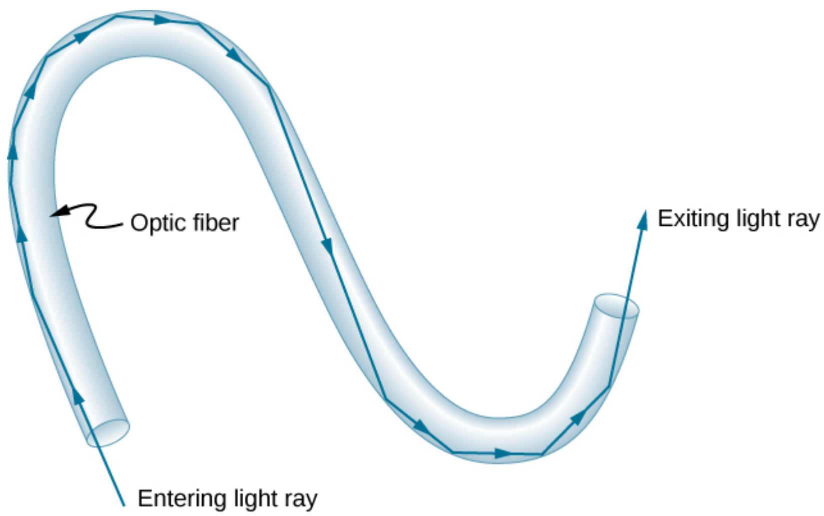


Figure 30.15 Light entering a thin optic fiber may strike the inside surface at large or grazing angles and is completely reflected if these angles exceed the critical angle. Such rays continue down the fiber, even following it around corners, since the angles of reflection and incidence remain large.

Bundles of fibers can be used to transmit an image without a lens, as illustrated in **Figure 30.16**. The output of a device called an endoscope is shown in **Figure 30.16**(b). Endoscopes are used to explore the interior of the body through its natural orifices or minor incisions. Light is transmitted down one fiber bundle to illuminate internal parts, and the reflected light is transmitted back out through another bundle to be observed.

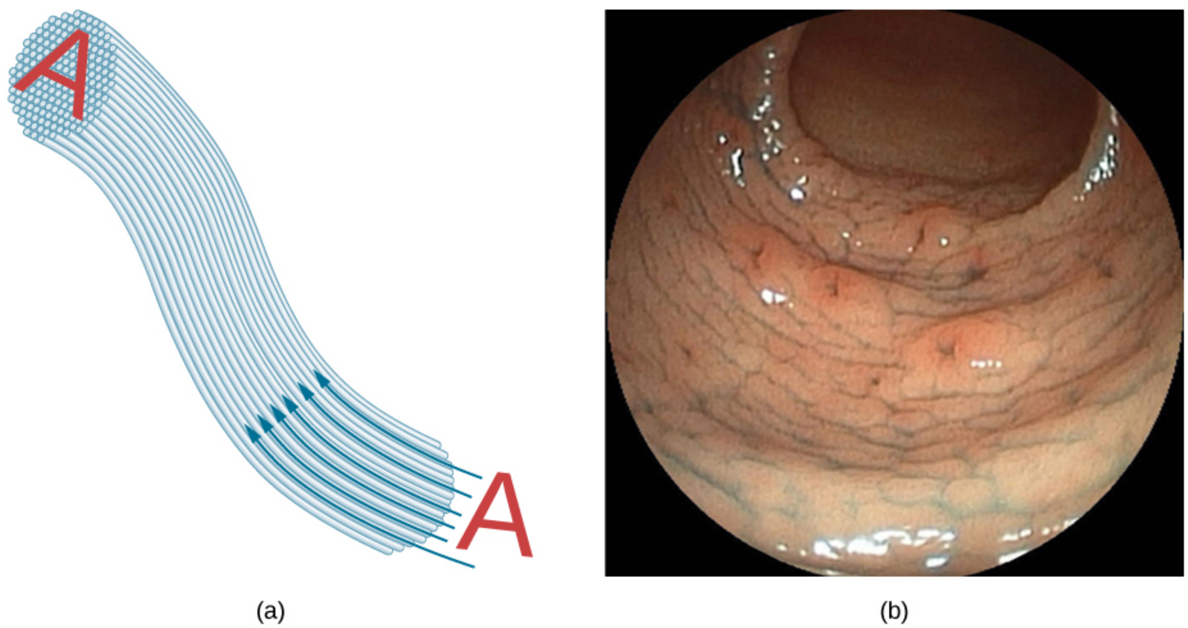


Figure 30.16 (a) An image "A" is transmitted by a bundle of optical fibers. (b) An endoscope is used to probe the body, both transmitting light to the interior and returning an image such as the one shown of a human epiglottis (a structure at the base of the tongue). (credit b: modification of work by "Med_Chaos"/Wikimedia Commons)

Fiber optics has revolutionized surgical techniques and observations within the body, with a host of medical diagnostic and therapeutic uses. Surgery can be performed, such as arthroscopic surgery on a knee or shoulder joint, employing cutting tools attached to and observed with the endoscope. Samples can also be obtained, such as by lassoing an intestinal polyp for external examination. The flexibility of the fiber optic bundle allows doctors to navigate it around small and difficult-to-reach regions in the body, such as the intestines, the heart, blood vessels, and joints. Transmission of an intense laser beam to burn away obstructing plaques in major arteries, as well as delivering light to activate chemotherapy drugs, are becoming commonplace. Optical fibers have in fact enabled microsurgery and remote surgery where the incisions are small and the surgeon's fingers do not need to touch the diseased tissue.

Optical fibers in bundles are surrounded by a cladding material that has a lower index of refraction than the core (**Figure 30.17**). The cladding prevents light from being transmitted between fibers in a bundle. Without cladding, light could pass between fibers in contact, since their indices of refraction are identical. Since no light gets into the cladding (there is total internal reflection back into the core), none can be transmitted between clad fibers that are in contact with one another. Instead, the light is propagated along the length of the fiber, minimizing the loss of signal and ensuring that a quality image is formed at the other end. The cladding and an additional protective layer make optical fibers durable as well as flexible.

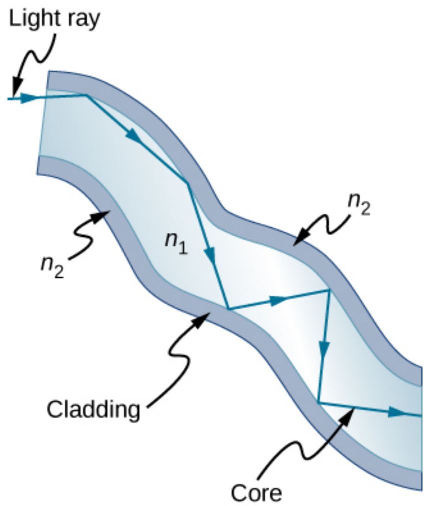


Figure 30.17 Fibers in bundles are clad by a material that has a lower index of refraction than the core to ensure total internal reflection, even when fibers are in contact with one another.

Special tiny lenses that can be attached to the ends of bundles of fibers have been designed and fabricated. Light emerging from a fiber bundle can be focused through such a lens, imaging a tiny spot. In some cases, the spot can be scanned, allowing quality imaging of a region inside the body. Special minute optical filters inserted at the end of the fiber bundle have the capacity to image the interior of organs located tens of microns below the surface without cutting the surface—an area known as nonintrusive diagnostics. This is particularly useful for determining the extent of cancers in the stomach and bowel.

In another type of application, optical fibers are commonly used to carry signals for telephone conversations and internet communications. Extensive optical fiber cables have been placed on the ocean floor and underground to enable optical communications. Optical fiber communication systems offer several advantages over electrical (copper)-based systems, particularly for long distances. The fibers can be made so transparent that light can travel many kilometers before it becomes dim enough to require amplification—much superior to copper conductors. This property of optical fibers is called low loss. Lasers emit light with characteristics that allow far more conversations in one fiber than are possible with electric signals on a single conductor. This property of optical fibers is called high bandwidth. Optical signals in one fiber do not produce undesirable effects in other adjacent fibers. This property of optical fibers is called reduced crosstalk. We shall explore the unique characteristics of laser radiation in a later chapter.

Corner Reflectors and Diamonds

Corner reflectors (**The Law of Reflection**) are perfectly efficient when the conditions for total internal reflection are satisfied. With common materials, it is easy to obtain a critical angle that is less than 45°. One use of these perfect mirrors is in binoculars, as shown in **Figure 30.18**. Another use is in periscopes found in submarines.

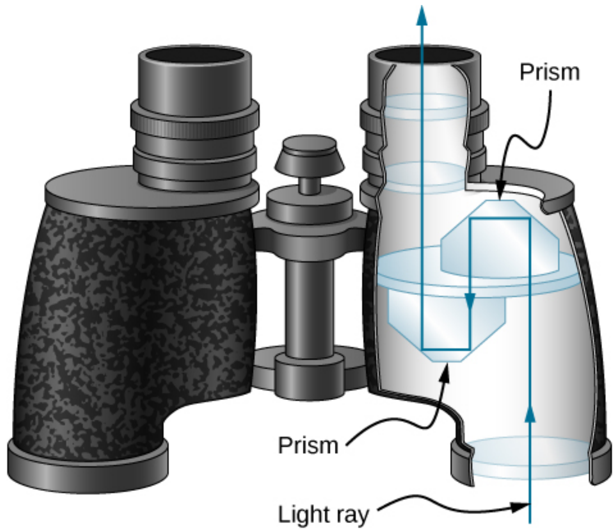


Figure 30.18 These binoculars employ corner reflectors (prisms) with total internal reflection to get light to the observer's eyes.

Total internal reflection, coupled with a large index of refraction, explains why diamonds sparkle more than other materials. The critical angle for a diamond-to-air surface is only 24.4°, so when light enters a diamond, it has trouble getting back out (**Figure 30.19**). Although light freely enters the diamond, it can exit only if it makes an angle less than 24.4°. Facets on diamonds are specifically intended to make this unlikely. Good diamonds are very clear, so that the light makes many internal reflections and is concentrated before exiting—hence the bright sparkle. (Zircon is a natural gemstone that has an exceptionally large index of refraction, but it is not as large as diamond, so it is not as highly prized. Cubic zirconia is manufactured and has an even higher index of refraction (≈ 2.17), but it is still less than that of diamond.) The colors you see emerging from a clear diamond are not due to the diamond's color, which is usually nearly colorless. The colors result from dispersion, which we discuss in **Dispersion**. Colored diamonds get their color from structural defects of the crystal lattice and the inclusion of minute quantities of graphite and other materials. The Argyle Mine in Western Australia produces around 90% of the world's pink, red, champagne, and cognac diamonds, whereas around 50% of the world's clear diamonds come from central and southern Africa.

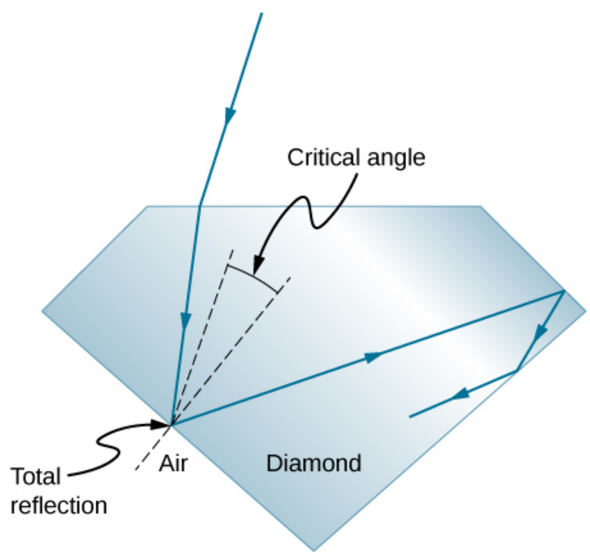


Figure 30.19 Light cannot easily escape a diamond, because its critical angle with air is so small. Most reflections are total, and the facets are placed so that light can exit only in particular ways—thus concentrating the light and making the diamond sparkle brightly.

Explore **refraction and reflection of light (https://openstax.org/l/21bendoflight)** between two media with different indices of refraction. Try to make the refracted ray disappear with total internal reflection. Use the protractor tool to measure the critical angle and compare with the prediction from **Equation 30.5**.

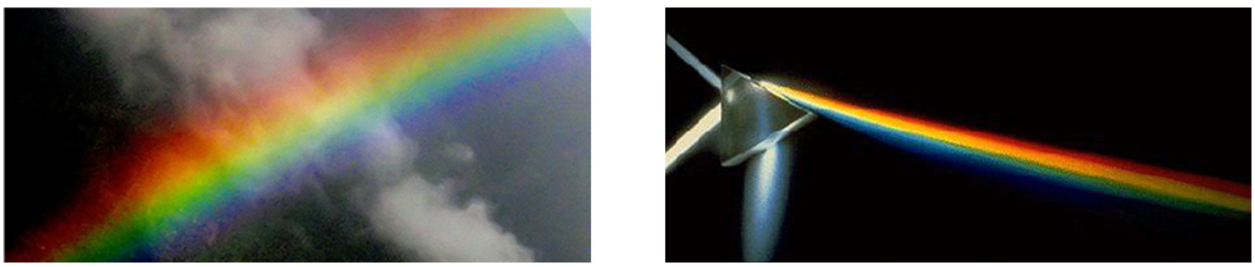
30.5 | Dispersion

Learning Objectives

By the end of this section, you will be able to:

- Explain the cause of dispersion in a prism
- Describe the effects of dispersion in producing rainbows
- Summarize the advantages and disadvantages of dispersion

Everyone enjoys the spectacle of a rainbow glimmering against a dark stormy sky. How does sunlight falling on clear drops of rain get broken into the rainbow of colors we see? The same process causes white light to be broken into colors by a clear glass prism or a diamond (**Figure 30.20**).



(a) (b)

Figure 30.20 The colors of the rainbow (a) and those produced by a prism (b) are identical. (credit a: modification of work by “Alfredo55”/Wikimedia Commons; credit b: modification of work by NASA)

We see about six colors in a rainbow—red, orange, yellow, green, blue, and violet; sometimes indigo is listed, too. These colors are associated with different wavelengths of light, as shown in **Figure 30.21**. When our eye receives pure-wavelength light, we tend to see only one of the six colors, depending on wavelength. The thousands of other hues we can sense in other situations are our eye’s response to various mixtures of wavelengths. White light, in particular, is a fairly uniform mixture of all visible wavelengths. Sunlight, considered to be white, actually appears to be a bit yellow, because of its mixture of wavelengths, but it does contain all visible wavelengths. The sequence of colors in rainbows is the same sequence as the colors shown in the figure. This implies that white light is spread out in a rainbow according to wavelength. **Dispersion** is defined as the spreading of white light into its full spectrum of wavelengths. More technically, dispersion occurs whenever the propagation of light depends on wavelength.

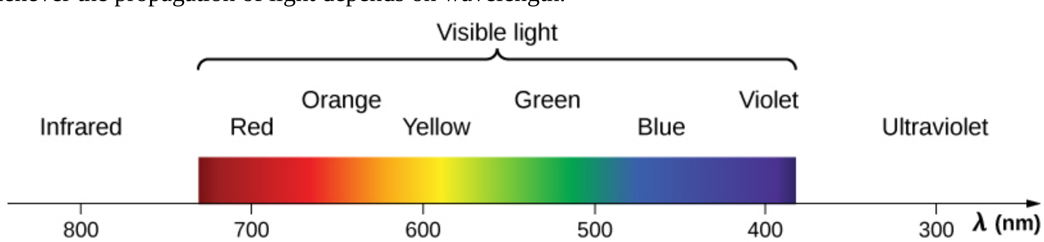


Figure 30.21 Even though rainbows are associated with six colors, the rainbow is a continuous distribution of colors according to wavelengths.

Any type of wave can exhibit dispersion. For example, sound waves, all types of electromagnetic waves, and water waves can be dispersed according to wavelength. Dispersion may require special circumstances and can result in spectacular displays such as in the production of a rainbow. This is also true for sound, since all frequencies ordinarily travel at the same speed. If you listen to sound through a long tube, such as a vacuum cleaner hose, you can easily hear it dispersed by interaction with the tube. Dispersion, in fact, can reveal a great deal about what the wave has encountered that disperses its wavelengths. The dispersion of electromagnetic radiation from outer space, for example, has revealed much about what exists between the stars—the so-called interstellar medium.

Nick Moore's **video (https://openstax.org/l/21nickmoorevid)** discusses dispersion of a pulse as he taps a long spring. Follow his explanation as Moore replays the high-speed footage showing high frequency waves outrunning the lower frequency waves.

Refraction is responsible for dispersion in rainbows and many other situations. The angle of refraction depends on the index of refraction, as we know from Snell's law. We know that the index of refraction n depends on the medium. But for a given medium, n also depends on wavelength (**Table 30.2**). Note that for a given medium, n increases as wavelength decreases and is greatest for violet light. Thus, violet light is bent more than red light, as shown for a prism in **Figure 30.22**(b). White light is dispersed into the same sequence of wavelengths as seen in **Figure 30.20** and **Figure 30.21**.

Medium	Red (660 nm)	Orange (610 nm)	Yellow (580 nm)	Green (550 nm)	Blue (470 nm)	Violet (410 nm)
Water	1.331	1.332	1.333	1.335	1.338	1.342
Diamond	2.410	2.415	2.417	2.426	2.444	2.458
Glass, crown	1.512	1.514	1.518	1.519	1.524	1.530
Glass, flint	1.662	1.665	1.667	1.674	1.684	1.698
Polystyrene	1.488	1.490	1.492	1.493	1.499	1.506
Quartz, fused	1.455	1.456	1.458	1.459	1.462	1.468

Table 30.2 Index of Refraction *n* in Selected Media at Various Wavelengths

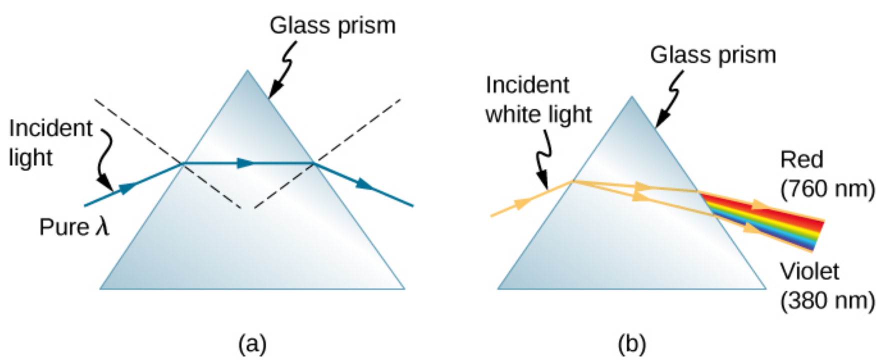
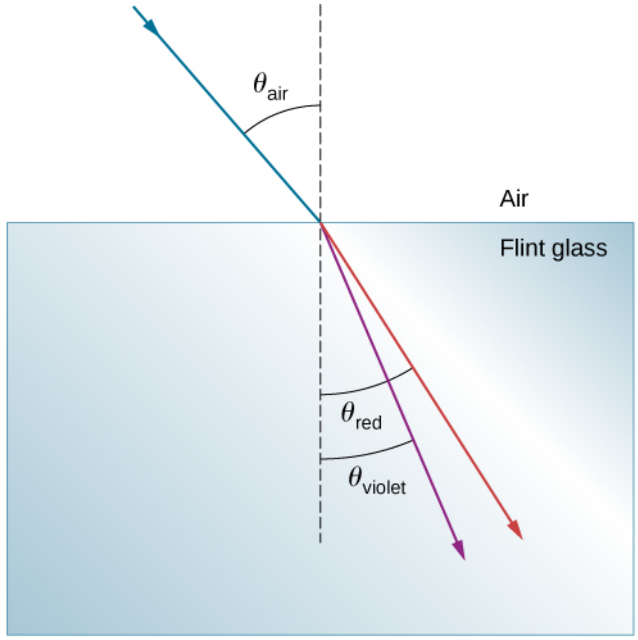


Figure 30.22 (a) A pure wavelength of light falls onto a prism and is refracted at both surfaces. (b) White light is dispersed by the prism (shown exaggerated). Since the index of refraction varies with wavelength, the angles of refraction vary with wavelength. A sequence of red to violet is produced, because the index of refraction increases steadily with decreasing wavelength.

Example 30.5

Dispersion of White Light by Flint Glass

A beam of white light goes from air into flint glass at an incidence angle of $43.2°$. What is the angle between the red (660 nm) and violet (410 nm) parts of the refracted light?



Strategy

Values for the indices of refraction for flint glass at various wavelengths are listed in **Table 30.2**. Use these values for calculate the angle of refraction for each color and then take the difference to find the dispersion angle.

Solution

Applying the law of refraction for the red part of the beam

$$n_{\text{air}} \sin\theta_{\text{air}} = n_{\text{red}} \sin\theta_{\text{red}},$$

we can solve for the angle of refraction as

$$\theta_{\text{red}} = \sin^{-1}\left(\frac{n_{\text{air}} \sin\theta_{\text{air}}}{n_{\text{red}}}\right) = \sin^{-1}\left[\frac{(1.000)\sin 43.2°}{(1.662)}\right] = 27.0°.$$

Similarly, the angle of incidence for the violet part of the beam is

$$\theta_{\text{violet}} = \sin^{-1}\left(\frac{n_{\text{air}} \sin\theta_{\text{air}}}{n_{\text{violet}}}\right) = \sin^{-1}\left[\frac{(1.000)\sin 43.2°}{(1.698)}\right] = 26.4°.$$

The difference between these two angles is

$$\theta_{\text{red}} - \theta_{\text{violet}} = 27.0° - 26.4° = 0.6°.$$

Significance

Although $0.6°$ may seem like a negligibly small angle, if this beam is allowed to propagate a long enough distance, the dispersion of colors becomes quite noticeable.

30.4 Check Your Understanding In the preceding example, how much distance inside the block of flint glass would the red and the violet rays have to progress before they are separated by 1.0 mm?

Rainbows are produced by a combination of refraction and reflection. You may have noticed that you see a rainbow only when you look away from the Sun. Light enters a drop of water and is reflected from the back of the drop (**Figure 30.23**). The light is refracted both as it enters and as it leaves the drop. Since the index of refraction of water varies with wavelength, the light is dispersed, and a rainbow is observed (**Figure 30.24**(a)). (No dispersion occurs at the back surface, because the law of reflection does not depend on wavelength.) The actual rainbow of colors seen by an observer depends on the myriad rays being refracted and reflected toward the observer's eyes from numerous drops of water. The effect is most spectacular when the background is dark, as in stormy weather, but can also be observed in waterfalls and lawn sprinklers. The arc of a rainbow comes from the need to be looking at a specific angle relative to the direction of the Sun, as illustrated in part (b). If two reflections of light occur within the water drop, another "secondary" rainbow is produced. This rare event produces an arc that lies above the primary rainbow arc, as in part (c), and produces colors in the reverse order of the primary rainbow, with red at the lowest angle and violet at the largest angle.

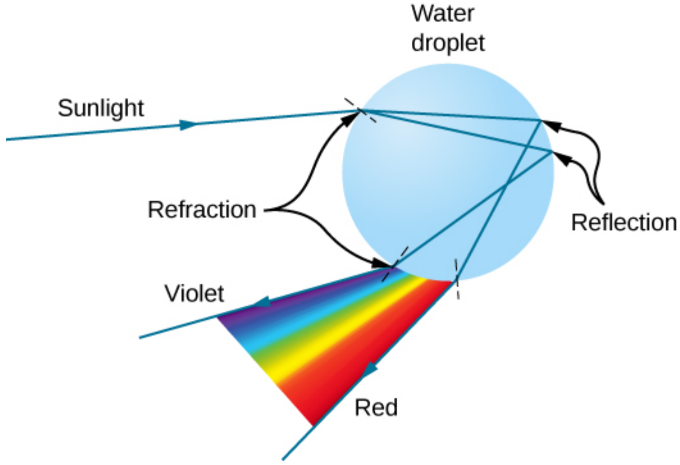


Figure 30.23 A ray of light falling on this water drop enters and is reflected from the back of the drop. This light is refracted and dispersed both as it enters and as it leaves the drop.

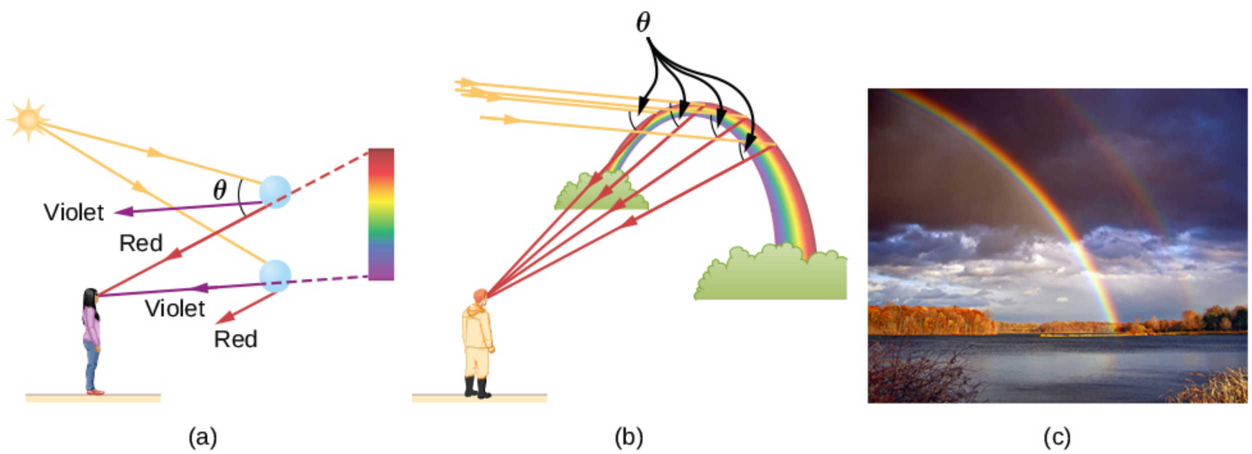


Figure 30.24 (a) Different colors emerge in different directions, and so you must look at different locations to see the various colors of a rainbow. (b) The arc of a rainbow results from the fact that a line between the observer and any point on the arc must make the correct angle with the parallel rays of sunlight for the observer to receive the refracted rays. (c) Double rainbow. (credit c: modification of work by "Nicholas"/Wikimedia Commons)

Dispersion may produce beautiful rainbows, but it can cause problems in optical systems. White light used to transmit messages in a fiber is dispersed, spreading out in time and eventually overlapping with other messages. Since a laser produces a nearly pure wavelength, its light experiences little dispersion, an advantage over white light for transmission of information. In contrast, dispersion of electromagnetic waves coming to us from outer space can be used to determine the amount of matter they pass through.

CHAPTER 30 REVIEW

KEY TERMS

corner reflector object consisting of two (or three) mutually perpendicular reflecting surfaces, so that the light that enters is reflected back exactly parallel to the direction from which it came

critical angle incident angle that produces an angle of refraction of $90°$

dispersion spreading of light into its spectrum of wavelengths

fiber optics field of study of the transmission of light down fibers of plastic or glass, applying the principle of total internal reflection

geometric optics part of optics dealing with the ray aspect of light

index of refraction for a material, the ratio of the speed of light in a vacuum to that in a material

law of reflection angle of reflection equals the angle of incidence

law of refraction when a light ray crosses from one medium to another, it changes direction by an amount that depends on the index of refraction of each medium and the sines of the angle of incidence and angle of refraction

ray straight line that originates at some point

refraction changing of a light ray's direction when it passes through variations in matter

total internal reflection phenomenon at the boundary between two media such that all the light is reflected and no refraction occurs

SUMMARY

30.1 The Propagation of Light

- The speed of light in a vacuum is $c = 2.99792458 \times 10^8 \text{ m/s} \approx 3.00 \times 10^8 \text{ m/s}$.
- The index of refraction of a material is $n = c/v$, where v is the speed of light in a material and c is the speed of light in a vacuum.
- The ray model of light describes the path of light as straight lines. The part of optics dealing with the ray aspect of light is called geometric optics.
- Light can travel in three ways from a source to another location: (1) directly from the source through empty space; (2) through various media; and (3) after being reflected from a mirror.

30.2 The Law of Reflection

- When a light ray strikes a smooth surface, the angle of reflection equals the angle of incidence.
- A mirror has a smooth surface and reflects light at specific angles.
- Light is diffused when it reflects from a rough surface.

30.3 Refraction

- The change of a light ray's direction when it passes through variations in matter is called refraction.
- The law of refraction, also called Snell's law, relates the indices of refraction for two media at an interface to the change in angle of a light ray passing through that interface.

30.4 Total Internal Reflection

- The incident angle that produces an angle of refraction of $90°$ is called the critical angle.
- Total internal reflection is a phenomenon that occurs at the boundary between two media, such that if the incident

angle in the first medium is greater than the critical angle, then all the light is reflected back into that medium.

- Fiber optics involves the transmission of light down fibers of plastic or glass, applying the principle of total internal reflection.
- Cladding prevents light from being transmitted between fibers in a bundle.
- Diamonds sparkle due to total internal reflection coupled with a large index of refraction.

30.5 Dispersion

- The spreading of white light into its full spectrum of wavelengths is called dispersion.
- Rainbows are produced by a combination of refraction and reflection, and involve the dispersion of sunlight into a continuous distribution of colors.
- Dispersion produces beautiful rainbows but also causes problems in certain optical systems.

CONCEPTUAL QUESTIONS

30.1 The Propagation of Light

1. Under what conditions can light be modeled like a ray? Like a wave?

2. Why is the index of refraction always greater than or equal to 1?

3. Does the fact that the light flash from lightning reaches you before its sound prove that the speed of light is extremely large or simply that it is greater than the speed of sound? Discuss how you could use this effect to get an estimate of the speed of light.

4. Speculate as to what physical process might be responsible for light traveling more slowly in a medium than in a vacuum.

30.2 The Law of Reflection

5. Using the law of reflection, explain how powder takes the shine off of a person's nose. What is the name of the optical effect?

30.3 Refraction

6. Diffusion by reflection from a rough surface is described in this chapter. Light can also be diffused by refraction. Describe how this occurs in a specific situation, such as light interacting with crushed ice.

7. Will light change direction toward or away from the perpendicular when it goes from air to water? Water to glass? Glass to air?

8. Explain why an object in water always appears to be at a depth shallower than it actually is?

9. Explain why a person's legs appear very short when wading in a pool. Justify your explanation with a ray diagram showing the path of rays from the feet to the eye of an observer who is out of the water.

10. Explain why an oar that is partially submerged in water appears bent.

30.4 Total Internal Reflection

11. A ring with a colorless gemstone is dropped into water. The gemstone becomes invisible when submerged. Can it be a diamond? Explain.

12. The most common type of mirage is an illusion that light from faraway objects is reflected by a pool of water that is not really there. Mirages are generally observed in deserts, when there is a hot layer of air near the ground. Given that the refractive index of air is lower for air at higher temperatures, explain how mirages can be formed.

13. How can you use total internal reflection to estimate the index of refraction of a medium?

30.5 Dispersion

14. Is it possible that total internal reflection plays a role in rainbows? Explain in terms of indices of refraction and angles, perhaps referring to that shown below. Some of us have seen the formation of a double rainbow; is it physically possible to observe a triple rainbow?

(credit: "Chad"/Flickr)

15. A high-quality diamond may be quite clear and colorless, transmitting all visible wavelengths with little absorption. Explain how it can sparkle with flashes of brilliant color when illuminated by white light.

PROBLEMS

30.1 The Propagation of Light

16. What is the speed of light in water? In glycerine?

17. What is the speed of light in air? In crown glass?

18. Calculate the index of refraction for a medium in which the speed of light is 2.012×10^8 m/s, and identify the most likely substance based on **Table 30.1**.

19. In what substance in **Table 30.1** is the speed of light 2.290×10^8 m/s?

20. There was a major collision of an asteroid with the Moon in medieval times. It was described by monks at Canterbury Cathedral in England as a red glow on and around the Moon. How long after the asteroid hit the Moon, which is 3.84×10^5 km away, would the light first arrive on Earth?

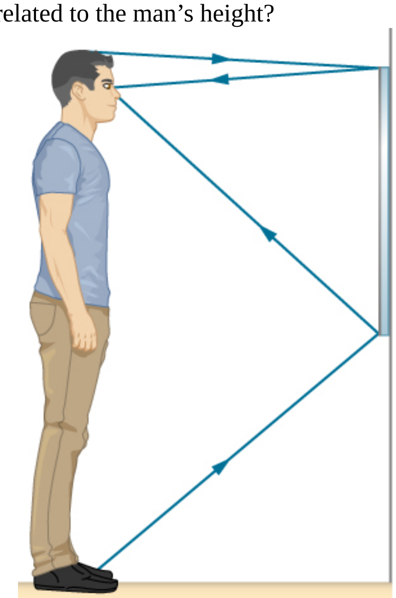
21. Components of some computers communicate with each other through optical fibers having an index of refraction $n = 1.55$. What time in nanoseconds is required for a signal to travel 0.200 m through such a fiber?

22. Compare the time it takes for light to travel 1000 m on the surface of Earth and in outer space.

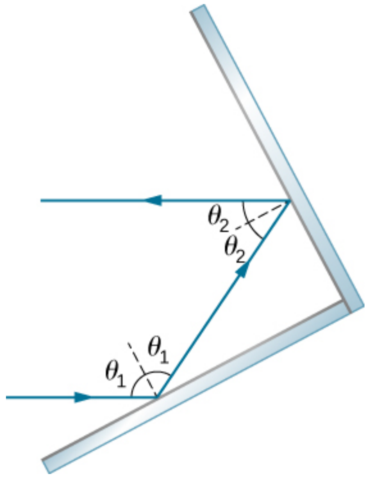
23. How far does light travel underwater during a time interval of 1.50×10^{-6} s?

30.2 The Law of Reflection

24. Suppose a man stands in front of a mirror as shown below. His eyes are 1.65 m above the floor and the top of his head is 0.13 m higher. Find the height above the floor of the top and bottom of the smallest mirror in which he can see both the top of his head and his feet. How is this distance related to the man's height?

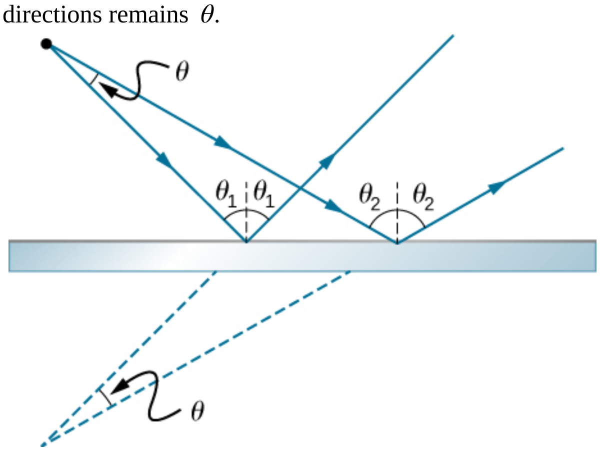


25. Show that when light reflects from two mirrors that meet each other at a right angle, the outgoing ray is parallel to the incoming ray, as illustrated below.



26. On the Moon's surface, lunar astronauts placed a corner reflector, off which a laser beam is periodically reflected. The distance to the Moon is calculated from the round-trip time. What percent correction is needed to account for the delay in time due to the slowing of light in Earth's atmosphere? Assume the distance to the Moon is precisely 3.84×10^8 m and Earth's atmosphere (which varies in density with altitude) is equivalent to a layer 30.0 km thick with a constant index of refraction $n = 1.000293$.

27. A flat mirror is neither converging nor diverging. To prove this, consider two rays originating from the same point and diverging at an angle θ (see below). Show that after striking a plane mirror, the angle between their directions remains θ.



30.3 Refraction

Unless otherwise specified, for problems 1 through 10, the indices of refraction of glass and water should be taken to be 1.50 and 1.333, respectively.

28. A light beam in air has an angle of incidence of $35°$ at the surface of a glass plate. What are the angles of reflection and refraction?

29. A light beam in air is incident on the surface of a pond, making an angle of $20°$ with respect to the surface. What are the angles of reflection and refraction?

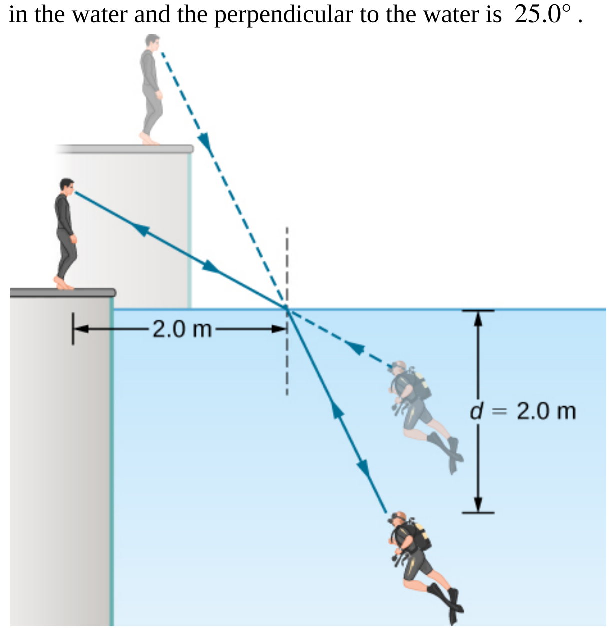
30. When a light ray crosses from water into glass, it emerges at an angle of $30°$ with respect to the normal of the interface. What is its angle of incidence?

31. A pencil flashlight submerged in water sends a light beam toward the surface at an angle of incidence of $30°$. What is the angle of refraction in air?

32. Light rays from the Sun make a $30°$ angle to the vertical when seen from below the surface of a body of water. At what angle above the horizon is the Sun?

33. The path of a light beam in air goes from an angle of incidence of $35°$ to an angle of refraction of $22°$ when it enters a rectangular block of plastic. What is the index of refraction of the plastic?

34. A scuba diver training in a pool looks at his instructor as shown below. What angle does the ray from the instructor's face make with the perpendicular to the water at the point where the ray enters? The angle between the ray in the water and the perpendicular to the water is $25.0°$.



35. (a) Using information in the preceding problem, find the height of the instructor's head above the water, noting that you will first have to calculate the angle of incidence. (b) Find the apparent depth of the diver's head below water as seen by the instructor.

30.4 Total Internal Reflection

36. Verify that the critical angle for light going from water to air is $48.6°$, as discussed at the end of **Example 30.4**, regarding the critical angle for light traveling in a polystyrene (a type of plastic) pipe surrounded by air.

37. (a) At the end of **Example 30.4**, it was stated that the critical angle for light going from diamond to air is $24.4°$. Verify this. (b) What is the critical angle for light going from zircon to air?

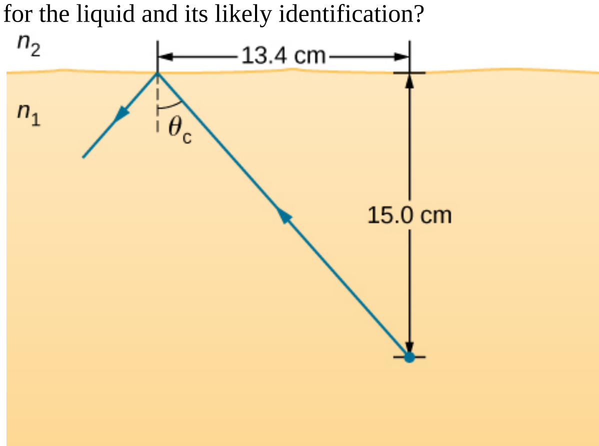
38. An optical fiber uses flint glass clad with crown glass. What is the critical angle?

39. At what minimum angle will you get total internal reflection of light traveling in water and reflected from ice?

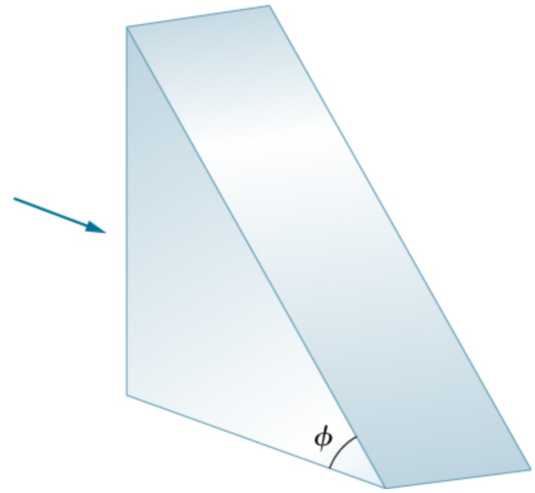
40. Suppose you are using total internal reflection to make an efficient corner reflector. If there is air outside and the incident angle is $45.0°$, what must be the minimum index of refraction of the material from which the reflector is made?

41. You can determine the index of refraction of a substance by determining its critical angle. (a) What is the index of refraction of a substance that has a critical angle of $68.4°$ when submerged in water? What is the substance, based on **Table 30.1**? (b) What would the critical angle be for this substance in air?

42. A ray of light, emitted beneath the surface of an unknown liquid with air above it, undergoes total internal reflection as shown below. What is the index of refraction for the liquid and its likely identification?



43. Light rays fall normally on the vertical surface of the glass prism $(n = 1.50)$ shown below. (a) What is the largest value for ϕ such that the ray is totally reflected at the slanted face? (b) Repeat the calculation of part (a) if the prism is immersed in water.



30.5 Dispersion

44. (a) What is the ratio of the speed of red light to violet light in diamond, based on **Table 30.2**? (b) What is this ratio in polystyrene? (c) Which is more dispersive?

45. A beam of white light goes from air into water at an incident angle of $75.0°$. At what angles are the red (660 nm) and violet (410 nm) parts of the light refracted?

46. By how much do the critical angles for red (660 nm) and violet (410 nm) light differ in a diamond surrounded by air?

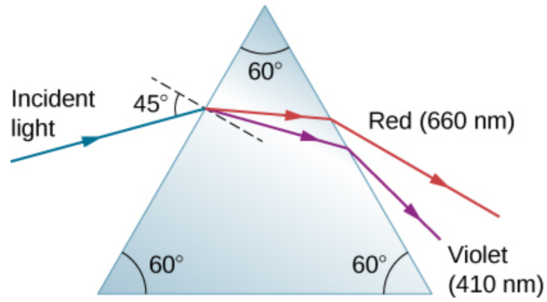
47. (a) A narrow beam of light containing yellow (580 nm) and green (550 nm) wavelengths goes from polystyrene to air, striking the surface at a $30.0°$ incident angle. What is the angle between the colors when they emerge? (b) How far would they have to travel to be separated by 1.00 mm?

48. A parallel beam of light containing orange (610 nm) and violet (410 nm) wavelengths goes from fused quartz to water, striking the surface between them at a $60.0°$ incident angle. What is the angle between the two colors in water?

49. A ray of 610-nm light goes from air into fused quartz at an incident angle of $55.0°$. At what incident angle must 470 nm light enter flint glass to have the same angle of refraction?

50. A narrow beam of light containing red (660 nm) and blue (470 nm) wavelengths travels from air through a 1.00-cm-thick flat piece of crown glass and back to air again. The beam strikes at a $30.0°$ incident angle. (a) At what angles do the two colors emerge? (b) By what distance are the red and blue separated when they emerge?

51. A narrow beam of white light enters a prism made of crown glass at a 45.0° incident angle, as shown below. At what angles, θ_R and θ_V, do the red (660 nm) and violet (410 nm) components of the light emerge from the prism?



31 | IMAGE FORMATION

Figure 31.1 *Cloud Gate* is a public sculpture by Anish Kapoor located in Millennium Park in Chicago. Its stainless steel plates reflect and distort images around it, including the Chicago skyline. Dedicated in 2006, it has become a popular tourist attraction, illustrating how art can use the principles of physical optics to startle and entertain. (credit: modification of work by Dhilung Kirat)

Chapter Outline

Introduction

This chapter introduces the major ideas of geometric optics, which describe the formation of images due to reflection and refraction. It is called "geometric" optics because the images can be characterized using geometric constructions, such as ray diagrams. We have seen that visible light is an electromagnetic wave; however, its wave nature becomes evident only when light interacts with objects with dimensions comparable to the wavelength (about 500 nm for visible light). Therefore, the laws of geometric optics only apply to light interacting with objects much larger than the wavelength of the light.

31.1 | Images Formed by Plane Mirrors

Learning Objectives

By the end of this section, you will be able to:

- Describe how an image is formed by a plane mirror.
- Distinguish between real and virtual images.
- Find the location and characterize the orientation of an image created by a plane mirror.

You only have to look as far as the nearest bathroom to find an example of an image formed by a mirror. Images in a **plane mirror** are the same size as the object, are located behind the mirror, and are oriented in the same direction as the object (i.e., "upright").

To understand how this happens, consider **Figure 31.2**. Two rays emerge from point *P*, strike the mirror, and reflect into the observer's eye. Note that we use the law of reflection to construct the reflected rays. If the reflected rays are extended backward behind the mirror (see dashed lines in **Figure 31.2**), they seem to originate from point *Q*. This is where the image of point *P* is located. If we repeat this process for point P', we obtain its image at point Q'. You should convince yourself by using basic geometry that the image height (the distance from *Q* to Q') is the same as the object height (the distance from *P* to P'). By forming images of all points of the object, we obtain an upright image of the object behind the mirror.

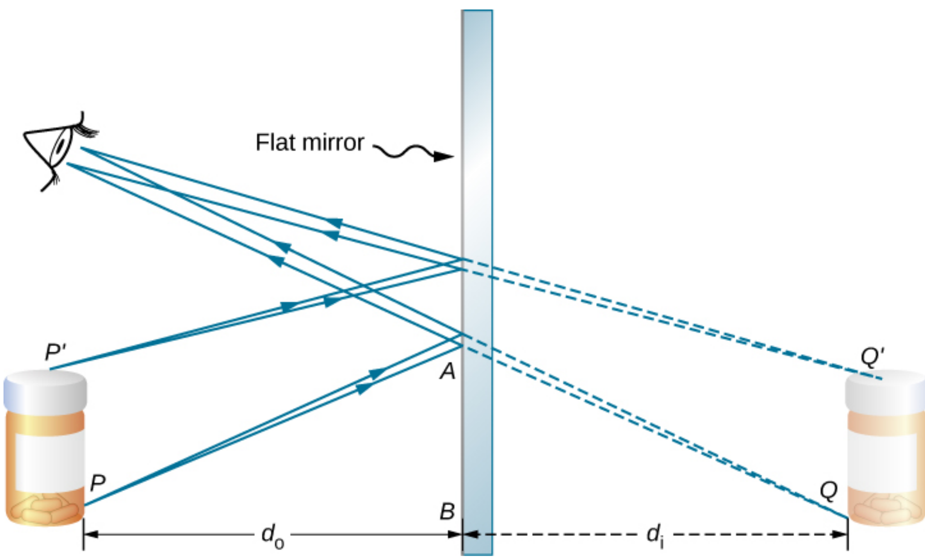


Figure 31.2 Two light rays originating from point *P* on an object are reflected by a flat mirror into the eye of an observer. The reflected rays are obtained by using the law of reflection. Extending these reflected rays backward, they seem to come from point *Q* behind the mirror, which is where the virtual image is located. Repeating this process for point P' gives the image point Q'. The image height is thus the same as the object height, the image is upright, and the object distance d_o is the same as the image distance d_i. (credit: modification of work by Kevin Dufendach)

Notice that the reflected rays appear to the observer to come directly from the image behind the mirror. In reality, these rays come from the points on the mirror where they are reflected. The image behind the mirror is called a **virtual image** because it cannot be projected onto a screen—the rays only appear to originate from a common point behind the mirror. If you walk behind the mirror, you cannot see the image, because the rays do not go there. However, in front of the mirror, the rays behave exactly as if they come from behind the mirror, so that is where the virtual image is located.

Later in this chapter, we discuss real images; a **real image** can be projected onto a screen because the rays physically go through the image. You can certainly see both real and virtual images. The difference is that a virtual image cannot be projected onto a screen, whereas a real image can.

Locating an Image in a Plane Mirror

The law of reflection tells us that the angle of incidence is the same as the angle of reflection. Applying this to triangles *PAB* and *QAB* in **Figure 31.2** and using basic geometry shows that they are congruent triangles. This means that the distance *PB* from the object to the mirror is the same as the distance *BQ* from the mirror to the image. The **object distance** (denoted d_o) is the distance from the mirror to the object (or, more generally, from the center of the optical element that creates its image). Similarly, the **image distance** (denoted d_i) is the distance from the mirror to the image (or, more generally, from the center of the optical element that creates it). If we measure distances from the mirror, then the object and image are in opposite directions, so for a plane mirror, the object and image distances should have the opposite signs:

$$d_o = -d_i. \quad (31.1)$$

An extended object such as the container in **Figure 31.2** can be treated as a collection of points, and we can apply the method above to locate the image of each point on the extended object, thus forming the extended image.

Multiple Images

If an object is situated in front of two mirrors, you may see images in both mirrors. In addition, the image in the first mirror may act as an object for the second mirror, so the second mirror may form an image of the image. If the mirrors are placed parallel to each other and the object is placed at a point other than the midpoint between them, then this process of image-of-an-image continues without end, as you may have noticed when standing in a hallway with mirrors on each side. This is shown in **Figure 31.3**, which shows three images produced by the blue object. Notice that each reflection reverses front and back, just like pulling a right-hand glove inside out produces a left-hand glove (this is why a reflection of your right hand is a left hand). Thus, the fronts and backs of images 1 and 2 are both inverted with respect to the object, and the front and back of image 3 is inverted with respect to image 2, which is the object for image 3.

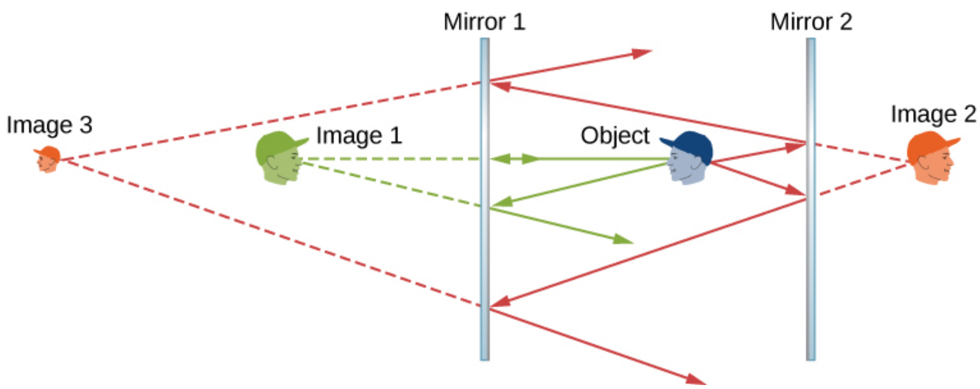


Figure 31.3 Two parallel mirrors can produce, in theory, an infinite number of images of an object placed off center between the mirrors. Three of these images are shown here. The front and back of each image is inverted with respect to its object. Note that the colors are only to identify the images. For normal mirrors, the color of an image is essentially the same as that of its object.

You may have noticed that image 3 is smaller than the object, whereas images 1 and 2 are the same size as the object. The ratio of the image height with respect to the object height is called **magnification**. More will be said about magnification in the next section.

Infinite reflections may terminate. For instance, two mirrors at right angles form three images, as shown in part (a) of **Figure 31.4**. Images 1 and 2 result from rays that reflect from only a single mirror, but image 1,2 is formed by rays that reflect from both mirrors. This is shown in the ray-tracing diagram in part (b) of **Figure 31.4**. To find image 1,2, you have to look behind the corner of the two mirrors.

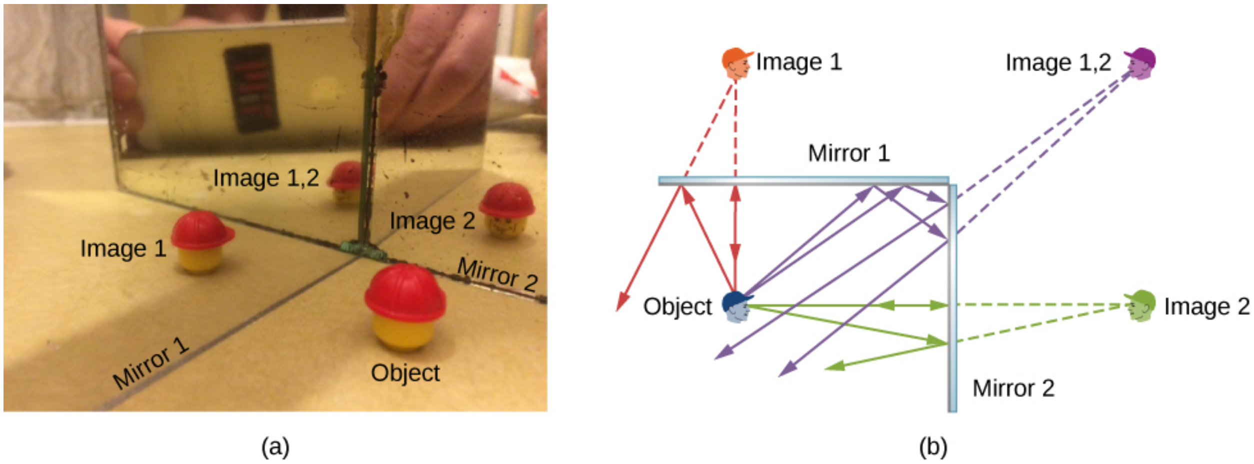


Figure 31.4 Two mirrors can produce multiple images. (a) Three images of a plastic head are visible in the two mirrors at a right angle. (b) A single object reflecting from two mirrors at a right angle can produce three images, as shown by the green, purple, and red images.

31.2 | Spherical Mirrors

Learning Objectives

By the end of this section, you will be able to:

- Describe image formation by spherical mirrors.
- Use ray diagrams and the mirror equation to calculate the properties of an image in a spherical mirror.

The image in a plane mirror has the same size as the object, is upright, and is the same distance behind the mirror as the object is in front of the mirror. A **curved mirror**, on the other hand, can form images that may be larger or smaller than the object and may form either in front of the mirror or behind it. In general, any curved surface will form an image, although some images may be so distorted as to be unrecognizable (think of fun house mirrors).

Because curved mirrors can create such a rich variety of images, they are used in many optical devices that find many uses. We will concentrate on spherical mirrors for the most part, because they are easier to manufacture than mirrors such as parabolic mirrors and so are more common.

Curved Mirrors

We can define two general types of spherical mirrors. If the reflecting surface is the outer side of the sphere, the mirror is called a **convex mirror**. If the inside surface is the reflecting surface, it is called a **concave mirror**.

Symmetry is one of the major hallmarks of many optical devices, including mirrors and lenses. The symmetry axis of such optical elements is often called the principal axis or **optical axis**. For a spherical mirror, the optical axis passes through the mirror's center of curvature and the mirror's **vertex**, as shown in **Figure 31.5**.

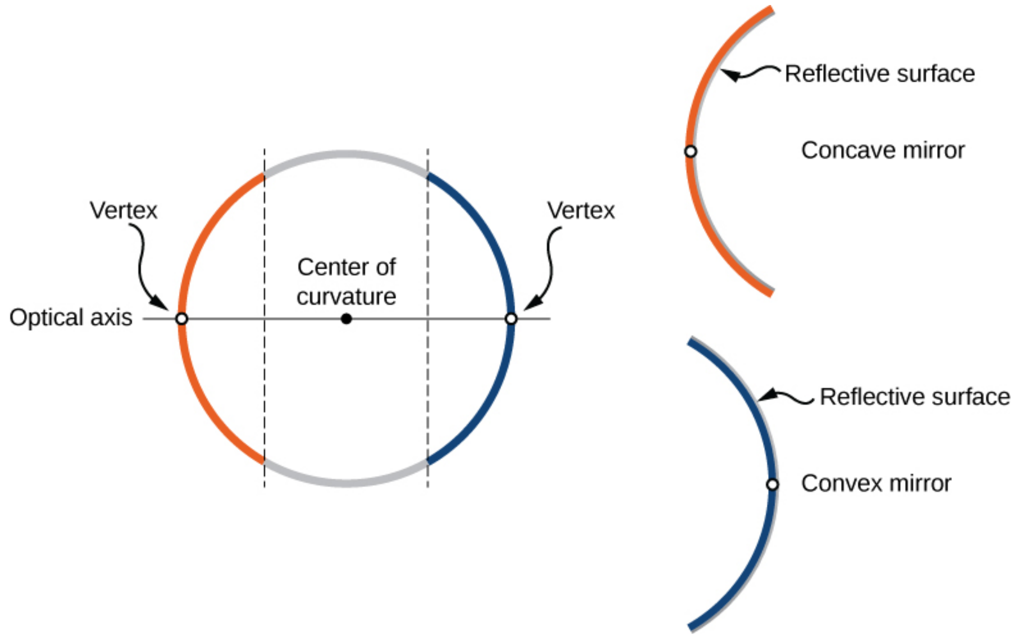


Figure 31.5 A spherical mirror is formed by cutting out a piece of a sphere and silvering either the inside or outside surface. A concave mirror has silvering on the interior surface (think “cave”), and a convex mirror has silvering on the exterior surface.

Consider rays that are parallel to the optical axis of a parabolic mirror, as shown in part (a) of **Figure 31.6**. Following the law of reflection, these rays are reflected so that they converge at a point, called the **focal point**. Part (b) of this figure shows a spherical mirror that is large compared with its radius of curvature. For this mirror, the reflected rays do not cross at the same point, so the mirror does not have a well-defined focal point. This is called spherical aberration and results in a blurred image of an extended object. Part (c) shows a spherical mirror that is small compared to its radius of curvature. This mirror is a good approximation of a parabolic mirror, so rays that arrive parallel to the optical axis are reflected to a well-defined focal point. The distance along the optical axis from the mirror to the focal point is called the **focal length** of the mirror.

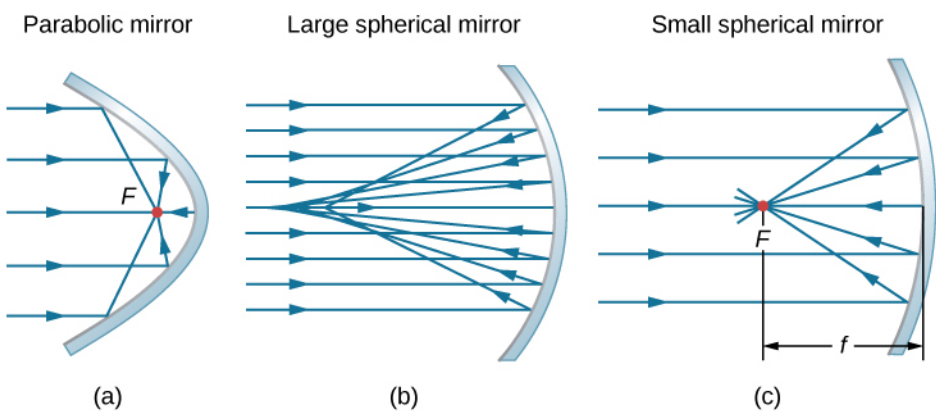
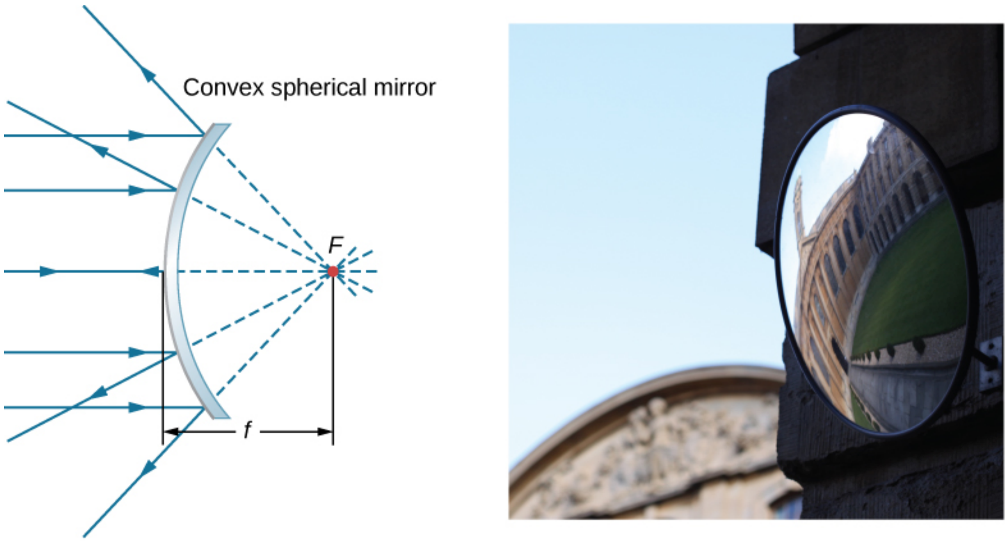


Figure 31.6 (a) Parallel rays reflected from a parabolic mirror cross at a single point called the focal point *F*. (b) Parallel rays reflected from a large spherical mirror do not cross at a common point. (c) If a spherical mirror is small compared with its radius of curvature, it better approximates the central part of a parabolic mirror, so parallel rays essentially cross at a common point. The distance along the optical axis from the mirror to the focal point is the focal length *f* of the mirror.

A convex spherical mirror also has a focal point, as shown in **Figure 31.7**. Incident rays parallel to the optical axis are reflected from the mirror and seem to originate from point *F* at focal length *f* behind the mirror. Thus, the focal point is virtual because no real rays actually pass through it; they only appear to originate from it.



(a) (b)

Figure 31.7 (a) Rays reflected by a convex spherical mirror: Incident rays of light parallel to the optical axis are reflected from a convex spherical mirror and seem to originate from a well-defined focal point at focal distance *f* on the opposite side of the mirror. The focal point is virtual because no real rays pass through it. (b) Photograph of a virtual image formed by a convex mirror. (credit b: modification of work by Jenny Downing)

How does the focal length of a mirror relate to the mirror's radius of curvature? **Figure 31.8** shows a single ray that is reflected by a spherical concave mirror. The incident ray is parallel to the optical axis. The point at which the reflected ray crosses the optical axis is the focal point. Note that all incident rays that are parallel to the optical axis are reflected through the focal point—we only show one ray for simplicity. We want to find how the focal length *FP* (denoted by *f*) relates to the radius of curvature of the mirror, *R*, whose length is $R = CF + FP$. The law of reflection tells us that angles *OXC* and *CXF* are the same, and because the incident ray is parallel to the optical axis, angles *OXC* and *XCP* are also the same. Thus, triangle *CXF* is an isosceles triangle with $CF = FX$. If the angle θ is small (so that $\sin\theta \approx \theta$; this is called the "small-angle approximation"), then $FX \approx FP$ or $CF \approx FP$. Inserting this into the equation for the radius *R*, we get

$$R = CF + FP = FP + FP = 2FP = 2f$$

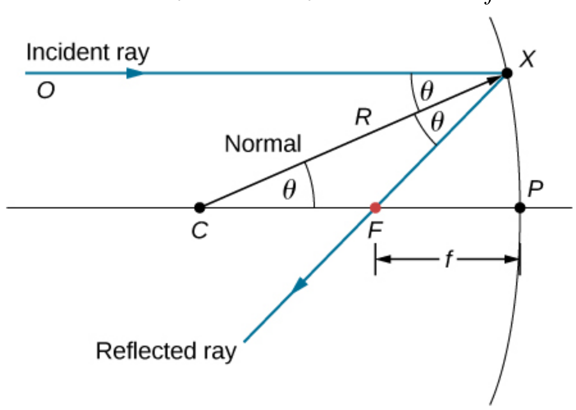


Figure 31.8 Reflection in a concave mirror. In the small-angle approximation, a ray that is parallel to the optical axis *CP* is reflected through the focal point *F* of the mirror.

In other words, in the small-angle approximation, the focal length *f* of a concave spherical mirror is half of its radius of curvature, *R*:

$$f = \frac{R}{2}. \quad (31.2)$$

In this chapter, we assume that the **small-angle approximation** (also called the paraxial approximation) is always valid. In this approximation, all rays are paraxial rays, which means that they make a small angle with the optical axis and are at a distance much less than the radius of curvature from the optical axis. In this case, their angles θ of reflection are small angles, so $\sin\theta \approx \tan\theta \approx \theta$.

Using Ray Tracing to Locate Images

To find the location of an image formed by a spherical mirror, we first use ray tracing, which is the technique of drawing rays and using the law of reflection to determine the reflected rays (later, for lenses, we use the law of refraction to determine refracted rays). Combined with some basic geometry, we can use ray tracing to find the focal point, the image location, and other information about how a mirror manipulates light. In fact, we already used ray tracing above to locate the focal point of spherical mirrors, or the image distance of flat mirrors. To locate the image of an object, you must locate at least two points of the image. Locating each point requires drawing at least two rays from a point on the object and constructing their reflected rays. The point at which the reflected rays intersect, either in real space or in virtual space, is where the corresponding point of the image is located. To make ray tracing easier, we concentrate on four "principal" rays whose reflections are easy to construct.

Figure 31.9 shows a concave mirror and a convex mirror, each with an arrow-shaped object in front of it. These are the objects whose images we want to locate by ray tracing. To do so, we draw rays from point *Q* that is on the object but not on the optical axis. We choose to draw our ray from the tip of the object. Principal ray 1 goes from point *Q* and travels parallel to the optical axis. The reflection of this ray must pass through the focal point, as discussed above. Thus, for the concave mirror, the reflection of principal ray 1 goes through focal point *F*, as shown in part (b) of the figure. For the convex mirror, the backward extension of the reflection of principal ray 1 goes through the focal point (i.e., a virtual focus). Principal ray 2 travels first on the line going through the focal point and then is reflected back along a line parallel to the optical axis. Principal ray 3 travels toward the center of curvature of the mirror, so it strikes the mirror at normal incidence and is reflected back along the line from which it came. Finally, principal ray 4 strikes the vertex of the mirror and is reflected symmetrically about the optical axis.

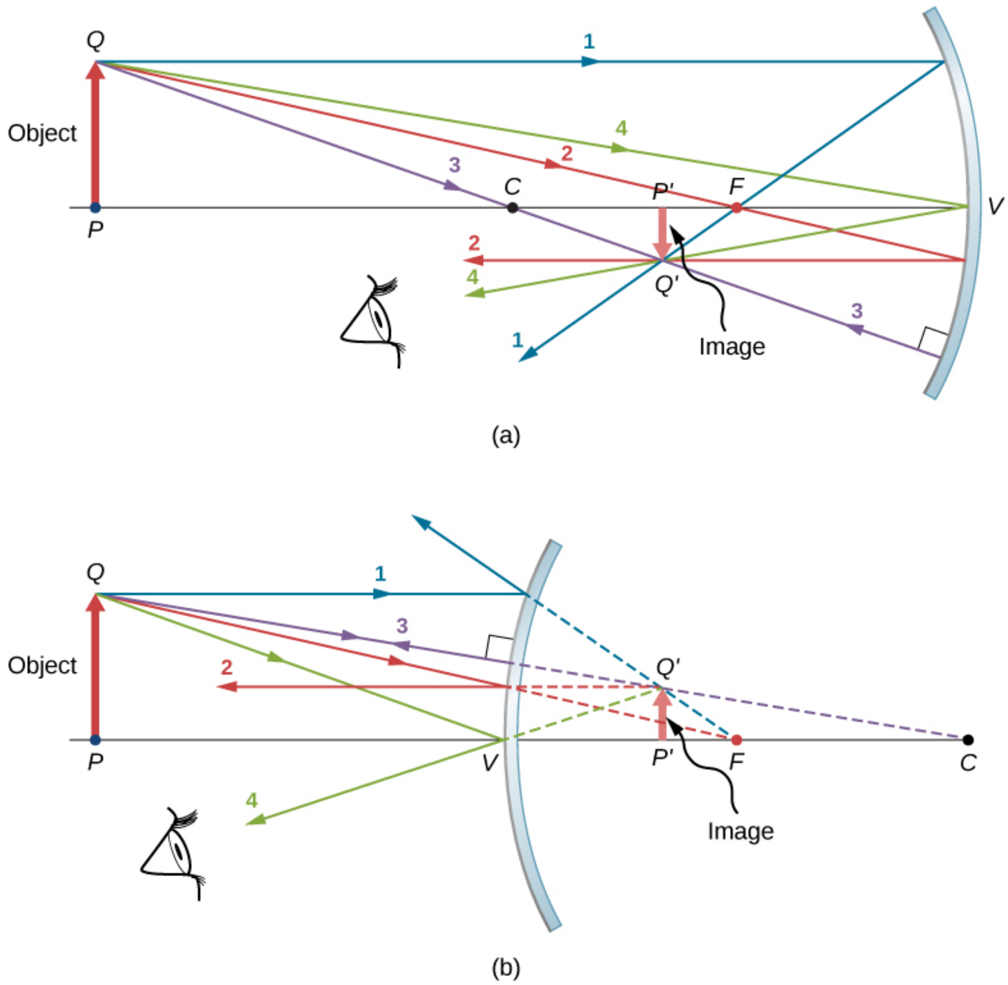


Figure 31.9 The four principal rays shown for both (a) a concave mirror and (b) a convex mirror. The image forms where the rays intersect (for real images) or where their backward extensions intersect (for virtual images).

The four principal rays intersect at point Q', which is where the image of point *Q* is located. To locate point Q', drawing any two of these principle rays would suffice. We are thus free to choose whichever of the principal rays we desire to locate the image. Drawing more than two principal rays is sometimes useful to verify that the ray tracing is correct.

To completely locate the extended image, we need to locate a second point in the image, so that we know how the image is oriented. To do this, we trace the principal rays from the base of the object. In this case, all four principal rays run along the optical axis, reflect from the mirror, and then run back along the optical axis. The difficulty is that, because these rays are collinear, we cannot determine a unique point where they intersect. All we know is that the base of the image is on the optical axis. However, because the mirror is symmetrical from top to bottom, it does not change the vertical orientation of the object. Thus, because the object is vertical, the image must be vertical. Therefore, the image of the base of the object is on the optical axis directly above the image of the tip, as drawn in the figure.

For the concave mirror, the extended image in this case forms between the focal point and the center of curvature of the mirror. It is inverted with respect to the object, is a real image, and is smaller than the object. Were we to move the object closer to or farther from the mirror, the characteristics of the image would change. For example, we show, as a later exercise, that an object placed between a concave mirror and its focal point leads to a virtual image that is upright and larger than the object. For the convex mirror, the extended image forms between the focal point and the mirror. It is upright with respect to the object, is a virtual image, and is smaller than the object.

Summary of Ray-Tracing Rules

Ray tracing is very useful for mirrors. The rules for ray tracing are summarized here for reference:

- A ray travelling parallel to the optical axis of a spherical mirror is reflected along a line that goes through the focal point of the mirror (ray 1 in **Figure 31.9**).

- A ray travelling along a line that goes through the focal point of a spherical mirror is reflected along a line parallel to the optical axis of the mirror (ray 2 in **Figure 31.9**).
- A ray travelling along a line that goes through the center of curvature of a spherical mirror is reflected back along the same line (ray 3 in **Figure 31.9**).
- A ray that strikes the vertex of a spherical mirror is reflected symmetrically about the optical axis of the mirror (ray 4 in **Figure 31.9**).

We use ray tracing to illustrate how images are formed by mirrors and to obtain numerical information about optical properties of the mirror. If we assume that a mirror is small compared with its radius of curvature, we can also use algebra and geometry to derive a mirror equation, which we do in the next section. Combining ray tracing with the mirror equation is a good way to analyze mirror systems.

Image Formation by Reflection—The Mirror Equation

For a plane mirror, we showed that the image formed has the same height and orientation as the object, and it is located at the same distance behind the mirror as the object is in front of the mirror. Although the situation is a bit more complicated for curved mirrors, using geometry leads to simple formulas relating the object and image distances to the focal lengths of concave and convex mirrors.

Consider the object *OP* shown in **Figure 31.10**. The center of curvature of the mirror is labeled *C* and is a distance *R* from the vertex of the mirror, as marked in the figure. The object and image distances are labeled d_o and d_i, and the object and image heights are labeled h_o and h_i, respectively. Because the angles ϕ and ϕ' are alternate interior angles, we know that they have the same magnitude. However, they must differ in sign if we measure angles from the optical axis, so $\phi = -\phi'$. An analogous scenario holds for the angles θ and θ'. The law of reflection tells us that they have the same magnitude, but their signs must differ if we measure angles from the optical axis. Thus, $\theta = -\theta'$. Taking the tangent of the angles θ and θ', and using the property that $\tan(-\theta) = -\tan\theta$, gives us

$$\left.\begin{aligned} \tan\theta &= \frac{h_o}{d_o} \\ \tan\theta' &= -\tan\theta = \frac{h_i}{d_i} \end{aligned}\right\} \frac{h_o}{d_o} = -\frac{h_i}{d_i} \text{ or } -\frac{h_o}{h_i} = \frac{d_o}{d_i}. \quad (31.3)$$

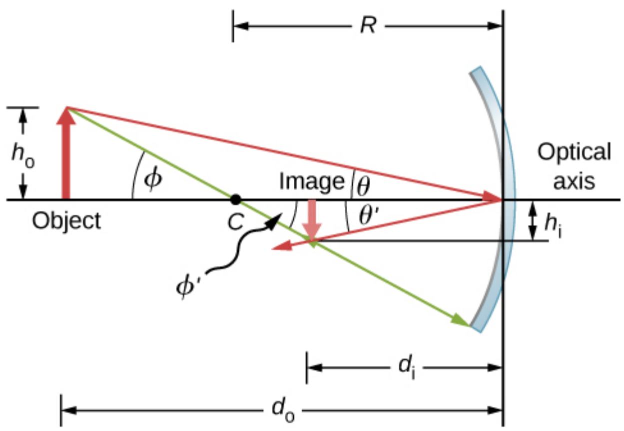


Figure 31.10 Image formed by a concave mirror.

Similarly, taking the tangent of ϕ and ϕ' gives

$$\left.\begin{aligned} \tan\phi &= \frac{h_o}{d_o - R} \\ \tan\phi' &= -\tan\phi = \frac{h_i}{R - d_i} \end{aligned}\right\} \frac{h_o}{d_o - R} = -\frac{h_i}{R - d_i} \text{ or } -\frac{h_o}{h_i} = \frac{d_o - R}{R - d_i}.$$

Combining these two results gives

$$\frac{d_o}{d_i} = \frac{d_o - R}{R - d_i}.$$

After a little algebra, this becomes

$$\frac{1}{d_o} + \frac{1}{d_i} = \frac{2}{R}. \quad (31.4)$$

No approximation is required for this result, so it is exact. However, as discussed above, in the small-angle approximation, the focal length of a spherical mirror is one-half the radius of curvature of the mirror, or $f = R/2$. Inserting this into **Equation 31.3** gives the *mirror equation*:

$$\frac{1}{d_o} + \frac{1}{d_i} = \frac{1}{f}. \quad (31.5)$$

The mirror equation relates the image and object distances to the focal distance and is valid only in the small-angle approximation. Although it was derived for a concave mirror, it also holds for convex mirrors (proving this is left as an exercise). We can extend the mirror equation to the case of a plane mirror by noting that a plane mirror has an infinite radius of curvature. This means the focal point is at infinity, so the mirror equation simplifies to

$$d_o = -d_i \quad (31.6)$$

which is the same as **Equation 31.1** obtained earlier.

Notice that we have been very careful with the signs in deriving the mirror equation. For a plane mirror, the image distance has the opposite sign of the object distance. Also, the real image formed by the concave mirror in **Figure 31.10** is on the opposite side of the optical axis with respect to the object. In this case, the image height should have the opposite sign of the object height. To keep track of the signs of the various quantities in the mirror equation, we now introduce a sign convention.

Sign convention for spherical mirrors

Using a consistent sign convention is very important in geometric optics. It assigns positive or negative values for the quantities that characterize an optical system. Understanding the sign convention allows you to describe an image without constructing a ray diagram. This text uses the following sign convention:

1. The focal length f is positive for concave mirrors and negative for convex mirrors.
2. The image distance d_i is positive for real images and negative for virtual images.

Notice that rule 1 means that the radius of curvature of a spherical mirror can be positive or negative. What does it mean to have a negative radius of curvature? This means simply that the radius of curvature for a convex mirror is defined to be negative.

Image magnification

Let's use the sign convention to further interpret the derivation of the mirror equation. In deriving this equation, we found that the object and image heights are related by

$$-\frac{h_o}{h_i} = \frac{d_o}{d_i}. \quad (31.7)$$

See **Equation 31.3**. Both the object and the image formed by the mirror in **Figure 31.10** are real, so the object and image distances are both positive. The highest point of the object is above the optical axis, so the object height is positive. The image, however, is below the optical axis, so the image height is negative. Thus, this sign convention is consistent with our derivation of the mirror equation.

Equation 31.7 in fact describes the **linear magnification** (often simply called "magnification") of the image in terms of the object and image distances. We thus define the dimensionless magnification m as follows:

$$m = \frac{h_i}{h_o}. \quad (31.8)$$

If m is positive, the image is upright, and if m is negative, the image is inverted. If $|m| > 1$, the image is larger than the object, and if $|m| < 1$, the image is smaller than the object. With this definition of magnification, we get the following relation between the vertical and horizontal object and image distances:

$$m = \frac{h_i}{h_o} = -\frac{d_i}{d_o}. \quad (31.9)$$

This is a very useful relation because it lets you obtain the magnification of the image from the object and image distances, which you can obtain from the mirror equation.

Example 31.1

Solar Electric Generating System

One of the solar technologies used today for generating electricity involves a device (called a parabolic trough or concentrating collector) that concentrates sunlight onto a blackened pipe that contains a fluid. This heated fluid is pumped to a heat exchanger, where the thermal energy is transferred to another system that is used to generate steam and eventually generates electricity through a conventional steam cycle. **Figure 31.11** shows such a working system in southern California. The real mirror is a parabolic cylinder with its focus located at the pipe; however, we can approximate the mirror as exactly one-quarter of a circular cylinder.

Figure 31.11 Parabolic trough collectors are used to generate electricity in southern California. (credit: “kjkolb”/Wikimedia Commons)

a. If we want the rays from the sun to focus at 40.0 cm from the mirror, what is the radius of the mirror?
b. What is the amount of sunlight concentrated onto the pipe, per meter of pipe length, assuming the insolation (incident solar radiation) is 900 W/m^2 ?
c. If the fluid-carrying pipe has a 2.00-cm diameter, what is the temperature increase of the fluid per meter of pipe over a period of 1 minute? Assume that all solar radiation incident on the reflector is absorbed by the pipe, and that the fluid is mineral oil.

Strategy

First identify the physical principles involved. Part (a) is related to the optics of spherical mirrors. Part (b) involves a little math, primarily geometry. Part (c) requires an understanding of heat and density.

Solution

a. The sun is the object, so the object distance is essentially infinity: $d_o = \infty$. The desired image distance is $d_i = 40.0\,\text{cm}$. We use the mirror equation to find the focal length of the mirror:

$$\begin{aligned}\frac{1}{d_o}+\frac{1}{d_i} &= \frac{1}{f}\\ f &= \left(\frac{1}{d_o}+\frac{1}{d_i}\right)^{-1}\\ &= \left(\frac{1}{\infty}+\frac{1}{40.0\text{ cm}}\right)^{-1}\\ &= 40.0\text{ cm}\end{aligned}$$

Thus, the radius of the mirror is $R = 2f = 80.0\text{ cm}$.

b. The insolation is 900 W/m^2. You must find the cross-sectional area A of the concave mirror, since the power delivered is 900 $\text{W/m}^2 \times A$. The mirror in this case is estimated as a quarter-section of a cylinder, so the area for a length L of the mirror is $A = \frac{1}{4}(2\pi R)L$. The area for a length of 1.00 m is then

$$A = \frac{\pi}{2}R(1.00\text{ m}) = \frac{(3.14)}{2}(0.800\text{ m})(1.00\text{ m}) = 1.26\text{ m}^2.$$

The insolation on the 1.00-m length of pipe is then

$$\left(9.00\times10^2\,\frac{\text{W}}{\text{m}^2}\right)\left(1.26\text{ m}^2\right) = 1130\text{ W}.$$

c. The increase in temperature is given by $Q = mc\Delta T$. The mass m of the mineral oil in the one-meter section of pipe is

$$\begin{aligned}m &= \rho V = \rho\pi\left(\frac{d}{2}\right)^2(1.00\text{ m})\\ &= \left(8.00\times10^2\text{ kg/m}^3\right)(3.14)(0.0100\text{ m})^2(1.00\text{ m})\\ &= 0.251\text{ kg}\end{aligned}$$

Therefore, the increase in temperature in one minute is

$$\begin{aligned}\Delta T &= Q/mc\\ &= \frac{(1130\text{ W})(60.0\text{ s})}{(0.251\text{ kg})(1670\text{ J}\cdot\text{kg/°C})}\\ &= 162\text{°C}\end{aligned}$$

Significance

An array of such pipes in the California desert can provide a thermal output of 250 MW on a sunny day, with fluids reaching temperatures as high as 400°C. We are considering only one meter of pipe here and ignoring heat losses along the pipe.

Example 31.2

Image in a Convex Mirror

A keratometer is a device used to measure the curvature of the cornea of the eye, particularly for fitting contact lenses. Light is reflected from the cornea, which acts like a convex mirror, and the keratometer measures the magnification of the image. The smaller the magnification, the smaller the radius of curvature of the cornea. If the light source is 12 cm from the cornea and the image magnification is 0.032, what is the radius of curvature of the cornea?

Strategy

If you find the focal length of the convex mirror formed by the cornea, then you know its radius of curvature (it's twice the focal length). The object distance is $d_o = 12\text{ cm}$ and the magnification is $m = 0.032$. First find the image distance d_i and then solve for the focal length *f*.

Solution

Start with the equation for magnification, $m = -d_i/d_o$. Solving for d_i and inserting the given values yields

$$d_i = -md_o = -(0.032)(12\text{ cm}) = -0.384\text{ cm}$$

where we retained an extra significant figure because this is an intermediate step in the calculation. Solve the mirror equation for the focal length *f* and insert the known values for the object and image distances. The result is

$$\begin{aligned} \frac{1}{d_o} + \frac{1}{d_i} &= \frac{1}{f} \\ f &= \left(\frac{1}{d_o} + \frac{1}{d_i}\right)^{-1} \\ &= \left(\frac{1}{12\text{ cm}} + \frac{1}{-0.384\text{ cm}}\right)^{-1} \\ &= -0.40\text{ cm} \end{aligned}$$

The radius of curvature is twice the focal length, so

$$R = 2f = -0.80\text{ cm}$$

Significance

The focal length is negative, so the focus is virtual, as expected for a concave mirror and a real object. The radius of curvature found here is reasonable for a cornea. The distance from cornea to retina in an adult eye is about 2.0 cm. In practice, corneas may not be spherical, which complicates the job of fitting contact lenses. Note that the image distance here is negative, consistent with the fact that the image is behind the mirror. Thus, the image is virtual because no rays actually pass through it. In the problems and exercises, you will show that, for a fixed object distance, a smaller radius of curvature corresponds to a smaller the magnification.

Problem-Solving Strategy: Spherical Mirrors

Step 1. First make sure that image formation by a spherical mirror is involved.

Step 2. Determine whether ray tracing, the mirror equation, or both are required. A sketch is very useful even if ray tracing is not specifically required by the problem. Write symbols and known values on the sketch.

Step 3. Identify exactly what needs to be determined in the problem (identify the unknowns).

Step 4. Make a list of what is given or can be inferred from the problem as stated (identify the knowns).

Step 5. If ray tracing is required, use the ray-tracing rules listed near the beginning of this section.

Step 6. Most quantitative problems require using the mirror equation. Use the examples as guides for using the mirror equation.

Step 7. Check to see whether the answer makes sense. Do the signs of object distance, image distance, and focal length correspond with what is expected from ray tracing? Is the sign of the magnification correct? Are the object and image distances reasonable?

Departure from the Small-Angle Approximation

The small-angle approximation is a cornerstone of the above discussion of image formation by a spherical mirror. When this approximation is violated, then the image created by a spherical mirror becomes distorted. Such distortion is called **aberration**. Here we briefly discuss two specific types of aberrations: spherical aberration and coma.

Spherical aberration

Consider a broad beam of parallel rays impinging on a spherical mirror, as shown in **Figure 31.12**.

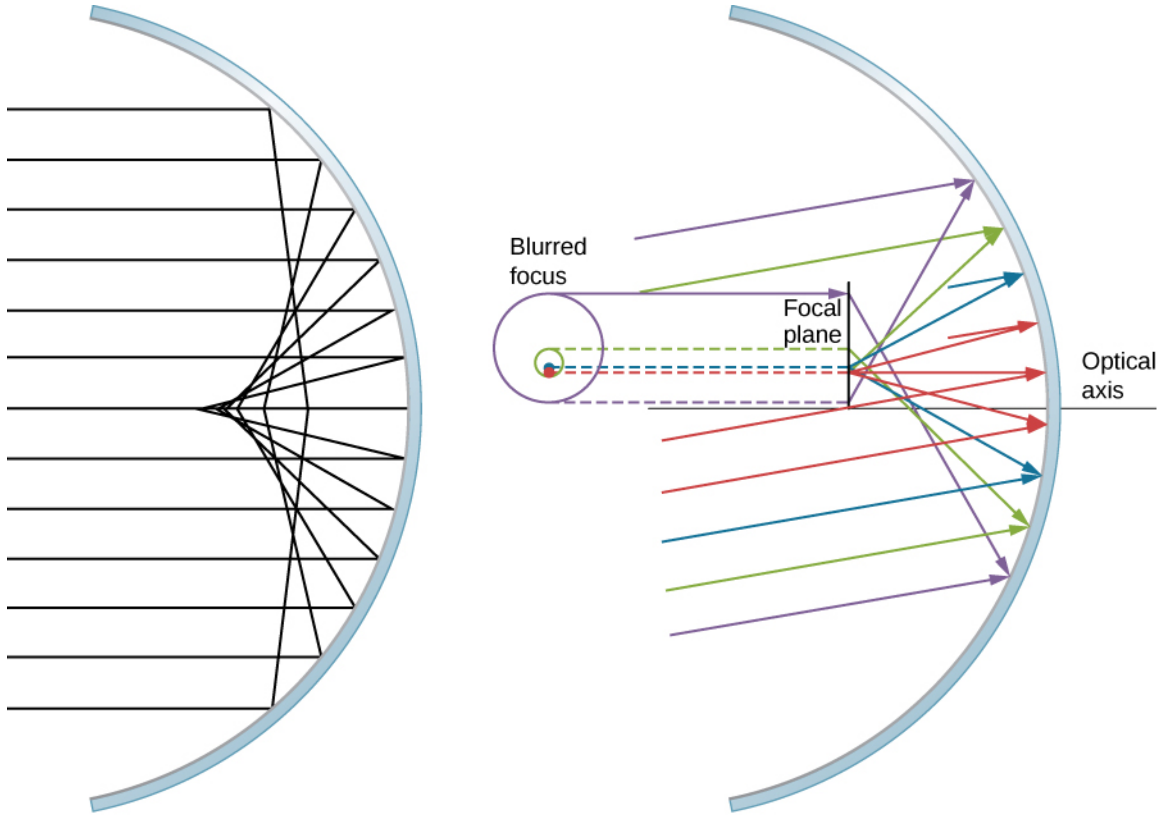


Figure 31.12 (a) With spherical aberration, the rays that are farther from the optical axis and the rays that are closer to the optical axis are focused at different points. Notice that the aberration gets worse for rays farther from the optical axis. (b) For comatic aberration, parallel rays that are not parallel to the optical axis are focused at different heights and at different focal lengths, so the image contains a "tail" like a comet (which is "coma" in Latin). Note that the colored rays are only to facilitate viewing; the colors do not indicate the color of the light.

The farther from the optical axis the rays strike, the worse the spherical mirror approximates a parabolic mirror. Thus, these rays are not focused at the same point as rays that are near the optical axis, as shown in the figure. Because of **spherical aberration**, the image of an extended object in a spherical mirror will be blurred. Spherical aberrations are characteristic of the mirrors and lenses that we consider in the following section of this chapter (more sophisticated mirrors and lenses are needed to eliminate spherical aberrations).

Coma or comatic aberration

Coma is similar to spherical aberration, but arises when the incoming rays are not parallel to the optical axis, as shown in part (b) of **Figure 31.12**. Recall that the small-angle approximation holds for spherical mirrors that are small compared to their radius. In this case, spherical mirrors are good approximations of parabolic mirrors. Parabolic mirrors focus all rays that are parallel to the optical axis at the focal point. However, parallel rays that are *not* parallel to the optical axis are focused at different heights and at different focal lengths, as show in part (b) of **Figure 31.12**. Because a spherical mirror is symmetric about the optical axis, the various colored rays in this figure create circles of the corresponding color on the focal plane.

Although a spherical mirror is shown in part (b) of **Figure 31.12**, comatic aberration occurs also for parabolic mirrors—it does not result from a breakdown in the small-angle approximation. Spherical aberration, however, occurs only for spherical mirrors and is a result of a breakdown in the small-angle approximation. We will discuss both coma and spherical aberration later in this chapter, in connection with telescopes.

31.3 | Images Formed by Refraction

Learning Objectives

By the end of this section, you will be able to:

- Describe image formation by a single refracting surface
- Determine the location of an image and calculate its properties by using a ray diagram
- Determine the location of an image and calculate its properties by using the equation for a single refracting surface

When rays of light propagate from one medium to another, these rays undergo refraction, which is when light waves are bent at the interface between two media. The refracting surface can form an image in a similar fashion to a reflecting surface, except that the law of refraction (Snell's law) is at the heart of the process instead of the law of reflection.

Refraction at a Plane Interface—Apparent Depth

If you look at a straight rod partially submerged in water, it appears to bend at the surface (**Figure 31.13**). The reason behind this curious effect is that the image of the rod inside the water forms a little closer to the surface than the actual position of the rod, so it does not line up with the part of the rod that is above the water. The same phenomenon explains why a fish in water appears to be closer to the surface than it actually is.

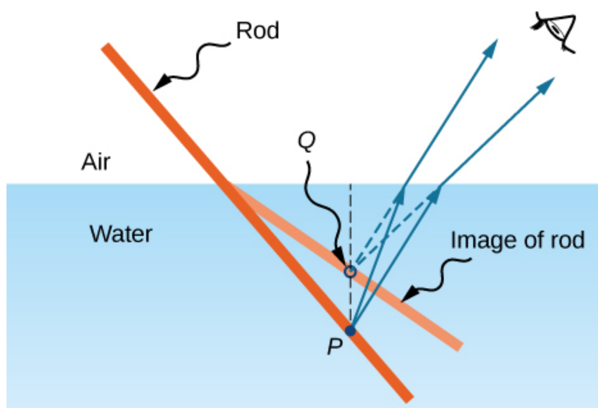


Figure 31.13 Bending of a rod at a water-air interface. Point *P* on the rod appears to be at point *Q*, which is where the image of point *P* forms due to refraction at the air-water interface.

To study image formation as a result of refraction, consider the following questions:

1. What happens to the rays of light when they enter or pass through a different medium?
2. Do the refracted rays originating from a single point meet at some point or diverge away from each other?

To be concrete, we consider a simple system consisting of two media separated by a plane interface (**Figure 31.14**). The object is in one medium and the observer is in the other. For instance, when you look at a fish from above the water surface, the fish is in medium 1 (the water) with refractive index 1.33, and your eye is in medium 2 (the air) with refractive index 1.00, and the surface of the water is the interface. The depth that you "see" is the image height h_i and is called the **apparent depth**. The actual depth of the fish is the object height h_o.

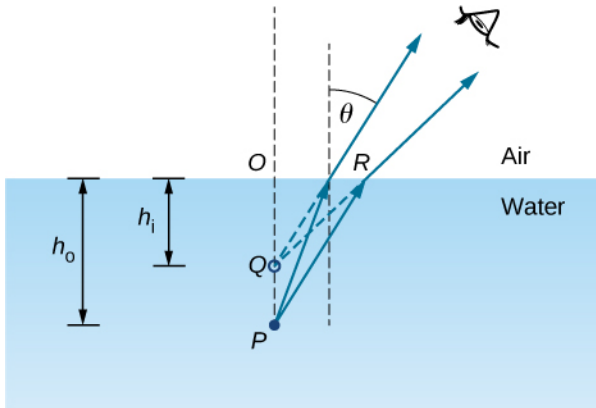


Figure 31.14 Apparent depth due to refraction. The real object at point *P* creates an image at point *Q*. The image is not at the same depth as the object, so the observer sees the image at an "apparent depth."

The apparent depth h_i depends on the angle at which you view the image. For a view from above (the so-called "normal" view), we can approximate the refraction angle θ to be small, and replace sin θ in Snell's law by tan θ. With this approximation, you can use the triangles ΔOPR and ΔOQR to show that the apparent depth is given by

$$h_i = \left(\frac{n_2}{n_1}\right)h_o. \tag{31.10}$$

The derivation of this result is left as an exercise. Thus, a fish appears at 3/4 of the real depth when viewed from above.

Refraction at a Spherical Interface

Spherical shapes play an important role in optics primarily because high-quality spherical shapes are far easier to manufacture than other curved surfaces. To study refraction at a single spherical surface, we assume that the medium with the spherical surface at one end continues indefinitely (a "semi-infinite" medium).

Refraction at a convex surface

Consider a point source of light at point *P* in front of a convex surface made of glass (see **Figure 31.15**). Let *R* be the radius of curvature, n_1 be the refractive index of the medium in which object point *P* is located, and n_2 be the refractive index of the medium with the spherical surface. We want to know what happens as a result of refraction at this interface.

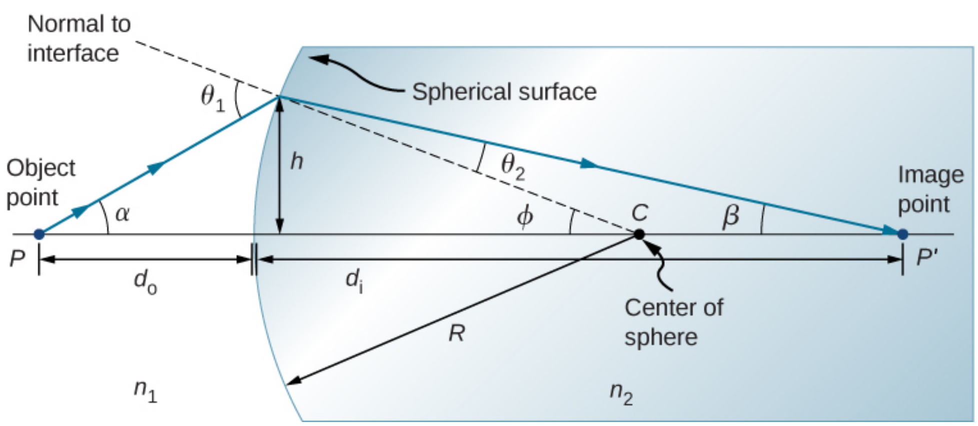


Figure 31.15 Refraction at a convex surface $(n_2 > n_1)$.

Because of the symmetry involved, it is sufficient to examine rays in only one plane. The figure shows a ray of light that starts at the object point *P*, refracts at the interface, and goes through the image point P'. We derive a formula relating the object distance d_o, the image distance d_i, and the radius of curvature *R*.

Applying Snell's law to the ray emanating from point *P* gives $n_1 \sin\theta_1 = n_2 \sin\theta_2$. We work in the small-angle approximation, so $\sin\theta \approx \theta$ and Snell's law then takes the form

$$n_1\theta_1 \approx n_2\theta_2.$$

From the geometry of the figure, we see that

$$\theta_1 = \alpha + \phi, \quad \theta_2 = \phi - \beta.$$

Inserting these expressions into Snell's law gives

$$n_1(\alpha + \phi) \approx n_2(\phi - \beta).$$

Using the diagram, we calculate the tangent of the angles α, β, and ϕ:

$$\tan\alpha \approx \frac{h}{d_o}, \quad \tan\beta \approx \frac{h}{d_i}, \quad \tan\phi \approx \frac{h}{R}.$$

Again using the small-angle approximation, we find that $\tan\theta \approx \theta$, so the above relationships become

$$\alpha \approx \frac{h}{d_o}, \quad \beta \approx \frac{h}{d_i}, \quad \phi \approx \frac{h}{R}.$$

Putting these angles into Snell's law gives

$$n_1\left(\frac{h}{d_o} + \frac{h}{R}\right) = n_2\left(\frac{h}{R} - \frac{h}{d_i}\right).$$

We can write this more conveniently as

$$\frac{n_1}{d_o} + \frac{n_2}{d_i} = \frac{n_2 - n_1}{R}. \qquad (31.11)$$

If the object is placed at a special point called the **first focus**, or the **object focus** F_1, then the image is formed at infinity, as shown in part (a) of Figure 31.16.

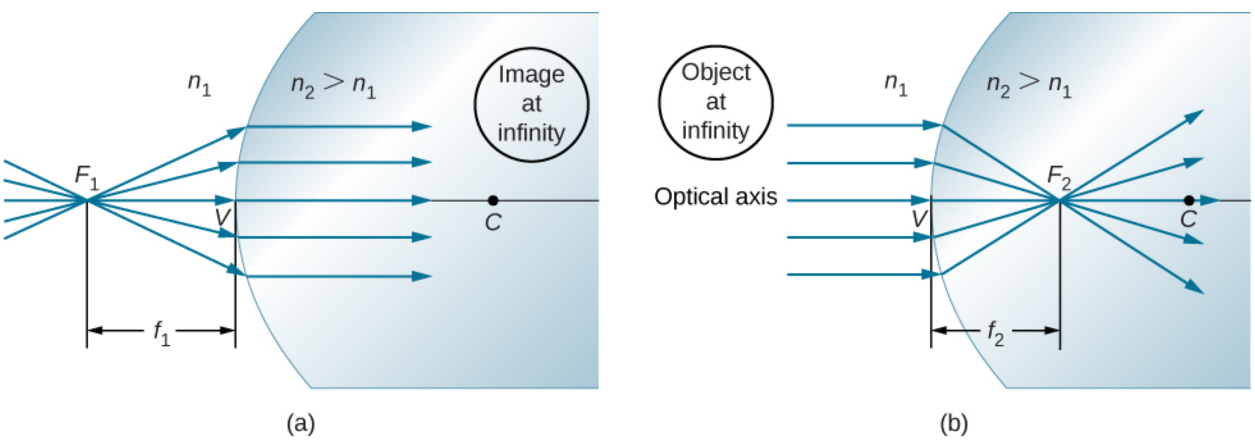


Figure 31.16 (a) First focus (called the "object focus") for refraction at a convex surface. (b) Second focus (called "image focus") for refraction at a convex surface.

We can find the location f_1 of the first focus F_1 by setting $d_i = \infty$ in the preceding equation.

$$\frac{n_1}{f_1} + \frac{n_2}{\infty} = \frac{n_2 - n_1}{R} \tag{31.12}$$

$$f_1 = \frac{n_1 R}{n_2 - n_1} \tag{31.13}$$

Similarly, we can define a **second focus** or **image focus** F_2 where the image is formed for an object that is far away [part (b)]. The location of the second focus F_2 is obtained from **Equation 31.11** by setting $d_o = \infty$:

$$\frac{n_1}{\infty} + \frac{n_2}{f_2} = \frac{n_2 - n_1}{R}$$

$$f_2 = \frac{n_2 R}{n_2 - n_1}.$$

Note that the object focus is at a different distance from the vertex than the image focus because $n_1 \neq n_2$.

Sign convention for single refracting surfaces

Although we derived this equation for refraction at a convex surface, the same expression holds for a concave surface, provided we use the following sign convention:

1. $R > 0$ if surface is convex toward object; otherwise, $R < 0$.
2. $d_i > 0$ if image is real and on opposite side from the object; otherwise, $d_i < 0$.

31.4 | Thin Lenses

Learning Objectives

By the end of this section, you will be able to:

- Use ray diagrams to locate and describe the image formed by a lens
- Employ the thin-lens equation to describe and locate the image formed by a lens

Lenses are found in a huge array of optical instruments, ranging from a simple magnifying glass to a camera's zoom lens to the eye itself. In this section, we use the Snell's law to explore the properties of lenses and how they form images.

The word "lens" derives from the Latin word for a lentil bean, the shape of which is similar to a convex lens. However, not all lenses have the same shape. **Figure 31.17** shows a variety of different lens shapes. The vocabulary used to describe lenses is the same as that used for spherical mirrors: The axis of symmetry of a lens is called the optical axis, where this axis intersects the lens surface is called the vertex of the lens, and so forth.

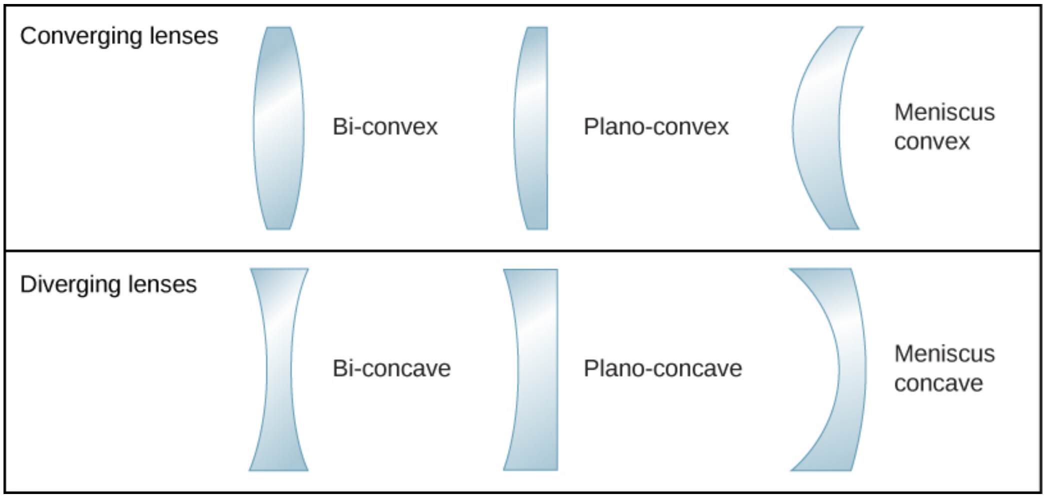


Figure 31.17 Various types of lenses: Note that a converging lens has a thicker "waist," whereas a diverging lens has a thinner waist.

A **convex** or **converging lens** is shaped so that all light rays that enter it parallel to its optical axis intersect (or focus) at a single point on the optical axis on the opposite side of the lens, as shown in part (a) of Figure 31.18. Likewise, a **concave** or **diverging lens** is shaped so that all rays that enter it parallel to its optical axis diverge, as shown in part (b). To understand more precisely how a lens manipulates light, look closely at the top ray that goes through the converging lens in part (a). Because the index of refraction of the lens is greater than that of air, Snell's law tells us that the ray is bent toward the perpendicular to the interface as it enters the lens. Likewise, when the ray exits the lens, it is bent away from the perpendicular. The same reasoning applies to the diverging lenses, as shown in part (b). The overall effect is that light rays are bent toward the optical axis for a converging lens and away from the optical axis for diverging lenses. For a converging lens, the point at which the rays cross is the focal point F of the lens. For a diverging lens, the point from which the rays appear to originate is the (virtual) focal point. The distance from the center of the lens to its focal point is the focal length f of the lens.

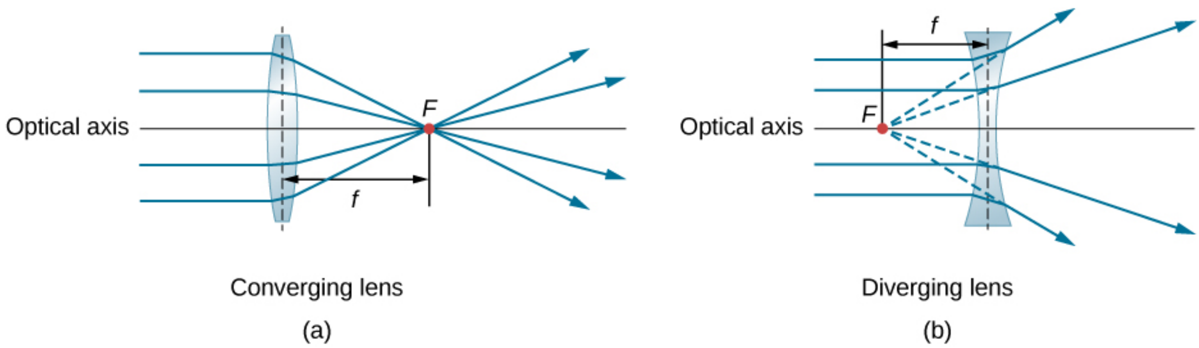


Figure 31.18 Rays of light entering (a) a converging lens and (b) a diverging lens, parallel to its axis, converge at its focal point F. The distance from the center of the lens to the focal point is the lens's focal length f. Note that the light rays are bent upon entering and exiting the lens, with the overall effect being to bend the rays toward the optical axis.

A lens is considered to be thin if its thickness t is much less than the radii of curvature of both surfaces, as shown in Figure 31.19. In this case, the rays may be considered to bend once at the center of the lens. For the case drawn in the figure, light ray 1 is parallel to the optical axis, so the outgoing ray is bent once at the center of the lens and goes through the focal point. Another important characteristic of thin lenses is that light rays that pass through the center of the lens are undeviated, as shown by light ray 2.

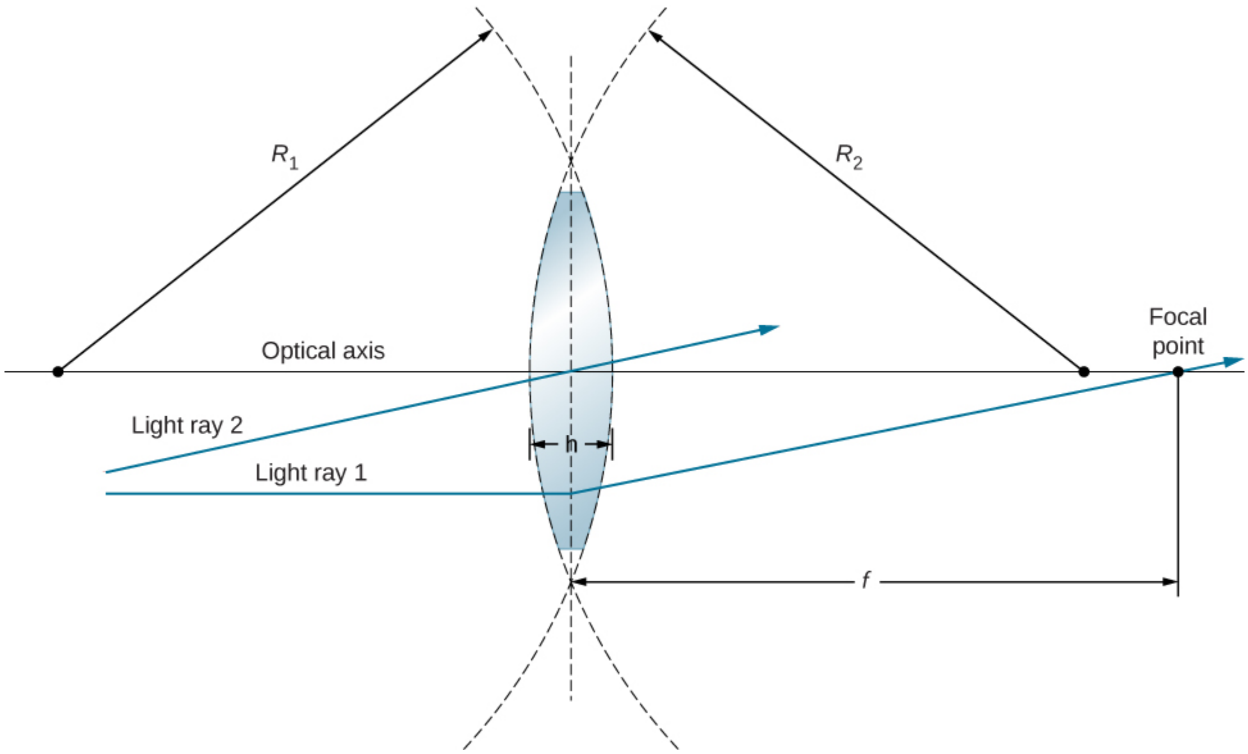


Figure 31.19 In the thin-lens approximation, the thickness *d* of the lens is much, much less than the radii R_1 and R_2 of curvature of the surfaces of the lens. Light rays are considered to bend at the center of the lens, such as light ray 1. Light ray 2 passes through the center of the lens and is undeviated in the thin-lens approximation.

As noted in the initial discussion of Snell's law, the paths of light rays are exactly reversible. This means that the direction of the arrows could be reversed for all of the rays in **Figure 31.18**. For example, if a point-light source is placed at the focal point of a convex lens, as shown in **Figure 31.20**, parallel light rays emerge from the other side.

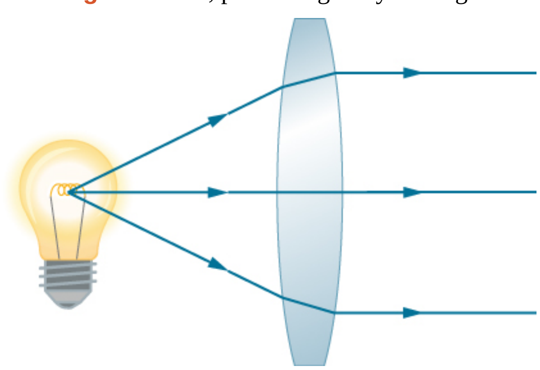


Figure 31.20 A small light source, like a light bulb filament, placed at the focal point of a convex lens results in parallel rays of light emerging from the other side. The paths are exactly the reverse of those shown in **Figure 31.18** in converging and diverging lenses. This technique is used in lighthouses and sometimes in traffic lights to produce a directional beam of light from a source that emits light in all directions.

Ray Tracing and Thin Lenses

Ray tracing is the technique of determining or following (tracing) the paths taken by light rays.

Ray tracing for thin lenses is very similar to the technique we used with spherical mirrors. As for mirrors, ray tracing can accurately describe the operation of a lens. The rules for ray tracing for thin lenses are similar to those of spherical mirrors:

1. A ray entering a converging lens parallel to the optical axis passes through the focal point on the other side of the

lens (ray 1 in part (a) of **Figure 31.21**). A ray entering a diverging lens parallel to the optical axis exits along the line that passes through the focal point on the *same* side of the lens (ray 1 in part (b) of the figure).

2. A ray passing through the center of either a converging or a diverging lens is not deviated (ray 2 in parts (a) and (b)).
3. For a converging lens, a ray that passes through the focal point exits the lens parallel to the optical axis (ray 3 in part (a)). For a diverging lens, a ray that approaches along the line that passes through the focal point on the opposite side exits the lens parallel to the axis (ray 3 in part (b)).

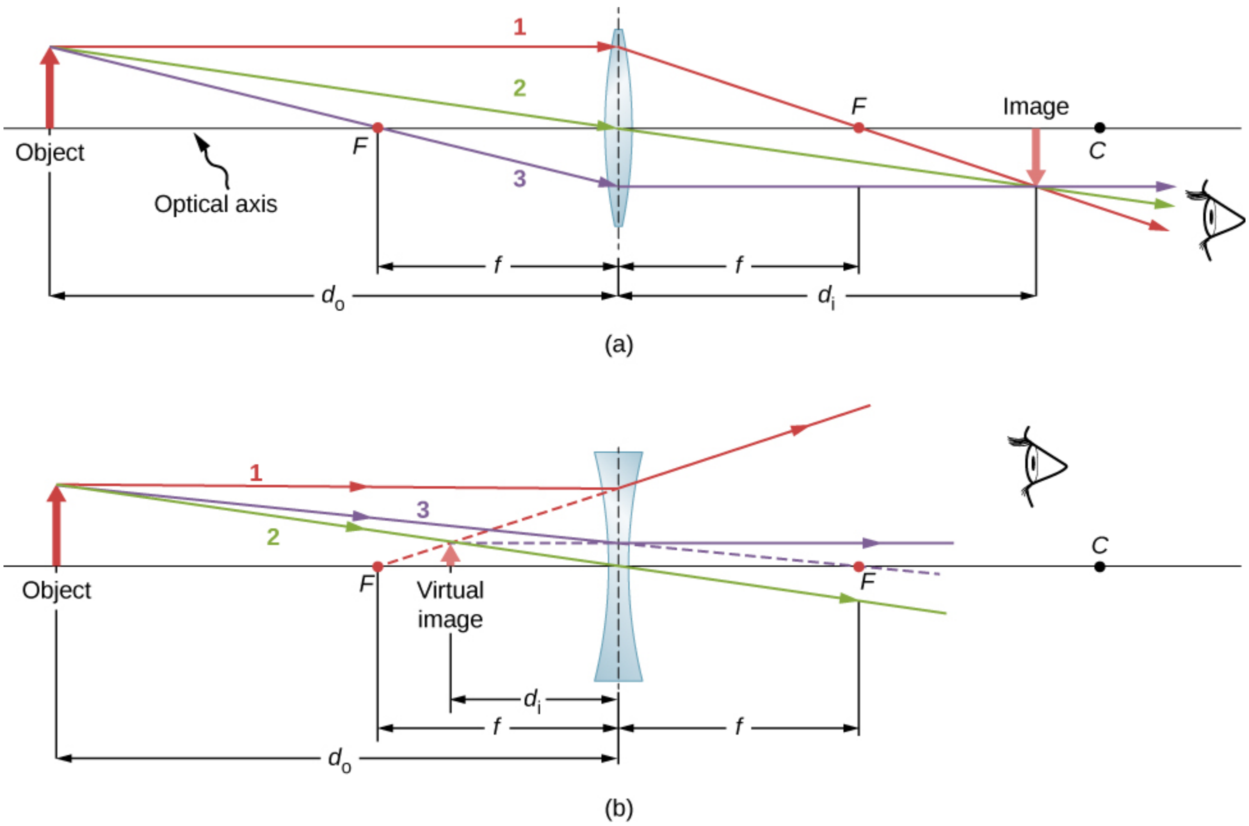


Figure 31.21 Thin lenses have the same focal lengths on either side. (a) Parallel light rays from the object toward a converging lens cross at its focal point on the right. (b) Parallel light rays from the object entering a diverging lens from the left seem to come from the focal point on the left.

Thin lenses work quite well for monochromatic light (i.e., light of a single wavelength). However, for light that contains several wavelengths (e.g., white light), the lenses work less well. The problem is that, as we learned in the previous chapter, the index of refraction of a material depends on the wavelength of light. This phenomenon is responsible for many colorful effects, such as rainbows. Unfortunately, this phenomenon also leads to aberrations in images formed by lenses. In particular, because the focal distance of the lens depends on the index of refraction, it also depends on the wavelength of the incident light. This means that light of different wavelengths will focus at different points, resulting is so-called “chromatic aberrations.” In particular, the edges of an image of a white object will become colored and blurred. Special lenses called doublets are capable of correcting chromatic aberrations. A doublet is formed by gluing together a converging lens and a diverging lens. The combined doublet lens produces significantly reduced chromatic aberrations.

Image Formation by Thin Lenses

We use ray tracing to investigate different types of images that can be created by a lens. In some circumstances, a lens forms a real image, such as when a movie projector casts an image onto a screen. In other cases, the image is a virtual image, which cannot be projected onto a screen. Where, for example, is the image formed by eyeglasses? We use ray tracing for thin lenses to illustrate how they form images, and then we develop equations to analyze quantitatively the properties of thin lenses.

Consider an object some distance away from a converging lens, as shown in **Figure 31.22**. To find the location and size of the image, we trace the paths of selected light rays originating from one point on the object, in this case, the tip of the arrow.

The figure shows three rays from many rays that emanate from the tip of the arrow. These three rays can be traced by using the ray-tracing rules given above.

- Ray 1 enters the lens parallel to the optical axis and passes through the focal point on the opposite side (rule 1).
- Ray 2 passes through the center of the lens and is not deviated (rule 2).
- Ray 3 passes through the focal point on its way to the lens and exits the lens parallel to the optical axis (rule 3).

The three rays cross at a single point on the opposite side of the lens. Thus, the image of the tip of the arrow is located at this point. All rays that come from the tip of the arrow and enter the lens are refracted and cross at the point shown.

After locating the image of the tip of the arrow, we need another point of the image to orient the entire image of the arrow. We chose to locate the image base of the arrow, which is on the optical axis. As explained in the section on spherical mirrors, the base will be on the optical axis just above the image of the tip of the arrow (due to the top-bottom symmetry of the lens). Thus, the image spans the optical axis to the (negative) height shown. Rays from another point on the arrow, such as the middle of the arrow, cross at another common point, thus filling in the rest of the image.

Although three rays are traced in this figure, only two are necessary to locate a point of the image. It is best to trace rays for which there are simple ray-tracing rules.

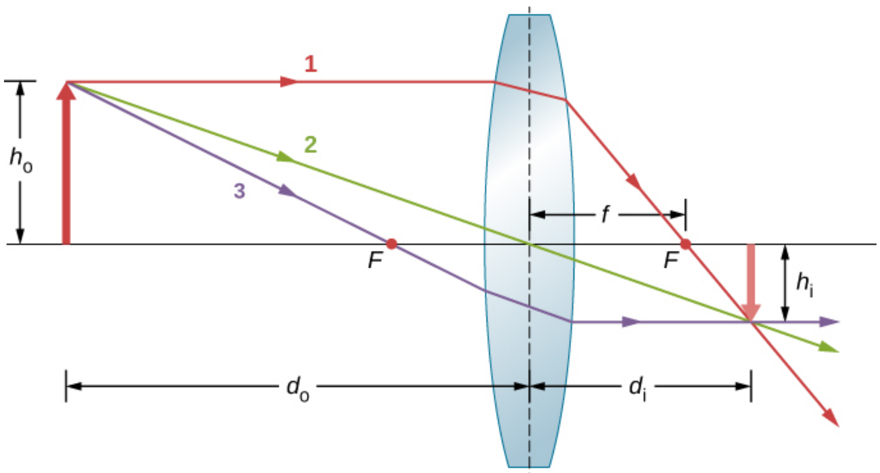


Figure 31.22 Ray tracing is used to locate the image formed by a lens. Rays originating from the same point on the object are traced—the three chosen rays each follow one of the rules for ray tracing, so that their paths are easy to determine. The image is located at the point where the rays cross. In this case, a real image—one that can be projected on a screen—is formed.

Several important distances appear in the figure. As for a mirror, we define d_o to be the object distance, or the distance of an object from the center of a lens. The image distance d_i is defined to be the distance of the image from the center of a lens. The height of the object and the height of the image are indicated by h_o and h_i, respectively. Images that appear upright relative to the object have positive heights, and those that are inverted have negative heights. By using the rules of ray tracing and making a scale drawing with paper and pencil, like that in Figure 31.22, we can accurately describe the location and size of an image. But the real benefit of ray tracing is in visualizing how images are formed in a variety of situations.

Oblique Parallel Rays and Focal Plane

We have seen that rays parallel to the optical axis are directed to the focal point of a converging lens. In the case of a diverging lens, they come out in a direction such that they appear to be coming from the focal point on the opposite side of the lens (i.e., the side from which parallel rays enter the lens). What happens to parallel rays that are not parallel to the optical axis (Figure 31.23)? In the case of a converging lens, these rays do not converge at the focal point. Instead, they come together on another point in the plane called the **focal plane**. The focal plane contains the focal point and is perpendicular to the optical axis. As shown in the figure, parallel rays focus where the ray through the center of the lens crosses the focal plane.

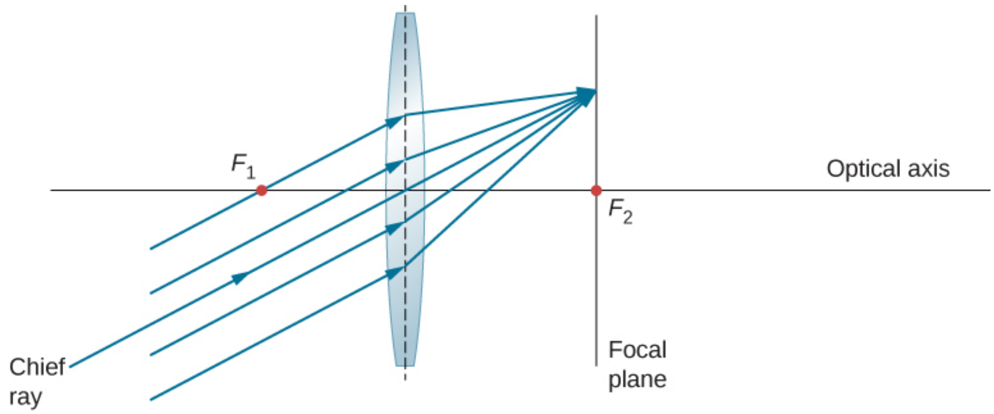


Figure 31.23 Parallel oblique rays focus on a point in a focal plane.

Thin-Lens Equation

Ray tracing allows us to get a qualitative picture of image formation. To obtain numeric information, we derive a pair of equations from a geometric analysis of ray tracing for thin lenses. These equations, called the thin-lens equation and the lens maker's equation, allow us to quantitatively analyze thin lenses.

Consider the thick bi-convex lens shown in **Figure 31.24**. The index of refraction of the surrounding medium is n_1 (if the lens is in air, then $n_1 = 1.00$) and that of the lens is n_2. The radii of curvatures of the two sides are R_1 and R_2. We wish to find a relation between the object distance d_o, the image distance d_i, and the parameters of the lens.

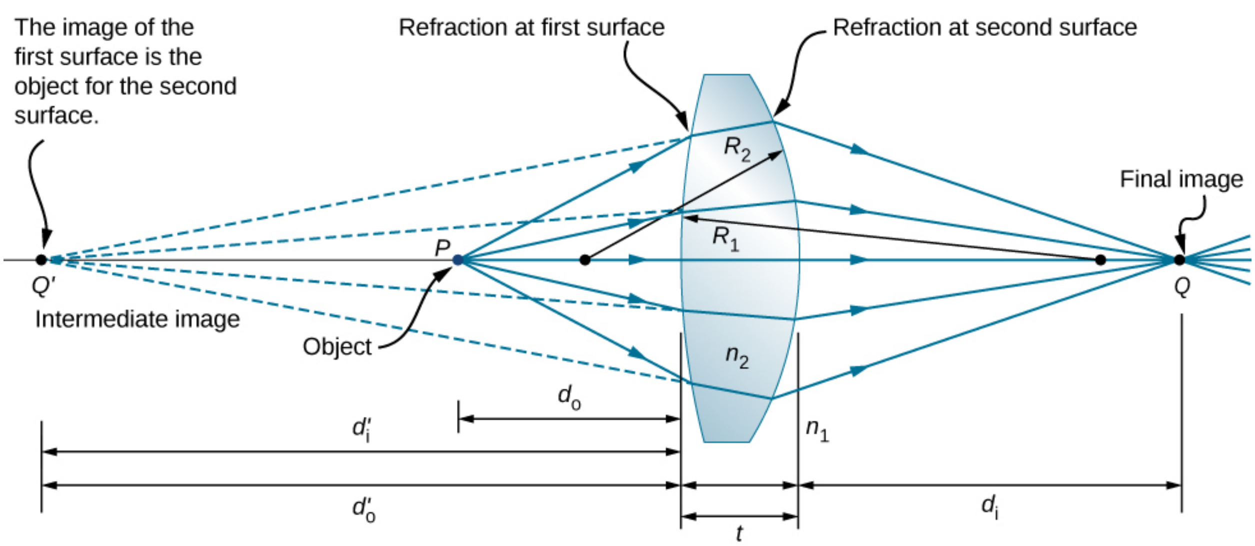


Figure 31.24 Figure for deriving the lens maker's equation. Here, t is the thickness of lens, n_1 is the index of refraction of the exterior medium, and n_2 is the index of refraction of the lens. We take the limit of $t \to 0$ to obtain the formula for a thin lens.

To derive the thin-lens equation, we consider the image formed by the first refracting surface (i.e., left surface) and then use this image as the object for the second refracting surface. In the figure, the image from the first refracting surface is Q', which is formed by extending backwards the rays from inside the lens (these rays result from refraction at the first surface). This is shown by the dashed lines in the figure. Notice that this image is virtual because no rays actually pass through the point Q'. To find the image distance d'_i corresponding to the image Q', we use **Equation 31.11**. In this case, the object distance is d_o, the image distance is d'_i, and the radius of curvature is R_1. Inserting these into **Equation 31.3** gives

$$\frac{n_1}{d_o} + \frac{n_2}{d'_i} = \frac{n_2 - n_1}{R_1}. \quad (31.14)$$

The image is virtual and on the same side as the object, so $d'_i < 0$ and $d_o > 0$. The first surface is convex toward the object, so $R_1 > 0$.

To find the object distance for the object Q formed by refraction from the second interface, note that the role of the indices of refraction n_1 and n_2 are interchanged in **Equation 31.11**. In **Figure 31.24**, the rays originate in the medium with index n_2, whereas in **Figure 31.15**, the rays originate in the medium with index n_1. Thus, we must interchange n_1 and n_2 in **Equation 31.11**. In addition, by consulting again **Figure 31.24**, we see that the object distance is d'_o and the image distance is d_i. The radius of curvature is R_2 Inserting these quantities into **Equation 31.11** gives

$$\frac{n_2}{d'_o} + \frac{n_1}{d_i} = \frac{n_1 - n_2}{R_2}. \tag{31.15}$$

The image is real and on the opposite side from the object, so $d_i > 0$ and $d'_o > 0$. The second surface is convex away from the object, so $R_2 < 0$. **Equation 31.15** can be simplified by noting that $d'_o = |d'_i| + t$, where we have taken the absolute value because d'_i is a negative number, whereas both d'_o and t are positive. We can dispense with the absolute value if we negate d'_i, which gives $d'_o = -d'_i + t$. Inserting this into **Equation 31.15** gives

$$\frac{n_2}{-d'_i + t} + \frac{n_1}{d_i} = \frac{n_1 - n_2}{R_2}. \tag{31.16}$$

Summing **Equation 31.14** and **Equation 31.16** gives

$$\frac{n_1}{d_o} + \frac{n_1}{d_i} + \frac{n_2}{d'_i} + \frac{n_2}{-d'_i + t} = (n_2 - n_1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right). \tag{31.17}$$

In the **thin-lens approximation**, we assume that the lens is very thin compared to the first image distance, or $t \ll d'_i$ (or, equivalently, $t \ll R_1$ and R_2). In this case, the third and fourth terms on the left-hand side of **Equation 31.17** cancel, leaving us with

$$\frac{n_1}{d_o} + \frac{n_1}{d_i} = (n_2 - n_1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right).$$

Dividing by n_1 gives us finally

$$\frac{1}{d_o} + \frac{1}{d_i} = \left(\frac{n_2}{n_1} - 1\right)\left(\frac{1}{R_1} - \frac{1}{R_2}\right). \tag{31.18}$$

The left-hand side looks suspiciously like the mirror equation that we derived above for spherical mirrors. As done for spherical mirrors, we can use ray tracing and geometry to show that, for a thin lens,

$$\frac{1}{d_o} + \frac{1}{d_i} = \frac{1}{f} \tag{31.19}$$

where f is the focal length of the thin lens (this derivation is left as an exercise). This is the thin-lens equation. The focal length of a thin lens is the same to the left and to the right of the lens. Combining **Equation 31.18** and **Equation 31.19** gives

$$\frac{1}{f} = \left(\frac{n_2}{n_1} - 1\right)\left(\frac{1}{R_1} - \frac{1}{R_2}\right) \tag{31.20}$$

which is called the lens maker's equation. It shows that the focal length of a thin lens depends only of the radii of curvature and the index of refraction of the lens and that of the surrounding medium. For a lens in air, $n_1 = 1.0$ and $n_2 \equiv n$, so the lens maker's equation reduces to

$$\frac{1}{f} = (n - 1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right). \quad (31.21)$$

Sign conventions for lenses

To properly use the thin-lens equation, the following sign conventions must be obeyed:

1. d_i is positive if the image is on the side opposite the object (i.e., real image); otherwise, d_i is negative (i.e., virtual image).
2. f is positive for a converging lens and negative for a diverging lens.
3. R is positive for a surface convex toward the object, and negative for a surface concave toward object.

Magnification

By using a finite-size object on the optical axis and ray tracing, you can show that the magnification m of an image is

$$m \equiv \frac{h_i}{h_o} = -\frac{d_i}{d_o} \quad (31.22)$$

(where the three lines mean "is defined as"). This is exactly the same equation as we obtained for mirrors (see **Equation 31.8**). If $m > 0$, then the image has the same vertical orientation as the object (called an "upright" image). If $m < 0$, then the image has the opposite vertical orientation as the object (called an "inverted" image).

Using the Thin-Lens Equation

The thin-lens equation and the lens maker's equation are broadly applicable to situations involving thin lenses. We explore many features of image formation in the following examples.

Consider a thin converging lens. Where does the image form and what type of image is formed as the object approaches the lens from infinity? This may be seen by using the thin-lens equation for a given focal length to plot the image distance as a function of object distance. In other words, we plot

$$d_i = \left(\frac{1}{f} - \frac{1}{d_o}\right)^{-1}$$

for a given value of f. For $f = 1$ cm, the result is shown in part (a) of **Figure 31.25**.

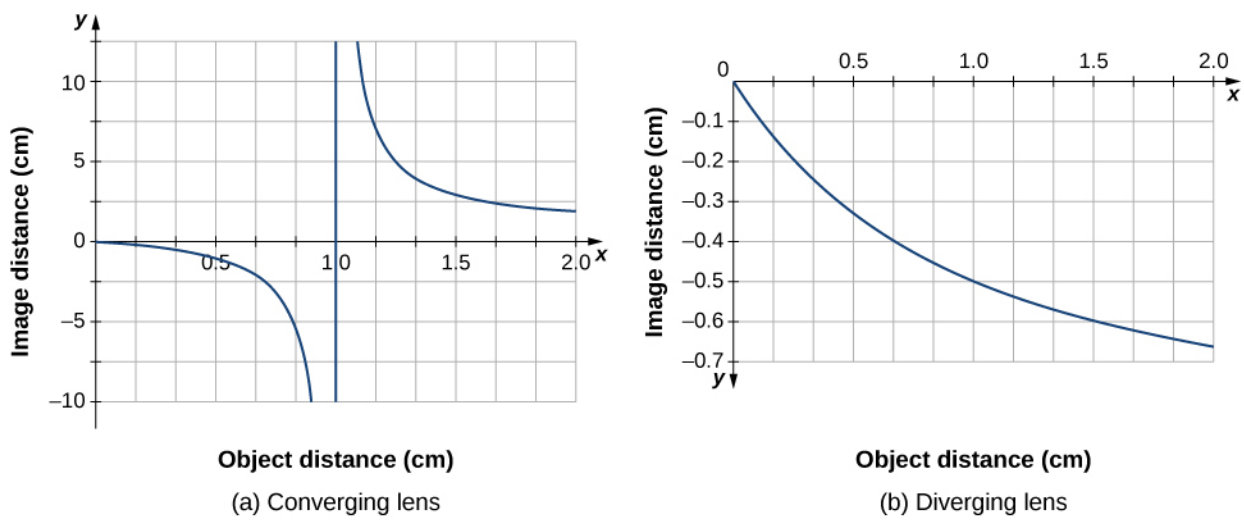


Figure 31.25 (a) Image distance for a thin converging lens with $f = 1.0$ cm as a function of object distance. (b) Same thing but for a diverging lens with $f = -1.0$ cm.

An object much farther than the focal length f from the lens should produce an image near the focal plane, because the second term on the right-hand side of the equation above becomes negligible compared to the first term, so we have $d_\text{i} \approx f$. This can be seen in the plot of part (a) of the figure, which shows that the image distance approaches asymptotically the focal length of 1 cm for larger object distances. As the object approaches the focal plane, the image distance diverges to positive infinity. This is expected because an object at the focal plane produces parallel rays that form an image at infinity (i.e., very far from the lens). When the object is farther than the focal length from the lens, the image distance is positive, so the image is real, on the opposite side of the lens from the object, and inverted (because $m = -d_\text{i}/d_\text{o}$). When the object is closer than the focal length from the lens, the image distance becomes negative, which means that the image is virtual, on the same side of the lens as the object, and upright.

For a thin diverging lens of focal length $f = -1.0\text{ cm}$, a similar plot of image distance vs. object distance is shown in part (b). In this case, the image distance is negative for all positive object distances, which means that the image is virtual, on the same side of the lens as the object, and upright. These characteristics may also be seen by ray-tracing diagrams (see **Figure 31.26**).

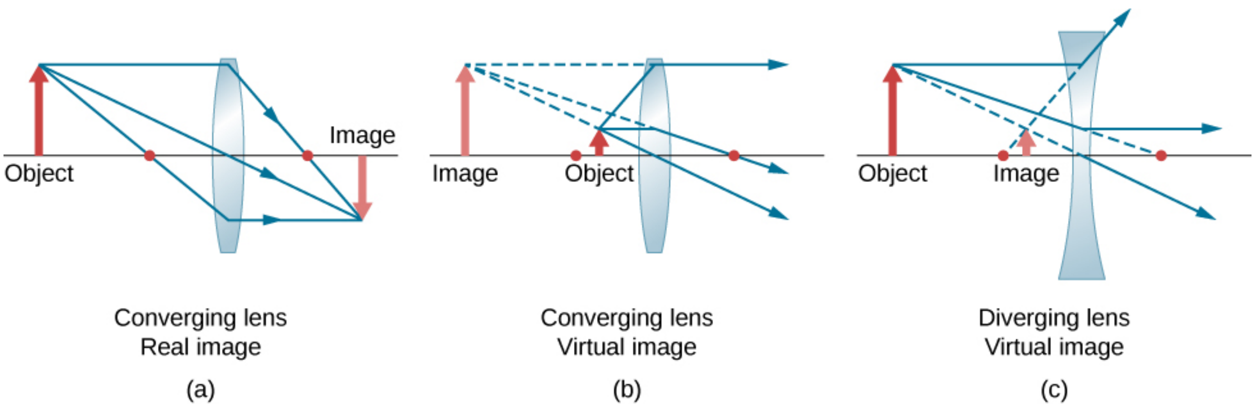


Figure 31.26 The red dots show the focal points of the lenses. (a) A real, inverted image formed from an object that is farther than the focal length from a converging lens. (b) A virtual, upright image formed from an object that is closer than a focal length from the lens. (c) A virtual, upright image formed from an object that is farther than a focal length from a diverging lens.

To see a concrete example of upright and inverted images, look at **Figure 31.27**, which shows images formed by converging lenses when the object (the person's face in this case) is place at different distances from the lens. In part (a) of the figure, the person's face is farther than one focal length from the lens, so the image is inverted. In part (b), the person's face is closer than one focal length from the lens, so the image is upright.


(a) (b)

Figure 31.27 (a) When a converging lens is held farther than one focal length from the man's face, an inverted image is formed. Note that the image is in focus but the face is not, because the image is much closer to the camera taking this photograph than the face. (b) An upright image of the man's face is produced when a converging lens is held at less than one focal length from his face. (credit a: modification of work by "DaMongMan"/Flickr; credit b: modification of work by Casey Fleser)

Work through the following examples to better understand how thin lenses work.

Problem-Solving Strategy: Lenses

Step 1. Determine whether ray tracing, the thin-lens equation, or both would be useful. Even if ray tracing is not used, a careful sketch is always very useful. Write symbols and values on the sketch.

Step 2. Identify what needs to be determined in the problem (identify the unknowns).

Step 3. Make a list of what is given or can be inferred from the problem (identify the knowns).

Step 4. If ray tracing is required, use the ray-tracing rules listed near the beginning of this section.

Step 5. Most quantitative problems require the use of the thin-lens equation and/or the lens maker's equation. Solve these for the unknowns and insert the given quantities or use both together to find two unknowns.

Step 7. Check to see if the answer is reasonable. Are the signs correct? Is the sketch or ray tracing consistent with the calculation?

Example 31.3

Using the Lens Maker's Equation

Find the radius of curvature of a biconcave lens symmetrically ground from a glass with index of refractive 1.55 so that its focal length in air is 20 cm (for a biconcave lens, both surfaces have the same radius of curvature).

Strategy

Use the thin-lens form of the lens maker's equation:

$$\frac{1}{f} = \left(\frac{n_2}{n_1} - 1\right)\left(\frac{1}{R_1} - \frac{1}{R_2}\right)$$

where $R_1 < 0$ and $R_2 > 0$. Since we are making a symmetric biconcave lens, we have $|R_1| = |R_2|$.

Solution

We can determine the radius R of curvature from

$$\frac{1}{f} = \left(\frac{n_2}{n_1} - 1\right)\left(-\frac{2}{R}\right).$$

Solving for R and inserting $f = -20\text{ cm}$, $n_2 = 1.55$, and $n_1 = 1.00$ gives

$$R = -2f\left(\frac{n_2}{n_1} - 1\right) = -2(-20\text{ cm})\left(\frac{1.55}{1.00} - 1\right) = 22\text{ cm}.$$

Example 31.4

Converging Lens and Different Object Distances

Find the location, orientation, and magnification of the image for an 3.0 cm high object at each of the following positions in front of a convex lens of focal length 10.0 cm. (a) $d_o = 50.0\text{ cm}$, (b) $d_o = 5.00\text{ cm}$, and (c) $d_o = 20.0\text{ cm}$.

Strategy

We start with the thin-lens equation $\frac{1}{d_i} + \frac{1}{d_o} = \frac{1}{f}$. Solve this for the image distance d_i and insert the given object distance and focal length.

Solution

a. For $d_o = 50\text{ cm}, f = +10\text{ cm}$, this gives

$$\begin{aligned} d_i &= \left(\frac{1}{f} - \frac{1}{d_o}\right)^{-1} \\ &= \left(\frac{1}{10.0\text{ cm}} - \frac{1}{50.0\text{ cm}}\right)^{-1} \\ &= 12.5\text{ cm} \end{aligned}$$

The image is positive, so the image, is real, is on the opposite side of the lens from the object, and is 12.6 cm from the lens. To find the magnification and orientation of the image, use

$$m = -\frac{d_i}{d_o} = -\frac{12.5\text{ cm}}{50.0\text{ cm}} = -0.250.$$

The negative magnification means that the image is inverted. Since $|m| < 1$, the image is smaller than the object. The size of the image is given by

$$|h_i| = |m|h_o = (0.250)(3.0\text{ cm}) = 0.75\text{ cm}$$

b. For $d_o = 5.00\text{ cm}, f = +10.0\text{ cm}$

$$\begin{aligned} d_i &= \left(\frac{1}{f} - \frac{1}{d_o}\right)^{-1} \\ &= \left(\frac{1}{10.0\text{ cm}} - \frac{1}{5.00\text{ cm}}\right)^{-1} \\ &= -10.0\text{ cm} \end{aligned}$$

The image distance is negative, so the image is virtual, is on the same side of the lens as the object, and is 10 cm from the lens. The magnification and orientation of the image are found from

$$m = -\frac{d_i}{d_o} = -\frac{-10.0\text{ cm}}{5.00\text{ cm}} = +2.00.$$

The positive magnification means that the image is upright (i.e., it has the same orientation as the object). Since $|m| > 0$, the image is larger than the object. The size of the image is

$$|h_i| = |m|h_o = (2.00)(3.0\text{ cm}) = 6.0\text{ cm}.$$

c. For $d_o = 20\text{ cm}, f = +10\text{ cm}$

$$\begin{aligned} d_i &= \left(\frac{1}{f} - \frac{1}{d_o}\right)^{-1} \\ &= \left(\frac{1}{10.0\text{ cm}} - \frac{1}{20.0\text{ cm}}\right)^{-1} \\ &= 20.0\text{ cm} \end{aligned}$$

The image distance is positive, so the image is real, is on the opposite side of the lens from the object, and is 20.0 cm from the lens. The magnification is

$$m = -\frac{d_i}{d_o} = -\frac{20.0\text{ cm}}{20.0\text{ cm}} = -1.00.$$

The negative magnification means that the image is inverted. Since $|m| = 1$, the image is the same size as the object.

When solving problems in geometric optics, we often need to combine ray tracing and the lens equations. The following example demonstrates this approach.

Example 31.5

Choosing the Focal Length and Type of Lens

To project an image of a light bulb on a screen 1.50 m away, you need to choose what type of lens to use (converging or diverging) and its focal length (**Figure 31.28**). The distance between the lens and the lightbulb is fixed at 0.75 m. Also, what is the magnification and orientation of the image?

Strategy

The image must be real, so you choose to use a converging lens. The focal length can be found by using the thin-lens equation and solving for the focal length. The object distance is $d_o = 0.75\text{ m}$ and the image distance is $d_i = 1.5\text{ m}$.

Solution

Solve the thin lens for the focal length and insert the desired object and image distances:

$$\begin{aligned} \frac{1}{d_o} + \frac{1}{d_i} &= \frac{1}{f} \\ f &= \left(\frac{1}{d_o} + \frac{1}{d_i}\right)^{-1} \\ &= \left(\frac{1}{0.75\text{ m}} + \frac{1}{1.5\text{ m}}\right)^{-1} \\ &= 0.50\text{ m} \end{aligned}$$

The magnification is

$$m = -\frac{d_i}{d_o} = -\frac{1.5\text{ m}}{0.75\text{ m}} = -2.0.$$

Significance

The minus sign for the magnification means that the image is inverted. The focal length is positive, as expected for a converging lens. Ray tracing can be used to check the calculation (see **Figure 31.28**). As expected, the image is inverted, is real, and is larger than the object.

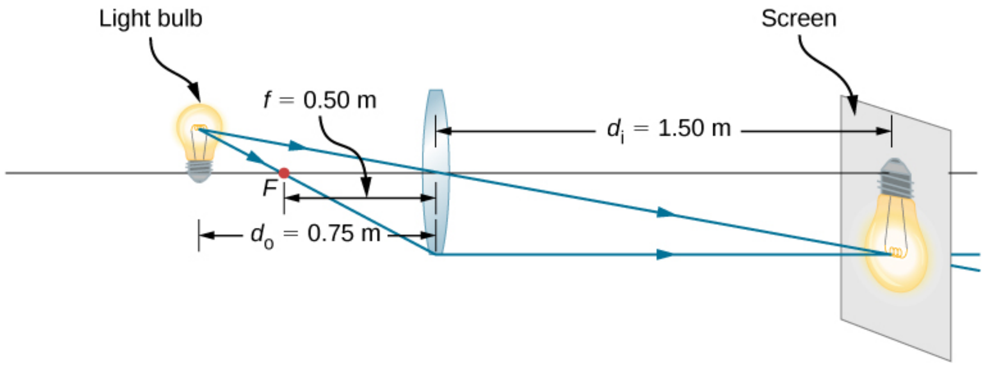


Figure 31.28 A light bulb placed 0.75 m from a lens having a 0.50-m focal length produces a real image on a screen, as discussed in the example. Ray tracing predicts the image location and size.

31.5 | The Camera

Learning Objectives

By the end of this section, you will be able to:

- Describe the optics of a camera
- Characterize the image created by a camera

Cameras are very common in our everyday life. Between 1825 and 1827, French inventor Nicéphore Niépce successfully photographed an image created by a primitive camera. Since then, enormous progress has been achieved in the design of cameras and camera-based detectors.

Initially, photographs were recorded by using the light-sensitive reaction of silver-based compounds such as silver chloride or silver bromide. Silver-based photographic paper was in common use until the advent of digital photography in the 1980s, which is intimately connected to **charge-coupled device (CCD)** detectors. In a nutshell, a CCD is a semiconductor chip that records images as a matrix of tiny pixels, each pixel located in a "bin" in the surface. Each pixel is capable of detecting the intensity of light impinging on it. Color is brought into play by putting red-, blue-, and green-colored filters over the pixels, resulting in colored digital images (**Figure 31.29**). At its best resolution, one CCD pixel corresponds to one pixel of the image. To reduce the resolution and decrease the size of the file, we can "bin" several CCD pixels into one, resulting in a smaller but "pixelated" image.

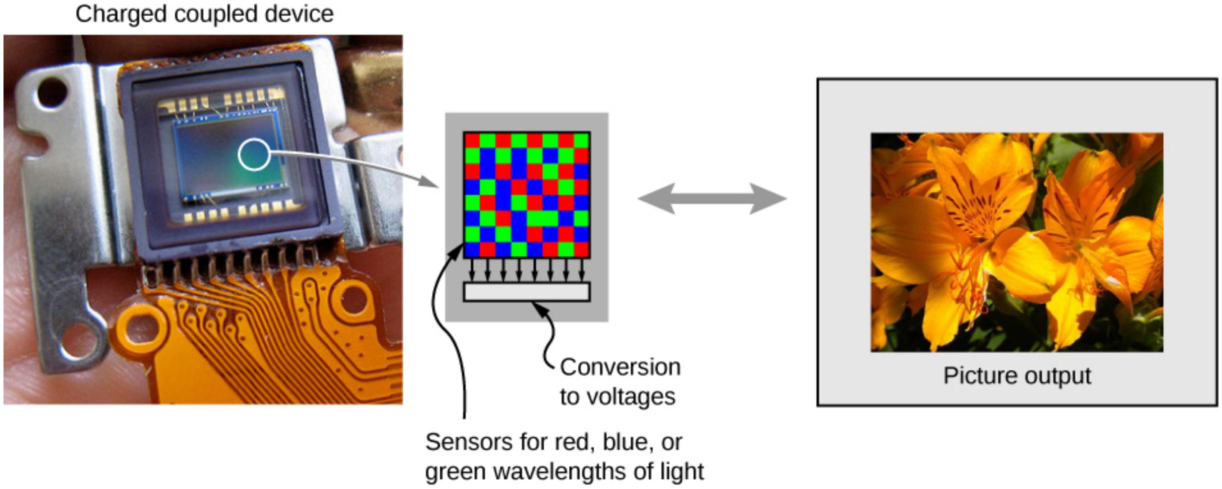


Figure 31.29 A charge-coupled device (CCD) converts light signals into electronic signals, enabling electronic processing and storage of visual images. This is the basis for electronic imaging in all digital cameras, from cell phones to movie cameras. (credit left: modification of work by Bruce Turner)

Clearly, electronics is a big part of a digital camera; however, the underlying physics is basic optics. As a matter of fact, the optics of a camera are pretty much the same as those of a single lens with an object distance that is significantly larger than the lens's focal distance (**Figure 31.30**).

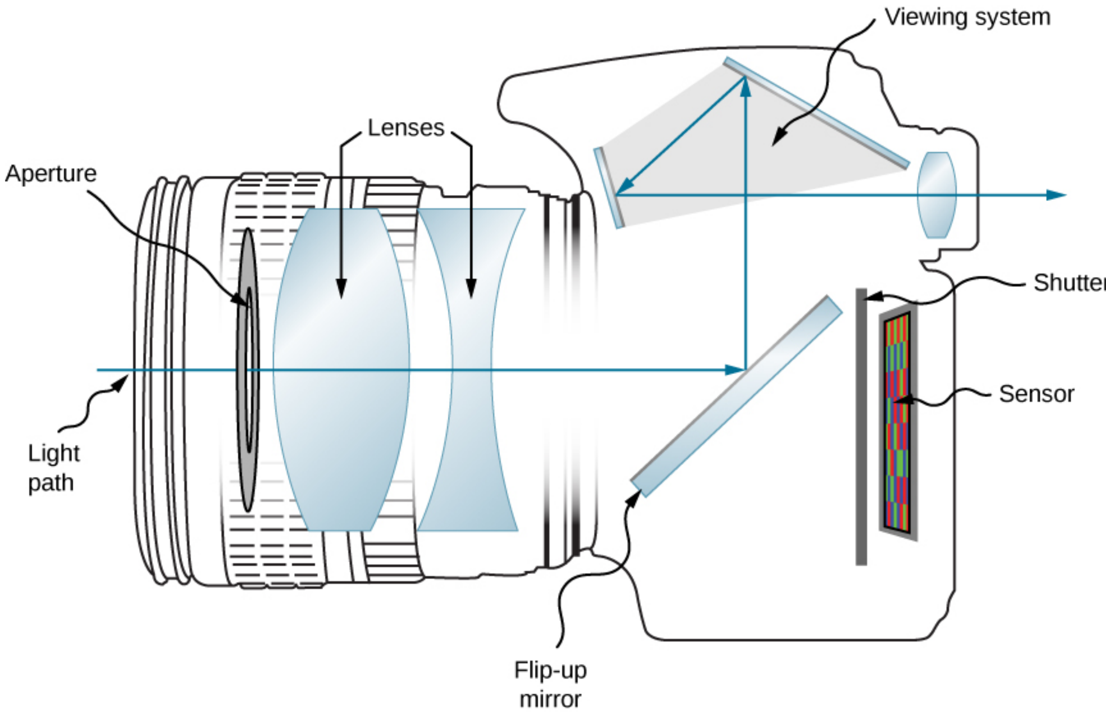


Figure 31.30 Modern digital cameras have several lenses to produce a clear image with minimal aberration and use red, blue, and green filters to produce a color image.

For instance, let us consider the camera in a smartphone. An average smartphone camera is equipped with a stationary wide-angle lens with a focal length of about 4–5 mm. (This focal length is about equal to the thickness of the phone.) The image created by the lens is focused on the CCD detector mounted at the opposite side of the phone. In a cell phone, the lens and the CCD cannot move relative to each other. So how do we make sure that both the images of a distant and a close object are in focus?

Recall that a human eye can accommodate for distant and close images by changing its focal distance. A cell phone camera cannot do that because the distance from the lens to the detector is fixed. Here is where the small focal distance becomes important. Let us assume we have a camera with a 5-mm focal distance. What is the image distance for a selfie? The object distance for a selfie (the length of the hand holding the phone) is about 50 cm. Using the thin-lens equation, we can write

$$\frac{1}{5\text{ mm}} = \frac{1}{500\text{ mm}} + \frac{1}{d_i}$$

We then obtain the image distance:

$$\frac{1}{d_\text{i}} = \frac{1}{5\text{ mm}} - \frac{1}{500\text{ mm}}$$

Note that the object distance is 100 times larger than the focal distance. We can clearly see that the 1/(500 mm) term is significantly smaller than 1/(5 mm), which means that the image distance is pretty much equal to the lens's focal length. An actual calculation gives us the image distance $d_i = 5.05\text{ mm}$. This value is extremely close to the lens's focal distance.

Now let us consider the case of a distant object. Let us say that we would like to take a picture of a person standing about 5 m from us. Using the thin-lens equation again, we obtain the image distance of 5.005 mm. The farther the object is from the lens, the closer the image distance is to the focal distance. At the limiting case of an infinitely distant object, we obtain the image distance exactly equal to the focal distance of the lens.

As you can see, the difference between the image distance for a selfie and the image distance for a distant object is just about 0.05 mm or 50 microns. Even a short object distance such as the length of your hand is two orders of magnitude larger than the lens's focal length, resulting in minute variations of the image distance. (The 50-micron difference is smaller than the thickness of an average sheet of paper.) Such a small difference can be easily accommodated by the same detector, positioned at the focal distance of the lens. Image analysis software can help improve image quality.

Conventional point-and-shoot cameras often use a movable lens to change the lens-to-image distance. Complex lenses of the more expensive mirror reflex cameras allow for superb quality photographic images. The optics of these camera lenses is beyond the scope of this textbook.

31.6 | The Simple Magnifier

Learning Objectives

By the end of this section, you will be able to:

- Understand the optics of a simple magnifier
- Characterize the image created by a simple magnifier

The apparent size of an object perceived by the eye depends on the angle the object subtends from the eye. As shown in **Figure 31.31**, the object at *A* subtends a larger angle from the eye than when it is position at point *B*. Thus, the object at *A* forms a larger image on the retina (see OA') than when it is positioned at *B* (see OB'). Thus, objects that subtend large angles from the eye appear larger because they form larger images on the retina.

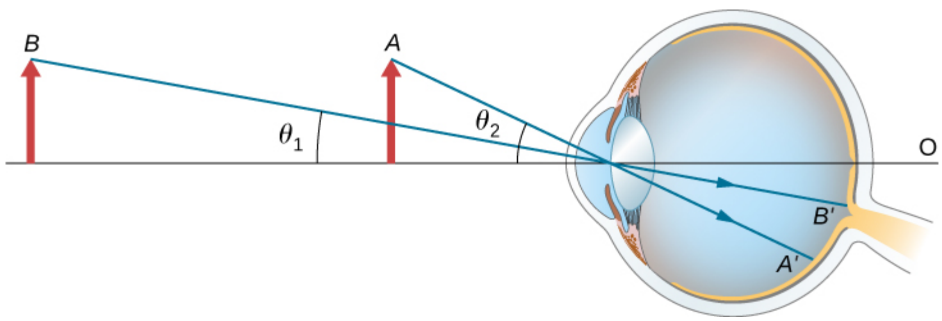


Figure 31.31 Size perceived by an eye is determined by the angle subtended by the object. An image formed on the retina by an object at *A* is larger than an image formed on the retina by the same object positioned at *B* (compared image heights OA' to OB').

We have seen that, when an object is placed within a focal length of a convex lens, its image is virtual, upright, and larger than the object (see part (b) of **Figure 31.26**). Thus, when such an image produced by a convex lens serves as the object for the eye, as shown in **Figure 31.32**, the image on the retina is enlarged, because the image produced by the lens subtends a larger angle in the eye than does the object. A convex lens used for this purpose is called a **magnifying glass** or a **simple magnifier**.

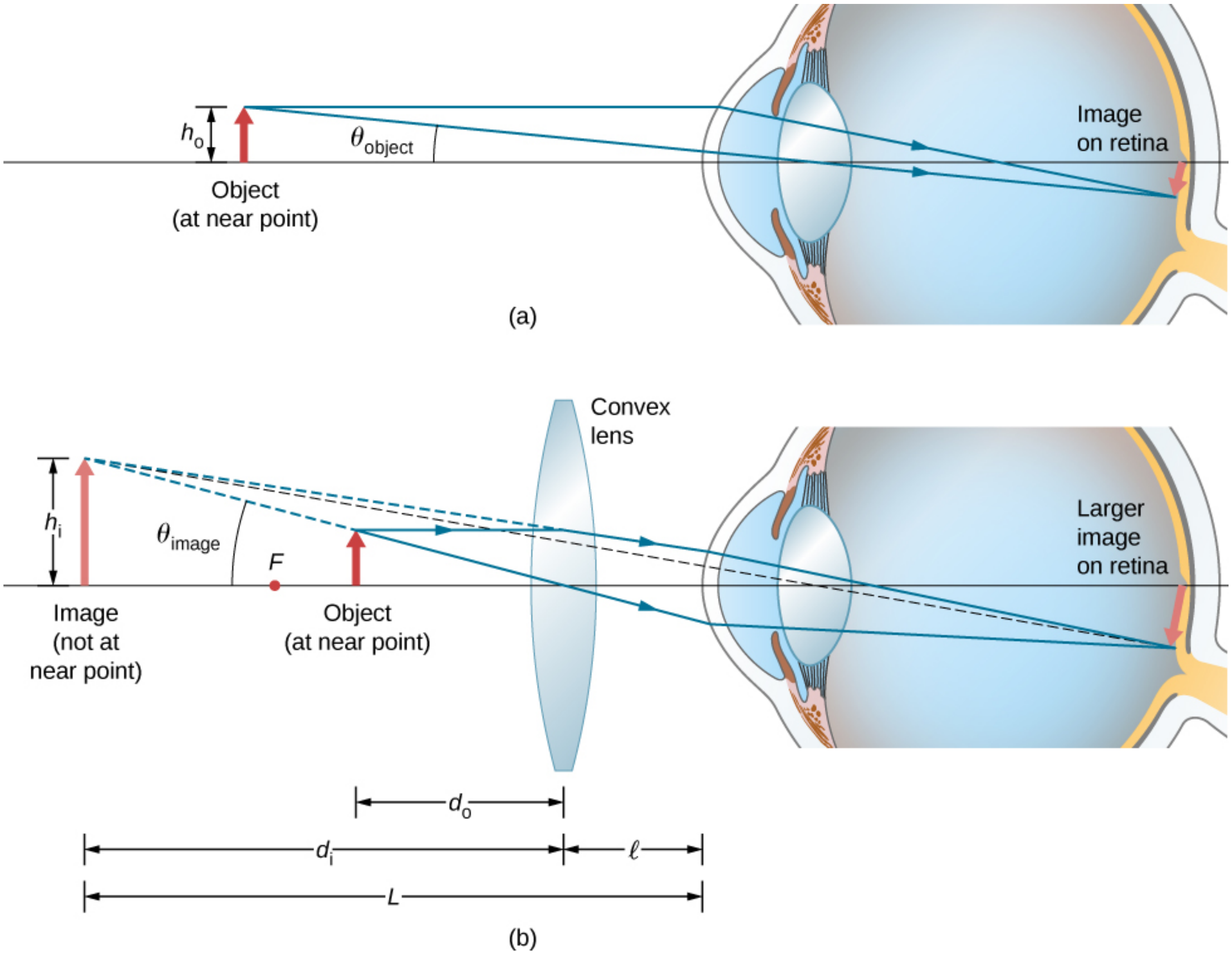


Figure 31.32 The simple magnifier is a convex lens used to produce an enlarged image of an object on the retina. (a) With no convex lens, the object subtends an angle θ_{object} from the eye. (b) With the convex lens in place, the image produced by the convex lens subtends an angle θ_{image} from the eye, with $\theta_{\text{image}} > \theta_{\text{object}}$. Thus, the image on the retina is larger with the convex lens in place.

To account for the magnification of a magnifying lens, we compare the angle subtended by the image (created by the lens) with the angle subtended by the object (viewed with no lens), as shown in **Figure 31.32**. We assume that the object is situated at the near point of the eye, because this is the object distance at which the unaided eye can form the largest image on the retina. We will compare the magnified images created by a lens with this maximum image size for the unaided eye. The magnification of an image when observed by the eye is the **angular magnification** *M*, which is defined by the ratio of the angle θ_{image} subtended by the image to the angle θ_{object} subtended by the object:

$$M = \frac{\theta_{\text{image}}}{\theta_{\text{object}}}. \tag{31.23}$$

Consider the situation shown in **Figure 31.32**. The magnifying lens is held a distance ℓ from the eye, and the image produced by the magnifier forms a distance *L* from the eye. We want to calculate the angular magnification for any arbitrary

L and ℓ. In the small-angle approximation, the angular size θ_{image} of the image is h_i/L. The angular size θ_{object} of the object at the near point is $\theta_{\text{object}} = h_o/25\text{ cm}$. The angular magnification is then

$$M = \frac{\theta_{\text{image}}}{\theta_{\text{object}}} = \frac{h_i(25\text{ cm})}{Lh_o}. \qquad (31.24)$$

Using **Equation 31.8** for linear magnification

$$m = -\frac{d_i}{d_o} = \frac{h_i}{h_o}$$

and the thin-lens equation

$$\frac{1}{d_o} + \frac{1}{d_i} = \frac{1}{f}$$

in **Equation 31.24**, we arrive at the following expression for the angular magnification of a magnifying lens:

$$\begin{aligned} M &= \left(-\frac{d_i}{d_o}\right)\left(\frac{25\text{ cm}}{L}\right) \\ &= -d_i\left(\frac{1}{f} - \frac{1}{d_i}\right)\left(\frac{25\text{ cm}}{L}\right) \\ &= \left(1 - \frac{d_i}{f}\right)\left(\frac{25\text{ cm}}{L}\right) \end{aligned} \qquad (31.25)$$

From part (b) of the figure, we see that the absolute value of the image distance is $|d_i| = L - \ell$. Note that $d_i < 0$ because the image is virtual, so we can dispense with the absolute value by explicitly inserting the minus sign: $-d_i = L - \ell$. Inserting this into **Equation 31.25** gives us the final equation for the angular magnification of a magnifying lens:

$$M = \left(\frac{25\text{ cm}}{L}\right)\left(1 + \frac{L - \ell}{f}\right). \qquad (31.26)$$

Note that all the quantities in this equation have to be expressed in centimeters. Often, we want the image to be at the near-point distance ($L = 25\text{ cm}$) to get maximum magnification, and we hold the magnifying lens close to the eye ($\ell = 0$). In this case, **Equation 31.26** gives

$$M = 1 + \frac{25\text{ cm}}{f} \qquad (31.27)$$

which shows that the greatest magnification occurs for the lens with the shortest focal length. In addition, when the image is at the near-point distance and the lens is held close to the eye $(\ell = 0)$, then $L = d_i = 25\text{ cm}$ and **Equation 31.24** becomes

$$M = \frac{h_i}{h_o} = m \qquad (31.28)$$

where m is the linear magnification (**Equation 31.29**) derived for spherical mirrors and thin lenses. Another useful situation is when the image is at infinity $(L = \infty)$. **Equation 31.26** then takes the form

$$M(L = \infty) = \frac{25\text{ cm}}{f}. \qquad (31.29)$$

The resulting magnification is simply the ratio of the near-point distance to the focal length of the magnifying lens, so a lens with a shorter focal length gives a stronger magnification. Although this magnification is smaller by 1 than the magnification obtained with the image at the near point, it provides for the most comfortable viewing conditions, because the eye is relaxed when viewing a distant object.

By comparing **Equation 31.26** with **Equation 31.29**, we see that the range of angular magnification of a given converging lens is

$$\frac{25\text{ cm}}{f} \leq M \leq 1 + \frac{25\text{ cm}}{f}. \quad (31.30)$$

Example 31.6

Magnifying a Diamond

A jeweler wishes to inspect a 3.0-mm-diameter diamond with a magnifier. The diamond is held at the jeweler's near point (25 cm), and the jeweler holds the magnifying lens close to his eye.

(a) What should the focal length of the magnifying lens be to see a 15-mm-diameter image of the diamond?

(b) What should the focal length of the magnifying lens be to obtain $10\times$ magnification?

Strategy

We need to determine the requisite magnification of the magnifier. Because the jeweler holds the magnifying lens close to his eye, we can use **Equation 31.27** to find the focal length of the magnifying lens.

Solution

a. The required linear magnification is the ratio of the desired image diameter to the diamond's actual diameter (**Equation 31.29**). Because the jeweler holds the magnifying lens close to his eye and the image forms at his near point, the linear magnification is the same as the angular magnification, so

$$M = m = \frac{h_i}{h_o} = \frac{15\text{ mm}}{3.0\text{ mm}} = 5.0.$$

The focal length f of the magnifying lens may be calculated by solving **Equation 31.27** for f, which gives

$$M = 1 + \frac{25\text{ cm}}{f}$$

$$f = \frac{25\text{ cm}}{M - 1} = \frac{25\text{ cm}}{5.0 - 1} = 6.3\text{ cm}$$

b. To get an image magnified by a factor of ten, we again solve **Equation 31.27** for f, but this time we use $M = 10$. The result is

$$f = \frac{25\text{ cm}}{M - 1} = \frac{25\text{ cm}}{10 - 1} = 2.8\text{ cm}.$$

Significance

Note that a greater magnification is achieved by using a lens with a smaller focal length. We thus need to use a lens with radii of curvature that are less than a few centimeters and hold it very close to our eye. This is not very convenient. A compound microscope, explored in the following section, can overcome this drawback.

31.7 | Microscopes and Telescopes

Learning Objectives

By the end of this section, you will be able to:

- Explain the physics behind the operation of microscopes and telescopes
- Describe the image created by these instruments and calculate their magnifications

Microscopes and telescopes are major instruments that have contributed hugely to our current understanding of the micro- and macroscopic worlds. The invention of these devices led to numerous discoveries in disciplines such as physics, astronomy, and biology, to name a few. In this section, we explain the basic physics that make these instruments work.

Microscopes

Although the eye is marvelous in its ability to see objects large and small, it obviously is limited in the smallest details it can detect. The desire to see beyond what is possible with the naked eye led to the use of optical instruments. We have seen that a simple convex lens can create a magnified image, but it is hard to get large magnification with such a lens. A magnification greater than $5\times$ is difficult without distorting the image. To get higher magnification, we can combine the simple magnifying glass with one or more additional lenses. In this section, we examine microscopes that enlarge the details that we cannot see with the naked eye.

Microscopes were first developed in the early 1600s by eyeglass makers in The Netherlands and Denmark. The simplest **compound microscope** is constructed from two convex lenses (Figure 31.33). The **objective** lens is a convex lens of short focal length (i.e., high power) with typical magnification from $5\times$ to $100\times$. The **eyepiece**, also referred to as the ocular, is a convex lens of longer focal length.

The purpose of a microscope is to create magnified images of small objects, and both lenses contribute to the final magnification. Also, the final enlarged image is produced sufficiently far from the observer to be easily viewed, since the eye cannot focus on objects or images that are too close (i.e., closer than the near point of the eye).

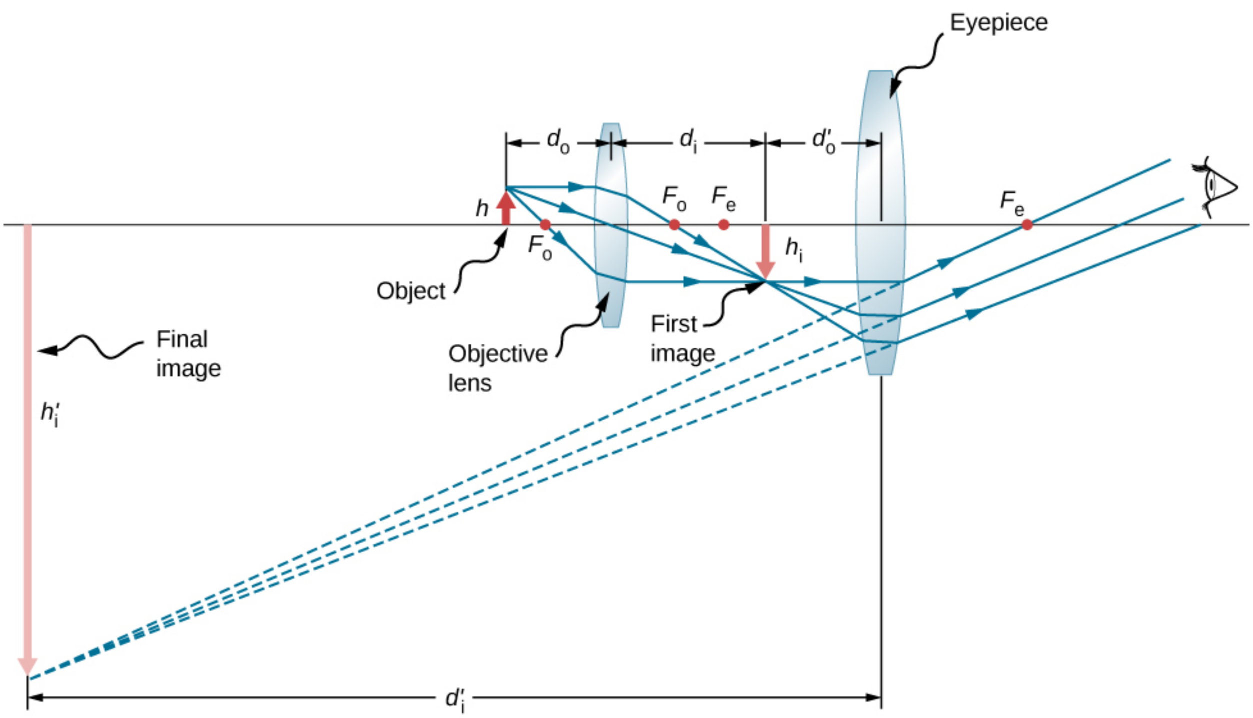


Figure 31.33 A compound microscope is composed of two lenses: an objective and an eyepiece. The objective forms the first image, which is larger than the object. This first image is inside the focal length of the eyepiece and serves as the object for the eyepiece. The eyepiece forms final image that is further magnified.

To see how the microscope in Figure 31.33 forms an image, consider its two lenses in succession. The object is just beyond the focal length f^{obj} of the objective lens, producing a real, inverted image that is larger than the object. This first image serves as the object for the second lens, or eyepiece. The eyepiece is positioned so that the first image is within its focal length f^{eye}, so that it can further magnify the image. In a sense, it acts as a magnifying glass that magnifies the intermediate image produced by the objective. The image produced by the eyepiece is a magnified virtual image. The final image remains inverted but is farther from the observer than the object, making it easy to view.

The eye views the virtual image created by the eyepiece, which serves as the object for the lens in the eye. The virtual image formed by the eyepiece is well outside the focal length of the eye, so the eye forms a real image on the retina.

The magnification of the microscope is the product of the linear magnification m^{obj} by the objective and the angular magnification M^{eye} by the eyepiece. These are given by

$$m^{\text{obj}} = -\frac{d_{\text{i}}^{\text{obj}}}{d_{\text{o}}^{\text{obj}}} \approx -\frac{d_{\text{i}}^{\text{obj}}}{f^{\text{obj}}} \text{ (linear magnification by objective)}$$

$$M^{\text{eye}} = 1 + \frac{25\text{ cm}}{f^{\text{eye}}} \text{ (angular magnification by eyepiece)}$$

Here, f^{obj} and f^{eye} are the focal lengths of the objective and the eyepiece, respectively. We assume that the final image is formed at the near point of the eye, providing the largest magnification. Note that the angular magnification of the eyepiece is the same as obtained earlier for the simple magnifying glass. This should not be surprising, because the eyepiece is essentially a magnifying glass, and the same physics applies here. The **net magnification** M_{net} of the compound microscope is the product of the linear magnification of the objective and the angular magnification of the eyepiece:

$$M_{\text{net}} = m^{\text{obj}} M^{\text{eye}} = -\frac{d_{\text{i}}^{\text{obj}}\left(f^{\text{eye}} + 25\text{ cm}\right)}{f^{\text{obj}} f^{\text{eye}}}. \tag{31.31}$$

Example 31.7

Microscope Magnification

Calculate the magnification of an object placed 6.20 mm from a compound microscope that has a 6.00 mm-focal length objective and a 50.0 mm-focal length eyepiece. The objective and eyepiece are separated by 23.0 cm.

Strategy

This situation is similar to that shown in **Figure 31.33**. To find the overall magnification, we must know the linear magnification of the objective and the angular magnification of the eyepiece. We can use **Equation 31.31**, but we need to use the thin-lens equation to find the image distance $d_{\text{i}}^{\text{obj}}$ of the objective.

Solution

Solving the thin-lens equation for $d_{\text{i}}^{\text{obj}}$ gives

$$\begin{aligned} d_{\text{i}}^{\text{obj}} &= \left(\frac{1}{f^{\text{obj}}} - \frac{1}{d_{\text{o}}^{\text{obj}}}\right)^{-1} \\ &= \left(\frac{1}{6.00\text{ mm}} - \frac{1}{6.20\text{ mm}}\right)^{-1} = 186\text{ mm} = 18.6\text{ cm} \end{aligned}$$

Inserting this result into **Equation 31.31** along with the known values $f^{\text{obj}} = 6.00\text{ mm} = 0.600\text{ cm}$ and $f^{\text{eye}} = 50.0\text{ mm} = 5.00\text{ cm}$ gives

$$\begin{aligned} M_{\text{net}} &= -\frac{d_{\text{i}}^{\text{obj}}(f^{\text{eye}} + 25\text{ cm})}{f^{\text{obj}} f^{\text{eye}}} \\ &= -\frac{(18.6\text{ cm})(5.00\text{ cm} + 25\text{ cm})}{(0.600\text{ cm})(5.00\text{ cm})} \\ &= -186 \end{aligned}$$

Significance

Both the objective and the eyepiece contribute to the overall magnification, which is large and negative, consistent with **Figure 31.33**, where the image is seen to be large and inverted. In this case, the image is virtual and inverted, which cannot happen for a single element (see **Figure 31.26**).

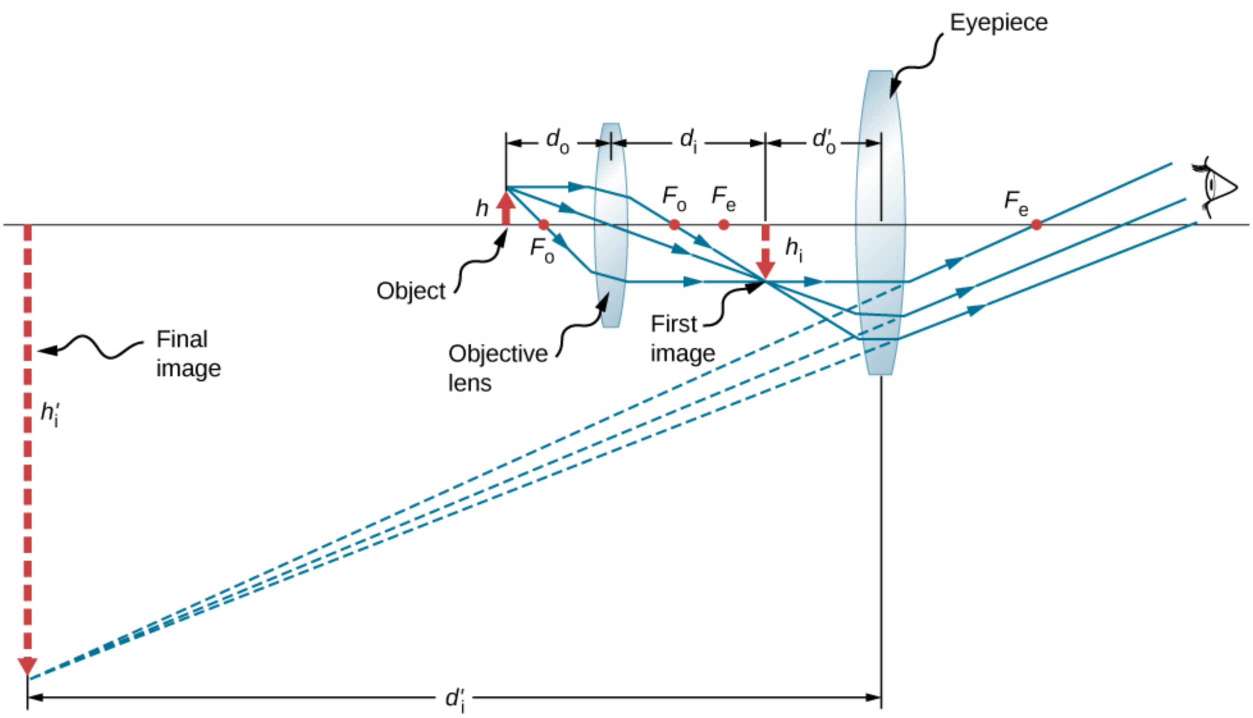


Figure 31.34 A compound microscope with the image created at infinity.

We now calculate the magnifying power of a microscope when the image is at infinity, as shown in **Figure 31.34**, because this makes for the most relaxed viewing. The magnifying power of the microscope is the product of linear magnification m^{obj} of the objective and the angular magnification M^{eye} of the eyepiece. We know that $m^{\text{obj}} = -d_i^{\text{obj}}/d_o^{\text{obj}}$ and from the thin-lens equation we obtain

$$m^{\text{obj}} = -\frac{d_i^{\text{obj}}}{d_o^{\text{obj}}} = 1 - \frac{d_i^{\text{obj}}}{f^{\text{obj}}} = \frac{f^{\text{obj}} - d_i^{\text{obj}}}{f^{\text{obj}}}. \tag{31.32}$$

If the final image is at infinity, then the image created by the objective must be located at the focal point of the eyepiece. This may be seen by considering the thin-lens equation with $d_i = \infty$ or by recalling that rays that pass through the focal point exit the lens parallel to each other, which is equivalent to focusing at infinity. For many microscopes, the distance between the image-side focal point of the objective and the object-side focal point of the eyepiece is standardized at $L = 16\,\text{cm}$. This distance is called the tube length of the microscope. From **Figure 31.34**, we see that $L = f^{\text{obj}} - d_i^{\text{obj}}$. Inserting this into **Equation 31.32** gives

$$m^{\text{obj}} = \frac{L}{f^{\text{obj}}} = \frac{16\,\text{cm}}{f^{\text{obj}}}. \tag{31.33}$$

We now need to calculate the angular magnification of the eyepiece with the image at infinity. To do so, we take the ratio of the angle θ_{image} subtended by the image to the angle θ_{object} subtended by the object at the near point of the eye (this is the closest that the unaided eye can view the object, and thus this is the position where the object will form the largest image on the retina of the unaided eye). Using **Figure 31.34** and working in the small-angle approximation, we have $\theta_{\text{image}} \approx h_i^{\text{obj}}/f^{\text{eye}}$ and $\theta_{\text{object}} \approx h_i^{\text{obj}}/25\,\text{cm}$, where h_i^{obj} is the height of the image formed by the objective, which is the object of the eyepiece. Thus, the angular magnification of the eyepiece is

$$M^{\text{eye}} = \frac{\theta_{\text{image}}}{\theta_{\text{object}}} = \frac{h_i^{\text{obj}}}{f^{\text{eye}}}\frac{25\,\text{cm}}{h_i^{\text{obj}}} = \frac{25\,\text{cm}}{f^{\text{eye}}}. \tag{31.34}$$

The net magnifying power of the compound microscope with the image at infinity is therefore

$$M_{\text{net}} = m^{\text{obj}} M^{\text{eye}} = -\frac{(16\text{ cm})(25\text{ cm})}{f^{\text{obj}} f^{\text{eye}}}. \tag{31.35}$$

The focal distances must be in centimeters. The minus sign indicates that the final image is inverted. Note that the only variables in the equation are the focal distances of the eyepiece and the objective, which makes this equation particularly useful.

Telescopes

Telescopes are meant for viewing distant objects and produce an image that is larger than the image produced in the unaided eye. Telescopes gather far more light than the eye, allowing dim objects to be observed with greater magnification and better resolution. Telescopes were invented around 1600, and Galileo was the first to use them to study the heavens, with monumental consequences. He observed the moons of Jupiter, the craters and mountains on the moon, the details of sunspots, and the fact that the Milky Way is composed of a vast number of individual stars.

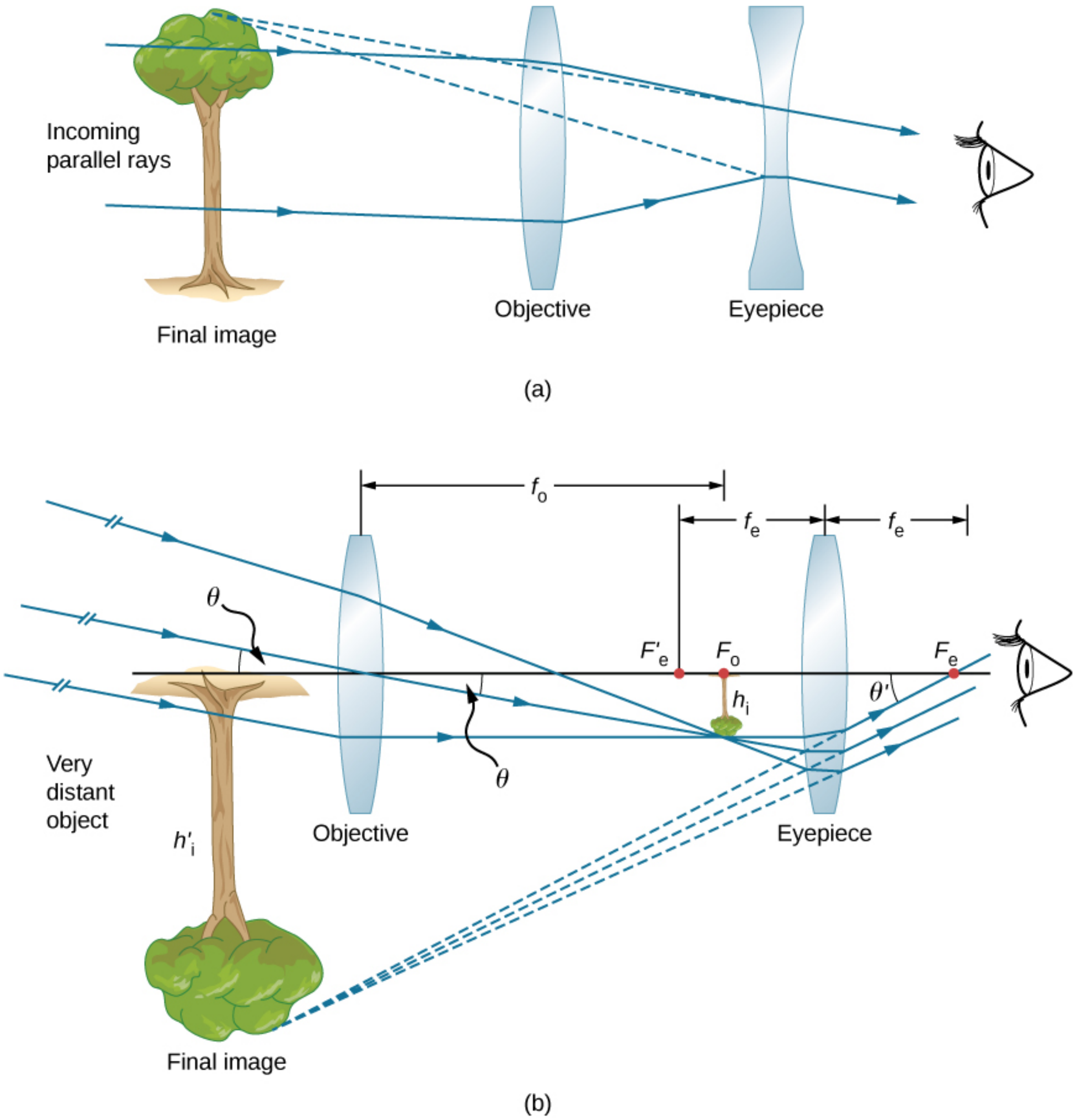


Figure 31.35 (a) Galileo made telescopes with a convex objective and a concave eyepiece. These produce an upright image and are used in spyglasses. (b) Most simple refracting telescopes have two convex lenses. The objective forms a real, inverted image at (or just within) the focal plane of the eyepiece. This image serves as the object for the eyepiece. The eyepiece forms a virtual, inverted image that is magnified.

Part (a) of **Figure 31.35** shows a refracting telescope made of two lenses. The first lens, called the objective, forms a real image within the focal length of the second lens, which is called the eyepiece. The image of the objective lens serves as the object for the eyepiece, which forms a magnified virtual image that is observed by the eye. This design is what Galileo used to observe the heavens.

Although the arrangement of the lenses in a refracting telescope looks similar to that in a microscope, there are important differences. In a telescope, the real object is far away and the intermediate image is smaller than the object. In a microscope, the real object is very close and the intermediate image is larger than the object. In both the telescope and the microscope, the eyepiece magnifies the intermediate image; in the telescope, however, this is the only magnification.

The most common two-lens telescope is shown in part (b) of the figure. The object is so far from the telescope that it is essentially at infinity compared with the focal lengths of the lenses $(d_o^{obj} \approx \infty)$, so the incoming rays are essentially parallel and focus on the focal plane. Thus, the first image is produced at $d_i^{obj} = f^{obj}$, as shown in the figure, and is not large compared with what you might see by looking directly at the object. However, the eyepiece of the telescope eyepiece (like the microscope eyepiece) allows you to get nearer than your near point to this first image and so magnifies it (because you are near to it, it subtends a larger angle from your eye and so forms a larger image on your retina). As for a simple magnifier, the angular magnification of a telescope is the ratio of the angle subtended by the image [θ_{image} in part (b)] to the angle subtended by the real object [θ_{object} in part (b)]:

$$M = \frac{\theta_{image}}{\theta_{object}}. \tag{31.36}$$

To obtain an expression for the magnification that involves only the lens parameters, note that the focal plane of the objective lens lies very close to the focal plan of the eyepiece. If we assume that these planes are superposed, we have the situation shown in **Figure 31.36**.

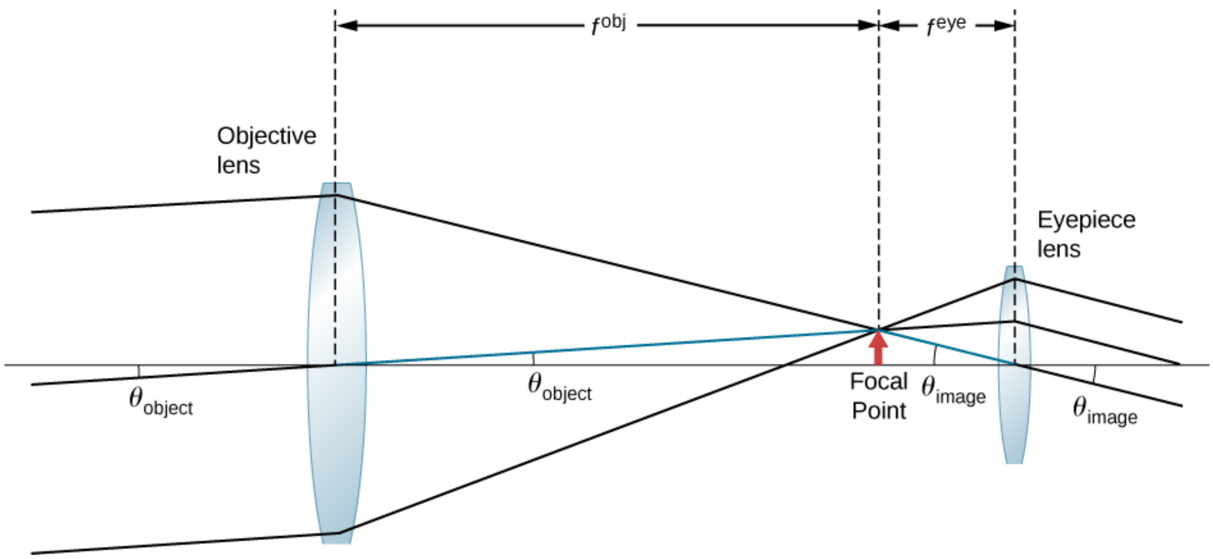


Figure 31.36 The focal plane of the objective lens of a telescope is very near to the focal plane of the eyepiece. The angle θ_{image} subtended by the image viewed through the eyepiece is larger than the angle θ_{object} subtended by the object when viewed with the unaided eye.

We further assume that the angles θ_{object} and θ_{image} are small, so that the small-angle approximation holds ($\tan\theta \approx \theta$). If the image formed at the focal plane has height *h*, then

$$\theta_{object} \approx \tan\theta_{object} = \frac{h}{f^{obj}}$$

$$\theta_{image} \approx \tan\theta_{image} = \frac{-h}{f^{eye}}$$

where the minus sign is introduced because the height is negative if we measure both angles in the counterclockwise direction. Inserting these expressions into **Equation 31.36** gives

$$M = \frac{-h_i}{f^{eye}}\frac{f^{obj}}{h_i} = -\frac{f^{obj}}{f^{eye}}. \quad (31.37)$$

Thus, to obtain the greatest angular magnification, it is best to have an objective with a long focal length and an eyepiece with a short focal length. The greater the angular magnification *M*, the larger an object will appear when viewed through a telescope, making more details visible. Limits to observable details are imposed by many factors, including lens quality and atmospheric disturbance. Typical eyepieces have focal lengths of 2.5 cm or 1.25 cm. If the objective of the telescope has a focal length of 1 meter, then these eyepieces result in magnifications of $40\times$ and $80\times$, respectively. Thus, the angular magnifications make the image appear 40 times or 80 times closer than the real object.

The minus sign in the magnification indicates the image is inverted, which is unimportant for observing the stars but is a real problem for other applications, such as telescopes on ships or telescopic gun sights. If an upright image is needed, Galileo's arrangement in part (a) of **Figure 31.35** can be used. But a more common arrangement is to use a third convex lens as an eyepiece, increasing the distance between the first two and inverting the image once again, as seen in **Figure 31.37**.

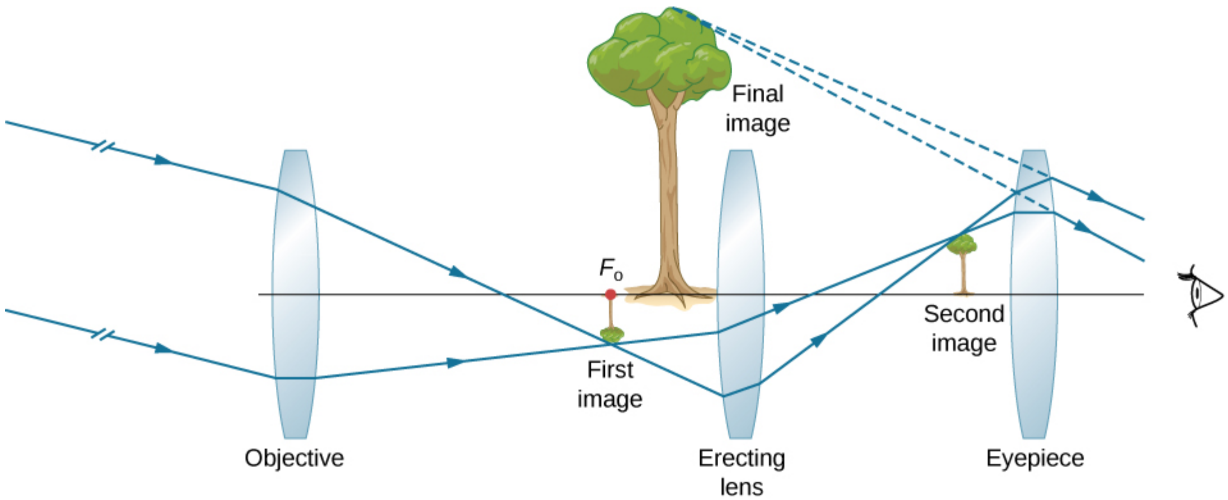


Figure 31.37 This arrangement of three lenses in a telescope produces an upright final image. The first two lenses are far enough apart that the second lens inverts the image of the first. The third lens acts as a magnifier and keeps the image upright and in a location that is easy to view.

The largest refracting telescope in the world is the 40-inch diameter Yerkes telescope located at Lake Geneva, Wisconsin (**Figure 31.38**), and operated by the University of Chicago.

It is very difficult and expensive to build large refracting telescopes. You need large defect-free lenses, which in itself is a technically demanding task. A refracting telescope basically looks like a tube with a support structure to rotate it in different directions. A refracting telescope suffers from several problems. The aberration of lenses causes the image to be blurred. Also, as the lenses become thicker for larger lenses, more light is absorbed, making faint stars more difficult to observe. Large lenses are also very heavy and deform under their own weight. Some of these problems with refracting telescopes are addressed by avoiding refraction for collecting light and instead using a curved mirror in its place, as devised by Isaac Newton. These telescopes are called reflecting telescopes.

Figure 31.38 In 1897, the Yerkes Observatory in Wisconsin (USA) built a large refracting telescope with an objective lens that is 40 inches in diameter and has a tube length of 62 feet. (credit: Yerkes Observatory, University of Chicago)

Reflecting Telescopes

Isaac Newton designed the first reflecting telescope around 1670 to solve the problem of chromatic aberration that happens in all refracting telescopes. In chromatic aberration, light of different colors refracts by slightly different amounts in the lens. As a result, a rainbow appears around the image and the image appears blurred. In the reflecting telescope, light rays from a distant source fall upon the surface of a concave mirror fixed at the bottom end of the tube. The use of a mirror instead of a lens eliminates chromatic aberration. The concave mirror focuses the rays on its focal plane. The design problem is how to observe the focused image. Newton used a design in which the focused light from the concave mirror was reflected to one side of the tube into an eyepiece [part (a) of **Figure 31.39**]. This arrangement is common in many amateur telescopes and is called the **Newtonian design**.

Some telescopes reflect the light back toward the middle of the concave mirror using a convex mirror. In this arrangement, the light-gathering concave mirror has a hole in the middle [part (b) of the figure]. The light then is incident on an eyepiece lens. This arrangement of the objective and eyepiece is called the **Cassegrain design**. Most big telescopes, including the Hubble space telescope, are of this design. Other arrangements are also possible. In some telescopes, a light detector is placed right at the spot where light is focused by the curved mirror.

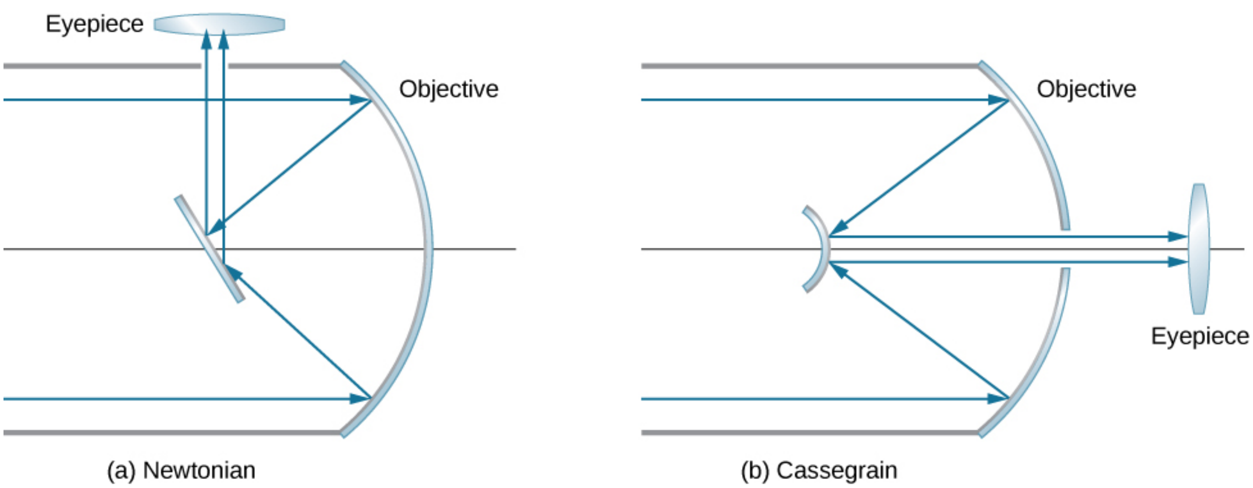


Figure 31.39 Reflecting telescopes: (a) In the Newtonian design, the eyepiece is located at the side of the telescope; (b) in the Cassegrain design, the eyepiece is located past a hole in the primary mirror.

Most astronomical research telescopes are now of the reflecting type. One of the earliest large telescopes of this kind is the Hale 200-inch (or 5-meter) telescope built on Mount Palomar in southern California, which has a 200 inch-diameter mirror. One of the largest telescopes in the world is the 10-meter Keck telescope at the Keck Observatory on the summit of the dormant Mauna Kea volcano in Hawaii. The Keck Observatory operates two 10-meter telescopes. Each is not a single mirror, but is instead made up of 36 hexagonal mirrors. Furthermore, the two telescopes on the Keck can work together, which increases their power to an effective 85-meter mirror. The Hubble telescope (**Figure 31.40**) is another large reflecting telescope with a 2.4 meter-diameter primary mirror. The Hubble was put into orbit around Earth in 1990.

Figure 31.40 The Hubble space telescope as seen from the Space Shuttle Discovery. (credit: modification of work by NASA)

The angular magnification *M* of a reflecting telescope is also given by **Equation 31.33**. For a spherical mirror, the focal length is half the radius of curvature, so making a large objective mirror not only helps the telescope collect more light but also increases the magnification of the image.

CHAPTER 31 REVIEW

KEY TERMS

aberration distortion in an image caused by departures from the small-angle approximation

angular magnification ratio of the angle subtended by an object observed with a magnifier to that observed by the naked eye

apparent depth depth at which an object is perceived to be located with respect to an interface between two media

Cassegrain design arrangement of an objective and eyepiece such that the light-gathering concave mirror has a hole in the middle, and light then is incident on an eyepiece lens

charge-coupled device (CCD) semiconductor chip that converts a light image into tiny pixels that can be converted into electronic signals of color and intensity

coma similar to spherical aberration, but arises when the incoming rays are not parallel to the optical axis

compound microscope microscope constructed from two convex lenses, the first serving as the eyepiece and the second serving as the objective lens

concave mirror spherical mirror with its reflecting surface on the inner side of the sphere; the mirror forms a “cave”

converging (or convex) lens lens in which light rays that enter it parallel converge into a single point on the opposite side

convex mirror spherical mirror with its reflecting surface on the outer side of the sphere

curved mirror mirror formed by a curved surface, such as spherical, elliptical, or parabolic

diverging (or concave) lens lens that causes light rays to bend away from its optical axis

eyepiece lens or combination of lenses in an optical instrument nearest to the eye of the observer

first focus or object focus object located at this point will result in an image created at infinity on the opposite side of a spherical interface between two media

focal length distance along the optical axis from the focal point to the optical element that focuses the light rays

focal plane plane that contains the focal point and is perpendicular to the optical axis

focal point for a converging lens or mirror, the point at which converging light rays cross; for a diverging lens or mirror, the point from which diverging light rays appear to originate

image distance distance of the image from the central axis of the optical element that produces the image

linear magnification ratio of image height to object height

magnification ratio of image size to object size

net magnification (M_{net}) of the compound microscope is the product of the linear magnification of the objective and the angular magnification of the eyepiece

Newtonian design arrangement of an objective and eyepiece such that the focused light from the concave mirror was reflected to one side of the tube into an eyepiece

object distance distance of the object from the central axis of the optical element that produces its image

objective lens nearest to the object being examined.

optical axis axis about which the mirror is rotationally symmetric; you can rotate the mirror about this axis without changing anything

plane mirror plane (flat) reflecting surface

ray tracing technique that uses geometric constructions to find and characterize the image formed by an optical system

real image image that can be projected onto a screen because the rays physically go through the image

second focus or image focus for a converging interface, the point where a bundle of parallel rays refracting at a spherical interface; for a diverging interface, the point at which the backward continuation of the refracted rays will converge between two media will focus

simple magnifier (or magnifying glass) converging lens that produces a virtual image of an object that is within the focal length of the lens

small-angle approximation approximation that is valid when the size of a spherical mirror is significantly smaller than the mirror's radius; in this approximation, spherical aberration is negligible and the mirror has a well-defined focal point

spherical aberration distortion in the image formed by a spherical mirror when rays are not all focused at the same point

thin-lens approximation assumption that the lens is very thin compared to the first image distance

vertex point where the mirror's surface intersects with the optical axis

virtual image image that cannot be projected on a screen because the rays do not physically go through the image, they only appear to originate from the image

KEY EQUATIONS

Image distance in a plane mirror	$d_o = -d_i$
Focal length for a spherical mirror	$f = \frac{R}{2}$
Mirror equation	$\frac{1}{d_o} + \frac{1}{d_i} = \frac{1}{f}$
Magnification of a spherical mirror	$m = \frac{h_i}{h_o} = -\frac{d_i}{d_o}$
Sign convention for mirrors	
Focal length f	+ for concave mirror − for convex mirror
Object distance d_o	+ for real object − for virtual object
Image distance d_i	+ for real image − for virtual image
Magnification m	+ for upright image − for inverted image
Apparent depth equation	$h_i = \left(\frac{n_2}{n_1}\right)h_o$
Spherical interface equation	$\frac{n_1}{d_o} + \frac{n_2}{d_i} = \frac{n_2 - n_1}{R}$
The thin-lens equation	$\frac{1}{d_o} + \frac{1}{d_i} = \frac{1}{f}$
The lens maker's equation	$\frac{1}{f} = \left(\frac{n_2}{n_1} - 1\right)\left(\frac{1}{R_1} - \frac{1}{R_2}\right)$
The magnification m of an object	$m \equiv \frac{h_i}{h_o} = -\frac{d_i}{d_o}$

Optical power	$P = \frac{1}{f}$
Optical power of thin, closely spaced lenses	$P_{\text{total}} = P_{\text{lens1}} + P_{\text{lens2}} + P_{\text{lens3}} + \cdots$
Angular magnification M of a simple magnifier	$M = \frac{\theta_{\text{image}}}{\theta_{\text{object}}}$
Angular magnification of an object a distance L from the eye for a convex lens of focal length f held a distance ℓ from the eye	$M = \left(\frac{25\text{ cm}}{L}\right)\left(1 + \frac{L - \ell}{f}\right)$
Range of angular magnification for a given lens for a person with a near point of 25 cm	$\frac{25\text{ cm}}{f} \le M \le 1 + \frac{25\text{ cm}}{f}$
Net magnification of compound microscope	$M_{\text{net}} = m^{\text{obj}} M^{\text{eye}} = -\frac{d_{\text{i}}^{\text{obj}}\left(f^{\text{eye}} + 25\text{ cm}\right)}{f^{\text{obj}} f^{\text{eye}}}$

SUMMARY

31.1 Images Formed by Plane Mirrors

- A plane mirror always forms a virtual image (behind the mirror).
- The image and object are the same distance from a flat mirror, the image size is the same as the object size, and the image is upright.

31.2 Spherical Mirrors

- Spherical mirrors may be concave (converging) or convex (diverging).
- The focal length of a spherical mirror is one-half of its radius of curvature: $f = R/2$.
- The mirror equation and ray tracing allow you to give a complete description of an image formed by a spherical mirror.
- Spherical aberration occurs for spherical mirrors but not parabolic mirrors; comatic aberration occurs for both types of mirrors.

31.3 Images Formed by Refraction

This section explains how a single refracting interface forms images.

- When an object is observed through a plane interface between two media, then it appears at an apparent distance h_{i} that differs from the actual distance h_{o}: $h_{\text{i}} = (n_2/n_1)h_{\text{o}}$.
- An image is formed by the refraction of light at a spherical interface between two media of indices of refraction n_1 and n_2.
- Image distance depends on the radius of curvature of the interface, location of the object, and the indices of refraction of the media.

31.4 Thin Lenses

- Two types of lenses are possible: converging and diverging. A lens that causes light rays to bend toward (away from) its optical axis is a converging (diverging) lens.
- For a converging lens, the focal point is where the converging light rays cross; for a diverging lens, the focal point is the point from which the diverging light rays appear to originate.
- The distance from the center of a thin lens to its focal point is called the focal length *f*.

- Ray tracing is a geometric technique to determine the paths taken by light rays through thin lenses.
- A real image can be projected onto a screen.
- A virtual image cannot be projected onto a screen.
- A converging lens forms either real or virtual images, depending on the object location; a diverging lens forms only virtual images.

31.5 The Camera

- Cameras use combinations of lenses to create an image for recording.
- Digital photography is based on charge-coupled devices (CCDs) that break an image into tiny "pixels" that can be converted into electronic signals.

31.6 The Simple Magnifier

- A simple magnifier is a converging lens and produces a magnified virtual image of an object located within the focal length of the lens.
- Angular magnification accounts for magnification of an image created by a magnifier. It is equal to the ratio of the angle subtended by the image to that subtended by the object when the object is observed by the unaided eye.
- Angular magnification is greater for magnifying lenses with smaller focal lengths.
- Simple magnifiers can produce as great as tenfold ($10\times$) magnification.

31.7 Microscopes and Telescopes

- Many optical devices contain more than a single lens or mirror. These are analyzed by considering each element sequentially. The image formed by the first is the object for the second, and so on. The same ray-tracing and thin-lens techniques developed in the previous sections apply to each lens element.
- The overall magnification of a multiple-element system is the product of the linear magnifications of its individual elements times the angular magnification of the eyepiece. For a two-element system with an objective and an eyepiece, this is

  $$M = m^{\text{obj}} M^{\text{eye}}.$$

  where m^{obj} is the linear magnification of the objective and M^{eye} is the angular magnification of the eyepiece.
- The microscope is a multiple-element system that contains more than a single lens or mirror. It allows us to see detail that we could not to see with the unaided eye. Both the eyepiece and objective contribute to the magnification. The magnification of a compound microscope with the image at infinity is

  $$M_{\text{net}} = -\frac{(16\text{ cm})(25\text{ cm})}{f^{\text{obj}} f^{\text{eye}}}.$$

  In this equation, 16 cm is the standardized distance between the image-side focal point of the objective lens and the object-side focal point of the eyepiece, 25 cm is the normal near point distance, f^{obj} and f^{eye} are the focal distances for the objective lens and the eyepiece, respectively.
- Simple telescopes can be made with two lenses. They are used for viewing objects at large distances.
- The angular magnification *M* for a telescope is given by

  $$M = -\frac{f^{\text{obj}}}{f^{\text{eye}}},$$

  where f^{obj} and f^{eye} are the focal lengths of the objective lens and the eyepiece, respectively.

CONCEPTUAL QUESTIONS

31.1 Images Formed by Plane Mirrors

1. What are the differences between real and virtual images? How can you tell (by looking) whether an image formed by a single lens or mirror is real or virtual?

2. Can you see a virtual image? Explain your response.

3. Can you photograph a virtual image?

4. Can you project a virtual image onto a screen?

5. Is it necessary to project a real image onto a screen to see it?

6. Devise an arrangement of mirrors allowing you to see the back of your head. What is the minimum number of mirrors needed for this task?

7. If you wish to see your entire body in a flat mirror (from head to toe), how tall should the mirror be? Does its size depend upon your distance away from the mirror? Provide a sketch.

31.2 Spherical Mirrors

8. At what distance is an image always located: at d_o, d_i, or f?

9. Under what circumstances will an image be located at the focal point of a spherical lens or mirror?

10. What is meant by a negative magnification? What is meant by a magnification whose absolute value is less than one?

11. Can an image be larger than the object even though its magnification is negative? Explain.

31.3 Images Formed by Refraction

12. Derive the formula for the apparent depth of a fish in a fish tank using Snell's law.

13. Use a ruler and a protractor to find the image by refraction in the following cases. Assume an air-glass interface. Use a refractive index of 1 for air and of 1.5 for glass. (*Hint*: Use Snell's law at the interface.)

(a) A point object located on the axis of a concave interface located at a point within the focal length from the vertex.

(b) A point object located on the axis of a concave interface located at a point farther than the focal length from the vertex.

(c) A point object located on the axis of a convex interface located at a point within the focal length from the vertex.

(d) A point object located on the axis of a convex interface located at a point farther than the focal length from the vertex.

(e) Repeat (a)–(d) for a point object off the axis.

31.4 Thin Lenses

14. You can argue that a flat piece of glass, such as in a window, is like a lens with an infinite focal length. If so, where does it form an image? That is, how are d_i and d_o related?

15. When you focus a camera, you adjust the distance of the lens from the film. If the camera lens acts like a thin lens, why can it not be a fixed distance from the film for both near and distant objects?

16. A thin lens has two focal points, one on either side of the lens at equal distances from its center, and should behave the same for light entering from either side. Look backward and forward through a pair of eyeglasses and comment on whether they are thin lenses.

17. Will the focal length of a lens change when it is submerged in water? Explain.

31.7 Microscopes and Telescopes

18. Geometric optics describes the interaction of light with macroscopic objects. Why, then, is it correct to use geometric optics to analyze a microscope's image?

19. The image produced by the microscope in **Figure 31.33** cannot be projected. Could extra lenses or mirrors project it? Explain.

20. If you want your microscope or telescope to project a real image onto a screen, how would you change the placement of the eyepiece relative to the objective?

PROBLEMS

31.1 Images Formed by Plane Mirrors

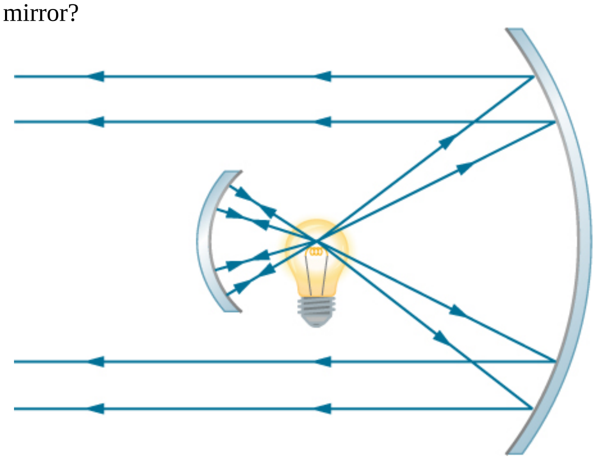
21. Consider a pair of flat mirrors that are positioned so that they form an angle of $120°$. An object is placed on the bisector between the mirrors. Construct a ray diagram as in **Figure 31.4** to show how many images are formed.

22. Consider a pair of flat mirrors that are positioned so that they form an angle of $60°$. An object is placed on the bisector between the mirrors. Construct a ray diagram as in **Figure 31.4** to show how many images are formed.

23. By using more than one flat mirror, construct a ray diagram showing how to create an inverted image.

31.2 Spherical Mirrors

24. The following figure shows a light bulb between two spherical mirrors. One mirror produces a beam of light with parallel rays; the other keeps light from escaping without being put into the beam. Where is the filament of the light in relation to the focal point or radius of curvature of each mirror?



25. Why are diverging mirrors often used for rearview mirrors in vehicles? What is the main disadvantage of using such a mirror compared with a flat one?

26. Some telephoto cameras use a mirror rather than a lens. What radius of curvature mirror is needed to replace a 800 mm-focal length telephoto lens?

27. Calculate the focal length of a mirror formed by the shiny back of a spoon that has a 3.00 cm radius of curvature.

28. Electric room heaters use a concave mirror to reflect infrared (IR) radiation from hot coils. Note that IR radiation follows the same law of reflection as visible light. Given that the mirror has a radius of curvature of 50.0 cm and produces an image of the coils 3.00 m away from the mirror, where are the coils?

29. Find the magnification of the heater element in the previous problem. Note that its large magnitude helps spread out the reflected energy.

30. What is the focal length of a makeup mirror that produces a magnification of 1.50 when a person's face is 12.0 cm away? Explicitly show how you follow the steps in the **Problem-Solving Strategy: Spherical Mirrors**.

31. A shopper standing 3.00 m from a convex security mirror sees his image with a magnification of 0.250. (a) Where is his image? (b) What is the focal length of the mirror? (c) What is its radius of curvature?

32. An object 1.50 cm high is held 3.00 cm from a person's cornea, and its reflected image is measured to be 0.167 cm high. (a) What is the magnification? (b) Where is the image? (c) Find the radius of curvature of the convex mirror formed by the cornea. (Note that this technique is used by optometrists to measure the curvature of the cornea for contact lens fitting. The instrument used is called a keratometer, or curve measurer.)

33. Ray tracing for a flat mirror shows that the image is located a distance behind the mirror equal to the distance of the object from the mirror. This is stated as $d_\text{i} = -d_\text{o}$, since this is a negative image distance (it is a virtual image). What is the focal length of a flat mirror?

34. Show that, for a flat mirror, $h_\text{i} = h_\text{o}$, given that the image is the same distance behind the mirror as the distance of the object from the mirror.

35. Use the law of reflection to prove that the focal length of a mirror is half its radius of curvature. That is, prove that $f = R/2$. Note this is true for a spherical mirror only if its diameter is small compared with its radius of curvature.

36. Referring to the electric room heater considered in problem 5, calculate the intensity of IR radiation in W/m^2 projected by the concave mirror on a person 3.00 m away. Assume that the heating element radiates 1500 W and has an area of $100\ \text{cm}^2$, and that half of the radiated power is reflected and focused by the mirror.

37. Two mirrors are inclined at an angle of $60°$ and an object is placed at a point that is equidistant from the two mirrors. Use a protractor to draw rays accurately and locate all images. You may have to draw several figures so that

that rays for different images do not clutter your drawing.

38. Two parallel mirrors are facing each other and are separated by a distance of 3 cm. A point object is placed between the mirrors 1 cm from one of the mirrors. Find the coordinates of all the images.

31.3 Images Formed by Refraction

39. An object is located in air 30 cm from the vertex of a concave surface made of glass with a radius of curvature 10 cm. Where does the image by refraction form and what is its magnification? Use $n_{\text{air}} = 1$ and $n_{\text{glass}} = 1.5$.

40. An object is located in air 30 cm from the vertex of a convex surface made of glass with a radius of curvature 80 cm. Where does the image by refraction form and what is its magnification?

41. An object is located in water 15 cm from the vertex of a concave surface made of glass with a radius of curvature 10 cm. Where does the image by refraction form and what is its magnification? Use $n_{\text{water}} = 4/3$ and $n_{\text{glass}} = 1.5$.

42. An object is located in water 30 cm from the vertex of a convex surface made of Plexiglas with a radius of curvature of 80 cm. Where does the image form by refraction and what is its magnification? $n_{\text{water}} = 4/3$ and $n_{\text{Plexiglas}} = 1.65$.

43. An object is located in air 5 cm from the vertex of a concave surface made of glass with a radius of curvature 20 cm. Where does the image form by refraction and what is its magnification? Use $n_{\text{air}} = 1$ and $n_{\text{glass}} = 1.5$.

44. Derive the spherical interface equation for refraction at a concave surface. (*Hint*: Follow the derivation in the text for the convex surface.)

31.4 Thin Lenses

45. How far from the lens must the film in a camera be, if the lens has a 35.0-mm focal length and is being used to photograph a flower 75.0 cm away? Explicitly show how you follow the steps in the **Problem-Solving Strategy: Lenses**.

46. A certain slide projector has a 100 mm-focal length lens. (a) How far away is the screen if a slide is placed 103 mm from the lens and produces a sharp image? (b) If the slide is 24.0 by 36.0 mm, what are the dimensions of the image? Explicitly show how you follow the steps in the **Problem-Solving Strategy: Lenses**.

47. A doctor examines a mole with a 15.0-cm focal length magnifying glass held 13.5 cm from the mole. (a) Where is the image? (b) What is its magnification? (c) How big is the image of a 5.00 mm diameter mole?

48. A camera with a 50.0-mm focal length lens is being used to photograph a person standing 3.00 m away. (a) How far from the lens must the film be? (b) If the film is 36.0 mm high, what fraction of a 1.75-m-tall person will fit on it? (c) Discuss how reasonable this seems, based on your experience in taking or posing for photographs.

49. A camera lens used for taking close-up photographs has a focal length of 22.0 mm. The farthest it can be placed from the film is 33.0 mm. (a) What is the closest object that can be photographed? (b) What is the magnification of this closest object?

50. Suppose your 50.0 mm-focal length camera lens is 51.0 mm away from the film in the camera. (a) How far away is an object that is in focus? (b) What is the height of the object if its image is 2.00 cm high?

51. What is the focal length of a magnifying glass that produces a magnification of 3.00 when held 5.00 cm from an object, such as a rare coin?

52. The magnification of a book held 7.50 cm from a 10.0 cm-focal length lens is 3.00. (a) Find the magnification for the book when it is held 8.50 cm from the magnifier. (b) Repeat for the book held 9.50 cm from the magnifier. (c) Comment on how magnification changes as the object distance increases as in these two calculations.

53. Suppose a 200 mm-focal length telephoto lens is being used to photograph mountains 10.0 km away. (a) Where is the image? (b) What is the height of the image of a 1000 m high cliff on one of the mountains?

54. A camera with a 100 mm-focal length lens is used to photograph the sun. What is the height of the image of the sun on the film, given the sun is 1.40×10^6 km in diameter and is 1.50×10^8 km away?

55. Use the thin-lens equation to show that the magnification for a thin lens is determined by its focal length and the object distance and is given by $m = f/(f - d_{\text{o}})$.

56. An object of height 3.0 cm is placed 5.0 cm in front of a converging lens of focal length 20 cm and observed from the other side. Where and how large is the image?

57. An object of height 3.0 cm is placed at 5.0 cm in front of a diverging lens of focal length 20 cm and observed from the other side. Where and how large is the image?

58. An object of height 3.0 cm is placed at 25 cm in front of a diverging lens of focal length 20 cm. Behind the diverging lens, there is a converging lens of focal length 20 cm. The distance between the lenses is 5.0 cm. Find the location and size of the final image.

59. Two convex lenses of focal lengths 20 cm and 10 cm are placed 30 cm apart, with the lens with the longer focal length on the right. An object of height 2.0 cm is placed midway between them and observed through each lens from the left and from the right. Describe what you will see, such as where the image(s) will appear, whether they will be upright or inverted and their magnifications.

31.6 The Simple Magnifier

60. If the image formed on the retina subtends an angle of $30°$ and the object subtends an angle of $5°$, what is the magnification of the image?

61. What is the magnification of a magnifying lens with a focal length of 10 cm if it is held 3.0 cm from the eye and the object is 12 cm from the eye?

62. How far should you hold a 2.1 cm-focal length magnifying glass from an object to obtain a magnification of $10\times$? Assume you place your eye 5.0 cm from the magnifying glass.

63. You hold a 5.0 cm-focal length magnifying glass as close as possible to your eye. If you have a normal near point, what is the magnification?

64. You view a mountain with a magnifying glass of focal length $f = 10\,\text{cm}$. What is the magnification?

65. You view an object by holding a 2.5 cm-focal length magnifying glass 10 cm away from it. How far from your eye should you hold the magnifying glass to obtain a magnification of $10\times$?

66. A magnifying glass forms an image 10 cm on the opposite side of the lens from the object, which is 10 cm away. What is the magnification of this lens for a person with a normal near point if their eye 12 cm from the object?

67. An object viewed with the naked eye subtends a $2°$ angle. If you view the object through a $10\times$ magnifying glass, what angle is subtended by the image formed on your retina?

68. For a normal, relaxed eye, a magnifying glass produces an angular magnification of 4.0. What is the largest magnification possible with this magnifying glass?

69. What range of magnification is possible with a 7.0 cm-focal length converging lens?

70. A magnifying glass produces an angular magnification of 4.5 when used by a young person with a near point of 18 cm. What is the maximum angular magnification obtained by an older person with a near point of 45 cm?

31.7 Microscopes and Telescopes

71. A microscope with an overall magnification of 800 has an objective that magnifies by 200. (a) What is the angular magnification of the eyepiece? (b) If there are two other objectives that can be used, having magnifications of 100 and 400, what other total magnifications are possible?

72. (a) What magnification is produced by a 0.150 cm-focal length microscope objective that is 0.155 cm from the object being viewed? (b) What is the overall magnification if an $8\times$ eyepiece (one that produces an angular magnification of 8.00) is used?

73. Where does an object need to be placed relative to a microscope for its 0.50 cm-focal length objective to produce a magnification of −400?

74. An amoeba is 0.305 cm away from the 0.300 cm-focal length objective lens of a microscope. (a) Where is the image formed by the objective lens? (b) What is this image's magnification? (c) An eyepiece with a 2.00-cm focal length is placed 20.0 cm from the objective. Where is the final image? (d) What angular magnification is produced by the eyepiece? (e) What is the overall magnification? (See **Figure 31.34**.)

75. **Unreasonable Results** Your friends show you an image through a microscope. They tell you that the microscope has an objective with a 0.500-cm focal length and an eyepiece with a 5.00-cm focal length. The resulting overall magnification is 250,000. Are these viable values for a microscope?

Unless otherwise stated, the lens-to-retina distance is 2.00 cm.

76. What is the angular magnification of a telescope that has a 100 cm-focal length objective and a 2.50 cm-focal length eyepiece?

77. Find the distance between the objective and eyepiece lenses in the telescope in the above problem needed to produce a final image very far from the observer, where vision is most relaxed. Note that a telescope is normally used to view very distant objects.

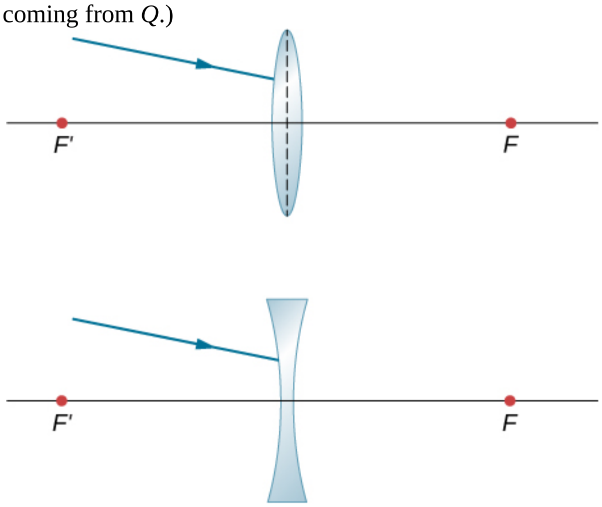
78. A large reflecting telescope has an objective mirror with a 10.0-m radius of curvature. What angular magnification does it produce when a 3.00 m-focal length eyepiece is used?

79. A small telescope has a concave mirror with a 2.00-m radius of curvature for its objective. Its eyepiece is a 4.00 cm-focal length lens. (a) What is the telescope's angular magnification? (b) What angle is subtended by a 25,000 km-diameter sunspot? (c) What is the angle of its telescopic image?

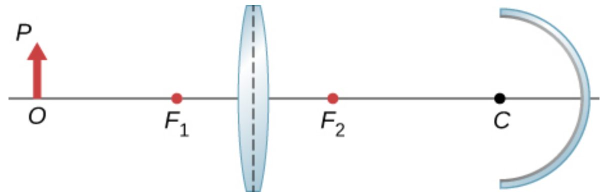
80. A $7.5\times$ binocular produces an angular magnification of −7.50, acting like a telescope. (Mirrors are used to make the image upright.) If the binoculars have objective lenses with a 75.0-cm focal length, what is the focal length of the eyepiece lenses?

81. Construct Your Own Problem Consider a telescope of the type used by Galileo, having a convex objective and a concave eyepiece as illustrated in part (a) of **Figure 31.35**. Construct a problem in which you calculate the location and size of the image produced. Among the things to be considered are the focal lengths of the lenses and their relative placements as well as the size and location of the object. Verify that the angular magnification is greater than one. That is, the angle subtended at the eye by the image is greater than the angle subtended by the object.

82. Trace rays to find which way the given ray will emerge after refraction through the thin lens in the following figure. Assume thin-lens approximation. (*Hint*: Pick a point *P* on the given ray in each case. Treat that point as an object. Now, find its image *Q*. Use the rule: All rays on the other side of the lens will either go through *Q* or appear to be coming from *Q*.)



83. Copy and draw rays to find the final image in the following diagram. (*Hint*: Find the intermediate image through lens alone. Use the intermediate image as the object for the mirror and work with the mirror alone to find the final image.)



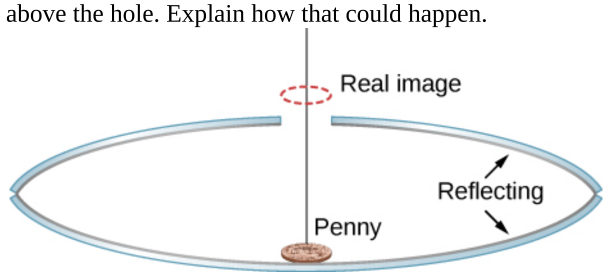
84. A concave mirror of radius of curvature 10 cm is placed 30 cm from a thin convex lens of focal length 15 cm. Find the location and magnification of a small bulb sitting 50 cm from the lens by using the algebraic method.

85. An object of height 3 cm is placed at 25 cm in front of a converging lens of focal length 20 cm. Behind the lens there is a concave mirror of focal length 20 cm. The distance between the lens and the mirror is 5 cm. Find the location, orientation and size of the final image.

86. An object of height 3 cm is placed at a distance of 25 cm in front of a converging lens of focal length 20 cm, to be referred to as the first lens. Behind the lens there is another converging lens of focal length 20 cm placed 10 cm from the first lens. There is a concave mirror of focal length 15 cm placed 50 cm from the second lens. Find the location, orientation, and size of the final image.

87. An object of height 2 cm is placed at 50 cm in front of a diverging lens of focal length 40 cm. Behind the lens, there is a convex mirror of focal length 15 cm placed 30 cm from the converging lens. Find the location, orientation, and size of the final image.

88. Two concave mirrors are placed facing each other. One of them has a small hole in the middle. A penny is placed on the bottom mirror (see the following figure). When you look from the side, a real image of the penny is observed above the hole. Explain how that could happen.



89. A lamp of height 5 cm is placed 40 cm in front of a converging lens of focal length 20 cm. There is a plane mirror 15 cm behind the lens. Where would you find the image when you look in the mirror?

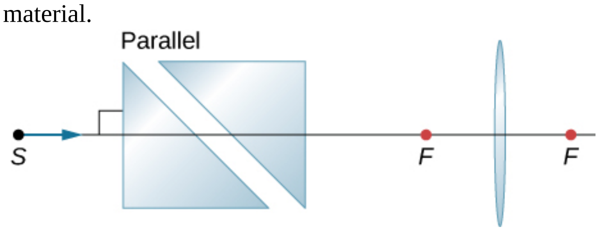
90. Parallel rays from a faraway source strike a converging lens of focal length 20 cm at an angle of 15 degrees with the horizontal direction. Find the vertical position of the real image observed on a screen in the focal plane.

91. Parallel rays from a faraway source strike a diverging lens of focal length 20 cm at an angle of 10 degrees with the horizontal direction. As you look through the lens, where in the vertical plane the image would appear?

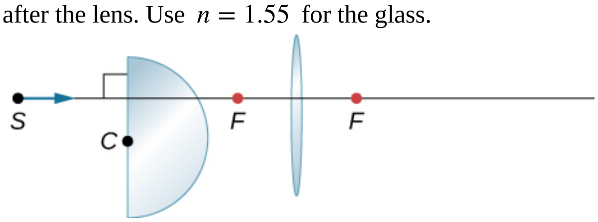
92. A light bulb is placed 10 cm from a plane mirror, which faces a convex mirror of radius of curvature 8 cm. The plane mirror is located at a distance of 30 cm from the vertex of the convex mirror. Find the location of two images in the convex mirror. Are there other images? If so, where are they located?

93. A point source of light is 50 cm in front of a converging lens of focal length 30 cm. A concave mirror with a focal length of 20 cm is placed 25 cm behind the lens. Where does the final image form, and what are its orientation and magnification?

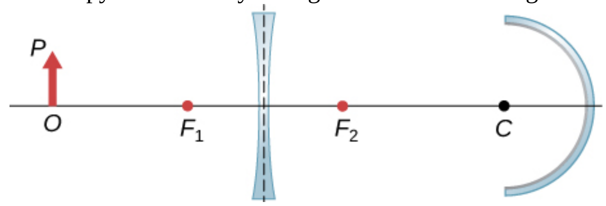
94. Copy and trace to find how a horizontal ray from *S* comes out after the lens. Use $n_{\text{glass}} = 1.5$ for the prism material.



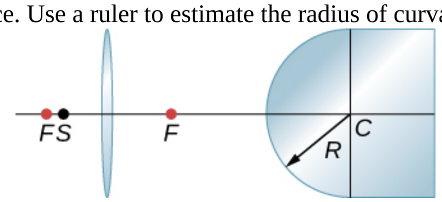
95. Copy and trace how a horizontal ray from *S* comes out after the lens. Use $n = 1.55$ for the glass.



96. Copy and draw rays to figure out the final image.



97. By ray tracing or by calculation, find the place inside the glass where rays from *S* converge as a result of refraction through the lens and the convex air-glass interface. Use a ruler to estimate the radius of curvature.



98. A diverging lens has a focal length of 20 cm. What is the power of the lens in diopters?

99. Two lenses of focal lengths of f_1 and f_2 are glued together with transparent material of negligible thickness. Show that the total power of the two lenses simply add.

100. What will be the angular magnification of a convex lens with the focal length 2.5 cm?

101. What will be the formula for the angular magnification of a convex lens of focal length *f* if the eye is very close to the lens and the near point is located a distance *D* from the eye?

ADDITIONAL PROBLEMS

102. Use a ruler and a protractor to draw rays to find images in the following cases.

(a) A point object located on the axis of a concave mirror located at a point within the focal length from the vertex.
(b) A point object located on the axis of a concave mirror located at a point farther than the focal length from the vertex.
(c) A point object located on the axis of a convex mirror located at a point within the focal length from the vertex.
(d) A point object located on the axis of a convex mirror located at a point farther than the focal length from the vertex.
(e) Repeat (a)–(d) for a point object off the axis.

103. Where should a 3 cm tall object be placed in front of a concave mirror of radius 20 cm so that its image is real and 2 cm tall?

104. A 3 cm tall object is placed 5 cm in front of a convex mirror of radius of curvature 20 cm. Where is the image formed? How tall is the image? What is the orientation of the image?

105. You are looking for a mirror so that you can see a four-fold magnified virtual image of an object when the object is placed 5 cm from the vertex of the mirror. What kind of mirror you will need? What should be the radius of curvature of the mirror?

106. Derive the following equation for a convex mirror:

$$\frac{1}{VO} - \frac{1}{VI} = -\frac{1}{VF},$$

where *VO* is the distance to the object *O* from vertex *V*, *VI* the distance to the image *I* from *V*, and *VF* is the distance to the focal point *F* from *V*. (*Hint*: use two sets of similar triangles.)

107. (a) Draw rays to form the image of a vertical object on the optical axis and farther than the focal point from a converging lens. (b) Use plane geometry in your figure and prove that the magnification *m* is given by $m = \frac{h_i}{h_o} = -\frac{d_i}{d_o}$.

108. Use another ray-tracing diagram for the same situation as given in the previous problem to derive the thin-lens equation, $\frac{1}{d_o} + \frac{1}{d_i} = \frac{1}{f}$.

109. You photograph a 2.0-m-tall person with a camera that has a 5.0 cm-focal length lens. The image on the film must be no more than 2.0 cm high. (a) What is the closest distance the person can stand to the lens? (b) For this distance, what should be the distance from the lens to the film?

110. Find the focal length of a thin plano-convex lens. The front surface of this lens is flat, and the rear surface has a radius of curvature of $R_2 = -35\ \text{cm}$. Assume that the index of refraction of the lens is 1.5.

111. Find the focal length of a meniscus lens with $R_1 = 20\ \text{cm}$ and $R_2 = 15\ \text{cm}$. Assume that the index of refraction of the lens is 1.5.

112. A nearsighted man cannot see objects clearly beyond 20 cm from his eyes. How close must he stand to a mirror in order to see what he is doing when he shaves?

113. A mother sees that her child's contact lens prescription is 0.750 D. What is the child's near point?

114. Repeat the previous problem for glasses that are 2.20 cm from the eyes.

115. The contact-lens prescription for a nearsighted person is −4.00 D and the person has a far point of 22.5 cm. What is the power of the tear layer between the cornea and the lens if the correction is ideal, taking the tear layer into account?

116. **Unreasonable Results** A boy has a near point of 50 cm and a far point of 500 cm. Will a −4.00 D lens correct his far point to infinity?

117. Find the angular magnification of an image by a magnifying glass of $f = 5.0\ \text{cm}$ if the object is placed $d_o = 4.0\ \text{cm}$ from the lens and the lens is close to the eye.

118. Let objective and eyepiece of a compound microscope have focal lengths of 2.5 cm and 10 cm, respectively and be separated by 12 cm. A $70\text{-}\mu\text{m}$ object is placed 6.0 cm from the objective. How large is the virtual image formed by the objective-eyepiece system?

119. Draw rays to scale to locate the image at the retina if the eye lens has a focal length 2.5 cm and the near point is 24 cm. (*Hint*: Place an object at the near point.)

120. The objective and the eyepiece of a microscope have the focal lengths 3 cm and 10 cm respectively. Decide about the distance between the objective and the eyepiece if we need a $10\times$ magnification from the objective/eyepiece compound system.

121. A far-sighted person has a near point of 100 cm. How far in front or behind the retina does the image of an object placed 25 cm from the eye form? Use the cornea to retina distance of 2.5 cm.

122. A near-sighted person has afar point of 80 cm. (a) What kind of corrective lens the person will need if the lens is to be placed 1.5 cm from the eye? (b) What would be the power of the contact lens needed? Assume distance to contact lens from the eye to be zero.

123. In a reflecting telescope the objective is a concave mirror of radius of curvature 2 m and an eyepiece is a convex lens of focal length 5 cm. Find the apparent size of a 25-m tree at a distance of 10 km that you would perceive when looking through the telescope.

124. Two stars that are $10^9\ \text{km}$ apart are viewed by a telescope and found to be separated by an angle of $10^{-5}\ \text{radians}$. If the eyepiece of the telescope has a focal length of 1.5 cm and the objective has a focal length of 3 meters, how far away are the stars from the observer?

125. What is the angular size of the Moon if viewed from a binocular that has a focal length of 1.2 cm for the eyepiece and a focal length of 8 cm for the objective? Use the radius of the moon $1.74 \times 10^6\ \text{m}$ and the distance of the moon from the observer to be $3.8 \times 10^8\ \text{m}$.

126. An unknown planet at a distance of $10^{12}\ \text{m}$ from Earth is observed by a telescope that has a focal length of the eyepiece of 1 cm and a focal length of the objective of 1 m. If the far away planet is seen to subtend an angle of 10^{-5} radian at the eyepiece, what is the size of the planet?

32 | PHYSICAL OPTICS: LIGHT AS WAVES

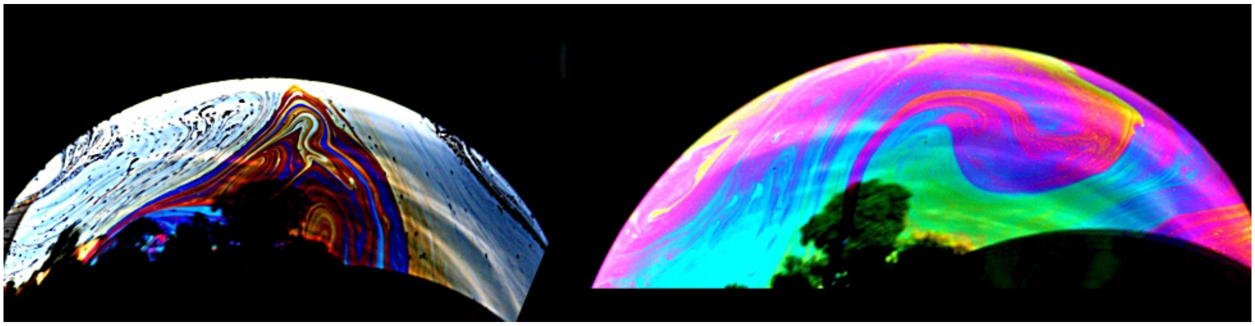


Figure 32.1 Soap bubbles are blown from clear fluid into very thin films. The colors we see are not due to any pigmentation but are the result of light interference, which enhances specific wavelengths for a given thickness of the film.

Chapter Outline

Introduction

The most certain indication of a wave is interference. This wave characteristic is most prominent when the wave interacts with an object that is not large compared with the wavelength. Interference is observed for water waves, sound waves, light waves, and, in fact, all types of waves.

If you have ever looked at the reds, blues, and greens in a sunlit soap bubble and wondered how straw-colored soapy water could produce them, you have hit upon one of the many phenomena that can only be explained by the wave character of light (see **Figure 32.1**). The same is true for the colors seen in an oil slick or in the light reflected from a DVD disc. These and other interesting phenomena cannot be explained fully by geometric optics. In these cases, light interacts with objects and exhibits wave characteristics. The branch of optics that considers the behavior of light when it exhibits wave characteristics is called wave optics (sometimes called physical optics). It is the topic of this chapter.

32.1 | Young's Double-Slit Experiment

Learning Objectives

By the end of this section, you will be able to:

- Explain the phenomenon of interference
- Define constructive and destructive interference for a double slit

The Dutch physicist Christiaan Huygens (1629–1695) thought that light was a wave, but Isaac Newton did not. Newton thought that there were other explanations for color, and for the interference and diffraction effects that were observable at the time. Owing to Newton's tremendous reputation, his view generally prevailed; the fact that Huygens's principle worked was not considered direct evidence proving that light is a wave. The acceptance of the wave character of light came many years later in 1801, when the English physicist and physician Thomas Young (1773–1829) demonstrated optical interference with his now-classic double-slit experiment.

If there were not one but two sources of waves, the waves could be made to interfere, as in the case of waves on water (**Figure 32.2**). If light is an electromagnetic wave, it must therefore exhibit interference effects under appropriate circumstances. In Young's experiment, sunlight was passed through a pinhole on a board. The emerging beam fell on two pinholes on a second board. The light emanating from the two pinholes then fell on a screen where a pattern of bright and dark spots was observed. This pattern, called fringes, can only be explained through interference, a wave phenomenon.

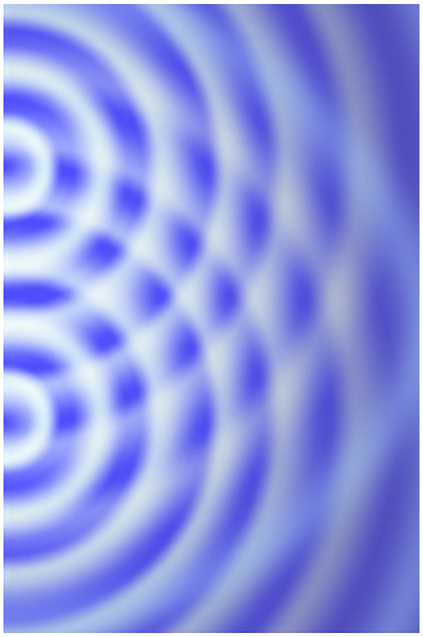


Figure 32.2 Photograph of an interference pattern produced by circular water waves in a ripple tank. Two thin plungers are vibrated up and down in phase at the surface of the water. Circular water waves are produced by and emanate from each plunger. The points where the water is calm (corresponding to destructive interference) are clearly visible.

We can analyze double-slit interference with the help of **Figure 32.3**, which depicts an apparatus analogous to Young's. Light from a monochromatic source falls on a slit S_0. The light emanating from S_0 is incident on two other slits S_1 and S_2 that are equidistant from S_0. A pattern of *interference fringes* on the screen is then produced by the light emanating from S_1 and S_2. All slits are assumed to be so narrow that they can be considered secondary point sources for Huygens' wavelets (**Huygens' Principle**). Slits S_1 and S_2 are a distance *d* apart ($d \le 1\text{ mm}$), and the distance between the screen and the slits is $D(\approx 1\text{ m})$, which is much greater than *d*.

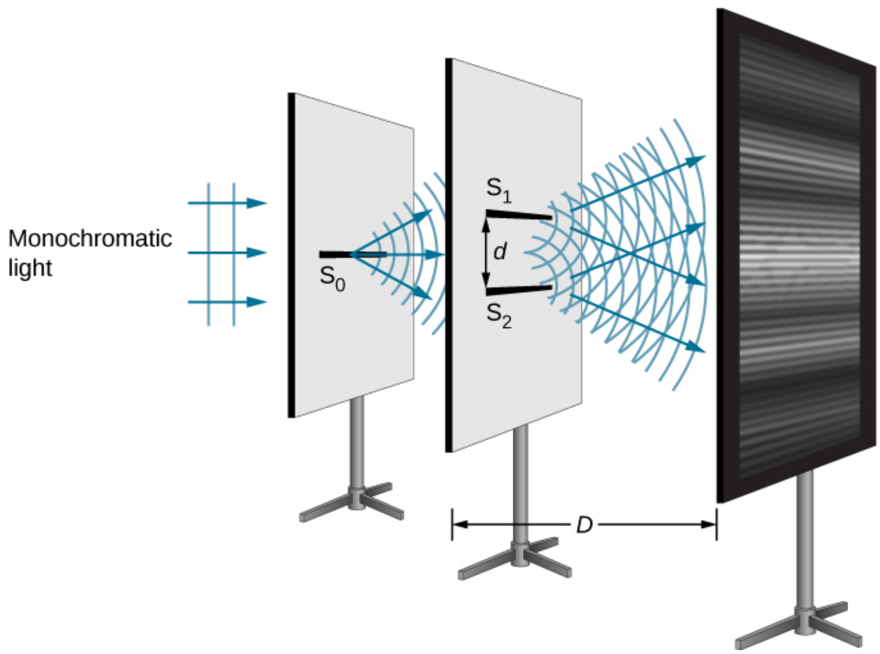


Figure 32.3 The double-slit interference experiment using monochromatic light and narrow slits. Fringes produced by interfering Huygens wavelets from slits S_1 and S_2 are observed on the screen.

Since S_0 is assumed to be a point source of monochromatic light, the secondary Huygens wavelets leaving S_1 and S_2 always maintain a constant phase difference (zero in this case because S_1 and S_2 are equidistant from S_0) and have the same frequency. The sources S_1 and S_2 are then said to be coherent. By **coherent waves**, we mean the waves are in phase or have a definite phase relationship. The term **incoherent** means the waves have random phase relationships, which would be the case if S_1 and S_2 were illuminated by two independent light sources, rather than a single source S_0. Two independent light sources (which may be two separate areas within the same lamp or the Sun) would generally not emit their light in unison, that is, not coherently. Also, because S_1 and S_2 are the same distance from S_0, the amplitudes of the two Huygens wavelets are equal.

Young used sunlight, where each wavelength forms its own pattern, making the effect more difficult to see. In the following discussion, we illustrate the double-slit experiment with **monochromatic** light (single λ) to clarify the effect. **Figure 32.4** shows the pure constructive and destructive interference of two waves having the same wavelength and amplitude.

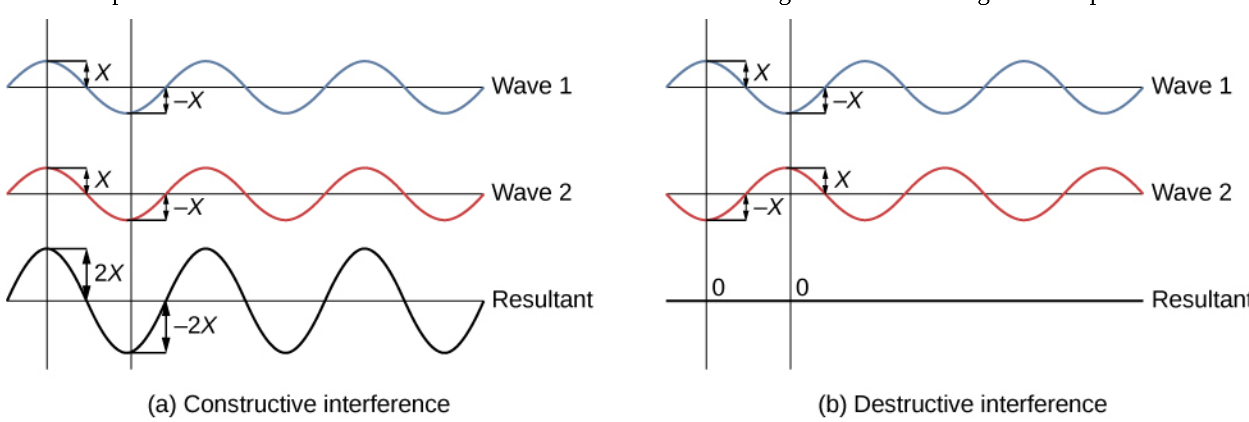


Figure 32.4 The amplitudes of waves add. (a) Pure constructive interference is obtained when identical waves are in phase. (b) Pure destructive interference occurs when identical waves are exactly out of phase, or shifted by half a wavelength.

When light passes through narrow slits, the slits act as sources of coherent waves and light spreads out as semicircular waves, as shown in **Figure 32.5**(a). Pure *constructive interference* occurs where the waves are crest to crest or trough to trough. Pure *destructive interference* occurs where they are crest to trough. The light must fall on a screen and be scattered into our eyes for us to see the pattern. An analogous pattern for water waves is shown in **Figure 32.2**. Note that regions of constructive and destructive interference move out from the slits at well-defined angles to the original beam. These angles depend on wavelength and the distance between the slits, as we shall see below.

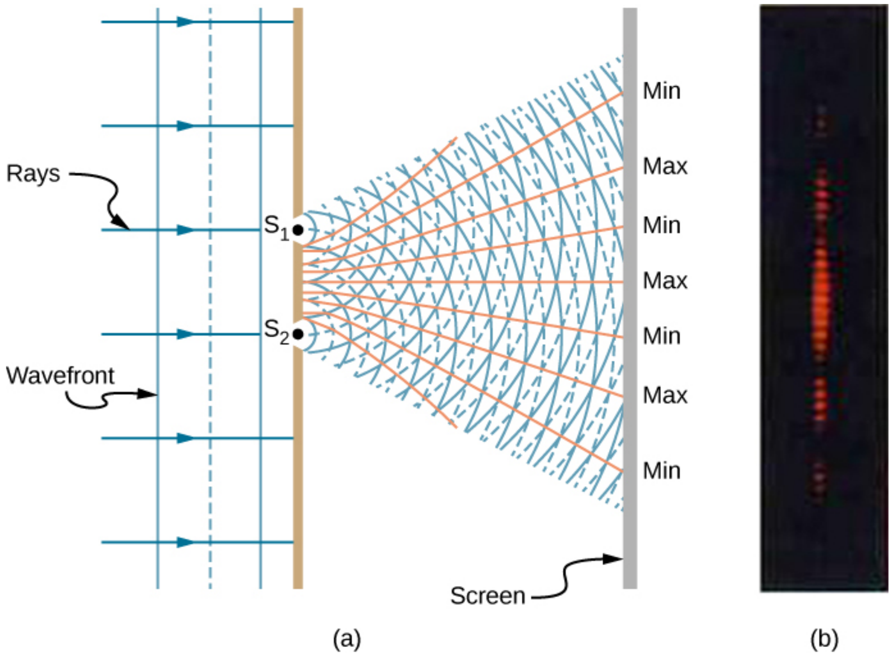


Figure 32.5 Double slits produce two coherent sources of waves that interfere. (a) Light spreads out (diffracts) from each slit, because the slits are narrow. These waves overlap and interfere constructively (bright lines) and destructively (dark regions). We can only see this if the light falls onto a screen and is scattered into our eyes. (b) When light that has passed through double slits falls on a screen, we see a pattern such as this.

To understand the double-slit interference pattern, consider how two waves travel from the slits to the screen (**Figure 32.6**). Each slit is a different distance from a given point on the screen. Thus, different numbers of wavelengths fit into each path. Waves start out from the slits in phase (crest to crest), but they may end up out of phase (crest to trough) at the screen if the paths differ in length by half a wavelength, interfering destructively. If the paths differ by a whole wavelength, then the waves arrive in phase (crest to crest) at the screen, interfering constructively. More generally, if the path length difference Δl between the two waves is any half-integral number of wavelengths [$(1/2)\lambda$, $(3/2)\lambda$, $(5/2)\lambda$, etc.], then destructive interference occurs. Similarly, if the path length difference is any integral number of wavelengths (λ, 2λ, 3λ, etc.), then constructive interference occurs. These conditions can be expressed as equations:

$$\Delta l = m\lambda, \quad \text{for } m = 0, \pm 1, \pm 2, \pm 3 \ \ldots \ \text{(constructive interference)} \tag{32.1}$$

$$\Delta l = \left(m + \frac{1}{2}\right)\lambda, \quad \text{for } m = 0, \pm 1, \pm 2, \pm 3 \ \ldots \ \text{(destructive interference)} \tag{32.2}$$

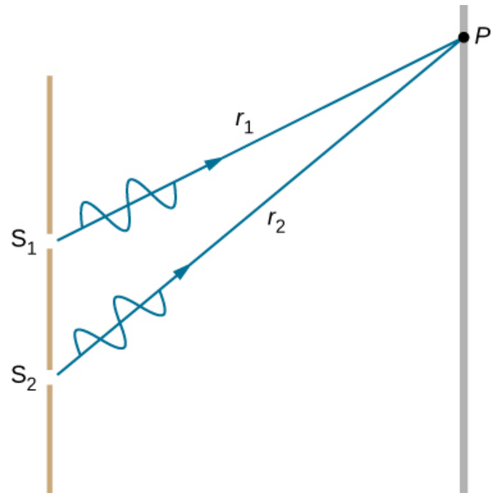


Figure 32.6 Waves follow different paths from the slits to a common point *P* on a screen. Destructive interference occurs where one path is a half wavelength longer than the other—the waves start in phase but arrive out of phase. Constructive interference occurs where one path is a whole wavelength longer than the other—the waves start out and arrive in phase.

32.2 | The Mathematics of Interference

Learning Objectives

By the end of this section, you will be able to:

- Determine the angles for bright and dark fringes for double slit interference
- Calculate the positions of bright fringes on a screen

Figure 32.7(a) shows how to determine the path length difference Δl for waves traveling from two slits to a common point on a screen. If the screen is a large distance away compared with the distance between the slits, then the angle θ between the path and a line from the slits to the screen [part (b)] is nearly the same for each path. In other words, r_1 and r_2 are essentially parallel. The lengths of r_1 and r_2 differ by Δl, as indicated by the two dashed lines in the figure. Simple trigonometry shows

$$\Delta l = d \sin\theta \tag{32.3}$$

where *d* is the distance between the slits. Combining this result with **Equation 32.1**, we obtain constructive interference for a double slit when the path length difference is an integral multiple of the wavelength, or

$$d \sin\theta = m\lambda, \ \text{for } m = 0, \pm1, \pm2, \pm3, \ldots \text{ (constructive interference).} \tag{32.4}$$

Similarly, to obtain destructive interference for a double slit, the path length difference must be a half-integral multiple of the wavelength, or

$$d \sin\theta = (m + \tfrac{1}{2})\lambda, \ \text{for } m = 0, \pm1, \pm2, \pm3, \ldots \text{ (destructive interference)} \tag{32.5}$$

where λ is the wavelength of the light, *d* is the distance between slits, and θ is the angle from the original direction of the beam as discussed above. We call *m* the **order** of the interference. For example, $m = 4$ is fourth-order interference.

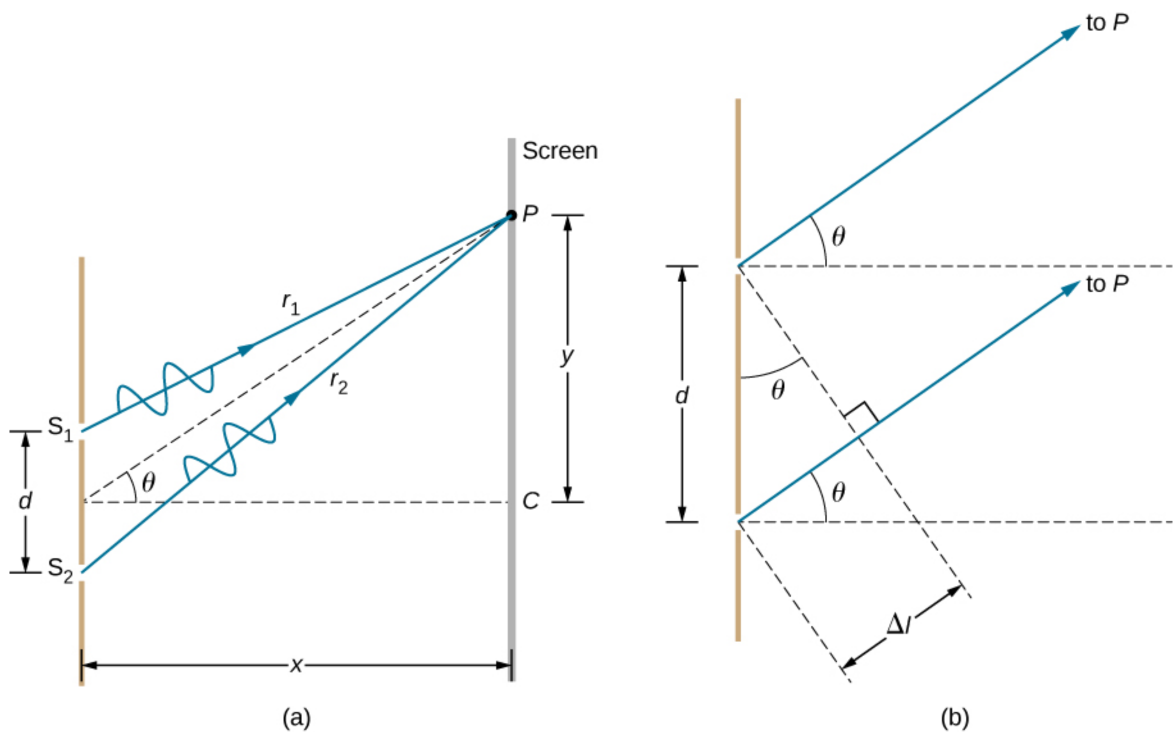


Figure 32.7 (a) To reach *P*, the light waves from S_1 and S_2 must travel different distances. (b) The path difference between the two rays is Δl.

The equations for double-slit interference imply that a series of bright and dark lines are formed. For vertical slits, the light spreads out horizontally on either side of the incident beam into a pattern called interference **fringes** (**Figure 32.8**). The closer the slits are, the more the bright fringes spread apart. We can see this by examining the equation

$d \sin \theta = m\lambda,$ for $m = 0, \pm 1, \pm 2, \pm 3\ldots$. For fixed λ and *m*, the smaller *d* is, the larger θ must be, since $\sin \theta = m\lambda/d$. This is consistent with our contention that wave effects are most noticeable when the object the wave encounters (here, slits a distance *d* apart) is small. Small *d* gives large θ , hence, a large effect.

Referring back to part (a) of the figure, θ is typically small enough that $\sin \theta \approx \tan \theta \approx y_m/D$, where y_m is the distance from the central maximum to the *m*th bright fringe and *D* is the distance between the slit and the screen. **Equation 32.4** may then be written as

$$d\frac{y_m}{D} = m\lambda$$

or

$$y_m = \frac{m\lambda D}{d}. \tag{32.6}$$

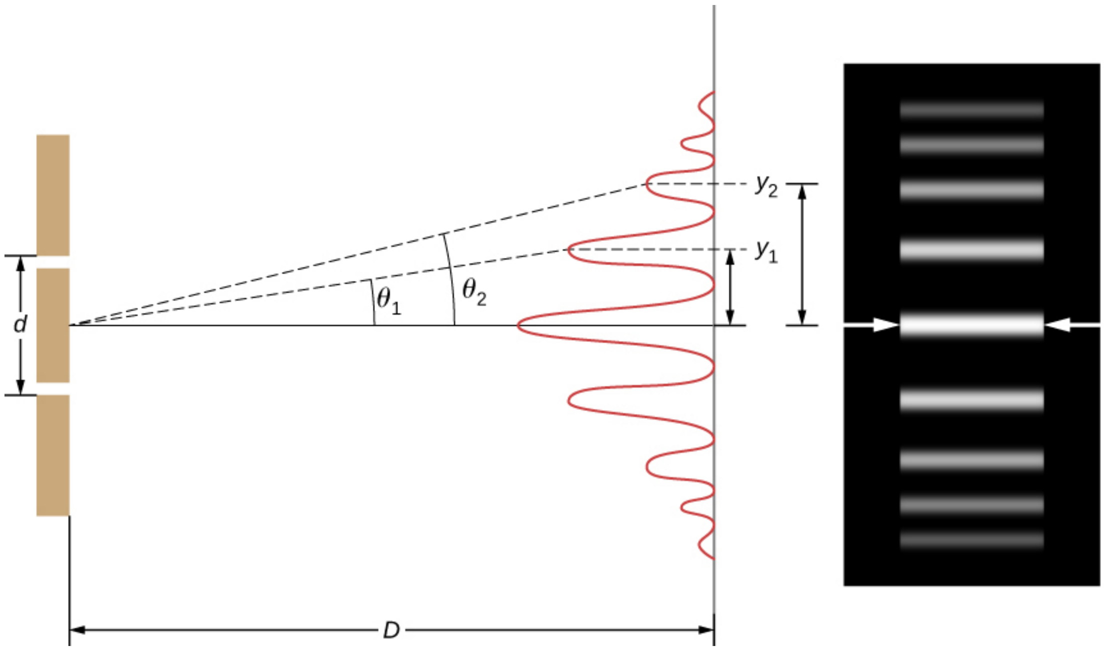


Figure 32.8 The interference pattern for a double slit has an intensity that falls off with angle. The image shows multiple bright and dark lines, or fringes, formed by light passing through a double slit.

Example 32.1

Finding a Wavelength from an Interference Pattern

Suppose you pass light from a He-Ne laser through two slits separated by 0.0100 mm and find that the third bright line on a screen is formed at an angle of $10.95°$ relative to the incident beam. What is the wavelength of the light?

Strategy

The phenomenon is two-slit interference as illustrated in **Figure 32.8** and the third bright line is due to third-order constructive interference, which means that $m = 3$. We are given $d = 0.0100\text{ mm}$ and $\theta = 10.95°$. The wavelength can thus be found using the equation $d \sin \theta = m\lambda$ for constructive interference.

Solution

Solving $d \sin \theta = m\lambda$ for the wavelength λ gives

$$\lambda = \frac{d \sin \theta}{m}.$$

Substituting known values yields

$$\lambda = \frac{(0.0100\text{ mm})(\sin 10.95°)}{3} = 6.33 \times 10^{-4}\text{ mm} = 633\text{ nm}.$$

Significance

To three digits, this is the wavelength of light emitted by the common He-Ne laser. Not by coincidence, this red color is similar to that emitted by neon lights. More important, however, is the fact that interference patterns can be used to measure wavelength. Young did this for visible wavelengths. This analytical techinque is still widely used to measure electromagnetic spectra. For a given order, the angle for constructive interference increases with λ , so that spectra (measurements of intensity versus wavelength) can be obtained.

Example 32.2

Calculating the Highest Order Possible

Interference patterns do not have an infinite number of lines, since there is a limit to how big m can be. What is the highest-order constructive interference possible with the system described in the preceding example?

Strategy

The equation $d \sin \theta = m\lambda$ (for $m = 0, \pm 1, \pm 2, \pm 3\ldots$) describes constructive interference from two slits. For fixed values of d and λ, the larger m is, the larger $\sin \theta$ is. However, the maximum value that $\sin \theta$ can have is 1, for an angle of $90°$. (Larger angles imply that light goes backward and does not reach the screen at all.) Let us find what value of m corresponds to this maximum diffraction angle.

Solution

Solving the equation $d \sin \theta = m\lambda$ for m gives

$$m = \frac{d \sin \theta}{\lambda}.$$

Taking $\sin \theta = 1$ and substituting the values of d and λ from the preceding example gives

$$m = \frac{(0.0100 \text{ mm})(1)}{633 \text{ nm}} \approx 15.8.$$

Therefore, the largest integer m can be is 15, or $m = 15$.

Significance

The number of fringes depends on the wavelength and slit separation. The number of fringes is very large for large slit separations. However, recall (see **The Propagation of Light** and the introduction for this chapter) that wave interference is only prominent when the wave interacts with objects that are not large compared to the wavelength. Therefore, if the slit separation and the sizes of the slits become much greater than the wavelength, the intensity pattern of light on the screen changes, so there are simply two bright lines cast by the slits, as expected, when light behaves like rays. We also note that the fringes get fainter farther away from the center. Consequently, not all 15 fringes may be observable.

32.1 Check Your Understanding In the system used in the preceding examples, at what angles are the first and the second bright fringes formed?

32.3 | Multiple-Slit Interference

Learning Objectives

By the end of this section, you will be able to:

- Describe the locations and intensities of secondary maxima for multiple-slit interference

Analyzing the interference of light passing through two slits lays out the theoretical framework of interference and gives us a historical insight into Thomas Young's experiments. However, much of the modern-day application of slit interference uses not just two slits but many, approaching infinity for practical purposes. The key optical element is called a diffraction grating, an important tool in optical analysis, which we discuss in detail in **Diffraction**. Here, we start the analysis of multiple-slit interference by taking the results from our analysis of the double slit ($N = 2$) and extending it to configurations with three, four, and much larger numbers of slits.

Figure 32.9 shows the simplest case of multiple-slit interference, with three slits, or $N = 3$. The spacing between slits is d, and the path length difference between adjacent slits is $d \sin \theta$, same as the case for the double slit. What is new is that the path length difference for the first and the third slits is $2d \sin \theta$. The condition for constructive interference is the same as for the double slit, that is

$$d \sin \theta = m\lambda.$$

When this condition is met, $2d \sin \theta$ is automatically a multiple of λ, so all three rays combine constructively, and the bright fringes that occur here are called **principal maxima**. But what happens when the path length difference between adjacent slits is only $\lambda/2$? We can think of the first and second rays as interfering destructively, but the third ray remains unaltered. Instead of obtaining a dark fringe, or a minimum, as we did for the double slit, we see a **secondary maximum** with intensity lower than the principal maxima.

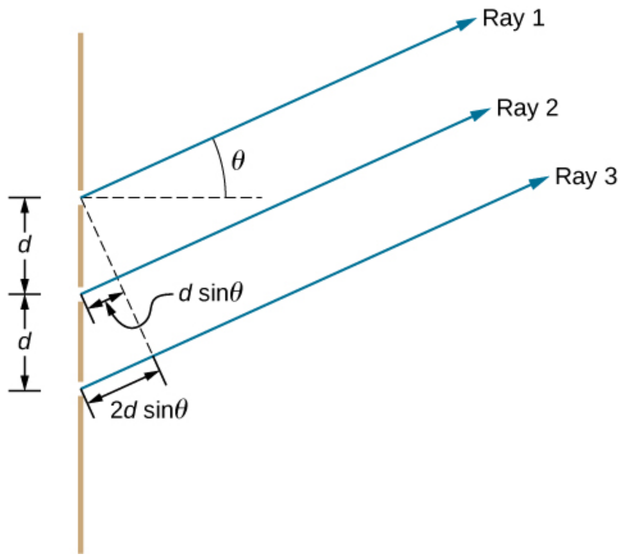


Figure 32.9 Interference with three slits. Different pairs of emerging rays can combine constructively or destructively at the same time, leading to secondary maxima.

In general, for N slits, these secondary maxima occur whenever an unpaired ray is present that does not go away due to destructive interference. This occurs at $(N - 2)$ evenly spaced positions between the principal maxima. The amplitude of the electromagnetic wave is correspondingly diminished to $1/N$ of the wave at the principal maxima, and the light intensity, being proportional to the square of the wave amplitude, is diminished to $1/N^2$ of the intensity compared to the principal maxima. As **Figure 32.10** shows, a dark fringe is located between every maximum (principal or secondary). As N grows larger and the number of bright and dark fringes increase, the widths of the maxima become narrower due to the closely located neighboring dark fringes. Because the total amount of light energy remains unaltered, narrower maxima require that each maximum reaches a correspondingly higher intensity.

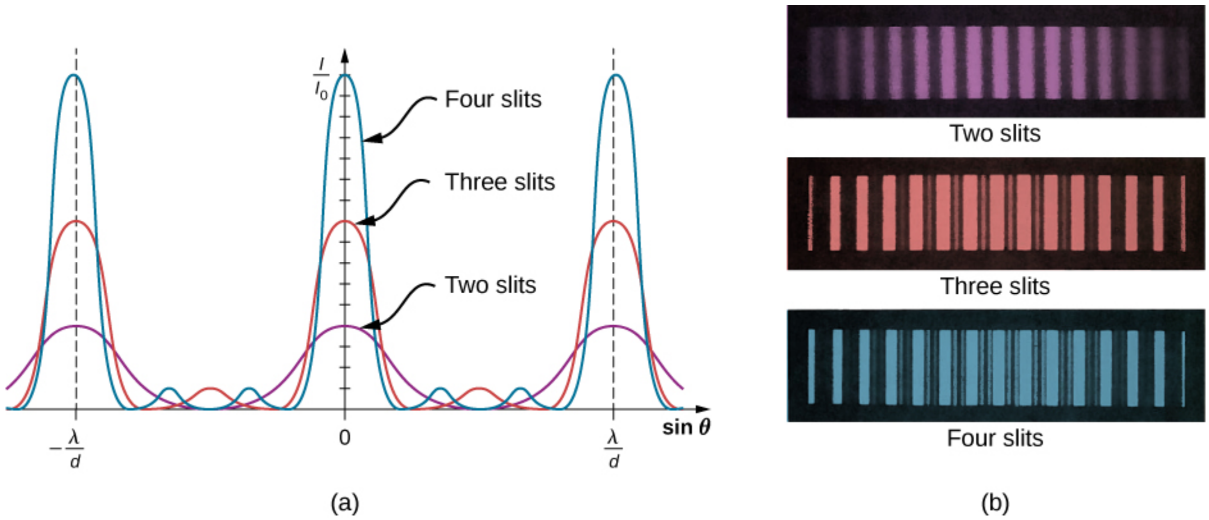


Figure 32.10 Interference fringe patterns for two, three and four slits. As the number of slits increases, more secondary maxima appear, but the principal maxima become brighter and narrower. (a) Graph and (b) photographs of fringe patterns.

32.4 | Interference in Thin Films

Learning Objectives

By the end of this section, you will be able to:

- Describe the phase changes that occur upon reflection
- Describe fringes established by reflected rays of a common source
- Explain the appearance of colors in thin films

The bright colors seen in an oil slick floating on water or in a sunlit soap bubble are caused by interference. The brightest colors are those that interfere constructively. This interference is between light reflected from different surfaces of a thin film; thus, the effect is known as **thin-film interference**.

As we noted before, interference effects are most prominent when light interacts with something having a size similar to its wavelength. A thin film is one having a thickness t smaller than a few times the wavelength of light, λ. Since color is associated indirectly with λ and because all interference depends in some way on the ratio of λ to the size of the object involved, we should expect to see different colors for different thicknesses of a film, as in **Figure 32.11**.

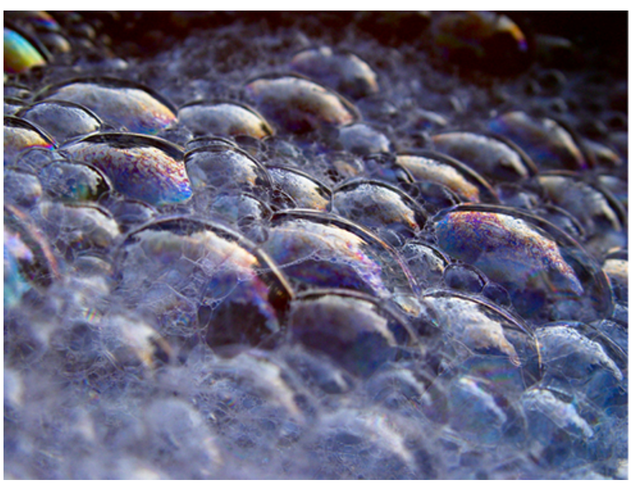


Figure 32.11 These soap bubbles exhibit brilliant colors when exposed to sunlight. (credit: Scott Robinson)

What causes thin-film interference? **Figure 32.12** shows how light reflected from the top and bottom surfaces of a film can interfere. Incident light is only partially reflected from the top surface of the film (ray 1). The remainder enters the film and is itself partially reflected from the bottom surface. Part of the light reflected from the bottom surface can emerge from the top of the film (ray 2) and interfere with light reflected from the top (ray 1). The ray that enters the film travels a greater distance, so it may be in or out of phase with the ray reflected from the top. However, consider for a moment, again, the bubbles in **Figure 32.11**. The bubbles are darkest where they are thinnest. Furthermore, if you observe a soap bubble carefully, you will note it gets dark at the point where it breaks. For very thin films, the difference in path lengths of rays 1 and 2 in **Figure 32.12** is negligible, so why should they interfere destructively and not constructively? The answer is that a phase change can occur upon reflection, as discussed next.

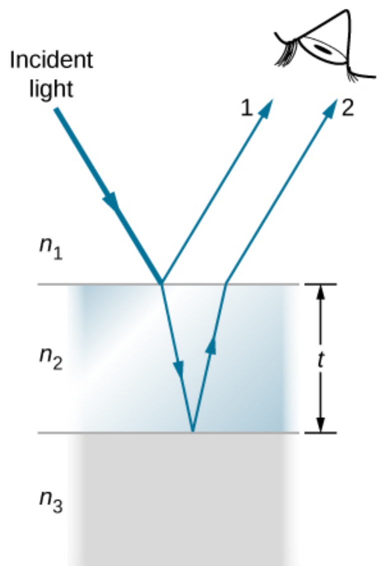


Figure 32.12 Light striking a thin film is partially reflected (ray 1) and partially refracted at the top surface. The refracted ray is partially reflected at the bottom surface and emerges as ray 2. These rays interfere in a way that depends on the thickness of the film and the indices of refraction of the various media.

Changes in Phase due to Reflection

Reflection of mechanical waves can involve a $180°$ phase change. For example, a traveling wave on a string is inverted (i.e., a $180°$ phase change) upon reflection at a boundary to which a heavier string is tied. However, if the second string is lighter (or more precisely, of a lower linear density), no inversion occurs. Light waves produce the same effect, but the deciding parameter for light is the index of refraction. Light waves undergo a $180°$ or π radians phase change upon reflection at an interface beyond which is a medium of higher index of refraction. No phase change takes place when reflecting from a medium of lower refractive index (**Figure 32.13**). Because of the periodic nature of waves, this phase change or inversion is equivalent to $\pm\lambda/2$ in distance travelled, or path length. Both the path length and refractive indices are important factors in thin-film interference.

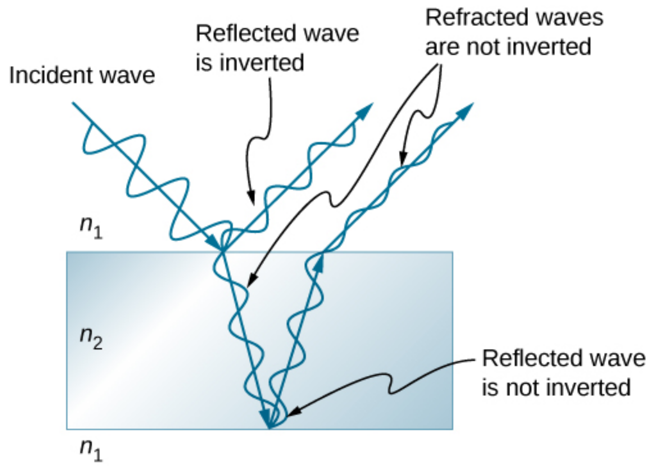


Figure 32.13 Reflection at an interface for light traveling from a medium with index of refraction n_1 to a medium with index of refraction n_2, $n_1 < n_2$, causes the phase of the wave to change by π radians.

If the film in **Figure 32.12** is a **soap bubble** (essentially water with air on both sides), then a phase shift of $\lambda/2$ occurs for ray 1 but not for ray 2. Thus, when the film is very thin and the path length difference between the two rays is negligible, they are exactly out of phase, and destructive interference occurs at all wavelengths. Thus, the soap bubble is dark here. The thickness of the film relative to the wavelength of light is the other crucial factor in thin-film interference. Ray 2 in **Figure 32.12** travels a greater distance than ray 1. For light incident perpendicular to the surface, ray 2 travels a distance approximately $2t$ farther than ray 1. When this distance is an integral or half-integral multiple of the wavelength in the medium ($\lambda_n = \lambda/n$, where λ is the wavelength in vacuum and n is the index of refraction), constructive or destructive interference occurs, depending also on whether there is a phase change in either ray.

Example 32.3

Calculating the Thickness of a Nonreflective Lens Coating

Sophisticated cameras use a series of several lenses. Light can reflect from the surfaces of these various lenses and degrade image clarity. To limit these reflections, lenses are coated with a thin layer of magnesium fluoride, which causes destructive thin-film interference. What is the thinnest this film can be, if its index of refraction is 1.38 and it is designed to limit the reflection of 550-nm light, normally the most intense visible wavelength? Assume the index of refraction of the glass is 1.52.

Strategy

Refer to **Figure 32.12** and use $n_1 = 1.00$ for air, $n_2 = 1.38$, and $n_3 = 1.52$. Both ray 1 and ray 2 have a $\lambda/2$ shift upon reflection. Thus, to obtain destructive interference, ray 2 needs to travel a half wavelength farther than ray 1. For rays incident perpendicularly, the path length difference is $2t$.

Solution

To obtain destructive interference here,

$$2t = \frac{\lambda_{n2}}{2}$$

where λ_{n2} is the wavelength in the film and is given by $\lambda_{n2} = \lambda/n_2$. Thus,

$$2t = \frac{\lambda/n_2}{2}.$$

Solving for t and entering known values yields

$$t = \frac{\lambda/n_2}{4} = \frac{(500\text{ nm})/1.38}{4} = 90.6\text{ nm}.$$

Significance

Films such as the one in this example are most effective in producing destructive interference when the thinnest layer is used, since light over a broader range of incident angles is reduced in intensity. These films are called nonreflective coatings; this is only an approximately correct description, though, since other wavelengths are only partially cancelled. Nonreflective coatings are also used in car windows and sunglasses.

Combining Path Length Difference with Phase Change

Thin-film interference is most constructive or most destructive when the path length difference for the two rays is an integral or half-integral wavelength. That is, for rays incident perpendicularly,

$$2t = \lambda_n, 2\lambda_n, 3\lambda_n, \ldots \text{ or } 2t = \lambda_n/2, 3\lambda_n/2, 5\lambda_n/2, \ldots.$$

To know whether interference is constructive or destructive, you must also determine if there is a phase change upon reflection. Thin-film interference thus depends on film thickness, the wavelength of light, and the refractive indices. For white light incident on a film that varies in thickness, you can observe rainbow colors of constructive interference for various wavelengths as the thickness varies.

Example 32.4

Soap Bubbles

(a) What are the three smallest thicknesses of a soap bubble that produce constructive interference for red light with a wavelength of 650 nm? The index of refraction of soap is taken to be the same as that of water. (b) What three smallest thicknesses give destructive interference?

Strategy

Use **Figure 32.12** to visualize the bubble, which acts as a thin film between two layers of air. Thus $n_1 = n_3 = 1.00$ for air, and $n_2 = 1.333$ for soap (equivalent to water). There is a $\lambda/2$ shift for ray 1 reflected from the top surface of the bubble and no shift for ray 2 reflected from the bottom surface. To get constructive interference, then, the path length difference ($2t$) must be a half-integral multiple of the wavelength—the first three being $\lambda_n/2, 3\lambda_n/2$, and $5\lambda_n/2$. To get destructive interference, the path length difference must be an integral multiple of the wavelength—the first three being 0, λ_n, and $2\lambda_n$.

Solution

a. Constructive interference occurs here when

$$2t_\text{c} = \frac{\lambda_n}{2}, \frac{3\lambda_n}{2}, \frac{5\lambda_n}{2}, \ \ldots\ .$$

Thus, the smallest constructive thickness t_c is

$$t_\text{c} = \frac{\lambda_n}{4} = \frac{\lambda/n}{4} = \frac{(650\text{ nm})/1.333}{4} = 122\text{ nm}.$$

The next thickness that gives constructive interference is $t'_\text{c} = 3\lambda_n/4$, so that

$$t'_\text{c} = 366\text{ nm}.$$

Finally, the third thickness producing constructive interference is $t'_\text{c} = 5\lambda_n/4$, so that

$$t'_\text{c} = 610\text{ nm}.$$

b. For destructive interference, the path length difference here is an integral multiple of the wavelength. The first occurs for zero thickness, since there is a phase change at the top surface, that is,

$$t_d = 0,$$

the very thin (or negligibly thin) case discussed above. The first non-zero thickness producing destructive interference is

$$2t'_d = \lambda_n.$$

Substituting known values gives

$$t'_d = \frac{\lambda}{2} = \frac{\lambda/n}{2} = \frac{(650\text{ nm})/1.333}{2} = 244\text{ nm}.$$

Finally, the third destructive thickness is $2t''_d = 2\lambda_n$, so that

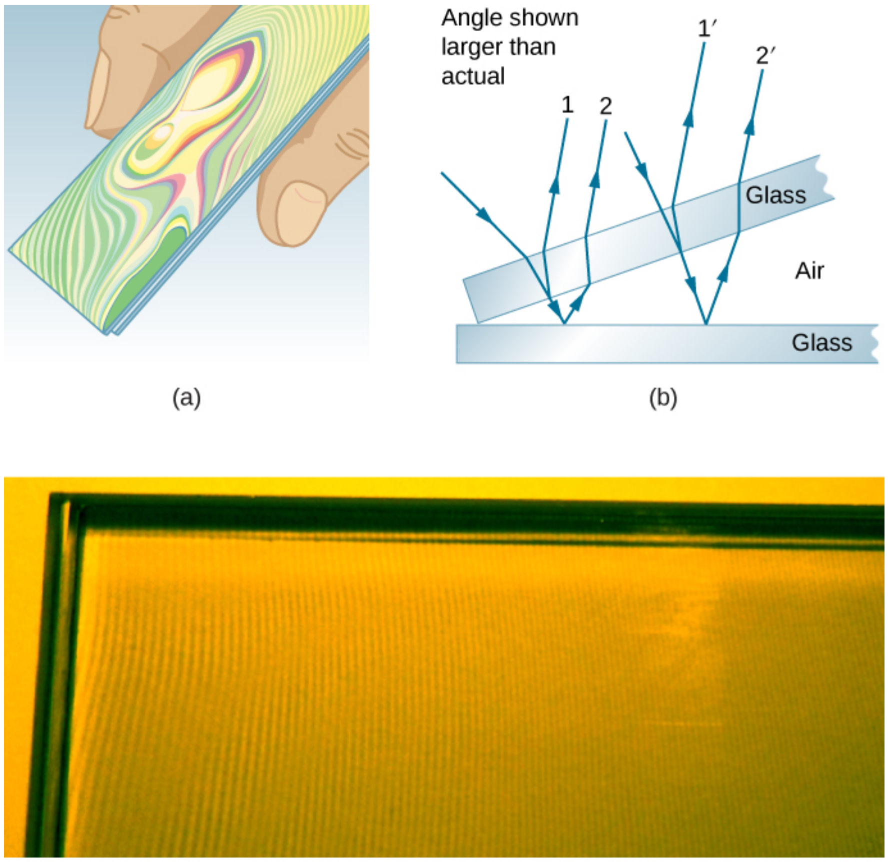
$$t''_d = \lambda_n = \frac{\lambda}{n} = \frac{650\text{ nm}}{1.333} = 488\text{ nm}.$$

Significance

If the bubble were illuminated with pure red light, we would see bright and dark bands at very uniform increases in thickness. First would be a dark band at 0 thickness, then bright at 122 nm thickness, then dark at 244 nm, bright at 366 nm, dark at 488 nm, and bright at 610 nm. If the bubble varied smoothly in thickness, like a smooth wedge, then the bands would be evenly spaced.

32.2 Check Your Understanding Going further with **Example 32.4**, what are the next two thicknesses of soap bubble that would lead to (a) constructive interference, and (b) destructive interference?

Another example of thin-film interference can be seen when microscope slides are separated (see **Figure 32.14**). The slides are very flat, so that the wedge of air between them increases in thickness very uniformly. A phase change occurs at the second surface but not the first, so a dark band forms where the slides touch. The rainbow colors of constructive interference repeat, going from violet to red again and again as the distance between the slides increases. As the layer of air increases, the bands become more difficult to see, because slight changes in incident angle have greater effects on path length differences. If monochromatic light instead of white light is used, then bright and dark bands are obtained rather than repeating rainbow colors.



(a) (b)

(c)

Figure 32.14 (a) The rainbow-color bands are produced by thin-film interference in the air between the two glass slides. (b) Schematic of the paths taken by rays in the wedge of air between the slides. (c) If the air wedge is illuminated with monochromatic light, bright and dark bands are obtained rather than repeating rainbow colors.

An important application of thin-film interference is found in the manufacturing of optical instruments. A lens or mirror can be compared with a master as it is being ground, allowing it to be shaped to an accuracy of less than a wavelength over its entire surface. **Figure 32.15** illustrates the phenomenon called **Newton's rings**, which occurs when the plane surfaces of two lenses are placed together. (The circular bands are called Newton's rings because Isaac Newton described them and their use in detail. Newton did not discover them; Robert Hooke did, and Newton did not believe they were due to the wave character of light.) Each successive ring of a given color indicates an increase of only half a wavelength in the distance between the lens and the blank, so that great precision can be obtained. Once the lens is perfect, no rings appear.

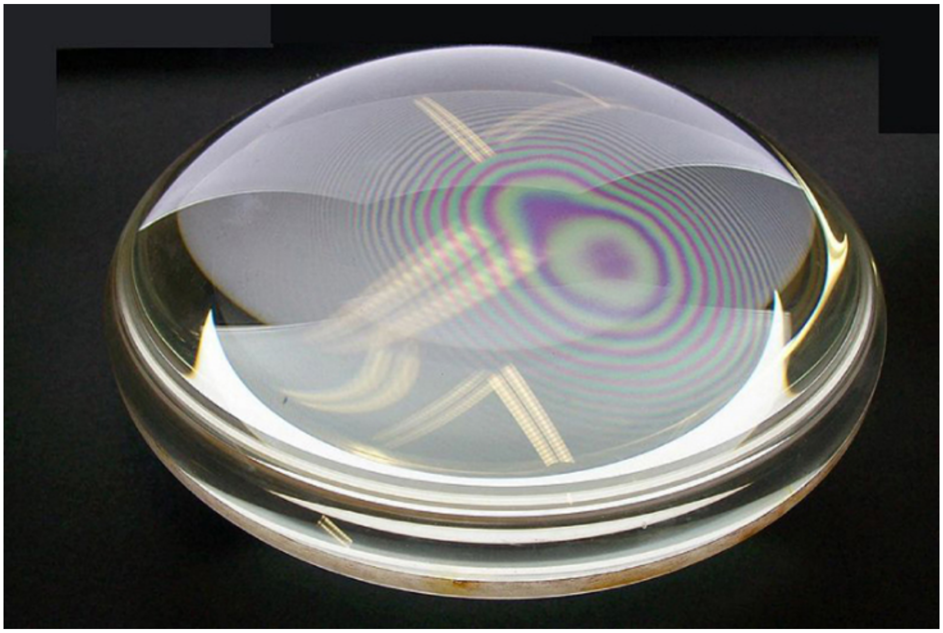


Figure 32.15 “Newton’s rings” interference fringes are produced when two plano-convex lenses are placed together with their plane surfaces in contact. The rings are created by interference between the light reflected off the two surfaces as a result of a slight gap between them, indicating that these surfaces are not precisely plane but are slightly convex. (credit: Ulf Seifert)

Thin-film interference has many other applications, both in nature and in manufacturing. The wings of certain moths and butterflies have nearly iridescent colors due to thin-film interference. In addition to pigmentation, the wing’s color is affected greatly by constructive interference of certain wavelengths reflected from its film-coated surface. Some car manufacturers offer special paint jobs that use thin-film interference to produce colors that change with angle. This expensive option is based on variation of thin-film path length differences with angle. Security features on credit cards, banknotes, driving licenses, and similar items prone to forgery use thin-film interference, diffraction gratings, or holograms. As early as 1998, Australia led the way with dollar bills printed on polymer with a diffraction grating security feature, making the currency difficult to forge. Other countries, such as Canada, New Zealand, and Taiwan, are using similar technologies, while US currency includes a thin-film interference effect.

32.5 | The Michelson Interferometer

Learning Objectives

By the end of this section, you will be able to:

- Explain changes in fringes observed with a Michelson interferometer caused by mirror movements
- Explain changes in fringes observed with a Michelson interferometer caused by changes in medium

The Michelson **interferometer** (invented by the American physicist Albert A. Michelson, 1852–1931) is a precision instrument that produces interference fringes by splitting a light beam into two parts and then recombining them after they have traveled different optical paths. Figure 32.16 depicts the interferometer and the path of a light beam from a single point on the extended source S, which is a ground-glass plate that diffuses the light from a monochromatic lamp of wavelength λ_0. The beam strikes the half-silvered mirror M, where half of it is reflected to the side and half passes through the mirror. The reflected light travels to the movable plane mirror M_1, where it is reflected back through M to the observer. The transmitted half of the original beam is reflected back by the stationary mirror M_2 and then toward the observer by M.

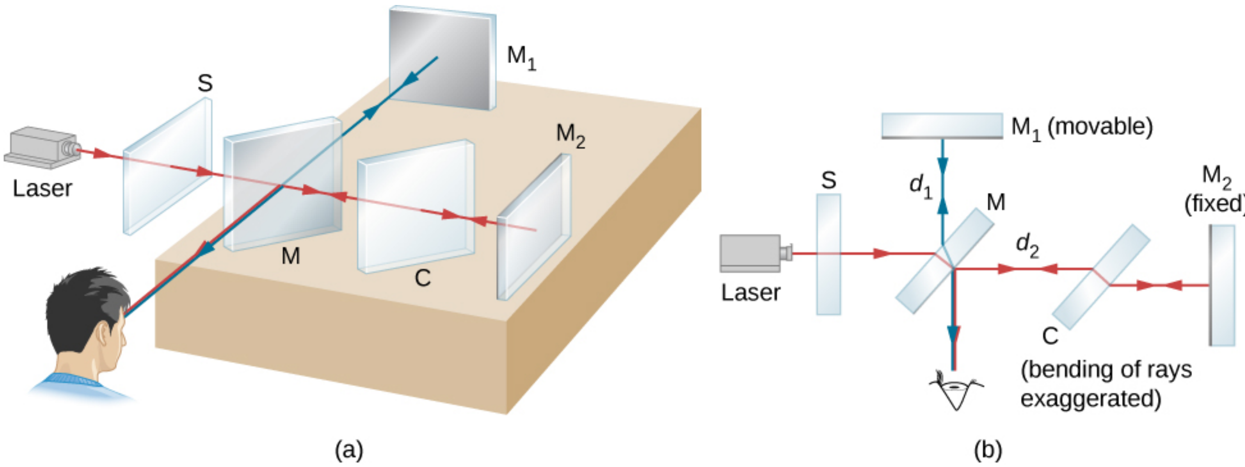


Figure 32.16 (a) The Michelson interferometer. The extended light source is a ground-glass plate that diffuses the light from a laser. (b) A planar view of the interferometer.

Because both beams originate from the same point on the source, they are coherent and therefore interfere. Notice from the figure that one beam passes through M three times and the other only once. To ensure that both beams traverse the same thickness of glass, a compensator plate C of transparent glass is placed in the arm containing M_2. This plate is a duplicate of M (without the silvering) and is usually cut from the same piece of glass used to produce M. With the compensator in place, any phase difference between the two beams is due solely to the difference in the distances they travel.

The path difference of the two beams when they recombine is $2d_1 - 2d_2$, where d_1 is the distance between M and M_1, and d_2 is the distance between M and M_2. Suppose this path difference is an integer number of wavelengths $m\lambda_0$. Then, constructive interference occurs and a bright image of the point on the source is seen at the observer. Now the light from any other point on the source whose two beams have this same path difference also undergoes constructive interference and produces a bright image. The collection of these point images is a bright fringe corresponding to a path difference of $m\lambda_0$ (**Figure 32.17**). When M_1 is moved a distance $\Delta d = \lambda_0/2$, this path difference changes by λ_0, and each fringe moves to the position previously occupied by an adjacent fringe. Consequently, by counting the number of fringes *m* passing a given point as M_1 is moved, an observer can measure minute displacements that are accurate to a fraction of a wavelength, as shown by the relation

$$\Delta d = m\frac{\lambda_0}{2}. \quad (32.7)$$

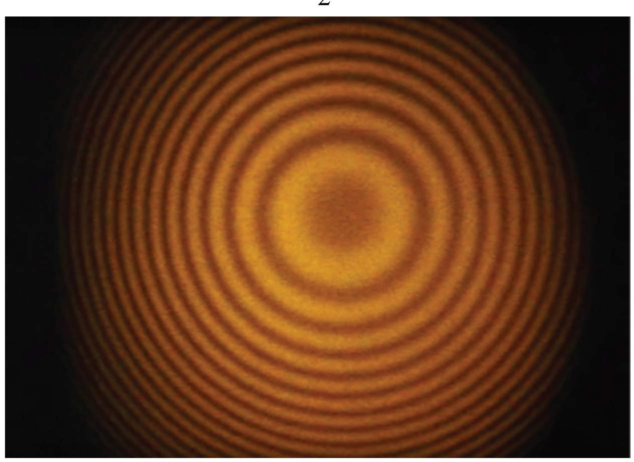


Figure 32.17 Fringes produced with a Michelson interferometer. (credit: “SILLAGESvideos”/YouTube)

Example 32.5

Precise Distance Measurements by Michelson Interferometer

A red laser light of wavelength 630 nm is used in a Michelson interferometer. While keeping the mirror M_1 fixed, mirror M_2 is moved. The fringes are found to move past a fixed cross-hair in the viewer. Find the distance the mirror M_2 is moved for a single fringe to move past the reference line.

Strategy

Refer to **Figure 32.16** for the geometry. We use the result of the Michelson interferometer interference condition to find the distance moved, Δd .

Solution

For a 630-nm red laser light, and for each fringe crossing $(m = 1)$, the distance traveled by M_2 if you keep M_1 fixed is

$$\Delta d = m\frac{\lambda_0}{2} = 1 \times \frac{630\text{ nm}}{2} = 315\text{ nm} = 0.315\,\mu\text{m}.$$

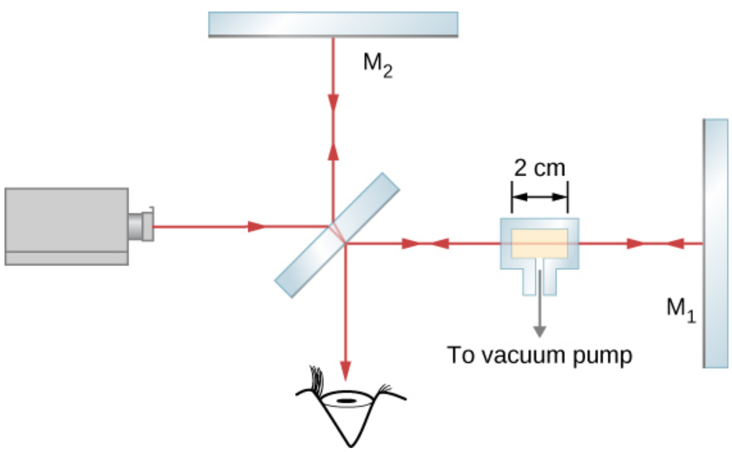
Significance

An important application of this measurement is the definition of the standard meter. As mentioned in **Introducing Astrophysics**, the length of the standard meter was once defined as the mirror displacement in a Michelson interferometer corresponding to 1,650,763.73 wavelengths of the particular fringe of krypton-86 in a gas discharge tube.

Example 32.6

Measuring the Refractive Index of a Gas

In one arm of a Michelson interferometer, a glass chamber is placed with attachments for evacuating the inside and putting gases in it. The space inside the container is 2 cm wide. Initially, the container is empty. As gas is slowly let into the chamber, you observe that dark fringes move past a reference line in the field of observation. By the time the chamber is filled to the desired pressure, you have counted 122 fringes move past the reference line. The wavelength of the light used is 632.8 nm. What is the refractive index of this gas?



Strategy

The $m = 122$ fringes observed compose the difference between the number of wavelengths that fit within the empty chamber (vacuum) and the number of wavelengths that fit within the same chamber when it is gas-filled. The wavelength in the filled chamber is shorter by a factor of n, the index of refraction.

Solution

The ray travels a distance $t = 2\text{ cm}$ to the right through the glass chamber and another distance t to the left upon reflection. The total travel is $L = 2t$. When empty, the number of wavelengths that fit in this chamber is

$$N_0 = \frac{L}{\lambda_0} = \frac{2t}{\lambda_0}$$

where $\lambda_0 = 632.8\text{ nm}$ is the wavelength in vacuum of the light used. In any other medium, the wavelength is $\lambda = \lambda_0/n$ and the number of wavelengths that fit in the gas-filled chamber is

$$N = \frac{L}{\lambda} = \frac{2t}{\lambda_0/n}.$$

The number of fringes observed in the transition is

$$\begin{aligned} m &= N - N_0, \\ &= \frac{2t}{\lambda_0/n} - \frac{2t}{\lambda_0}, \\ &= \frac{2t}{\lambda_0}(n - 1). \end{aligned}$$

Solving for $(n - 1)$ gives

$$n - 1 = m\left(\frac{\lambda_0}{2t}\right) = 122\left(\frac{632.8 \times 10^{-9}\text{ m}}{2(2 \times 10^{-2}\text{ m})}\right) = 0.0019$$

and $n = 1.0019$.

Significance

The indices of refraction for gases are so close to that of vacuum, that we normally consider them equal to 1. The difference between 1 and 1.0019 is so small that measuring it requires a correspondingly sensitive technique such as interferometry. We cannot, for example, hope to measure this value using techniques based simply on Snell's law.

32.3 Check Your Understanding Although m, the number of fringes observed, is an integer, which is often regarded as having zero uncertainty, in practical terms, it is all too easy to lose track when counting fringes. In **Example 32.6**, if you estimate that you might have missed as many as five fringes when you reported $m = 122$ fringes, (a) is the value for the index of refraction worked out in **Example 32.6** too large or too small? (b) By how much?

Problem-Solving Strategy: Wave Optics

Step 1. *Examine the situation to determine that interference is involved*. Identify whether slits, thin films, or interferometers are considered in the problem.

Step 2. *If slits are involved*, note that diffraction gratings and double slits produce very similar interference patterns, but that gratings have narrower (sharper) maxima. Single-slit patterns are characterized by a large central maximum and smaller maxima to the sides.

Step 3. *If thin-film interference or an interferometer is involved, take note of the path length difference between the two rays that interfere*. Be certain to use the wavelength in the medium involved, since it differs from the wavelength in vacuum. Note also that there is an additional $\lambda/2$ phase shift when light reflects from a medium with a greater index of refraction.

Step 4. *Identify exactly what needs to be determined in the problem (identify the unknowns)*. A written list is useful. Draw a diagram of the situation. Labeling the diagram is useful.

Step 5. *Make a list of what is given or can be inferred from the problem as stated (identify the knowns)*.

Step 6. *Solve the appropriate equation for the quantity to be determined (the unknown) and enter the knowns*. Slits, gratings, and the Rayleigh limit involve equations.

Step 7. *For thin-film interference, you have constructive interference for a total shift that is an integral number of wavelengths. You have destructive interference for a total shift of a half-integral number of wavelengths*. Always keep in mind that crest to crest is constructive whereas crest to trough is destructive.

Step 8. *Check to see if the answer is reasonable: Does it make sense?* Angles in interference patterns cannot be greater than 90°, for example.

CHAPTER 32 REVIEW

KEY TERMS

coherent waves waves are in phase or have a definite phase relationship

fringes bright and dark patterns of interference

incoherent waves have random phase relationships

interferometer instrument that uses interference of waves to make measurements

monochromatic light composed of one wavelength only

Newton's rings circular interference pattern created by interference between the light reflected off two surfaces as a result of a slight gap between them

order integer *m* used in the equations for constructive and destructive interference for a double slit

principal maximum brightest interference fringes seen with multiple slits

secondary maximum bright interference fringes of intensity lower than the principal maxima

thin-film interference interference between light reflected from different surfaces of a thin film

KEY EQUATIONS

Constructive interference	$\Delta l = m\lambda$, for m = 0, ±1, ±2, ±3…
Destructive interference	$\Delta l = (m + \frac{1}{2})\lambda$, for m = 0, ±1, ±2, ±3…
Path length difference for waves from two slits to a common point on a screen	$\Delta l = d \sin \theta$
Constructive interference	$d \sin \theta = m\lambda$, for $m = 0, \pm 1, \pm 2, \pm 3, \ldots$
Destructive interference	$d \sin \theta = (m + \frac{1}{2})\lambda$, for $m = 0, \pm 1, \pm 2, \pm 3, \ldots$
Distance from central maximum to the *m*th bright fringe	$y_m = \frac{m\lambda D}{d}$
Displacement measured by a Michelson interferometer	$\Delta d = m\frac{\lambda_0}{2}$

SUMMARY

32.1 Young's Double-Slit Experiment

- Young's double-slit experiment gave definitive proof of the wave character of light.
- An interference pattern is obtained by the superposition of light from two slits.

32.2 The Mathematics of Interference

- In double-slit diffraction, constructive interference occurs when $d \sin \theta = m\lambda (\text{for } m = 0, \pm 1, \pm 2, \pm 3 \ldots)$, where d is the distance between the slits, θ is the angle relative to the incident direction, and m is the order of the interference.
- Destructive interference occurs when $d \sin \theta = (m + \frac{1}{2})\lambda$ for $m = 0, \pm 1, \pm 2, \pm 3, \ldots$.

32.3 Multiple-Slit Interference

- Interference from multiple slits ($N > 2$) produces principal as well as secondary maxima.
- As the number of slits is increased, the intensity of the principal maxima increases and the width decreases.

32.4 Interference in Thin Films

- When light reflects from a medium having an index of refraction greater than that of the medium in which it is traveling, a $180°$ phase change (or a $\lambda/2$ shift) occurs.
- Thin-film interference occurs between the light reflected from the top and bottom surfaces of a film. In addition to the path length difference, there can be a phase change.

32.5 The Michelson Interferometer

- When the mirror in one arm of the interferometer moves a distance of $\lambda/2$ each fringe in the interference pattern moves to the position previously occupied by the adjacent fringe.

CONCEPTUAL QUESTIONS

32.1 Young's Double-Slit Experiment

1. Young's double-slit experiment breaks a single light beam into two sources. Would the same pattern be obtained for two independent sources of light, such as the headlights of a distant car? Explain.

2. Is it possible to create a experimental setup in which there is only destructive interference? Explain.

3. Why won't two small sodium lamps, held close together, produce an interference pattern on a distant screen? What if the sodium lamps were replaced by two laser pointers held close together?

32.2 The Mathematics of Interference

4. Suppose you use the same double slit to perform Young's double-slit experiment in air and then repeat the experiment in water. Do the angles to the same parts of the interference pattern get larger or smaller? Does the color of the light change? Explain.

5. Why is monochromatic light used in the double slit experiment? What would happen if white light were used?

32.4 Interference in Thin Films

6. What effect does increasing the wedge angle have on the spacing of interference fringes? If the wedge angle is too large, fringes are not observed. Why?

7. How is the difference in paths taken by two originally in-phase light waves related to whether they interfere constructively or destructively? How can this be affected by reflection? By refraction?

8. Is there a phase change in the light reflected from either surface of a contact lens floating on a person's tear layer? The index of refraction of the lens is about 1.5, and its top surface is dry.

9. In placing a sample on a microscope slide, a glass cover is placed over a water drop on the glass slide. Light incident from above can reflect from the top and bottom of the glass cover and from the glass slide below the water drop. At which surfaces will there be a phase change in the reflected light?

10. Answer the above question if the fluid between the two pieces of crown glass is carbon disulfide.

11. While contemplating the food value of a slice of ham, you notice a rainbow of color reflected from its moist surface. Explain its origin.

12. An inventor notices that a soap bubble is dark at its thinnest and realizes that destructive interference is taking place for all wavelengths. How could she use this knowledge to make a nonreflective coating for lenses that is effective at all wavelengths? That is, what limits would there be on the index of refraction and thickness of the coating? How might this be impractical?

13. A nonreflective coating like the one described in **Example 32.3** works ideally for a single wavelength and for perpendicular incidence. What happens for other wavelengths and other incident directions? Be specific.

14. Why is it much more difficult to see interference fringes for light reflected from a thick piece of glass than from a thin film? Would it be easier if monochromatic light were used?

32.5 The Michelson Interferometer

15. Describe how a Michelson interferometer can be used to measure the index of refraction of a gas (including air).

PROBLEMS

32.2 The Mathematics of Interference

16. At what angle is the first-order maximum for 450-nm wavelength blue light falling on double slits separated by 0.0500 mm?

17. Calculate the angle for the third-order maximum of 580-nm wavelength yellow light falling on double slits separated by 0.100 mm.

18. What is the separation between two slits for which 610-nm orange light has its first maximum at an angle of $30.0°$?

19. Find the distance between two slits that produces the first minimum for 410-nm violet light at an angle of $45.0°$.

20. Calculate the wavelength of light that has its third minimum at an angle of $30.0°$ when falling on double slits separated by $3.00\ \mu\text{m}$. Explicitly show how you follow the steps from the **Problem-Solving Strategy: Wave Optics**, located at the end of the chapter.

21. What is the wavelength of light falling on double slits separated by $2.00\ \mu\text{m}$ if the third-order maximum is at an angle of $60.0°$?

22. At what angle is the fourth-order maximum for the situation in the preceding problem?

23. What is the highest-order maximum for 400-nm light falling on double slits separated by $25.0\ \mu\text{m}$?

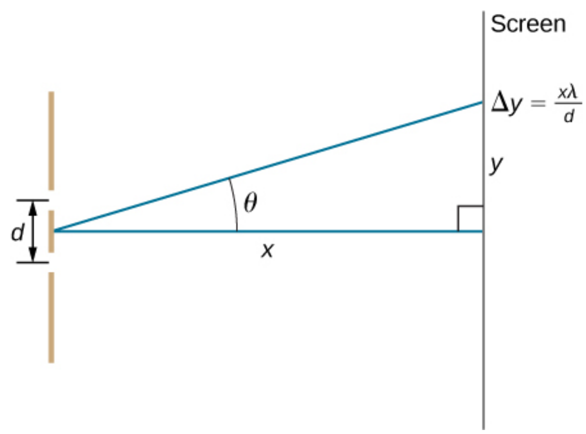
24. Find the largest wavelength of light falling on double slits separated by $1.20\ \mu\text{m}$ for which there is a first-order maximum. Is this in the visible part of the spectrum?

25. What is the smallest separation between two slits that will produce a second-order maximum for 720-nm red light?

26. (a) What is the smallest separation between two slits that will produce a second-order maximum for any visible light? (b) For all visible light?

27. (a) If the first-order maximum for monochromatic light falling on a double slit is at an angle of $10.0°$, at what angle is the second-order maximum? (b) What is the angle of the first minimum? (c) What is the highest-order maximum possible here?

28. Shown below is a double slit located a distance x from a screen, with the distance from the center of the screen given by y. When the distance d between the slits is relatively large, numerous bright spots appear, called fringes. Show that, for small angles (where $\sin\theta \approx \theta$, with θ in radians), the distance between fringes is given by $\Delta y = x\lambda/d$



29. Using the result of the preceding problem, (a) calculate the distance between fringes for 633-nm light falling on double slits separated by 0.0800 mm, located 3.00 m from a screen. (b) What would be the distance between fringes if the entire apparatus were submersed in water, whose index of refraction is 1.33?

30. Using the result of the problem two problems prior, find the wavelength of light that produces fringes 7.50 mm apart on a screen 2.00 m from double slits separated by 0.120 mm.

31. In a double-slit experiment, the fifth maximum is 2.8 cm from the central maximum on a screen that is 1.5 m away from the slits. If the slits are 0.15 mm apart, what is the wavelength of the light being used?

32. The source in Young's experiment emits at two wavelengths. On the viewing screen, the fourth maximum for one wavelength is located at the same spot as the fifth maximum for the other wavelength. What is the ratio of the two wavelengths?

33. If 500-nm and 650-nm light illuminates two slits that are separated by 0.50 mm, how far apart are the second-order maxima for these two wavelengths on a screen 2.0 m away?

34. Red light of wavelength of 700 nm falls on a double slit separated by 400 nm. (a) At what angle is the first-order maximum in the diffraction pattern? (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

32.3 Multiple-Slit Interference

35. Ten narrow slits are equally spaced 0.25 mm apart and illuminated with yellow light of wavelength 580 nm. (a) What are the angular positions of the third and fourth principal maxima? (b) What is the separation of these maxima on a screen 2.0 m from the slits?

36. The width of bright fringes can be calculated as the separation between the two adjacent dark fringes on either side. Find the angular widths of the third- and fourth-order bright fringes from the preceding problem.

37. For a three-slit interference pattern, find the ratio of the peak intensities of a secondary maximum to a principal maximum.

38. What is the angular width of the central fringe of the interference pattern of (a) 20 slits separated by $d = 2.0 \times 10^{-3}\,\text{mm}$? (b) 50 slits with the same separation? Assume that $\lambda = 600\,\text{nm}$.

32.4 Interference in Thin Films

39. A soap bubble is 100 nm thick and illuminated by white light incident perpendicular to its surface. What wavelength and color of visible light is most constructively reflected, assuming the same index of refraction as water?

40. An oil slick on water is 120 nm thick and illuminated by white light incident perpendicular to its surface. What color does the oil appear (what is the most constructively reflected wavelength), given its index of refraction is 1.40?

41. Calculate the minimum thickness of an oil slick on water that appears blue when illuminated by white light perpendicular to its surface. Take the blue wavelength to be 470 nm and the index of refraction of oil to be 1.40.

42. Find the minimum thickness of a soap bubble that appears red when illuminated by white light perpendicular to its surface. Take the wavelength to be 680 nm, and assume the same index of refraction as water.

43. A film of soapy water ($n = 1.33$) on top of a plastic cutting board has a thickness of 233 nm. What color is most strongly reflected if it is illuminated perpendicular to its surface?

44. What are the three smallest non-zero thicknesses of soapy water ($n = 1.33$) on Plexiglas if it appears green (constructively reflecting 520-nm light) when illuminated perpendicularly by white light?

45. Suppose you have a lens system that is to be used primarily for 700-nm red light. What is the second thinnest coating of fluorite (magnesium fluoride) that would be nonreflective for this wavelength?

46. (a) As a soap bubble thins it becomes dark, because the path length difference becomes small compared with the wavelength of light and there is a phase shift at the top surface. If it becomes dark when the path length difference is less than one-fourth the wavelength, what is the thickest the bubble can be and appear dark at all visible wavelengths? Assume the same index of refraction as water. (b) Discuss the fragility of the film considering the thickness found.

47. To save money on making military aircraft invisible to radar, an inventor decides to coat them with a nonreflective material having an index of refraction of 1.20, which is between that of air and the surface of the plane. This, he reasons, should be much cheaper than designing Stealth bombers. (a) What thickness should the coating be to inhibit the reflection of 4.00-cm wavelength radar? (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

32.5 The Michelson Interferometer

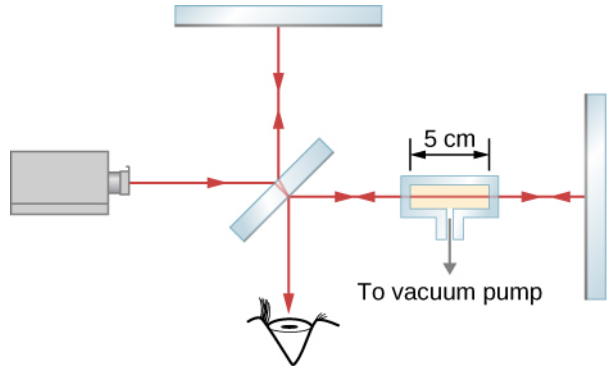
48. A Michelson interferometer has two equal arms. A mercury light of wavelength 546 nm is used for the interferometer and stable fringes are found. One of the arms is moved by $1.5\mu\text{m}$. How many fringes will cross the observing field?

49. What is the distance moved by the traveling mirror of a Michelson interferometer that corresponds to 1500 fringes passing by a point of the observation screen? Assume that the interferometer is illuminated with a 606 nm spectral line of krypton-86.

50. When the traveling mirror of a Michelson interferometer is moved 2.40×10^{-5} m, 90 fringes pass by a point on the observation screen. What is the wavelength of the light used?

51. In a Michelson interferometer, light of wavelength 632.8 nm from a He-Ne laser is used. When one of the mirrors is moved by a distance *D*, 8 fringes move past the field of view. What is the value of the distance *D*?

52. A chamber 5.0 cm long with flat, parallel windows at the ends is placed in one arm of a Michelson interferometer (see below). The light used has a wavelength of 500 nm in a vacuum. While all the air is being pumped out of the chamber, 29 fringes pass by a point on the observation screen. What is the refractive index of the air?



ADDITIONAL PROBLEMS

53. For 600-nm wavelength light and a slit separation of 0.12 mm, what are the angular positions of the first and third maxima in the double slit interference pattern?

54. If the light source in the preceding problem is changed, the angular position of the third maximum is found to be $0.57°$. What is the wavelength of light being used now?

55. Red light ($\lambda = 710.\,\text{nm}$) illuminates double slits separated by a distance $d = 0.150\,\text{mm}$. The screen and the slits are 3.00 m apart. (a) Find the distance on the screen between the central maximum and the third maximum. (b) What is the distance between the second and the fourth maxima?

56. Two sources as in phase and emit waves with $\lambda = 0.42\,\text{m}$. Determine whether constructive or destructive interference occurs at points whose distances from the two sources are (a) 0.84 and 0.42 m, (b) 0.21 and 0.42 m, (c) 1.26 and 0.42 m, (d) 1.87 and 1.45 m, (e) 0.63 and 0.84 m and (f) 1.47 and 1.26 m.

57. Two slits 4.0×10^{-6} m apart are illuminated by light of wavelength 600 nm. What is the highest order fringe in the interference pattern?

58. Suppose that the highest order fringe that can be observed is the eighth in a double-slit experiment where 550-nm wavelength light is used. What is the minimum separation of the slits?

59. The interference pattern of a He-Ne laser light $(\lambda = 632.9\,\text{nm})$ passing through two slits 0.031 mm apart is projected on a screen 10.0 m away. Determine the distance between the adjacent bright fringes.

60. Young's double-slit experiment is performed immersed in water ($n = 1.333$). The light source is a He-Ne laser, $\lambda = 632.9\,\text{nm}$ in vacuum. (a) What is the wavelength of this light in water? (b) What is the angle for the third order maximum for two slits separated by 0.100 mm.

61. A double-slit experiment is to be set up so that the bright fringes appear 1.27 cm apart on a screen 2.13 m away from the two slits. The light source was wavelength 500 nm. What should be the separation between the two slits?

62. An effect analogous to two-slit interference can occur with sound waves, instead of light. In an open field, two speakers placed 1.30 m apart are powered by a single-function generator producing sine waves at 1200-Hz frequency. A student walks along a line 12.5 m away and parallel to the line between the speakers. She hears an alternating pattern of loud and quiet, due to constructive and destructive interference. What is (a) the wavelength of this sound and (b) the distance between the central maximum and the first maximum (loud) position along this line?

63. A hydrogen gas discharge lamp emits visible light at four wavelengths, $\lambda =$ 410, 434, 486, and 656 nm. (a) If light from this lamp falls on a *N* slits separated by 0.025 mm, how far from the central maximum are the third maxima when viewed on a screen 2.0 m from the slits? (b) By what distance are the second and third maxima separated for $l = 486\,\text{nm}$?

64. Monochromatic light of frequency 5.5×10^{14} Hz falls on 10 slits separated by 0.020 mm. What is the separation between the first and third maxima on a screen that is 2.0 m from the slits?

65. Eight slits equally separated by 0.149 mm is uniformly illuminated by a monochromatic light at $\lambda = 523\text{ nm}$. What is the width of the central principal maximum on a screen 2.35 m away?

66. Eight slits equally separated by 0.149 mm is uniformly illuminated by a monochromatic light at $\lambda = 523\text{ nm}$. What is the intensity of a secondary maxima compared to that of the principal maxima?

67. A transparent film of thickness 250 nm and index of refraction of 1.40 is surrounded by air. What wavelength in a beam of white light at near-normal incidence to the film undergoes destructive interference when reflected?

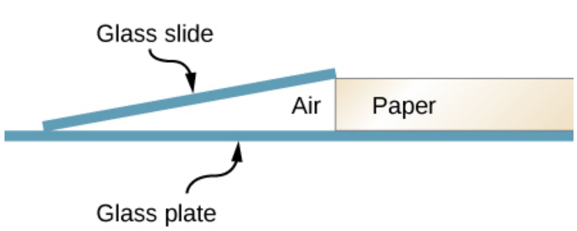
68. An intensity minimum is found for 450 nm light transmitted through a transparent film $(n = 1.20)$ in air. (a) What is minimum thickness of the film? (b) If this wavelength is the longest for which the intensity minimum occurs, what are the next three lower values of λ for which this happens?

69. A thin film with $n = 1.32$ is surrounded by air. What is the minimum thickness of this film such that the reflection of normally incident light with $\lambda = 500\text{ nm}$ is minimized?

70. Repeat your calculation of the previous problem with the thin film placed on a flat glass ($n = 1.50$) surface.

71. After a minor oil spill, a think film of oil ($n = 1.40$) of thickness 450 nm floats on the water surface in a bay. (a) What predominant color is seen by a bird flying overhead? (b) What predominant color is seen by a seal swimming underwater?

72. A microscope slide 10 cm long is separated from a glass plate at one end by a sheet of paper. As shown below, the other end of the slide is in contact with the plate. The slide is illuminated from above by light from a sodium lamp ($\lambda = 589\text{ nm}$), and 14 fringes per centimeter are seen along the slide. What is the thickness of the piece of paper?



73. Suppose that the setup of the preceding problem is immersed in an unknown liquid. If 18 fringes per centimeter are now seen along the slide, what is the index of refraction of the liquid?

74. A thin wedge filled with air is produced when two flat glass plates are placed on top of one another and a slip of paper is inserted between them at one edge. Interference fringes are observed when monochromatic light falling vertically on the plates are seen in reflection. Is the first fringe near the edge where the plates are in contact a bright fringe or a dark fringe? Explain.

75. Two identical pieces of rectangular plate glass are used to measure the thickness of a hair. The glass plates are in direct contact at one edge and a single hair is placed between them hear the opposite edge. When illuminated with a sodium lamp ($\lambda = 589\text{ nm}$), the hair is seen between the 180th and 181st dark fringes. What are the lower and upper limits on the hair's diameter?

76. Two microscope slides made of glass are illuminated by monochromatic ($\lambda = 589\text{ nm}$) light incident perpendicularly. The top slide touches the bottom slide at one end and rests on a thin copper wire at the other end, forming a wedge of air. The diameter of the copper wire is $29.45\ \mu\text{m}$. How many bright fringes are seen across these slides?

77. A good quality camera "lens" is actually a system of lenses, rather than a single lens, but a side effect is that a reflection from the surface of one lens can bounce around many times within the system, creating artifacts in the photograph. To counteract this problem, one of the lenses in such a system is coated with a thin layer of material ($n = 1.28$) on one side. The index of refraction of the lens glass is 1.68. What is the smallest thickness of the coating that reduces the reflection at 640 nm by destructive interference? (In other words, the coating's effect is to be optimized for $\lambda = 640\text{ nm}$.)

78. Constructive interference is observed from directly above an oil slick for wavelengths (in air) 440 nm and 616 nm. The index of refraction of this oil is $n = 1.54$. What is the film's minimum possible thickness?

79. A soap bubble is blown outdoors. What colors (indicate by wavelengths) of the reflected sunlight are seen enhanced? The soap bubble has index of refraction 1.36 and thickness 380 nm.

80. A Michelson interferometer with a He-Ne laser light source ($\lambda = 632.8\text{ nm}$) projects its interference pattern on a screen. If the movable mirror is caused to move by $8.54\ \mu\text{m}$, how many fringes will be observed shifting through a reference point on a screen?

81. An experimenter detects 251 fringes when the movable mirror in a Michelson interferometer is displaced. The light source used is a sodium lamp, wavelength 589 nm. By what distance did the movable mirror move?

82. A Michelson interferometer is used to measure the wavelength of light put through it. When the movable mirror is moved by exactly 0.100 mm, the number of fringes observed moving through is 316. What is the wavelength of the light?

83. A 5.08-cm-long rectangular glass chamber is inserted into one arm of a Michelson interferometer using a 633-nm light source. This chamber is initially filled with air $(n = 1.000293)$ at standard atmospheric pressure but the air is gradually pumped out using a vacuum pump until a near perfect vacuum is achieved. How many fringes are observed moving by during the transition?

84. Into one arm of a Michelson interferometer, a plastic sheet of thickness $75\ \mu\text{m}$ is inserted, which causes a shift in the interference pattern by 86 fringes. The light source has wavelength of 610 nm in air. What is the index of refraction of this plastic?

85. The thickness of an aluminum foil is measured using a Michelson interferometer that has its movable mirror mounted on a micrometer. There is a difference of 27 fringes in the observed interference pattern when the micrometer clamps down on the foil compared to when the micrometer is empty. Calculate the thickness of the foil?

86. The movable mirror of a Michelson interferometer is attached to one end of a thin metal rod of length 23.3 mm. The other end of the rod is anchored so it does not move. As the temperature of the rod changes from $15\ °\text{C}$ to $25\,\text{C}$, a change of 14 fringes is observed. The light source is a He Ne laser, $\lambda = 632.8\ \text{nm}$. What is the change in length of the metal bar, and what is its thermal expansion coefficient?

87. In a thermally stabilized lab, a Michelson interferometer is used to monitor the temperature to ensure it stays constant. The movable mirror is mounted on the end of a 1.00-m-long aluminum rod, held fixed at the other end. The light source is a He Ne laser, $\lambda = 632.8\ \text{nm}$. The resolution of this apparatus corresponds to the temperature difference when a change of just one fringe is observed. What is this temperature difference?

88. A 65-fringe shift results in a Michelson interferometer when a $42.0\text{-}\mu\text{m}$ film made of an unknown material is placed in one arm. The light source has wavelength 632.9 nm. Identify the material using the indices of refraction found in **Table 30.1**.

CHALLENGE PROBLEMS

89. Determine what happens to the double-slit interference pattern if one of the slits is covered with a thin, transparent film whose thickness is $\lambda/[2(n-1)]$, where λ is the wavelength of the incident light and n is the index of refraction of the film.

90. Fifty-one narrow slits are equally spaced and separated by 0.10 mm. The slits are illuminated by blue light of wavelength 400 nm. What is angular position of the twenty-fifth secondary maximum? What is its peak intensity in comparison with that of the primary maximum?

91. A film of oil on water will appear dark when it is very thin, because the path length difference becomes small compared with the wavelength of light and there is a phase shift at the top surface. If it becomes dark when the path length difference is less than one-fourth the wavelength, what is the thickest the oil can be and appear dark at all visible wavelengths? Oil has an index of refraction of 1.40.

92. **Figure 32.14** shows two glass slides illuminated by monochromatic light incident perpendicularly. The top slide touches the bottom slide at one end and rests on a 0.100-mm-diameter hair at the other end, forming a wedge of air. (a) How far apart are the dark bands, if the slides are 7.50 cm long and 589-nm light is used? (b) Is there any difference if the slides are made from crown or flint glass? Explain.

93. **Figure 32.14** shows two 7.50-cm-long glass slides illuminated by pure 589-nm wavelength light incident perpendicularly. The top slide touches the bottom slide at one end and rests on some debris at the other end, forming a wedge of air. How thick is the debris, if the dark bands are 1.00 mm apart?

94. A soap bubble is 100 nm thick and illuminated by white light incident at a $45°$ angle to its surface. What wavelength and color of visible light is most constructively reflected, assuming the same index of refraction as water?

95. An oil slick on water is 120 nm thick and illuminated by white light incident at a $45°$ angle to its surface. What color does the oil appear (what is the most constructively reflected wavelength), given its index of refraction is 1.40?

33 | DIFFRACTION AND POLARIZATION

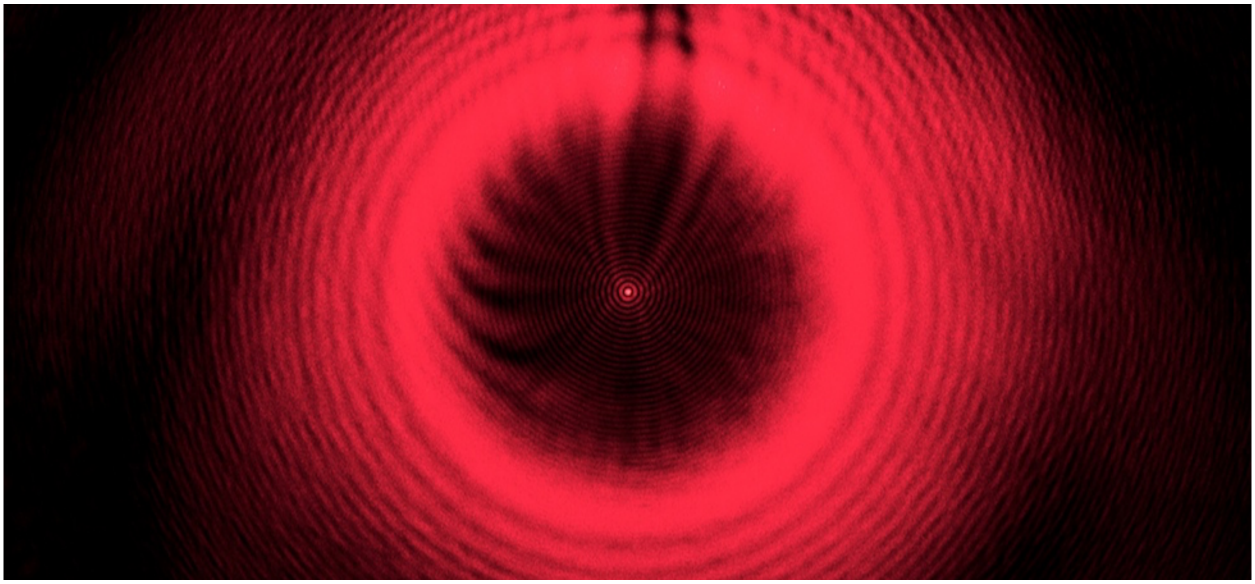


Figure 33.1 A steel ball bearing illuminated by a laser does not cast a sharp, circular shadow. Instead, a series of diffraction fringes and a central bright spot are observed. Known as Poisson’s spot, the effect was first predicted by Augustin-Jean Fresnel (1788–1827) as a consequence of diffraction of light waves. Based on principles of ray optics, Siméon-Denis Poisson (1781–1840) argued against Fresnel’s prediction. (credit: modification of work by Harvard Natural Science Lecture Demonstrations)

Chapter Outline

Introduction

Imagine passing a monochromatic light beam through a narrow opening—a slit just a little wider than the wavelength of the light. Instead of a simple shadow of the slit on the screen, you will see that an interference pattern appears, even though there is only one slit.

In the chapter on interference, we saw that you need two sources of waves for interference to occur. How can there be an interference pattern when we have only one slit? In **Huygens' Principle**, we learned that, due to Huygens’s principle, we can imagine a wave front as equivalent to infinitely many point sources of waves. Thus, a wave from a slit can behave not as one wave but as an infinite number of point sources. These waves can interfere with each other, resulting in an interference pattern without the presence of a second slit. This phenomenon is called *diffraction*.

Another way to view this is to recognize that a slit has a small but finite width. In the preceding chapter, we implicitly regarded slits as objects with positions but no size. The widths of the slits were considered negligible. When the slits have finite widths, each point along the opening can be considered a point source of light—a foundation of Huygens’s principle.

Because real-world optical instruments must have finite apertures (otherwise, no light can enter), diffraction plays a major role in the way we interpret the output of these optical instruments. For example, diffraction places limits on our ability to resolve images or objects. This is a problem that we will study later in this chapter.

33.1 | Single-Slit Diffraction

Learning Objectives

By the end of this section, you will be able to:

- Explain the phenomenon of diffraction and the conditions under which it is observed
- Describe diffraction through a single slit

After passing through a narrow aperture (opening), a wave propagating in a specific direction tends to spread out. For example, sound waves that enter a room through an open door can be heard even if the listener is in a part of the room where the geometry of ray propagation dictates that there should only be silence. Similarly, ocean waves passing through an opening in a breakwater can spread throughout the bay inside. (Figure 33.2). The spreading and bending of sound and ocean waves are two examples of **diffraction**, which is the bending of a wave around the edges of an opening or an obstacle—a phenomenon exhibited by all types of waves.

Figure 33.2 Because of the diffraction of waves, ocean waves entering through an opening in a breakwater can spread throughout the bay. (credit: modification of map data from Google Earth)

The diffraction of sound waves is apparent to us because wavelengths in the audible region are approximately the same size as the objects they encounter, a condition that must be satisfied if diffraction effects are to be observed easily. Since the wavelengths of visible light range from approximately 390 to 770 nm, most objects do not diffract light significantly. However, situations do occur in which apertures are small enough that the diffraction of light is observable. For example, if you place your middle and index fingers close together and look through the opening at a light bulb, you can see a rather clear diffraction pattern, consisting of light and dark lines running parallel to your fingers.

Diffraction through a Single Slit

Light passing through a single slit forms a diffraction pattern somewhat different from those formed by double slits or diffraction gratings, which we discussed in the chapter on interference. Figure 33.3 shows a single-slit diffraction pattern. Note that the central maximum is larger than maxima on either side and that the intensity decreases rapidly on either side. In contrast, a diffraction grating (Diffraction Gratings) produces evenly spaced lines that dim slowly on either side of the center.

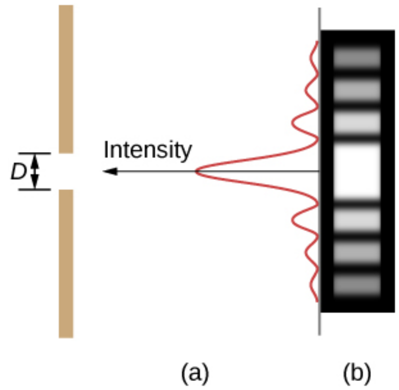


Figure 33.3 Single-slit diffraction pattern. (a) Monochromatic light passing through a single slit has a central maximum and many smaller and dimmer maxima on either side. The central maximum is six times higher than shown. (b) The diagram shows the bright central maximum, and the dimmer and thinner maxima on either side.

The analysis of single-slit diffraction is illustrated in **Figure 33.4**. Here, the light arrives at the slit, illuminating it uniformly and is in phase across its width. We then consider light propagating onwards from different parts of the *same* slit. According to Huygens's principle, every part of the wave front in the slit emits wavelets, as we discussed in **Huygens' Principle**. These are like rays that start out in phase and head in all directions. (Each ray is perpendicular to the wave front of a wavelet.) Assuming the screen is very far away compared with the size of the slit, rays heading toward a common destination are nearly parallel. When they travel straight ahead, as in part (a) of the figure, they remain in phase, and we observe a central maximum. However, when rays travel at an angle θ relative to the original direction of the beam, each ray travels a different distance to a common location, and they can arrive in or out of phase. In part (b), the ray from the bottom travels a distance of one wavelength λ farther than the ray from the top. Thus, a ray from the center travels a distance $\lambda/2$ less than the one at the bottom edge of the slit, arrives out of phase, and interferes destructively. A ray from slightly above the center and one from slightly above the bottom also cancel one another. In fact, each ray from the slit interferes destructively with another ray. In other words, a pair-wise cancellation of all rays results in a dark minimum in intensity at this angle. By symmetry, another minimum occurs at the same angle to the right of the incident direction (toward the bottom of the figure) of the light.

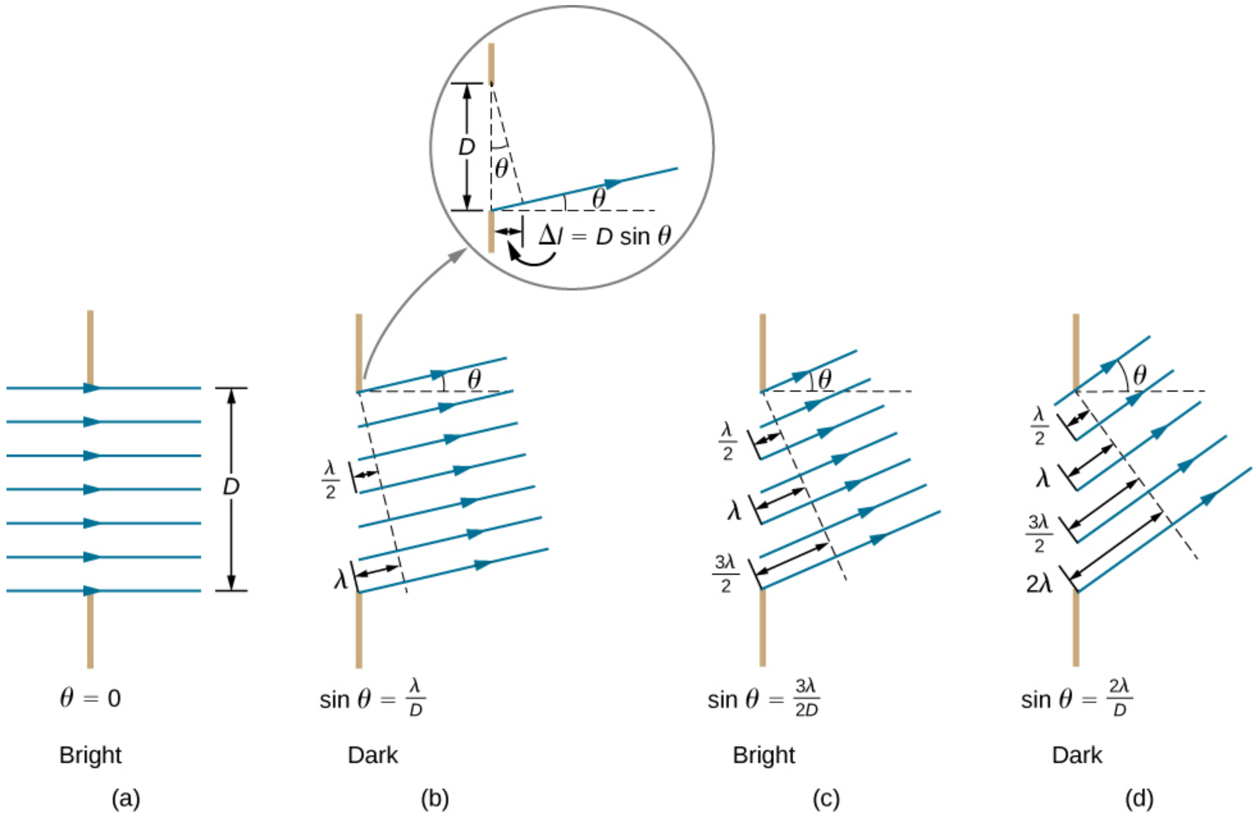


Figure 33.4 Light passing through a single slit is diffracted in all directions and may interfere constructively or destructively, depending on the angle. The difference in path length for rays from either side of the slit is seen to be $D \sin \theta$.

At the larger angle shown in part (c), the path lengths differ by $3\lambda/2$ for rays from the top and bottom of the slit. One ray travels a distance λ different from the ray from the bottom and arrives in phase, interfering constructively. Two rays, each from slightly above those two, also add constructively. Most rays from the slit have another ray to interfere with constructively, and a maximum in intensity occurs at this angle. However, not all rays interfere constructively for this situation, so the maximum is not as intense as the central maximum. Finally, in part (d), the angle shown is large enough to produce a second minimum. As seen in the figure, the difference in path length for rays from either side of the slit is $D \sin \theta$, and we see that a destructive minimum is obtained when this distance is an integral multiple of the wavelength.

Thus, to obtain **destructive interference for a single slit**,

$$D \sin \theta = m\lambda, \text{ for } m = \pm 1, \pm 2, \pm 3, \ldots \text{(destructive)}, \tag{33.1}$$

where D is the slit width, λ is the light's wavelength, θ is the angle relative to the original direction of the light, and m is the order of the minimum. **Figure 33.5** shows a graph of intensity for single-slit interference, and it is apparent that the maxima on either side of the central maximum are much less intense and not as wide. This effect is explored in **Double-Slit Diffraction**.

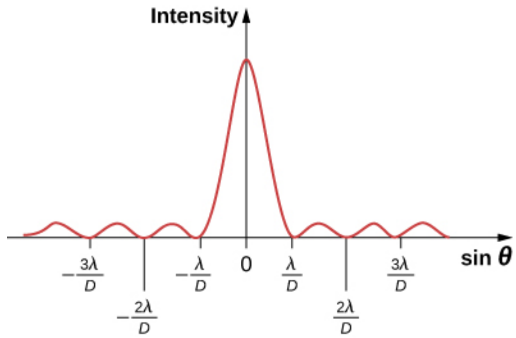


Figure 33.5 A graph of single-slit diffraction intensity showing the central maximum to be wider and much more intense than those to the sides. In fact, the central maximum is six times higher than shown here.

Example 33.1

Calculating Single-Slit Diffraction

Visible light of wavelength 550 nm falls on a single slit and produces its second diffraction minimum at an angle of $45.0°$ relative to the incident direction of the light, as in **Figure 33.6**. (a) What is the width of the slit? (b) At what angle is the first minimum produced?

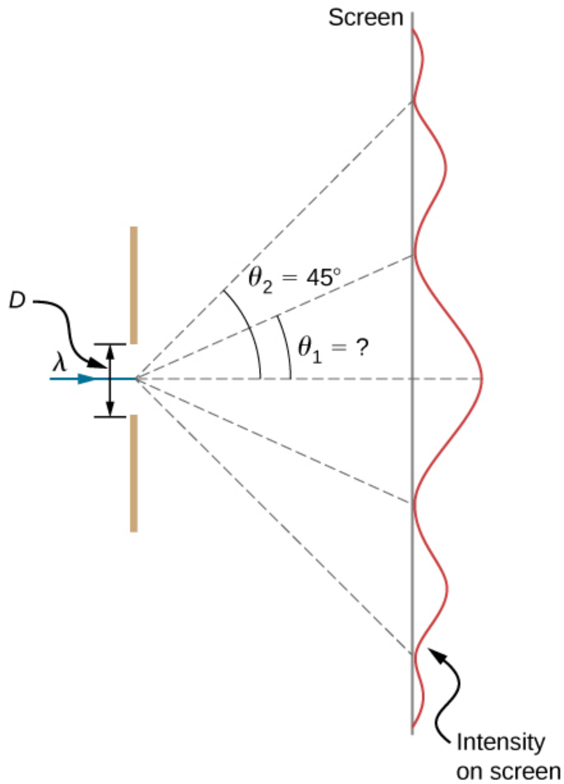


Figure 33.6 In this example, we analyze a graph of the single-slit diffraction pattern.

Strategy

From the given information, and assuming the screen is far away from the slit, we can use the equation $D \sin\theta = m\lambda$ first to find *D*, and again to find the angle for the first minimum θ_1.

Solution

a. We are given that $\lambda = 550\text{ nm}$, $m = 2$, and $\theta_2 = 45.0°$. Solving the equation $D \sin\theta = m\lambda$ for D and substituting known values gives

$$D = \frac{m\lambda}{\sin\theta_2} = \frac{2(550\text{ nm})}{\sin 45.0°} = \frac{1100 \times 10^{-9}\text{ m}}{0.707} = 1.56 \times 10^{-6}\text{ m}.$$

b. Solving the equation $D \sin\theta = m\lambda$ for $\sin\theta_1$ and substituting the known values gives

$$\sin\theta_1 = \frac{m\lambda}{D} = \frac{1(550 \times 10^{-9}\text{ m})}{1.56 \times 10^{-6}\text{ m}}.$$

Thus the angle θ_1 is

$$\theta_1 = \sin^{-1} 0.354 = 20.7°.$$

Significance

We see that the slit is narrow (it is only a few times greater than the wavelength of light). This is consistent with the fact that light must interact with an object comparable in size to its wavelength in order to exhibit significant wave effects such as this single-slit diffraction pattern. We also see that the central maximum extends $20.7°$ on either side of the original beam, for a width of about $41°$. The angle between the first and second minima is only about $24°$ $(45.0° - 20.7°)$. Thus, the second maximum is only about half as wide as the central maximum.

33.1 Check Your Understanding Suppose the slit width in **Example 33.1** is increased to 1.8×10^{-6} m. What are the new angular positions for the first, second, and third minima? Would a fourth minimum exist?

33.2 | Intensity in Single-Slit Diffraction

Learning Objectives

By the end of this section, you will be able to:

- Calculate the intensity relative to the central maximum of the single-slit diffraction peaks
- Calculate the intensity relative to the central maximum of an arbitrary point on the screen

If we consider that there are N Huygens sources across the slit shown in **Figure 33.4**, with each source separated by a distance D/N from its adjacent neighbors, the path difference between waves from adjacent sources reaching the arbitrary point P on the screen is $(D/N)\sin\theta$. This distance is equivalent to a phase difference of $(2\pi D/\lambda N)\sin\theta$. .

$$\phi = \left(\frac{2\pi}{\lambda}\right) D \sin\theta.$$

Now defining

$$\beta = \frac{\phi}{2} = \frac{\pi D \sin\theta}{\lambda} \tag{33.2}$$

we can add the waves (electric fields) from each source to obtain a total electric field amplitude of

$$E = N\Delta E_0 \frac{\sin\beta}{\beta} \tag{33.3}$$

This equation relates the amplitude of the resultant field at any point in the diffraction pattern to the amplitude $N\Delta E_0$ at the central maximum. The intensity is proportional to the square of the amplitude, so

$$I = I_0 \left(\frac{\sin\beta}{\beta}\right)^2 \tag{33.4}$$

where $I_0 = (N\Delta E_0)^2 / 2\mu_0 c$ is the intensity at the center of the pattern.

For the central maximum, $\phi = 0$, β is also zero and we see from l'Hôpital's rule that $\lim_{\beta \to 0}(\sin\beta/\beta) = 1,$ so that $\lim_{\phi \to 0} I = I_0.$ For the next maximum, $\phi = 3\pi$ rad, we have $\beta = 3\pi/2$ rad and when substituted into **Equation 33.4**, it yields

$$I_1 = I_0 \left(\frac{\sin 3\pi/2}{3\pi/2}\right)^2 \approx 0.045 I_0,$$

in agreement with what we found earlier in this section using the diameters and circumferences of phasor diagrams. Substituting $\phi = 5\pi$ rad into **Equation 33.4** yields a similar result for I_2.

A plot of **Equation 33.4** is shown in **Figure 33.7** and directly below it is a photograph of an actual diffraction pattern. Notice that the central peak is much brighter than the others, and that the zeros of the pattern are located at those points where $\sin\beta = 0,$ which occurs when $\beta = m\pi$ rad. This corresponds to

$$\frac{\pi D \sin\theta}{\lambda} = m\pi,$$

or

$$D \sin\theta = m\lambda,$$

which is **Equation 33.1**.

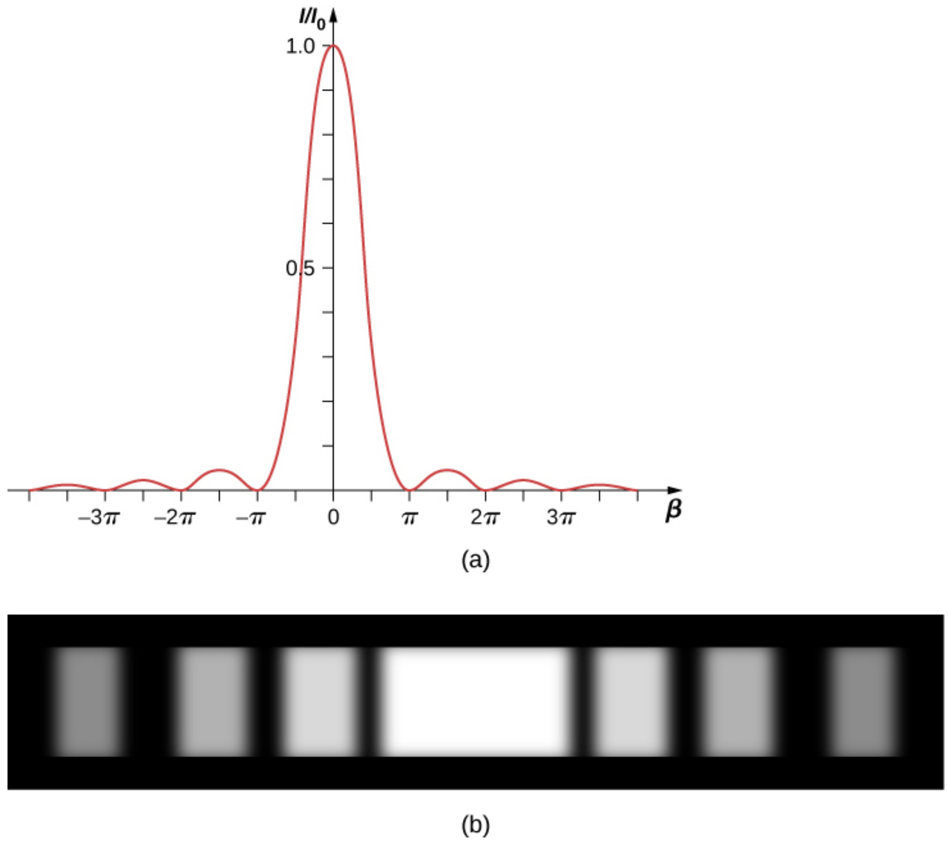


Figure 33.7 (a) The calculated intensity distribution of a single-slit diffraction pattern. (b) The actual diffraction pattern.

Example 33.2

Intensity in Single-Slit Diffraction

Light of wavelength 550 nm passes through a slit of width $2.00\ \mu\text{m}$ and produces a diffraction pattern similar to that shown in **Figure 33.7**. (a) Find the locations of the first two minima in terms of the angle from the central maximum and (b) determine the intensity relative to the central maximum at a point halfway between these two minima.

Strategy

The minima are given by **Equation 33.1**, $D\sin\theta = m\lambda$. The first two minima are for $m = 1$ and $m = 2$. **Equation 33.4** and **Equation 33.2** can be used to determine the intensity once the angle has been worked out.

Solution

a. Solving **Equation 33.1** for θ gives us $\theta_m = \sin^{-1}(m\lambda/D)$, so that

$$\theta_1 = \sin^{-1}\left(\frac{(+1)(550 \times 10^{-9}\ \text{m})}{2.00 \times 10^{-6}\ \text{m}}\right) = +16.0°$$

and

$$\theta_2 = \sin^{-1}\left(\frac{(+2)(550 \times 10^{-9}\ \text{m})}{2.00 \times 10^{-6}\ \text{m}}\right) = +33.4°.$$

b. The halfway point between θ_1 and θ_2 is

$$\theta = (\theta_1 + \theta_2)/2 = (16.0° + 33.4°)/2 = 24.7°.$$

Equation 33.2 gives

$$\beta = \frac{\pi D \sin\theta}{\lambda} = \frac{\pi\left(2.00 \times 10^{-6}\text{ m}\right)\sin(24.7°)}{\left(550 \times 10^{-9}\text{ m}\right)} = 1.52\pi \text{ or } 4.77\text{ rad}.$$

From **Equation 33.4**, we can calculate

$$\frac{I}{I_o} = \left(\frac{\sin\beta}{\beta}\right)^2 = \left(\frac{\sin(4.77)}{4.77}\right)^2 = \left(\frac{-0.9985}{4.77}\right)^2 = 0.044.$$

Significance

This position, halfway between two minima, is very close to the location of the maximum, expected near $\beta = 3\pi/2$, or 1.5π.

33.2 Check Your Understanding For the experiment in **Example 33.2**, at what angle from the center is the third maximum and what is its intensity relative to the central maximum?

If the slit width D is varied, the intensity distribution changes, as illustrated in **Figure 33.8**. The central peak is distributed over the region from $\sin\theta = -\lambda/D$ to $\sin\theta = +\lambda/D$. For small θ, this corresponds to an angular width $\Delta\theta \approx 2\lambda/D$. Hence, an increase in the slit width results in a decrease in the **width of the central peak**. For a slit with $D \gg \lambda$, the central peak is very sharp, whereas if $D \approx \lambda$, it becomes quite broad.

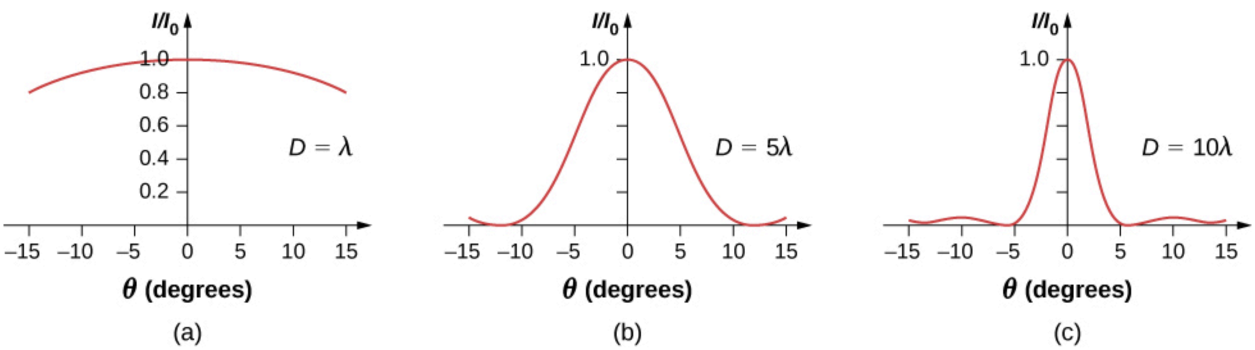


Figure 33.8 Single-slit diffraction patterns for various slit widths. As the slit width D increases from $D = \lambda$ to 5λ and then to 10λ, the width of the central peak decreases as the angles for the first minima decrease as predicted by **Equation 33.1**.

A diffraction experiment in optics can require a lot of preparation but **this simulation (https://openstax.org/l/21diffrexpoptsi)** by Andrew Duffy offers not only a quick set up but also the ability to change the slit width instantly. Run the simulation and select "Single slit." You can adjust the slit width and see the effect on the diffraction pattern on a screen and as a graph.

33.3 | Double-Slit Diffraction

Learning Objectives

By the end of this section, you will be able to:

- Describe the combined effect of interference and diffraction with two slits, each with finite width
- Determine the relative intensities of interference fringes within a diffraction pattern
- Identify missing orders, if any

When we studied interference in Young's double-slit experiment, we ignored the diffraction effect in each slit. We assumed that the slits were so narrow that on the screen you saw only the interference of light from just two point sources. If the slit is smaller than the wavelength, then Figure 33.8(a) shows that there is just a spreading of light and no peaks or troughs on the screen. Therefore, it was reasonable to leave out the diffraction effect in that chapter. However, if you make the slit wider, Figure 33.8(b) and (c) show that you cannot ignore diffraction. In this section, we study the complications to the double-slit experiment that arise when you also need to take into account the diffraction effect of each slit.

To calculate the diffraction pattern for two (or any number of) slits, we need to generalize the method we just used for a single slit. That is, across each slit, we place a uniform distribution of point sources that radiate Huygens wavelets, and then we sum the wavelets from all the slits. This gives the intensity at any point on the screen. Although the details of that calculation can be complicated, the final result is quite simple:

Two-Slit Diffraction Pattern

The diffraction pattern of two slits of width D that are separated by a distance d is the interference pattern of two point sources separated by d multiplied by the diffraction pattern of a slit of width D.

In other words, the *locations* of the interference fringes are given by the equation $d\sin\theta = m\lambda$, the same as when we considered the slits to be point sources, but the *intensities* of the fringes are now reduced by diffraction effects, according to Equation 33.4. [Note that in the chapter on interference, we wrote $d\sin\theta = m\lambda$ and used the integer m to refer to interference fringes. Equation 33.1 also uses m, but this time to refer to diffraction minima. If both equations are used simultaneously, it is good practice to use a different variable (such as n) for one of these integers in order to keep them distinct.]

Interference and diffraction effects operate simultaneously and generally produce minima at different angles. This gives rise to a complicated pattern on the screen, in which some of the maxima of interference from the two slits are missing if the maximum of the interference is in the same direction as the minimum of the diffraction. We refer to such a missing peak as a **missing order**. One example of a diffraction pattern on the screen is shown in Figure 33.9. The solid line with multiple peaks of various heights is the intensity observed on the screen. It is a product of the interference pattern of waves from separate slits and the diffraction of waves from within one slit.

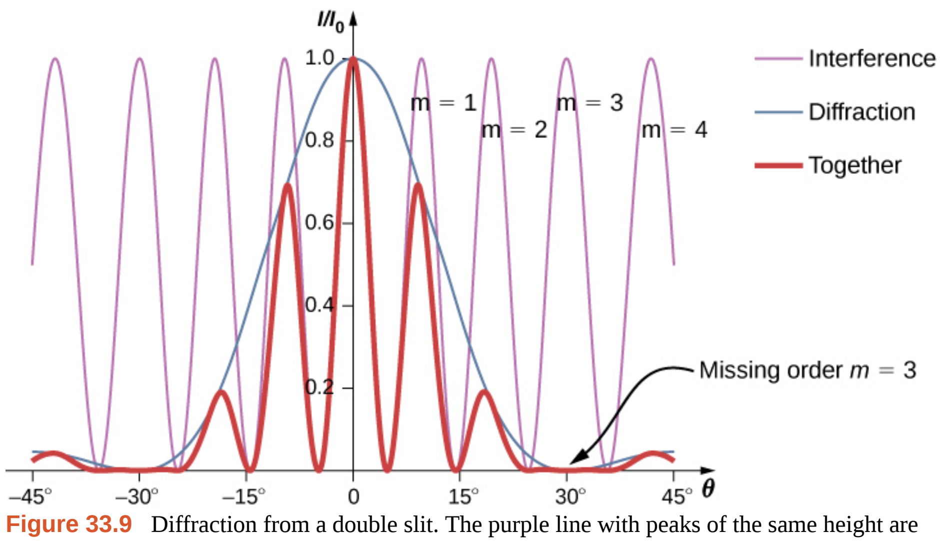


Figure 33.9 Diffraction from a double slit. The purple line with peaks of the same height are from the interference of the waves from two slits; the blue line with one big hump in the middle is the diffraction of waves from within one slit; and the thick red line is the product of the two, which is the pattern observed on the screen. The plot shows the expected result for a slit width $D = 2\lambda$ and slit separation $d = 6\lambda$. The maximum of $m = \pm 3$ order for the interference is missing because the minimum of the diffraction occurs in the same direction.

Example 33.3

Intensity of the Fringes

Figure 33.9 shows that the intensity of the fringe for $m = 3$ is zero, but what about the other fringes? Calculate the intensity for the fringe at $m = 1$ relative to I_0, the intensity of the central peak.

Strategy

Determine the angle for the double-slit interference fringe, using the equation from Interference, then determine the relative intensity in that direction due to diffraction by using Equation 33.4.

Solution

From the chapter on interference, we know that the bright interference fringes occur at $d \sin\theta = m\lambda$, or

$$\sin\theta = \frac{m\lambda}{d}.$$

From Equation 33.4,

$$I = I_0\left(\frac{\sin\beta}{\beta}\right)^2, \text{ where } \beta = \frac{\phi}{2} = \frac{\pi D \sin\theta}{\lambda}.$$

Substituting from above,

$$\beta = \frac{\pi D \sin\theta}{\lambda} = \frac{\pi D}{\lambda}\cdot\frac{m\lambda}{d} = \frac{m\pi D}{d}.$$

For $D = 2\lambda$, $d = 6\lambda$, and $m = 1$,

$$\beta = \frac{(1)\pi(2\lambda)}{(6\lambda)} = \frac{\pi}{3}.$$

Then, the intensity is

$$I = I_0\left(\frac{\sin\beta}{\beta}\right)^2 = I_0\left(\frac{\sin(\pi/3)}{\pi/3}\right)^2 = 0.684 I_0.$$

Significance

Note that this approach is relatively straightforward and gives a result that is almost exactly the same as the more complicated analysis using phasors to work out the intensity values of the double-slit interference (thin line in Figure 33.9). The phasor approach accounts for the downward slope in the diffraction intensity (blue line) so that the peak *near* $m = 1$ occurs at a value of θ ever so slightly smaller than we have shown here.

Example 33.4

Two-Slit Diffraction

Suppose that in Young's experiment, slits of width 0.020 mm are separated by 0.20 mm. If the slits are illuminated by monochromatic light of wavelength 500 nm, how many bright fringes are observed in the central peak of the diffraction pattern?

Solution

From Equation 33.1, the angular position of the first diffraction minimum is $\theta \approx \sin\theta = \frac{\lambda}{D} = \frac{5.0 \times 10^{-7}\text{ m}}{2.0 \times 10^{-5}\text{ m}} = 2.5 \times 10^{-2}\text{ rad}.$

Using $d \sin\theta = m\lambda$ for $\theta = 2.5 \times 10^{-2}$ rad, we find

$$m = \frac{d \sin \theta}{\lambda} = \frac{(0.20\ \text{mm})\left(2.5 \times 10^{-2}\ \text{rad}\right)}{\left(5.0 \times 10^{-7}\ \text{m}\right)} = 10,$$

which is the maximum interference order that fits inside the central peak. We note that $m = \pm 10$ are missing orders as θ matches exactly. Accordingly, we observe bright fringes for

$$m = -9,\ -8,\ -7,\ -6,\ -5,\ -4,\ -3,\ -2,\ -1,\ 0,\ +1,\ +2,\ +3,\ +4,\ +5,\ +6,\ +7,\ +8,\ \text{and}\ +9$$

for a total of 19 bright fringes.

33.3 Check Your Understanding For the experiment in **Example 33.4**, show that $m = 20$ is also a missing order.

Explore the effects of double-slit diffraction. In **this simulation (https://openstax.org/l/21doubslitdiff)** written by Fu-Kwun Hwang, select $N = 2$ using the slider and see what happens when you control the slit width, slit separation and the wavelength. Can you make an order go "missing?"

33.4 | Diffraction Gratings

Learning Objectives

By the end of this section, you will be able to:

- Discuss the pattern obtained from diffraction gratings
- Explain diffraction grating effects

Analyzing the interference of light passing through two slits lays out the theoretical framework of interference and gives us a historical insight into Thomas Young's experiments. However, most modern-day applications of slit interference use not just two slits but many, approaching infinity for practical purposes. The key optical element is called a diffraction grating, an important tool in optical analysis.

Diffraction Gratings: An Infinite Number of Slits

The analysis of multi-slit interference in **Interference** allows us to consider what happens when the number of slits N approaches infinity. Recall that $N - 2$ secondary maxima appear between the principal maxima. We can see there will be an infinite number of secondary maxima that appear, and an infinite number of dark fringes between them. This makes the spacing between the fringes, and therefore the width of the maxima, infinitesimally small. Furthermore, because the intensity of the secondary maxima is proportional to $1/N^2$, it approaches zero so that the secondary maxima are no longer seen. What remains are only the principal maxima, now very bright and very narrow (**Figure 33.10**).

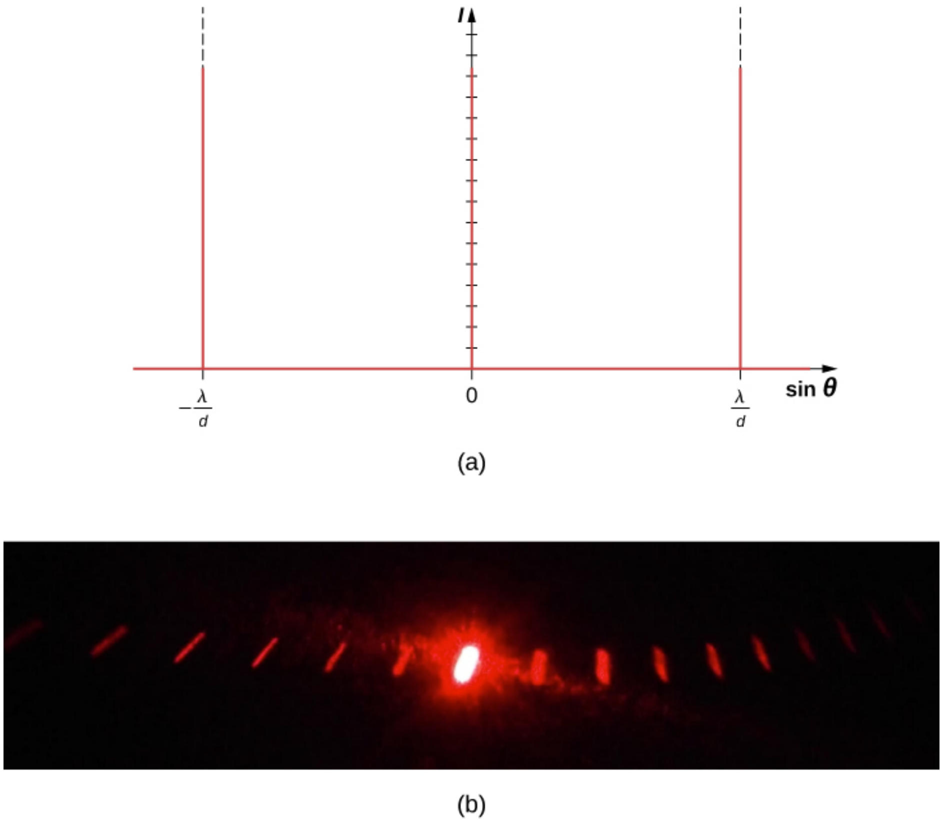


Figure 33.10 (a) Intensity of light transmitted through a large number of slits. When *N* approaches infinity, only the principal maxima remain as very bright and very narrow lines. (b) A laser beam passed through a diffraction grating. (credit b: modification of work by Sebastian Stapelberg)

In reality, the number of slits is not infinite, but it can be very large—large enough to produce the equivalent effect. A prime example is an optical element called a **diffraction grating**. A diffraction grating can be manufactured by carving glass with a sharp tool in a large number of precisely positioned parallel lines, with untouched regions acting like slits (**Figure 33.11**). This type of grating can be photographically mass produced rather cheaply. Because there can be over 1000 lines per millimeter across the grating, when a section as small as a few millimeters is illuminated by an incoming ray, the number of illuminated slits is effectively infinite, providing for very sharp principal maxima.

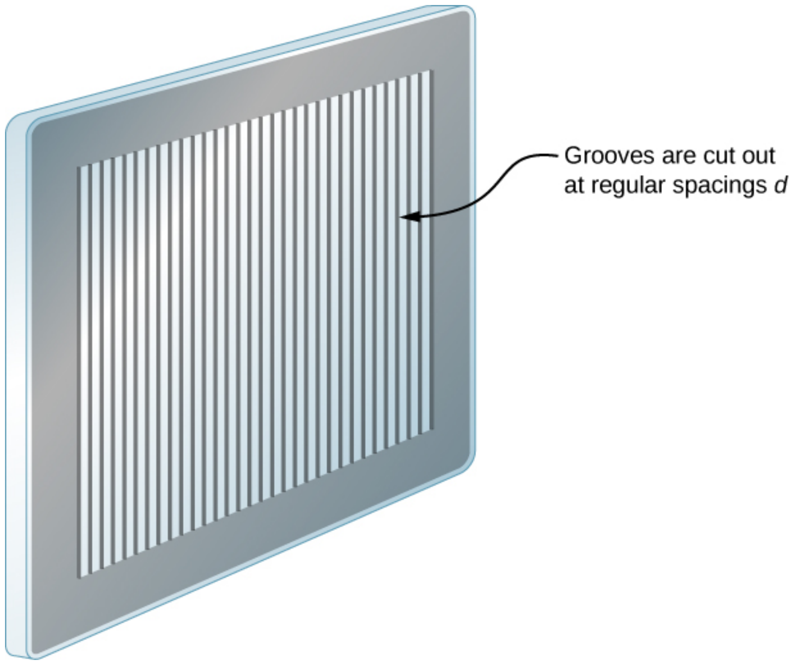


Figure 33.11 A diffraction grating can be manufactured by carving glass with a sharp tool in a large number of precisely positioned parallel lines.

Diffraction gratings work both for transmission of light, as in **Figure 33.12**, and for reflection of light, as on butterfly wings and the Australian opal in **Figure 33.13**. Natural diffraction gratings also occur in the feathers of certain birds such as the hummingbird. Tiny, finger-like structures in regular patterns act as reflection gratings, producing constructive interference that gives the feathers colors not solely due to their pigmentation. This is called iridescence.

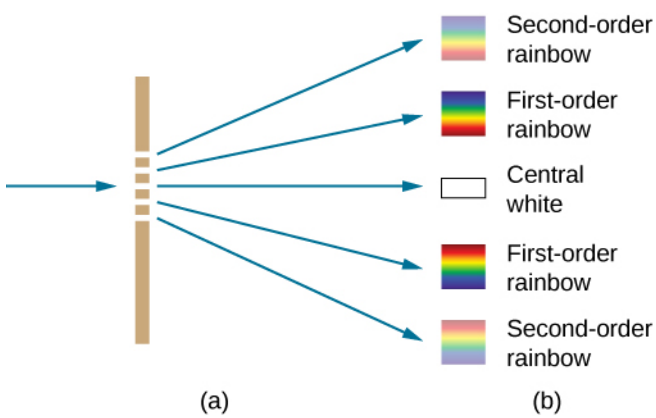


Figure 33.12 (a) Light passing through a diffraction grating is diffracted in a pattern similar to a double slit, with bright regions at various angles. (b) The pattern obtained for white light incident on a grating. The central maximum is white, and the higher-order maxima disperse white light into a rainbow of colors.

(a) (b)

Figure 33.13 (a) This Australian opal and (b) butterfly wings have rows of reflectors that act like reflection gratings, reflecting different colors at different angles. (credit a: modification of work by "Opals-On-Black"/Flickr; credit b: modification of work by “whologwhy”/Flickr)

Applications of Diffraction Gratings

Where are diffraction gratings used in applications? Diffraction gratings are commonly used for spectroscopic dispersion and analysis of light. What makes them particularly useful is the fact that they form a sharper pattern than double slits do. That is, their bright fringes are narrower and brighter while their dark regions are darker. Diffraction gratings are key components of monochromators used, for example, in optical imaging of particular wavelengths from biological or medical samples. A diffraction grating can be chosen to specifically analyze a wavelength emitted by molecules in diseased cells in a biopsy sample or to help excite strategic molecules in the sample with a selected wavelength of light. Another vital use is in optical fiber technologies where fibers are designed to provide optimum performance at specific wavelengths. A range of diffraction gratings are available for selecting wavelengths for such use.

Example 33.5

Calculating Typical Diffraction Grating Effects

Diffraction gratings with 10,000 lines per centimeter are readily available. Suppose you have one, and you send a beam of white light through it to a screen 2.00 m away. (a) Find the angles for the first-order diffraction of the shortest and longest wavelengths of visible light (380 and 760 nm, respectively). (b) What is the distance between the ends of the rainbow of visible light produced on the screen for first-order interference? (See **Figure 33.14**.)

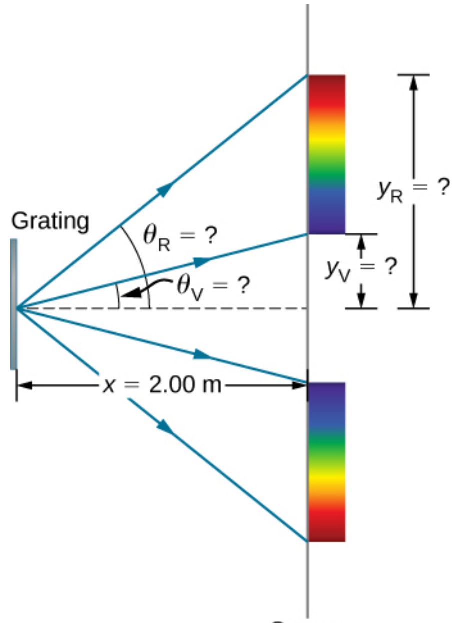


Figure 33.14 (a) The diffraction grating considered in this example produces a rainbow of colors on a screen a distance $x = 2.00 \text{ m}$ from the grating. The distances along the screen are measured perpendicular to the *x*-direction. In other words, the rainbow pattern extends out of the page.
(b) In a bird's-eye view, the rainbow pattern can be seen on a table where the equipment is placed.

Strategy

Once a value for the diffraction grating's slit spacing *d* has been determined, the angles for the sharp lines can be found using the equation

$$d \sin \theta = m\lambda \text{ for } m = 0, \ \pm 1, \ \pm 2, \ \ldots .$$

Since there are 10,000 lines per centimeter, each line is separated by 1/10,000 of a centimeter. Once we know the angles, we an find the distances along the screen by using simple trigonometry.

Solution

a. The distance between slits is $d = (1 \text{ cm})/10,000 = 1.00 \times 10^{-4} \text{ cm}$ or $1.00 \times 10^{-6} \text{ m}$. Let us call the two angles θ_V for violet (380 nm) and θ_R for red (760 nm). Solving the equation $d \sin \theta_V = m\lambda$ for $\sin \theta_V$,

$$\sin \theta_V = \frac{m\lambda_V}{d},$$

where $m = 1$ for the first-order and $\lambda_V = 380 \text{ nm} = 3.80 \times 10^{-7} \text{ m}$. Substituting these values gives

$$\sin \theta_V = \frac{3.80 \times 10^{-7} \text{ m}}{1.00 \times 10^{-6} \text{ m}} = 0.380.$$

Thus the angle θ_V is

$$\theta_V = \sin^{-1} 0.380 = 22.33°.$$

Similarly,

$$\sin\theta_{\text{R}} = \frac{7.60 \times 10^{-7}\ \text{m}}{1.00 \times 10^{-6}\ \text{m}} = 0.760.$$

Thus the angle θ_{R} is

$$\theta_{\text{R}} = \sin^{-1} 0.760 = 49.46°.$$

Notice that in both equations, we reported the results of these intermediate calculations to four significant figures to use with the calculation in part (b).

b. The distances on the secreen are labeled y_{V} and y_{R} in **Figure 33.14**. Notice that $\tan\theta = y/x$. We can solve for y_{V} and y_{R}. That is,

$$y_{\text{V}} = x\tan\theta_V = (2.00\ \text{m})(\tan 22.33°) = 0.815\ \text{m}$$

and

$$y_{\text{R}} = x\tan\theta_R = (2.00\ \text{m})(\tan 49.46°) = 2.338\ \text{m}.$$

The distance between them is therefore

$$y_{\text{R}} - y_{\text{V}} = 1.523\ \text{m}.$$

Significance

The large distance between the red and violet ends of the rainbow produced from the white light indicates the potential this diffraction grating has as a spectroscopic tool. The more it can spread out the wavelengths (greater dispersion), the more detail can be seen in a spectrum. This depends on the quality of the diffraction grating—it must be very precisely made in addition to having closely spaced lines.

33.4 Check Your Understanding If the line spacing of a diffraction grating *d* is not precisely known, we can use a light source with a well-determined wavelength to measure it. Suppose the first-order constructive fringe of the H_β emission line of hydrogen $(\lambda = 656.3\ \text{nm})$ is measured at $11.36°$ using a spectrometer with a diffraction grating. What is the line spacing of this grating?

Take **the same simulation (https://openstax.org/l/21doubslitdiff)** we used for double-slit diffraction and try increasing the number of slits from $N = 2$ to $N = 3, 4, 5...$. The primary peaks become sharper, and the secondary peaks become less and less pronounced. By the time you reach the maximum number of $N = 20$, the system is behaving much like a diffraction grating.

33.5 | Circular Apertures and Resolution

Learning Objectives

By the end of this section, you will be able to:

- Describe the diffraction limit on resolution
- Describe the diffraction limit on beam propagation

Light diffracts as it moves through space, bending around obstacles, interfering constructively and destructively. This can be used as a spectroscopic tool—a diffraction grating disperses light according to wavelength, for example, and is used to produce spectra—but diffraction also limits the detail we can obtain in images.

Figure 33.15(a) shows the effect of passing light through a small circular aperture. Instead of a bright spot with sharp edges, we obtain a spot with a fuzzy edge surrounded by circles of light. This pattern is caused by diffraction, similar to that produced by a single slit. Light from different parts of the circular aperture interferes constructively and destructively. The effect is most noticeable when the aperture is small, but the effect is there for large apertures as well.

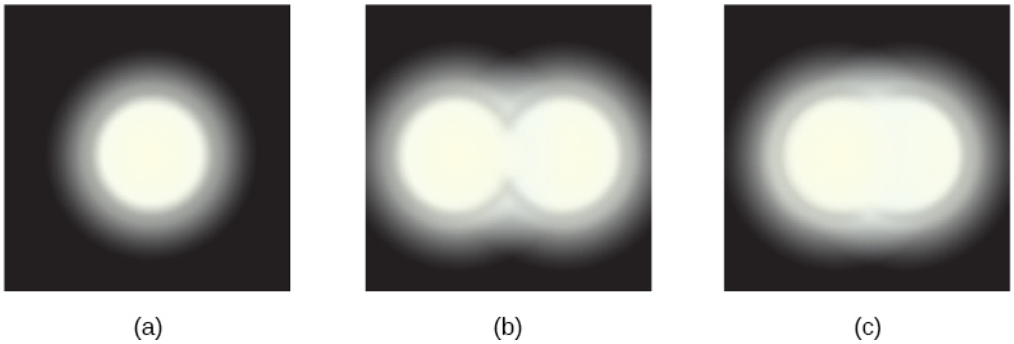


Figure 33.15 (a) Monochromatic light passed through a small circular aperture produces this diffraction pattern. (b) Two point-light sources that are close to one another produce overlapping images because of diffraction. (c) If the sources are closer together, they cannot be distinguished or resolved.

How does diffraction affect the detail that can be observed when light passes through an aperture? **Figure 33.15**(b) shows the diffraction pattern produced by two point-light sources that are close to one another. The pattern is similar to that for a single point source, and it is still possible to tell that there are two light sources rather than one. If they are closer together, as in **Figure 33.15**(c), we cannot distinguish them, thus limiting the detail or **resolution** we can obtain. This limit is an inescapable consequence of the wave nature of light.

Diffraction limits the resolution in many situations. The acuity of our vision is limited because light passes through the pupil, which is the circular aperture of the eye. Be aware that the diffraction-like spreading of light is due to the limited diameter of a light beam, not the interaction with an aperture. Thus, light passing through a lens with a diameter D shows this effect and spreads, blurring the image, just as light passing through an aperture of diameter D does. Thus, diffraction limits the resolution of any system having a lens or mirror. Telescopes are also limited by diffraction, because of the finite diameter D of the primary mirror.

Just what is the limit? To answer that question, consider the diffraction pattern for a circular aperture, which has a central maximum that is wider and brighter than the maxima surrounding it (similar to a slit) (**Figure 33.16**(a)). It can be shown that, for a circular aperture of diameter D, the first minimum in the diffraction pattern occurs at $\theta = 1.22\lambda/D$ (providing the aperture is large compared with the wavelength of light, which is the case for most optical instruments). The accepted criterion for determining the **diffraction limit** to resolution based on this angle is known as the **Rayleigh criterion**, which was developed by Lord Rayleigh in the nineteenth century.

Rayleigh Criterion

The diffraction limit to resolution states that two images are just resolvable when the center of the diffraction pattern of one is directly over the first minimum of the diffraction pattern of the other (**Figure 33.16**(b)).

The first minimum is at an angle of $\theta = 1.22\lambda/D$, so that two point objects are just resolvable if they are separated by the angle

$$\theta = 1.22\frac{\lambda}{D} \tag{33.5}$$

where λ is the wavelength of light (or other electromagnetic radiation) and D is the diameter of the aperture, lens, mirror, etc., with which the two objects are observed. In this expression, θ has units of radians. This angle is also commonly known as the diffraction limit.

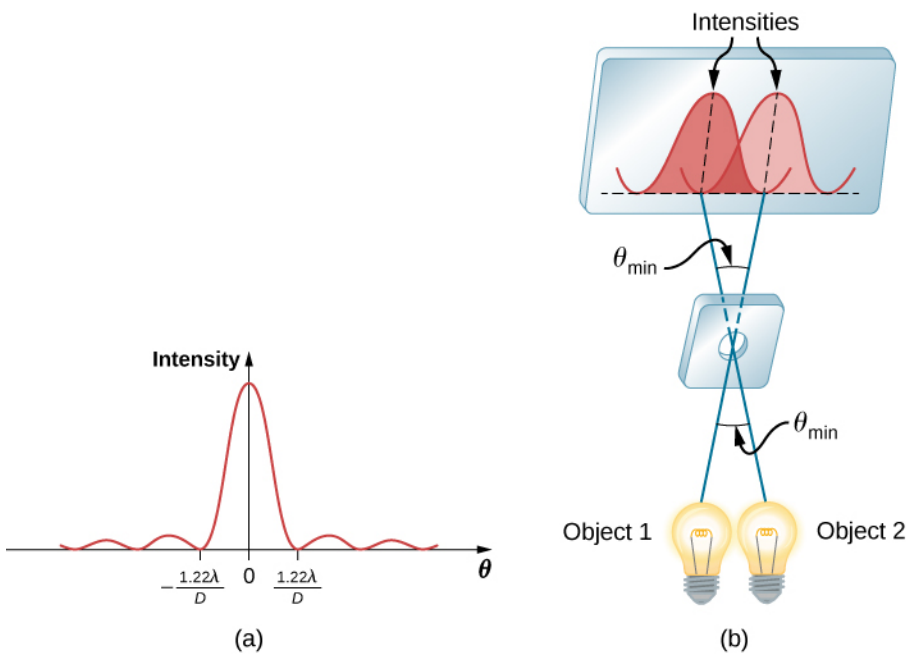


Figure 33.16 (a) Graph of intensity of the diffraction pattern for a circular aperture. Note that, similar to a single slit, the central maximum is wider and brighter than those to the sides. (b) Two point objects produce overlapping diffraction patterns. Shown here is the Rayleigh criterion for being just resolvable. The central maximum of one pattern lies on the first minimum of the other.

All attempts to observe the size and shape of objects are limited by the wavelength of the probe. Even the small wavelength of light prohibits exact precision. When extremely small wavelength probes are used, as with an electron microscope, the system is disturbed, still limiting our knowledge. Heisenberg's uncertainty principle asserts that this limit is fundamental and inescapable, as we shall see in the chapter on quantum mechanics.

Example 33.6

Calculating Diffraction Limits of the Hubble Space Telescope

The primary mirror of the orbiting Hubble Space Telescope has a diameter of 2.40 m. Being in orbit, this telescope avoids the degrading effects of atmospheric distortion on its resolution. (a) What is the angle between two just-resolvable point light sources (perhaps two stars)? Assume an average light wavelength of 550 nm. (b) If these two stars are at a distance of 2 million light-years, which is the distance of the Andromeda Galaxy, how close together can they be and still be resolved? (A light-year, or ly, is the distance light travels in 1 year.)

Strategy

The Rayleigh criterion stated in **Equation 33.5**, $\theta = 1.22\lambda/D$, gives the smallest possible angle θ between point sources, or the best obtainable resolution. Once this angle is known, we can calculate the distance between the stars, since we are given how far away they are.

Solution

a. The Rayleigh criterion for the minimum resolvable angle is

$$\theta = 1.22\frac{\lambda}{D}.$$

Entering known values gives

$$\theta = 1.22\frac{550 \times 10^{-9}\ \text{m}}{2.40\ \text{m}} = 2.80 \times 10^{-7}\ \text{rad}.$$

b. The distance s between two objects a distance r away and separated by an angle θ is $s = r\theta$.

Substituting known values gives

$$s = \left(2.0 \times 10^6 \text{ ly}\right)\left(2.80 \times 10^{-7} \text{ rad}\right) = 0.56 \text{ ly}.$$

Significance

The angle found in part (a) is extraordinarily small (less than 1/50,000 of a degree), because the primary mirror is so large compared with the wavelength of light. As noticed, diffraction effects are most noticeable when light interacts with objects having sizes on the order of the wavelength of light. However, the effect is still there, and there is a diffraction limit to what is observable. The actual resolution of the Hubble Telescope is not quite as good as that found here. As with all instruments, there are other effects, such as nonuniformities in mirrors or aberrations in lenses that further limit resolution. However, **Figure 33.17** gives an indication of the extent of the detail observable with the Hubble because of its size and quality, and especially because it is above Earth's atmosphere.

Figure 33.17 These two photographs of the M82 Galaxy give an idea of the observable detail using (a) a ground-based telescope and (b) the Hubble Space Telescope. (credit a: modification of work by "Ricnun"/Wikimedia Commons; credit b: modification of work by NASA, ESA, and The Hubble Heritage Team (STScI/AURA))

The answer in part (b) indicates that two stars separated by about half a light-year can be resolved. The average distance between stars in a galaxy is on the order of five light-years in the outer parts and about one light-year near the galactic center. Therefore, the Hubble can resolve most of the individual stars in Andromeda Galaxy, even though it lies at such a huge distance that its light takes 2 million years to reach us. **Figure 33.18** shows another mirror used to observe radio waves from outer space.

Figure 33.18 A 305-m-diameter paraboloid at Arecibo in Puerto Rico is lined with reflective material, making it into a radio telescope. It is the largest curved focusing dish in the world. Although *D* for Arecibo is much larger than for the Hubble Telescope, it detects radiation of a much longer wavelength and its diffraction limit is significantly poorer than Hubble's. The Arecibo telescope is still very useful, because important information is carried by radio waves that is not carried by visible light. (credit: Jeff Hitchcock)

33.5 Check Your Understanding What is the angular resolution of the Arecibo telescope shown in Figure 33.18 when operated at 21-cm wavelength? How does it compare to the resolution of the Hubble Telescope?

Diffraction is not only a problem for optical instruments but also for the electromagnetic radiation itself. Any beam of light having a finite diameter *D* and a wavelength λ exhibits diffraction spreading. The beam spreads out with an angle θ given by Equation 33.5, $\theta = 1.22\lambda/D$. Take, for example, a laser beam made of rays as parallel as possible (angles between rays as close to $\theta = 0°$ as possible) instead spreads out at an angle $\theta = 1.22\lambda/D$, where *D* is the diameter of the beam and λ is its wavelength. This spreading is impossible to observe for a flashlight because its beam is not very parallel to start with. However, for long-distance transmission of laser beams or microwave signals, diffraction spreading can be significant (Figure 33.19). To avoid this, we can increase *D*. This is done for laser light sent to the moon to measure its distance from Earth. The laser beam is expanded through a telescope to make *D* much larger and θ smaller.

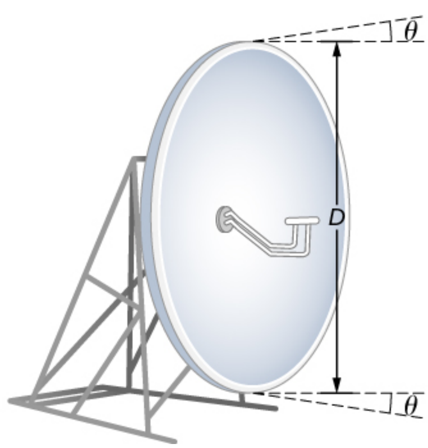


Figure 33.19 The beam produced by this microwave transmission antenna spreads out at a minimum angle $\theta = 1.22\lambda/D$ due to diffraction. It is impossible to produce a near-parallel beam because the beam has a limited diameter.

In most biology laboratories, resolution is an issue when the use of the microscope is introduced. The smaller the distance x by which two objects can be separated and still be seen as distinct, the greater the resolution. The resolving power of a lens is defined as that distance x. An expression for resolving power is obtained from the Rayleigh criterion. **Figure 33.20**(a) shows two point objects separated by a distance x. According to the Rayleigh criterion, resolution is possible when the minimum angular separation is

$$\theta = 1.22\frac{\lambda}{D} = \frac{x}{d},$$

where d is the distance between the specimen and the objective lens, and we have used the small angle approximation (i.e., we have assumed that x is much smaller than d), so that $\tan\theta \approx \sin\theta \approx \theta$. Therefore, the resolving power is

$$x = 1.22\frac{\lambda d}{D}.$$

Another way to look at this is by the concept of numerical aperture (NA), which is a measure of the maximum acceptance angle at which a lens will take light and still contain it within the lens. **Figure 33.20**(b) shows a lens and an object at point P. The NA here is a measure of the ability of the lens to gather light and resolve fine detail. The angle subtended by the lens at its focus is defined to be $\theta = 2\alpha$. From the figure and again using the small angle approximation, we can write

$$\sin\alpha = \frac{D/2}{d} = \frac{D}{2d}.$$

The NA for a lens is $NA = n\sin\alpha$, where n is the index of refraction of the medium between the objective lens and the object at point P. From this definition for NA, we can see that

$$x = 1.22\frac{\lambda d}{D} = 1.22\frac{\lambda}{2\sin\alpha} = 0.61\frac{\lambda n}{NA}.$$

In a microscope, NA is important because it relates to the resolving power of a lens. A lens with a large NA is able to resolve finer details. Lenses with larger NA are also able to collect more light and so give a brighter image. Another way to describe this situation is that the larger the NA, the larger the cone of light that can be brought into the lens, so more of the diffraction modes are collected. Thus the microscope has more information to form a clear image, and its resolving power is higher.

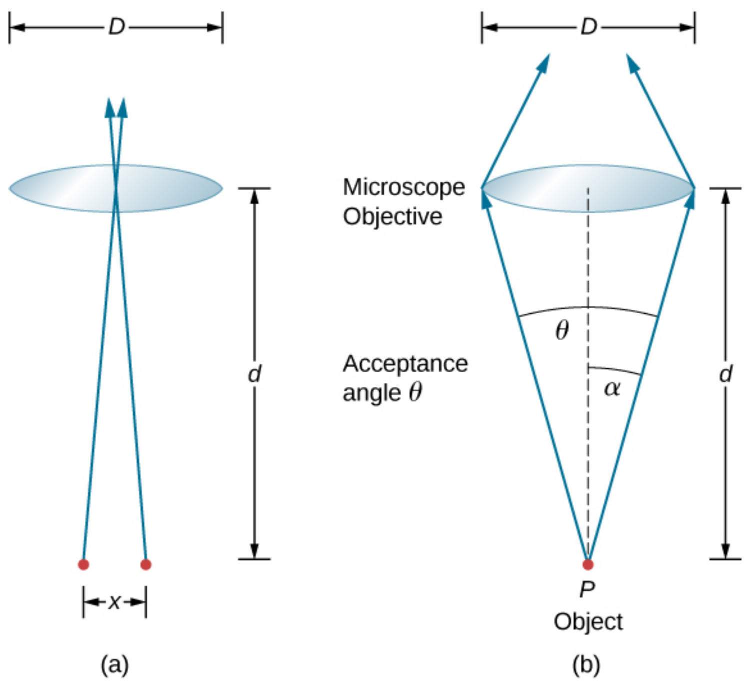


Figure 33.20 (a) Two points separated by a distance x and positioned a distance d away from the objective. (b) Terms and symbols used in discussion of resolving power for a lens and an object at point P (credit a: modification of work by “Infopro”/Wikimedia Commons).

One of the consequences of diffraction is that the focal point of a beam has a finite width and intensity distribution. Imagine focusing when only considering geometric optics, as in **Figure 33.21**(a). The focal point is regarded as an infinitely small point with a huge intensity and the capacity to incinerate most samples, irrespective of the *NA* of the objective lens—an unphysical oversimplification. For wave optics, due to diffraction, we take into account the phenomenon in which the focal point spreads to become a focal spot (**Figure 33.21**(b)) with the size of the spot decreasing with increasing *NA*. Consequently, the intensity in the focal spot increases with increasing *NA*. The higher the *NA*, the greater the chances of photodegrading the specimen. However, the spot never becomes a true point.

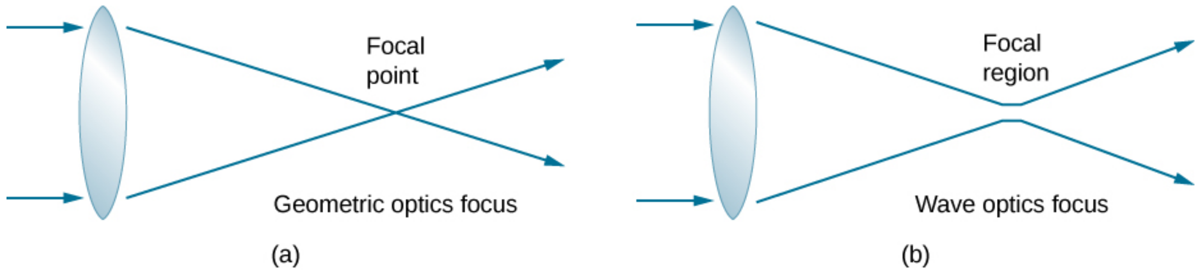


Figure 33.21 (a) In geometric optics, the focus is modelled as a point, but it is not physically possible to produce such a point because it implies infinite intensity. (b) In wave optics, the focus is an extended region.

In a different type of microscope, molecules within a specimen are made to emit light through a mechanism called fluorescence. By controlling the molecules emitting light, it has become possible to construct images with resolution much finer than the Rayleigh criterion, thus circumventing the diffraction limit. The development of super-resolved fluorescence microscopy led to the 2014 Nobel Prize in Chemistry.

In this Optical Resolution Model, two diffraction patterns for light through two circular apertures are shown side by side in **this simulation (https://openstax.org/l/21optresmodsim)** by Fu-Kwun Hwang. Watch the patterns merge as you decrease the aperture diameters.

33.6 | Polarization

Learning Objectives

By the end of this section, you will be able to:

- Explain the change in intensity as polarized light passes through a polarizing filter
- Calculate the effect of polarization by reflection and Brewster's angle
- Describe the effect of polarization by scattering
- Explain the use of polarizing materials in devices such as LCDs

Another phenomenon that arises from the wave nature of light is **polarization**. This follows directly from our understanding of light as a transverse **electromagnetic wave**.

Polarizing sunglasses are familiar to most of us. They have a special ability to cut the glare of light reflected from water or glass (**Figure 33.22**). They have this ability because of a wave characteristic of light called polarization. What is polarization? How is it produced? What are some of its uses? The answers to these questions are related to the wave character of light.

Figure 33.22 These two photographs of a river show the effect of a polarizing filter in reducing glare in light reflected from the surface of water. Part (b) of this figure was taken with a polarizing filter and part (a) was not. As a result, the reflection of clouds and sky observed in part (a) is not observed in part (b). Polarizing sunglasses are particularly useful on snow and water. (credit a and credit b: modifications of work by "Amithshs"/Wikimedia Commons)

Malus's Law

Light is one type of electromagnetic (EM) wave. EM waves are *transverse waves* consisting of varying electric and magnetic fields that oscillate perpendicular to the direction of propagation (**Figure 33.23**). However, in general, there are no specific directions for the oscillations of the electric and magnetic fields; they vibrate in any randomly oriented plane perpendicular to the direction of propagation. **Polarization** is the attribute that a wave's oscillations do have a definite direction relative to the direction of propagation of the wave. (This is not the same type of polarization as that discussed for the separation of charges.) Waves having such a direction are said to be **polarized**. For an EM wave, we define the **direction of polarization** to be the direction parallel to the electric field. Thus, we can think of the electric field arrows as showing the direction of polarization, as in **Figure 33.23**.

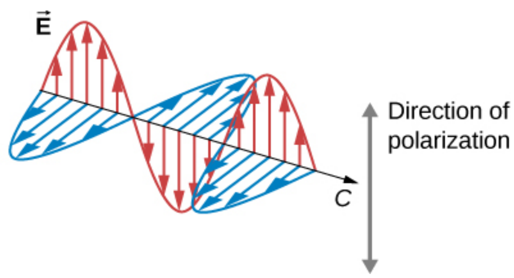


Figure 33.23 An EM wave, such as light, is a transverse wave. The electric $(\vec{\mathbf{E}})$ and magnetic $(\vec{\mathbf{B}})$ fields are perpendicular to the direction of propagation. The direction of polarization of the wave is the direction of the electric field.

To examine this further, consider the transverse waves in the ropes shown in **Figure 33.24**. The oscillations in one rope are in a vertical plane and are said to be **vertically polarized**. Those in the other rope are in a horizontal plane and are **horizontally polarized**. If a vertical slit is placed on the first rope, the waves pass through. However, a vertical slit blocks the horizontally polarized waves. For EM waves, the direction of the electric field is analogous to the disturbances on the ropes.

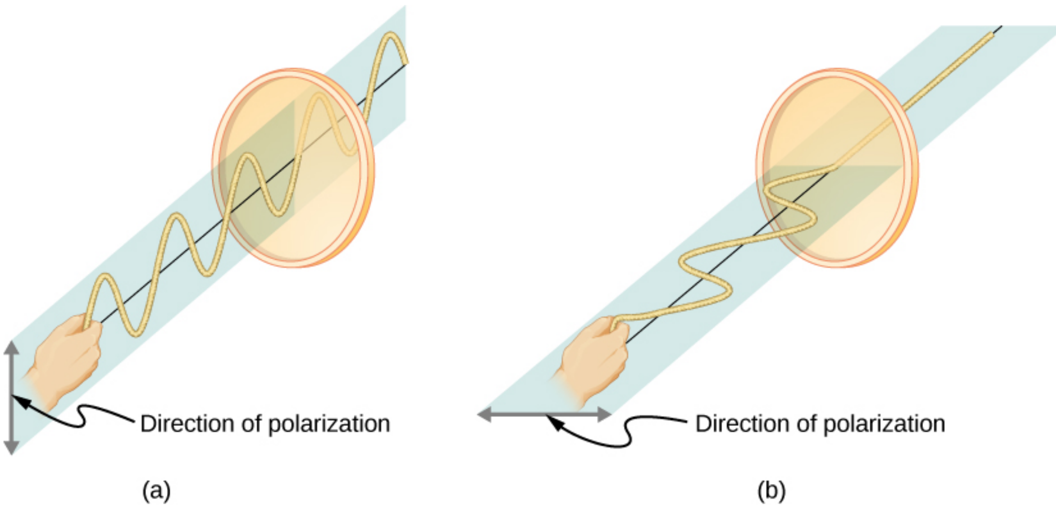


Figure 33.24 The transverse oscillations in one rope (a) are in a vertical plane, and those in the other rope (b) are in a horizontal plane. The first is said to be vertically polarized, and the other is said to be horizontally polarized. Vertical slits pass vertically polarized waves and block horizontally polarized waves.

The Sun and many other light sources produce waves that have the electric fields in random directions (**Figure 33.25**(a)). Such light is said to be **unpolarized**, because it is composed of many waves with all possible directions of polarization. Polaroid materials—which were invented by the founder of the Polaroid Corporation, Edwin Land—act as a polarizing slit for light, allowing only polarization in one direction to pass through. Polarizing filters are composed of long molecules aligned in one direction. If we think of the molecules as many slits, analogous to those for the oscillating ropes, we can understand why only light with a specific polarization can get through. The axis of a polarizing filter is the direction along which the filter passes the electric field of an EM wave.

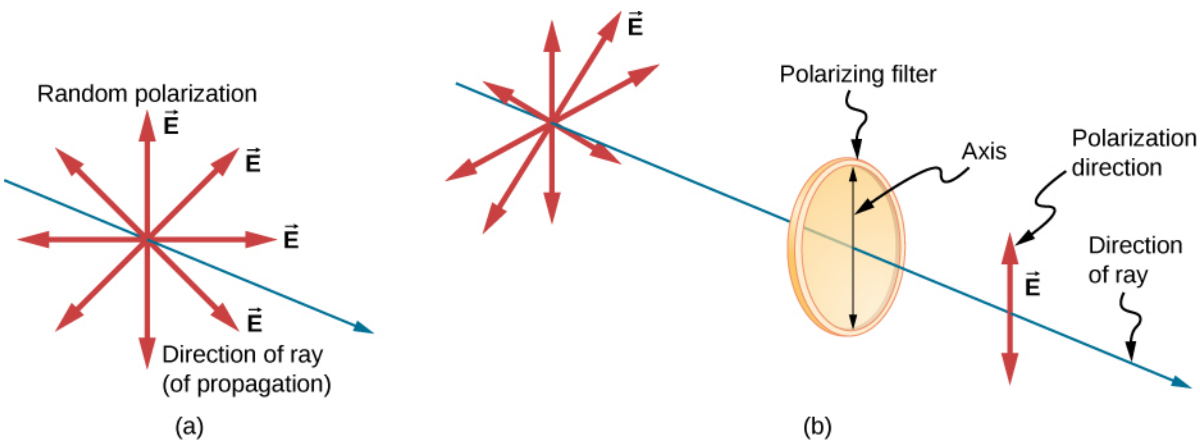


Figure 33.25 The slender arrow represents a ray of unpolarized light. The bold arrows represent the direction of polarization of the individual waves composing the ray. (a) If the light is unpolarized, the arrows point in all directions. (b) A polarizing filter has a polarization axis that acts as a slit passing through electric fields parallel to its direction. The direction of polarization of an EM wave is defined to be the direction of its electric field.

Figure 33.26 shows the effect of two polarizing filters on originally unpolarized light. The first filter polarizes the light along its axis. When the axes of the first and second filters are aligned (parallel), then all of the polarized light passed by the first filter is also passed by the second filter. If the second polarizing filter is rotated, only the component of the light parallel to the second filter's axis is passed. When the axes are perpendicular, no light is passed by the second filter.

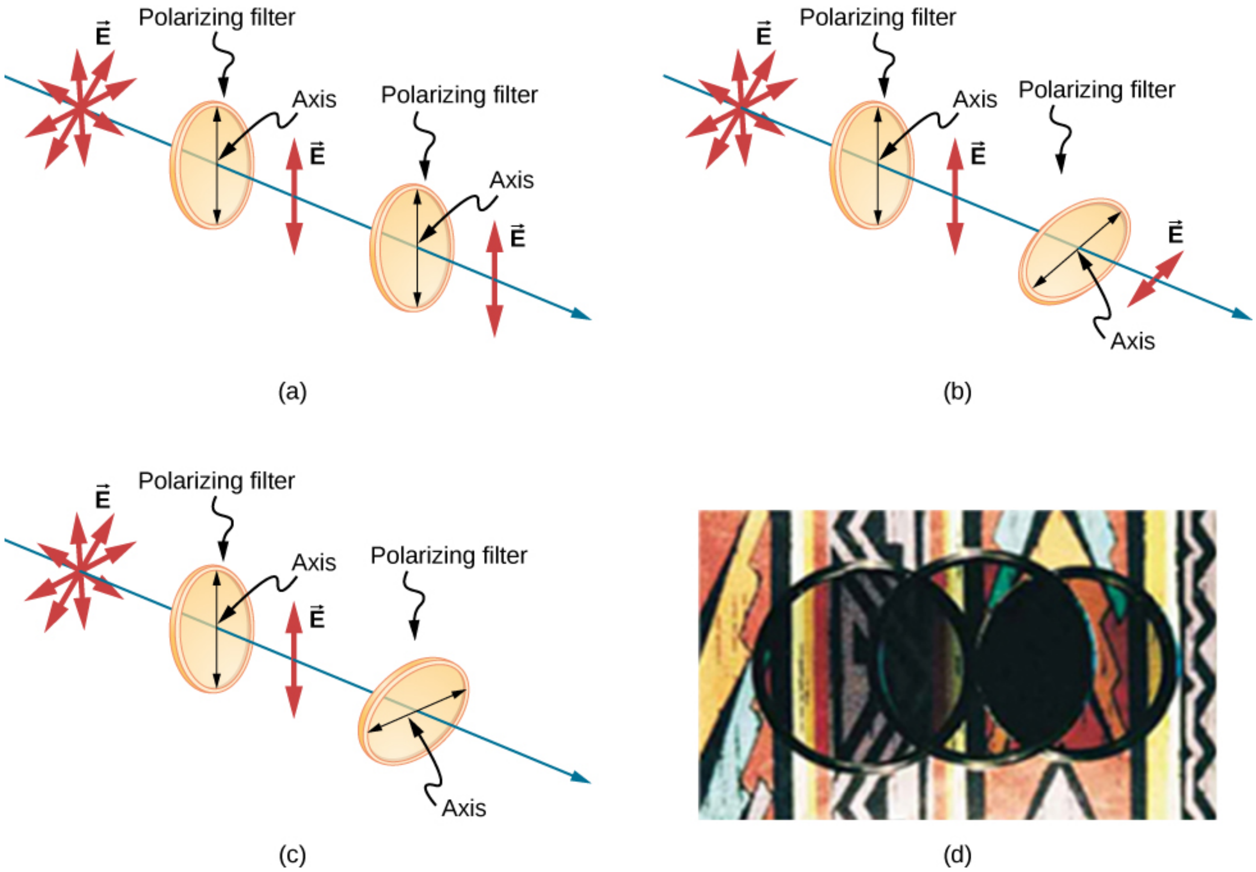


Figure 33.26 The effect of rotating two polarizing filters, where the first polarizes the light. (a) All of the polarized light is passed by the second polarizing filter, because its axis is parallel to the first. (b) As the second filter is rotated, only part of the light is passed. (c) When the second filter is perpendicular to the first, no light is passed. (d) In this photograph, a polarizing filter is placed above two others. Its axis is perpendicular to the filter on the right (dark area) and parallel to the filter on the left (lighter area). (credit d: modification of work by P.P. Urone)

Only the component of the EM wave parallel to the axis of a filter is passed. Let us call the angle between the direction of polarization and the axis of a filter θ. If the electric field has an amplitude *E*, then the transmitted part of the wave has an amplitude $E\cos\theta$ (**Figure 33.27**). Since the intensity of a wave is proportional to its amplitude squared, the intensity *I* of the transmitted wave is related to the incident wave by

$$I = I_0 \cos^2 \theta \tag{33.6}$$

where I_0 is the intensity of the polarized wave before passing through the filter. This equation is known as **Malus's law**.

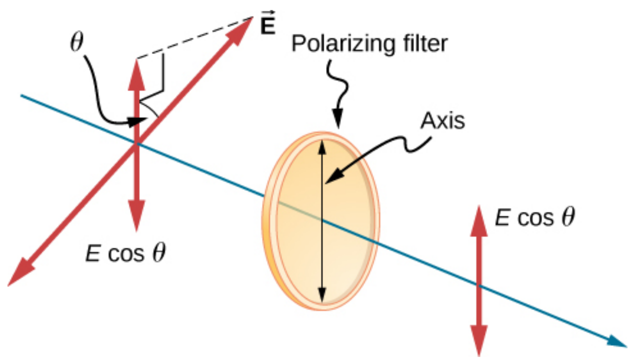


Figure 33.27 A polarizing filter transmits only the component of the wave parallel to its axis, reducing the intensity of any light not polarized parallel to its axis.

This **Open Source Physics animation (https://openstax.org/l/21phyanielefie)** helps you visualize the electric field vectors as light encounters a polarizing filter. You can rotate the filter—note that the angle displayed is in radians. You can also rotate the animation for 3D visualization.

Example 33.7

Calculating Intensity Reduction by a Polarizing Filter

What angle is needed between the direction of polarized light and the axis of a polarizing filter to reduce its intensity by 90.0% ?

Strategy

When the intensity is reduced by 90.0% , it is 10.0% or 0.100 times its original value. That is, $I = 0.100\,I_0$. Using this information, the equation $I = I_0 \cos^2 \theta$ can be used to solve for the needed angle.

Solution

Solving the equation $I = I_0 \cos^2 \theta$ for $\cos \theta$ and substituting with the relationship between I and I_0 gives

$$\cos \theta = \sqrt{\frac{I}{I_0}} = \sqrt{\frac{0.100\,I_0}{I_0}} = 0.3162.$$

Solving for θ yields

$$\theta = \cos^{-1} 0.3162 = 71.6°.$$

Significance

A fairly large angle between the direction of polarization and the filter axis is needed to reduce the intensity to 10.0% of its original value. This seems reasonable based on experimenting with polarizing films. It is interesting that at an angle of $45°$, the intensity is reduced to 50% of its original value. Note that $71.6°$ is $18.4°$ from reducing the intensity to zero, and that at an angle of $18.4°$, the intensity is reduced to 90.0% of its original value, giving evidence of symmetry.

33.6 Check Your Understanding Although we did not specify the direction in **Example 33.7**, let's say the polarizing filter was rotated clockwise by $71.6°$ to reduce the light intensity by 90.0% . What would be the intensity reduction if the polarizing filter were rotated counterclockwise by $71.6°$?

Polarization by Reflection

By now, you can probably guess that polarizing sunglasses cut the glare in reflected light, because that light is polarized.

You can check this for yourself by holding polarizing sunglasses in front of you and rotating them while looking at light reflected from water or glass. As you rotate the sunglasses, you will notice the light gets bright and dim, but not completely black. This implies the reflected light is partially polarized and cannot be completely blocked by a polarizing filter.

Figure 33.28 illustrates what happens when unpolarized light is reflected from a surface. Vertically polarized light is preferentially refracted at the surface, so the reflected light is left more horizontally polarized. The reasons for this phenomenon are beyond the scope of this text, but a convenient mnemonic for remembering this is to imagine the polarization direction to be like an arrow. Vertical polarization is like an arrow perpendicular to the surface and is more likely to stick and not be reflected. Horizontal polarization is like an arrow bouncing on its side and is more likely to be reflected. Sunglasses with vertical axes thus block more reflected light than unpolarized light from other sources.

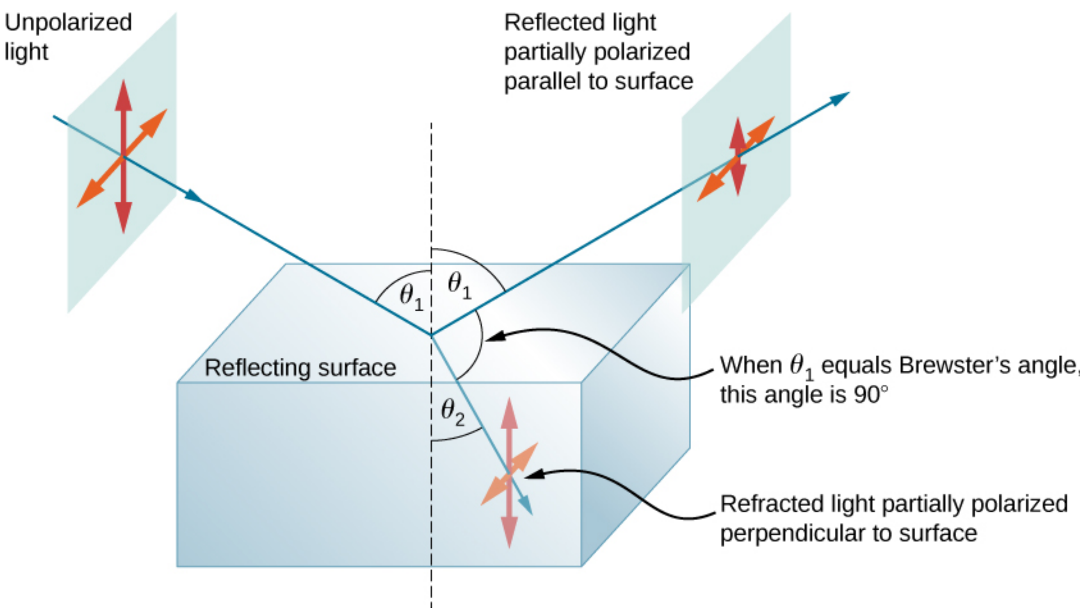


Figure 33.28 Polarization by reflection. Unpolarized light has equal amounts of vertical and horizontal polarization. After interaction with a surface, the vertical components are preferentially absorbed or refracted, leaving the reflected light more horizontally polarized. This is akin to arrows striking on their sides and bouncing off, whereas arrows striking on their tips go into the surface.

Since the part of the light that is not reflected is refracted, the amount of polarization depends on the indices of refraction of the media involved. It can be shown that reflected light is completely polarized at an angle of reflection θ_b given by

$$\tan\theta_b = \frac{n_2}{n_1} \qquad (33.7)$$

where n_1 is the medium in which the incident and reflected light travel and n_2 is the index of refraction of the medium that forms the interface that reflects the light. This equation is known as **Brewster's law** and θ_b is known as **Brewster's angle**, named after the nineteenth-century Scottish physicist who discovered them.

This **Open Source Physics animation (https://openstax.org/l/21phyaniincref)** shows incident, reflected, and refracted light as rays and EM waves. Try rotating the animation for 3D visualization and also change the angle of incidence. Near Brewster's angle, the reflected light becomes highly polarized.

Example 33.8

Calculating Polarization by Reflection

(a) At what angle will light traveling in air be completely polarized horizontally when reflected from water? (b) From glass?

Strategy

All we need to solve these problems are the indices of refraction. Air has $n_1 = 1.00,$ water has $n_2 = 1.333,$ and crown glass has $n'_2 = 1.520.$ The equation $\tan \theta_b = \frac{n_2}{n_1}$ can be directly applied to find θ_b in each case.

Solution

a. Putting the known quantities into the equation

$$\tan \theta_b = \frac{n_2}{n_1}$$

gives

$$\tan \theta_b = \frac{n_2}{n_1} = \frac{1.333}{1.00} = 1.333.$$

Solving for the angle θ_b yields

$$\theta_b = \tan^{-1} 1.333 = 53.1°.$$

b. Similarly, for crown glass and air,

$$\tan \theta'_b = \frac{n'_2}{n_1} = \frac{1.520}{1.00} = 1.52.$$

Thus,

$$\theta'_b = \tan^{-1} 1.52 = 56.7°.$$

Significance

Light reflected at these angles could be completely blocked by a good polarizing filter held with its axis vertical. Brewster's angle for water and air are similar to those for glass and air, so that sunglasses are equally effective for light reflected from either water or glass under similar circumstances. Light that is not reflected is refracted into these media. Therefore, at an incident angle equal to Brewster's angle, the refracted light is slightly polarized vertically. It is not completely polarized vertically, because only a small fraction of the incident light is reflected, so a significant amount of horizontally polarized light is refracted.

33.7 Check Your Understanding What happens at Brewster's angle if the original incident light is already 100% vertically polarized?

Atomic Explanation of Polarizing Filters

Polarizing filters have a polarization axis that acts as a slit. This slit passes EM waves (often visible light) that have an electric field parallel to the axis. This is accomplished with long molecules aligned perpendicular to the axis, as shown in **Figure 33.29**.

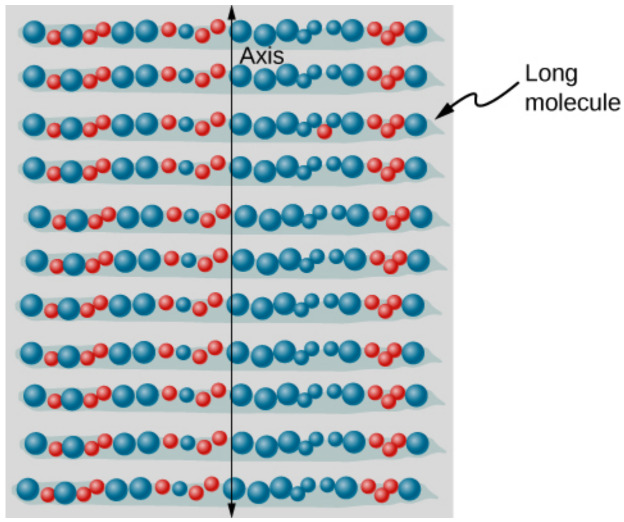


Figure 33.29 Long molecules are aligned perpendicular to the axis of a polarizing filter. In an EM wave, the component of the electric field perpendicular to these molecules passes through the filter, whereas the component parallel to the molecules is absorbed.

Figure 33.30 illustrates how the component of the electric field parallel to the long molecules is absorbed. An EM wave is composed of oscillating electric and magnetic fields. The electric field is strong compared with the magnetic field and is more effective in exerting force on charges in the molecules. The most affected charged particles are the electrons, since electron masses are small. If an electron is forced to oscillate, it can absorb energy from the EM wave. This reduces the field in the wave and, hence, reduces its intensity. In long molecules, electrons can more easily oscillate parallel to the molecule than in the perpendicular direction. The electrons are bound to the molecule and are more restricted in their movement perpendicular to the molecule. Thus, the electrons can absorb EM waves that have a component of their electric field parallel to the molecule. The electrons are much less responsive to electric fields perpendicular to the molecule and allow these fields to pass. Thus, the axis of the polarizing filter is perpendicular to the length of the molecule.

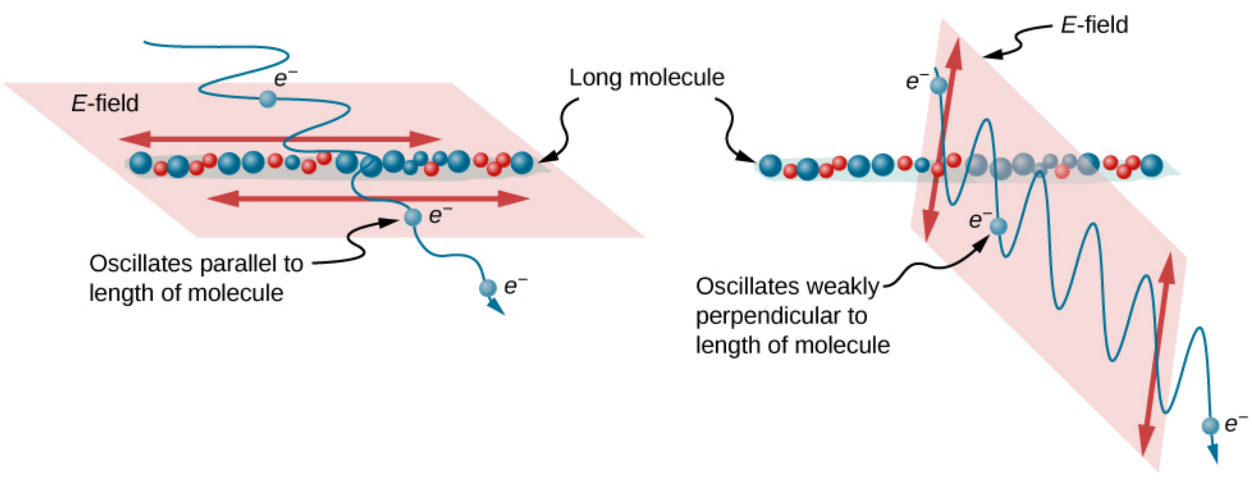


Figure 33.30 Diagram of an electron in a long molecule oscillating parallel to the molecule. The oscillation of the electron absorbs energy and reduces the intensity of the component of the EM wave that is parallel to the molecule.

Polarization by Scattering

If you hold your polarizing sunglasses in front of you and rotate them while looking at blue sky, you will see the sky get bright and dim. This is a clear indication that light scattered by air is partially polarized. **Figure 33.31** helps illustrate how this happens. Since light is a transverse EM wave, it vibrates the electrons of air molecules perpendicular to the direction that it is traveling. The electrons then radiate like small antennae. Since they are oscillating perpendicular to the direction of the light ray, they produce EM radiation that is polarized perpendicular to the direction of the ray. When viewing the light along a line perpendicular to the original ray, as in the figure, there can be no polarization in the scattered light parallel to the original ray, because that would require the original ray to be a longitudinal wave. Along other directions, a component of the other polarization can be projected along the line of sight, and the scattered light is only partially polarized. Furthermore, multiple scattering can bring light to your eyes from other directions and can contain different polarizations.

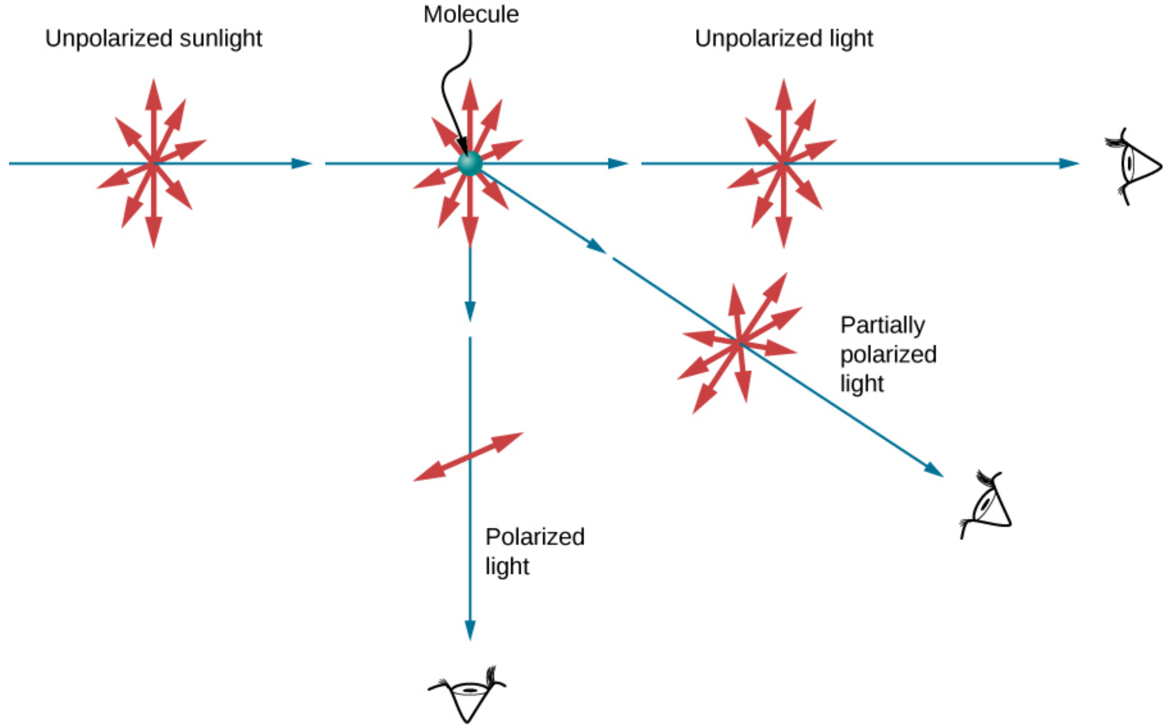


Figure 33.31 Polarization by scattering. Unpolarized light scattering from air molecules shakes their electrons perpendicular to the direction of the original ray. The scattered light therefore has a polarization perpendicular to the original direction and none parallel to the original direction.

Photographs of the sky can be darkened by polarizing filters, a trick used by many photographers to make clouds brighter by contrast. Scattering from other particles, such as smoke or dust, can also polarize light. Detecting polarization in scattered EM waves can be a useful analytical tool in determining the scattering source.

A range of optical effects are used in sunglasses. Besides being polarizing, sunglasses may have colored pigments embedded in them, whereas others use either a nonreflective or reflective coating. A recent development is photochromic lenses, which darken in the sunlight and become clear indoors. Photochromic lenses are embedded with organic microcrystalline molecules that change their properties when exposed to UV in sunlight, but become clear in artificial lighting with no UV.

Liquid Crystals and Other Polarization Effects in Materials

Although you are undoubtedly aware of liquid crystal displays (LCDs) found in watches, calculators, computer screens, cellphones, flat screen televisions, and many other places, you may not be aware that they are based on polarization. Liquid crystals are so named because their molecules can be aligned even though they are in a liquid. Liquid crystals have the property that they can rotate the polarization of light passing through them by $90°$. Furthermore, this property can be turned off by the application of a voltage, as illustrated in **Figure 33.32**. It is possible to manipulate this characteristic quickly and in small, well-defined regions to create the contrast patterns we see in so many LCD devices.

In flat screen LCD televisions, a large light is generated at the back of the TV. The light travels to the front screen through millions of tiny units called pixels (picture elements). One of these is shown in **Figure 33.32**(a) and (b). Each unit has three cells, with red, blue, or green filters, each controlled independently. When the voltage across a liquid crystal is switched off, the liquid crystal passes the light through the particular filter. We can vary the picture contrast by varying the strength of the voltage applied to the liquid crystal.

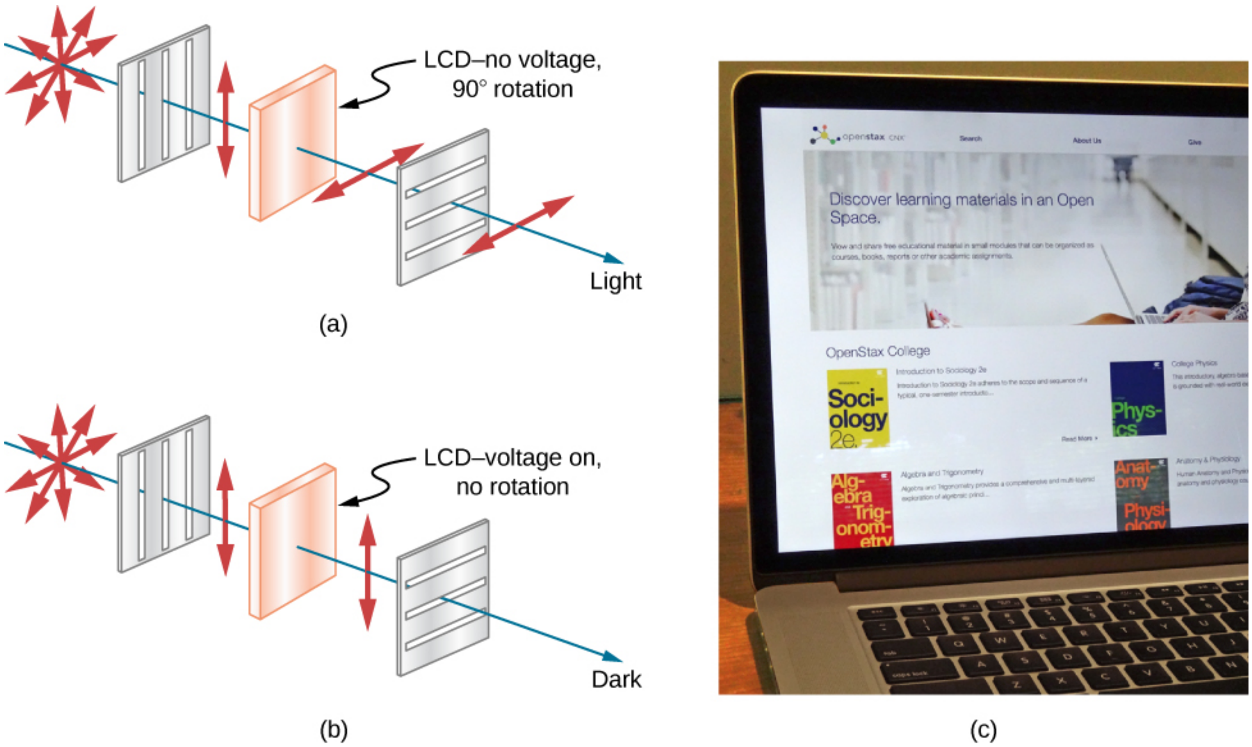


Figure 33.32 (a) Polarized light is rotated $90°$ by a liquid crystal and then passed by a polarizing filter that has its axis perpendicular to the direction of the original polarization. (b) When a voltage is applied to the liquid crystal, the polarized light is not rotated and is blocked by the filter, making the region dark in comparison with its surroundings. (c) LCDs can be made color specific, small, and fast enough to use in laptop computers and TVs. (credit c: modification of work by Jane Whitney)

Many crystals and solutions rotate the plane of polarization of light passing through them. Such substances are said to be **optically active**. Examples include sugar water, insulin, and collagen (**Figure 33.33**). In addition to depending on the type of substance, the amount and direction of rotation depend on several other factors. Among these is the concentration of the substance, the distance the light travels through it, and the wavelength of light. Optical activity is due to the asymmetrical shape of molecules in the substance, such as being helical. Measurements of the rotation of polarized light passing through substances can thus be used to measure concentrations, a standard technique for sugars. It can also give information on the shapes of molecules, such as proteins, and factors that affect their shapes, such as temperature and pH.

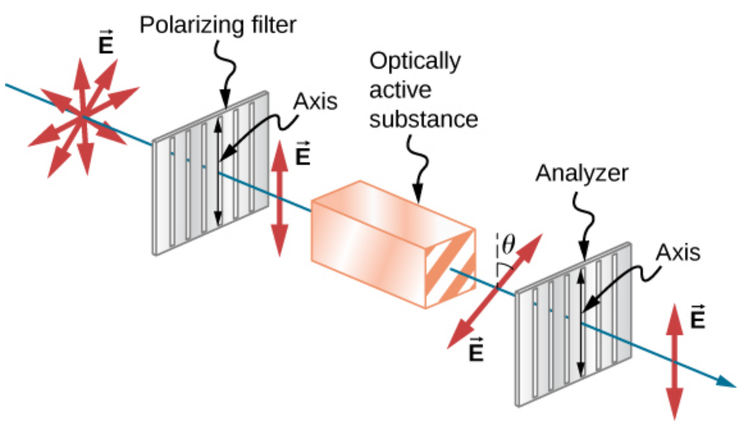


Figure 33.33 Optical activity is the ability of some substances to rotate the plane of polarization of light passing through them. The rotation is detected with a polarizing filter or analyzer.

Glass and plastic become optically active when stressed: the greater the stress, the greater the effect. Optical stress analysis on complicated shapes can be performed by making plastic models of them and observing them through crossed filters, as seen in **Figure 33.34**. It is apparent that the effect depends on wavelength as well as stress. The wavelength dependence is sometimes also used for artistic purposes.

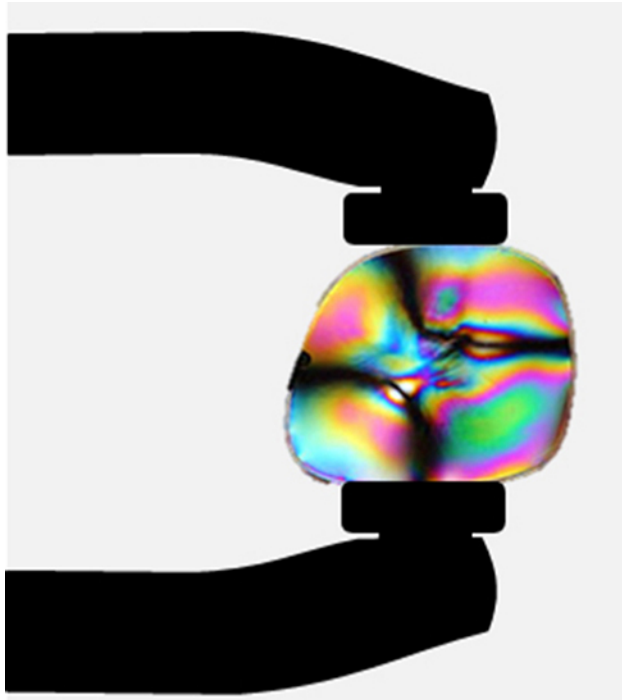


Figure 33.34 Optical stress analysis of a plastic lens placed between crossed polarizers. (credit: "Infopro"/Wikimedia Commons)

Another interesting phenomenon associated with polarized light is the ability of some crystals to split an unpolarized beam of light into two polarized beams. This occurs because the crystal has one value for the index of refraction of polarized light but a different value for the index of refraction of light polarized in the perpendicular direction, so that each component has its own angle of refraction. Such crystals are said to be **birefringent**, and, when aligned properly, two perpendicularly polarized beams will emerge from the crystal (Figure 33.35). Birefringent crystals can be used to produce polarized beams from unpolarized light. Some birefringent materials preferentially absorb one of the polarizations. These materials are called dichroic and can produce polarization by this preferential absorption. This is fundamentally how polarizing filters and other polarizers work.

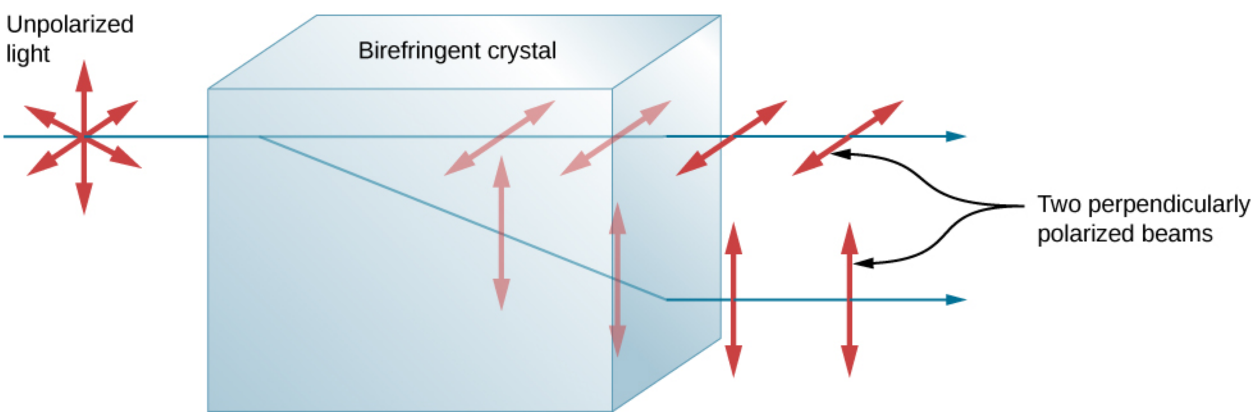


Figure 33.35 Birefringent materials, such as the common mineral calcite, split unpolarized beams of light into two with two different values of index of refraction.

CHAPTER 33 REVIEW

KEY TERMS

birefringent refers to crystals that split an unpolarized beam of light into two beams

Brewster's angle angle of incidence at which the reflected light is completely polarized

Brewster's law $\tan\theta_b = \frac{n_2}{n_1}$, where n_1 is the medium in which the incident and reflected light travel and n_2 is the index of refraction of the medium that forms the interface that reflects the light

destructive interference for a single slit occurs when the width of the slit is comparable to the wavelength of light illuminating it

diffraction bending of a wave around the edges of an opening or an obstacle

diffraction grating large number of evenly spaced parallel slits

diffraction limit fundamental limit to resolution due to diffraction

direction of polarization direction parallel to the electric field for EM waves

horizontally polarized oscillations are in a horizontal plane

Malus's law where I_0 is the intensity of the polarized wave before passing through the filter

missing order interference maximum that is not seen because it coincides with a diffraction minimum

optically active substances that rotate the plane of polarization of light passing through them

polarization attribute that wave oscillations have a definite direction relative to the direction of propagation of the wave

polarized refers to waves having the electric and magnetic field oscillations in a definite direction

Rayleigh criterion two images are just-resolvable when the center of the diffraction pattern of one is directly over the first minimum of the diffraction pattern of the other

resolution ability, or limit thereof, to distinguish small details in images

two-slit diffraction pattern diffraction pattern of two slits of width D that are separated by a distance d is the interference pattern of two point sources separated by d multiplied by the diffraction pattern of a slit of width D

unpolarized refers to waves that are randomly polarized

vertically polarized oscillations are in a vertical plane

width of the central peak angle between the minimum for $m = 1$ and $m = -1$

KEY EQUATIONS

Locations of diffraction minima	$D\sin\theta = m\lambda, \text{ for } m = \pm 1, \pm 2, \pm 3, \ldots \text{(destructive)},$
Phase difference between adjacent Huygens sources	$\phi = \left(\frac{2\pi}{\lambda}\right)D\sin\theta.$
Intensity in single-slit diffraction	$I = I_0\left(\frac{\sin\frac{\phi}{2}}{\frac{\phi}{2}}\right)^2$
Diffraction grating maxima	$d\sin\theta = m\lambda \text{ for } m = 0, \pm 1, \pm 2, \ldots.$
Rayleigh's criterion for resolution of a circular aperture	$\theta = 1.22\frac{\lambda}{D}$
Speed of light	$c = 2.99792458 \times 10^8 \text{ m/s} \approx 3.00 \times 10^8 \text{ m/s}$

Index of refraction	$n = \frac{c}{v}$
Law of reflection	$\theta_r = \theta_i$
Law of refraction (Snell's law)	$n_1 \sin\theta_1 = n_2 \sin\theta_2$
Critical angle	$\theta_c = \sin^{-1}\left(\frac{n_2}{n_1}\right)$ for $n_1 > n_2$
Malus's law	$I = I_0 \cos^2\theta$
Brewster's law	$\tan\theta_b = \frac{n_2}{n_1}$

SUMMARY

33.1 Single-Slit Diffraction

- Diffraction can send a wave around the edges of an opening or other obstacle.
- A single slit produces an interference pattern characterized by a broad central maximum with narrower and dimmer maxima to the sides.

33.2 Intensity in Single-Slit Diffraction

- The intensity pattern for diffraction due to a single slit can be calculated using phasors as

$$I = I_0\left(\frac{\sin\beta}{\beta}\right)^2,$$

where $\beta = \frac{\phi}{2} = \frac{\pi D \sin\theta}{\lambda}$, D is the slit width, λ is the wavelength, and θ is the angle from the central peak.

33.3 Double-Slit Diffraction

- With real slits with finite widths, the effects of interference and diffraction operate simultaneously to form a complicated intensity pattern.
- Relative intensities of interference fringes within a diffraction pattern can be determined.
- Missing orders occur when an interference maximum and a diffraction minimum are located together.

33.4 Diffraction Gratings

- A diffraction grating consists of a large number of evenly spaced parallel slits that produce an interference pattern similar to but sharper than that of a double slit.
- Constructive interference occurs when $d\sin\theta = m\lambda$ for $m = 0, \pm 1, \pm 2, \ldots,$ where d is the distance between the slits, θ is the angle relative to the incident direction, and m is the order of the interference.

33.5 Circular Apertures and Resolution

- Diffraction limits resolution.
- The Rayleigh criterion states that two images are just resolvable when the center of the diffraction pattern of one is directly over the first minimum of the diffraction pattern of the other.

33.6 Polarization

- Polarization is the attribute that wave oscillations have a definite direction relative to the direction of propagation of the wave. The direction of polarization is defined to be the direction parallel to the electric field of the EM wave.

- Unpolarized light is composed of many rays having random polarization directions.
- Unpolarized light can be polarized by passing it through a polarizing filter or other polarizing material. The process of polarizing light decreases its intensity by a factor of 2.
- The intensity, *I*, of polarized light after passing through a polarizing filter is $I = I_0 \cos^2 \theta$, where I_0 is the incident intensity and θ is the angle between the direction of polarization and the axis of the filter.
- Polarization is also produced by reflection.
- Brewster's law states that reflected light is completely polarized at the angle of reflection θ_b, known as Brewster's angle.
- Polarization can also be produced by scattering.
- Several types of optically active substances rotate the direction of polarization of light passing through them.

CONCEPTUAL QUESTIONS

33.1 Single-Slit Diffraction

1. As the width of the slit producing a single-slit diffraction pattern is reduced, how will the diffraction pattern produced change?

2. Compare interference and diffraction.

3. If you and a friend are on opposite sides of a hill, you can communicate with walkie-talkies but not with flashlights. Explain.

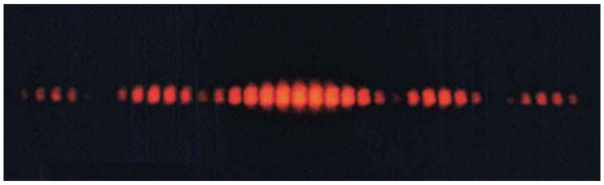
4. What happens to the diffraction pattern of a single slit when the entire optical apparatus is immersed in water?

5. In our study of diffraction by a single slit, we assume that the length of the slit is much larger than the width. What happens to the diffraction pattern if these two dimensions were comparable?

6. A rectangular slit is twice as wide as it is high. Is the central diffraction peak wider in the vertical direction or in the horizontal direction?

33.3 Double-Slit Diffraction

7. Shown below is the central part of the interference pattern for a pure wavelength of red light projected onto a double slit. The pattern is actually a combination of single- and double-slit interference. Note that the bright spots are evenly spaced. Is this a double- or single-slit characteristic? Note that some of the bright spots are dim on either side of the center. Is this a single- or double-slit characteristic? Which is smaller, the slit width or the separation between slits? Explain your responses.



(credit: PASCO)

33.5 Circular Apertures and Resolution

8. Is higher resolution obtained in a microscope with red or blue light? Explain your answer.

9. The resolving power of refracting telescope increases with the size of its objective lens. What other advantage is gained with a larger lens?

10. The distance between atoms in a molecule is about 10^{-8} cm . Can visible light be used to "see" molecules?

11. A beam of light always spreads out. Why can a beam not be created with parallel rays to prevent spreading? Why can lenses, mirrors, or apertures not be used to correct the spreading?

33.6 Polarization

12. Can a sound wave in air be polarized? Explain.

13. No light passes through two perfect polarizing filters with perpendicular axes. However, if a third polarizing filter is placed between the original two, some light can pass. Why is this? Under what circumstances does most of the light pass?

14. Explain what happens to the energy carried by light that it is dimmed by passing it through two crossed polarizing filters.

15. When particles scattering light are much smaller than its wavelength, the amount of scattering is proportional to $\frac{1}{\lambda}$. Does this mean there is more scattering for small λ than large λ? How does this relate to the fact that the sky is blue?

16. Using the information given in the preceding question, explain why sunsets are red.

17. When light is reflected at Brewster's angle from a smooth surface, it is 100% polarized parallel to the surface. Part of the light will be refracted into the surface. Describe how you would do an experiment to determine the polarization of the refracted light. What direction would you expect the polarization to have and would you expect it to be 100%?

18. If you lie on a beach looking at the water with your head tipped slightly sideways, your polarized sunglasses do not work very well. Why not?

PROBLEMS

33.1 Single-Slit Diffraction

19. (a) At what angle is the first minimum for 550-nm light falling on a single slit of width $1.00\mu\text{m}$? (b) Will there be a second minimum?

20. (a) Calculate the angle at which a $2.00\text{-}\mu\text{m}$-wide slit produces its first minimum for 410-nm violet light. (b) Where is the first minimum for 700-nm red light?

21. (a) How wide is a single slit that produces its first minimum for 633-nm light at an angle of $28.0°$? (b) At what angle will the second minimum be?

22. (a) What is the width of a single slit that produces its first minimum at $60.0°$ for 600-nm light? (b) Find the wavelength of light that has its first minimum at $62.0°$.

23. Find the wavelength of light that has its third minimum at an angle of $48.6°$ when it falls on a single slit of width $3.00\mu\text{m}$.

24. (a) Sodium vapor light averaging 589 nm in wavelength falls on a single slit of width $7.50\mu\text{m}$. At what angle does it produces its second minimum? (b) What is the highest-order minimum produced?

25. Consider a single-slit diffraction pattern for $\lambda = 589\text{ nm}$, projected on a screen that is 1.00 m from a slit of width 0.25 mm. How far from the center of the pattern are the centers of the first and second dark fringes?

26. (a) Find the angle between the first minima for the two sodium vapor lines, which have wavelengths of 589.1 and 589.6 nm, when they fall upon a single slit of width $2.00\mu\text{m}$. (b) What is the distance between these minima if the diffraction pattern falls on a screen 1.00 m from the slit? (c) Discuss the ease or difficulty of measuring such a distance.

27. (a) What is the minimum width of a single slit (in multiples of λ) that will produce a first minimum for a wavelength λ? (b) What is its minimum width if it produces 50 minima? (c) 1000 minima?

28. (a) If a single slit produces a first minimum at $14.5°$, at what angle is the second-order minimum? (b) What is the angle of the third-order minimum? (c) Is there a fourth-order minimum? (d) Use your answers to illustrate how the angular width of the central maximum is about twice the angular width of the next maximum (which is the angle between the first and second minima).

29. If the separation between the first and the second minima of a single-slit diffraction pattern is 6.0 mm, what is the distance between the screen and the slit? The light wavelength is 500 nm and the slit width is 0.16 mm.

30. A water break at the entrance to a harbor consists of a rock barrier with a 50.0-m-wide opening. Ocean waves of 20.0-m wavelength approach the opening straight on. At what angles to the incident direction are the boats inside the harbor most protected against wave action?

31. An aircraft maintenance technician walks past a tall hangar door that acts like a single slit for sound entering the hangar. Outside the door, on a line perpendicular to the opening in the door, a jet engine makes a 600-Hz sound. At what angle with the door will the technician observe the first minimum in sound intensity if the vertical opening is 0.800 m wide and the speed of sound is 340 m/s?

33.2 Intensity in Single-Slit Diffraction

32. A single slit of width $3.0\,\mu\text{m}$ is illuminated by a sodium yellow light of wavelength 589 nm. Find the intensity at a $15°$ angle to the axis in terms of the intensity of the central maximum.

33. A single slit of width 0.1 mm is illuminated by a mercury light of wavelength 576 nm. Find the intensity at a $10°$ angle to the axis in terms of the intensity of the central maximum.

34. The width of the central peak in a single-slit diffraction pattern is 5.0 mm. The wavelength of the light is 600 nm, and the screen is 2.0 m from the slit. (a) What is the width of the slit? (b) Determine the ratio of the intensity at 4.5 mm from the center of the pattern to the intensity at the center.

35. Consider the single-slit diffraction pattern for $\lambda = 600\text{ nm}$, $D = 0.025\text{ mm}$, and $x = 2.0\text{ m}$. Find the intensity in terms of I_o at $\theta = 0.5°$, $1.0°$, $1.5°$, $3.0°$, and $10.0°$.

33.3 Double-Slit Diffraction

36. Two slits of width $2\,\mu\text{m}$, each in an opaque material, are separated by a center-to-center distance of $6\,\mu\text{m}$. A monochromatic light of wavelength 450 nm is incident on the double-slit. One finds a combined interference and diffraction pattern on the screen.

(a) How many peaks of the interference will be observed in the central maximum of the diffraction pattern?

(b) How many peaks of the interference will be observed if the slit width is doubled while keeping the distance between the slits same?

(c) How many peaks of interference will be observed if the slits are separated by twice the distance, that is, $12\,\mu\text{m}$, while keeping the widths of the slits same?

(d) What will happen in (a) if instead of 450-nm light another light of wavelength 680 nm is used?

(e) What is the value of the ratio of the intensity of the central peak to the intensity of the next bright peak in (a)?

(f) Does this ratio depend on the wavelength of the light?

(g) Does this ratio depend on the width or separation of the slits?

37. A double slit produces a diffraction pattern that is a combination of single- and double-slit interference. Find the ratio of the width of the slits to the separation between them, if the first minimum of the single-slit pattern falls on the fifth maximum of the double-slit pattern. (This will greatly reduce the intensity of the fifth maximum.)

38. For a double-slit configuration where the slit separation is four times the slit width, how many interference fringes lie in the central peak of the diffraction pattern?

39. Light of wavelength 500 nm falls normally on 50 slits that are $2.5 \times 10^{-3}\text{ mm}$ wide and spaced $5.0 \times 10^{-3}\text{ mm}$ apart. How many interference fringes lie in the central peak of the diffraction pattern?

40. A monochromatic light of wavelength 589 nm incident on a double slit with slit width $2.5\,\mu\text{m}$ and unknown separation results in a diffraction pattern containing nine interference peaks inside the central maximum. Find the separation of the slits.

41. When a monochromatic light of wavelength 430 nm incident on a double slit of slit separation $5\,\mu\text{m}$, there are 11 interference fringes in its central maximum. How many interference fringes will be in the central maximum of a light of wavelength 632.8 nm for the same double slit?

42. Determine the intensities of two interference peaks other than the central peak in the central maximum of the diffraction, if possible, when a light of wavelength 628 nm is incident on a double slit of width 500 nm and separation 1500 nm. Use the intensity of the central spot to be 1 mW/cm^2.

33.4 Diffraction Gratings

43. A diffraction grating has 2000 lines per centimeter. At what angle will the first-order maximum be for 520-nm-wavelength green light?

44. Find the angle for the third-order maximum for 580-nm-wavelength yellow light falling on a difraction grating having 1500 lines per centimeter.

45. How many lines per centimeter are there on a diffraction grating that gives a first-order maximum for 470-nm blue light at an angle of $25.0°$?

46. What is the distance between lines on a diffraction grating that produces a second-order maximum for 760-nm red light at an angle of $60.0°$?

47. Calculate the wavelength of light that has its second-order maximum at $45.0°$ when falling on a diffraction grating that has 5000 lines per centimeter.

48. An electric current through hydrogen gas produces several distinct wavelengths of visible light. What are the wavelengths of the hydrogen spectrum, if they form first-order maxima at angles $24.2°$, $25.7°$, $29.1°$, and $41.0°$ when projected on a diffraction grating having 10,000 lines per centimeter?

49. (a) What do the four angles in the preceding problem become if a 5000-line per centimeter diffraction grating is used? (b) Using this grating, what would the angles be for the second-order maxima? (c) Discuss the relationship between integral reductions in lines per centimeter and the new angles of various order maxima.

50. What is the spacing between structures in a feather that acts as a reflection grating, giving that they produce a first-order maximum for 525-nm light at a 30.0° angle?

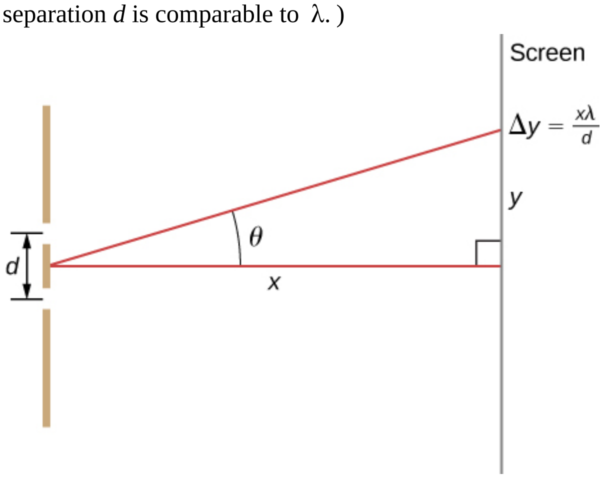
51. An opal such as that shown in **Figure 33.13** acts like a reflection grating with rows separated by about $8\ \mu\text{m}$. If the opal is illuminated normally, (a) at what angle will red light be seen and (b) at what angle will blue light be seen?

52. At what angle does a diffraction grating produce a second-order maximum for light having a first-order maximum at 20.0° ?

53. (a) Find the maximum number of lines per centimeter a diffraction grating can have and produce a maximum for the smallest wavelength of visible light. (b) Would such a grating be useful for ultraviolet spectra? (c) For infrared spectra?

54. (a) Show that a 30,000 line per centimeter grating will not produce a maximum for visible light. (b) What is the longest wavelength for which it does produce a first-order maximum? (c) What is the greatest number of line per centimeter a diffraction grating can have and produce a complete second-order spectrum for visible light?

55. The analysis shown below also applies to diffraction gratings with lines separated by a distance *d*. What is the distance between fringes produced by a diffraction grating having 125 lines per centimeter for 600-nm light, if the screen is 1.50 m away? (*Hint:* The distance between adjacent fringes is $\Delta y = x\lambda/d$, assuming the slit separation *d* is comparable to λ.)



33.5 Circular Apertures and Resolution

56. The 305-m-diameter Arecibo radio telescope pictured in **Figure 33.18** detects radio waves with a 4.00-cm average wavelength. (a) What is the angle between two just-resolvable point sources for this telescope? (b) How close together could these point sources be at the 2 million light-year distance of the Andromeda Galaxy?

57. Assuming the angular resolution found for the Hubble Telescope in **Example 33.6**, what is the smallest detail that could be observed on the moon?

58. Diffraction spreading for a flashlight is insignificant compared with other limitations in its optics, such as spherical aberrations in its mirror. To show this, calculate the minimum angular spreading of a flashlight beam that is originally 5.00 cm in diameter with an average wavelength of 600 nm.

59. (a) What is the minimum angular spread of a 633-nm wavelength He-Ne laser beam that is originally 1.00 mm in diameter? (b) If this laser is aimed at a mountain cliff 15.0 km away, how big will the illuminated spot be? (c) How big a spot would be illuminated on the moon, neglecting atmospheric effects? (This might be done to hit a corner reflector to measure the round-trip time and, hence, distance.)

60. A telescope can be used to enlarge the diameter of a laser beam and limit diffraction spreading. The laser beam is sent through the telescope in opposite the normal direction and can then be projected onto a satellite or the moon. (a) If this is done with the Mount Wilson telescope, producing a 2.54-m-diameter beam of 633-nm light, what is the minimum angular spread of the beam? (b) Neglecting atmospheric effects, what is the size of the spot this beam would make on the moon, assuming a lunar distance of 3.84×10^8 m ?

61. The limit to the eye's acuity is actually related to diffraction by the pupil. (a) What is the angle between two just-resolvable points of light for a 3.00-mm-diameter pupil, assuming an average wavelength of 550 nm? (b) Take your result to be the practical limit for the eye. What is the greatest possible distance a car can be from you if you can resolve its two headlights, given they are 1.30 m apart? (c) What is the distance between two just-resolvable points held at an arm's length (0.800 m) from your eye? (d) How does your answer to (c) compare to details you normally observe in everyday circumstances?

62. What is the minimum diameter mirror on a telescope that would allow you to see details as small as 5.00 km on the moon some 384,000 km away? Assume an average wavelength of 550 nm for the light received.

63. Find the radius of a star's image on the retina of an eye if its pupil is open to 0.65 cm and the distance from the pupil to the retina is 2.8 cm. Assume $\lambda = 550\text{ nm}$.

64. (a) The dwarf planet Pluto and its moon, Charon, are separated by 19,600 km. Neglecting atmospheric effects, should the 5.08-m-diameter Palomar Mountain telescope be able to resolve these bodies when they are $4.50 \times 10^9\text{ km}$ from Earth? Assume an average wavelength of 550 nm. (b) In actuality, it is just barely possible to discern that Pluto and Charon are separate bodies using a ground-based telescope. What are the reasons for this?

65. A spy satellite orbits Earth at a height of 180 km. What is the minimum diameter of the objective lens in a telescope that must be used to resolve columns of troops marching 2.0 m apart? Assume $\lambda = 550\text{ nm}$.

66. What is the minimum angular separation of two stars that are just-resolvable by the 8.1-m Gemini South telescope, if atmospheric effects do not limit resolution? Use 550 nm for the wavelength of the light from the stars.

67. The headlights of a car are 1.3 m apart. What is the maximum distance at which the eye can resolve these two headlights? Take the pupil diameter to be 0.40 cm.

68. When dots are placed on a page from a laser printer, they must be close enough so that you do not see the individual dots of ink. To do this, the separation of the dots must be less than Raleigh's criterion. Take the pupil of the eye to be 3.0 mm and the distance from the paper to the eye of 35 cm; find the minimum separation of two dots such that they cannot be resolved. How many dots per inch (dpi) does this correspond to?

69. Suppose you are looking down at a highway from a jetliner flying at an altitude of 6.0 km. How far apart must two cars be if you are able to distinguish them? Assume that $\lambda = 550\text{ nm}$ and that the diameter of your pupils is 4.0 mm.

70. Can an astronaut orbiting Earth in a satellite at a distance of 180 km from the surface distinguish two skyscrapers that are 20 m apart? Assume that the pupils of the astronaut's eyes have a diameter of 5.0 mm and that most of the light is centered around 500 nm.

71. The characters of a stadium scoreboard are formed with closely spaced lightbulbs that radiate primarily yellow light. (Use $\lambda = 600\text{ nm}$.) How closely must the bulbs be spaced so that an observer 80 m away sees a display of continuous lines rather than the individual bulbs? Assume that the pupil of the observer's eye has a diameter of 5.0 mm.

72. If a microscope can accept light from objects at angles as large as $\alpha = 70°$, what is the smallest structure that can be resolved when illuminated with light of wavelength 500 nm and (a) the specimen is in air? (b) When the specimen is immersed in oil, with index of refraction of 1.52?

73. A camera uses a lens with aperture 2.0 cm. What is the angular resolution of a photograph taken at 700 nm wavelength? Can it resolve the millimeter markings of a ruler placed 35 m away?

33.6 Polarization

74. What angle is needed between the direction of polarized light and the axis of a polarizing filter to cut its intensity in half?

75. The angle between the axes of two polarizing filters is $45.0°$. By how much does the second filter reduce the intensity of the light coming through the first?

76. Two polarizing sheets P_1 and P_2 are placed together with their transmission axes oriented at an angle θ to each other. What is θ when only 25% of the maximum transmitted light intensity passes through them?

77. Suppose that in the preceding problem the light incident on P_1 is unpolarized. At the determined value of θ, what fraction of the incident light passes through the combination?

78. If you have completely polarized light of intensity 150 W/m^2, what will its intensity be after passing through a polarizing filter with its axis at an $89.0°$ angle to the light's polarization direction?

79. What angle would the axis of a polarizing filter need to make with the direction of polarized light of intensity 1.00 kW/m^2 to reduce the intensity to 10.0 W/m^2?

80. At the end of **Example 33.7**, it was stated that the intensity of polarized light is reduced to 90.0% of its original value by passing through a polarizing filter with its axis at an angle of $18.4°$ to the direction of polarization. Verify this statement.

81. Show that if you have three polarizing filters, with the second at an angle of $45.0°$ to the first and the third at an angle of $90.0°$ to the first, the intensity of light passed by the first will be reduced to 25.0% of its value. (This is in contrast to having only the first and third, which reduces the intensity to zero, so that placing the second between them increases the intensity of the transmitted light.)

82. Three polarizing sheets are placed together such that the transmission axis of the second sheet is oriented at $25.0°$ to the axis of the first, whereas the transmission axis of the third sheet is oriented at $40.0°$ (in the same sense) to the axis of the first. What fraction of the intensity of an incident unpolarized beam is transmitted by the combination?

83. In order to rotate the polarization axis of a beam of linearly polarized light by $90.0°$, a student places sheets P_1 and P_2 with their transmission axes at $45.0°$ and $90.0°$, respectively, to the beam's axis of polarization. (a) What fraction of the incident light passes through P_1 and (b) through the combination? (c) Repeat your calculations for part (b) for transmission-axis angles of $30.0°$ and $90.0°$, respectively.

84. It is found that when light traveling in water falls on a plastic block, Brewster's angle is $50.0°$. What is the refractive index of the plastic?

85. At what angle will light reflected from diamond be completely polarized?

86. What is Brewster's angle for light traveling in water that is reflected from crown glass?

87. A scuba diver sees light reflected from the water's surface. At what angle relative to the water's surface will this light be completely polarized?

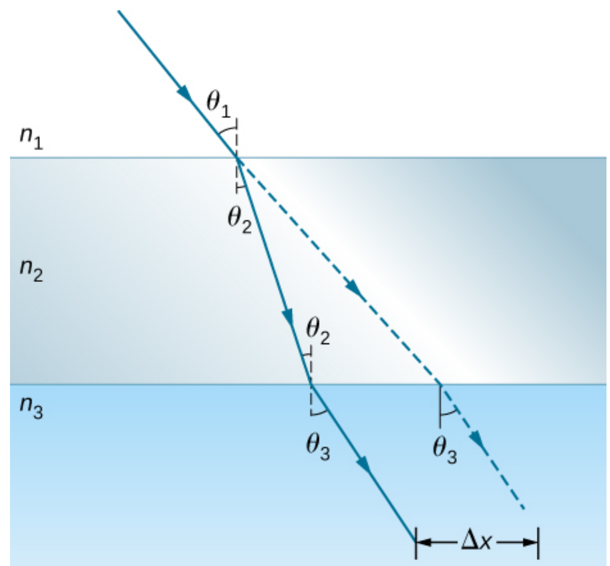
ADDITIONAL PROBLEMS

88. From his measurements, Roemer estimated that it took 22 min for light to travel a distance equal to the diameter of Earth's orbit around the Sun. (a) Use this estimate along with the known diameter of Earth's orbit to obtain a rough value of the speed of light. (b) Light actually takes 16.5 min to travel this distance. Use this time to calculate the speed of light.

89. Cornu performed Fizeau's measurement of the speed of light using a wheel of diameter 4.00 cm that contained 180 teeth. The distance from the wheel to the mirror was 22.9 km. Assuming he measured the speed of light accurately, what was the angular velocity of the wheel?

90. Suppose you have an unknown clear substance immersed in water, and you wish to identify it by finding its index of refraction. You arrange to have a beam of light enter it at an angle of $45.0°$, and you observe the angle of refraction to be $40.3°$. What is the index of refraction of the substance and its likely identity?

91. Shown below is a ray of light going from air through crown glass into water, such as going into a fish tank. Calculate the amount the ray is displaced by the glass (Δx), given that the incident angle is $40.0°$ and the glass is 1.00 cm thick.



92. Considering the previous problem, show that θ_3 is the same as it would be if the second medium were not present.

93. At what angle is light inside crown glass completely polarized when reflected from water, as in a fish tank?

94. Light reflected at $55.6°$ from a window is completely polarized. What is the window's index of refraction and the likely substance of which it is made?

95. (a) Light reflected at $62.5°$ from a gemstone in a ring is completely polarized. Can the gem be a diamond? (b) At what angle would the light be completely polarized if the gem was in water?

96. If θ_b is Brewster's angle for light reflected from the top of an interface between two substances, and θ'_b is Brewster's angle for light reflected from below, prove that $\theta_b + \theta'_b = 90.0°$.

97. Unreasonable results Suppose light travels from water to another substance, with an angle of incidence of $10.0°$ and an angle of refraction of $14.9°$. (a) What is the index of refraction of the other substance? (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

98. Unreasonable results Light traveling from water to a gemstone strikes the surface at an angle of $80.0°$ and has an angle of refraction of $15.2°$. (a) What is the speed of light in the gemstone? (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

99. If a polarizing filter reduces the intensity of polarized light to 50.0% of its original value, by how much are the electric and magnetic fields reduced?

100. Suppose you put on two pairs of polarizing sunglasses with their axes at an angle of $15.0°$. How much longer will it take the light to deposit a given amount of energy in your eye compared with a single pair of sunglasses? Assume the lenses are clear except for their polarizing characteristics.

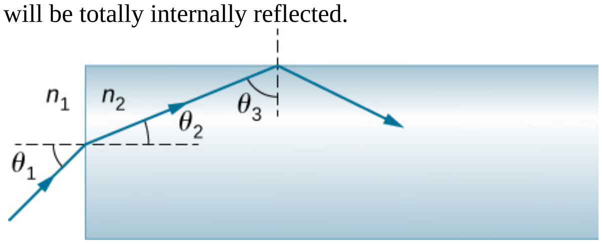
101. (a) On a day when the intensity of sunlight is $1.00\ \text{kW/m}^2$, a circular lens 0.200 m in diameter focuses light onto water in a black beaker. Two polarizing sheets of plastic are placed in front of the lens with their axes at an angle of $20.0°$. Assuming the sunlight is unpolarized and the polarizers are 100% efficient, what is the initial rate of heating of the water in $°\text{C/s}$, assuming it is 80.0% absorbed? The aluminum beaker has a mass of 30.0 grams and contains 250 grams of water. (b) Do the polarizing filters get hot? Explain.

CHALLENGE PROBLEMS

102. Light shows staged with lasers use moving mirrors to swing beams and create colorful effects. Show that a light ray reflected from a mirror changes direction by 2θ when the mirror is rotated by an angle θ.

103. Consider sunlight entering Earth's atmosphere at sunrise and sunset—that is, at a $90.0°$ incident angle. Taking the boundary between nearly empty space and the atmosphere to be sudden, calculate the angle of refraction for sunlight. This lengthens the time the Sun appears to be above the horizon, both at sunrise and sunset. Now construct a problem in which you determine the angle of refraction for different models of the atmosphere, such as various layers of varying density. Your instructor may wish to guide you on the level of complexity to consider and on how the index of refraction varies with air density.

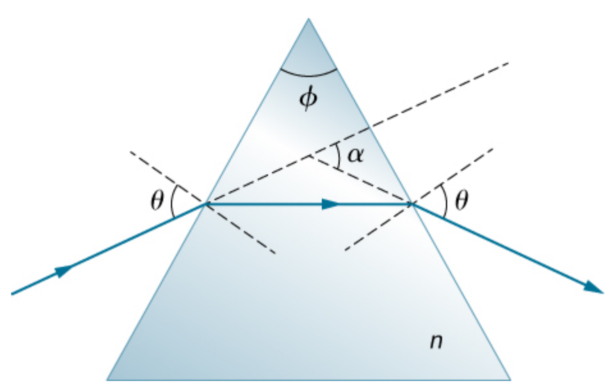
104. A light ray entering an optical fiber surrounded by air is first refracted and then reflected as shown below. Show that if the fiber is made from crown glass, any incident ray will be totally internally reflected.



105. A light ray falls on the left face of a prism (see below) at the angle of incidence θ for which the emerging beam has an angle of refraction θ at the right face. Show that the index of refraction n of the glass prism is given by

$$n = \frac{\sin\frac{1}{2}(\alpha + \phi)}{\sin\frac{1}{2}\phi}$$

where ϕ is the vertex angle of the prism and α is the angle through which the beam has been deviated. If $\alpha = 37.0°$ and the base angles of the prism are each $50.0°$, what is n?



106. If the apex angle ϕ in the previous problem is $20.0°$ and $n = 1.50$, what is the value of α?

107. The light incident on polarizing sheet P_1 is linearly polarized at an angle of $30.0°$ with respect to the transmission axis of P_1. Sheet P_2 is placed so that its axis is parallel to the polarization axis of the incident light, that is, also at $30.0°$ with respect to P_1. (a) What fraction of the incident light passes through P_1? (b) What fraction of the incident light is passed by the combination? (c) By rotating P_2, a maximum in transmitted intensity is obtained. What is the ratio of this maximum intensity to the intensity of transmitted light when P_2 is at $30.0°$ with respect to P_1?

108. Prove that if I is the intensity of light transmitted by two polarizing filters with axes at an angle θ and I' is the intensity when the axes are at an angle $90.0° - \theta$, then $I + I' = I_0$, the original intensity. (*Hint:* Use the trigonometric identities $\cos 90.0° - \theta = \sin \theta$ and $\cos^2 \theta + \sin^2 \theta = 1$.)

34 | ASTRONOMICAL INSTRUMENTS

Figure 34.1 This artist's impression shows the Hubble above Earth, with the rectangular solar panels that provide it with power seen to the left and right.

Chapter Outline

Introduction

Thus far in our studies, we have used a lot of information (e.g. images and spectra) that astronomers obtain from the solar system, the stars, and the galaxies. That information was then used to build up our current models in astrophysics. But up until now, we have not talked about the science that goes into actually obtaining precise, useful information (images or spectra) from the light reaching Earth (or reaching an instrument like the Hubble Space telescope).

In this final unit, we have studied many of the basic principles of optics that must be understood to construct modern-day astronomical instruments and telescopes. In this, the final chapter of the book, we will look at how those principles come together in some current (and soon-to-be-constructed) instruments. Along the way, we will introduce a few more important ideas, like **light-collecting area** and **interferometry**).

If you look at the sky when you are far away from city lights, there seem to be an overwhelming number of stars up there. In reality, only about 9000 stars are visible to the unaided eye (from both hemispheres of our planet). The light from most stars is so weak that by the time it reaches Earth, it cannot be detected by the human eye. How can we learn about the vast majority of objects in the universe that our unaided eyes simply cannot see?

In this chapter, we describe the tools astronomers use to extend their vision into space. We have learned almost everything we know about the universe from studying electromagnetic radiation, as discussed in the chapter on **Spectroscopy**. In the twentieth century, our exploration of space made it possible to detect electromagnetic radiation at all wavelengths, from gamma rays to radio waves. The different wavelengths carry different kinds of information, and the appearance of any given object often depends on the wavelength at which the observations are made.

34.1 | Telescopes

Learning Objectives

After completing this section, you will be able to:

- Describe the three basic components of a modern system for measuring astronomical sources
- Describe the main functions of a telescope
- Describe the two basic types of visible-light telescopes and how they form images

Systems for Measuring Radiation

There are three basic components of a modern system for measuring radiation from astronomical sources. First, there is a **telescope**, which serves as a "bucket" for collecting visible light (or radiation at other wavelengths, as shown in (**Figure 34.2**). Just as you can catch more rain with a garbage can than with a coffee cup, large telescopes gather much more light than your eye can. Second, there is an instrument attached to the telescope that sorts the incoming radiation by wavelength. Sometimes the sorting is fairly crude. For example, we might simply want to separate blue light from red light so that we can determine the temperature of a star. But at other times, we want to see individual spectral lines to determine what an object is made of, or to measure its speed (as explained in the **Spectroscopy** chapter). Third, we need some type of **detector**, a device that senses the radiation in the wavelength regions we have chosen and permanently records the observations.

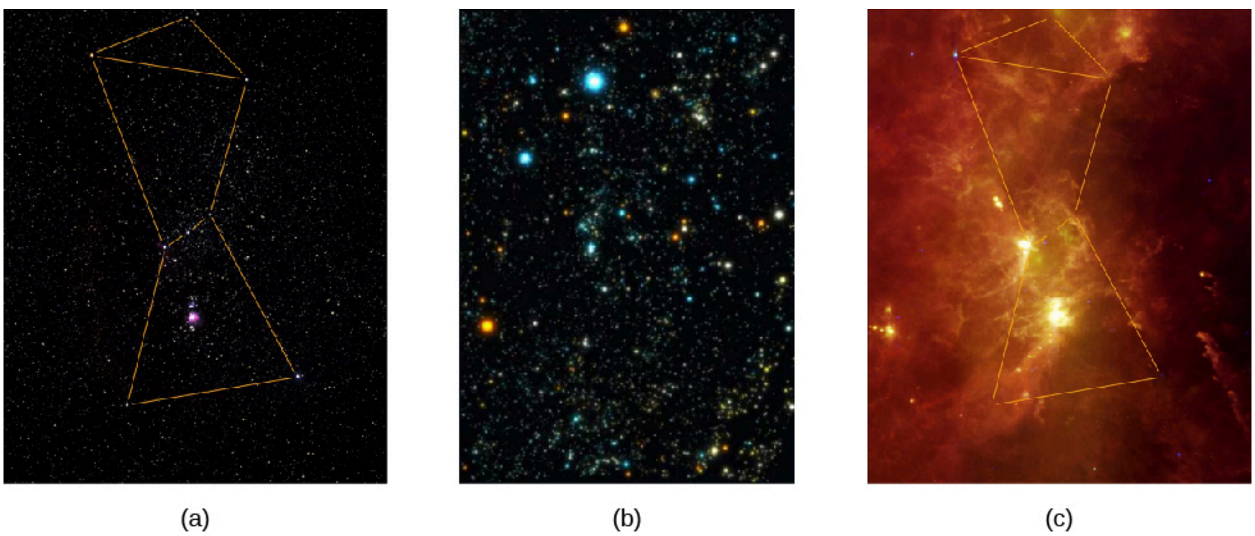


Figure 34.2 The same part of the sky looks different when observed with instruments that are sensitive to different bands of the spectrum. (a) Visible light: this shows part of the Orion region as the human eye sees it, with dotted lines added to show the figure of the mythical hunter, Orion. (b) X-rays: here, the view emphasizes the point-like X-ray sources nearby. The colors are artificial, changing from yellow to white to blue with increasing energy of the X-rays. The bright, hot stars in Orion are still seen in this image, but so are many other objects located at very different distances, including other stars, star corpses, and galaxies at the edge of the observable universe. (c) Infrared radiation: here, we mainly see the glowing dust in this region. (credit a: modification of work by Howard McCallon/NASA/IRAS; credit b: modification of work by Howard McCallon/NASA/IRAS; credit c: modification of work by Michael F. Corcoran)

The history of the development of astronomical telescopes is about how new technologies have been applied to improve the efficiency of these three basic components: the telescopes, the wavelength-sorting device, and the detectors. Let's first look at the development of the telescope.

Many ancient cultures built special sites for observing the sky (**Figure 34.3**). At these ancient *observatories*, they could measure the positions of celestial objects, mostly to keep track of time and date. Many of these ancient observatories had religious and ritual functions as well. The eye was the only device available to gather light, all of the colors in the light were observed at once, and the only permanent record of the observations was made by human beings writing down or sketching what they saw.

Figure 34.3 (a) Machu Picchu is a fifteenth century Incan site located in Peru. (b) Stonehenge, a prehistoric site (3000–2000 BCE), is located in England. (credit a: modification of work by Allard Schmidt)

While Hans Lippershey, Zaccharias Janssen, and Jacob Metius are all credited with the invention of the telescope around 1608—applying for patents within weeks of each other—it was Galileo who, in 1610, used this simple tube with lenses (which he called a spyglass) to observe the sky and gather more light than his eyes alone could. Even his small telescope—used over many nights—revolutionized ideas about the nature of the planets and the position of Earth.

How Telescopes Work

Telescopes have come a long way since Galileo's time. Now they tend to be huge devices; the most expensive cost hundreds of millions to billions of dollars. (To provide some reference point, however, keep in mind that just renovating college football stadiums typically costs hundreds of millions of dollars—with the most expensive recent renovation, at Texas A&M University's Kyle Field, costing $450 million.) The reason astronomers keep building bigger and bigger telescopes is that celestial objects—such as planets, stars, and galaxies—send much more light to Earth than any human eye (with its tiny opening) can catch, and bigger telescopes can detect fainter objects. If you have ever watched the stars with a group of friends, you know that there's plenty of starlight to go around; each of you can see each of the stars. If a thousand more people were watching, each of them would also catch a bit of each star's light. Yet, as far as you are concerned, the light not shining into your eye is wasted. It would be great if some of this "wasted" light could also be captured and brought to your eye. This is precisely what a telescope does.

The most important functions of a telescope are (1) to *collect* the faint light from an astronomical source and (2) to *focus* all the light into a point or an image. Most objects of interest to astronomers are extremely faint: the more light we can collect, the better we can study such objects. (And remember, even though we are focusing on visible light first, there are many telescopes that collect other kinds of electromagnetic radiation.)

Telescopes that collect visible radiation use a lens or mirror to gather the light. Other types of telescopes may use collecting devices that look very different from the lenses and mirrors with which we are familiar, but they serve the same function. In all types of telescopes, the light-gathering ability is determined by the area of the device acting as the light-gathering "bucket." Since most telescopes have mirrors or lenses, we can compare their light-gathering power by comparing the **apertures**, or diameters, of the opening through which light travels or reflects.

The amount of light a telescope can collect increases with the size of the aperture. A telescope with a mirror that is 4 meters in diameter can collect 16 times as much light as a telescope that is 1 meter in diameter. (The diameter is squared because the area of a circle equals $\pi d^2/4$, where d is the diameter of the circle.)

Example 34.1

Calculating the Light-Collecting Area

What is the area of a 1-m diameter telescope? A 4-m diameter one?

Solution

Using the equation for the area of a circle,

$$A = \frac{\pi d^2}{4}$$

the area of a 1-m telescope is

$$\frac{\pi d^2}{4} = \frac{\pi (1\text{ m})^2}{4} = 0.79\text{ m}^2$$

and the area of a 4-m telescope is

$$\frac{\pi d^2}{4} = \frac{\pi (4\text{ m})^2}{4} = 12.6\text{ m}^2$$

Check Your Learning

Show that the ratio of the two areas is 16:1.

Answer:

$\frac{12.6\text{ m}^2}{0.79\text{ m}^2} = 16.$ Therefore, with 16 times the area, a 4-m telescope collects 16 times the light of a 1-m telescope.

After the telescope forms an image, we need some way to detect and record it so that we can measure, reproduce, and analyze the image in various ways. Before the nineteenth century, astronomers simply viewed images with their eyes and wrote descriptions of what they saw. This was very inefficient and did not lead to a very reliable long-term record; you know from crime shows on television that eyewitness accounts are often inaccurate.

In the nineteenth century, the use of photography became widespread. In those days, photographs were a chemical record of an image on a specially treated glass plate. Today, the image is generally detected with sensors similar to those in digital cameras, recorded electronically, and stored in computers. This permanent record can then be used for detailed and quantitative studies. Professional astronomers rarely look through the large telescopes that they use for their research.

Formation of an Image by a Lens or a Mirror

Whether or not you wear glasses, you see the world through lenses; they are key elements of your eyes. A lens is a transparent piece of material that bends the rays of light passing through it. If the light rays are parallel as they enter, the lens brings them together in one place to form an image (Figure 34.4). If the curvatures of the lens surfaces are just right, all parallel rays of light (say, from a star) are bent, or *refracted*, in such a way that they converge toward a point, called the **focus** of the lens. At the focus, an image of the light source appears. In the case of parallel light rays, the distance from the lens to the location where the light rays focus, or image, behind the lens is called the *focal length* of the lens.

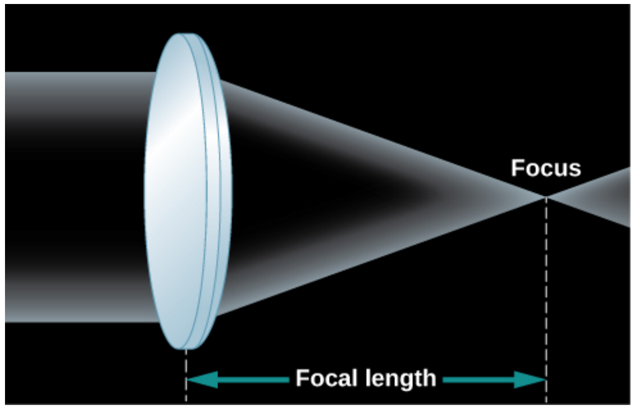


Figure 34.4 Parallel rays from a distant source are bent by the convex lens so that they all come together in a single place (the focus) to form an image.

As you look at **Figure 34.4**, you may ask why two rays of light from the same star would be parallel to each other. After all, if you draw a picture of star shining in all directions, the rays of light coming from the star don't look parallel at all. But remember that the stars (and other astronomical objects) are all extremely far away. By the time the few rays of light pointed toward us actually arrive at Earth, they are, for all practical purposes, parallel to each other. Put another way, any rays that were *not* parallel to the ones pointed at Earth are now heading in some very different direction in the universe.

To view the image formed by the lens in a telescope, we use an additional lens called an **eyepiece**. The eyepiece focuses the image at a distance that is either directly viewable by a human or at a convenient place for a detector. Using different eyepieces, we can change the *magnification* (or size) of the image and also redirect the light to a more accessible location. Stars look like points of light, and magnifying them makes little difference, but the image of a planet or a galaxy, which has structure, can often benefit from being magnified.

Many people, when thinking of a telescope, picture a long tube with a large glass lens at one end. This design, which uses a lens as its main optical element to form an image, as we have been discussing, is known as a *refractor* (**Figure 34.5**), and a telescope based on this design is called a **refracting telescope**. Galileo's telescopes were refractors, as are today's binoculars and field glasses. However, there is a limit to the size of a refracting telescope. The largest one ever built was a 49-inch refractor built for the Paris 1900 Exposition, and it was dismantled after the Exposition. Currently, the largest refracting telescope is the 40-inch refractor at Yerkes Observatory in Wisconsin.

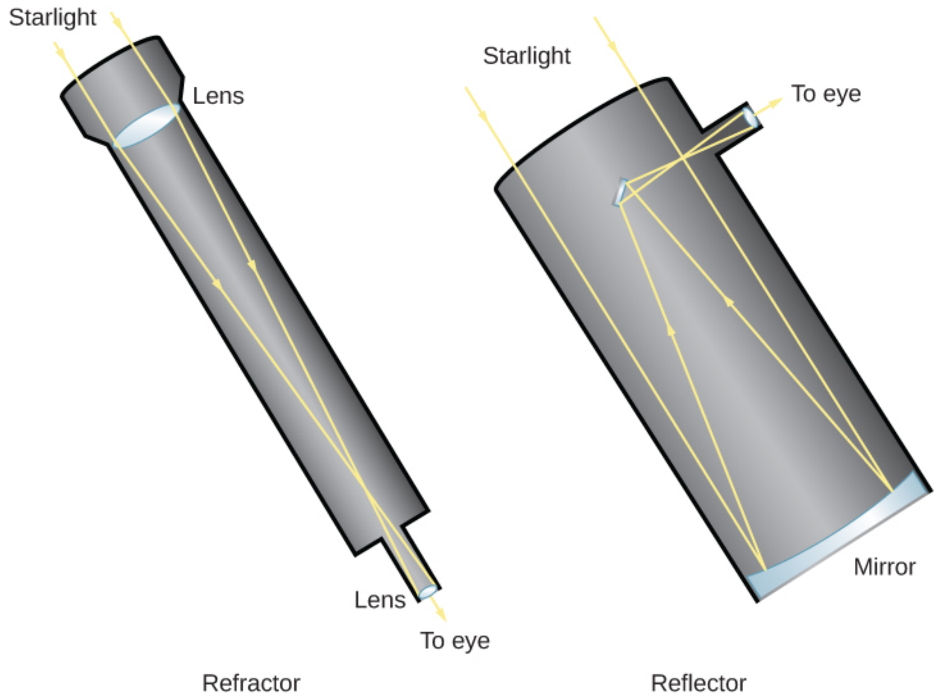


Figure 34.5 Light enters a refracting telescope through a lens at the upper end, which focuses the light near the bottom of the telescope. An eyepiece then magnifies the image so that it can be viewed by the eye, or a detector like a photographic plate can be placed at the focus. The upper end of a reflecting telescope is open, and the light passes through to the mirror located at the bottom of the telescope. The mirror then focuses the light at the top end, where it can be detected. Alternatively, as in this sketch, a second mirror may reflect the light to a position outside the telescope structure, where an observer can have easier access to it. Professional astronomers' telescopes are more complicated than this, but they follow the same principles of reflection and refraction.

One problem with a refracting telescope is that the light must pass *through* the lens of a refractor. That means the glass must be perfect all the way through, and it has proven very difficult to make large pieces of glass without flaws and bubbles in them. Also, optical properties of transparent materials change a little bit with the wavelengths (or colors) of light, so there is some additional distortion, known as **chromatic aberration**. Each wavelength focuses at a slightly different spot, causing the image to appear blurry.

In addition, since the light must pass through the lens, the lens can only be supported around its edges (just like the frames of our eyeglasses). The force of gravity will cause a large lens to sag and distort the path of the light rays as they pass through it. Finally, because the light passes through it, both sides of the lens must be manufactured to precisely the right shape in order to produce a sharp image.

A different type of telescope uses a concave *primary mirror* as its main optical element. The mirror is curved like the inner surface of a sphere, and it reflects light in order to form an image (**Figure 34.5**). Telescope mirrors are coated with a shiny metal, usually silver, aluminum, or, occasionally, gold, to make them highly reflective. If the mirror has the correct shape, all parallel rays are reflected back to the same point, the focus of the mirror. Thus, images are produced by a mirror exactly as they are by a lens.

Telescopes designed with mirrors avoid the problems of refracting telescopes. Because the light is reflected from the front surface only, flaws and bubbles within the glass do not affect the path of the light. In a telescope designed with mirrors, only the front surface has to be manufactured to a precise shape, and the mirror can be supported from the back. For these reasons, most astronomical telescopes today (both amateur and professional) use a mirror rather than a lens to form an image; this type of telescope is called a **reflecting telescope**. The first successful reflecting telescope was built by Isaac Newton in 1668.

In a reflecting telescope, the concave mirror is placed at the bottom of a tube or open framework. The mirror reflects the light back up the tube to form an image near the front end at a location called the **prime focus**. The image can be observed at the prime focus, or additional mirrors can intercept the light and redirect it to a position where the observer can view it more easily (**Figure 34.6**). Since an astronomer at the prime focus can block much of the light coming to the main mirror, the use of a small *secondary mirror* allows more light to get through the system.

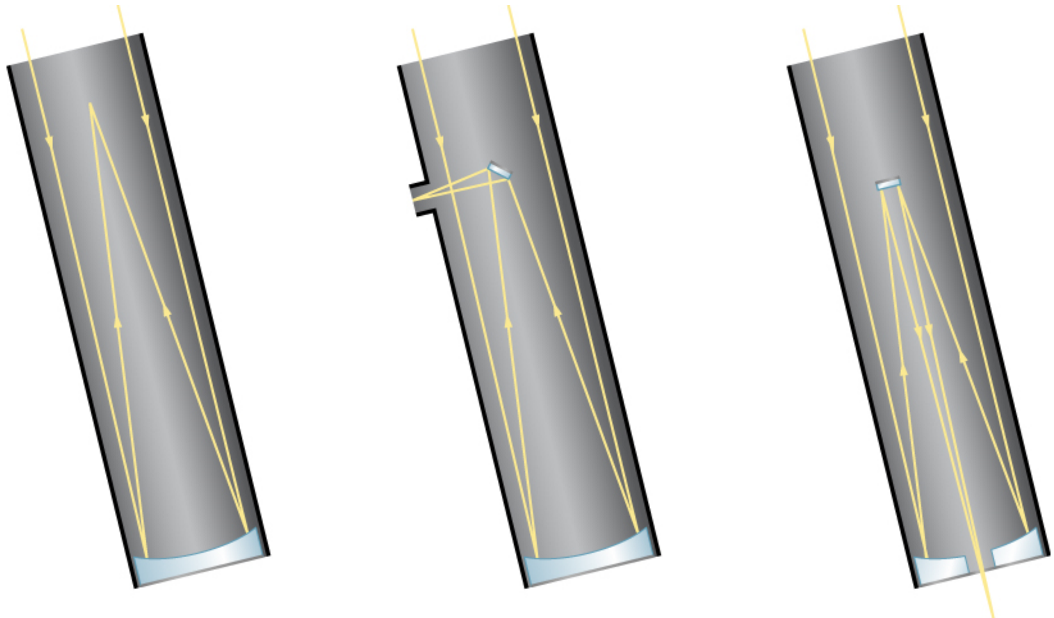


Figure 34.6 Reflecting telescopes have different options for where the light is brought to a focus. With prime focus, light is detected where it comes to a focus after reflecting from the primary mirror. With Newtonian focus, light is reflected by a small secondary mirror off to one side, where it can be detected (see also **Figure 34.5**). Most large professional telescopes have a Cassegrain focus in which light is reflected by the secondary mirror down through a hole in the primary mirror to an observing station below the telescope.

Choosing Your Own Telescope

If this course whets your appetite for exploring the sky further, you may be thinking about buying your own telescope. Many excellent amateur telescopes are available, and some research is required to find the best model for your needs. Some good sources of information about personal telescopes are the two popular US magazines aimed at amateur astronomers: *Sky & Telescope* and *Astronomy*. Both carry regular articles with advice, reviews, and advertisements from reputable telescope dealers. Monthly issues of both these magazines are available in the Physics Library, Olin Hall Room 215.

Some of the factors that determine which telescope is right for you depend upon your preferences:

- Will you be setting up the telescope in one place and leaving it there, or do you want an instrument that is portable and can come with you on outdoor excursions? How portable should it be, in terms of size and weight?
- Do you want to observe the sky with your eyes only, or do you want to take photographs? (Long-exposure photography, for example, requires a good clock drive to turn your telescope to compensate for Earth's rotation.)
- What types of objects will you be observing? Are you interested primarily in comets, planets, star clusters, or galaxies, or do you want to observe all kinds of celestial sights?

You may not know the answers to some of these questions yet. For this reason, you may want to "test-drive" some telescopes first. Many communities have amateur astronomy clubs that sponsor star parties open to the public. The members of those clubs often know a lot about telescopes and can share their ideas with you. The Olin Observatory at Gustavus is home to several different kinds of telescope. You can become familiar with those and get an idea of the advantages and disadvantages of each

Furthermore, you may already have an instrument like a telescope at home (or have access to one through a relative or friend). Many amateur astronomers recommend starting your survey of the sky with a good pair of binoculars. These are easily carried around and can show you many objects not visible (or clear) to the unaided eye.

When you are ready to purchase a telescope, you might find the following ideas useful:

- The key characteristic of a telescope is the aperture of the main mirror or lens; when someone says they have a 6-inch or 8-inch telescope, they mean the diameter of the collecting surface. The larger the aperture, the more light you can gather, and the fainter the objects you can see or photograph.
- Telescopes of a given aperture that use lenses (refractors) are typically more expensive than those using mirrors (reflectors) because both sides of a lens must be polished to great accuracy. And, because the light passes through it, the lens must be made of high-quality glass throughout. In contrast, only the front surface of a mirror must be accurately polished.
- Magnification is not one of the criteria on which to base your choice of a telescope. As we discussed, the magnification of the image is done by a smaller eyepiece, so the magnification can be adjusted by changing eyepieces. However, a telescope will magnify not only the astronomical object you are viewing but also the turbulence of Earth’s atmosphere. If the magnification is too high, your image will shimmer and shake and be difficult to view. A good telescope will come with a variety of eyepieces that stay within the range of useful magnification.
- The mount of a telescope (the structure on which it rests) is one of its most critical elements. Because a telescope shows a tiny field of view, which is magnified significantly, even the smallest vibration or jarring of the telescope can move the object you are viewing around or out of your field of view. A sturdy and stable mount is essential for serious viewing or photography (although it clearly affects how portable your telescope can be).
- A telescope requires some practice to set up and use effectively. Don’t expect everything to go perfectly on your first try. Take some time to read the instructions. If a local amateur astronomy club is nearby, use it as a resource.

34.2 | Telescopes Today

Learning Objectives

After completing this section, you will be able to:

- Recognize the largest visible-light and infrared telescopes in operation today
- Discuss the factors relevant to choosing an appropriate telescope site
- Define the technique of adaptive optics and describe the effects of the atmosphere on astronomical observations

Since Newton’s time, when the sizes of the mirrors in telescopes were measured in inches, reflecting telescopes have grown ever larger. In 1948, US astronomers built a telescope with a 5-meter (200-inch) diameter mirror on Palomar Mountain in Southern California. It remained the largest visible-light telescope in the world for several decades. The giants of today, however, have primary mirrors (the largest mirrors in the telescope) that are 8- to 10-meters in diameter, and larger ones are being built (**Figure 34.7**).

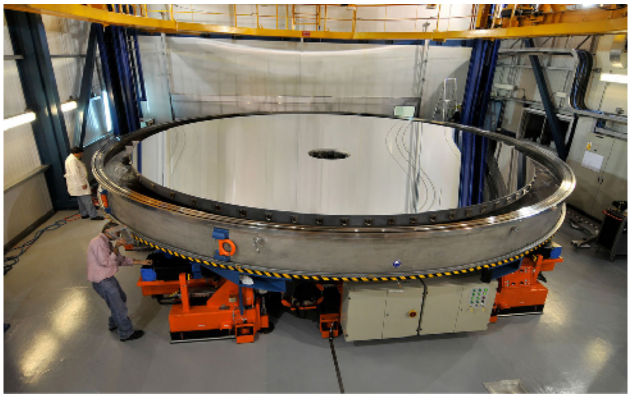


Figure 34.7 This image shows one of the primary mirrors of the European Southern Observatory's Very Large Telescope, named Yepun, just after it was recoated with aluminum. The mirror is a little over 8 meters in diameter. (credit: ESO/G. Huedepohl)

Modern Visible-Light and Infrared Telescopes

The decades starting in 1990 saw telescope building around the globe grow at an unprecedented rate. (See **Table 34.1**, which also includes websites for each telescope in case you want to visit or learn more about them.) Technological advancements had finally made it possible to build telescopes significantly larger than the 5-meter telescope at Palomar at a reasonable cost. New technologies have also been designed to work well in the infrared, and not just visible, wavelengths.

Large Single-Dish Visible-Light and Infrared Telescopes

Aperture (m)	Telescope Name	Location	Status	Website
39	European Extremely Large Telescope (E-ELT)	Cerro Armazonas, Chile	First light 2025 (estimated)	www.eso.org/sci/facilities/eelt
30	Thirty-Meter Telescope (TMT)	Mauna Kea, HI	First light 2025 (estimated)	www.tmt.org
24.5	Giant Magellan Telescope (GMT)	Las Campanas Observatory, Chile	First light 2025 (estimated)	www.gmto.org
11.1 × 9.9	Southern African Large Telescope (SALT)	Sutherland, South Africa	2005	www.salt.ac.za
10.4	Gran Telescopio Canarias (GTC)	La Palma, Canary Islands	First light 2007	http://www.gtc.iac.es
10.0	Keck I and II (two telescopes)	Mauna Kea, HI	Completed 1993–96	www.keckobservatory.org
9.1	Hobby–Eberly Telescope (HET)	Mount Locke, TX	Completed 1997	www.as.utexas.edu/mcdonald/het
8.4	Large Binocular Telescope (LBT) (two telescopes)	Mount Graham, AZ	First light 2004	www.lbto.org

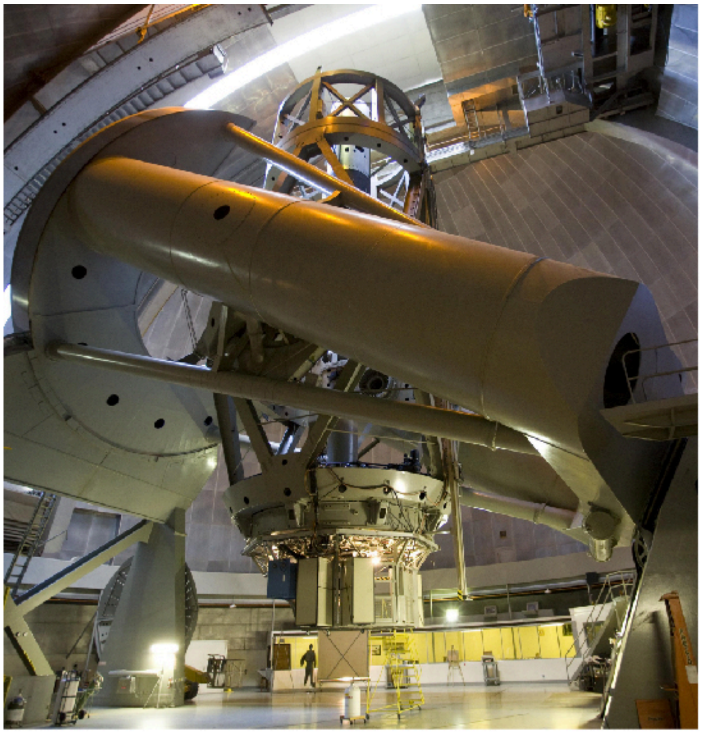
Table 34.1

Large Single-Dish Visible-Light and Infrared Telescopes

Aperture (m)	Telescope Name	Location	Status	Website
8.4	Large Synoptic Survey Telescope (LSST)	The Cerro Pachón, Chile	First light 2021	www.lsst.org
8.3	Subaru Telescope	Mauna Kea, HI	First light 1998	www.naoj.org
8.2	Very Large Telescope (VLT)	Cerro Paranal, Chile	All four telescopes completed 2000	www.eso.org/public/teles-instr/paranal
8.1	Gemini North and Gemini South	Mauna Kea, HI (North) and Cerro Pachón, Chile (South)	First light 1999 (North), First light 2000 (South)	www.gemini.edu
6.5	Magellan Telescopes (two telescopes: Baade and Landon Clay)	Las Campanas, Chile	First light 2000 and 2002	obs.carnegiescience.edu/Magellan
6.5	Multi-Mirror Telescope (MMT)	Mount Hopkins, AZ	Completed 1979	www.mmto.org
6.0	Big Telescope Altazimuth (BTA-6)	Mount Pastukhov, Russia	Completed 1976	w0.sao.ru/Doc-en/Telescopes/bta/descrip.html
5.1	Hale Telescope	Mount Palomar, CA	Completed 1948	www.astro.caltech.edu/palomar/about/telescopes/hale.html

Table 34.1

The differences between the Palomar telescope and the modern Gemini North telescope (to take an example) are easily seen in **Figure 34.8**. The Palomar telescope is a massive steel structure designed to hold the 14.5-ton primary mirror with a 5-meter diameter. Glass tends to sag under its own weight; hence, a huge steel structure is needed to hold the mirror. A mirror 8 meters in diameter, the size of the Gemini North telescope, if it were built using the same technology as the Palomar telescope, would have to weigh at least eight times as much and would require an enormous steel structure to support it.



(a)

(b)

Figure 34.8 (a) The Palomar 5-meter reflector: The Hale telescope on Palomar Mountain has a complex mounting structure that enables the telescope (in the open "tube" pointing upward in this photo) to swing easily into any position. (b) The Gemini North 8-meter telescope: The Gemini North mirror has a larger area than the Palomar mirror, but note how much less massive the whole instrument seems. (credit a: modification of work by Caltech/Palomar Observatory; credit b: modification of work by Gemini Observatory/AURA)

The 8-meter Gemini North telescope looks like a featherweight by contrast, and indeed it is. The mirror is only about 8 inches thick and weighs 24.5 tons, less than twice as much as the Palomar mirror. The Gemini North telescope was completed about 50 years after the Palomar telescope. Engineers took advantage of new technologies to build a telescope that is much lighter in weight relative to the size of the primary mirror. The Gemini mirror does sag, but with modern computers, it is possible to measure that sag many times each second and apply forces at 120 different locations to the back of the mirror to correct the sag, a process called *active control*. Seventeen telescopes with mirrors 6.5 meters in diameter and larger have been constructed since 1990.

The twin 10-meter Keck telescopes on Mauna Kea, which were the first of these new-technology instruments, use precision control in an entirely novel way. Instead of a single primary mirror 10 meters in diameter, each Keck telescope achieves its larger aperture by combining the light from 36 separate hexagonal mirrors, each 1.8 meters wide (**Figure 34.9**). Computer-controlled actuators (motors) constantly adjust these 36 mirrors so that the overall reflecting surface acts like a single mirror with just the right shape to collect and focus the light into a sharp image.

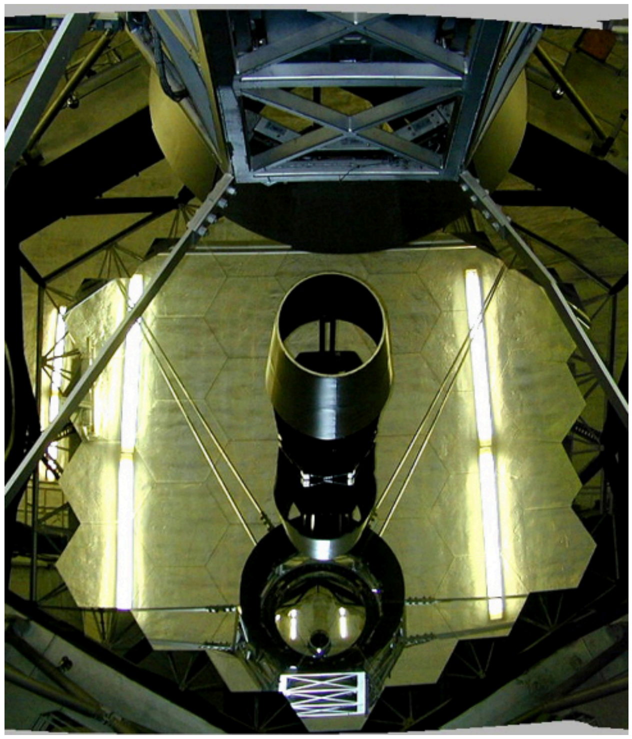


Figure 34.9 The mirror of the 10-meter Keck telescope is composed of 36 hexagonal sections. (credit: NASA)

Learn more about the Keck Observatory on Mauna Kea (https://openstaxcollege.org/l/30KeckObserv) through this History Channel clip on the telescopes and the work that they do.

In addition to holding the mirror, the steel structure of a telescope is designed so that the entire telescope can be pointed quickly toward any object in the sky. Since Earth is rotating, the telescope must have a motorized drive system that moves it very smoothly from east to west at exactly the same rate that Earth is rotating from west to east, so it can continue to point at the object being observed. All this machinery must be housed in a dome to protect the telescope from the elements. The dome has an opening in it that can be positioned in front of the telescope and moved along with it, so that the light from the objects being observed is not blocked.

George Ellery Hale: Master Telescope Builder

George Ellery Hale (Figure 34.10) was a giant among early telescope builders. Not once, but four times, he initiated projects that led to the construction of what was the world's largest telescope at the time. And he was a master at winning over wealthy benefactors to underwrite the construction of these new instruments.

Figure 34.10 Hale's work led to the construction of several major telescopes, including the 40-inch refracting telescope at Yerkes Observatory, and three reflecting telescopes: the 60-inch Hale and 100-inch Hooker telescopes at Mount Wilson Observatory, and the 200-inch Hale Telescope at Palomar Observatory.

Hale's training and early research were in solar physics. In 1892, at age 24, he was named associate professor of astral physics and director of the astronomical observatory at the University of Chicago. At the time, the largest telescope in the world was the 36-inch refractor at the Lick Observatory near San Jose, California. Taking advantage of an existing glass blank for a 40-inch telescope, Hale set out to raise money for a larger telescope than the one at Lick. One prospective donor was Charles T. Yerkes, who, among other things, ran the trolley system in Chicago.

Hale wrote to Yerkes, encouraging him to support the construction of the giant telescope by saying that "the donor could have no more enduring monument. It is certain that Mr. Lick's name would not have been nearly so widely known today were it not for the famous observatory established as a result of his munificence." Yerkes agreed, and the new telescope was completed in May 1897; it remains the largest refractor in the world (**Figure 34.11**).

Figure 34.11 The Yerkes 40-inch (1-meter) telescope.

Even before the completion of the Yerkes refractor, Hale was not only dreaming of building a still larger telescope but was also taking concrete steps to achieve that goal. In the 1890s, there was a major controversy about the relative quality of refracting and reflecting telescopes. Hale realized that 40 inches was close to the maximum feasible aperture for refracting telescopes. If telescopes with significantly larger apertures were to be built, they would have to be reflecting telescopes.

Using funds borrowed from his own family, Hale set out to construct a 60-inch reflector. For a site, he left the Midwest for the much better conditions on Mount Wilson—at the time, a wilderness peak above the small city of Los Angeles. In 1904, at the age of 36, Hale received funds from the Carnegie Foundation to establish the Mount Wilson Observatory. The 60-inch mirror was placed in its mount in December 1908.

Two years earlier, in 1906, Hale had already approached John D. Hooker, who had made his fortune in hardware and steel pipe, with a proposal to build a 100-inch telescope. The technological risks were substantial. The 60-inch telescope was not yet complete, and the usefulness of large reflectors for astronomy had yet to be demonstrated. George Ellery Hale's brother called him "the greatest gambler in the world." Once again, Hale successfully obtained funds, and the 100-inch telescope was completed in November 1917. (It was with this telescope that Edwin Hubble was able to establish that the spiral nebulae were separate islands of stars—or galaxies—quite removed from our own Milky Way.)

Hale was not through dreaming. In 1926, he wrote an article in *Harper's Magazine* about the scientific value of a still larger telescope. This article came to the attention of the Rockefeller Foundation, which granted $6 million for the construction of a 200-inch telescope. Hale died in 1938, but the 200-inch (5-meter) telescope on Palomar Mountain was dedicated 10 years later and is now named in Hale's honor.

Picking the Best Observing Sites

A telescope like the Gemini or Keck telescope costs about $100 million to build. That kind of investment demands that the telescope be placed in the best possible site. Since the end of the nineteenth century, astronomers have realized that the best observatory sites are on mountains, far from the lights and pollution of cities. Although a number of urban observatories remain, especially in the large cities of Europe, they have become administrative centers or museums. The real action takes place far away, often on desert mountains or isolated peaks in the Atlantic and Pacific Oceans, where we find the staff's living quarters, computers, electronic and machine shops, and of course the telescopes themselves. A large observatory today requires a supporting staff of 20 to 100 people in addition to the astronomers.

The performance of a telescope is determined not only by the size of its mirror but also by its location. Earth's atmosphere, so vital to life, presents challenges for the observational astronomer. In at least four ways, our air imposes limitations on the usefulness of telescopes:

1. The most obvious limitation is weather conditions such as clouds, wind, and rain. At the best sites, the weather is clear as much as 75% of the time.
2. Even on a clear night, the atmosphere filters out a certain amount of starlight, especially in the infrared, where the absorption is due primarily to water vapor. Astronomers therefore prefer dry sites, generally found at high altitudes.
3. The sky above the telescope should be dark. Near cities, the air scatters the glare from lights, producing an illumination that hides the faintest stars and limits the distances that can be probed by telescopes. (Astronomers call this effect *light pollution*.) Observatories are best located at least 100 miles from the nearest large city.
4. Finally, the air is often unsteady; light passing through this turbulent air is disturbed, resulting in blurred star images. Astronomers call these effects "bad **seeing**." When seeing is bad, images of celestial objects are distorted by the constant twisting and bending of light rays by turbulent air.

The best observatory sites are therefore high, dark, and dry. The world's largest telescopes are found in such remote mountain locations as the Andes Mountains of Chile (**Figure 34.12**), the desert peaks of Arizona, the Canary Islands in the Atlantic Ocean, and Mauna Kea in Hawaii, a dormant volcano with an altitude of 13,700 feet (4200 meters).

Light pollution is a problem not just for professional astronomers but for everyone who wants to enjoy the beauty of the night sky. In addition research is now showing that it can disrupt the life cycle of animals with whom we share the urban and suburban landscape. And the light wasted shining into the sky leads to unnecessary municipal expenses and use of fossil fuels. Concerned people have formed an organization, the International Dark-Sky Association, whose **website (https://openstaxcollege.org/l/30IntDSA)** is full of good information. A citizen science project called **Globe at Night (https://openstaxcollege.org/l/30GlbatNght)** allows you to measure the light levels in your community by counting stars and to compare it to others around the world. And, if you get interested in this topic and want to do a paper for your astronomy course or another course while you are in college, the **Dark Night Skies guide (https://openstaxcollege.org/l/30DNSGuide)** can point you to a variety of resources on the topic.

Figure 34.12 Cerro Paranal, a mountain summit 2.7 kilometers above sea level in Chile's Atacama Desert, is the site of the European Southern Observatory's Very Large Telescope. This photograph shows the four 8-meter telescope buildings on the site and vividly illustrates that astronomers prefer high, dry sites for their instruments. The 4.1-meter Visible and Infrared Survey Telescope for Astronomy (VISTA) can be seen in the distance on the next mountain peak. (credit: ESO)

The Resolution of a Telescope

In addition to gathering as much light as they can, astronomers also want to have the sharpest images possible. **Resolution** refers to the precision of detail present in an image: that is, the smallest features that can be distinguished. Astronomers are always eager to make out more detail in the images they study, whether they are following the weather on Jupiter or trying to peer into the violent heart of a "cannibal galaxy" that recently ate its neighbor for lunch.

One factor that determines how good the resolution will be is the size of the telescope. Larger apertures produce sharper images. As we saw in **Section 33.5**, Rayleigh's Criterion (embodied in **Equation 33.5**) shows that the theoretical angle of minimum resolution is inversely proportional to the diameter of the aperture. Until very recently, however, visible-light and infrared telescopes on Earth's surface could not produce images as sharp as the theory of light said they should.

The problem—as we saw earlier in this chapter—is our planet's atmosphere, which is turbulent. It contains many small-scale blobs or cells of gas that range in size from inches to several feet. Each cell has a slightly different temperature from its neighbor, and each cell acts like a lens, bending (refracting) the path of the light by a small amount. This bending slightly changes the position where each light ray finally reaches the detector in a telescope. The cells of air are in motion, constantly being blown through the light path of the telescope by winds, often in different directions at different altitudes. As a result, the path followed by the light is constantly changing.

For an analogy, think about watching a parade from a window high up in a skyscraper. You decide to throw some confetti down toward the marchers. Even if you drop a handful all at the same time and in the same direction, air currents will toss the pieces around, and they will reach the ground at different places. As we described earlier, we can think of the light from the stars as a series of parallel beams, each making its way through the atmosphere. Each path will be slightly different, and each will reach the detector of the telescope at a slightly different place. The result is a blurred image, and because the cells are being blown by the wind, the nature of the blur will change many times each second. You have probably noticed this effect as the "twinkling" of stars seen from Earth. The light beams are bent enough that part of the time they reach your eye, and part of the time some of them miss, thereby making the star seem to vary in brightness. In space, however, the light of the stars is steady.

Astronomers search the world for locations where the amount of atmospheric blurring, or turbulence, is as small as possible. It turns out that the best sites are in coastal mountain ranges and on isolated volcanic peaks in the middle of an ocean. Air that has flowed long distances over water before it encounters land is especially stable.

The resolution of an image is measured in units of angle on the sky, typically in units of arcseconds. One arcsecond is 1/3600 degree, and there are 360 degrees in a full circle. So we are talking about tiny angles on the sky. To give you a sense of just how tiny, we might note that 1 arcsecond is how big a quarter would look when seen from a distance of 5 kilometers. The best images obtained from the ground with traditional techniques reveal details as small as several tenths of an arcsecond across. This image size is remarkably good. One of the main reasons for launching the Hubble Space Telescope was to escape Earth's atmosphere and obtain even sharper images.

But since we can't put every telescope into space, astronomers have devised a technique called **adaptive optics** that can beat Earth's atmosphere at its own game of blurring. This technique (which is most effective in the infrared region of the spectrum with our current technology) makes use of a small flexible mirror placed in the beam of a telescope. A sensor measures how much the atmosphere has distorted the image, and as often as 500 times per second, it sends instructions to the flexible mirror on how to change shape in order to compensate for distortions produced by the atmosphere. The light is thus brought back to an almost perfectly sharp focus at the detector. Figure 34.13 shows just how effective this technique is. With adaptive optics, ground-based telescopes can achieve resolutions of 0.1 arcsecond or a little better in the infrared region of the spectrum. This impressive figure is the equivalent of the resolution that the Hubble Space Telescope achieves in the visible-light region of the spectrum.

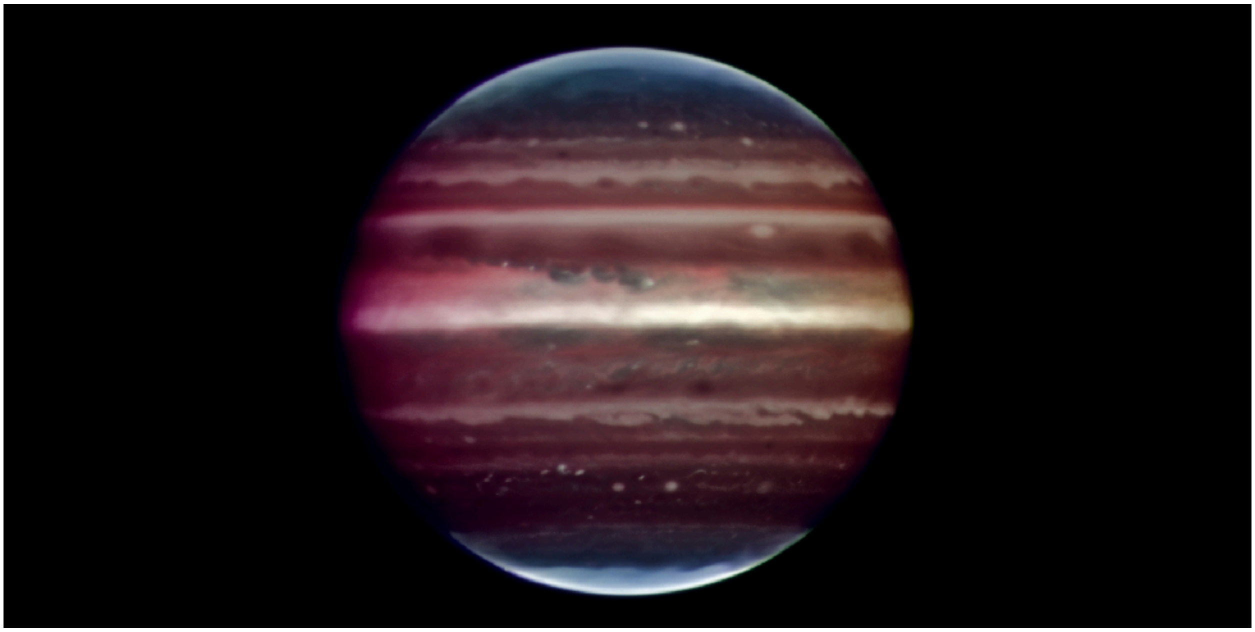


Figure 34.13 One of the clearest pictures of Jupiter ever taken from the ground, this image was produced with adaptive optics using an 8-meter-diameter telescope at the Very Large Telescope in Chile. Adaptive optics uses infrared wavelengths to remove atmospheric blurring, resulting in a much clearer image. (credit: modification of work by ESO, F.Marchis, M.Wong (UC Berkeley); E.Marchetti, P.Amico, S.Tordo (ESO))

How Astronomers Really Use Telescopes

In the popular view (and some bad movies), an astronomer spends most nights in a cold observatory peering through a telescope, but this is not very accurate today. Most astronomers do not live at observatories, but near the universities or laboratories where they work. An astronomer might spend only a week or so each year observing at the telescope and the rest of the time measuring or analyzing the data acquired from large project collaborations and dedicated surveys. Many astronomers use radio telescopes for space experiments, which work just as well during the daylight hours. Still others work at purely theoretical problems using supercomputers and never observe at a telescope of any kind.

Even when astronomers are observing with large telescopes, they seldom peer through them. Electronic detectors permanently record the data for detailed analysis later. At some observatories, observations may be made remotely, with the astronomer sitting at a computer thousands of miles away from the telescope.

Time on major telescopes is at a premium, and an observatory director will typically receive many more requests for telescope time than can be accommodated during the year. Astronomers must therefore write a convincing proposal explaining how they would like to use the telescope and why their observations will be important to the progress of astronomy. A committee of astronomers is then asked to judge and rank the proposals, and time is assigned only to those with the greatest merit. Even if your proposal is among the high-rated ones, you may have to wait many months for your turn. If the skies are cloudy on the nights you have been assigned, it may be more than a year before you get another chance.

Some older astronomers still remember long, cold nights spent alone in an observatory dome, with only music from a tape recorder or an all-night radio station for company. The sight of the stars shining brilliantly hour after hour through the open slit in the observatory dome was unforgettable. So, too, was the relief as the first pale light of dawn announced the end of a 12-hour observation session. Astronomy is much easier today, with teams of observers working together, often at their computers, in a warm room. Those who are more nostalgic, however, might argue that some of the romance has gone from the field, too.

34.3 | Visible-Light Detectors and Instruments

Learning Objectives

After completing this section, you will be able to:

- Describe the difference between photographic plates and charge-coupled devices
- Describe the unique difficulties associated with infrared observations and their solutions
- Describe how a spectrometer works

After a telescope collects radiation from an astronomical source, the radiation must be *detected* and measured. The first detector used for astronomical observations was the human eye, but it suffers from being connected to an imperfect recording and retrieving device—the human brain. Photography and modern electronic detectors have eliminated the quirks of human memory by making a permanent record of the information from the cosmos.

The eye also suffers from having a very short *integration time*; it takes only a fraction of a second to add light energy together before sending the image to the brain. One important advantage of modern detectors is that the light from astronomical objects can be collected by the detector over longer periods of time; this technique is called “taking a long exposure.” Exposures of several hours are required to detect very faint objects in the cosmos.

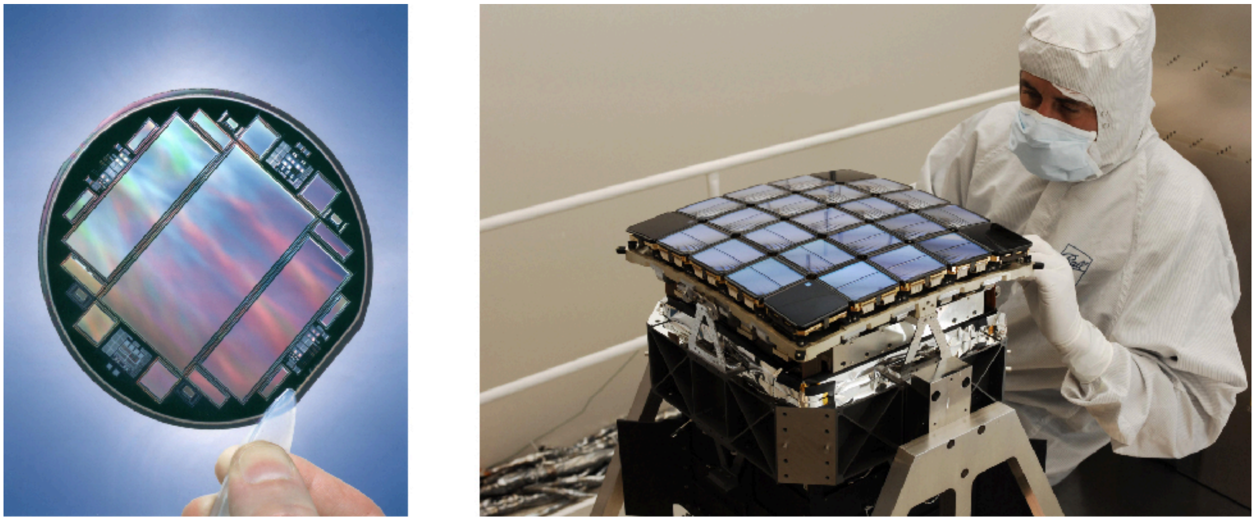
Before the light reaches the detector, astronomers today normally use some type of instrument to sort the light according to wavelength. The instrument may be as simple as colored filters, which transmit light within a specified range of wavelengths. A red transparent plastic is an everyday example of a filter that transmits only the red light and blocks the other colors. After the light passes through a filter, it forms an image that astronomers can then use to measure the apparent brightness and color of objects. We will show you many examples of such images in the later chapters of this book, and we will describe what we can learn from them.

Alternatively, the instrument between telescope and detector may be one of several devices that spread the light out into its full rainbow of colors so that astronomers can measure individual lines in the spectrum. Such an instrument (which you learned about in the chapter on **Spectroscopy**) is called a *spectrometer* because it allows astronomers to measure (to meter) the spectrum of a source of radiation. Whether a filter or a spectrometer, both types of wavelength-sorting instruments still have to use detectors to record and measure the properties of light.

Photographic and Electronic Detectors

Throughout most of the twentieth century, photographic film or *glass plates* served as the prime astronomical detectors, whether for photographing spectra or direct images of celestial objects. In a photographic plate, a light-sensitive chemical coating is applied to a piece of glass that, when developed, provides a lasting record of the image. At observatories around the world, vast collections of photographs preserve what the sky has looked like during the past 100 years. Photography represents a huge improvement over the human eye, but it still has limitations. Photographic films are inefficient: only about 1% of the light that actually falls on the film contributes to the chemical change that makes the image; the rest is wasted.

Astronomers today have much more efficient electronic detectors to record astronomical images. Most often, these are **charge-coupled devices (CCDs)**, which are similar to the detectors used in video camcorders or in digital cameras (like the one more and more students have on their cell phones) (see **Figure 34.14**). In a CCD, photons of radiation hitting any part of the detector generate a stream of charged particles (electrons) that are stored and counted at the end of the exposure. Each place where the radiation is counted is called a pixel (picture element), and modern detectors can count the photons in millions of pixels (megapixels, or MPs).



(a) (b)

Figure 34.14 (a) This CCD is a mere 300-micrometers thick (thinner than a human hair) yet holds more than 21 million pixels. (b) This matrix of 42 CCDs serves the Kepler telescope. (credit a: modification of work by US Department of Energy; credit b: modification of work by NASA and Ball Aerospace)

Because CCDs typically record as much as 60–70% of all the photons that strike them, and the best silicon and infrared CCDs exceed 90% sensitivity, we can detect much fainter objects. Among these are many small moons around the outer planets, icy dwarf planets beyond Pluto, and dwarf galaxies of stars. CCDs also provide more accurate measurements of the brightness of astronomical objects than photography, and their output is digital—in the form of numbers that can go directly into a computer for analysis.

Infrared Observations

Observing the universe in the infrared band of the spectrum presents some additional challenges. The infrared region extends from wavelengths near 1 micrometer (μm), which is about the long wavelength sensitivity limit of both CCDs and photography, to 100 micrometers or longer. Recall from the discussion on radiation and spectra that infrared is “heat radiation” (given off at temperatures that we humans are comfortable with). The main challenge to astronomers using infrared is to distinguish between the tiny amount of heat radiation that reaches Earth from stars and galaxies, and the much greater heat radiated by the telescope itself and our planet’s atmosphere.

Typical temperatures on Earth’s surface are near 300 K, and the atmosphere through which observations are made is only a little cooler. According to Wien’s law (from the section on **Blackbody Radiation**), the telescope, the observatory, and even the sky are radiating infrared energy with a peak wavelength of about 10 micrometers. To infrared eyes, everything on Earth is brightly aglow—including the telescope and camera (**Figure 34.15**). The challenge is to detect faint cosmic sources against this sea of infrared light. Another way to look at this is that an astronomer using infrared must always contend with the situation that a visible-light observer would face if working in broad daylight with a telescope and optics lined with bright fluorescent lights.

Figure 34.15 Infrared waves can penetrate places in the universe from which light is blocked, as shown in this infrared image where the plastic bag blocks visible light but not infrared. (credit: NASA/JPL-Caltech/R. Hurt (SSC))

To solve this problem, astronomers must protect the infrared detector from nearby radiation, just as you would shield photographic film from bright daylight. Since anything warm radiates infrared energy, the detector must be isolated in very cold surroundings; often, it is held near absolute zero (1 to 3 K) by immersing it in liquid helium. The second step is to reduce the radiation emitted by the telescope structure and optics, and to block this heat from reaching the infrared detector.

Check out **The Infrared Zoo (https://openstaxcollege.org/l/30IFZoo)** to get a sense of what familiar objects look like with infrared radiation. Slide the slider to change the wavelength of radiation for the picture, and click the arrow to see other animals.

Spectroscopy

Spectroscopy is one of the astronomer's most powerful tools, providing information about the composition, temperature, motion, and other characteristics of celestial objects. More than half of the time spent on most large telescopes is used for spectroscopy.

The many different wavelengths present in light can be separated by passing them through a spectrometer to form a spectrum. The design of a simple spectrometer is illustrated in Figure 34.16. Light from the source (actually, the image of a source produced by the telescope) enters the instrument through a small hole or narrow slit, and is collimated (made into a beam of parallel rays) by a lens. The light then passes through a prism, producing a spectrum: different wavelengths leave the prism in different directions because each wavelength is bent by a different amount when it enters and leaves the prism. A second lens placed behind the prism focuses the many different images of the slit or entrance hole onto a CCD or other detecting device. This collection of images (spread out by color) is the spectrum that astronomers can then analyze at a later point. As spectroscopy spreads the light out into more and more collecting bins, fewer photons go into each bin, so either a larger telescope is needed or the integration time must be greatly increased—usually both.

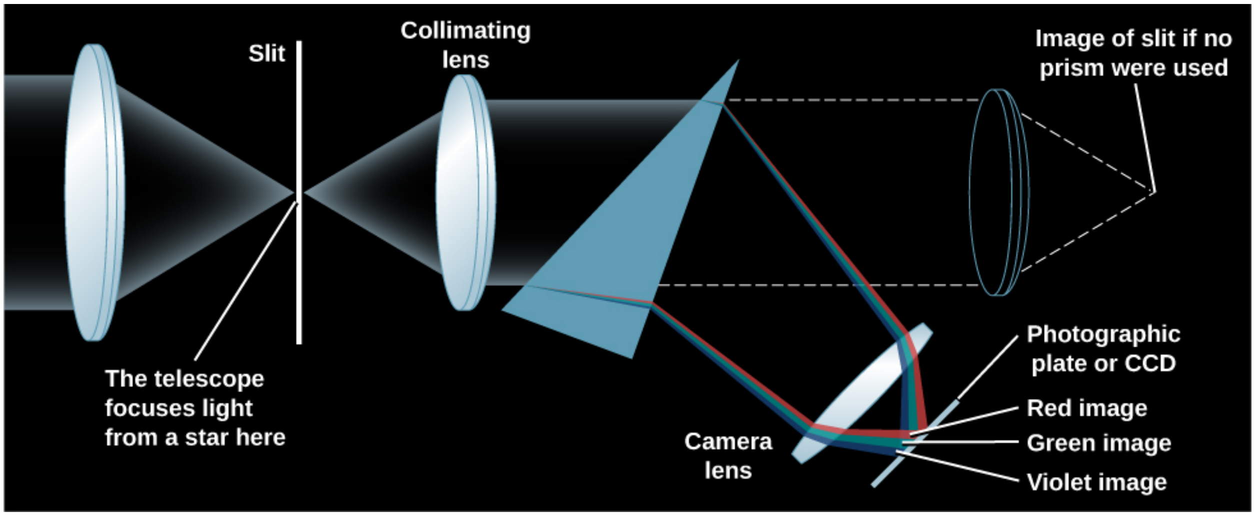


Figure 34.16 The light from the telescope is focused on a slit. A prism (or grating) disperses the light into a spectrum, which is then photographed or recorded electronically.

In practice, astronomers today are more likely to use a different device, called a *grating*, to disperse the spectrum. A grating is a piece of material with thousands of grooves on its surface. While it functions completely differently, a grating, like a prism, also spreads light out into a spectrum.

34.4 | Radio Telescopes

Learning Objectives

After completing this section, you will be able to:

- Describe how radio waves from space are detected
- Identify the world's largest radio telescopes
- Define the technique of interferometry and discuss the benefits of interferometers over single-dish telescopes

In addition to visible and infrared radiation, radio waves from astronomical objects can also be detected from the surface of Earth. In the early 1930s, Karl G. Jansky, an engineer at Bell Telephone Laboratories, was experimenting with antennas for long-range radio communication when he encountered some mysterious static—radio radiation coming from an unknown source (**Figure 34.17**). He discovered that this radiation came in strongest about four minutes earlier on each successive day and correctly concluded that since Earth's sidereal rotation period (how long it takes us to rotate relative to the stars) is four minutes shorter than a solar day, the radiation must be originating from some region fixed on the celestial sphere. Subsequent investigation showed that the source of this radiation was part of the Milky Way Galaxy; Jansky had discovered the first source of cosmic radio waves.

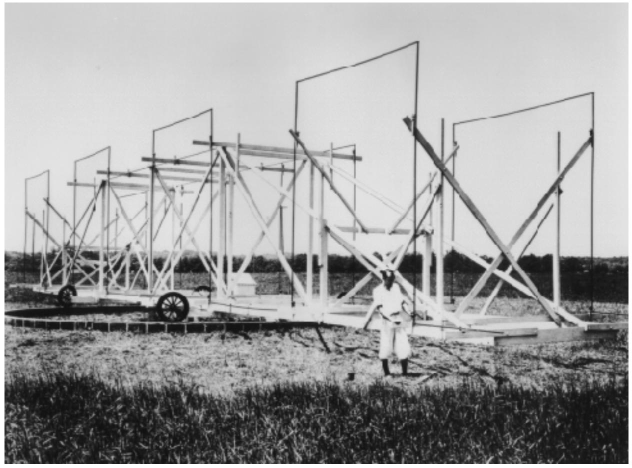


Figure 34.17 This rotating radio antenna was used by Jansky in his serendipitous discovery of radio radiation from the Milky Way.

In 1936, Grote Reber, who was an amateur astronomer interested in radio communications, used galvanized iron and wood to build the first antenna specifically designed to receive cosmic radio waves. Over the years, Reber built several such antennas and used them to carry out pioneering surveys of the sky for celestial radio sources; he remained active in radio astronomy for more than 30 years. During the first decade, he worked practically alone because professional astronomers had not yet recognized the vast potential of radio astronomy.

Detection of Radio Energy from Space

It is important to understand that radio waves cannot be "heard": they are not the sound waves you hear coming out of the radio receiver in your home or car. Like light, radio waves are a form of electromagnetic radiation, but unlike light, we cannot detect them with our senses—we must rely on electronic equipment to pick them up. In commercial radio broadcasting, we encode sound information (music or a newscaster's voice) into radio waves. These must be decoded at the other end and then turned back into sound by speakers or headphones.

The radio waves we receive from space do not, of course, have music or other program information encoded in them. If cosmic radio signals were translated into sound, they would sound like the static you hear when scanning between stations. Nevertheless, there is information in the radio waves we receive—information that can tell us about the chemistry and physical conditions of the sources of the waves.

Just as vibrating charged particles can produce electromagnetic waves (see the **The Nature of Light** chapter), electromagnetic waves can make charged particles move back and forth. Radio waves can produce a current in conductors of electricity such as metals. An antenna is such a conductor: it intercepts radio waves, which create a feeble current in it. The current is then amplified in a radio receiver until it is strong enough to measure or record. Like your television or radio, receivers can be tuned to select a single frequency (channel). In astronomy, however, it is more common to use sophisticated data-processing techniques that allow thousands of separate frequency bands to be detected simultaneously. Thus, the astronomical radio receiver operates much like a spectrometer on a visible-light or infrared telescope, providing information about how much radiation we receive at each wavelength or frequency. After computer processing, the radio signals are recorded on magnetic disks for further analysis.

Radio waves are reflected by conducting surfaces, just as light is reflected from a shiny metallic surface, and according to the same laws of optics. A radio-reflecting telescope consists of a concave metal reflector (called a *dish*), analogous to a telescope mirror. The radio waves collected by the dish are reflected to a focus, where they can then be directed to a receiver and analyzed. Because humans are such visual creatures, radio astronomers often construct a pictorial representation of the radio sources they observe. **Figure 34.18** shows such a radio image of a distant galaxy, where radio telescopes reveal vast jets and complicated regions of radio emissions that are completely invisible in photographs taken with light.

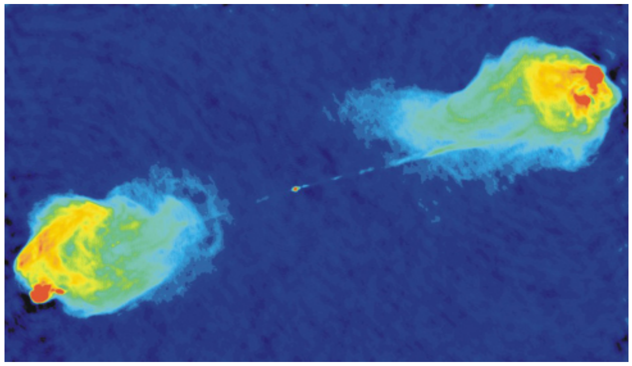


Figure 34.18 This image has been constructed of radio observations at the Very Large Array of a galaxy called Cygnus A. Colors have been added to help the eye sort out regions of different radio intensities. Red regions are the most intense, blue the least. The visible galaxy would be a small dot in the center of the image. The radio image reveals jets of expelled material (more than 160,000 light-years long) on either side of the galaxy. (credit: NRAO/AUI)

Radio astronomy is a young field compared with visible-light astronomy, but it has experienced tremendous growth in recent decades. The world's largest radio reflectors that can be pointed to any direction in the sky have apertures of 100 meters. One of these has been built at the US National Radio Astronomy Observatory in West Virginia (**Figure 34.19**). **Table 34.2** lists some of the major radio telescopes of the world.

Figure 34.19 This fully steerable radio telescope in West Virginia went into operation in August 2000. Its dish is about 100 meters across. (credit: modification of work by "b3nscott"/Flickr)

Major Radio Observatories of the World

Observatory	Location	Description	Website
Individual Radio Dishes			
Five-hundred-meter Aperture Spherical radio Telescope (FAST)	Guizhou, China	500-m fixed dish	fast.bao.ac.cn/en/

Table 34.2

Major Radio Observatories of the World

Observatory	Location	Description	Website
Arecibo Observatory	Arecibo, Puerto Rico	305-m fixed dish	www.naic.edu
Green Bank Telescope (GBT)	Green Bank, WV	110 × 100-m steerable dish	www.science.nrao.edu/facilities/gbt
Effelsberg 100-m Telescope	Bonn, Germany	100-m steerable dish	www.mpifr-bonn.mpg.de/en/effelsberg
Lovell Telescope	Manchester, England	76-m steerable dish	www.jb.man.ac.uk/aboutus/lovell
Canberra Deep Space Communication Complex (CDSCC)	Tidbinbilla, Australia	70-m steerable dish	www.cdscc.nasa.gov
Goldstone Deep Space Communications Complex (GDSCC)	Barstow, CA	70-m steerable dish	www.gdscc.nasa.gov
Parkes Observatory	Parkes, Australia	64-m steerable dish	www.parkes.atnf.csiro.au
Arrays of Radio Dishes			
Square Kilometre Array (SKA)	South Africa and Western Australia	Thousands of dishes, km^2 collecting area, partial array in 2020	www.skatelescope.org
Atacama Large Millimeter/submillimeter Array (ALMA)	Atacama desert, Northern Chile	66 7-m and 12-m dishes	www.almaobservatory.org
Jansky Very Large Array (VLA)	Socorro, New Mexico	27-element array of 25-m dishes (36-km baseline)	www.science.nrao.edu/facilities/vla
Westerbork Synthesis Radio Telescope (WSRT)	Westerbork, the Netherlands	12-element array of 25-m dishes (1.6-km baseline)	www.astron.nl/radio-observatory/public/public-0
Very Long Baseline Array (VLBA)	Ten US sites, HI to the Virgin Islands	10-element array of 25-m dishes (9000 km baseline)	www.science.nrao.edu/facilities/vlba
Australia Telescope Compact Array (ATCA)	Several sites in Australia	8-element array (seven 22-m dishes plus Parkes 64 m)	www.narrabri.atnf.csiro.au
Multi-Element Radio Linked Interferometer Network (MERLIN)	Cambridge, England, and other British sites	Network of seven dishes (the largest is 32 m)	www.e-merlin.ac.uk
Millimeter-wave Telescopes			
IRAM	Granada, Spain	30-m steerable mm-wave dish	www.iram-institute.org

Table 34.2

Major Radio Observatories of the World

Observatory	Location	Description	Website
James Clerk Maxwell Telescope (JCMT)	Mauna Kea, HI	15-m steerable mm-wave dish	www.eaobservatory.org/jcmt
Nobeyama Radio Observatory (NRO)	Minamimaki, Japan	6-element array of 10-m wave dishes	www.nro.nao.ac.jp/en
Hat Creek Radio Observatory (HCRO)	Cassel, CA	6-element array of 5-m wave dishes	www.sri.com/research-development/specialized-facilities/hat-creek-radio-observatory

Table 34.2

Radio Interferometry

As we discussed earlier, a telescope's ability to show us fine detail (its resolution) depends upon its aperture, but it also depends upon the wavelength of the radiation that the telescope is gathering. The longer the waves, the harder it is to resolve fine detail in the images or maps we make. Because radio waves have such long wavelengths, they present tremendous challenges for astronomers who need good resolution. In fact, even the largest radio dishes on Earth, operating alone, cannot make out as much detail as the typical small visible-light telescope used in a college astronomy lab. To overcome this difficulty, radio astronomers have learned to sharpen their images by linking two or more radio telescopes together electronically. Two or more telescopes linked together in this way are called an **interferometer**.

"Interferometer" may seem like a strange term because the telescopes in an interferometer work cooperatively; they don't "interfere" with each other. **Interference**, however, is a technical term for the way that multiple waves interact with each other when they arrive in our instruments, and this interaction allows us to coax more detail out of our observations. The resolution of an interferometer depends upon the separation of the telescopes, not upon their individual apertures. Two telescopes separated by 1 kilometer provide the same resolution as would a single dish 1 kilometer across (although they are not, of course, able to collect as much radiation as a radio-wave bucket that is 1 kilometer across).

To get even better resolution, astronomers combine a large number of radio dishes into an **interferometer array**. In effect, such an array works like a large number of two-dish interferometers, all observing the same part of the sky together. Computer processing of the results permits the reconstruction of a high-resolution radio image. The most extensive such instrument in the United States is the National Radio Astronomy Observatory's Jansky Very Large Array (VLA) near Socorro, New Mexico. It consists of 27 movable radio telescopes (on railroad tracks), each having an aperture of 25 meters, spread over a total span of about 36 kilometers. By electronically combining the signals from all of its individual telescopes, this array permits the radio astronomer to make pictures of the sky at radio wavelengths comparable to those obtained with a visible-light telescope, with a resolution of about 1 arcsecond.

The Atacama Large Millimeter/submillimeter array (ALMA) in the Atacama Desert of Northern Chile (Figure 34.20), at an altitude of 16,400 feet, consists of 12 7-meter and 54 12-meter telescopes, and can achieve baselines up to 16 kilometers. Since it became operational in 2013, it has made observations at resolutions down to 6 milliarcseconds (0.006 arcseconds), a remarkable achievement for radio astronomy.

Figure 34.20 Located in the Atacama Desert of Northern Chile, ALMA currently provides the highest resolution for radio observations. (credit: ESO/S. Guisard)

Watch this **documentary (https://openstax.org/l/30ALMAdoc)** that explains the work that went into designing and building ALMA, discusses some of its first images, and explores its future.

Initially, the size of interferometer arrays was limited by the requirement that all of the dishes be physically wired together. The maximum dimensions of the array were thus only a few tens of kilometers. However, larger interferometer separations can be achieved if the telescopes do not require a physical connection. Astronomers, with the use of current technology and computing power, have learned to time the arrival of electromagnetic waves coming from space very precisely at each telescope and combine the data later. If the telescopes are as far apart as California and Australia, or as West Virginia and Crimea in Ukraine, the resulting resolution far surpasses that of visible-light telescopes.

The United States operates the Very Long Baseline Array (VLBA), made up of 10 individual telescopes stretching from the Virgin Islands to Hawaii (**Figure 34.21**). The VLBA, completed in 1993, can form astronomical images with a resolution of 0.0001 arcseconds, permitting features as small as 10 astronomical units (AU) to be distinguished at the center of our Galaxy.

Figure 34.21 This map shows the distribution of 10 antennas that constitute an array of radio telescopes stretching across the United States and its territories.

Recent advances in technology have also made it possible to do interferometry at visible-light and infrared wavelengths. At the beginning of the twenty-first century, three observatories with multiple telescopes each began using their dishes as interferometers, combining their light to obtain a much greater resolution. In addition, a dedicated interferometric array was built on Mt. Wilson in California. Just as in radio arrays, these observations allow astronomers to make out more detail than a single telescope could provide.

Visible-Light Interferometers

Longest Baseline (m)	Telescope Name	Location	Mirrors	Status
400	CHARA Array (Center for High Angular Resolution Astronomy)	Mount Wilson, CA	Six 1-m telescopes	Operational since 2004
200	Very Large Telescope	Cerro Paranal, Chile	Four 8.2-m telescopes	Completed 2000
85	Keck I and II telescopes	Mauna Kea, HI	Two 10-m telescopes	Operated from 2001 to 2012
22.8	Large Binocular Telescope	Mount Graham, AZ	Two 8.4-m telescopes	First light 2004

Table 34.3

Radar Astronomy

Radar is the technique of transmitting radio waves to an object in our solar system and then detecting the radio radiation that the object reflects back. The time required for the round trip can be measured electronically with great precision.

Because we know the speed at which radio waves travel (the speed of light), we can determine the distance to the object or a particular feature on its surface (such as a mountain).

Radar observations have been used to determine the distances to planets and how fast things are moving in the solar system (using the **Doppler effect.**). Radar waves have played important roles in navigating spacecraft throughout the solar system. In addition, as will be discussed in later chapters, radar observations have determined the rotation periods of Venus and Mercury, probed tiny Earth-approaching asteroids, and allowed us to investigate the mountains and valleys on the surfaces of Mercury, Venus, Mars, and the large moons of Jupiter.

Any radio dish can be used as a radar telescope if it is equipped with a powerful transmitter as well as a receiver. The most spectacular facility in the world for radar astronomy is the 1000-foot (305-meter) telescope at Arecibo in Puerto Rico (**Figure 34.22**). The Arecibo telescope is too large to be pointed directly at different parts of the sky. Instead, it is constructed in a huge natural "bowl" (more than a mere dish) formed by several hills, and it is lined with reflecting metal panels. A limited ability to track astronomical sources is achieved by moving the receiver system, which is suspended on cables 100 meters above the surface of the bowl. An even larger (500-meter) radar telescope is currently under construction. It is the Five-hundred-meter Aperture Spherical Telescope (FAST) in China and is expected to be completed in 2016.

Figure 34.22 The Arecibo Observatory, with its 1000-foot radio dish-filling valley in Puerto Rico, is part of the National Astronomy and Ionosphere Center, operated by SRI International, USRA, and UMET under a cooperative agreement with the National Science Foundation. (credit: National Astronomy and Ionosphere Center, Cornell U., NSF)

34.5 | Observations Outside Earth's Atmosphere

Learning Objectives

After completing this section, you will be able to:

- List the advantages of making astronomical observations from space
- Explain the importance of the Hubble Space Telescope
- Describe some of the major space-based observatories astronomers use

Earth's atmosphere blocks most radiation at wavelengths shorter than visible light, so we can only make direct ultraviolet, X-ray, and gamma ray observations from space (though indirect gamma ray observations can be made from Earth). Getting above the distorting effects of the atmosphere is also an advantage at visible and infrared wavelengths. The stars don't "twinkle" in space, so the amount of detail you can observe is limited only by the size of your instrument. On the other hand, it is expensive to place telescopes into space, and repairs can present a major challenge. This is why astronomers continue to build telescopes for use on the ground as well as for launching into space.

Airborne and Space Infrared Telescopes

Water vapor, the main source of atmospheric interference for making infrared observations, is concentrated in the lower part of Earth's atmosphere. For this reason, a gain of even a few hundred meters in elevation can make an important difference in the quality of an infared observatory site. Given the limitations of high mountains, most of which attract clouds and violent storms, and the fact that the ability of humans to perform complex tasks degrades at high altitudes, it was natural for astronomers to investigate the possibility of observing infrared waves from airplanes and ultimately from space.

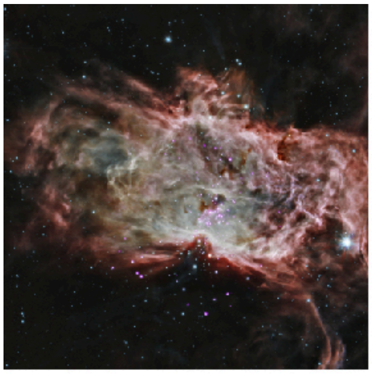
Infrared observations from airplanes have been made since the 1960s, starting with a 15-centimeter telescope on board a Learjet. From 1974 through 1995, NASA operated a 0.9-meter airborne telescope flying regularly out of the Ames Research Center south of San Francisco. Observing from an altitude of 12 kilometers, the telescope was above 99% of the atmospheric water vapor. More recently, NASA (in partnership with the German Aerospace Center) has constructed a much larger 2.5-meter telescope, called the Stratospheric Observatory for Infrared Astronomy (SOFIA), which flies in a modified Boeing 747SP (**Figure 34.23**).

Figure 34.23 SOFIA allows observations to be made above most of Earth's atmospheric water vapor. (credit: NASA)

To find out more about SOFIA, watch this **video (https://openstaxcollege.org/l/30SOFIAvid)** provided by NASA's Armstrong Flight Research Center.

Getting even higher and making observations from space itself have important advantages for infrared astronomy. First is the elimination of all interference from the atmosphere. Equally important is the opportunity to cool the entire optical system of the instrument in order to nearly eliminate infrared radiation from the telescope itself. If we tried to cool a telescope within the atmosphere, it would quickly become coated with condensing water vapor and other gases, making it useless. Only in the vacuum of space can optical elements be cooled to hundreds of degrees below freezing and still remain operational.

The first orbiting infrared observatory, launched in 1983, was the Infrared Astronomical Satellite (IRAS), built as a joint project by the United States, the Netherlands, and Britain. IRAS was equipped with a 0.6-meter telescope cooled to a temperature of less than 10 K. For the first time, the infrared sky could be seen as if it were night, rather than through a bright foreground of atmospheric and telescope emissions. IRAS carried out a rapid but comprehensive survey of the entire infrared sky over a 10-month period, cataloging about 350,000 sources of infrared radiation. Since then, several other infrared telescopes have operated in space with much better sensitivity and resolution due to improvements in infrared detectors. The most powerful of these infrared telescopes is the 0.85-meter Spitzer Space Telescope, which launched in 2003. A few of its observations are shown in **Figure 34.24**. With infrared observations, astronomers can detect cooler parts of cosmic objects, such as the dust clouds around star nurseries and the remnants of dying stars, that visible-light images don't reveal.



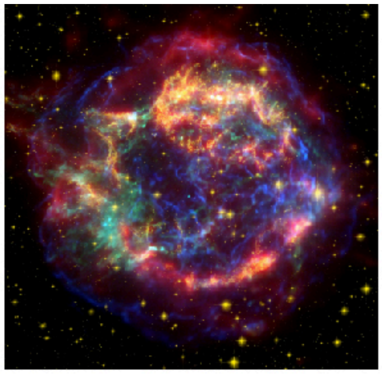


Figure 34.24 These infrared images—a region of star formation, the remnant of an exploded star, and a region where an old star is losing its outer shell—show just a few of the observations made and transmitted back to Earth from the SST. Since our eyes are not sensitive to infrared rays, we don't perceive colors from them. The colors in these images have been selected by astronomers to highlight details like the composition or temperature in these regions. (credit "Flame nebula": modification of work by NASA (X-ray: NASA/CXC/PSU/K.Getman, E.Feigelson, M.Kuhn & the MYStIX team; Infrared:NASA/JPL-Caltech); credit "Cassiopeia A": modification of work by NASA/JPL-Caltech; credit "Helix nebula": modification of work by NASA/JPL-Caltech)

Hubble Space Telescope

In April 1990, a great leap forward in astronomy was made with the launch of the Hubble Space Telescope (HST). With an aperture of 2.4 meters, this is the largest telescope put into space so far. (Its aperture was limited by the size of the payload bay in the Space Shuttle that served as its launch vehicle.) It was named for Edwin Hubble, the astronomer who discovered the expansion of the universe in the 1920s (whose work we will discuss in the chapters on **Galaxies**).

HST is operated jointly by NASA's Goddard Space Flight Center and the Space Telescope Science Institute in Baltimore. It was the first orbiting observatory designed to be serviced by Shuttle astronauts and, over the years since it was launched, they made several visits to improve or replace its initial instruments and to repair some of the systems that operate the spacecraft (**Figure 34.1**)—though this repair program has now been discontinued, and no more visits or improvements will be made.

With the Hubble, astronomers have obtained some of the most detailed images of astronomical objects from the solar system outward to the most distant galaxies. Among its many great achievements is the Hubble Ultra-Deep Field, an image of a small region of the sky observed for almost 100 hours. It contains views of about 10,000 galaxies, some of which formed when the universe was just a few percent of its current age (**Figure 34.25**).

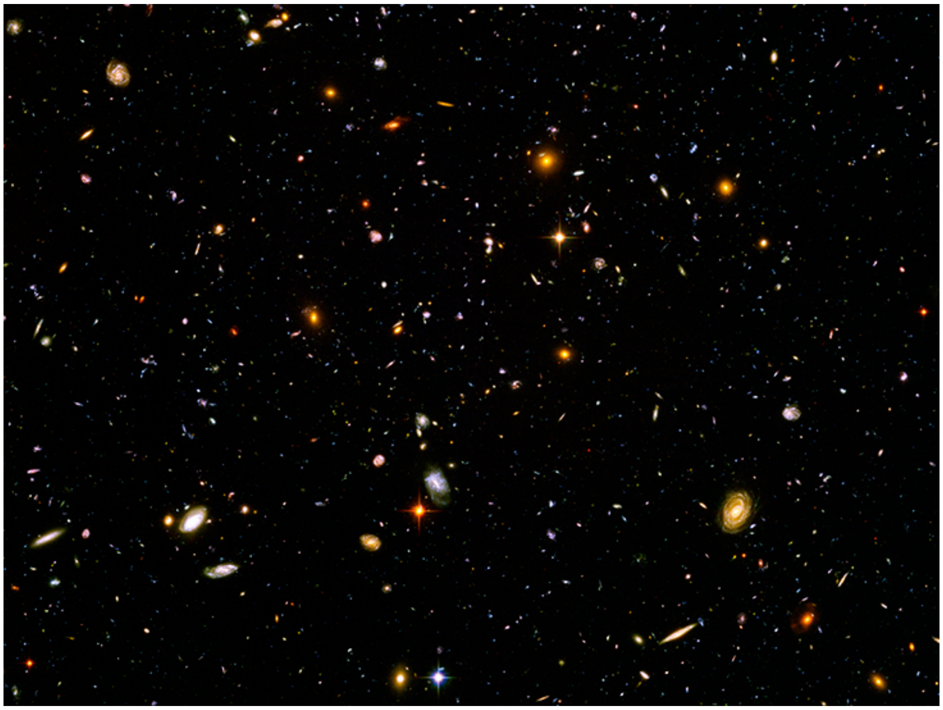


Figure 34.25 The Hubble Space Telescope has provided an image of a specific region of space built from data collected between September 24, 2003, and January 16, 2004. These data allow us to search for galaxies that existed approximately 13 billion years ago. (credit: modification of work by NASA)

The HST's mirror was ground and polished to a remarkable degree of accuracy. If we were to scale up its 2.4-meter mirror to the size of the entire continental United States, there would be no hill or valley larger than about 6 centimeters in its smooth surface. Unfortunately, after it was launched, scientists discovered that the primary mirror had a slight error in its *shape*, equal to roughly 1/50 the width of a human hair. Small as that sounds, it was enough to ensure that much of the light entering the telescope did not come to a clear focus and that all the images were blurry. (In a misplaced effort to save money, a complete test of the optical system had not been carried out before launch, so the error was not discovered until HST was in orbit.)

The solution was to do something very similar to what we do for astronomy students with blurry vision: put corrective optics in front of their eyes. In December 1993, in one of the most exciting and difficult space missions ever flown, astronauts captured the orbiting telescope and brought it back into the shuttle payload bay. There they installed a package containing compensating optics as well as a new, improved camera before releasing HST back into orbit. The telescope now works as it was intended to, and further missions to it were able to install even more advanced instruments to take advantage of its capabilities.

High-Energy Observatories

Ultraviolet, X-ray, and direct gamma-ray (high-energy electromagnetic wave) observations can be made only from space. Such observations first became possible in 1946, with V2 rockets captured from Germany after World War II. The US Naval Research Laboratory put instruments on these rockets for a series of pioneering flights, used initially to detect ultraviolet radiation from the Sun. Since then, many other rockets have been launched to make X-ray and ultraviolet observations of the Sun, and later of other celestial objects.

Beginning in the 1960s, a steady stream of high-energy observatories has been launched into orbit to reveal and explore the universe at short wavelengths. Among recent X-ray telescopes is the Chandra X-ray Observatory, which was launched in 1999 (**Figure 34.26**). It is producing X-ray images with unprecedented resolution and sensitivity. Designing instruments that can collect and focus energetic radiation like X-rays and gamma rays is an enormous technological challenge. The 2002 Nobel Prize in physics was awarded to Riccardo Giacconi, a pioneer in the field of building and launching sophisticated X-ray instruments. In 2008, NASA launched the Fermi Gamma-ray Space Telescope, designed to measure cosmic gamma rays at energies greater than any previous telescope, and thus able to collect radiation from some of the most energetic events in the universe.

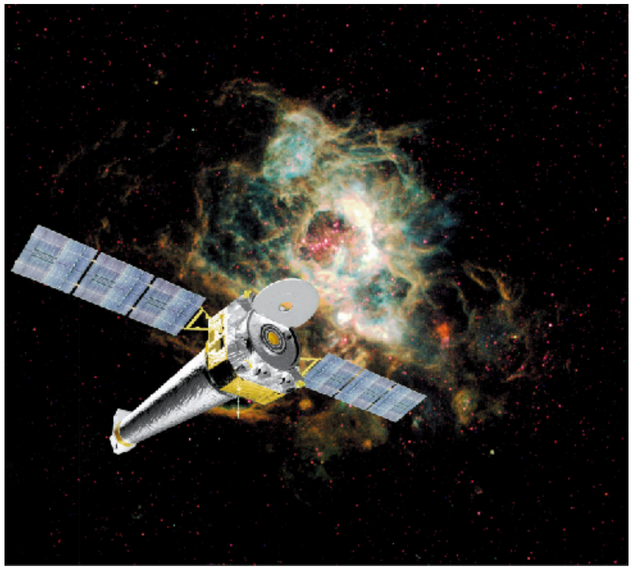


Figure 34.26 Chandra, the world's most powerful X-ray telescope, was developed by NASA and launched in July 1999. (credit: modification of work by NASA)

One major challenge is to design "mirrors" to reflect such penetrating radiation as X-rays and gamma rays, which normally pass straight through matter. However, although the technical details of design are more complicated, the three basic components of an observing system, as we explained earlier in this chapter, are the same at all wavelengths: a telescope to gather up the radiation, filters or instruments to sort the radiation according to wavelength, and some method of detecting and making a permanent record of the observations. **Table 34.4** lists some of the most important active space observatories that humanity has launched.

Gamma-ray detections can also be made from Earth's surface by using the atmosphere as the primary detector. When a gamma ray hits our atmosphere, it accelerates charged particles (mostly electrons) in the atmosphere. Those energetic particles hit other particles in the atmosphere and give off their own radiation. The effect is a cascade of light and energy that can be detected on the ground. The VERITAS array in Arizona and the H.E.S.S. array in Namibia are two such ground-based gamma-ray observatories.

Recent Observatories in Space

Observatory	Date Operation Began	Bands of the Spectrum	Notes	Website
Hubble Space Telescope (HST)	1990	visible, UV, IR	2.4-m mirror; images and spectra	www.hubblesite.org
Chandra X-Ray Observatory	1999	X-rays	X-ray images and spectra	www.chandra.si.edu
XMM-Newton	1999	X-rays	X-ray spectroscopy	http://www.cosmos.esa.int/web/xmm-newton
International Gamma-Ray Astrophysics Laboratory (INTEGRAL)	2002	X- and gamma-rays	higher resolution gamma-ray images	http://sci.esa.int/integral/
Spitzer Space Telescope	2003	IR	0.85-m telescope	www.spitzer.caltech.edu

Table 34.4

Recent Observatories in Space

Observatory	Date Operation Began	Bands of the Spectrum	Notes	Website
Fermi Gamma-ray Space Telescope	2008	gamma-rays	first high-energy gamma-ray observations	fermi.gsfc.nasa.gov
Kepler	2009	visible-light	planet finder	http://kepler.nasa.gov
Wide-field Infrared Survey Explorer (WISE)	2009	IR	whole-sky map, asteroid searches	www.nasa.gov/mission_pages/WISE/main
Gaia	2013	visible-light	Precise map of the Milky Way	http://sci.esa.int/gaia/
Transiting Exoplanet Survey Satellite (TESS)	2018	visible-light	Planet finder	http://tess.mit.edu

Table 34.4

34.6 | The Future of Large Telescopes

Learning Objectives

After completing this section, you will be able to:

- Describe the next generation of ground- and space-based observatories
- Explain some of the challenges involved in building these observatories

If you've ever gone on a hike, you have probably been eager to see what lies just around the next bend in the path. Researchers are no different, and astronomers and engineers are working on the technologies that will allow us to explore even more distant parts of the universe and to see them more clearly.

The premier space facility planned for the next decade is the James Webb Space Telescope (**Figure 34.27**), which (in a departure from tradition) is named after one of the early administrators of NASA instead of a scientist. This telescope will have a mirror 6 meters in diameter, made up, like the Keck telescopes, of 36 small hexagons. These will have to unfold into place once the telescope reaches its stable orbit point, some 1.5 million kilometers from Earth (where no astronauts can currently travel if it needs repair.) The telescope is scheduled for launch in 2021 and should have the sensitivity needed to detect the very first generation of stars, formed when the universe was only a few hundred million years old. With the ability to measure both visible and infrared wavelengths, it will serve as the successor to both HST and the Spitzer Space Telescope.

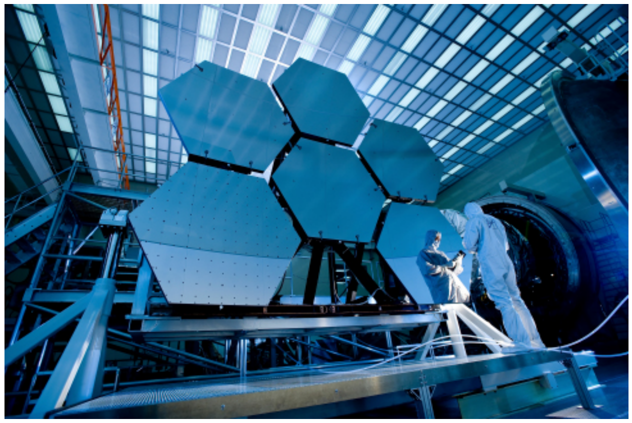


Figure 34.27 This image shows some of the mirrors of the JWST as they underwent cryogenic testing. The mirrors were exposed to extreme temperatures in order to gather accurate measurements on changes in their shape as they heated and cooled. (credit: NASA/MSFC/David Higginbotham/Emmett Given)

Watch this **video (https://openstaxcollege.org/l/30JWSTvid)** to learn more about the James Webb Space Telescope and how it will build upon the work that Hubble has allowed us to begin in exploring the universe.

On the ground, astronomers have started building the Large Synoptic Survey Telescope (LSST), an 8.4-meter telescope with a significantly larger field of view than any existing telescopes. It will rapidly scan the sky to find *transients*, phenomena that change quickly, such as exploding stars and chunks of rock that orbit near Earth. The LSST is expected to see first light in 2021.

The international gamma-ray community is planning the Cherenkov Telescope Array (CTA), two arrays of telescopes, one in each hemisphere, which will indirectly measure gamma rays from the ground. The CTA will measure gamma-ray energies a thousand times as great as the Fermi telescope can detect.

Several groups of astronomers around the globe interested in studying visible light and infrared are exploring the feasibility of building ground-based telescopes with mirrors larger than 30 meters across. Stop and think what this means: 30 meters is one-third the length of a football field. It is technically impossible to build and transport a single astronomical mirror that is 30 meters or larger in diameter. The primary mirror of these giant telescopes will consist of smaller mirrors, all aligned so that they act as a very large mirror in combination. These include the Thirty-Meter Telescope for which construction has begun at the top of Mauna Kea in Hawaii.

The most ambitious of these projects is the European Extremely Large Telescope (E-ELT) (**Figure 34.28**). (Astronomers try to outdo each other not only with the size of these telescopes, but also their names!) The design of the E-ELT calls for a 39.3-meter primary mirror, which will follow the Keck design and be made up of 798 hexagonal mirrors, each 1.4 meters in diameter and all held precisely in position so that they form a continuous surface.

Construction on the site in the Atacama Desert in Northern Chile started in 2014. The E-ELT, along with the Thirty Meter Telescope and the Giant Magellan Telescope, which are being built by international consortia led by US astronomers, will combine light-gathering power with high-resolution imaging. These powerful new instruments will enable astronomers to tackle many important astronomical problems. For example, they should be able to tell us when, where, and how often planets form around other stars. They should even be able to provide us images and spectra of such planets and thus, perhaps, give us the first real evidence (from the chemistry of these planets' atmospheres) that life exists elsewhere.

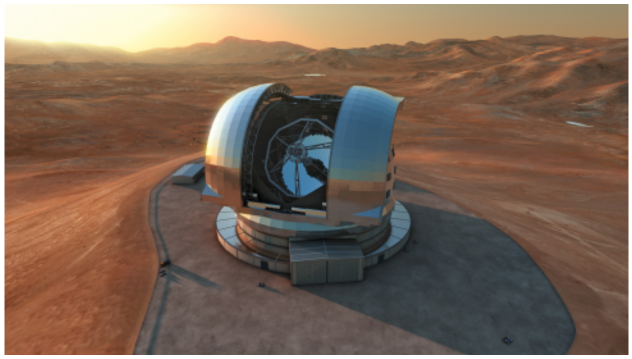


Figure 34.28 The primary mirror in this telescope is 39.3 meters across. The telescope is under construction in the Atacama Desert in Northern Chile. (credit: ESO/L. Calçada)

Check out this **fun diagram (https://openstaxcollege.org/l/30JWSTdiag)** comparing the sizes of the largest planned and existing telescopes to a regulation basketball and tennis court.

Summary

- New and even larger telescopes are on the drawing boards.
- The James Webb Space Telescope, a 6-meter successor to Hubble, is currently scheduled for launch in 2021.
- Gamma-ray astronomers are planning to build the CTA to measure very energetic gamma rays.
- Astronomers are building the LSST to observe with an unprecedented field of view and a new generation of visible-light/infrared telescopes with apertures of 24.5 to 39 meters in diameter.

For Further Exploration

Articles

Blades, J. C. “Fixing the Hubble One Last Time.” *Sky & Telescope* (October 2008): 26. On the last Shuttle service mission and what the Hubble was then capable of doing.

Brown, A. “How Gaia will Map a Billion Stars.” *Astronomy* (December 2014): 32. Nice review of the mission to do photometry and spectroscopy of all stars above a certain brightness.

Irion, R. “Prime Time.” *Astronomy* (February 2001): 46. On how time is allotted on the major research telescopes.

Jedicke, Peter & Robert. “The Coming Giant Sky Patrols.” *Sky & Telescope* (September 2008): 30. About giant telescopes to survey the sky continuously.

Lazio, Joseph, et al. “Tuning in to the Universe: 21st Century Radio Astronomy.” *Sky & Telescope* (July 2008): 21. About ALMA and the Square Kilometer Array.

Lowe, Jonathan. “Mirror, Mirror.” *Sky & Telescope* (December 2007): 22. On the Large Binocular Telescope in Arizona.

Lowe, Jonathan. “Next Light: Tomorrow’s Monster Telescopes.” *Sky & Telescope* (April 2008): 20. About plans for extremely large telescopes on the ground.

Mason, Todd & Robin. “Palomar’s Big Eye.” *Sky & Telescope* (December 2008): 36. On the Hale 200-inch telescope.

Subinsky, Raymond. “Who Really Invented the Telescope.” *Astronomy* (August 2008): 84. Brief historical introduction, focusing on Hans Lippershey.

Websites

Websites for major telescopes are given in **Table 34.1**, **Table 34.2**, **Table 34.3**, and **Table 34.4**.

Videos

Astronomy from the Stratosphere: SOFIA: **https://www.youtube.com/watch?v=NV98BcBBA9c (https://www.youtube.com/watch?v=NV98BcBBA9c)** . A talk by Dr. Dana Backman (1:15:32)

Galaxies Viewed in Full Spectrum of Light: **https://www.youtube.com/watch?v=368K0iQv8nE**

(https://www.youtube.com/watch?v=368K0iQv8nE) . Scientists with the Spitzer Observatory show how a galaxy looks different at different wavelengths (6:22)

Lifting the Cosmic Veil: Highlights from a Decade of the Spitzer Space Telescope: **https://www.youtube.com/watch?v=nkrNQcwkY78 (https://www.youtube.com/watch?v=nkrNQcwkY78)** . A talk by Dr. Michael Bicay (1:42:44)

CHAPTER 34 REVIEW

KEY TERMS

adaptive optics systems used with telescopes that can compensate for distortions in an image introduced by the atmosphere, thus resulting in sharper images

aperture diameter of the primary lens or mirror of a telescope

charge-coupled device (CCD) array of high-sensitivity electronic detectors of electromagnetic radiation, used at the focus of a telescope (or camera lens) to record an image or spectrum

chromatic aberration distortion that causes an image to appear fuzzy when each wavelength coming into a transparent material focuses at a different spot

detector device sensitive to electromagnetic radiation that makes a record of astronomical observations

eyepiece magnifying lens used to view the image produced by the objective lens or primary mirror of a telescope

focus (of telescope) point where the rays of light converged by a mirror or lens meet

interference process in which waves mix together such that their crests and troughs can alternately reinforce and cancel one another

interferometer instrument that combines electromagnetic radiation from one or more telescopes to obtain a resolution equivalent to what would be obtained with a single telescope with a diameter equal to the baseline separating the individual separate telescopes

interferometer array combination of multiple radio dishes to, in effect, work like a large number of two-dish interferometers

prime focus point in a telescope where the objective lens or primary mirror focuses the light

radar technique of transmitting radio waves to an object and then detecting the radiation that the object reflects back to the transmitter; used to measure the distance to, and motion of, a target object or to form images of it

reflecting telescope telescope in which the principal light collector is a concave mirror

refracting telescope telescope in which the principal light collector is a lens or system of lenses

resolution detail in an image; specifically, the smallest angular (or linear) features that can be distinguished

seeing unsteadiness of Earth’s atmosphere, which blurs telescopic images; good seeing means the atmosphere is steady

telescope instrument for collecting visible-light or other electromagnetic radiation

SUMMARY

34.1 Telescopes

- A telescope collects the faint light from astronomical sources and brings it to a focus, where an instrument can sort the light according to wavelength. Light is then directed to a detector, where a permanent record is made.
- The light-gathering power of a telescope is determined by the diameter of its aperture, or opening—that is, by the area of its largest or primary lens or mirror.
- The primary optical element in a telescope is either a convex lens (in a refracting telescope) or a concave mirror (in a reflector) that brings the light to a focus.
- Most large telescopes are reflectors; it is easier to manufacture and support large mirrors because the light does not have to pass through glass.

34.2 Telescopes Today

- New technologies for creating and supporting lightweight mirrors have led to the construction of a number of large telescopes since 1990.

- The site for an astronomical observatory must be carefully chosen for clear weather, dark skies, low water vapor, and excellent atmospheric seeing (low atmospheric turbulence).
- The resolution of a visible-light or infrared telescope is degraded by turbulence in Earth's atmosphere. The technique of adaptive optics, however, can make corrections for this turbulence in real time and produce exquisitely detailed images.

34.3 Visible-Light Detectors and Instruments

- Visible-light detectors include the human eye, photographic film, and charge-coupled devices (CCDs).
- Detectors that are sensitive to infrared radiation must be cooled to very low temperatures since everything in and near the telescope gives off infrared waves.
- A spectrometer disperses the light into a spectrum to be recorded for detailed analysis.

34.4 Radio Telescopes

- In the 1930s, radio astronomy was pioneered by Karl G. Jansky and Grote Reber.
- A radio telescope is basically a radio antenna (often a large, curved dish) connected to a receiver.
- Significantly enhanced resolution can be obtained with interferometers, including interferometer arrays like the 27-element VLA and the 66-element ALMA.
- Expanding to very long baseline interferometers, radio astronomers can achieve resolutions as precise as 0.0001 arcsecond.
- Radar astronomy involves transmitting as well as receiving.
- The largest radar telescope currently in operation is a 305-meter bowl at Arecibo.

34.5 Observations Outside Earth's Atmosphere

- Infrared observations are made with telescopes aboard aircraft and in space, as well as from ground-based facilities on dry mountain peaks.
- Ultraviolet, X-ray, and gamma-ray observations must be made from above the atmosphere.
- Many orbiting observatories have been flown to observe in these bands of the spectrum in the last few decades.
- The largest-aperture telescope in space is the Hubble Space telescope (HST), the most significant infrared telescope is Spitzer, and Chandra and Fermi are the premier X-ray and gamma-ray observatories, respectively.

CONCEPTUAL QUESTIONS

34.1 Telescopes

1. What are the three basic components of a modern astronomical instrument? Describe each in one to two sentences.

2. The Hooker telescope at Palomar Observatory has a diameter of 5 m, and the Keck I telescope has a diameter of 10 m. How much more light can the Keck telescope collect than the Hooker telescope in the same amount of time?

3. What is meant by "reflecting" and "refracting" telescopes?

4. Why are the largest visible-light telescopes in the world made with mirrors rather than lenses?

5. What happens to the image produced by a lens if the lens is "stopped down" (the aperture reduced, thereby reducing the amount of light passing through the lens) with an iris diaphragm—a device that covers its periphery?

6. The dean of a university located near the ocean (who was not a science major in college) proposes building an infrared telescope right on campus and operating it in a nice heated dome so that astronomers will be comfortable on cold winter nights. Criticize this proposal, giving your reasoning.

34.2 Telescopes Today

7. Name the two spectral windows through which electromagnetic radiation easily reaches the surface of Earth and describe the largest-aperture telescope currently in use for each window.

8. List the largest-aperture single telescope currently in use in each of the following bands of the electromagnetic spectrum: radio, X-ray, gamma ray.

9. When astronomers discuss the apertures of their telescopes, they say bigger is better. Explain why.

10. Many decades ago, the astronomers on the staff of Mount Wilson and Palomar Observatories each received about 60 nights per year for their observing programs. Today, an astronomer feels fortunate to get 10 nights per year on a large telescope. Can you suggest some reasons for this change?

34.3 Visible-Light Detectors and Instruments

11. Compare the eye, photographic film, and CCDs as detectors for light. What are the advantages and disadvantages of each?

12. What is a charge-coupled device (CCD), and how is it used in astronomy?

13. Why is it difficult to observe at infrared wavelengths? What do astronomers do to address this difficulty?

14. What would be the properties of an ideal astronomical detector? How closely do the actual properties of a CCD approach this ideal?

34.4 Radio Telescopes

15. Radio and radar observations are often made with the same antenna, but otherwise they are very different techniques. Compare and contrast radio and radar astronomy in terms of the equipment needed, the methods used, and the kind of results obtained.

16. Look back at **Figure 34.18** of Cygnus A and read its caption again. The material in the giant lobes at the edges of the image had to have been ejected from the center *at least* how many years ago?

34.5 Observations Outside Earth's Atmosphere

17. Why do astronomers place telescopes in Earth's orbit? What are the advantages for the different regions of the spectrum?

18. What was the problem with the Hubble Space Telescope and how was it solved?

34.6 The Future of Large Telescopes

19. What kind of visible-light and infrared telescopes on the ground are astronomers planning for the future? Why are they building them on the ground and not in space?

20. Describe one visible-light or infrared telescope that astronomers are planning to launch into space in the future.

21. Suppose you are looking for sites for a visible-light observatory, an infrared observatory, and a radio observatory. What are the main criteria of excellence for each? What sites are actually considered the best today?

22. Radio astronomy involves wavelengths much longer than those of visible light, and many orbiting observatories have probed the universe for radiation of very short wavelengths. What sorts of objects and physical conditions would you expect to be associated with emission of radiation at very long and very short wavelengths?

PROBLEMS

34.1 Telescopes

23. What is the area, in square meters, of a 10-m telescope?

24. Approximately 9000 stars are visible to the naked eye in the whole sky (imagine that you could see around the entire globe and both the northern and southern hemispheres), and there are about 41,200 square degrees on the sky. How many stars are visible per square degree? Per square arcsecond?

25. In broad daylight, the size of your pupil is typically 3 mm. In dark situations, it expands to about 7 mm. How much more light can it gather?

26. How much more light can be gathered by a telescope that is 8 m in diameter than by your fully dark-adapted eye at 7 mm?

27. How much more light can the Keck telescope (with its 10-m diameter mirror) gather than an amateur telescope whose mirror is 25 cm (0.25 m) across?

28. People are often bothered when they discover that reflecting telescopes have a second mirror in the middle to bring the light out to an accessible focus where big instruments can be mounted. "Don't you lose light?" people ask. Well, yes, you do, but there is no better alternative. You can estimate how much light is lost by such an arrangement. The primary mirror (the one at the bottom in **Figure 34.6**) of the Gemini North telescope is 8 m in diameter. The secondary mirror at the top is about 1 m in diameter. Use the formula for the area of a circle to estimate what fraction of the light is blocked by the secondary mirror.

29. Telescopes can now be operated remotely from a warm room, but until about 25 years ago, astronomers worked at the telescope to guide it so that it remained pointed in exactly the right place. In a large telescope, like the Palomar 200-inch telescope, astronomers sat in a cage at the top of the telescope, where the secondary mirror is located, as shown in **Figure 34.6**. Assume for the purpose of your calculation that the diameter of this cage was 40 inches. What fraction of the light is blocked?

30. The Palomar telescope's 5-m mirror weighs 14.5 tons. If a 10-m mirror were constructed of the same thickness as Palomar's (only bigger), how much would it weigh?

34.2 Telescopes Today

31. The HST cost about $1.7 billion for construction and $300 million for its shuttle launch, and it costs $250 million per year to operate. If the telescope lasts for 20 years, what is the total cost per year? Per day? If the telescope can be used just 30% of the time for actual observations, what is the cost per hour and per minute for the astronomer's observing time on this instrument? What is the cost per person in the United States? Was your investment in the Hubble Space telescope worth it?

32. How much more light can the James Webb Space Telescope (with its 6-m diameter mirror) gather than the Hubble Space Telescope (with a diameter of 2.4 m)?

34.4 Radio Telescopes

33. Theoretically (that is, if seeing were not an issue), the resolution of a telescope is inversely proportional to its diameter. How much better is the resolution of the ALMA when operating at its longest baseline than the resolution of the Arecibo telescope?

APPENDIX A | UNITS

Quantity	Common Symbol	Unit	Unit in Terms of Base SI Units
Acceleration	$\vec{\mathbf{a}}$	m/s^2	m/s^2
Amount of substance	*n*	**mole**	mol
Angle	θ, ϕ	radian (rad)	
Angular acceleration	$\vec{\alpha}$	rad/s^2	s^{-2}
Angular frequency	ω	rad/s	s^{-1}
Angular momentum	$\vec{\mathbf{L}}$	$\text{kg} \cdot \text{m}^2/\text{s}$	$\text{kg} \cdot \text{m}^2/\text{s}$
Angular velocity	$\vec{\omega}$	rad/s	s^{-1}
Area	*A*	m^2	m^2
Atomic number	*Z*		
Capacitance	*C*	farad (F)	$\text{A}^2 \cdot \text{s}^4/\text{kg} \cdot \text{m}^2$
Charge	*q*, *Q*, *e*	coulomb (C)	$\text{A} \cdot \text{s}$
Charge density:			
Line	λ	C/m	$\text{A} \cdot \text{s/m}$
Surface	σ	C/m^2	$\text{A} \cdot \text{s/m}^2$
Volume	ρ	C/m^3	$\text{A} \cdot \text{s/m}^3$
Conductivity	σ	$1/\Omega \cdot \text{m}$	$\text{A}^2 \cdot \text{s}^3/\text{kg} \cdot \text{m}^3$
Current	*I*	**ampere**	A
Current density	$\vec{\mathbf{J}}$	A/m^2	A/m^2
Density	ρ	kg/m^3	kg/m^3
Dielectric constant	κ		
Electric dipole moment	$\vec{\mathbf{p}}$	$\text{C} \cdot \text{m}$	$\text{A} \cdot \text{s} \cdot \text{m}$
Electric field	$\vec{\mathbf{E}}$	N/C	$\text{kg} \cdot \text{m/A} \cdot \text{s}^3$
Electric flux	Φ	$\text{N} \cdot \text{m}^2/\text{C}$	$\text{kg} \cdot \text{m}^3/\text{A} \cdot \text{s}^3$
Electromotive force	ε	volt (V)	$\text{kg} \cdot \text{m}^2/\text{A} \cdot \text{s}^3$
Energy	*E*,*U*,*K*	joule (J)	$\text{kg} \cdot \text{m}^2/\text{s}^2$
Entropy	*S*	J/K	$\text{kg} \cdot \text{m}^2/\text{s}^2 \cdot \text{K}$

Table A1 Units Used in Physics (Fundamental units in bold)

Quantity	Common Symbol	Unit	Unit in Terms of Base SI Units
Force	$\vec{\mathbf{F}}$	newton (N)	$\text{kg} \cdot \text{m/s}^2$
Frequency	f	hertz (Hz)	s^{-1}
Heat	Q	joule (J)	$\text{kg} \cdot \text{m}^2/\text{s}^2$
Inductance	L	henry (H)	$\text{kg} \cdot \text{m}^2/\text{A}^2 \cdot \text{s}^2$
Length:	ℓ, L	**meter**	m
Displacement	$\Delta x, \Delta \vec{\mathbf{r}}$		
Distance	d, h		
Position	$x, y, z, \vec{\mathbf{r}}$		
Magnetic dipole moment	$\vec{\mu}$	$\text{N} \cdot \text{J/T}$	$\text{A} \cdot \text{m}^2$
Magnetic field	$\vec{\mathbf{B}}$	$\text{tesla(T)} = (\text{Wb/m}^2)$	$\text{kg/A} \cdot \text{s}^2$
Magnetic flux	Φ_m	weber (Wb)	$\text{kg} \cdot \text{m}^2/\text{A} \cdot \text{s}^2$
Mass	m, M	**kilogram**	kg
Molar specific heat	C	$\text{J/mol} \cdot \text{K}$	$\text{kg} \cdot \text{m}^2/\text{s}^2 \cdot \text{mol} \cdot \text{K}$
Moment of inertia	I	$\text{kg} \cdot \text{m}^2$	$\text{kg} \cdot \text{m}^2$
Momentum	$\vec{\mathbf{p}}$	$\text{kg} \cdot \text{m/s}$	$\text{kg} \cdot \text{m/s}$
Period	T	s	s
Permeability of free space	μ_0	$\text{N/A}^2 = (\text{H/m})$	$\text{kg} \cdot \text{m/A}^2 \cdot \text{s}^2$
Permittivity of free space	ε_0	$\text{C}^2/\text{N} \cdot \text{m}^2 = (\text{F/m})$	$\text{A}^2 \cdot \text{s}^4/\text{kg} \cdot \text{m}^3$
Potential	V	$\text{volt(V)} = (\text{J/C})$	$\text{kg} \cdot \text{m}^2/\text{A} \cdot \text{s}^3$
Power	P	$\text{watt(W)} = (\text{J/s})$	$\text{kg} \cdot \text{m}^2/\text{s}^3$
Pressure	p	$\text{pascal(Pa)} = (\text{N/m}^2)$	$\text{kg/m} \cdot \text{s}^2$
Resistance	R	$\text{ohm}(\Omega) = (\text{V/A})$	$\text{kg} \cdot \text{m}^2/\text{A}^2 \cdot \text{s}^3$
Specific heat	c	$\text{J/kg} \cdot \text{K}$	$\text{m}^2/\text{s}^2 \cdot \text{K}$
Speed	v	m/s	m/s
Temperature	T	**kelvin**	K
Time	t	**second**	s
Torque	$\vec{\tau}$	$\text{N} \cdot \text{m}$	$\text{kg} \cdot \text{m}^2/\text{s}^2$

Table A1 Units Used in Physics (Fundamental units in bold)

Quantity	Common Symbol	Unit	Unit in Terms of Base SI Units
Velocity	$\vec{\mathbf{v}}$	m/s	m/s
Volume	*V*	m^3	m^3
Wavelength	λ	m	m
Work	*W*	$\text{joule(J)} = (\text{N} \cdot \text{m})$	$\text{kg} \cdot \text{m}^2/\text{s}^2$

Table A1 Units Used in Physics (Fundamental units in bold)

APPENDIX B | UNIT CONVERSIONS

	m	cm	km
1 meter	1	10^2	10^{-3}
1 centimeter	10^{-2}	1	10^{-5}
1 kilometer	10^3	10^5	1
1 inch	2.540×10^{-2}	2.540	2.540×10^{-5}
1 foot	0.3048	30.48	3.048×10^{-4}
1 mile	1609	1.609×10^4	1.609
1 angstrom	10^{-10}		
1 fermi	10^{-15}		
1 light-year	9.461×10^{15}		9.461×10^{12}
1 parsec	3.084×10^{16}		3.084×10^{13}
	in.	**ft**	**mi**
1 meter	39.37	3.281	6.214×10^{-4}
1 centimeter	0.3937	3.281×10^{-2}	6.214×10^{-6}
1 kilometer	3.937×10^4	3.281×10^3	0.6214
1 inch	1	8.333×10^{-2}	1.578×10^{-5}
1 foot	12	1	1.894×10^{-4}
1 mile	6.336×10^4	5280	1

Table B1 Length

Area

$1\ \text{cm}^2 = 0.155\ \text{in.}^2$

$1\ \text{m}^2 = 10^4\ \text{cm}^2 = 10.76\ \text{ft}^2$

$1\ \text{in.}^2 = 6.452\ \text{cm}^2$

$1\ \text{ft}^2 = 144\ \text{in.}^2 = 0.0929\ \text{m}^2$

Volume

$1\ \text{liter} = 1000\ \text{cm}^3 = 10^{-3}\ \text{m}^3 = 0.03531\ \text{ft}^3 = 61.02\ \text{in.}^3$

$1\ \text{ft}^3 = 0.02832\ \text{m}^3 = 28.32\ \text{liters} = 7.477\ \text{gallons}$

$1\ \text{gallon} = 3.788\ \text{liters}$

	s	min	h	day	yr
1 second	1	1.667×10^{-2}	2.778×10^{-4}	1.157×10^{-5}	3.169×10^{-8}
1 minute	60	1	1.667×10^{-2}	6.944×10^{-4}	1.901×10^{-6}
1 hour	3600	60	1	4.167×10^{-2}	1.141×10^{-4}
1 day	8.640×10^{4}	1440	24	1	2.738×10^{-3}
1 year	3.156×10^{7}	5.259×10^{5}	8.766×10^{3}	365.25	1

Table B2 Time

	m/s	cm/s	ft/s	mi/h
1 meter/second	1	10^{2}	3.281	2.237
1 centimeter/second	10^{-2}	1	3.281×10^{-2}	2.237×10^{-2}
1 foot/second	0.3048	30.48	1	0.6818
1 mile/hour	0.4470	44.70	1.467	1

Table B3 Speed

Acceleration

$1 \text{ m/s}^2 = 100 \text{ cm/s}^2 = 3.281 \text{ ft/s}^2$

$1 \text{ cm/s}^2 = 0.01 \text{ m/s}^2 = 0.03281 \text{ ft/s}^2$

$1 \text{ ft/s}^2 = 0.3048 \text{ m/s}^2 = 30.48 \text{ cm/s}^2$

$1 \text{ mi/h} \cdot \text{s} = 1.467 \text{ ft/s}^2$

	kg	g	slug	u
1 kilogram	1	10^{3}	6.852×10^{-2}	6.024×10^{26}
1 gram	10^{-3}	1	6.852×10^{-5}	6.024×10^{23}
1 slug	14.59	1.459×10^{4}	1	8.789×10^{27}
1 atomic mass unit	1.661×10^{-27}	1.661×10^{-24}	1.138×10^{-28}	1
1 metric ton	1000			

Table B4 Mass

	N	dyne	lb
1 newton	1	10^{5}	0.2248
1 dyne	10^{-5}	1	2.248×10^{-6}
1 pound	4.448	4.448×10^{5}	1

Table B5 Force

	Pa	dyne/cm^2	atm	cmHg	lb/in.^{2}
1 pascal	1	10	9.869×10^{-6}	7.501×10^{-4}	1.450×10^{-4}
1 dyne/ centimeter^2	10^{-1}	1	9.869×10^{-7}	7.501×10^{-5}	1.450×10^{-5}
1 atmosphere	1.013×10^{5}	1.013×10^{6}	1	76	14.70
1 centimeter mercury*	1.333×10^{3}	1.333×10^{4}	1.316×10^{-2}	1	0.1934
1 pound/inch^2	6.895×10^{3}	6.895×10^{4}	6.805×10^{-2}	5.171	1
1 bar	10^5				
1 torr				1 (mmHg)	

***Where the acceleration due to gravity is $9.80665\ m/s^2$ and the temperature is $0°C$**

Table B6 Pressure

	J	erg	ft.lb
1 joule	1	10^7	0.7376
1 erg	10^{-7}	1	7.376×10^{-8}
1 foot-pound	1.356	1.356×10^{7}	1
1 electron-volt	1.602×10^{-19}	1.602×10^{-12}	1.182×10^{-19}
1 calorie	4.186	4.186×10^{7}	3.088
1 British thermal unit	1.055×10^{3}	1.055×10^{10}	7.779×10^{2}
1 kilowatt-hour	3.600×10^{6}		
	eV	**cal**	**Btu**
1 joule	6.242×10^{18}	0.2389	9.481×10^{-4}
1 erg	6.242×10^{11}	2.389×10^{-8}	9.481×10^{-11}
1 foot-pound	8.464×10^{18}	0.3239	1.285×10^{-3}
1 electron-volt	1	3.827×10^{-20}	1.519×10^{-22}
1 calorie	2.613×10^{19}	1	3.968×10^{-3}
1 British thermal unit	6.585×10^{21}	2.520×10^{2}	1

Table B7 Work, Energy, Heat

Power

$1\ W = 1\ J/s$

$1\ hp = 746\ W = 550\ ft \cdot lb/s$

$1\ Btu/h = 0.293\ W$

Angle

$$1 \text{ rad} = 57.30° = 180°/\pi$$

$$1° = 0.01745 \text{ rad} = \pi/180 \text{ rad}$$

$$1 \text{ revolution} = 360° = 2\pi \text{ rad}$$

$$1 \text{ rev/min(rpm)} = 0.1047 \text{ rad/s}$$

APPENDIX C | FUNDAMENTAL PHYSICS CONSTANTS

Quantity	Symbol	Value
Atomic mass unit	u	$1.660\ 538\ 782\ (83) \times 10^{-27}$ kg $931.494\ 028\ (23)$ MeV/c^2
Avogadro's number	N_A	$6.022\ 141\ 79\ (30) \times 10^{23}$ particles/mol
Bohr magneton	$\mu_B = \frac{e\hbar}{2m_e}$	$9.274\ 009\ 15\ (23) \times 10^{-24}$ J/T
Bohr radius	$a_0 = \frac{\hbar^2}{m_e e^2 k_e}$	$5.291\ 772\ 085\ 9\ (36) \times 10^{-11}$ m
Boltzmann's constant	$k_B = \frac{R}{N_A}$	$1.380\ 650\ 4\ (24) \times 10^{-23}$ J/K
Compton wavelength	$\lambda_C = \frac{h}{m_e c}$	$2.426\ 310\ 217\ 5\ (33) \times 10^{-12}$ m
Coulomb constant	$k_e = \frac{1}{4\pi\varepsilon_0}$	$8.987\ 551\ 788... \times 10^9\ \text{N} \cdot \text{m}^2/\text{C}^2$ (exact)
Deuteron mass	m_d	$3.343\ 583\ 20\ (17) \times 10^{-27}$ kg $2.013\ 553\ 212\ 724(78)$ u $1875.612\ 859$ MeV/c^2
Electron mass	m_e	$9.109\ 382\ 15\ (45) \times 10^{-31}$ kg $5.485\ 799\ 094\ 3(23) \times 10^{-4}$ u $0.510\ 998\ 910\ (13)$ MeV/c^2
Electron volt	eV	$1.602\ 176\ 487\ (40) \times 10^{-19}$ J
Elementary charge	e	$1.602\ 176\ 487\ (40) \times 10^{-19}$ C
Gas constant	R	$8.314\ 472\ (15)$ J/mol · K
Gravitational constant	G	$6.674\ 28\ (67) \times 10^{-11}\ \text{N} \cdot \text{m}^2/\text{kg}^2$
Hydrogen atom mass	m_H	1.673×10^{-27} kg

Table C1 Fundamental Constants *Note:* These constants are the values recommended in 2006 by CODATA, based on a least-squares adjustment of data from different measurements. The numbers in parentheses for the values represent the uncertainties of the last two digits.

Quantity	Symbol	Value
Neutron mass	m_n	$1.674\,927\,211\,(84) \times 10^{-27}$ kg $1.008\,664\,915\,97\,(43)$ u $939.565\,346\,(23)\ \text{MeV}/c^2$
Nuclear magneton	$\mu_n = \frac{e\hbar}{2m_p}$	$5.050\,783\,24\,(13) \times 10^{-27}$ J/T
Permeability of free space	μ_0	$4\pi \times 10^{-7}\ \text{T} \cdot \text{m/A}$(exact)
Permittivity of free space	$\varepsilon_0 = \frac{1}{\mu_0 c^2}$	$8.854\,187\,817... \times 10^{-12}\ \text{C}^2/\text{N} \cdot \text{m}^2$ (exact)
Planck's constant	h $\hbar = \frac{h}{2\pi}$	$6.626\,068\,96\,(33) \times 10^{-34}\ \text{J} \cdot \text{s}$ $1.054\,571\,628\,(53) \times 10^{-34}\ \text{J} \cdot \text{s}$
Proton mass	m_p	$1.672\,621\,637\,(83) \times 10^{-27}$ kg $1.007\,276\,466\,77\,(10)$ u $938.272\,013\,(23)\ \text{MeV}/c^2$
Rydberg constant	R_H	$1.097\,373\,156\,852\,7\,(73) \times 10^{7}\ \text{m}^{-1}$
Speed of light in vacuum	c	$2.997\,924\,58 \times 10^{8}$ m/s (exact)
Stefan-Boltzmann constant	σ	$5.670 \times 10^{-8}\ \text{W/m}^2 \cdot \text{K}^4$
Wien's Law constant	$\lambda_{\text{max}}\text{T}$	2.898×10^{-3} m/K

Table C1 Fundamental Constants *Note:* These constants are the values recommended in 2006 by CODATA, based on a least-squares adjustment of data from different measurements. The numbers in parentheses for the values represent the uncertainties of the last two digits.

Useful combinations of constants for calculations:

$$hc = 12{,}400\ \text{eV} \cdot \text{Å} = 1240\ \text{eV} \cdot \text{nm} = 1240\ \text{MeV} \cdot \text{fm}$$

$$\hbar c = 1973\ \text{eV} \cdot \text{Å} = 197.3\ \text{eV} \cdot \text{nm} = 197.3\ \text{MeV} \cdot \text{fm}$$

$$k_e e^2 = 14.40\ \text{eV} \cdot \text{Å} = 1.440\ \text{eV} \cdot \text{nm} = 1.440\ \text{MeV} \cdot \text{fm}$$

$$k_\text{B} T = 0.02585\ \text{eV at } T = 300\ \text{K}$$

APPENDIX D | ASTRONOMICAL DATA

Astronomical Constants

Name	Value
astronomical unit (AU)	1.496×10^{11} m
Light-year (ly)	9.461×10^{15} m
parsec (pc)	3.086×10^{16} m = 3.262 light-years
sidereal year (y)	3.156×10^{7} s
mass of Earth (M_{Earth})	5.974×10^{24} kg
equatorial radius of Earth (R_{Earth})	6.378×10^{6} m
obliquity of ecliptic	23.4° 26’
surface gravity of Earth (g)	9.807 m/s^2
escape velocity of Earth (v_{Earth})	1.119×10^{4} m/s
mass of Sun (M_{Sun})	1.989×10^{30} kg
equatorial radius of Sun (R_{Sun})	6.960×10^{8} m
luminosity of Sun (L_{Sun})	3.85×10^{26} W
solar constant (flux of energy received at Earth) (S)	1.368×10^{3} W/m^2
Hubble constant (H_0)	approximately 20 km/s per million light-years, or approximately 70 km/s per megaparsec

Table D1

Celestial Object	Mean Distance from Sun (million km)	Period of Revolution (d = days) (y = years)	Period of Rotation at Equator	Eccentricity of Orbit
Sun	–	–	27 d	–
Mercury	57.9	88 d	59 d	0.206
Venus	108.2	224.7 d	243 d	0.007
Earth	149.6	365.26 d	23 h 56 min 4 s	0.017
Mars	227.9	687 d	24 h 37 min 23 s	0.093
Jupiter	778.4	11.9 y	9 h 50 min 30 s	0.048
Saturn	1426.7	29.5 6	10 h 14 min	0.054
Uranus	2871.0	84.0 y	17 h 14 min	0.047
Neptune	4498.3	164.8 y	16 h	0.009

Table D2 Comparative Planetary Data

Celestial Object	Mean Distance from Sun (million km)	Period of Revolution (d = days) (y = years)	Period of Rotation at Equator	Eccentricity of Orbit
Earth's Moon	149.6 (0.386 from Earth)	27.3 d	27.3 d	0.055

Celestial Object	Equatorial Diameter (km)	Mass (Earth = 1)	Density (g/cm^3)
Sun	1,392,000	333,000.00	1.4
Mercury	4879	0.06	5.4
Venus	12,104	0.82	5.2
Earth	12,756	1.00	5.5
Mars	6794	0.11	3.9
Jupiter	142,984	317.83	1.3
Saturn	120,536	95.16	0.7
Uranus	51,118	14.54	1.3
Neptune	49,528	17.15	1.6
Earth's Moon	3476	0.01	3.3

Table D2 Comparative Planetary Data

D1 Physical and Orbital Data for the Planets

Physical Data for the Major Planets

Major Planet	Mean Diameter (km)	Mean Diameter (Earth = 1)	Mass (Earth = 1)	Mean Density (g/cm^3)	Rotation Period (d)	Inclination of Equator to Orbit (°)	Surface Gravity (Earth = 1[g])	Velocity of Escape (km/s)
Mercury	4879	0.38	0.055	5.43	58.	0.0	0.38	4.3
Venus	12,104	0.95	0.815	5.24	−243.	177	0.90	10.4
Earth	12,756	1.00	1.00	5.51	1.000	23.4	1.00	11.2
Mars	6779	0.53	0.11	3.93	1.026	25.2	0.38	5.0
Jupiter	140,000	10.9	318	1.33	0.414	3.1	2.53	60.
Saturn	117,000	9.13	95.2	0.69	0.440	26.7	1.07	36.
Uranus	50,700	3.98	14.5	1.27	−0.718	97.9	0.89	21.
Neptune	49,200	3.86	17.2	1.64	0.671	29.6	1.14	23.

Table D3

Physical Data for Well-Studied Dwarf Planets

Well-Studied Dwarf Planet	Diameter (km)	Diameter (Earth = 1)	Mass (Earth = 1)	Mean Density (g/cm^3)	Rotation Period (d)	Inclination of Equator to Orbit (°)	Surface Gravity (Earth = 1[g])	Velocity of Escape (km/s)
Ceres	950	0.07	0.0002	2.2	0.378	3	0.03	0.5
Pluto	2470	0.18	0.0024	1.9	−6.387	122	0.06	1.3
Haumea	1700	0.13	0.0007	3	0.163	—	—	0.8
Makemake	1400	0.11	0.0005	2	0.321	—	—	0.8
Eris	2326	0.18	0.0028	2.5	1.25[1]	—	—	1.4

Table D4

Orbital Data for the Major Planets

Major Planet	Semimajor Axis (AU)	Semimajor Axis (10^6 km)	Sidereal Period (y)	Sidereal Period (d)	Mean Orbital Speed (km/s)	Orbital Eccentricity	Inclination of Orbit to Ecliptic (°)
Mercury	0.39	58	0.24	88.0	47.9	0.206	7.0
Venus	0.72	108	0.6	224.7	35.0	0.007	3.4
Earth	1.00	149	1.00	365.2	29.8	0.017	0.0
Mars	1.52	228	1.88	687.0	24.1	0.093	1.9
Jupiter	5.20	778	11.86	—	13.1	0.048	1.3
Saturn	9.54	1427	29.46	—	9.6	0.056	2.5
Uranus	19.19	2871	84.01	—	6.8	0.046	0.8
Neptune	30.06	4497	164.82	—	5.4	0.010	1.8

Table D5

Orbital Data for Well-Studied Dwarf Planets

Well-Studied Dwarf Planet	Semimajor Axis (AU)	Semimajor Axis (10^6 km)	Sidereal Period (y)	Mean Orbital Speed (km/s)	Orbital Eccentricity	Inclination of Orbit to Ecliptic (°)
Ceres	2.77	414.0	4.6	18	0.08	11
Pluto	39.5	5915	248.6	4.7	0.25	17
Haumea	43.1	6452	283.3	4.5	0.19	28
Makemake	45.8	6850	309.9	4.4	0.16	29
Eris	68.0	10,120	560.9	3.4	0.44	44

Table D6

D2 Selected Moons of the Planets

Note: As this book goes to press, nearly two hundred moons are now known in the solar system and more are being

1. This measurement is quite uncertain.

discovered on a regular basis. Of the major planets, only Mercury and Venus do not have moons. In addition to moons of the planets, there are many moons of asteroids. In this appendix, we list only the largest and most interesting objects that orbit each planet (including dwarf planets). The number given for each planet is discoveries through 2015. For further information see https://solarsystem.nasa.gov/planets/solarsystem/moons and https://en.wikipedia.org/wiki/List_of_natural_satellites.

Selected Moons of the Planets

Planet (moons)	Satellite Name	Discovery	Semimajor Axis (km × 1000)	Period (d)	Diameter (km)	Mass (10^{20} kg)	Density (g/cm^3)
Earth (1)	Moon	—	384	27.32	3476	735	3.3
Mars (2)	Phobos	Hall (1877)	9.4	0.32	23	1×10^{-4}	2.0
	Deimos	Hall (1877)	23.5	1.26	13	2×10^{-5}	1.7
Jupiter (67)	Amalthea	Barnard (1892)	181	0.50	200	—	—
	Thebe	Voyager (1979)	222	0.67	90	—	—
	Io	Galileo (1610)	422	1.77	3630	894	3.6
	Europa	Galileo (1610)	671	3.55	3138	480	3.0
	Ganymede	Galileo (1610)	1070	7.16	5262	1482	1.9
	Callisto	Galileo (1610)	1883	16.69	4800	1077	1.9
	Himalia	Perrine (1904)	11,460	251	170	—	—
Saturn (62)	Pan	Voyager (1985)	133.6	0.58	20	3×10^{-5}	—
	Atlas	Voyager (1980)	137.7	0.60	40	—	—
	Prometheus	Voyager (1980)	139.4	0.61	80	—	—
	Pandora	Voyager (1980)	141.7	0.63	100	—	—
	Janus	Dollfus (1966)	151.4	0.69	190	—	—
	Epimetheus	Fountain, Larson (1980)	151.4	0.69	120	—	—
	Mimas	Herschel (1789)	186	0.94	394	0.4	1.2
	Enceladus	Herschel (1789)	238	1.37	502	0.8	1.2
	Tethys	Cassini (1684)	295	1.89	1048	7.5	1.3

Table D7

Selected Moons of the Planets

Planet (moons)	Satellite Name	Discovery	Semimajor Axis (km × 1000)	Period (d)	Diameter (km)	Mass (10^{20} kg)	Density (g/cm^3)
	Dione	Cassini (1684)	377	2.74	1120	11	1.3
	Rhea	Cassini (1672)	527	4.52	1530	25	1.3
	Titan	Huygens (1655)	1222	15.95	5150	1346	1.9
	Hyperion	Bond, Lassell (1848)	1481	21.3	270	—	—
	Iapetus	Cassini (1671)	3561	79.3	1435	19	1.2
	Phoebe	Pickering (1898)	12,950	550 (R)[2]	220	—	—
Uranus (27)	Puck	Voyager (1985)	86.0	0.76	170	—	—
	Miranda	Kuiper (1948)	130	1.41	485	0.8	1.3
	Ariel	Lassell (1851)	191	2.52	1160	13	1.6
	Umbriel	Lassell (1851)	266	4.14	1190	13	1.4
	Titania	Herschel (1787)	436	8.71	1610	35	1.6
	Oberon	Herschel (1787)	583	13.5	1550	29	1.5
Neptune (14)	Despina	Voyager (1989)	53	0.33	150	—	—
	Galatea	Voyager (1989)	62	0.40	150	—	—
	Larissa	Voyager (1989)	118	1.12	400	—	—
	Triton	Lassell (1846)	355	5.88 (R)[3]	2720	220	2.1
	Nereid	Kuiper (1949)	5511	360	340	—	—
Pluto (5)	Charon	Christy (1978)	19.7	6.39	1200	—	1.7

Table D7

2. R stands for retrograde rotation (backward from the direction that most objects in the solar system revolve and rotate).
3. R stands for retrograde rotation (backward from the direction that most objects in the solar system revolve and rotate).

Selected Moons of the Planets

Planet (moons)	Satellite Name	Discovery	Semimajor Axis (km × 1000)	Period (d)	Diameter (km)	Mass (10^{20} kg)	Density (g/cm^3)
	Styx	Showalter et al (2012)	42	20	20	—	—
	Nix	Weaver et al (2005)	48	24	46	—	2.1
	Kerberos	Showalter et al (2011)	58	24	28	—	1.4
	Hydra	Weaver et al (2005)	65	38	61	—	0.8
Eris (1)	Dysnomea	Brown et al (2005)	38	16	684	—	—
Makemake (1)	(MK2)	Parker et al (2016)	—	—	160	—	—
Haumea (2)	Hi'iaka	Brown et al (2005)	50	49	400	—	—
	Namaka	Brown et al (2005)	39	35	200	—	—

Table D7

D3 The Nearest Stars, Brown Dwarfs, and White Dwarfs

The Nearest Stars, Brown Dwarfs, and White Dwarfs

Star	System	Discovery Name	Distance (light-year)	Spectral Type	Location: RA[4]	Location: Dec[5]	Luminosity (Sun = 1)
		Sun	—	G2 V	—	—	1
1	1	Proxima Centauri	4.2	M5.5 V	14 29	−62 40	5×10^{-5}
2	2	Alpha Centauri A	4.4	G2 V	14 39	−60 50	1.5
3		Alpha Centauri B	4.4	K2 IV	14 39	−60 50	0.5
4	3	Barnard's Star	6.0	M4 V	17 57	+04 42	4.4×10^{-4}
5	4	Wolf 359	7.8	M6 V	10 56	+07 00	2×10^{-5}
6	5	Lalande 21 185	8.3	M2 V	11 03	+35 58	5.7×10^{-3}
7	6	Sirius A	8.6	A1 V	06 45	−16 42	23.1

Table D8

4. Location (right ascension) given for Epoch 2000.0
5. Location (declination) given for Epoch 2000.0

The Nearest Stars, Brown Dwarfs, and White Dwarfs

Star	System	Discovery Name	Distance (light-year)	Spectral Type	Location: RA	Location: Dec	Luminosity (Sun = 1)
8		Sirius B	8.6	DA2[6]	06 45	−16 43	2.5×10^{-3}
9	7	Luyten 726-8 A	8.7	M5.5 V	01 39	−17 57	6×10^{-5}
10		Luyten 726-8 B (UV Ceti)	8.7	M6 V	01 39	−17 57	4×10^{-5}
11	8	Ross 154	9.7	M.05 V	18 49	−23 50	5×10^{-4}
12	9	Ross 248 (HH Andromedae)	10.3	M5.5 V	23 41	+44 10	1.0×10^{-4}
13	10	Epsilon Eridani	10.5	K2 V	03 32	−09 27	0.29
14	11	Lacaille 9352	10.7	M0.5 V	23 05	−35 51	0.011
15	12	Ross 128 (FI Virginis)	10.9	M4 V	11 47	+00 48	3.4×10^{-4}
16	13	Luyten 789-6 A (EZ Aquarii A)	11.3	M5 V	22 38	−15 17	5×10^{-5}
17		Luyten 789-6 B (EZ Aquarii B)	11.3	M5.5 V	22 38	−15 15	5×10^{-5}
18		Luyten 789-6 C (EZ Aquarii C)	11.3	M6.5 V	22 38	−15 17	2×10^{-5}
19	14	61 Cygni A	11.4	K5 V	21 06	+38 44	0.086
20		61 Cygni B	11.4	K7 V	21 06	+38 44	0.041
21	15	Procyon A	11.4	F51V	07 39	+05 13	7.38
22		Procyon B	11.4	wd[7]	07 39	+05 13	5.5×10^{-4}
23	16	Sigma 2398 A	11.5	M3 V	18 42	+59 37	0.003
24		Sigma 2398 B	11.5	M3.5 V	18 42	+59 37	1.4×10^{-3}
25	17	Groombridge 34 A (GX Andromedae)	11.6	M1.5 V	00 18	+44 01	6.4×10^{-3}
26		Groombridge 34 B (GQ Andromedae)	11.6	M3.5 V	00 18	+44 01	4.1×10^{-4}
27	18	Epsilon Indi A	11.8	K5 V	22 03	−56 46	0.150
28		Epsilon Indi Ba	11.7	T1[8]	22 04	−56 46	—

Table D8

6. White dwarf stellar remnant
7. White dwarf stellar remnant

The Nearest Stars, Brown Dwarfs, and White Dwarfs

Star	System	Discovery Name	Distance (light-year)	Spectral Type	Location: RA	Location: Dec	Luminosity (Sun = 1)
29		Epsilon Indi Bb	11.7	T6[9]	22 04	−56 46	—
30	19	G 51-15 (DX Cancri)	11.8	M6.5 V	08 29	+26 46	1×10^{-5}
31	20	Tau Ceti	11.9	G8.5 V	01 44	−15 56	0.458
32	21	Luyten 372-58	12.0	M5 V	03 35	−44 30	7×10^{-5}
33	22	Luyten 725-32 (YZ Ceti)	12.1	M4.5 V	01 12	−16 59	1.8×10^{-4}
34	23	Luyten's Star	12.4	M3.5 V	07 27	+05 13	1.4×10^{-3}
35	24	SCR J184-6357 A	12.6	M8.5 V	18 45	−63 57	1×10^{-6}
36		SCR J184-6357 B	12.7	T6[10]	18 45	−63 57	—
37	25	Teegarden's Star	12.5	M6 V	02 53	+16 52	1×10^{-5}
38	26	Kapteyn's Star	12.8	M1 V	05 11	−45 01	3.8×10^{-3}
39	27	Lacaille 8760 (AX Microscopium)	12.9	K7 V	21 17	−38 52	0.029

Table D8

D4 Stellar Data

Note: These are the stars that *appear* the brightest visually, as seen from our vantage point on Earth. They are not the stars that are intrinsically the most luminous.

8. Brown dwarf
9. Brown dwarf
10. Brown dwarf

The Brightest Twenty Stars										
Name					Proper Motion (arcsec/y)		Right Ascension		Declination	
Traditional	Bayer	Luminosity (Sun = 1)	Distance (light-years)	Spectral Type	RA	Dec	(h)	(m)	(deg)	(min)
Sirius	α Canis Majoris	22.5	8.6	A1 V	–0.5	–1.2	06	45.2	–16	43
Canopus	α Carinae	13,500	309	F0 II	+0.02	+0.02	06	24.0	–52	42
Rigil Kentaurus	α Centauri	1.94	4.32	G2 V + K IV	–3.7	+0.5	14	39.7	–60	50
Arcturus	α Bootis	120	36.72	K1.5 III	–1.1	–2.0	14	15.7	+19	11
Vega	α Lyrae	49	25.04	A0 V	+0.2	+0.3	18	36.9	+38	47
Capella	α Aurigae	140	42.80	G8 III + G0 III	+0.08	–0.4	05	16.7	+46	00
Rigel	β Orionis	50,600	863	B8 I	+0.00	+0.00	05	14.5	–08	12
Procyon	α Canis Minoris	7.31	11.46	F5 IV-V	–0.7	–1.0	07	39.3	+05	14
Achernar	α Eridani	1030	139	B3 V	+0.10	–0.04	01	37.7	–57	14
Betelgeuse	α Orionis	13,200	498	M2 I	+0.02	+0.01	05	55.2	+07	24
Hadar	β Centauri	7050	392	B1 III	–0.03	–0.02	14	03.8	–60	22
Altair	α Aquilae	11.2	16.73	A7 V	+0.5	+0.4	19	50.8	+08	52
Acrux	α Crucis	4090	322	B0.5 IV + B1V	–0.04	–0.01	12	26.6	–63	06
Aldebaran	α Tauri	160	66.64	K5 III	+0.1	–0.2	04	35.9	+16	31
Spica	α Virginis	2030	250	B1 III-IV + B2 V	–0.04	–0.03	13	25.2	–11	10
Antares	α Scorpii	9290	554	M1.5 I + B2.5 V	–0.01	–0.02	16	29.4	–26	26
Pollux	β Geminorum	31.6	33.78	K0 III	–0.6	–0.05	07	45.3	+28	02
Fomalhaut	α Piscis Austrini	17.2	25.13	A3 V	+0.03	–0.2	22	57.6	–29	37
Mimosa	β Crucis	1980	279	B0.5 III	–0.04	–0.02	12	47.7	–59	41
Deneb	α Cygni	50,600	1412	A2 I	+0.00	+0.00	20	41.4	+45	17

Figure D1 The brightest stars typically have names from antiquity. Next to each star's ancient name, we have added a column with its name in the system originated by Bayer (see the **Naming Stars (https://legacy.cnx.org/content/m59905/latest/#fs-id1165721974313)** feature box.) The distances of the more remote stars are estimated from their spectral types and apparent brightnesses and are only approximate. The luminosities for those stars are approximate to the same degree. Right ascension and declination is given for Epoch 2000.0.

D5 Galaxies Visible from 40°N Latitude

Galaxies

Name	RA (2000.0)	Dec.	Const.	Diam.	Magn.	Dist.(Mpc)	Type	Nuclear class	Common Name	View
M31	$00^h 42.7^m$	$+41^\circ 16^m$	Andromeda	$178 \times 63'$	3.4	0.78	Sb I-II		Andromeda Galaxy	
NGC 253	00 47.6	−25 17	Sculptor	25.1 x 7.4	7.3	3.5	Sc p	3, Extremely Small	Sculptor Galaxy	
NGC 300	00 54.9	−37 41	Sculptor	20.0 x 14.8	8.2	2.6	Sd III-IV	3, Extremely Small		
M33	01 33.9	+30 39	Triangulum	62 x 39	5.7	0.94	Sc II-III		Triangulum Galaxy	
M81	09 55.6	+69 04	Ursa Major	25.7 x 14.1	6.9	5.6	Sb I-II		Cigar Galaxy	
M82	09 55.8	+69 41	Ursa Major	11.2 x 4.6	8.3	2.2	P		Bode's Galaxy	
M106	12 19.0	+47 18	Canes Venatici	18.2 x 7.9	8.4	6.9	Sb+p			
M49	12 29.8	+08 00	Virgo	8.9 x 7.4	8.3	12	E4			
M104	12 40.0	−11 37	Virgo	8.9 x 4.1	8.1	19	Sb-	5	Sombrero Galaxy	
M94	12 50.9	+41 07	Canes Venatici	11.0 x 9.1	8.1	7.1	Sb-p II:	5		
M64	12 56.7	+21 41	Coma Berenices	9.3 x 5.4	8.4	7.0	Sb-		Black eye Galaxy	
NGC 4945	13 05.4	−49 28	Centaurus	20.0 x 4.4	8.6	2.8	SBc:	4, Very Small		
M63	13 15.8	+42 02	Canes Venatici	12.3 x 7.6	8.6	9.2	Sb+ II			
NGC 5128	13 25.5	−43 01	Centaurus	18.2 x 14.5	6.7	4.5	S0p		Centaurus A	
M51	13 29.9	+47 12	Canes Venatici	11.0 x 7.8	8.3	6.3	Sc I	5	Whirlpool Galaxy	
M83	13 37.0	−29 52	Hydra	11.2 x 10.2	7.2	3.7	Sc I-II		Southern Pinwheel Galaxy	
M101	14 03.2	+54 21	Ursa Major	26.9 x 26.3	7.8	5.8	Sc I	3, Diffuse	Pinwheel Galaxy	

Out of sight from 40°N

Name	RA (2000.0)	Dec.	Const.	Diam.	Magn.	Dist.(Mpc)	Type	Nuclear class	Common Name	View
NGC 292	$00^h 52.7^m$	$-72^\circ 50^m$	Tucana	$280 \times 160'$	2.3	0.08	SBmp IV		Small Magellanic Cloud	
(PGC 17223)	05 23.6	−69 45	Dorado	650 x 550	0.9	0.048	SB(s)m III-IV		Large Magellanic Cloud	
NGC 6744	19 09.8	−63 51	Pavo	15.5 x 10.2	8.4	6.7	S(B)b+ II			

Figure D2 Source: **Roberto Mura (https://commons.wikimedia.org/wiki/User:Roberto_Mura)** , **Galaxies table 40°N (https://commons.wikimedia.org/wiki/File:Galaxies_table_40°N.png)** , file type, **CC BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0/legalcode)**

APPENDIX E | USEFUL MATHEMATICAL FORMULAS

Quadratic formula

If $ax^2 + bx + c = 0$, then $x = \dfrac{-b \pm \sqrt{b^2 - 4ac}}{2a}$

Triangle of base b and height h	**Area $= \frac{1}{2}bh$**	
Circle of radius r	Circumference $= 2\pi r$	Area $= \pi r^2$
Sphere of radius r	Surface area $= 4\pi r^2$	Volume $= \frac{4}{3}\pi r^3$
Cylinder of radius r and height h	Area of curved surface $= 2\pi rh$	Volume $= \pi r^2 h$

Table E1 Geometry

Trigonometry

Trigonometric Identities

1. $\sin\theta = 1/\csc\theta$
2. $\cos\theta = 1/\sec\theta$
3. $\tan\theta = 1/\cot\theta$
4. $\sin\left(90^0 - \theta\right) = \cos\theta$
5. $\cos\left(90^0 - \theta\right) = \sin\theta$
6. $\tan\left(90^0 - \theta\right) = \cot\theta$
7. $\sin^2\theta + \cos^2\theta = 1$
8. $\sec^2\theta - \tan^2\theta = 1$
9. $\tan\theta = \sin\theta/\cos\theta$
10. $\sin(\alpha \pm \beta) = \sin\alpha\cos\beta \pm \cos\alpha\sin\beta$
11. $\cos(\alpha \pm \beta) = \cos\alpha\cos\beta \mp \sin\alpha\sin\beta$
12. $\tan(\alpha \pm \beta) = \dfrac{\tan\alpha \pm \tan\beta}{1 \mp \tan\alpha\tan\beta}$
13. $\sin 2\theta = 2\sin\theta\cos\theta$
14. $\cos 2\theta = \cos^2\theta - \sin^2\theta = 2\cos^2\theta - 1 = 1 - 2\sin^2\theta$

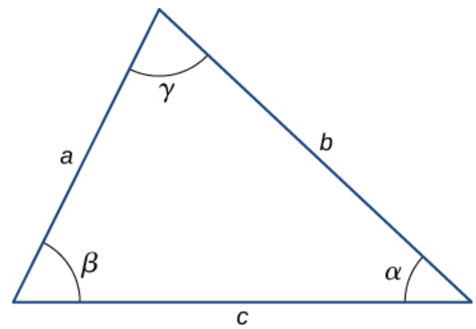
15. $\sin\alpha + \sin\beta = 2\sin\frac{1}{2}(\alpha+\beta)\cos\frac{1}{2}(\alpha-\beta)$

16. $\cos\alpha + \cos\beta = 2\cos\frac{1}{2}(\alpha+\beta)\cos\frac{1}{2}(\alpha-\beta)$

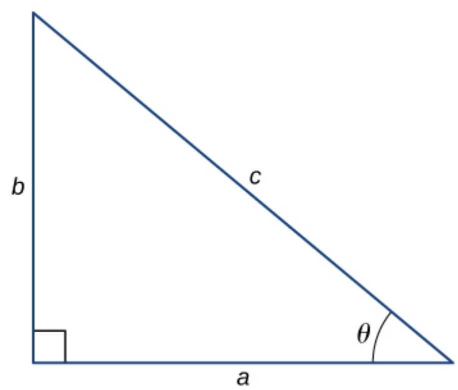
Triangles

1. Law of sines: $\frac{a}{\sin\alpha} = \frac{b}{\sin\beta} = \frac{c}{\sin\gamma}$

2. Law of cosines: $c^2 = a^2 + b^2 - 2ab\cos\gamma$



3. Pythagorean theorem: $a^2 + b^2 = c^2$



Series expansions

1. Binomial theorem: $(a+b)^n = a^n + na^{n-1}b + \frac{n(n-1)a^{n-2}b^2}{2!} + \frac{n(n-1)(n-2)a^{n-3}b^3}{3!} + \cdots$

2. $(1 \pm x)^n = 1 \pm \frac{nx}{1!} + \frac{n(n-1)x^2}{2!} \pm \cdots \left(x^2 < 1\right)$

3. $(1 \pm x)^{-n} = 1 \mp \frac{nx}{1!} + \frac{n(n+1)x^2}{2!} \mp \cdots \left(x^2 < 1\right)$

4. $\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots$

5. $\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots$

6. $\tan x = x + \frac{x^3}{3} + \frac{2x^5}{15} + \cdots$

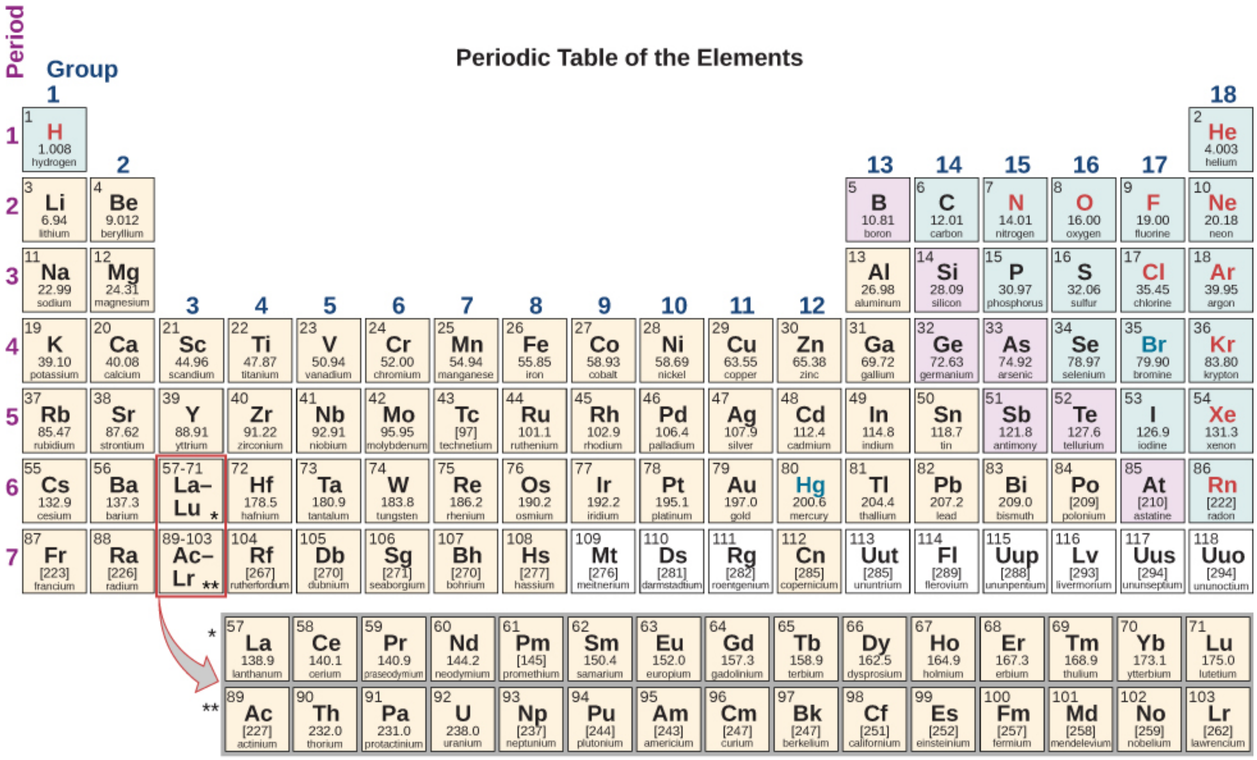
7. $e^x = 1 + x + \frac{x^2}{2!} + \cdots$

8. $\ln(1+x) = x - \frac{1}{2}x^2 + \frac{1}{3}x^3 - \cdots (|x| < 1)$

Derivatives

1. $\frac{d}{dx}[af(x)] = a\frac{d}{dx}f(x)$
2. $\frac{d}{dx}[f(x) + g(x)] = \frac{d}{dx}f(x) + \frac{d}{dx}g(x)$
3. $\frac{d}{dx}[f(x)g(x)] = f(x)\frac{d}{dx}g(x) + g(x)\frac{d}{dx}f(x)$
4. $\frac{d}{dx}f(u) = \left[\frac{d}{du}f(u)\right]\frac{du}{dx}$
5. $\frac{d}{dx}x^m = mx^{m-1}$
6. $\frac{d}{dx}\sin x = \cos x$
7. $\frac{d}{dx}\cos x = -\sin x$
8. $\frac{d}{dx}\tan x = \sec^2 x$
9. $\frac{d}{dx}\cot x = -\csc^2 x$
10. $\frac{d}{dx}\sec x = \tan x \sec x$
11. $\frac{d}{dx}\csc x = -\cot x \csc x$
12. $\frac{d}{dx}e^x = e^x$
13. $\frac{d}{dx}\ln x = \frac{1}{x}$
14. $\frac{d}{dx}\sin^{-1} x = \frac{1}{\sqrt{1 - x^2}}$
15. $\frac{d}{dx}\cos^{-1} x = -\frac{1}{\sqrt{1 - x^2}}$
16. $\frac{d}{dx}\tan^{-1} x = -\frac{1}{1 + x^2}$

APPENDIX F | THE CHEMICAL ELEMENTS



The Chemical Elements

Element	Symbol	Atomic Number	Atomic Weight[1]	Percentage of Naturally Occurring Elements in the Universe
Hydrogen	H	1	1.008	75
Helium	He	2	4.003	23
Lithium	Li	3	6.94	6×10^{-7}
Beryllium	Be	4	9.012	1×10^{-7}
Boron	B	5	10.821	1×10^{-7}
Carbon	C	6	12.011	0.5

Table F1

1. Where mean atomic weights have not been well determined, the atomic mass numbers of the most stable isotopes are given in parentheses.

The Chemical Elements

Element	Symbol	Atomic Number	Atomic Weight	Percentage of Naturally Occurring Elements in the Universe
Nitrogen	N	7	14.007	0.1
Oxygen	O	8	15.999	1
Fluorine	F	9	18.998	4×10^{-5}
Neon	Ne	10	20.180	0.13
Sodium	Na	11	22.990	0.002
Magnesium	Mg	12	24.305	0.06
Aluminum	Al	13	26.982	0.005
Silicon	Si	14	28.085	0.07
Phosphorus	P	15	30.974	7×10^{-4}
Sulfur	S	16	32.06	0.05
Chlorine	Cl	17	35.45	1×10^{-4}
Argon	Ar	18	39.948	0.02
Potassium	K	19	39.098	3×10^{-4}
Calcium	Ca	20	40.078	0.007
Scandium	Sc	21	44.956	3×10^{-6}
Titanium	Ti	22	47.867	3×10^{-4}
Vanadium	V	23	50.942	3×10^{-4}
Chromium	Cr	24	51.996	0.0015
Manganese	Mn	25	54.938	8×10^{-4}
Iron	Fe	26	55.845	0.11
Cobalt	Co	27	58.933	3×10^{-4}
Nickel	Ni	28	58.693	0.006
Copper	Cu	29	63.546	6×10^{-6}
Zinc	Zn	30	65.38	3×10^{-5}
Gallium	Ga	31	69.723	1×10^{-6}
Germanium	Ge	32	72.630	2×10^{-5}
Arsenic	As	33	74.922	8×10^{-7}
Selenium	Se	34	78.971	3×10^{-6}
Bromine	Br	35	79.904	7×10^{-7}
Krypton	Kr	36	83.798	4×10^{-6}
Rubidium	Rb	37	85.468	1×10^{-6}
’Strontium	Sr	38	87.62	4×10^{-6}
Yttrium	Y	39	88.906	7×10^{-7}
Zirconium	Zr	40	91.224	5×10^{-6}

Table F1

The Chemical Elements

Element	Symbol	Atomic Number	Atomic Weight	Percentage of Naturally Occurring Elements in the Universe
Niobium	Nb	41	92.906	2×10^{-7}
Molybdenum	Mo	42	95.95	5×10^{-7}
Technetium	Tc	43	(98)	—
Ruthenium	Ru	44	101.07	4×10^{-7}
Rhodium	Rh	45	102.906	6×10^{-8}
Palladium	Pd	46	106.42	2×10^{-7}
Silver	Ag	47	107.868	6×10^{-8}
Cadmium	Cd	48	112.414	2×10^{-7}
Indium	In	49	114.818	3×10^{-8}
Tin	Sn	50	118.710	4×10^{-7}
Antimony	Sb	51	121.760	4×10^{-8}
Tellurium	Te	52	127.60	9×10^{-7}
Iodine	I	53	126.904	1×10^{-7}
Xenon	Xe	54	131.293	1×10^{-6}
Cesium	Cs	55	132.905	8×10^{-8}
Barium	Ba	56	137.327	1×10^{-6}
Lanthanum	La	57	138.905	2×10^{-7}
Cerium	Ce	58	140.116	1×10^{-6}
Praseodymium	Pr	59	140.907	2×10^{-7}
Neodymium	Nd	60	144.242	1×10^{-6}
Promethium	Pm	61	(145)	—
Samarium	Sm	62	150.36	5×10^{-7}
Europium	Eu	63	151.964	5×10^{-8}
Gadolinium	Gd	64	157.25	2×10^{-7}
Terbium	Tb	65	158.925	5×10^{-8}
Dysprosium	Dy	66	162.500	2×10^{-7}
Holmium	Ho	67	164.930	5×10^{-8}
Erbium	Er	68	167.259	2×10^{-7}
Thulium	Tm	69	168.934	1×10^{-8}
Ytterbium	Yb	70	173.054	2×10^{-7}
Lutetium	Lu	71	174.967	1×10^{-8}
Hafnium	Hf	72	178.49	7×10^{-8}
Tantalum	Ta	73	180.948	8×10^{-9}
Tungsten	W	74	183.84	5×10^{-8}

Table F1

The Chemical Elements

Element	Symbol	Atomic Number	Atomic Weight	Percentage of Naturally Occurring Elements in the Universe
Rhenium	Re	75	186.207	2×10^{-8}
Osmium	Os	76	190.23	3×10^{-7}
Iridium	Ir	77	192.217	2×10^{-7}
Platinum	Pt	78	195.084	5×10^{-7}
Gold	Au	79	196.967	6×10^{-8}
Mercury	Hg	80	200.592	1×10^{-7}
Thallium	Tl	81	204.38	5×10^{-8}
Lead	Pb	82	207.2	1×10^{-6}
Bismuth	Bi	83	208.980	7×10^{-8}
Polonium	Po	84	(209)	—
Astatine	At	85	(210)	—
Radon	Rn	86	(222)	—
Francium	Fr	87	(223)	—
Radium	Ra	88	(226)	—
Actinium	Ac	89	(227)	—
Thorium	Th	90	232.038	4×10^{-8}
Protactinium	Pa	91	231.036	—
Uranium	U	92	238.029	2×10^{-8}
Neptunium	Np	93	(237)	—
Plutonium	Pu	94	(244)	—
Americium	Am	95	(243)	—
Curium	Cm	96	(247)	—
Berkelium	Bk	97	(247)	—
Californium	Cf	98	(251)	—
Einsteinium	Es	99	(252)	—
Fermium	Fm	100	(257)	—
Mendelevium	Md	101	(258)	—
Nobelium	No	102	(259)	—
Lawrencium	Lr	103	(262)	—
Rutherfordium	Rf	104	(267)	—
Dubnium	Db	105	(268)	—
Seaborgium	Sg	106	(271)	—
Bohrium	Bh	107	(272)	—
Hassium	Hs	108	(270)	—

Table F1

The Chemical Elements

Element	Symbol	Atomic Number	Atomic Weight	Percentage of Naturally Occurring Elements in the Universe
Meitnerium	Mt	109	(276)	—
Darmstadtium	Ds	110	(281)	—
Roentgenium	Rg	111	(280)	—
Copernicium	Cn	112	(285)	—
Nihonium	Nh	113	(284)	—
Flerovium	Fl	114	(289)	—
Moskovium	Mc	115	(288)	—
Livermorium	Lv	116	(293)	—
Tennessine	Ts	117	(294)	—
Oganesson	Og	118	(294)	—

Table F1

APPENDIX G | THE GREEK ALPHABET

Name	Capital	Lowercase	Name	Capital	Lowercase
Alpha	A	α	Nu	N	ν
Beta	B	β	Xi	Ξ	ξ
Gamma	Γ	γ	Omicron	O	o
Delta	Δ	δ	Pi	Π	π
Epsilon	E	ε	Rho	P	ρ
Zeta	Z	ζ	Sigma	Σ	σ
Eta	H	η	Tau	T	τ
Theta	Θ	θ	Upsilon	Υ	υ
Iota	I	ι	Phi	Φ	ϕ
Kappa	K	κ	Chi	X	χ
Lambda	Λ	λ	Psi	Ψ	ψ
Mu	M	μ	Omega	Ω	ω

Table G1 The Greek Alphabet

ANSWER KEY

CHAPTER 1

CHECK YOUR UNDERSTANDING

1.1. 4.79×10^2 Mg or 479 Mg
1.2. 3×10^8 m/s
1.3. 10^8 km^2
1.4. The numbers were too small, by a factor of 4.45.
1.5. $4\pi r^3/3$
1.6. yes
1.7. 3×10^4 m or 30 km. It is probably an underestimate because the density of the atmosphere decreases with altitude. (In fact, 30 km does not even get us out of the stratosphere.)
1.8. No, the coach's new stopwatch will not be helpful. The uncertainty in the stopwatch is too great to differentiate between the sprint times effectively.

CONCEPTUAL QUESTIONS

1. Physics is the science concerned with describing the interactions of energy, matter, space, and time to uncover the fundamental mechanisms that underlie every phenomenon.
3. No, neither of these two theories is more valid than the other. Experimentation is the ultimate decider. If experimental evidence does not suggest one theory over the other, then both are equally valid. A given physicist might prefer one theory over another on the grounds that one seems more simple, more natural, or more beautiful than the other, but that physicist would quickly acknowledge that he or she cannot say the other theory is invalid. Rather, he or she would be honest about the fact that more experimental evidence is needed to determine which theory is a better description of nature.
5. Probably not. As the saying goes, "Extraordinary claims require extraordinary evidence."
7. Conversions between units require factors of 10 only, which simplifies calculations. Also, the same basic units can be scaled up or down using metric prefixes to sizes appropriate for the problem at hand.
9. a. Base units are defined by a particular process of measuring a base quantity whereas derived units are defined as algebraic combinations of base units. b. A base quantity is chosen by convention and practical considerations. Derived quantities are expressed as algebraic combinations of base quantities. c. A base unit is a standard for expressing the measurement of a base quantity within a particular system of units. So, a measurement of a base quantity could be expressed in terms of a base unit in any system of units using the same base quantities. For example, length is a base quantity in both SI and the English system, but the meter is a base unit in the SI system only.
11. a. Uncertainty is a quantitative measure of precision. b. Discrepancy is a quantitative measure of accuracy.
13. Check to make sure it makes sense and assess its significance.

PROBLEMS

15. a. 10^3; b. 10^5; c. 10^2; d. 10^{15}; e. 10^2; f. 10^{57}
17. 10^2 generations
19. 10^{11} atoms
21. 10^3 nerve impulses/s
23. 10^{26} floating-point operations per human lifetime
25. a. 957 ks; b. 4.5 cs or 45 ms; c. 550 ns; d. 31.6 Ms
27. a. 75.9 Mm; b. 7.4 mm; c. 88 pm; d. 16.3 Tm
29. a. 3.8 cg or 38 mg; b. 230 Eg; c. 24 ng; d. 8 Eg e. 4.2 g
31. a. 27.8 m/s; b. 62 mi/h
33. a. 3.6 km/h; b. 2.2 mi/h
35. 1.05×10^5 ft^2
37. 8.847 km
39. a. 1.3×10^{-9} m; b. 40 km/My
41. 10^6 Mg/μL
43. 62.4 lbm/ft^3
45. 0.017 rad
47. 1 light-nanosecond
49. 3.6×10^{-4} m^3
51. a. Yes, both terms have dimension L^2T^{-2} b. No. c. Yes, both terms have dimension LT^{-1} d. Yes, both terms have dimension

LT^{-2}

53. a. $[v] = LT^{-1}$; b. $[a] = LT^{-2}$; c. $\left[\int vdt\right] = L$; d. $\left[\int adt\right] = LT^{-1}$; e. $\left[\frac{da}{dt}\right] = LT^{-3}$

55. a. L; b. L; c. $L^0 = 1$ (that is, it is dimensionless)
56. 25.7 miles per gallon
57. 24.0 miles per gallon
58. 25.0 miles per gallon
60. 10^{28} atoms
62. 10^{51} molecules
64. 10^{16} solar systems
66. a. Volume = 10^{27} m^3, diameter is 10^9 m.; b. 10^{11} m
68. a. A reasonable estimate might be one operation per second for a total of 10^9 in a lifetime.; b. about $(10^9)(10^{-17}\text{ s}) = 10^{-8}$ s, or about 10 ns
70. 2 kg
72. 4%
74. 67 mL
76. a. The number 99 has 2 significant figures; 100. has 3 significant figures. b. 1.00%; c. percent uncertainties
78. a. 2%; b. 1 mm Hg
80. 7.557 cm^2
82. a. 37.2 lb; because the number of bags is an exact value, it is not considered in the significant figures; b. 1.4 N; because the value 55 kg has only two significant figures, the final value must also contain two significant figures

ADDITIONAL PROBLEMS

84. a. $[s_0] = L$ and units are meters (m); b. $[v_0] = LT^{-1}$ and units are meters per second (m/s); c. $[a_0] = LT^{-2}$ and units are meters per second squared (m/s^2); d. $[j_0] = LT^{-3}$ and units are meters per second cubed (m/s^3); e. $[S_0] = LT^{-4}$ and units are m/s^4; f. $[c] = LT^{-5}$ and units are m/s^5.

86. a. 0.059%; b. 0.01%; c. 4.681 m/s; d. 0.07%, 0.003 m/s
88. a. 0.02%; b. 1×10^4 lbm
90. a. 143.6 cm^3; b. 0.1 cm^3 or 0.084%

CHALLENGE PROBLEMS

92. Since each term in the power series involves the argument raised to a different power, the only way that every term in the power series can have the same dimension is if the argument is dimensionless. To see this explicitly, suppose $[x] = L^aM^bT^c$. Then, $[x^n] = [x]^n = L^{an}M^{bn}T^{cn}$. If we want $[x] = [x^n]$, then an = a, bn = b, and cn = c for all n. The only way this can happen is if a = b = c = 0.

CHAPTER 3

CHECK YOUR UNDERSTANDING

3.1. (a) The rider's displacement is $\Delta x = x_f - x_0 = -1\text{ km}$. (The displacement is negative because we take east to be positive and west to be negative.) (b) The distance traveled is 3 km + 2 km = 5 km. (c) The magnitude of the displacement is 1 km.
3.2. Inserting the knowns, we have

$$\bar{a} = \frac{\Delta v}{\Delta t} = \frac{2.0\times10^7\text{ m/s} - 0}{10^{-4}\text{ s} - 0} = 2.0\times10^{11}\text{ m/s}^2.$$

3.3. If we take east to be positive, then the airplane has negative acceleration because it is accelerating toward the west. It is also decelerating; its acceleration is opposite in direction to its velocity.
3.4. To answer this, choose an equation that allows us to solve for time *t*, given only a, v_0, and *v*:

$$v = v_0 + at.$$

Rearrange to solve for *t*:

$$t = \frac{v - v_0}{a} = \frac{400\text{ m/s} - 0\text{ m/s}}{20\text{ m/s}^2} = 20\text{ s}.$$

3.5. $a = \frac{2}{3}\text{ m/s}^2$.

3.6. It takes 2.47 s to hit the water. The quantity distance traveled increases faster.

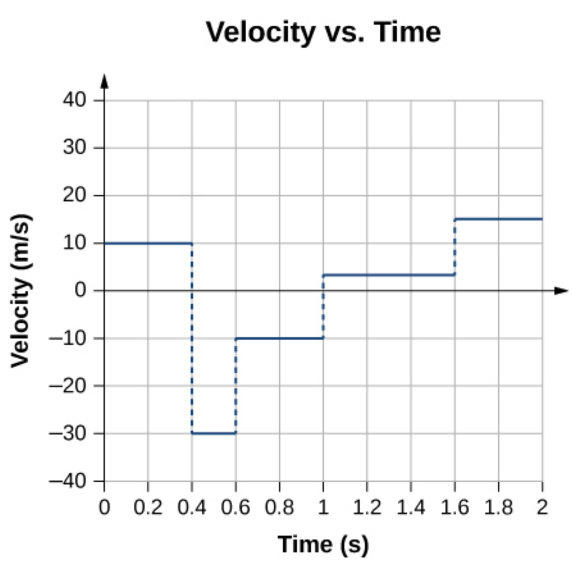
CONCEPTUAL QUESTIONS

1. You drive your car into town and return to drive past your house to a friend's house.

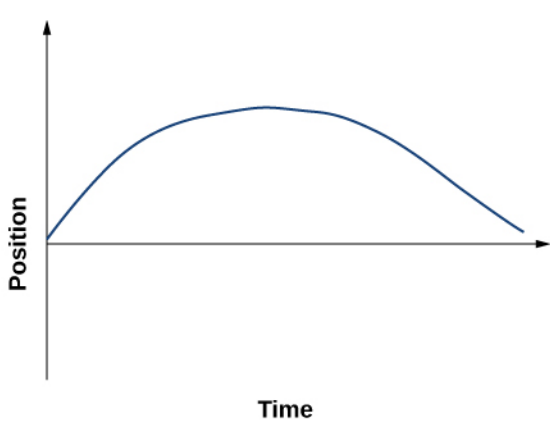
3. If the bacteria are moving back and forth, then the displacements are canceling each other and the final displacement is small.
5. Distance traveled
7. Average speed is the total distance traveled divided by the elapsed time. If you go for a walk, leaving and returning to your home, your average speed is a positive number. Since Average velocity = Displacement/Elapsed time, your average velocity is zero.
9. Average speed. They are the same if the car doesn't reverse direction.
11. No, in one dimension constant speed requires zero acceleration.
13. A ball is thrown into the air and its velocity is zero at the apex of the throw, but acceleration is not zero.
15. Plus, minus
17. If the acceleration, time, and displacement are the knowns, and the initial and final velocities are the unknowns, then two kinematic equations must be solved simultaneously. Also if the final velocity, time, and displacement are the knowns then two kinematic equations must be solved for the initial velocity and acceleration.
19. a. at the top of its trajectory; b. yes, at the top of its trajectory; c. yes
21. Earth $v = v_0 - gt = -gt$; Moon $v' = \frac{g}{6}t'$ $v = v'$ $-gt = -\frac{g}{6}t'$ $t' = 6t$; Earth $y = -\frac{1}{2}gt^2$ Moon $y' = -\frac{1}{2}\frac{g}{6}(6t)^2 = -\frac{1}{2}g6t^2 = -6\left(\frac{1}{2}gt^2\right) = -6y$

PROBLEMS

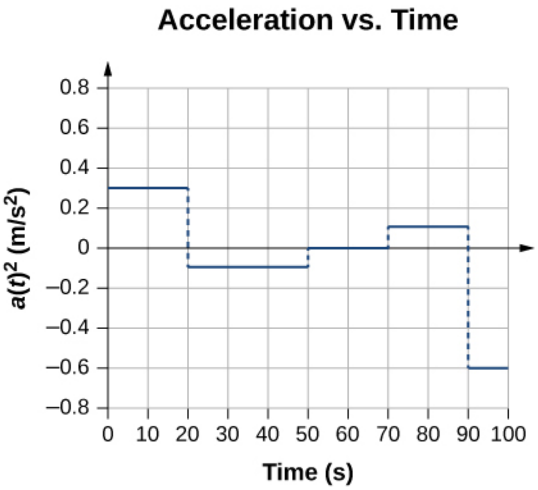
25. a. $\vec{\mathbf{x}}_1 = (-2.0\,\text{m})\,\hat{\mathbf{i}}$, $\vec{\mathbf{x}}_2 = (5.0\,\text{m})\,\hat{\mathbf{i}}$; b. 7.0 m east
27. a. $t = 2.0$ s; b. $x(6.0) - x(3.0) = -8.0 - (-2.0) = -6.0\,\text{m}$
29. a. 150.0 s, $\bar{v} = 156.7\,\text{m/s}$; b. 45.7% the speed of sound at sea level



31.



33.
34. $a = 4.29\text{m/s}^2$



36.

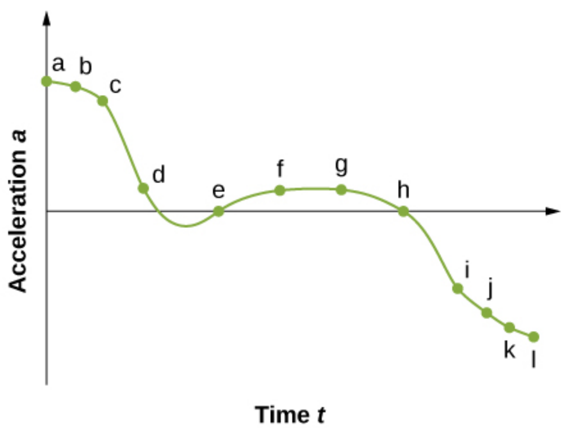
38. $a = 11.1g$

40. 150 m

42. a. 525 m;

b. $v = 180 \text{ m/s}$

44. a.



b. The acceleration has the greatest positive value at t_a

c. The acceleration is zero at t_e and t_h

d. The acceleration is negative at t_i, t_j, t_k, t_l

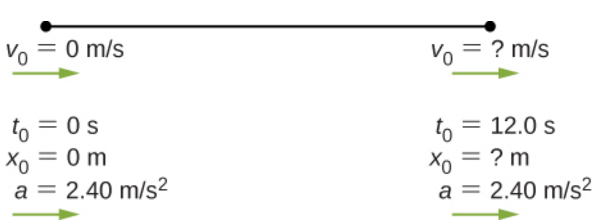
46. a. $a = -1.3 \text{ m/s}^2$;

b. $v_0 = 18 \text{ m/s}$;

c. $t = 13.8 \text{ s}$

48. $v = 502.20 \text{ m/s}$

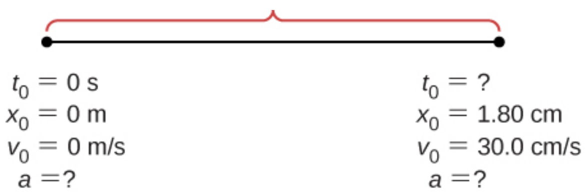
50. a.



b. Knowns: $a = 2.40 \text{ m/s}^2$, $t = 12.0 \text{ s}$, $v_0 = 0 \text{ m/s}$, and $x_0 = 0 \text{ m}$;

c. $x = x_0 + v_0 t + \frac{1}{2}at^2 = \frac{1}{2}at^2 = 2.40\ \text{m/s}^2 (12.0\ \text{s})^2 = 172.80\ \text{m}$, the answer seems reasonable at about 172.8 m; d. $v = 28.8\ \text{m/s}$

52. a.



b. Knowns: $v = 30.0\ \text{cm/s}, x = 1.80\ \text{cm}$;

c. $a = 250\ \text{cm/s}^2, \quad t = 0.12\ \text{s}$;

d. yes

54. a. 6.87 m/s^2; b. $x = 52.26\ \text{m}$

56. a. $a = 8450\ \text{m/s}^2$;

b. $t = 0.0077\ \text{s}$

58. a. $a = 9.18\ g$;

b. $t = 6.67 \times 10^{-3}\ \text{s}$;

c. $\begin{array}{l} a = -40.0\ \text{m/s}^2 \\ a = 4.08\ g \end{array}$

60. Knowns: $x = 3\ \text{m}, v = 0\ \text{m/s}, \ v_0 = 54\ \text{m/s}$. We want *a*, so we can use this equation: $a = -486\ \text{m/s}^2$.

62. a. $a = 32.58\ \text{m/s}^2$;

b. $v = 161.85\ \text{m/s}$;

c. $v > v_{\text{max}}$, because the assumption of constant acceleration is not valid for a dragster. A dragster changes gears and would have a greater acceleration in first gear than second gear than third gear, and so on. The acceleration would be greatest at the beginning, so it would not be accelerating at $32.6\ \text{m/s}^2$ during the last few meters, but substantially less, and the final velocity would be less than $162\ \text{m/s}$.

64. a. $\begin{array}{l} y = -8.23\ \text{m} \\ v_1 = -18.9\ \text{m/s} \end{array}$;

b. $\begin{array}{l} y = -18.9\ \text{m} \\ v_2 = 23.8\ \text{m/s} \end{array}$;

c. $\begin{array}{l} y = -32.0\ \text{m} \\ v_3 = -28.7\ \text{m/s} \end{array}$;

d. $\begin{array}{l} y = -47.6\ \text{m} \\ v_4 = -33.6\ \text{m/s} \end{array}$;

e. $\begin{array}{l} y = -65.6\ \text{m} \\ v_5 = -38.5\ \text{m/s} \end{array}$

66. a. Knowns: $a = -9.8\ \text{m/s}^2 \quad v_0 = -1.4\ \text{m/s} \quad t = 1.8\ \text{s} \quad y_0 = 0\ \text{m}$;

b. $y = y_0 + v_0 t - \frac{1}{2}gt^2 \quad y = v_0 t - \frac{1}{2}gt = -1.4\ \text{m/s}(1.8\ \text{sec}) - \frac{1}{2}(9.8)(1.8\ \text{s})^2 = -18.4\ \text{m}$ and the origin is at the rescuers, who are 18.4 m above the water.

68. a. $v^2 = v_0^2 - 2g(y - y_0) \quad y_0 = 0 \quad v = 0 \quad y = \frac{v_0^2}{2g} = \frac{(4.0\ \text{m/s})^2}{2(9.80)} = 0.82\ \text{m}$; b. to the apex $v = 0.41\ \text{s}$ times 2 to the

board = 0.82 s from the board to the water $y = y_0 + v_0 t - \frac{1}{2}gt^2$ $y = -1.80\ \text{m}$ $y_0 = 0$ $v_0 = 4.0\ \text{m/s}$ $-1.8 = 4.0t - 4.9t^2$ $4.9t^2 - 4.0t - 1.80 = 0$, solution to quadratic equation gives 1.13 s; c. $v^2 = v_0^2 - 2g(y - y_0)$ $y_0 = 0$ $v_0 = 4.0\ \text{m/s}$ $y = -1.80\ \text{m}$
$v = 7.16\ \text{m/s}$

70. Time to the apex: $t = 1.12\ \text{s}$ times 2 equals 2.24 s to a height of 2.20 m. To 1.80 m in height is an additional 0.40 m.

$y = y_0 + v_0 t - \frac{1}{2}gt^2$ $y = -0.40\ \text{m}$ $y_0 = 0$ $v_0 = -11.0\ \text{m/s}$

$y = y_0 + v_0 t - \frac{1}{2}gt^2$ $y = -0.40\ \text{m}$ $y_0 = 0$ $v_0 = -11.0\ \text{m/s}$.

$-0.40 = -11.0t - 4.9t^2$ or $4.9t^2 + 11.0t - 0.40 = 0$

Take the positive root, so the time to go the additional 0.4 m is 0.04 s. Total time is $2.24\ \text{s} + 0.04\ \text{s} = 2.28\ \text{s}$.

72. a. $v^2 = v_0^2 - 2g(y - y_0)$ $y_0 = 0$ $v = 0$ $y = 2.50\ \text{m}$
$v_0^2 = 2gy \Rightarrow v_0 = \sqrt{2(9.80)(2.50)} = 7.0\ \text{m/s}$; b. $t = 0.72\ \text{s}$ times 2 gives 1.44 s in the air

74. a. $v = 70.0\ \text{m/s}$; b. time heard after rock begins to fall: 0.75 s, time to reach the ground: 6.09 s

ADDITIONAL PROBLEMS

76. Take west to be the positive direction.
1st plane: $\bar{v} = 600\ \text{km/h}$
2nd plane $\bar{v} = 667.0\ \text{km/h}$

78. $a = \frac{v - v_0}{t - t_0}$, $t = 0$, $a = \frac{-3.4\ \text{cm/s} - v_0}{4\ \text{s}} = 1.2\ \text{cm/s}^2 \Rightarrow v_0 = -8.2\ \text{cm/s}$ $v = v_0 + at = -8.2 + 1.2\,t$;
$v = -7.0\ \text{cm/s}$ $v = -1.0\ \text{cm/s}$

79. $a = -3\ \text{m/s}^2$

81. a.
$v = 8.7 \times 10^5\ \text{m/s}$;
b. $t = 7.8 \times 10^{-8}\ \text{s}$

83. $1\ \text{km} = v_0(80.0\ \text{s}) + \frac{1}{2}a(80.0)^2$; $2\ \text{km} = v_0(200.0) + \frac{1}{2}a(200.0)^2$ solve simultaneously to get $a = -\frac{0.1}{2400.0}\text{km/s}^2$ and $v_0 = 0.014167\ \text{km/s}$, which is $51.0\ \text{km/h}$. Velocity at the end of the trip is $v = 21.0\ \text{km/h}$.

85. $a = -0.9\ \text{m/s}^2$

87. Equation for the speeding car: This car has a constant velocity, which is the average velocity, and is not accelerating, so use the equation for displacement with $x_0 = 0$: $x = x_0 + \bar{v}t = \bar{v}t$; Equation for the police car: This car is accelerating, so use the equation for displacement with $x_0 = 0$ and $v_0 = 0$, since the police car starts from rest: $x = x_0 + v_0 t + \frac{1}{2}at^2 = \frac{1}{2}at^2$; Now we have an equation of motion for each car with a common parameter, which can be eliminated to find the solution. In this case, we solve for t. Step 1, eliminating x: $x = \bar{v}t = \frac{1}{2}at^2$; Step 2, solving for t: $t = \frac{2\bar{v}}{a}$. The speeding car has a constant velocity of 40 m/s, which is its average velocity. The acceleration of the police car is 4 m/s^2. Evaluating *t*, the time for the police car to reach the speeding car, we have $t = \frac{2\bar{v}}{a} = \frac{2(40)}{4} = 20\ \text{s}$.

89. At this acceleration she comes to a full stop in $t = \frac{-v_0}{a} = \frac{8}{0.5} = 16\ \text{s}$, but the distance covered is $x = 8\ \text{m/s}(16\ \text{s}) - \frac{1}{2}(0.5)(16\ \text{s})^2 = 64\ \text{m}$, which is less than the distance she is away from the finish line, so she never finishes the race.

91. $x_1 = \frac{3}{2}v_0 t$

$x_2 = \frac{5}{3}x_1$

93. $v_0 = 7.9\ \text{m/s}$ velocity at the bottom of the window.

$v = 7.9 \text{ m/s}$
$v_0 = 14.1 \text{ m/s}$

95. a. $v = 5.42 \text{ m/s}$;
b. $v = 4.64 \text{ m/s}$;
c. $a = 2874.28 \text{ m/s}^2$;
d. $(x - x_0) = 5.11 \times 10^{-3} \text{ m}$

97. Consider the players fall from rest at the height 1.0 m and 0.3 m.
0.9 s
0.5 s

99. a. $t = 6.37 \text{ s}$ taking the positive root;
b. $v = 59.5 \text{ m/s}$

101. a. $y = 4.9 \text{ m}$;
b. $v = 38.3 \text{ m/s}$;
c. -33.3 m

103. $h = \frac{1}{2}gt^2$, h = total height and time to drop to ground

$\frac{2}{3}h = \frac{1}{2}g(t-1)^2$ in $t - 1$ seconds it drops $2/3h$

$\frac{2}{3}\left(\frac{1}{2}gt^2\right) = \frac{1}{2}g(t-1)^2$ or $\frac{t^2}{3} = \frac{1}{2}(t-1)^2$

$0 = t^2 - 6t + 3 \quad t = \frac{6 \pm \sqrt{6^2 - 4 \cdot 3}}{2} = 3 \pm \frac{\sqrt{24}}{2}$

t = 5.45 s and h = 145.5 m. Other root is less than 1 s. Check for t = 4.45 s $h = \frac{1}{2}gt^2 = 97.0 \text{ m} = \frac{2}{3}(145.5)$

CHALLENGE PROBLEMS

106. $96 \text{ km/h} = 26.67 \text{ m/s}, a = \frac{26.67 \text{ m/s}}{4.0 \text{ s}} = 6.67 \text{m/s}^2$, 295.38 km/h = 82.05 m/s, $t = 12.3 \text{ s}$ time to accelerate to maximum speed

$x = 504.55 \text{ m}$ distance covered during acceleration

7495.44 m at a constant speed

$\frac{7495.44 \text{ m}}{82.05 \text{ m/s}} = 91.35 \text{ s}$ so total time is $91.35 \text{ s} + 12.3 \text{ s} = 103.65 \text{ s}$.

CHAPTER 4

CHECK YOUR UNDERSTANDING

4.1. a. $40.0 \text{ rev/s} = 2\pi(40.0) \text{ rad/s}$, $\bar{\alpha} = \frac{\Delta\omega}{\Delta t} = \frac{2\pi(40.0) - 0 \text{ rad/s}}{20.0 \text{ s}} = 2\pi(2.0) = 4.0\pi \text{ rad/s}^2$; b. Since the angular velocity increases linearly, there has to be a constant acceleration throughout the indicated time. Therefore, the instantaneous angular acceleration at any time is the solution to $4.0\pi \text{ rad/s}^2$.

4.2. a. Using **Equation 4.22**, we have $7000 \text{ rpm} = \frac{7000.0(2\pi \text{ rad})}{60.0 \text{ s}} = 733.0 \text{ rad/s},$

$\alpha = \frac{\omega - \omega_0}{t} = \frac{733.0 \text{ rad/s}}{10.0 \text{ s}} = 73.3 \text{ rad/s}^2$;

b. Using **Equation 4.25**, we have

$$\omega^2 = \omega_0^2 + 2\alpha\Delta\theta \Rightarrow \Delta\theta = \frac{\omega^2 - \omega_0^2}{2\alpha} = \frac{0 - (733.0 \text{ rad/s})^2}{2(73.3 \text{ rad/s}^2)} = 3665.2 \text{ rad}$$

4.3. The angular acceleration is $\alpha = \frac{(5.0 - 0)\text{rad/s}}{20.0 \text{ s}} = 0.25 \text{ rad/s}^2$. Therefore, the total angle that the boy passes through is

$$\Delta\theta = \frac{\omega^2 - \omega_0^2}{2\alpha} = \frac{(5.0)^2 - 0}{2(0.25)} = 50 \text{ rad}.$$

Thus, we calculate

$s = r\theta = 5.0\,\text{m}(50.0\,\text{rad}) = 250.0\,\text{m}$.

CONCEPTUAL QUESTIONS

1. The second hand rotates clockwise, so by the right-hand rule, the angular velocity vector is into the wall.
3. They have the same angular velocity. Points further out on the bat have greater tangential speeds.
4. straight line, linear in time variable
6. constant

PROBLEMS

9. $\omega = \dfrac{2\pi\,\text{rad}}{45.0\,\text{s}} = 0.14\,\text{rad/s}$

11. a. $\theta = \frac{s}{r} = \dfrac{3.0\,\text{m}}{1.5\,\text{m}} = 2.0\,\text{rad}$; b. $\omega = \dfrac{2.0\,\text{rad}}{1.0\,\text{s}} = 2.0\,\text{rad/s}$

13. The propeller takes only $\Delta t = \dfrac{\Delta\omega}{\alpha} = \dfrac{0\,\text{rad/s} - 10.0(2\pi)\,\text{rad/s}}{-2.0\,\text{rad/s}^2} = 31.4\,\text{s}$ to come to rest, when the propeller is at 0 rad/s, it would start rotating in the opposite direction. This would be impossible due to the magnitude of forces involved in getting the propeller to stop and start rotating in the opposite direction.

15. a. $\omega = 25.0(2.0\,\text{s}) = 50.0\,\text{rad/s}$; b. $\alpha = \dfrac{d\omega}{dt} = 25.0\,\text{rad/s}^2$

16. a. $\omega = 54.8\,\text{rad/s}$;
b. $t = 11.0\,\text{s}$

18. a. $0.87\,\text{rad/s}^2$;
b. $\theta = 66{,}264\,\text{rad}$

20. a. $\omega = 42.0\,\text{rad/s}$;

b. $\theta = 200\,\text{rad}$; c. $\begin{matrix} v_t = 42\,\text{m/s} \\ a_t = 4.0\,\text{m/s}^2 \end{matrix}$

22. a. $\omega = 7.0\,\text{rad/s}$;
b. $\theta = 22.5\,\text{rad}$; c. $a_\text{t} = 0.1\,\text{m/s}$

24. $\alpha = 28.6\,\text{rad/s}^2$.

27. (a) 62.8 m/s
(b) $\alpha = -0.314\,\text{rad/s}^2$

CHAPTER 5

CHECK YOUR UNDERSTANDING

5.1. a. not equal because they are orthogonal; b. not equal because they have different magnitudes; c. not equal because they have different magnitudes and directions; d. not equal because they are antiparallel; e. equal.

5.2. 16 m; $\vec{\mathbf{D}} = -16\,\text{m}\,\hat{\mathbf{u}}$

5.3. G = 28.2 cm, $\theta_G = 291°$

5.4. $\vec{\mathbf{D}} = (-5.0\,\hat{\mathbf{i}} - 3.0\,\hat{\mathbf{j}})\text{cm}$; the fly moved 5.0 cm to the left and 3.0 cm down from its landing site.

5.5. 5.83 cm, $211°$

5.6. $\vec{\mathbf{D}} = (-20\,\text{m})\,\hat{\mathbf{j}}$

5.7. 35.1 m/s = 126.4 km/h

5.8. $\vec{\mathbf{G}} = (10.25\,\hat{\mathbf{i}} - 26.22\,\hat{\mathbf{j}})\text{cm}$

5.9. D = 55.7 N; direction $65.7°$ north of east

5.10. $\hat{\mathbf{v}} = 0.8\,\hat{\mathbf{i}} + 0.6\,\hat{\mathbf{j}}$, $36.87°$ north of east

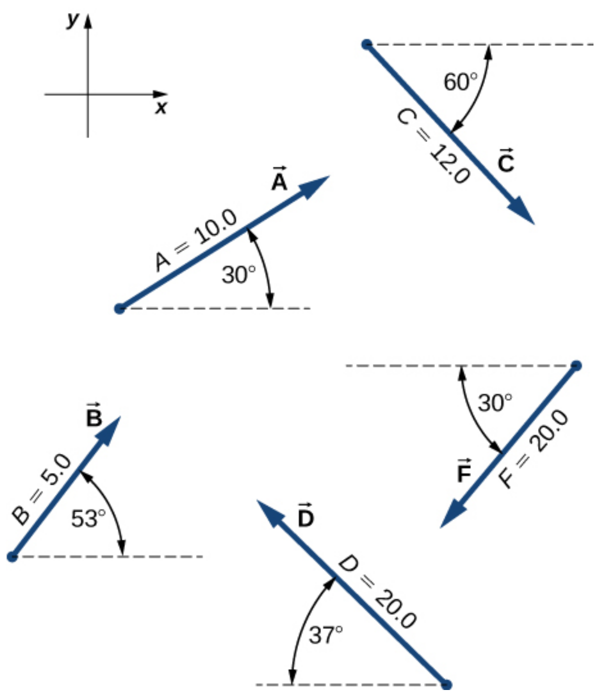
CONCEPTUAL QUESTIONS

1. scalar
3. answers may vary

5. parallel, sum of magnitudes, antiparallel, zero
7. no, yes
9. zero, yes
11. no
13. equal, equal, the same
15. a unit vector of the *x*-axis
17. They are equal.
19. yes

PROBLEMS

21. $\vec{\mathbf{h}} = -49\,\text{m}\,\hat{\mathbf{u}}$, 49 m
23. 30.8 m, $35.7°$ west of north
25. 134 km, $80°$
27. 7.34 km, $63.5°$ south of east
29. 3.8 km east, 3.2 km north, 7.0 km
31. 14.3 km, $65°$
33. a. $\vec{\mathbf{A}} = +8.66\,\hat{\mathbf{i}} + 5.00\,\hat{\mathbf{j}}$, b. $\vec{\mathbf{B}} = +30.09\,\hat{\mathbf{i}} + 39.93\,\hat{\mathbf{j}}$, c. $\vec{\mathbf{C}} = +6.00\,\hat{\mathbf{i}} - 10.39\,\hat{\mathbf{j}}$, d. $\vec{\mathbf{D}} = -15.97\,\hat{\mathbf{i}} + 12.04\,\hat{\mathbf{j}}$, f. $\vec{\mathbf{F}} = -17.32\,\hat{\mathbf{i}} - 10.00\,\hat{\mathbf{j}}$



35. a. 1.94 km, 7.24 km; b. proof
37. 3.8 km east, 3.2 km north, 2.0 km, $\vec{\mathbf{D}} = (3.8\,\hat{\mathbf{i}} + 3.2\,\hat{\mathbf{j}})\text{km}$
39. $P_1(2.165\text{ m},\ 1.250\text{ m})$, $P_2(-1.900\text{ m},\ 3.290\text{ m})$, 5.27 m
41. 8.60 m, $A(2\sqrt{5}\text{ m},\ 0.647\pi)$, $B(3\sqrt{2}\text{ m},\ 0.75\pi)$
43. a. $\vec{\mathbf{A}} + \vec{\mathbf{B}} = -4\,\hat{\mathbf{i}} - 6\,\hat{\mathbf{j}}$, $\left|\vec{\mathbf{A}} + \vec{\mathbf{B}}\right| = 7.211,\ \theta = 213.7°$; b. $\vec{\mathbf{A}} - \vec{\mathbf{B}} = 2\,\hat{\mathbf{i}} - 2\,\hat{\mathbf{j}}$, $\left|\vec{\mathbf{A}} - \vec{\mathbf{B}}\right| = 2\sqrt{2},\ \theta = -45°$
45. a. $\vec{\mathbf{C}} = (5.0\,\hat{\mathbf{i}} - 1.0\,\hat{\mathbf{j}} - 3.0\,\hat{\mathbf{k}})\text{m},\ C = 5.92\text{ m}$;
b. $\vec{\mathbf{D}} = (4.0\,\hat{\mathbf{i}} - 11.0\,\hat{\mathbf{j}} + 15.0\,\hat{\mathbf{k}})\text{m},\ D = 19.03\text{ m}$

47. $\vec{\mathbf{D}} = (3.3\,\hat{\mathbf{i}} - 6.6\,\hat{\mathbf{j}})\text{km}$, $\hat{\mathbf{i}}$ is to the east, 7.34 km, $-63.5°$

49. a. $\vec{\mathbf{R}} = -1.35\,\hat{\mathbf{i}} - 22.04\,\hat{\mathbf{j}}$, b. $\vec{\mathbf{R}} = -17.98\,\hat{\mathbf{i}} + 0.89\,\hat{\mathbf{j}}$

51. $\vec{\mathbf{D}} = (200\,\hat{\mathbf{i}} + 300\,\hat{\mathbf{j}})\text{yd}$, D = 360.5 yd, $56.3°$ north of east; The numerical answers would stay the same but the physical unit would be meters. The physical meaning and distances would be about the same because 1 yd is comparable with 1 m.

53. $\vec{\mathbf{R}} = -3\,\hat{\mathbf{i}} - 16\,\hat{\mathbf{j}}$

55. $\vec{\mathbf{E}} = E\hat{\mathbf{E}}$, $E_x = +178.9\text{V/m}$, $E_y = -357.8\text{V/m}$, $E_z = 0.0\text{V/m}$, $\theta_E = -\tan^{-1}(2)$

57. a. $\vec{\mathbf{R}}_B = (12.278\,\hat{\mathbf{i}} + 7.089\,\hat{\mathbf{j}} + 2.500\,\hat{\mathbf{k}})\text{km}$, $\vec{\mathbf{R}}_D = (-0.262\,\hat{\mathbf{i}} + 3.000\,\hat{\mathbf{k}})\text{km}$; b. $\left|\vec{\mathbf{R}}_B - \vec{\mathbf{R}}_D\right| = 14.414\text{ km}$

ADDITIONAL PROBLEMS

58. a. 18.4 km and 26.2 km, b. 31.5 km and 5.56 km
60. a. $(r, \varphi + \pi/2)$, b. $(2r, \varphi + 2\pi)$, (c) $(3r, -\varphi)$
62. $d_{\text{PM}} = 33.12\text{ nmi} = 61.34\text{ km}, \quad d_{\text{NP}} = 35.47\text{ nmi} = 65.69\text{ km}$
64. proof
66. a. 10.00 m, b. 5π m, c. 0
68. 22.2 km/h, $35.8°$ south of west
70. 240.2 m, $2.2°$ south of west

CHAPTER 6

CHECK YOUR UNDERSTANDING

6.1. (a) Taking the derivative with respect to time of the position function, we have $\vec{\mathbf{v}}(t) = 9.0t^2\,\hat{\mathbf{i}}$ and $\vec{\mathbf{v}}(3.0\text{s}) = 81.0\,\hat{\mathbf{i}}\text{ m/s}.$ (b) Since the velocity function is nonlinear, we suspect the average velocity is not equal to the instantaneous velocity. We check this and find

$$\vec{\mathbf{v}}_{\text{avg}} = \frac{\vec{\mathbf{r}}(t_2) - \vec{\mathbf{r}}(t_1)}{t_2 - t_1} = \frac{\vec{\mathbf{r}}(4.0\text{ s}) - \vec{\mathbf{r}}(2.0\text{ s})}{4.0\text{ s} - 2.0\text{ s}} = \frac{(144.0\,\hat{\mathbf{i}} - 36.0\,\hat{\mathbf{i}})\text{ m}}{2.0\text{ s}} = 54.0\,\hat{\mathbf{i}}\text{ m/s},$$

which is different from $\vec{\mathbf{v}}(3.0\text{s}) = 81.0\,\hat{\mathbf{i}}\text{ m/s}.$

6.2. (a) Choose the top of the cliff where the rock is thrown from the origin of the coordinate system. Although it is arbitrary, we typically choose time t = 0 to correspond to the origin. (b) The equation that describes the horizontal motion is $x = x_0 + v_x t.$ With $x_0 = 0,$ this equation becomes $x = v_x t.$ (c) m58290 (https://legacy.cnx.org/content/m58290/latest/#fs-id1165038300036) through m58290 (https://legacy.cnx.org/content/m58290/latest/#fs-id1165038323452) and Equation 6.42 describe the vertical motion, but since $y_0 = 0$ and $v_{0y} = 0,$ these equations simplify greatly to become $y = \frac{1}{2}(v_{0y} + v_y)t = \frac{1}{2}v_y t, \quad v_y = -gt, \quad y = -\frac{1}{2}gt^2,$ and $v_y^2 = -2gy.$ (d) We use the kinematic equations to find the x and y components of the velocity at the point of impact. Using $v_y^2 = -2gy$ and noting the point of impact is −100.0 m, we find the y component of the velocity at impact is $v_y = 44.3\text{ m/s}.$ We are given the x component, $v_x = 15.0\text{ m/s},$ so we can calculate the total velocity at impact: v = 46.8 m/s and $\theta = 71.3°$ below the horizontal.

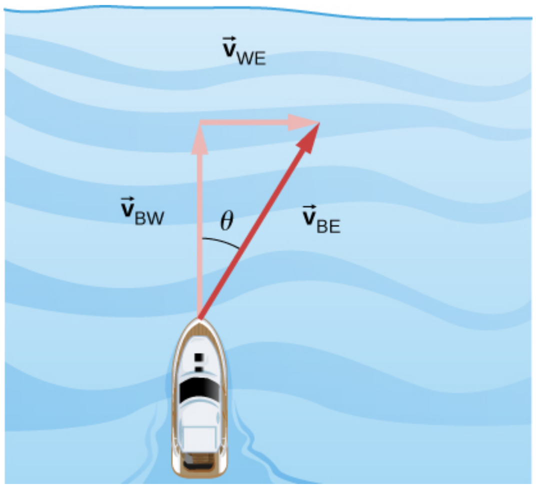
6.3. The golf shot at $30°$.

6.4. 134.0 cm/s

6.5. Labeling subscripts for the vector equation, we have B = boat, R = river, and E = Earth. The vector equation becomes $\vec{\mathbf{v}}_{\text{BE}} = \vec{\mathbf{v}}_{\text{BR}} + \vec{\mathbf{v}}_{\text{RE}}$. We have right triangle geometry shown in Figure 04_05_BoatRiv_img. Solving for $\vec{\mathbf{v}}_{\text{BE}}$, we have

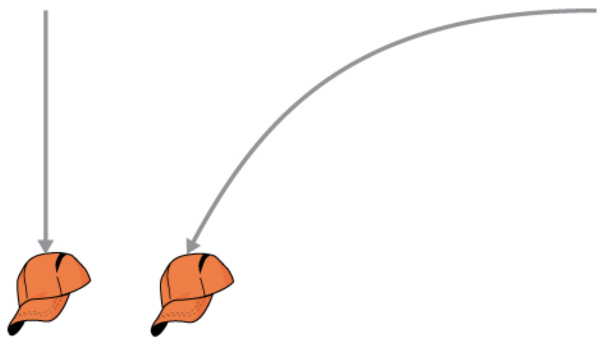
$$v_{\text{BE}} = \sqrt{v_{\text{BR}}^2 + v_{\text{RE}}^2} = \sqrt{4.5^2 + 3.0^2}$$

$$v_{\text{BE}} = 5.4\text{ m/s}, \quad \theta = \tan^{-1}\left(\frac{3.0}{4.5}\right) = 33.7°.$$



CONCEPTUAL QUESTIONS

1. straight line
3. The slope must be zero because the velocity vector is tangent to the graph of the position function.
5. No, motions in perpendicular directions are independent.
7. a. no; b. minimum at apex of trajectory and maximum at launch and impact; c. no, velocity is a vector; d. yes, where it lands
9. They both hit the ground at the same time.
11. yes
13. If he is going to pass the ball to another player, he needs to keep his eyes on the reference frame in which the other players on the team are located.



15.

PROBLEMS

17. $\vec{r} = 1.0\hat{i} - 4.0\hat{j} + 6.0\hat{k}$

19. $\Delta\vec{r}_{\text{Total}} = 472.0\text{ m}\hat{i} + 80.3\text{ m}\hat{j}$

21. Sum of displacements $= -6.4\text{ km}\hat{i} + 9.4\text{ km}\hat{j}$

23. a. $\vec{v}(t) = 8.0t\hat{i} + 6.0t^2\hat{k}$, $\vec{v}(0) = 0$, $\vec{v}(1.0) = 8.0\hat{i} + 6.0\hat{k}\text{ m/s}$,

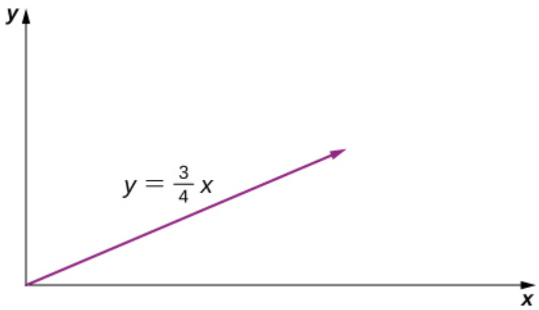
b. $\vec{v}_{\text{avg}} = 4.0\hat{i} + 2.0\hat{k}\text{ m/s}$

25. $\Delta\vec{r}_1 = 20.00\text{ m}\hat{j}$, $\Delta\vec{r}_2 = (2.000 \times 10^4\text{ m})(\cos 30°\hat{i} + \sin 30°\hat{j})$

$\Delta\vec{r} = 1.700 \times 10^4\text{ m}\hat{i} + 1.002 \times 10^4\text{ m}\hat{j}$

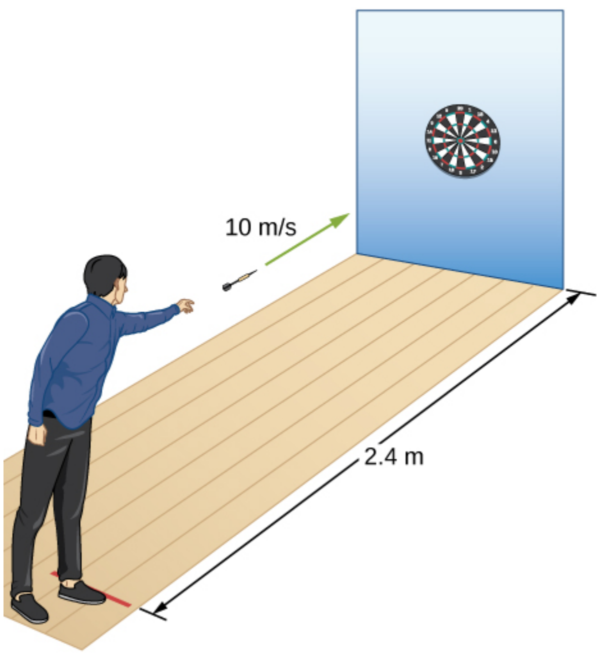
26. a. $\vec{\mathbf{v}}(t) = (4.0t\,\hat{\mathbf{i}} + 3.0t\,\hat{\mathbf{j}})\text{m/s}, \quad \vec{\mathbf{r}}(t) = (2.0t^2\,\hat{\mathbf{i}} + \frac{3}{2}t^2\,\hat{\mathbf{j}})\,\text{m}$,

b. $x(t) = 2.0t^2\,\text{m}, y(t) = \frac{3}{2}t^2\,\text{m}, t^2 = \frac{x}{2} \Rightarrow y = \frac{3}{4}x$



30. a. $t = 0.55\,\text{s}$, b. $x = 110\,\text{m}$

32. a. $t = 0.24\text{s}, \; d = 0.28\,\text{m}$, b. They aim high.



34. a., $t = 12.8\,\text{s}, \; x = 5619\,\text{m}$ b. $v_y = 125.0\,\text{m/s}, \; v_x = 439.0\,\text{m/s}, \; |\vec{v}| = 456.0\,\text{m/s}$

36. a. $v_y = v_{0y} - gt, \quad t = 10\text{s}, \; v_y = 0, \quad v_{0y} = 98.0\,\text{m/s}, \quad v_0 = 196.0\,\text{m/s}$, b. $h = 490.0\,\text{m}$,

c. $v_{0x} = 169.7\,\text{m/s}, \; x = 3394.0\,\text{m}$,

d. $$x = 2545.5\,\text{m}$$
$$y = 465.5\,\text{m}$$
$$\vec{\mathbf{s}} = 2545.5\,\text{m}\,\hat{\mathbf{i}} + 465.5\,\text{m}\,\hat{\mathbf{j}}$$

38. $-100\,\text{m} = (-2.0\,\text{m/s})t - (4.9\,\text{m/s}^2)t^2, \quad t = 4.3\,\text{s}, \; x = 86.0\,\text{m}$

40. $R_{Moon} = 48\,\text{m}$

42. a. $v_{0y} = 24\,\text{m/s} \; v_y^2 = v_{0y}^2 - 2gy \Rightarrow h = 23.4\,\text{m}$,

b. $t = 3\,\text{s} \; v_{0x} = 18\,\text{m/s} \; x = 54\,\text{m}$,

c. $y = -100 \text{ m}$ $y_0 = 0$ $y - y_0 = v_{0y}t - \frac{1}{2}gt^2$ $-100 = 24t - 4.9t^2$ $\Rightarrow t = 7.58 \text{ s}$,

d. $x = 136.44 \text{ m}$,

e. $t = 2.0 \text{ s}$ $y = 28.4 \text{ m}$ $x = 36 \text{ m}$

$t = 4.0 \text{ s}$ $y = 17.6 \text{ m}$ $x = 22.4 \text{ m}$

$t = 6.0 \text{ s}$ $y = -32.4 \text{ m}$ $x = 108 \text{ m}$

44. $v_{0y} = 12.9 \text{ m/s}$ $y - y_0 = v_{0y}t - \frac{1}{2}gt^2$ $-20.0 = 12.9t - 4.9t^2$

$t = 3.7 \text{ s}$ $v_{0x} = 15.3 \text{ m/s} \Rightarrow x = 56.7 \text{ m}$

So the golfer's shot lands 13.3 m short of the green.

46. a. $R = 60.8 \text{ m}$,

b. $R = 137.8 \text{ m}$

48. a. $v_y^2 = v_{0y}^2 - 2gy \Rightarrow y = 2.9 \text{ m/s}$

$y = 3.3 \text{ m/s}$

$y = \frac{v_{0y}^2}{2g} = \frac{(v_0 \sin\theta)^2}{2g} \Rightarrow \sin\theta = 0.91 \Rightarrow \theta = 65.5°$

50. $R = 18.5 \text{ m}$

52. $y = (\tan\theta_0)x - \left[\frac{g}{2(v_0 \cos\theta_0)^2}\right]x^2 \Rightarrow v_0 = 16.4 \text{ m/s}$

54. $R = \frac{v_0^2 \sin 2\theta_0}{g} \Rightarrow \theta_0 = 15.0°$

56. It takes the wide receiver 1.1 s to cover the last 10 m of his run.

$T_{\text{tof}} = \frac{2(v_0 \sin\theta)}{g} \Rightarrow \sin\theta = 0.27 \Rightarrow \theta = 15.6°$

58. $a_C = 40 \text{ m/s}^2$

60. $a_C = \frac{v^2}{r} \Rightarrow v^2 = r$ $a_C = 78.4$, $v = 8.85 \text{ m/s}$

$T = 5.68 \text{ s}$, which is $0.176 \text{ rev/s} = 10.6 \text{ rev/min}$

62. Venus is 108.2 million km from the Sun and has an orbital period of 0.6152 y.

$r = 1.082 \times 10^{11} \text{ m}$ $T = 1.94 \times 10^7 \text{ s}$

$v = 3.5 \times 10^4 \text{ m/s}$, $a_C = 1.135 \times 10^{-2} \text{ m/s}^2$

64. $360 \text{ rev/min} = 6 \text{ rev/s}$

$v = 3.8 \text{ m/s}$ $a_C = 144. \text{ m/s}^2$

66. a. $O'(t) = (4.0\hat{\mathbf{i}} + 3.0\hat{\mathbf{j}} + 5.0\hat{\mathbf{k}})t \text{ m}$,

b. $\vec{\mathbf{r}}_{PS} = \vec{\mathbf{r}}_{PS'} + \vec{\mathbf{r}}_{S'S}$, $\vec{\mathbf{r}}(t) = \vec{\mathbf{r}}'(t) + (4.0\hat{\mathbf{i}} + 3.0\hat{\mathbf{j}} + 5.0\hat{\mathbf{k}})t \text{ m}$,

c. $\vec{\mathbf{v}}(t) = \vec{\mathbf{v}}'(t) + (4.0\hat{\mathbf{i}} + 3.0\hat{\mathbf{j}} + 5.0\hat{\mathbf{k}}) \text{ m/s}$, d. The accelerations are the same.

68. $\vec{\mathbf{v}}_{PC} = (2.0\hat{\mathbf{i}} + 5.0\hat{\mathbf{j}} + 4.0\hat{\mathbf{k}}) \text{m/s}$

70. a. A = air, S = seagull, G = ground

$\vec{\mathbf{v}}_{\text{SA}} = 9.0 \text{ m/s}$ velocity of seagull with respect to still air

$\vec{\mathbf{v}}_{\text{AG}} = ?$ $\vec{\mathbf{v}}_{\text{SG}} = 5 \text{ m/s}$ $\vec{\mathbf{v}}_{\text{SG}} = \vec{\mathbf{v}}_{\text{SA}} + \vec{\mathbf{v}}_{\text{AG}} \Rightarrow \vec{\mathbf{v}}_{\text{AG}} = \vec{\mathbf{v}}_{\text{SG}} - \vec{\mathbf{v}}_{\text{SA}}$

$\vec{\mathbf{v}}_{\text{AG}} = -4.0 \text{ m/s}$

b. $\vec{\mathbf{v}}_{\text{SG}} = \vec{\mathbf{v}}_{\text{SA}} + \vec{\mathbf{v}}_{\text{AG}} \Rightarrow \vec{\mathbf{v}}_{\text{SG}} = -13.0 \text{ m/s}$

$\frac{-6000 \text{ m}}{-13.0 \text{ m/s}} = 7 \text{ min } 42 \text{ s}$

72. Take the positive direction to be the same direction that the river is flowing, which is east. S = shore/Earth, W = water, and B = boat.

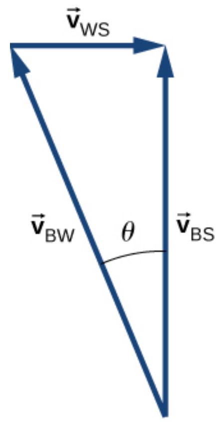
a. $\vec{v}_{BS} = 11\text{ km/h}$

$t = 8.2\text{ min}$

b. $\vec{v}_{BS} = -5\text{ km/h}$

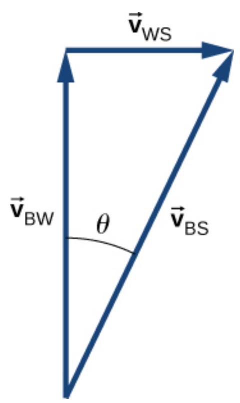
$t = 18\text{ min}$

c. $\vec{v}_{BS} = \vec{v}_{BW} + \vec{v}_{WS}$ $\theta = 22°$ west of north



d. $|\vec{v}_{BS}| = 7.4\text{ km/h}$ $t = 6.5\text{ min}$

e. $\vec{v}_{BS} = 8.54\text{ km/h},$ but only the component of the velocity straight across the river is used to get the time



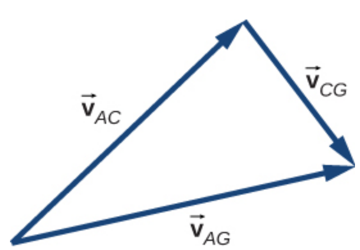
$t = 6.0\text{ min}$

Downstream = 0.3 km

74. $\vec{v}_{AG} = \vec{v}_{AC} + \vec{v}_{CG}$

$|\vec{v}_{AC}| = 25\text{ km/h}$ $|\vec{v}_{CG}| = 15\text{ km/h}$ $|\vec{v}_{AG}| = 29.15\text{ km/h}$ $\vec{v}_{AG} = \vec{v}_{AC} + \vec{v}_{CG}$

The angle between $\vec{v}_{AC}$ and $\vec{v}_{AG}$ is $31°,$ so the direction of the wind is $14°$ north of east.



ADDITIONAL PROBLEMS

76. $a_C = 39.6\text{ m/s}^2$

78. $90.0\text{ km/h} = 25.0\text{ m/s},\ 9.0\text{ km/h} = 2.5\text{ m/s},\quad 60.0\text{ km/h} = 16.7\text{ m/s}$

$a_T = -2.5\text{ m/s}^2,\ a_C = 1.86\text{ m/s}^2,\ a = 3.1\text{ m/s}^2$

80. The radius of the circle of revolution at latitude λ is $R_E \cos\lambda$. The velocity of the body is $\frac{2\pi r}{T}$. $a_C = \frac{4\pi^2 R_E \cos\lambda}{T^2}$ for $\lambda = 40°,\ a_C = 0.26\%\ g$

82. $a_T = 3.00\text{ m/s}^2$

$v(5\text{ s}) = 15.00\text{ m/s}\ \ a_C = 150.00\text{ m/s}^2\ \ \theta = 88.8°$ with respect to the tangent to the circle of revolution directed inward. $|\vec{\mathbf{a}}| = 150.03\text{ m/s}^2$

84. $\vec{\mathbf{a}}(t) = -A\omega^2 \cos\omega t\,\hat{\mathbf{i}} - A\omega^2 \sin\omega t\,\hat{\mathbf{j}}$

$a_C = 5.0\text{ m}\omega^2\ \ \omega = 0.89\text{ rad/s}$

$\vec{\mathbf{v}}(t) = -2.24\text{ m/s}\,\hat{\mathbf{i}} - 3.87\text{ m/s}\,\hat{\mathbf{j}}$

86. $\vec{\mathbf{r}}_1 = 1.5\hat{\mathbf{j}} + 4.0\hat{\mathbf{k}}\quad \vec{\mathbf{r}}_2 = \Delta\vec{\mathbf{r}} + \vec{\mathbf{r}}_1 = 2.5\hat{\mathbf{i}} + 4.7\hat{\mathbf{j}} + 2.8\hat{\mathbf{k}}$

88. $v_x(t) = 265.0\text{ m/s}$

$v_y(t) = 20.0\text{ m/s}$

$\vec{\mathbf{v}}(5.0\text{ s}) = (265.0\hat{\mathbf{i}} + 20.0\hat{\mathbf{j}})\text{m/s}$

90. $R = 1.07\text{ m}$

92. $v_0 = 20.1\text{ m/s}$

94. $v = 3072.5\text{ m/s}$

$a_C = 0.223\text{ m/s}^2$

CHALLENGE PROBLEMS

96. a. $-400.0\text{ m} = v_{0y}t - 4.9t^2\ \ 359.0\text{ m} = v_{0x}t\ \ t = \frac{359.0}{v_{0x}}\ \ -400.0 = 359.0\frac{v_{0y}}{v_{0x}} - 4.9\left(\frac{359.0}{v_{0x}}\right)^2$

$-400.0 = 359.0\tan 40 - \frac{631{,}516.9}{v_{0x}^2} \Rightarrow v_{0x}^2 = 900.6\ \ v_{0x} = 30.0\text{ m/s}\ \ v_{0y} = v_{0x}\tan 40 = 25.2\text{ m/s}$

$v = 39.2\text{ m/s}$, b. $t = 12.0\text{ s}$

98. a. $\vec{\mathbf{r}}_{TC} = (-32 + 80t)\hat{\mathbf{i}} + 50t\hat{\mathbf{j}},\ \ |\vec{\mathbf{r}}_{TC}|^2 = (-32 + 80t)^2 + (50t)^2$

$2r\frac{dr}{dt} = 2(-32 + 80t) + 100t\ \ \frac{dr}{dt} = \frac{2(-32 + 80t) + 100t}{2r} = 0$

$260t = 64 \Rightarrow t = 15\text{ min}$,

b. $|\vec{\mathbf{r}}_{TC}| = 17\text{ km}$

CHAPTER 7

CHECK YOUR UNDERSTANDING

7.1. For a sphere,

$$\text{density} = \frac{\text{mass}}{\left(\frac{4}{3}\pi R^3\right)}\ \text{kg/m}^3.$$

For Earth, then,

$$\text{density} = \frac{6\times 10^{24}\ \text{kg}}{4.2\times 2.6\times 10^{20}\ \text{m}^3} = 5.5\times 10^3\ \text{kg/m}^3.$$

This density is four to five times greater than Mimas'. In fact, Earth is the densest of the planets.

7.2. The semi-major axis for the highly elliptical orbit of Halley's comet is 17.8 AU and is the average of the perihelion and aphelion. This lies between the 9.5 AU and 19 AU orbital radii for Saturn and Uranus, respectively. The radius for a circular orbit is the same as the semi-major axis, and since the period increases with an increase of the semi-major axis, the fact that Halley's period is between the periods of Saturn and Uranus is expected.

CONCEPTUAL QUESTIONS

31. The speed is greatest where the satellite is closest to the large mass and least where farther away—at the periapsis and apoapsis, respectively. It is conservation of angular momentum that governs this relationship. But it can also be gleaned from conservation of energy, the kinetic energy must be greatest where the gravitational potential energy is the least (most negative). The force, and hence acceleration, is always directed towards M in the diagram, and the velocity is always tangent to the path at all points. The acceleration vector has a tangential component along the direction of the velocity at the upper location on the y-axis; hence, the satellite is speeding up. Just the opposite is true at the lower position.

PROBLEMS

37. 1.98×10^{30} kg ; The values are the same within 0.05%.

39. Compare **m58348 (https://legacy.cnx.org/content/m58348/latest/#fs-id1168327874347)** and **???** to see that they differ only in that the circular radius, r, is replaced by the semi-major axis, a. Therefore, the mean radius is one-half the sum of the aphelion and perihelion, the same as the semi-major axis.

41. The semi-major axis, 3.78 AU is found from the equation for the period. This is one-half the sum of the aphelion and perihelion, giving an aphelion distance of 4.95 AU.

43. 1.75 years

CHAPTER 8

CHECK YOUR UNDERSTANDING

8.1. 14 N, $56°$ measured from the positive x-axis

8.2. a. His weight acts downward, and the force of air resistance with the parachute acts upward. b. neither; the forces are equal in magnitude

8.3. 0.1 m/s^2

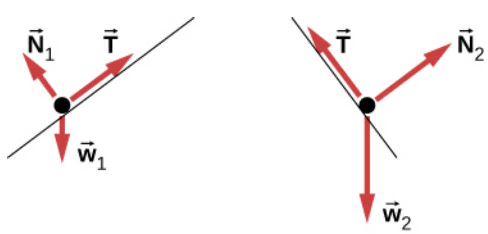
8.4. 40 m/s^2

8.5. $a = 2.78 \text{ m/s}^2$

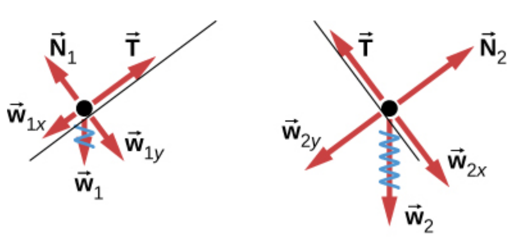
8.6. a. 3.0 m/s^2 ; b. 18 N

8.7. a. 1.7 m/s^2; b. 1.3 m/s^2

8.8.



;



8.9. a. 4.9 N; b. 0.98 m/s^2

8.10. -0.23 m/s^2 ; the negative sign indicates that the snowboarder is slowing down.

8.11. 0.40

8.12. The force of gravity on each object increases with the square of the inverse distance as they fall together, and hence so does the acceleration. For example, if the distance is halved, the force and acceleration are quadrupled. Our average is accurate only

for a linearly increasing acceleration, whereas the acceleration actually increases at a greater rate. So our calculated speed is too small. From Newton's third law (action-reaction forces), the force of gravity between any two objects must be the same. But the accelerations will not be if they have different masses.
8.13. The tallest buildings in the world are all less than 1 km. Since *g* is proportional to the distance squared from Earth's center, a simple ratio shows that the change in *g* at 1 km above Earth's surface is less than 0.0001%. There would be no need to consider this in structural design.
8.14. In **Equation 8.98**, the radius appears in the denominator inside the square root. So the radius must increase by a factor of 4, to decrease the orbital velocity by a factor of 2. The circumference of the orbit has also increased by this factor of 4, and so with half the orbital velocity, the period must be 8 times longer. That can also be seen directly from **Equation 8.99**.
8.15. The assumption is that orbiting object is much less massive than the body it is orbiting. This is not really justified in the case of the Moon and Earth. Both Earth and the Moon orbit about their common center of mass. We tackle this issue in the next example.
8.16. The stars on the "inside" of each galaxy will be closer to the other galaxy and hence will feel a greater gravitational force than those on the outside. Consequently, they will have a greater acceleration. Even without this force difference, the inside stars would be orbiting at a smaller radius, and, hence, there would develop an elongation or stretching of each galaxy. The force difference only increases this effect.

CONCEPTUAL QUESTIONS

1. Forces are directional and have magnitude.
3. The cupcake velocity before the braking action was the same as that of the car. Therefore, the cupcakes were unrestricted bodies in motion, and when the car suddenly stopped, the cupcakes kept moving forward according to Newton's first law.
5. No. If the force were zero at this point, then there would be nothing to change the object's momentary zero velocity. Since we do not observe the object hanging motionless in the air, the force could not be zero.
7. The astronaut is truly weightless in the location described, because there is no large body (planet or star) nearby to exert a gravitational force. Her mass is 70 kg regardless of where she is located.
9. The force you exert (a contact force equal in magnitude to your weight) is small. Earth is extremely massive by comparison. Thus, the acceleration of Earth would be incredibly small. To see this, use Newton's second law to calculate the acceleration you would cause if your weight is 600.0 N and the mass of Earth is 6.00×10^{24} kg.
11. a. action: Earth pulls on the Moon, reaction: Moon pulls on Earth; b. action: foot applies force to ball, reaction: ball applies force to foot; c. action: rocket pushes on gas, reaction: gas pushes back on rocket; d. action: car tires push backward on road, reaction: road pushes forward on tires; e. action: jumper pushes down on ground, reaction: ground pushes up on jumper; f. action: gun pushes forward on bullet, reaction: bullet pushes backward on gun.
13. a. The rifle (the shell supported by the rifle) exerts a force to expel the bullet; the reaction to this force is the force that the bullet exerts on the rifle (shell) in opposite direction. b. In a recoilless rifle, the shell is not secured in the rifle; hence, as the bullet is pushed to move forward, the shell is pushed to eject from the opposite end of the barrel. c. It is not safe to stand behind a recoilless rifle.
15. a. Yes, the force can be acting to the left; the particle would experience deceleration and lose speed. B. Yes, the force can be acting downward because its weight acts downward even as it moves to the right.
17. two forces of different types: weight acting downward and normal force acting upward
20. If you do not let up on the brake pedal, the car's wheels will lock so that they are not rolling; sliding friction is now involved and the sudden change (due to the larger force of static friction) causes the jerk.
22. 5.00 N
24. Centripetal force is defined as any net force causing uniform circular motion. The centripetal force is not a new kind of force. The label "centripetal" refers to *any* force that keeps something turning in a circle. That force could be tension, gravity, friction, electrical attraction, the normal force, or any other force. Any combination of these could be the source of centripetal force, for example, the centripetal force at the top of the path of a tetherball swung through a vertical circle is the result of both tension and gravity.
26. The driver who cuts the corner (on Path 2) has a more gradual curve, with a larger radius. That one will be the better racing line. If the driver goes too fast around a corner using a racing line, he will still slide off the track; the key is to stay at the maximum value of static friction. So, the driver wants maximum possible speed and maximum friction. Consider the equation for centripetal force: $F_c = m\frac{v^2}{r}$ where *v* is speed and *r* is the radius of curvature. So by decreasing the curvature (1/*r*) of the path that the car takes, we reduce the amount of force the tires have to exert on the road, meaning we can now increase the speed, *v*. Looking at this from the point of view of the driver on Path 1, we can reason this way: the sharper the turn, the smaller the turning circle; the smaller the turning circle, the larger is the required centripetal force. If this centripetal force is not exerted, the result is a skid.
28. The barrel of the dryer provides a centripetal force on the clothes (including the water droplets) to keep them moving in a circular path. As a water droplet comes to one of the holes in the barrel, it will move in a path tangent to the circle.
30. Since the radial friction with the tires supplies the centripetal force, and friction is nearly 0 when the car encounters the ice, the car will obey Newton's first law and go off the road in a straight line path, tangent to the curve. A common misconception is that the car will follow a curved path off the road.
31. Anna is correct. The satellite is freely falling toward Earth due to gravity, even though gravity is weaker at the altitude of the

satellite, and g is not $9.80\ \text{m/s}^2$. Free fall does not depend on the value of g; that is, you could experience free fall on Mars if you jumped off Olympus Mons (the tallest volcano in the solar system).

32. The ultimate truth is experimental verification. Field theory was developed to help explain how force is exerted without objects being in contact for both gravity and electromagnetic forces that act at the speed of light. It has only been since the twentieth century that we have been able to measure that the force is not conveyed immediately.

35. The period of the orbit must be 24 hours. But in addition, the satellite must be located in an equatorial orbit and orbiting in the same direction as Earth's rotation. All three criteria must be met for the satellite to remain in one position relative to Earth's surface. At least three satellites are needed, as two on opposite sides of Earth cannot communicate with each other. (This is not technically true, as a wavelength could be chosen that provides sufficient diffraction. But it would be totally impractical.)

PROBLEMS

36. a. $\vec{\mathbf{F}}_{\text{net}} = 5.0\,\hat{\mathbf{i}} + 10.0\,\hat{\mathbf{j}}\ \text{N}$; b. the magnitude is $F_{\text{net}} = 11\ \text{N}$, and the direction is $\theta = 63°$

38. a. $\vec{\mathbf{F}}_{\text{net}} = 660.0\,\hat{\mathbf{i}} + 150.0\,\hat{\mathbf{j}}\ \text{N}$; b. $F_{\text{net}} = 676.6\ \text{N}$ at $\theta = 12.8°$ from David's rope

40. a. $\vec{\mathbf{F}}_{\text{net}} = 95.0\,\hat{\mathbf{i}} + 283\,\hat{\mathbf{j}}\ \text{N}$; b. 299 N at $71°$ north of east; c. $\vec{\mathbf{F}}_{\text{DS}} = -\left(95.0\,\hat{\mathbf{i}} + 283\,\hat{\mathbf{j}}\right)\text{N}$

42. Running from rest, the sprinter attains a velocity of $v = 12.96\ \text{m/s}$, at end of acceleration. We find the time for acceleration using $x = 20.00\ \text{m} = 0 + 0.5at_1^{\ 2}$, or $t_1 = 3.086\ \text{s}$. For maintained velocity, $x_2 = vt_2$, or $t_2 = x_2/v = 80.00\ \text{m}/12.96\ \text{m/s} = 6.173\ \text{s}$. Total time $= 9.259\ \text{s}$.

44. a. $m = 56.0\ \text{kg}$; b. $a_{\text{meas}} = a_{\text{astro}} + a_{\text{ship}}$, where $a_{\text{ship}} = \dfrac{m_{\text{astro}}a_{\text{astro}}}{m_{\text{ship}}}$; c. If the force could be exerted on the astronaut by another source (other than the spaceship), then the spaceship would not experience a recoil.

45. $F_{\text{net}} = 4.12 \times 10^5\ \text{N}$

47. $a = 253\ \text{m/s}^2$

49. $F_{\text{net}} = F - f = ma \Rightarrow F = 1.26 \times 10^3\ \text{N}$

51. $$\begin{aligned} v^2 &= v_0^2 + 2ax \Rightarrow a = -7.80\ \text{m/s}^2 \\ F_{\text{net}} &= -7.80 \times 10^3\ \text{N} \end{aligned}$$

53. a. $\vec{\mathbf{F}}_{\text{net}} = m\,\vec{\mathbf{a}} \Rightarrow \vec{\mathbf{a}} = 9.0\,\hat{\mathbf{i}}\ \text{m/s}^2$; b. The acceleration has magnitude $9.0\ \text{m/s}^2$, so $x = 110\ \text{m}$.

54. a. $$\begin{aligned} w_{\text{Moon}} &= mg_{\text{Moon}} \\ m &= 150\ \text{kg} \\ w_{\text{Earth}} &= 1.5 \times 10^3\ \text{N} \end{aligned}$$; b. Mass does not change, so the suited astronaut's mass on both Earth and the Moon is $150\ \text{kg}$.

56. a. $$\begin{aligned} F_{\text{h}} &= 3.68 \times 10^3\ \text{N and} \\ w &= 7.35 \times 10^2\ \text{N} \\ \frac{F_{\text{h}}}{w} &= 5.00\ \text{times greater than weight} \end{aligned}$$;

b. $$\begin{aligned} F_{\text{net}} &= 3750\ \text{N} \\ \theta &= 11.3°\ \text{from horizontal} \end{aligned}$$

58. $$\begin{aligned} w &= 19.6\ \text{N} \\ F_{\text{net}} &= 5.40\ \text{N} \\ F_{\text{net}} &= ma \Rightarrow a = 2.70\ \text{m/s}^2 \end{aligned}$$

60. $0.60\,\hat{\mathbf{i}} - 8.4\,\hat{\mathbf{j}}\ \text{m/s}^2$

62. 497 N

64. a. $F_{\text{net}} = 2.64 \times 10^7\ \text{N}$; b. The force exerted on the ship is also $2.64 \times 10^7\ \text{N}$ because it is opposite the shell's direction of motion.

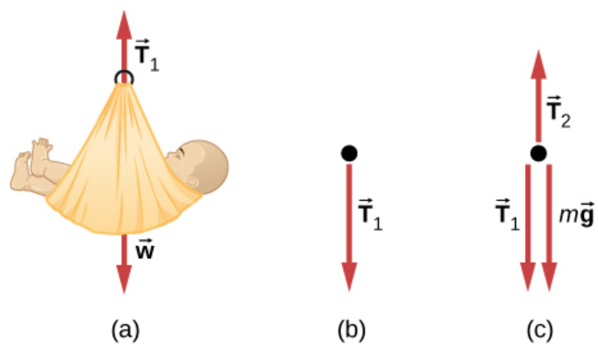
66. Because the weight of the history book is the force exerted by Earth on the history book, we represent it as

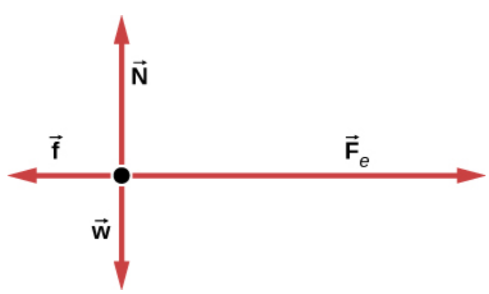
$\vec{F}_{EH} = -14\,\hat{j}\text{ N}.$ Aside from this, the history book interacts only with the physics book. Because the acceleration of the history book is zero, the net force on it is zero by Newton's second law: $\vec{F}_{PH} + \vec{F}_{EH} = \vec{0}$, where $\vec{F}_{PH}$ is the force exerted by the physics book on the history book. Thus, $\vec{F}_{PH} = -\vec{F}_{EH} = -\left(-14\,\hat{j}\right)\text{N} = 14\,\hat{j}\text{ N}.$ We find that the physics book exerts an upward force of magnitude 14 N on the history book. The physics book has three forces exerted on it: $\vec{F}_{EP}$ due to Earth, $\vec{F}_{HP}$ due to the history book, and $\vec{F}_{DP}$ due to the desktop. Since the physics book weighs 18 N, $\vec{F}_{EP} = -18\,\hat{j}\text{ N}.$ From Newton's third law, $\vec{F}_{HP} = -\vec{F}_{PH},$ so $\vec{F}_{HP} = -14\,\hat{j}\text{ N}.$ Newton's second law applied to the physics book gives $\sum \vec{F} = \vec{0}$, or $\vec{F}_{DP} + \vec{F}_{EP} + \vec{F}_{HP} = \vec{0}$, so $\vec{F}_{DP} = -\left(-18\,\hat{j}\right) - \left(-14\,\hat{j}\right) = 32\,\hat{j}\text{ N}.$ The desk exerts an upward force of 32 N on the physics book. To arrive at this solution, we apply Newton's second law twice and Newton's third law once.

69. $T = 1.96 \times 10^{-4}\text{ N}$

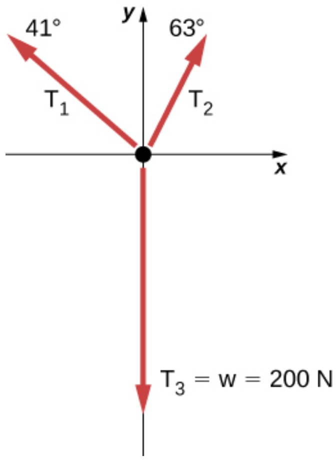
71. a. 5.6 kg; b. 55 N; c. $T_2 = 60\text{ N}$;

d.





73.



75.
77. a. 10.0 N; b. 97.0 N
79. a. $4.9\ \text{m/s}^2$; b. The cabinet will not slip. c. The cabinet will slip.

82.
$$\begin{aligned}\text{net } F_y &= 0 \Rightarrow N = mg\cos\theta \\ \text{net } F_x &= ma \\ a &= g(\sin\theta - \mu_\text{k}\cos\theta)\end{aligned}$$

84. a. $1.69\ \text{m/s}^2$; b. $5.71°$

86. a. $10.8\ \text{m/s}^2$; b. $7.85\ \text{m/s}^2$; c. $2.00\ \text{m/s}^2$

88. a. $9.09\ \text{m/s}^2$; b. $6.16\ \text{m/s}^2$; c. $0.294\ \text{m/s}^2$

91. a. 46.5 N; b. $0.629\ \text{m/s}^2$

93. a. 483 N; b. 17.4 N; c. 2.24, 0.0807
96. a. 179 N; b. 290 N; c. 8.3 m/s
99. 21 m/s
100. $7.4 \times 10^{-8}\ \text{N}$

102. a. $7.01 \times 10^{-7}\ \text{N}$; b. The mass of Jupiter is

$$m_\text{J} = 1.90 \times 10^{27}\ \text{kg}$$

$$F_\text{J} = 1.35 \times 10^{-6}\ \text{N}$$

$$\frac{F_\text{f}}{F_\text{J}} = 0.521$$

104. a. $9.25 \times 10^{-6}\ \text{N}$; b. Not very, as the ISS is not even symmetrical, much less spherically symmetrical.

106. a. $1.41 \times 10^{-15}\ \text{m/s}^2$; b. $1.69 \times 10^{-4}\ \text{m/s}^2$

108. a. $1.62\ \text{m/s}^2$; b. $3.75\ \text{m/s}^2$

110. a. 147 N; b. 25.5 N; c. 15 kg; d. 0; e. 15 kg
112. $12\ \text{m/s}^2$

114. $(3/2)R_\text{E}$

117. a. 0.25; b. 0.125
119. a. $5.08 \times 10^3\ \text{km}$; b. This less than the radius of Earth.

121. $1.89 \times 10^{27}\ \text{kg}$

123. a. $4.01 \times 10^{13}\ \text{kg}$; b. The satellite must be outside the radius of the asteroid, so it can’t be larger than this. If it were this size, then its density would be about $1200\ \text{kg/m}^3$. This is just above that of water, so this seems quite reasonable.

125. a. $1.66 \times 10^{-10}\ \text{m/s}^2$; Yes, the centripetal acceleration is so small it supports the contention that a nearly inertial frame of reference can be located at the Sun. b. $2.17 \times 10^5\ \text{m/s}$

ADDITIONAL PROBLEMS

127.

129. a. $F_{\text{net}} = \dfrac{m(v^2 - v_0{}^2)}{2x}$; b. 2590 N

130. 16 N

CHALLENGE PROBLEMS

132. a. 14.1 m/s; b. 601 N

133. $\frac{F}{m}t^2$

CHAPTER 9

CHECK YOUR UNDERSTANDING

9.1. $I_{\text{parallel-axis}} = I_{\text{center of mass}} + md^2 = mR^2 + mR^2 = 2mR^2$

9.2. The angle between the lever arm and the force vector is $80°$; therefore, $r_\perp = 100\text{m}(\sin 80°) = 98.5\ \text{m}$. The force gives the ship a negative or clockwise torque. The torque is then $\tau = -r_\perp F = -98.5\ \text{m}(5.0 \times 10^5\ \text{N}) = -4.9 \times 10^7\ \text{N} \cdot \text{m}$.

9.3. a. The angular acceleration is $\alpha = \dfrac{20.0(2\pi)\text{rad/s} - 0}{10.0\ \text{s}} = 12.56\ \text{rad/s}^2$. Solving for the torque, we have $\sum_i \tau_i = I\alpha = (30.0\ \text{kg} \cdot \text{m}^2)(12.56\ \text{rad/s}^2) = 376.80\ \text{N} \cdot \text{m}$; b. The angular acceleration is $\alpha = \dfrac{0 - 20.0(2\pi)\text{rad/s}}{20.0\ \text{s}} = -6.28\ \text{rad/s}^2$. Solving for the torque, we have $\sum_i \tau_i = I\alpha = (30.0\ \text{kg-m}^2)(-6.28\ \text{rad/s}^2) = -188.50\ \text{N} \cdot \text{m}$

CONCEPTUAL QUESTIONS

2. As each piece of matter moves closer to the center of rotation, the moment of inertia will decrease.
3. The hollow sphere, since the mass is distributed further away from the rotation axis.
5. a. It decreases. b. The arms could be approximated with rods and the discus with a disk. The torso is near the axis of rotation so it doesn't contribute much to the moment of inertia.
7. Because the moment of inertia varies as the square of the distance to the axis of rotation. The mass of the rod located at distances greater than $L/2$ would provide the larger contribution to make its moment of inertia greater than the point mass at $L/2$.
9. magnitude of the force, length of the lever arm, and angle of the lever arm and force vector
11. The moment of inertia of the wheels is reduced, so a smaller torque is needed to accelerate them.
13. yes
15. $|\vec{\mathbf{r}}|$ can be equal to the lever arm but never less than the lever arm

17. If the forces are along the axis of rotation, or if they have the same lever arm and are applied at a point on the rod.

PROBLEMS

19. $I = 0.315\ \text{kg} \cdot \text{m}^2$

20. $I = \frac{7}{36}mL^2$

21. $F = 30\ \text{N}$

23. a. $0.85\ \text{m}(55.0\ \text{N}) = 46.75\ \text{N} \cdot \text{m}$; b. It does not matter at what height you push.

25. $m_2 = \frac{4.9\,\text{N} \cdot \text{m}}{9.8(0.3\,\text{m})} = 1.67\,\text{kg}$

27. $\tau_{net} = -9.0\,\text{N} \cdot \text{m} + 3.46\,\text{N} \cdot \text{m} + 0 - 3.38\,\text{N} \cdot \text{m} = -8.92\,\text{N} \cdot \text{m}$

29. $\tau = 5.66\,\text{N} \cdot \text{m}$

32. a. $\tau = (0.280\,\text{m})(180.0\,\text{N}) = 50.4\,\text{N} \cdot \text{m}$; b. $\alpha = 17.14\,\text{rad/s}^2$;
c. $\alpha = 17.04\,\text{rad/s}^2$

34. $\tau = 8.0\,\text{N} \cdot \text{m}$

36. $\tau = -43.6\,\text{N} \cdot \text{m}$

38. a. $\alpha = 1.4 \times 10^{-10}\ \text{rad/s}^2$;
b. $\tau = 1.36 \times 10^{28}\,\text{N-m}$; c. $F = 2.1 \times 10^{21}\ \text{N}$

CHAPTER 10

CHECK YOUR UNDERSTANDING

10.1. No, only its magnitude can be constant; its direction must change, to be always opposite the relative displacement along the surface.
10.2. No, it's only approximately constant near Earth's surface.
10.3. a. the car; b. the truck
10.4. $\sqrt{3}$ m/s
10.5. 980 W
10.6. 35.3 kJ, 143 kJ, 0
10.7. The value of g drops by about 10% over this change in height. So $\Delta U = mg(y_2 - y_1)$ will give too large a value. If we use $g = 9.80\,\text{m/s}$, then we get $\Delta U = mg(y_2 - y_1) = 3.53 \times 10^{10}\ \text{J}$ which is about 6% greater than that found with the correct method.
10.8. 0.033 m
10.9. b. At any given height, the gravitational potential energy is the same going up or down, but the kinetic energy is less going down than going up, since air resistance is dissipative and does negative work. Therefore, at any height, the speed going down is less than the speed going up, so it must take a longer time to go down than to go up.
10.10. constant $U(x) = -1\ \text{J}$
10.11. The probe must overcome both the gravitational pull of Earth and the Sun. In the second calculation of our example, we found the speed necessary to escape the Sun from a distance of Earth's orbit, not from Earth itself. The proper way to find this value is to start with the energy equation, **Equation 10.57**, in which you would include a potential energy term for both Earth and the Sun.

CONCEPTUAL QUESTIONS

2. When you push on the wall, this "feels" like work; however, there is no displacement so there is no physical work. Energy is consumed, but no energy is transferred.
4. If you continue to push on a wall without breaking through the wall, you continue to exert a force with no displacement, so no work is done.
6. The total displacement of the ball is zero, so no work is done.
8. Both require the same gravitational work, but the stairs allow Tarzan to take this work over a longer time interval and hence gradually exert his energy, rather than dramatically by climbing a vine.
9. The first particle has a kinetic energy of $4(\frac{1}{2}mv^2)$ whereas the second particle has a kinetic energy of $2(\frac{1}{2}mv^2),$ so the first particle has twice the kinetic energy of the second particle.
11. The hollow sphere, since the mass is distributed further away from the rotation axis.
12. The mower would gain energy if $-90° < \theta < 90°.$ It would lose energy if $90° < \theta < 270°.$ The mower may also lose energy due to friction with the grass while pushing; however, we are not concerned with that energy loss for this problem.
14. The second marble has twice the kinetic energy of the first because kinetic energy is directly proportional to mass, like the work done by gravity.
16. Unless the environment is nearly frictionless, you are doing some positive work on the environment to cancel out the frictional work against you, resulting in zero total work producing a constant velocity.
18. Appliances are rated in terms of the energy consumed in a relatively small time interval. It does not matter how long the appliance is on, only the rate of change of energy per unit time.
20. The spark occurs over a relatively short time span, thereby delivering a very low amount of energy to your body.
22. If the force is antiparallel or points in an opposite direction to the velocity, the power expended can be negative.

24. The potential energy of a system can be negative because its value is relative to a defined point.
26. If the reference point of the ground is zero gravitational potential energy, the javelin first increases its gravitational potential energy, followed by a decrease in its gravitational potential energy as it is thrown until it hits the ground. The overall change in gravitational potential energy of the javelin is zero unless the center of mass of the javelin is lower than from where it is initially thrown, and therefore would have slightly less gravitational potential energy.
28. the vertical height from the ground to the object
30. As we move to larger orbits, the change in potential energy increases, whereas the orbital velocity decreases. Hence, the ratio is highest near Earth's surface (technically infinite if we orbit at Earth's surface with no elevation change), moving to zero as we reach infinitely far away.
32. The car experiences a change in gravitational potential energy as it goes down the hills because the vertical distance is decreasing. Some of this change of gravitational potential energy will be taken away by work done by friction. The rest of the energy results in a kinetic energy increase, making the car go faster. Lastly, the car brakes and will lose its kinetic energy to the work done by braking to a stop.
34. It states that total energy of the system E is conserved as long as there are no non-conservative forces acting on the object.
36. He puts energy into the system through his legs compressing and expanding.
38. Four times the original height would double the impact speed.
40. As we move to larger orbits, the change in potential energy increases, whereas the orbital velocity decreases. Hence, the ratio is highest near Earth's surface (technically infinite if we orbit at Earth's surface with no elevation change), moving to zero as we reach infinitely far away.

PROBLEMS

41. 3.00 J
43. a. 593 kJ; b. –589 kJ; c. 0
45. 3.14 kJ
47. a. –700 J; b. 0; c. 700 J; d. 38.6 N; e. 0
48. 100 J
50. a. 2.45 J; b. – 2.45 J; c. 0
52. a. 2.22 kJ; b. −2.22 kJ; c. 0
53. 18.6 kJ
55. a. 1.47 m/s; b. answers may vary
57. a. 772 kJ; b. 4.0 kJ; c. 1.8×10^{-16} J
59. a. 2.6 kJ; b. 640 J
60. a. $K = 2.56 \times 10^{29}$ J;
b. $K = 2.68 \times 10^{33}$ J
62. $K = 434.0\,\text{J}$
64. $K = 3.95 \times 10^{42}$ J
67. 2.72 kN
69. 102 N
71. 2.8 m/s
73. $W(\text{bullet}) = 20 \times W(\text{crate})$
75. 12.8 kN
77. 0.25
79. a. 24 m/s, −4.8 m/s^2; b. 29.4 m
81. 310 m/s
83. a. 40; b. 8 million
85. $149
87. a. 208 W; b. 141 s
89. a. 3.20 s; b. 4.04 s
91. a. 224 s; b. 24.8 MW; c. 49.7 kN
93. a. 1.57 kW; b. 6.28 kW
95. $6.83\,\mu\text{W}$
97. a. 8.51 J; b. 8.51 W
99. 1.7 kW
100. 40,000
102. a. $-200\,\text{J}$; b. $-200\,\text{J}$; c. $-100\,\text{J}$; d. $-300\,\text{J}$
104. a. $0.068\,\text{J}$; b. $-0.068\,\text{J}$; c. $0.068\,\text{J}$; d. $0.068\,\text{J}$; e. $-0.068\,\text{J}$; f. 46 cm
105. 14 J
107. proof

109. 9.7 m/s
112. 1900 J
114. –137 J
117. 5000 m/s
119. 1440 m/s
121. 11 km/s
123. a. 5.85×10^{10} J; b. -5.85×10^{10} J; No. It assumes the kinetic energy is recoverable. This would not even be reasonable if we had an elevator between Earth and the Moon.
129. a. 47.6 m; b. 1.88×10^5 J; c. 373 N

ADDITIONAL PROBLEMS

133. $15 \text{ N} \cdot \text{m}$
135. $39 \text{ N} \cdot \text{m}$
137. a. $208 \text{ N} \cdot \text{m}$; b. $240 \text{ N} \cdot \text{m}$
139. a. $-0.9 \text{ N} \cdot \text{m}$; b. $-0.83 \text{ N} \cdot \text{m}$
141. a. 10. J; b. 10. J; c. 380 N/m
143. 160 J/s
145. a. 10 N; b. 20 W
147. 3.6 m/s
149. proof
152. 18 m/s
153. $v_A = 24 \text{ m/s}; v_B = 14 \text{ m/s}; v_C = 31 \text{ m/s}$
155. a. Loss of energy is $240 \text{ N} \cdot \text{m}$; b. $F = 8 \text{ N}$
157. 89.7 m/s

CHALLENGE PROBLEMS

159. If crate goes up: a. 3.46 kJ; b. −1.89 kJ; c. −1.57 kJ; d. 0; If crate goes down: a. −0.39 kJ; b. −1.18 kJ; c. 1.57 kJ; d. 0
162. a. 40 hp; b. 39.8 MJ, independent of speed; c. 80 hp, 79.6 MJ at 30 m/s; d. If air resistance is proportional to speed, the car gets about 22 mpg at 34 mph and half that at twice the speed, closer to actual driving experience.

CHAPTER 11

CHECK YOUR UNDERSTANDING

11.1. To reach a final speed of $v_{\text{f}} = \frac{1}{4}\left(3.0 \times 10^8 \text{ m/s}\right)$ at an acceleration of 10*g*, the time required is

$$\begin{aligned} 10g &= \frac{v_{\text{f}}}{\Delta t} \\ \Delta t &= \frac{v_{\text{f}}}{10g} \frac{\frac{1}{4}\left(3.0 \times 10^8 \text{ m/s}\right)}{10g} = 7.7 \times 10^5 \text{ s} = 8.9 \text{ d} \end{aligned}$$

11.2. If the phone bounces up with approximately the same initial speed as its impact speed, the change in momentum of the phone will be $\Delta \vec{\mathbf{p}} = m\Delta \vec{\mathbf{v}} - \left(-m\Delta \vec{\mathbf{v}}\right) = 2m\Delta \vec{\mathbf{v}}$. This is twice the momentum change than when the phone does not bounce, so the impulse-momentum theorem tells us that more force must be applied to the phone.
11.3. If the smaller cart were rolling at 1.33 m/s to the left, then conservation of momentum gives

$$\begin{aligned} (m_1 + m_2)\vec{\mathbf{v}}_{\mathbf{f}} &= m_1 v_1 \hat{\mathbf{i}} - m_2 v_2 \hat{\mathbf{i}} \\ \vec{\mathbf{v}}_{\mathbf{f}} &= \left(\frac{m_1 v_1 - m_2 v_2}{m_1 + m_2}\right)\hat{\mathbf{i}} \\ &= \left[\frac{(0.675 \text{ kg})(0.75 \text{ m/s}) - (0.500 \text{ kg})(1.33 \text{ m/s})}{1.175 \text{ kg}}\right]\hat{\mathbf{i}} \\ &= -(0.135 \text{ m/s})\hat{\mathbf{i}} \end{aligned}$$

Thus, the final velocity is 0.135 m/s to the left.
11.4. If the ball does not bounce, its final momentum $\vec{\mathbf{p}}_{\mathbf{2}}$ is zero, so

$$\begin{aligned}\Delta \vec{\mathbf{p}} &= \vec{\mathbf{p}}_2 - \vec{\mathbf{p}}_1\\ &= (0)\hat{\mathbf{j}} - (-1.4\,\text{kg}\cdot\text{m/s})\hat{\mathbf{j}}\\ &= +(1.4\,\text{kg}\cdot\text{m/s})\hat{\mathbf{j}}\end{aligned}$$

11.5. Consider the impulse momentum theory, which is $\vec{J} = \Delta \vec{p}$. If $\vec{J} = 0$, we have the situation described in the example. If a force acts on the system, then $\vec{J} = \vec{F}_{\text{ave}} \Delta t$. Thus, instead of $\vec{\mathbf{p}}_f = \vec{\mathbf{p}}_i$, we have

$$\vec{F}_{\text{ave}} \Delta t = \Delta \vec{p} = \vec{p}_{\text{f}} - \vec{p}_{\text{i}}$$

where $\vec{\mathbf{F}}_{\text{ave}}$ is the force due to friction.

11.6. The impulse is the change in momentum multiplied by the time required for the change to occur. By conservation of momentum, the changes in momentum of the probe and the comment are of the same magnitude, but in opposite directions, and the interaction time for each is also the same. Therefore, the impulse each receives is of the same magnitude, but in opposite directions. Because they act in opposite directions, the impulses are not the same. As for the impulse, the force on each body acts in opposite directions, so the forces on each are not equal. However, the change in kinetic energy differs for each, because the collision is not elastic.

11.7. This solution represents the case in which no interaction takes place: the first puck misses the second puck and continues on with a velocity of 2.5 m/s to the left. This case offers no meaningful physical insights.

11.8. If zero friction acts on the car, then it will continue to slide indefinitely ($d \to \infty$), so we cannot use the work-kinetic-energy theorem as is done in the example. Thus, we could not solve the problem from the information given.

11.9. The average radius of Earth's orbit around the Sun is 1.496×10^9 m. Taking the Sun to be the origin, and noting that the mass of the Sun is approximately the same as the masses of the Sun, Earth, and Moon combined, the center of mass of the Earth + Moon system and the Sun is

$$\begin{aligned}R_{\text{CM}} &= \frac{m_{\text{Sun}} R_{\text{Sun}} + m_{\text{em}} R_{\text{em}}}{m_{\text{Sun}}}\\ &= \frac{(1.989 \times 10^{30}\ \text{kg})(0) + (5.97 \times 10^{24}\ \text{kg} + 7.36 \times 10^{22}\ \text{kg})(1.496 \times 10^9\ \text{m})}{1.989 \times 10^{30}\ \text{kg}}\\ &= 4.6\,\text{km}\end{aligned}$$

Thus, the center of mass of the Sun, Earth, Moon system is 4.6 km from the center of the Sun.

11.10. The explosions would essentially be spherically symmetric, because gravity would not act to distort the trajectories of the expanding projectiles.

CONCEPTUAL QUESTIONS

1. Since $K = p^2/2m$, then if the momentum is fixed, the object with smaller mass has more kinetic energy.

3. Yes; impulse is the force applied multiplied by the time during which it is applied ($J = F\Delta t$), so if a small force acts for a long time, it may result in a larger impulse than a large force acting for a small time.

5. By friction, the road exerts a horizontal force on the tires of the car, which changes the momentum of the car.

7. Momentum is conserved when the mass of the system of interest remains constant during the interaction in question and when no *net* external force acts on the system during the interaction.

9. To accelerate air molecules in the direction of motion of the car, the car must exert a force on these molecules by Newton's second law $\vec{F} = d\vec{p}/dt$. By Newton's third law, the air molecules exert a force of equal magnitude but in the opposite direction on the car. This force acts in the direction opposite the motion of the car and constitutes the force due to air resistance.

11. No, he is not a closed system because a net nonzero external force acts on him in the form of the starting blocks pushing on his feet.

13. Yes, all the kinetic energy can be lost if the two masses come to rest due to the collision (i.e., they stick together).

PROBLEMS

17. a. magnitude: $25\,\text{kg}\cdot\text{m/s}$; b. same as a.

19. $1.78 \times 10^{29}\ \text{kg}\cdot\text{m/s}$

21. $1.3 \times 10^9\ \text{kg}\cdot\text{m/s}$

23. a. 1.50×10^6 N; b. 1.00×10^5 N

25. 4.69×10^5 N

27. 2.10×10^3 N

29. $\vec{\mathbf{p}}(t) = \left(10\,\hat{\mathbf{i}} + 20t\,\hat{\mathbf{j}}\right)\text{kg} \cdot \text{m/s}$; $\vec{\mathbf{F}} = (20\,\text{N})\,\hat{\mathbf{j}}$

30. $(0.122\,\text{m/s})\,\hat{\mathbf{i}}$

32. a. 47 m/s in the bullet to block direction; b. $70.6\,\text{N} \cdot \text{s}$, toward the bullet; c. $70.6\,\text{N} \cdot \text{s}$, toward the block; d. magnitude is 2.35×10^4 N

34. 2:5
36. 5.9 m/s
38. 2.5 cm
40. the speed of the leading bumper car is 6.00 m/s and that of the trailing bumper car is 5.60 m/s
42. 6.6%
44. 1.8 m/s
46. With the origin defined to be at the position of the 150-g mass, $x_{\text{CM}} = -1.23\text{cm}$ and $y_{\text{CM}} = 0.69\text{cm}$

47. a. $R_1 = 4$ m, $R_2 = 2$ m; b. $X_{\text{CM}} = \dfrac{m_1 x_1 + m_2 x_2}{m_1 + m_2}$, $Y_{\text{CM}} = \dfrac{m_1 y_1 + m_2 y_2}{m_1 + m_2}$; c. yes, with $R = \dfrac{1}{m_1 + m_2}\sqrt{16m_1^2 + 4m_2^2}$

49. $(x_{\text{CM}}, y_{\text{CM}}, z_{\text{CM}}) = (0{,}0,\ h/4)$

51. $(x_{\text{CM}}, y_{\text{CM}}, z_{\text{CM}}) = (0,\ 4R/(3\pi),\ 0)$

CHAPTER 12

CHECK YOUR UNDERSTANDING

12.1. $I_{\text{sphere}} = \frac{2}{5}mr^2$, $I_{\text{cylinder}} = \frac{1}{2}mr^2$; Taking the ratio of the angular momenta, we have: $\dfrac{L_{\text{cylinder}}}{L_{\text{sphere}}} = \dfrac{I_{\text{cylinder}}\omega_0}{I_{\text{sphere}}\omega_0} = \dfrac{\frac{1}{2}mr^2}{\frac{2}{5}mr^2} = \dfrac{5}{4}$. Thus, the cylinder has 25% more angular momentum. This is because the cylinder has more mass distributed farther from the axis of rotation.
12.2. Using conservation of angular momentum, we have
$I(4.0\,\text{rev/min}) = 1.25I\omega_{\text{f}}$, $\omega_{\text{f}} = \frac{1.0}{1.25}(4.0\,\text{rev/min}) = 3.2\,\text{rev/min}$

CONCEPTUAL QUESTIONS

2. All points on the straight line will give zero angular momentum, because a vector crossed into a parallel vector is zero.
4. The particle must be moving on a straight line that passes through the chosen origin.
6. Without the small propeller, the body of the helicopter would rotate in the opposite sense to the large propeller in order to conserve angular momentum. The small propeller exerts a thrust at a distance *R* from the center of mass of the aircraft to prevent this from happening.
8. The angular velocity increases because the moment of inertia is decreasing.
10. More mass is concentrated near the rotational axis, which decreases the moment of inertia causing the star to increase its angular velocity.

PROBLEMS

16. $I = 720.0\,\text{kg} \cdot \text{m}^2$; $\alpha = 4.20\,\text{rad/s}^2$;
$\omega(10\,\text{s}) = 42.0\,\text{rad/s}$; $L = 3.02 \times 10^4\ \text{kg} \cdot \text{m}^2/\text{s}$;
$\omega(20\,\text{s}) = 84.0\,\text{rad/s}$;

18. $L = 1.131 \times 10^7\ \text{kg} \cdot \text{m}^2/\text{s}$;

20. $\omega = 28.6\,\text{rad/s} \Rightarrow L = 2.6\,\text{kg} \cdot \text{m}^2/\text{s}$

22. $L_f = \frac{2}{5}M_S(3.5 \times 10^3\,\text{km})^2\frac{2\pi}{T_f}$,

$$(7.0 \times 10^5\ \text{km})^2 \frac{2\pi}{28\ \text{days}} = (3.5 \times 10^3\ \text{km})^2 \frac{2\pi}{T_f} \qquad T_f$$
$$= 28\ \text{days} \frac{(3.5 \times 10^3\ \text{km})^2}{(7.0 \times 10^5\ \text{km})^2} = 7.0 \times 10^{-4}\ \text{day} = 60.5\ \text{s}$$

24. $f_f = 2.1\ \text{rev/s} \Rightarrow f_0 = 0.5\ \text{rev/s}$

26. $r_P m v_P = r_A m v_A \Rightarrow v_P = 18.3\ \text{km/s}$

27. a. $I_{\text{disk}} = 5.0 \times 10^{-4}\ \text{kg} \cdot \text{m}^2$,

$I_{\text{bug}} = 2.0 \times 10^{-4}\ \text{kg} \cdot \text{m}^2$,

$(I_{\text{disk}} + I_{\text{bug}})\omega_1 = I_{\text{disk}}\omega_2$,

$\omega_2 = 14.0\ \text{rad/s}$

b. $\Delta K = 0.014\ \text{J}$;

c. $\omega_3 = 10.0\ \text{rad/s}$ back to the original value;

d. $\frac{1}{2}(I_{\text{disk}} + I_{\text{bug}})\omega_3^2 = 0.035\ \text{J}$ back to the original value;

e. work of the bug crawling on the disk

30. $I_0 = 340.48\ \text{kg} \cdot \text{m}^2$,

$I_f = 268.8\ \text{kg} \cdot \text{m}^2$,

$\omega_f = 25.33\ \text{rpm}$

31. Moment of inertia in the record spin: $I_0 = 0.5\ \text{kg} \cdot \text{m}^2$,

$I_f = 1.1\ \text{kg} \cdot \text{m}^2$,

$\omega_f = \frac{I_0}{I_f}\omega_0 \Rightarrow f_f = 155.5\ \text{rev/min}$

33. Her spin rate in the air is: $f_f = 2.0\ \text{rev/s}$;

She can do four flips in the air.

35. Moment of inertia with all children aboard:

$I_0 = 2.4 \times 10^5\ \text{kg} \cdot \text{m}^2$;

$I_f = 1.5 \times 10^5\ \text{kg} \cdot \text{m}^2$;

$f_f = 0.3\text{rev/s}$

37. $I_0 = 1.00 \times 10^{10}\ \text{kg} \cdot \text{m}^2$,

$I_f = 9.94 \times 10^9\ \text{kg} \cdot \text{m}^2$,

$f_f = 3.32\ \text{rev/min}$

CHAPTER 13

CHECK YOUR UNDERSTANDING

13.1. AA' becomes longer, $A'B'$ tilts further away from the surface, and the refracted ray tilts away from the normal.

13.2. $T = \frac{2.9 \times 10^6\ \text{nm K}}{\lambda_{\text{max}}} = \frac{2.9 \times 10^6\ \text{nm K}}{290\ \text{nm}} = 10{,}000\ \text{K}$

13.3. The 8700 K star has triple the temperature, so it is $3^4 = 81$ times brighter.

13.4. $\frac{b_{\text{Sirius}}}{b_{\text{Polaris}}} = \left(100^{0.2}\right)^{2.0 - (-1.5)} = \left(100^{0.2}\right)^{3.5} = 100^{0.7} = 25$

CONCEPTUAL QUESTIONS

4. yes

10. Since Star 1 has a lower value of the color index, its color is bluer than that of Star 2, and therefore its temperature must be higher than that of Star 2.

PROBLEMS

29. 52.5

CHAPTER 14

CHECK YOUR UNDERSTANDING

14.1. Because the light is shifted to a longer wavelength, the star is moving away from us: $v = c \times \frac{\Delta\lambda}{\lambda} = (3.0 \times 10^8\ \text{m/s})\left(\frac{0.1\ \text{nm}}{500\ \text{nm}}\right) = (3.0 \times 10^8\ \text{m/s})\left(\frac{0.1 \times 10^{-9}\ \text{m}}{500 \times 10^{-9}\ \text{m}}\right) = 60{,}000\ \text{m/s}.$ Its speed is 60,000 m/s.

PROBLEMS

27. A measurable shift occurs when $\frac{\Delta\lambda}{\lambda} = \frac{v}{c} \geq 0.003$; therefore $v = 0.003c = 9.0 \text{ x } 10^5\,\text{m/s}$
28. v = 320 km/s
29. 20 cm

CHAPTER 15

CHECK YOUR UNDERSTANDING

15.1. Bunsen's burner
15.2. The wavelength of the radiation maximum decreases with increasing temperature.
15.3. $T_\alpha / T_\beta = 1/\sqrt{3} \cong 0.58,$ so the star β is hotter.
15.4. $3.3 \times 10^{-19}\,\text{J}$
15.5. No, because then $\Delta E/E \approx 10^{-21}$
15.6. 121.5 nm and 91.1 nm; no, these spectral bands are in the ultraviolet
15.7. $v_2 = 1.1 \times 10^6\,\text{m/s} \cong 0.0036c;$ $L_2 = 2\hbar$ $K_2 = 3.4\,\text{eV}$
15.8. frequency quadruples
15.9. the moment of inertia

CONCEPTUAL QUESTIONS

1. yellow
3. goes from red to violet through the rainbow of colors
5. would not differ
7. human eye does not see IR radiation
10. no
12. They are at ground state.
14. Answers may vary
16. increase
18. for larger *n*
20. Yes, the excess of 13.6 eV will become kinetic energy of a free electron.
22. no
24. Atomic and molecular spectra are said to be "discrete," because only certain spectral lines are observed. In contrast, spectra from a white light source (consisting of many photon frequencies) are continuous because a continuous "rainbow" of colors is observed.
26. UV light consists of relatively high frequency (short wavelength) photons. So the energy of the absorbed photon and the energy transition (ΔE) in the atom is relatively large. In comparison, visible light consists of relatively lower-frequency photons. Therefore, the energy transition in the atom and the energy of the emitted photon is relatively small.
28. For macroscopic systems, the quantum numbers are very large, so the energy difference (ΔE) between adjacent energy levels (orbits) is very small. The energy released in transitions between these closely space energy levels is much too small to be detected.
30. rotational energy, vibrational energy, and atomic energy

PROBLEMS

32. a. 0.81 eV; b. 2.1×10^{23}; c. 2 min 20 sec

34. a. 7245 K; b. 3.62 μm
36. about 3 K
38. 121.5 nm
40. a. 0.661 eV; b. –10.2 eV; c. 1.511 eV
42. 3038 THz
44. 97.33 nm
46. a. h/π; b. 3.4 eV; c. – 6.8 eV; d. – 3.4 eV
48. $n = 4$
50. 365 nm; UV
52. no
54. 7
55. For He^+, one electron "orbits" a nucleus with two protons and two neutrons ($Z = 2$). Ionization energy refers to the energy required to remove the electron from the atom. The energy needed to remove the electron in the ground state of He^+ ion to infinity is negative the value of the ground state energy, written:
$$E = -54.4\text{ eV}.$$
Thus, the energy to ionize the electron is $+54.4\text{ eV}$.
Similarly, the energy needed to remove an electron in the first excited state of Li^{2+} ion to infinity is negative the value of the first excited state energy, written:
$$E = -30.6\text{ eV}.$$
The energy to ionize the electron is 30.6 eV.
57. The wavelength of the laser is given by:
$$\lambda = \frac{hc}{-\Delta E},$$
where E_γ is the energy of the photon and ΔE is the magnitude of the energy difference. Solving for the latter, we get:
$$\Delta E = -2.795\text{ eV}.$$
The negative sign indicates that the electron lost energy in the transition.
59. $\Delta E_{L \to K} \approx (Z - 1)^2 (10.2\text{ eV}) = 3.68 \times 10^3\text{ eV}.$

61. According to the conservation of the energy, the potential energy of the electron is converted completely into kinetic energy. The initial kinetic energy of the electron is zero (the electron begins at rest). So, the kinetic energy of the electron just before it strikes the target is:
$$K = e\Delta V.$$
If *all* of this energy is converted into braking radiation, the frequency of the emitted radiation is a maximum, therefore:
$$f_{\text{max}} = \frac{e\Delta V}{h}.$$
When the emitted frequency is a maximum, then the emitted wavelength is a minimum, so:
$$\lambda_{\text{min}} = 0.1293\text{ nm}.$$
63. A muon is 200 times heavier than an electron, but the minimum wavelength does not depend on mass, so the result is unchanged.
65. 4.13×10^{-11} m
67. 72.5 keV
69. The atomic numbers for Cu and Au are $Z = 29$ and 79, respectively. The X-ray photon frequency for gold is greater than copper by a factor:
$$\left(\frac{f_{\text{Au}}}{f_{\text{Cu}}}\right)^2 = \left(\frac{79 - 1}{29 - 1}\right)^2 \approx 8.$$
Therefore, the X-ray wavelength of Au is about eight times shorter than for copper.
72. The measured value is 0.484 nm, and the actual value is close to 0.127 nm. The laboratory results are the same order of magnitude, but a factor 4 high.
74. 0.110 nm
76. a. $E = 2.2 \times 10^{-4}$ eV; b. $\Delta E = 4.4 \times 10^{-4}$ eV

CHAPTER 16

CHECK YOUR UNDERSTANDING

16.1. The actual amount (mass) of gasoline left in the tank when the gauge hits "empty" is less in the summer than in the winter. The gasoline has the same volume as it does in the winter when the "add fuel" light goes on, but because the gasoline has expanded, there is less mass.

16.2. To a good approximation, the heat transfer depends only on the temperature difference. Since the temperature differences are the same in both cases, the same 25 kJ is necessary in the second case. (As we will see in the next section, the answer would have been different if the object had been made of some substance that changes phase anywhere between $30\,°\text{C}$ and $50\,°\text{C}$.)

16.3. The ice and liquid water are in thermal equilibrium, so that the temperature stays at the freezing temperature as long as ice remains in the liquid. (Once all of the ice melts, the water temperature will start to rise.)

16.4. Snow is formed from ice crystals and thus is the solid phase of water. Because enormous heat is necessary for phase changes, it takes a certain amount of time for this heat to be transferred from the air, even if the air is above $0\,°\text{C}$.

16.5. Conduction: Heat transfers into your hands as you hold a hot cup of coffee. Convection: Heat transfers as the barista "steams" cold milk to make hot cocoa. Radiation: Heat transfers from the Sun to a jar of water with tea leaves in it to make "Sun tea." A great many other answers are possible.

16.6. Because area is the product of two spatial dimensions, it increases by a factor of four when each dimension is doubled $\left(A_{\text{final}} = (2d)^2 = 4d^2 = 4A_{\text{initial}}\right)$. The distance, however, simply doubles. Because the temperature difference and the coefficient of thermal conductivity are independent of the spatial dimensions, the rate of heat transfer by conduction increases by a factor of four divided by two, or two:

$$P_{\text{final}} = \frac{kA_{\text{final}}(T_{\text{h}} - T_{\text{c}})}{d_{\text{final}}} = \frac{k(4A_{\text{final}}(T_{\text{h}} - T_{\text{c}}))}{2d_{\text{initial}}} = 2\frac{kA_{\text{final}}(T_{\text{h}} - T_{\text{c}})}{d_{\text{initial}}} = 2P_{\text{initial}}.$$

16.7. Using a fan increases the flow of air: Warm air near your body is replaced by cooler air from elsewhere. Convection increases the rate of heat transfer so that moving air "feels" cooler than still air.

16.8. The radiated heat is proportional to the fourth power of the *absolute temperature*. Because $T_1 = 293\text{ K}$ and $T_2 = 313\text{ K}$, the rate of heat transfer increases by about 30% of the original rate.

CONCEPTUAL QUESTIONS

1. They are at the same temperature, and if they are placed in contact, no net heat flows between them.

3. The reading will change.

6. In principle, the lid expands more than the jar because metals have higher coefficients of expansion than glass. That should make unscrewing the lid easier. (In practice, getting the lid and jar wet may make gripping them more difficult.)

8. After being heated, the length is $(1 + 300\alpha)(1\text{ m})$. After being cooled, the length is $(1 - 300\,\alpha)(1 + 300\,\alpha)(1\text{ m})$. That answer is not 1 m, but it should be. The explanation is that even if α is exactly constant, the relation $\Delta L = \alpha L \Delta T$ is strictly true only in the limit of small ΔT. Since α values are small, the discrepancy is unimportant in practice.

9. Temperature differences cause heat transfer.

11. No, it is stored as thermal energy. A thermodynamic system does not have a well-defined quantity of heat.

13. It raises the boiling point, so the water, which the food gains heat from, is at a higher temperature.

15. Yes, by raising the pressure above 56 atm.

17. work

19. $0\,°\text{C}$ (at or near atmospheric pressure)

21. Condensation releases heat, so it speeds up the melting.

23. Because of water's high specific heat, it changes temperature less than land. Also, evaporation reduces temperature rises. The air tends to stay close to equilibrium with the water, so its temperature does not change much where there's a lot of water around, as in San Francisco but not Sacramento.

25. The liquid is oxygen, whose boiling point is above that of nitrogen but whose melting point is below the boiling point of liquid nitrogen. The crystals that sublime are carbon dioxide, which has no liquid phase at atmospheric pressure. The crystals that melt are water, whose melting point is above carbon dioxide's sublimation point. The water came from the instructor's breath.

27. Increasing circulation to the surface will warm the person, as the temperature of the water is warmer than human body temperature. Sweating will cause no evaporative cooling under water or in the humid air immediately above the tub.

29. It spread the heat over the area above the heating elements, evening the temperature there, but does not spread the heat much beyond the heating elements.

31. Heat is conducted from the fire through the fire box to the circulating air and then convected by the air into the room (forced convection).

33. The tent is heated by the Sun and transfers heat to you by all three processes, especially radiation.

35. If shielded, it measures the air temperature. If not, it measures the combined effect of air temperature and net radiative heat gain from the Sun.

37. Turn the thermostat down. To have the house at the normal temperature, the heating system must replace all the heat that was

lost. For all three mechanisms of heat transfer, the greater the temperature difference between inside and outside, the more heat is lost and must be replaced. So the house should be at the lowest temperature that does not allow freezing damage.
39. Air is a good insulator, so there is little conduction, and the heated air rises, so there is little convection downward.

PROBLEMS

41. That must be Celsius. Your Fahrenheit temperature is $102\ °\text{F}$. Yes, it is time to get treatment.

43. a. $\Delta T_\text{C} = 22.2\ °\text{C}$; b. We know that $\Delta T_\text{F} = T_\text{F2} - T_\text{F1}$. We also know that $T_\text{F2} = \frac{9}{5}T_\text{C2} + 32$ and $T_\text{F1} = \frac{9}{5}T_\text{C1} + 32$. So, substituting, we have $\Delta T_\text{F} = \left(\frac{9}{5}T_\text{C2} + 32\right) - \left(\frac{9}{5}T_\text{C1} + 32\right)$. Partially solving and rearranging the equation, we have $\Delta T_\text{F} = \frac{9}{5}(T_\text{C2} - T_\text{C1})$. Therefore, $\Delta T_\text{F} = \frac{9}{5}\Delta T_\text{C}$.

45. a. $-40°$; b. 575 K

47. Using **Table 16.2** to find the coefficient of thermal expansion of marble:

$$L = L_0 + \Delta L = L_0(1 + \alpha\Delta T) = 170\ \text{m}\left[1 + \left(2.5 \times 10^{-6}/°\text{C}\right)(-45.0\ °\text{C})\right] = 169.98\ \text{m}.$$

(Answer rounded to five significant figures to show the slight difference in height.)

49. Using **Table 16.2** to find the coefficient of thermal expansion of mercury:

$$\Delta L = \alpha L \Delta T = \left(6.0 \times 10^{-5}/°\text{C}\right)(0.0300\ \text{m})(3.00\ °\text{C}) = 5.4 \times 10^{-6}\ \text{m}.$$

51. On the warmer day, our tape measure will expand linearly. Therefore, each measured dimension will be smaller than the actual dimension of the land. Calling these measured dimensions l' and w', we will find a new area, *A*. Let's calculate these measured dimensions:

$$l' = l_0 - \Delta l = (20\ \text{m}) - (20\ °\text{C})(20\ \text{m})\left(\frac{1.2 \times 10^{-5}}{°\text{C}}\right) = 19.9952\ \text{m};$$

$$A' = l \times w' = (29.9928\ \text{m})(19.9952\ \text{m}) = 599.71\ \text{m}^2;$$

Cost change $= (A - A')\left(\frac{\$60{,}000}{\text{m}^2}\right) = \left((600 - 599.71)\text{m}^2\right)\left(\frac{\$60{,}000}{\text{m}^2}\right) = \$17{,}000.$

Because the area gets smaller, the price of the land *decreases* by about $17,000.

53. a. Use **Table 16.2** to find the coefficients of thermal expansion of steel and aluminum. Then $\Delta L_\text{Al} - \Delta L_\text{steel} = (\alpha_\text{Al} - \alpha_\text{steel})L_0\,\Delta T = \left(\frac{2.5 \times 10^{-5}}{°\text{C}} - \frac{1.2 \times 10^{-5}}{°\text{C}}\right)(1.00\ \text{m})(22\ °\text{C}) = 2.9 \times 10^{-4}\ \text{m}$.

b. By the same method with $L_0 = 30.0\ \text{m}$, we have $\Delta L = 8.6 \times 10^{-3}\ \text{m}$.

55. $\Delta V = 0.475\ \text{L}$

58. $m = 5.20 \times 10^8\ \text{J}$

60. $Q = mc\Delta T \ \Rightarrow \Delta T = \frac{Q}{mc}$; a. $21.0\ °\text{C}$; b. $25.0\ °\text{C}$; c. $29.3\ °\text{C}$; d. $50.0\ °\text{C}$

62. $Q = mc\Delta T \ \Rightarrow c = \frac{Q}{m\Delta T} = \frac{1.04\ \text{kcal}}{(0.250\ \text{kg})(45.0\ °\text{C})} = 0.0924\ \text{kcal/kg} \cdot °\text{C}$. It is copper.

64. a. $Q = m_\text{w}c_\text{w}\Delta T + m_\text{Al}c_\text{Al}\Delta T = (m_\text{w}c_\text{w} + m_\text{Al}c_\text{Al})\Delta T$;

$$Q = \begin{bmatrix}(0.500\ \text{kg})(1.00\ \text{kcal/kg} \cdot °\text{C}) + \\ (0.100\ \text{kg})(0.215\ \text{kcal/kg} \cdot °\text{C})\end{bmatrix}(54.9\ °\text{C}) = 28.63\ \text{kcal};$$

$\frac{Q}{m_p} = \frac{28.63\ \text{kcal}}{5.00\ \text{g}} = 5.73\ \text{kcal/g}$; b. $\frac{Q}{m_p} = \frac{200\ \text{kcal}}{33\ \text{g}} = 6\ \text{kcal/g}$, which is consistent with our results to part (a), to one significant figure.

66. $0.139\ °\text{C}$

68. It should be lower. The beaker will not make much difference: $16.3\ °\text{C}$

70. a. $1.00 \times 10^5\ \text{J}$; b. $3.68 \times 10^5\ \text{J}$; c. The ice is much more effective in absorbing heat because it first must be melted, which requires a lot of energy, and then it gains the same amount of heat as the bag that started with water. The first $2.67 \times 10^5\ \text{J}$ of heat is used to melt the ice, then it absorbs the $1.00 \times 10^5\ \text{J}$ of heat as water.

72. 58.1 g

74. Let *M* be the mass of pool water and *m* be the mass of pool water that evaporates.

$$Mc\Delta T = mL_{\text{V}(37\,°\text{C})} \Rightarrow \frac{m}{M} = \frac{c\Delta T}{L_{\text{V}(37\,°\text{C})}} = \frac{(1.00\text{ kcal/kg}\cdot°\text{C})(1.50\,°\text{C})}{580\text{ kcal/kg}} = 2.59\times10^{-3};$$

(Note that L_V for water at $37\,°\text{C}$ is used here as a better approximation than L_V for $100\,°\text{C}$ water.)

76. a. $1.47\times10^{15}\text{ kg}$; b. $4.90\times10^{20}\text{ J}$; c. 48.5 y

78. a. 9.67 L; b. Crude oil is less dense than water, so it floats on top of the water, thereby exposing it to the oxygen in the air, which it uses to burn. Also, if the water is under the oil, it is less able to absorb the heat generated by the oil.

80. a. 319 kcal; b. $2.00\,°\text{C}$

82. First bring the ice up to $0\,°\text{C}$ and melt it with heat Q_1: 4.74 kcal. This lowers the temperature of water by ΔT_2: $23.15\,°\text{C}$. Now, the heat lost by the hot water equals that gained by the cold water (T_f is the final temperature): $20.6\,°\text{C}$

84. Let the subscripts r, e, v, and w represent rock, equilibrium, vapor, and water, respectively.

$$m_\text{r}c_\text{r}(T_1 - T_\text{e}) = m_\text{V}L_\text{V} + m_\text{W}c_\text{W}(T_\text{e} - T_2);$$

$$\begin{aligned} m_\text{r} &= \frac{m_\text{V}L_\text{V} + m_\text{W}c_\text{W}(T_\text{e} - T_2)}{c_\text{r}(T_1 - T_\text{e})} \\ &= \frac{(0.0250\text{ kg})\left(2256\times10^3\text{ J/kg}\right) + (3.975\text{ kg})\left(4186\times10^3\text{ J/kg}\cdot°\text{C}\right)(100\,°\text{C} - 15\,°\text{C})}{(840\text{ J/kg}\cdot°\text{C})(500\,°\text{C} - 100\,°\text{C})} \\ &= 4.38\text{ kg} \end{aligned}$$

86. a. $1.01\times10^3\text{ W}$; b. One 1-kilowatt room heater is needed.

88. 84.0 W

90. 2.59 kg

92. a. 39.7 W; b. 820 kcal

94. $\frac{Q}{t} = \frac{kA(T_2 - T_1)}{d}$, so that

$$\frac{(Q/t)_\text{wall}}{(Q/t)_\text{window}} = \frac{k_\text{wall}A_\text{wall}d_\text{window}}{k_\text{window}A_\text{window}d_\text{wall}} = \frac{(2\times0.042\text{ J/s}\cdot\text{m}\cdot°\text{C})\left(10.0\text{ m}^2\right)\left(0.750\times10^{-2}\text{ m}\right)}{(0.84\text{ J/s}\cdot\text{m}\cdot°\text{C})\left(2.00\text{ m}^2\right)\left(13.0\times10^{-2}\text{ m}\right)}$$

This gives 0.0288 wall: window, or 35:1 window: wall

96. $\frac{Q}{t} = \frac{kA(T_2 - T_1)}{d} = \frac{kA\Delta T}{d} \Rightarrow$

$$\Delta T = \frac{d(Q/t)}{kA} = \frac{\left(6.00\times10^{-3}\text{ m}\right)(2256\text{ W})}{(0.84\text{ J/s}\cdot\text{m}\cdot°\text{C})\left(1.54\times10^{-2}\text{ m}^2\right)} = 1046\,°\text{C} = 1.05\times10^3\text{ K}$$

98. We found in the preceding problem that $P = 126\Delta T\text{ W}\cdot°\text{C}$ as baseline energy use. So the total heat loss during this period is $Q = (126\text{ J/s}\cdot°\text{C})(15.0\,°\text{C})(120\text{ days})\left(86.4\times10^3\text{ s/day}\right) = 1960\times10^6\text{ J}$. At the cost of \$1/MJ, the cost is \$1960. From an earlier problem, the savings is 12% or \$235/y. We need 150 m^2 of insulation in the attic. At $\$4/\text{m}^2$, this is a \$500 cost. So the payback period is $\$600/(\$235/\text{y}) = 2.6\text{ years}$ (excluding labor costs).

ADDITIONAL PROBLEMS

100. 7.39%

102. a. 1.06 cm; b. 1.11 cm

104. $1.7\text{ kJ/(kg}\cdot°\text{C)}$

106. a. $1.57\times10^4\text{ kcal}$; b. $18.3\text{ kW}\cdot\text{h}$; c. $1.29\times10^4\text{ kcal}$

108. $6.3\,°\text{C}$. All of the ice melted.

110. $63.9\,°\text{C}$, all the ice melted

112. a. 83 W; b. $1.97\times10^3\text{ W}$; The single-pane window has a rate of heat conduction equal to 1969/83, or 24 times that of a double-pane window.

114. The rate of heat transfer by conduction is 20.0 W. On a daily basis, this is 1,728 kJ/day. Daily food intake is $2400\text{ kcal/d}\times4186\text{ J/kcal} = 10{,}050\text{ kJ/day}$. So only 17.2% of energy intake goes as heat transfer by conduction to the environment at this ΔT.

116. 620 K

CHALLENGE PROBLEMS

118. Denoting the period by P, we know $P = 2\pi\sqrt{L/g}$. When the temperature increases by dT, the length increases by $\alpha L dT$. Then the new length is a. $P = 2\pi\sqrt{\frac{L + \alpha L dT}{g}} = 2\pi\sqrt{\frac{L}{g}(1 + \alpha dT)} = 2\pi\sqrt{\frac{L}{g}}\left(1 + \frac{1}{2}\alpha dT\right) = P\left(1 + \frac{1}{2}\alpha dT\right)$ by the binomial expansion. b. The clock runs slower, as its new period is 1.00019 s. It loses 16.4 s per day.

119. The amount of heat to melt the ice and raise it to $100\ ^\circ\text{C}$ is not enough to condense the steam, but it is more than enough to lower the steam's temperature by $50\ ^\circ\text{C}$, so the final state will consist of steam and liquid water in equilibrium, and the final temperature is $100\ ^\circ\text{C}$; 9.5 g of steam condenses, so the final state contains 49.5 g of steam and 40.5 g of liquid water.

122. a. $4\left(\pi R^2\right) T_\text{s}{}^4$; b. $4e\sigma\pi R^2 T_\text{s}{}^4$; c. $8e\sigma\pi R^2 T_\text{e}{}^4$; d. $T_s^4 = 2T_e^4$; e. $e\sigma T_\text{s}^4 + \frac{1}{4}(1 - A)S = \sigma T_\text{s}^4$; f. $288K$

CHAPTER 17

CHECK YOUR UNDERSTANDING

17.1. We first need to calculate the molar mass (the mass of one mole) of niacin. To do this, we must multiply the number of atoms of each element in the molecule by the element's molar mass.

$$(6 \text{ mol of carbon})(12.0 \text{ g/mol}) + (5 \text{ mol hydrogen})(1.0 \text{ g/mol})$$
$$+ (1 \text{ mol nitrogen})(14 \text{ g/mol}) + (2 \text{ mol oxygen})(16.0 \text{ g/mol}) = 123 \text{ g/mol}$$

Then we need to calculate the number of moles in 14 mg.

$$\left(\frac{14 \text{ mg}}{123 \text{ g/mol}}\right)\left(\frac{1 \text{ g}}{1000 \text{ mg}}\right) = 1.14 \times 10^{-4} \text{ mol.}$$

Then, we use Avogadro's number to calculate the number of molecules:

$$N = nN_A = \left(1.14 \times 10^{-4} \text{ mol}\right)\left(6.02 \times 10^{23} \text{ molecules/mol}\right) = 6.85 \times 10^{19} \text{ molecules.}$$

17.2. The density of a gas is equal to a constant, the average molecular mass, times the number density N/V. From the ideal gas law, $pV = Nk_\text{B}T$, we see that $N/V = p/k_\text{B}T$. Therefore, at constant temperature, if the density and, consequently, the number density are reduced by half, the pressure must also be reduced by half, and $p_\text{f} = 0.500 \text{ atm.}$

17.3. Density is mass per unit volume, and volume is proportional to the size of a body (such as the radius of a sphere) cubed. So if the distance between molecules increases by a factor of 10, then the volume occupied increases by a factor of 1000, and the density decreases by a factor of 1000. Since we assume molecules are in contact in liquids and solids, the distance between their centers is on the order of their typical size, so the distance in gases is on the order of 10 times as great.

17.4. Yes. Such fluctuations actually occur for a body of any size in a gas, but since the numbers of molecules are immense for macroscopic bodies, the fluctuations are a tiny percentage of the number of collisions, and the averages spoken of in this section vary imperceptibly. Roughly speaking, the fluctuations are inversely proportional to the square root of the number of collisions, so for small bodies, they can become significant. This was actually observed in the nineteenth century for pollen grains in water and is known as Brownian motion.

17.5. In a liquid, the molecules are very close together, constantly colliding with one another. For a gas to be nearly ideal, as air is under ordinary conditions, the molecules must be very far apart. Therefore the mean free path is much longer in the air.

CONCEPTUAL QUESTIONS

1. Mercury and water are liquid at room temperature and atmospheric pressure. Air is a gas at room temperature and atmospheric pressure. Glass is an amorphous solid (non-crystalline) material at room temperature and atmospheric pressure. At one time, it was thought that glass flowed, but flowed very slowly. This theory came from the observation that old glass planes were thicker at the bottom. It is now thought unlikely that this theory is accurate.

3. Pressure is force divided by area. If a knife is sharp, the force applied to the cutting surface is divided over a smaller area than the same force applied with a dull knife. This means that the pressure would be greater for the sharper knife, increasing its ability to cut.

5. If the two chunks of ice had the same volume, they would produce the same volume of water. The glacier would cause the greatest rise in the lake, however, because part of the floating chunk of ice is already submerged in the lake, and is thus already contributing to the lake's level.

7. The pressure is acting all around your body, assuming you are not in a vacuum.

9. 2 moles, as that will contain twice as many molecules as the 1 mole of oxygen

11. pressure

13. The flame contains hot gas (heated by combustion). The pressure is still atmospheric pressure, in mechanical equilibrium with the air around it (or roughly so). The density of the hot gas is proportional to its number density N/V (neglecting the difference in composition between the gas in the flame and the surrounding air). At higher temperature than the surrounding air, the ideal gas

law says that $N/V = p/k_B T$ is less than that of the surrounding air. Therefore the hot air has lower density than the surrounding air and is lifted by the buoyant force.
15. The mean free path is inversely proportional to the square of the radius, so it decreases by a factor of 4. The mean free time is proportional to the mean free path and inversely proportional to the rms speed, which in turn is inversely proportional to the square root of the mass. That gives a factor of $\sqrt{8}$ in the numerator, so the mean free time decreases by a factor of $\sqrt{2}$.
17. Since they're more massive, their gravity is stronger, so the escape velocity from them is higher. Since they're farther from the Sun, they're colder, so the speeds of atmospheric molecules including hydrogen and helium are lower. The combination of those facts means that relatively few hydrogen and helium molecules have escaped from the outer planets.
19. a. false; b. true; c. true; d. true
21. 1200 K

PROBLEMS

22. 1.610 cm^3
24. The mass is 2.58 g. The volume of your body increases by the volume of air you inhale. The average density of your body decreases when you take a deep breath because the density of air is substantially smaller than the average density of the body.
26. 3.99 cm
28. 2.86 times denser
32. a. 0.137 atm; b. $p_g = (1\text{ atm})\frac{T_2 V_1}{T_1 V_2} - 1\text{ atm}.$ Because of the expansion of the glass, $V_2 = 0.99973$. Multiplying by that factor does not make any significant difference.
34. a. 1.79×10^{-3} mol; b. 0.227 mol; c. 1.08×10^{21} molecules for the nitrogen, 1.37×10^{23} molecules for the carbon dioxide
36. 7.84×10^{-2} mol
38. 1.87×10^3
40. 2.47×10^7 molecules
42. 6.95×10^5 Pa; 6.86 atm
44. a. 9.14×10^6 Pa; b. 8.22×10^6 Pa; c. 2.15 K; d. no
46. 40.7 km
48. a. 0.61 N; b. 0.20 Pa
50. a. 5.88 m/s; b. 5.89 m/s
52. 177 m/s
54. 4.54×10^3
56. a. 0.0352 mol; b. 5.65×10^{-21} J; c. 139 J
59. 458 K
61. 3.22×10^3 K
63. a. 1.004; b. 764 K; c. This temperature is equivalent to $915\,°\text{F}$, which is high but not impossible to achieve. Thus, this process is feasible. At this temperature, however, there may be other considerations that make the process difficult. (In general, uranium enrichment by gaseous diffusion is indeed difficult and requires many passes.)
64. About 0.072. Answers may vary slightly. A more accurate answer is 0.074.
65. a. 419 m/s; b. 472 m/s; c. 513 m/s
67. 541 K
69. 2400 K for all three parts

ADDITIONAL PROBLEMS

71. a. 2.21×10^{27} molecules/m^3; b. 3.67×10^3 mol/m^3

CHALLENGE PROBLEMS

76. 29.5 N/m

CHAPTER 18

CHECK YOUR UNDERSTANDING

18.1. eight
18.2. harder

18.3. Half-life is inversely related to decay rate, so the half-life is short. Activity depends on both the number of decaying particles and the decay rate, so the activity can be great or small.

CONCEPTUAL QUESTIONS

3. The nucleus of an atom is made of one or more nucleons. A nucleon refers to either a proton or neutron. A nuclide is a stable nucleus.

5. A bound system should have less mass than its components because of energy-mass equivalence $\left(E = mc^2\right)$. If the energy of a system is reduced, the total mass of the system is reduced. If two bricks are placed next to one another, the attraction between them is purely gravitational, assuming the bricks are electrically neutral. The gravitational force between the bricks is relatively small (compared to the strong nuclear force), so the mass defect is much too small to be observed. If the bricks are glued together with cement, the mass defect is likewise small because the electrical interactions between the electrons involved in the bonding are still relatively small.

7. Nucleons at the surface of a nucleus interact with fewer nucleons. This reduces the binding energy per nucleon, which is based on an average over all the nucleons in the nucleus.

9. That it is constant.

PROBLEMS

11. Use the rule $A = Z + N$.

	Atomic Number (*Z*)	Neutron Number (*N*)	Mass Number (*A*)
(a)	29	29	58
(b)	11	13	24
(c)	84	126	210
(d)	20	25	45
(e)	82	124	206

13. a. $r = r_0 A^{1/3}, \rho = \dfrac{3\text{ u}}{4\pi r_0{}^3}$;

b. $\rho = 2.3 \times 10^{17} \text{ kg/m}^3$

15. side length $= 1.6\ \mu\text{m}$

17. 92.4 MeV

19. $8.790 \text{ MeV} \approx \text{graph's value}$

21. a. 7.570 MeV; b. $7.591 \text{ MeV} \approx \text{graph's value}$

23. The decay constant is equal to the negative value of the slope or 10^{-9} s^{-1}. The half-life of the nuclei, and thus the material, is $T_{1/2} = 693 \text{ million years}$.

25. a. The decay constant is $\lambda = 1.99 \times 10^{-5} \text{ s}^{-1}$. b. Since strontium-91 has an atomic mass of 90.90 g, the number of nuclei in a 1.00-g sample is initially

$N_0 = 6.63 \times 10^{21} \text{ nuclei}.$

The initial activity for strontium-91 is

$$\begin{aligned} A_0 &= \lambda N_0 \\ &= 1.32 \times 10^{17} \text{ decays/s} \end{aligned}$$

The activity at $t = 15.0 \text{ h} = 5.40 \times 10^4 \text{ s}$ is

$A = 4.51 \times 10^{16} \text{ decays/s}.$

27. $1.20 \times 10^{-2} \text{ mol}$; $6.00 \times 10^{-3} \text{ mol}$; $3.75 \times 10^{-4} \text{ mol}$

29. a. 0.988 Ci; b. The half-life of ^{226}Ra is more precisely known than it was when the Ci unit was established.

31. a. $2.73 \mu\text{g}$; b. $9.76 \times 10^4 \text{ Bq}$

33. a. $7.46 \times 10^5 \text{ Bq}$; b. $7.75 \times 10^5 \text{ Bq}$

CHAPTER 19

CHECK YOUR UNDERSTANDING

19.1. The Greenland ice sheet has about 2.85 times as much ice as in the polar ice caps on Mars. They are about the same to the nearest power of 10.

CONCEPTUAL QUESTIONS

32. Gases of low molecular mass, e.g. H_2 or He
33. High surface temperature and low surface gravity

CHAPTER 20

CHECK YOUR UNDERSTANDING

20.1. The radius of Earth is 6371 km. Therefore,

$$\left(\frac{R_{\text{planet}}}{R_{\text{star}}}\right)^2 = \left(\frac{6371\text{ km}}{695{,}700/2\text{ km}}\right)^2 = \left(\frac{6371\text{ km}}{347{,}850\text{ km}}\right)^2 = 0.0003$$, or significantly less than 1%.

PROBLEMS

15.
a. 1.06 AU
b. 0.00346 M_{Sun}

CHAPTER 21

CHECK YOUR UNDERSTANDING

21.1. the conversion of mass to energy

CONCEPTUAL QUESTIONS

5. The nuclei produced in the fusion process have a larger binding energy per nucleon than the nuclei that are fused. That is, nuclear fusion decreases average energy of the nucleons in the system. The energy difference is carried away as radiation.

PROBLEMS

31. i.
$$\begin{aligned} {}^1_1\text{H} + {}^1_1\text{H} &\rightarrow {}^2_1\text{H} + e^+ + \nu_e \\ A_\text{i} &= 1 + 1 = 2;\ A_\text{f} = 2\ Z_\text{i} = 1 + 1 = 2; \\ Z_\text{f} &= 1 + 1 = 2 \end{aligned}$$

ii.
$$\begin{aligned} {}^1_1\text{H} + {}^2_1\text{H} &\rightarrow {}^3_2\text{H} + \gamma \\ A_\text{i} &= 1 + 2 = 3;\ A_\text{f} = 3 + 0 = 3\ Z_\text{i} = 1 + 1 = 2; \\ Z_E &= 1 + 1 = 2 \end{aligned}$$

iii.
$$\begin{aligned} {}^3_2\text{H} + {}^3_2\text{H} &\rightarrow {}^4_2\text{H} + {}^1_1\text{H} + {}^1_1\text{H} \\ A_\text{i} &= 3 + 3 = 6;\ A_\text{f} = 4 + 1 + 1 = 6\ Z_\text{i} = 2 + 2 = 4 \\ Z_\text{f} &= 2 + 1 + 1 = 4 \end{aligned}$$

33. 26.73 MeV
35. a. 3×10^{38} protons/s ; b. 6×10^{14} neutrinos/m$^2 \cdot$ s ;

This huge number is indicative of how rarely a neutrino interacts, since large detectors observe very few per day.

CHAPTER 22

CHECK YOUR UNDERSTANDING

22.1. In **Appendix D**, Sirius is listed with a luminosity 23 times that of the Sun. This value can be inserted into the mass-luminosity relationship to get the mass of Sirius:

$$M/M_{\text{Sun}} = 23^{0.25} = 2.2$$

The mass of the companion star to Sirius is then 3.2 – 2.2 = 1.0 solar mass.

CHAPTER 23

CHECK YOUR UNDERSTANDING

23.1.

$$\begin{aligned} 1 \text{ light-year} &= 9.46 \times 10^{12} \text{ km} \\ &= 9.46 \times 10^{12} \text{ km} \times \frac{1 \text{ Earth diameter}}{12{,}700 \text{ km}} \\ &= 7.45 \times 10^{8} \text{ Earth diameters} \end{aligned}$$

That means that 1 light-year is about 745 million times the diameter of Earth.

23.2. 206,265 AU

23.3. using **Equation 23.9**: $d = 10 \times 10^{0.2(0.9 - (-3.6))} = 10 \times 10^{0.9} = 79.4$ pc So, Spica is located about 80 parsecs from Earth.

CONCEPTUAL QUESTIONS

26. The closest of the stars listed is Alpha Centauri A, because the value of the quantity (m-M) = -4.4, which is the lowest value for the numbers in the table. The farthest of the stars is Antares, because the value of the quantity (m-M) = +5.4 which is the largest value for the numbers in the table.

PROBLEMS

43. d = 2.75 pc

CHAPTER 24

CHECK YOUR UNDERSTANDING

24.1. The Jeans mass is now $M_{\text{Jeans}} = 18 M_{\text{Sun}} \sqrt{\frac{30^3}{30000}} = 17 M_{\text{Sun}}$ So, a cloud exceeding a mass of 17 solar masses would be$()$ sufficient to form a star.

24.2. The temperature would increase by a factor of $256^{0.25}$ (that is, the 4^{th} root of 256), or 4 times.

24.3. That the cluster is nearer to Earth than 10 parsecs.

PROBLEMS

51. 93 M_{Sun}

56. d = 136 pc

CHAPTER 25

CHECK YOUR UNDERSTANDING

25.1. $g(\text{white dwarf}) = \frac{(G \times 2M_{\text{Sun}})}{(0.5R_{\text{Earth}})^2} = \frac{\left(6.67 \times 10^{-11} \text{ m}^2/\text{kg s}^2 \times 4 \times 10^{30} \text{ kg}\right)}{\left(3.2 \times 10^6\right)^2} = 2.61 \times 10^7 \text{ m/s}^2$

25.2. Substituting the data into our equation gives

$$R_S = \frac{2GM}{c^2} = \frac{2(6.67 \times 10^{-11} \text{ N} \cdot \text{m}^2/\text{kg}^2)(3000 \text{ kg})}{(3.00 \times 10^8 \text{ m/s})^2} = 1.33 \times 10^{-23} \text{ m.}$$

For comparison, the size of a proton is usually considered to be about 8×10^{-16} m, which would be about ten million times larger.

CHAPTER 26

PROBLEMS

25. 510,000 M_{Sun}

CHAPTER 27

PROBLEMS

30. d = 263 kpc

CHAPTER 28

CHECK YOUR UNDERSTANDING

28.1. Because this is the same galaxy, we could pick any one of the four wavelengths and calculate how much it has shifted. If we use a rest wavelength of 410 nm and compare it to the shifted wavelength of 492 nm, we see that (28.4)

$$z = \frac{\Delta\lambda}{\lambda} = \frac{(492\text{ nm} - 410\text{ nm})}{410\text{ nm}} = \frac{82\text{ nm}}{410\text{ nm}} = 0.20$$

In the classical view, this galaxy is receding at 20% of the speed of light; however, at 20% of the speed of light, relativistic effects are starting to become important. So, using the relativistic Doppler equation, we compute the true recession rate as (28.5)

$$\frac{v}{c} = \frac{(z+1)^2 - 1}{(z+1)^2 + 1} = \frac{(0.2+1)^2 - 1}{(0.2+1)^2 + 1} = \frac{1.44 - 1}{1.44 + 1} = \frac{0.44}{2.44} = 0.18$$

Therefore, the actual recession speed is only 18% of the speed of light. While this may not initially seem like a big difference from the classical measurement, there is already an 11% deviation between the classical and the relativistic solutions; and at greater recession speeds, the divergence between the classical and relativistic speeds increases rapidly!

28.2. The total volume covered is $(4/3)\pi \times (90\text{ million light-years})^3 = 3.05 \times 10^{24}$ light-years^3. The survey reaches 3 times as far in distance, so it will cover a volume that is $3^3 = 27$ times larger.

CHAPTER 29

CHECK YOUR UNDERSTANDING

29.1. a. In this case, the average mass-energy in a volume V of space is $E = \rho_{\text{crit}}V$. Thus, for space with critical density, we require that (29.7)

$$V = \frac{E_{\text{grain}}}{\rho_{\text{crit}}} = \frac{1.1 \times 10^{-13}\text{ kg}}{9.6 \times 10^{-26}\text{ kg/m}^3} = 1.15 \times 10^{12}\text{ m}^3 = (10{,}500\text{ m})^3 \cong (10.5\text{ km})^3$$

Thus, the sides of a cube of space with mass-energy density averaging that of the critical density would need to be slightly greater than 10 km to contain the total energy equal to a single grain of dust! b. Since the critical density goes as the square of the Hubble constant, by doubling the Hubble parameter, the critical density would increase by a factor a four. So if the Hubble constant was 44 km/s per million light-years instead of 22 km/s per million light-years, the critical density would be $\rho_{\text{crit}} = 4 \times 9.6 \times 10^{-27}\text{ kg/m}^3 = 3.8 \times 10^{-26}\text{ kg/m}^3$.

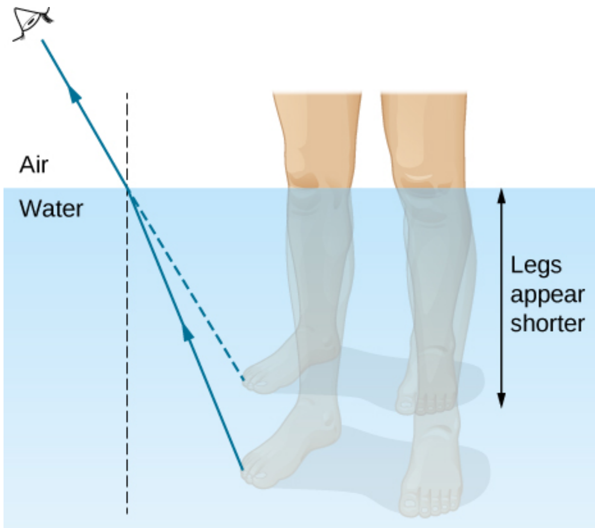
CHAPTER 30

CHECK YOUR UNDERSTANDING

30.1. 2.1% (to two significant figures)
30.2. $15.1°$
30.3. air to water, because the condition that the second medium must have a smaller index of refraction is not satisfied
30.4. 9.3 cm

CONCEPTUAL QUESTIONS

1. Light can be modeled as a ray when devices are large compared to wavelength, and as a wave when devices are comparable or small compared to wavelength.
3. This fact simply proves that the speed of light is greater than that of sound. If one knows the distance to the location of the lightning and the speed of sound, one could, in principle, determine the speed of light from the data. In practice, because the speed of light is so great, the data would have to be known to impractically high precision.
5. Powder consists of many small particles with randomly oriented surfaces. This leads to diffuse reflection, reducing shine.
7. “toward” when increasing n (air to water, water to glass); “away” when decreasing n (glass to air)
9. A ray from a leg emerges from water after refraction. The observer in air perceives an apparent location for the source, as if a ray traveled in a straight line. See the dashed ray below.



11. The gemstone becomes invisible when its index of refraction is the same, or at least similar to, the water surrounding it. Because diamond has a particularly high index of refraction, it can still sparkle as a result of total internal reflection, not invisible.
13. One can measure the critical angle by looking for the onset of total internal reflection as the angle of incidence is varied. Equation 30.14 can then be applied to compute the index of refraction.
15. In addition to total internal reflection, rays that refract into and out of diamond crystals are subject to dispersion due to varying values of n across the spectrum, resulting in a sparkling display of colors.

PROBLEMS

17. 2.99705×10^8 m/s ; 1.97×10^8 m/s
19. ice at $0°C$
21. 1.03 ns
23. 337 m
25. proof
27. proof
29. reflection, $70°$; refraction, $45°$
31. $42°$
33. 1.53
35. a. 2.9 m; b. 1.4 m
37. a. $24.42°$; b. $31.33°$
39. $79.11°$
41. a. 1.43, fluorite; b. $44.2°$
43. a. $48.2°$; b. $27.3°$
45. $46.5°$ for red, $46.0°$ for violet
47. a. $0.04°$; b. 1.3 m
49. $72.8°$
51. $53.5°$ for red, $55.2°$ for violet

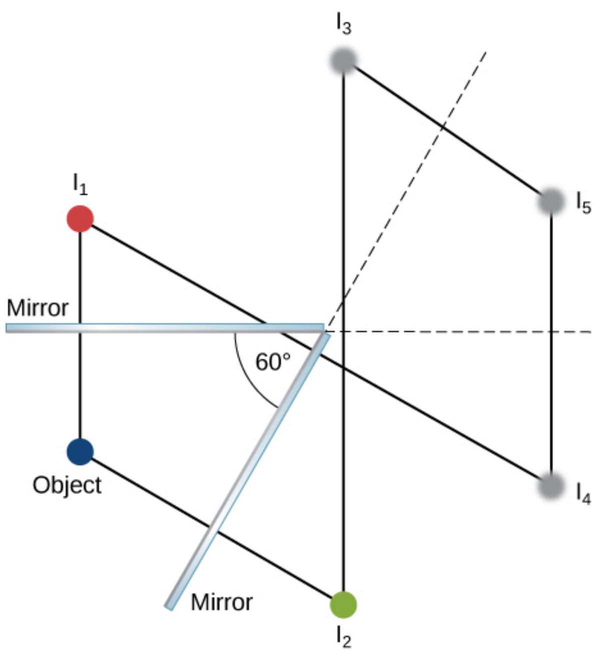
CHAPTER 31

CONCEPTUAL QUESTIONS

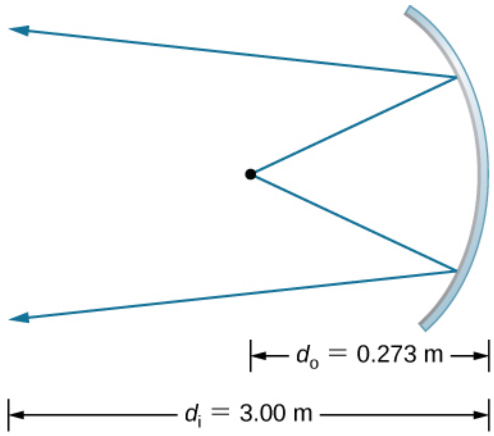
1. Virtual image cannot be projected on a screen. You cannot distinguish a real image from a virtual image simply by judging from the image perceived with your eye.
3. Yes, you can photograph a virtual image. For example, if you photograph your reflection from a plane mirror, you get a photograph of a virtual image. The camera focuses the light that enters its lens to form an image; whether the source of the light is a real object or a reflection from mirror (i.e., a virtual image) does not matter.
5. No, you can see the real image the same way you can see the virtual image. The retina of your eye effectively serves as a screen.
7. The mirror should be half your size and its top edge should be at the level of your eyes. The size does not depend on your distance from the mirror.

9. when the object is at infinity; see the mirror equation
11. Yes, negative magnification simply means that the image is upside down; this does not prevent the image from being larger than the object. For instance, for a concave mirror, if distance to the object is larger than one focal distance but smaller than two focal distances the image will be inverted and magnified.
13. answers may vary
15. The focal length of the lens is fixed, so the image distance changes as a function of object distance.
17. Yes, the focal length will change. The lens maker's equation shows that the focal length depends on the index of refraction of the medium surrounding the lens. Because the index of refraction of water differs from that of air, the focal length of the lens will change when submerged in water.
18. Microscopes create images of macroscopic size, so geometric optics applies.
20. The eyepiece would be moved slightly farther from the objective so that the image formed by the objective falls just beyond the focal length of the eyepiece.

PROBLEMS



22.
24. It is in the focal point of the big mirror and at the center of curvature of the small mirror.
26. $f=\frac{R}{2}\Rightarrow R=+1.60\text{ m}$
28. $d_o=27.3\text{ cm}$
30. Step 1: Image formation by a mirror is involved.
Step 2: Draw the problem set up when possible.
Step 3: Use thin-lens equations to solve this problem.
Step 4: Find f.
Step 5: Given: $m=1.50,\ d_o=0.120\text{ m}$.
Step 6: No ray tracing is needed.
Step 7: Using $m=\frac{d_i}{d_o},\ d_i=-0.180\text{ m}$. Then, $f=0.360\text{ m}$.
Step 8: The image is virtual because the image distance is negative. The focal length is positive, so the mirror is concave.
32. a. for a convex mirror $d_i<0\Rightarrow m>0.\ m=+0.111$; b. $d_i=-0.334\text{ cm}$ (behind the cornea);
c. $f=-0.376\text{ cm}$, so that $R=-0.752\text{ cm}$
34. $m=\frac{h_i}{h_o}=-\frac{d_i}{d_o}=-\frac{-d_o}{d_o}=\frac{d_o}{d_o}=1\Rightarrow h_i=h_o$



36.

$m = -11.0$
$A' = 0.110\ \text{m}^2$
$I = 6.82\ \text{kW/m}^2$

38. $x_{2m} = -x_{2m-1}, \quad (m = 1, 2, 3, ...),$
$x_{2m+1} = b - x_{2m}, \quad (m = 0, 1, 2, ...),$ with $x_0 = a.$

40. $d_i = -55$ cm; $m = +1.8$

42. $d_i = -41$ cm, $m = 1.4$

44. proof

46. a. $\frac{1}{d_i} + \frac{1}{d_o} = \frac{1}{f} \Rightarrow d_i = 3.43$ m;

b. $m = -33.33$, so that $(2.40 \times 10^{-2}\ \text{m})(33.33) = 80.0$ cm, and $(3.60 \times 10^{-2}\ \text{m})(33.33) = 1.20\ \text{m} \Rightarrow 0.800\ \text{m} \times 1.20\ \text{m}$ or $80.0\ \text{cm} \times 120\ \text{cm}$

48. a. $\frac{1}{d_o} + \frac{1}{d_i} = \frac{1}{f}$;
$d_i = 5.08$ cm

b. $m = -1.695 \times 10^{-2}$, so the maximum height is $\frac{0.036\ \text{m}}{1.695 \times 10^{-2}} = 2.12\ \text{m} \Rightarrow 100\%$;

c. This seems quite reasonable, since at 3.00 m it is possible to get a full length picture of a person.

50. a. $\frac{1}{d_o} + \frac{1}{d_i} = \frac{1}{f} \Rightarrow d_o = 2.55$ m;

b. $\frac{h_i}{h_o} = -\frac{d_i}{d_o} \Rightarrow h_o = 1.00$ m

52. a. Using $\frac{1}{d_o} + \frac{1}{d_i} = \frac{1}{f}$, $d_i = -56.67$ cm. Then we can determine the magnification, $m = 6.67$. b. $d_i = -190$ cm and $m = +20.0$; c. The magnification m increases rapidly as you increase the object distance toward the focal length.

54.

$$\begin{aligned} \frac{1}{d_o} + \frac{1}{d_i} &= \frac{1}{f} \\ d_i &= \frac{1}{(1/f) - (1/d_o)} \\ \frac{d_i}{d_o} &= 6.667 \times 10^{-13} = \frac{h_i}{h_o} \\ h_i &= -0.933\ \text{mm} \end{aligned}$$

56. $d_i = -6.7$ cm
$h_i = 4.0$ cm

58. 83 cm to the right of the converging lens, $m = -2.3$, $h_i = 6.9$ cm

60. $M = 6\times$

62.
$$\begin{aligned} M &= \left(\frac{25\text{ cm}}{L}\right)\left(1 + \frac{L-\ell}{f}\right) \\ L - \ell &= d_\text{o} \\ d_\text{o} &= 13\text{ cm} \end{aligned}$$

64. $M = 2.5\times$

66. $M \quad = -2.1\times$

68.
$$\begin{aligned} M &= \frac{25\text{ cm}}{f} \\ M_\text{max} &= 5 \end{aligned}$$

70.
$$\begin{aligned} M_\text{max}^\text{young} &= 1 + \frac{18\text{ cm}}{f} \Rightarrow f = \frac{18\text{ cm}}{M_\text{max}^\text{young} - 1} \\ M_\text{max}^\text{old} &= 9.8\times \end{aligned}$$

72. a.
$$\begin{aligned} \frac{1}{d_\text{o}} + \frac{1}{d_\text{i}} &= \frac{1}{f} \Rightarrow d_\text{i} = 4.65\text{ cm} \\ &\Rightarrow m = -30.0 \end{aligned}$$
;

b. $M_\text{net} = -240$

74. a. $\frac{1}{d_\text{o}^\text{obj}} + \frac{1}{d_\text{i}^\text{obj}} = \frac{1}{f^\text{obj}} \Rightarrow d_\text{i}^\text{obj} = 18.3\text{ cm}$ behind the objective lens;

b. $m^\text{obj} = -60.0$;

c.
$$\begin{aligned} d_\text{o}^\text{eye} &= 1.70\text{ cm} \\ d_\text{i}^\text{eye} &= -11.3\text{ cm} \end{aligned}$$

in front of the eyepiece; d. $M^\text{eye} = 13.5$;

e. $M_\text{net} = -810$

76. $M = -40.0$

78. $f^\text{obj} = \frac{R}{2}, M = -1.67$

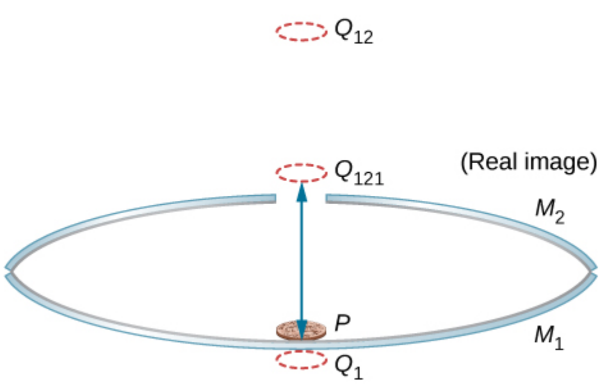
80. $M = -\frac{f^\text{obj}}{f^\text{eye}}, f^\text{eye} = +10.0\text{ cm}$

82. Answers will vary.

84. 12 cm to the left of the mirror, $m = 3/5$

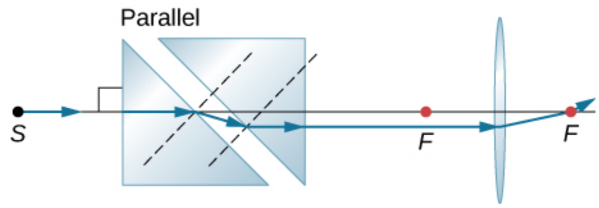
86. 27 cm in front of the mirror, $m = 0.6$, $h_i = 1.76\text{ cm}$, orientation upright

88. The following figure shows three successive images beginning with the image Q_1 in mirror M_1. Q_1 is the image in mirror M_1, whose image in mirror M_2 is Q_{12} whose image in mirror M_1 is the real image Q_{121}.

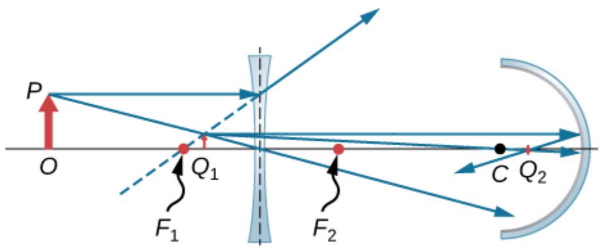


90. 5.4 cm from the axis

92. Let the vertex of the concave mirror be the origin of the coordinate system. Image 1 is at −10/3 cm (−3.3 cm), image 2 is at −40/11 cm (−3.6 cm). These serve as objects for subsequent images, which are at −310/83 cm (−3.7 cm), −9340/2501 cm (−3.7 cm), −140,720/37,681 cm (−3.7 cm). All remaining images are at approximately −3.7 cm.



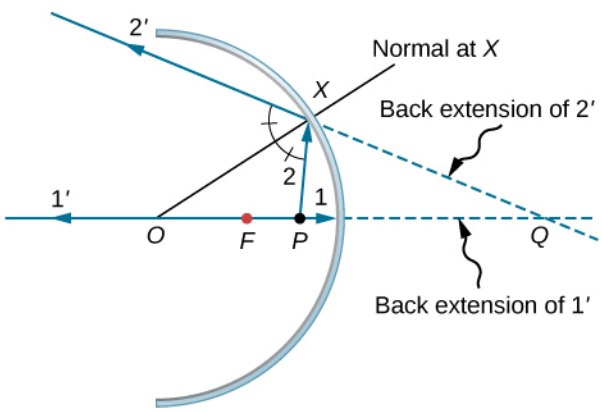
94.



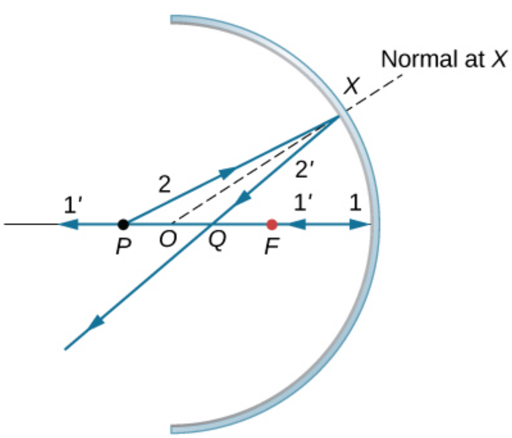
96.
98. −5 D
100. 11

ADDITIONAL PROBLEMS

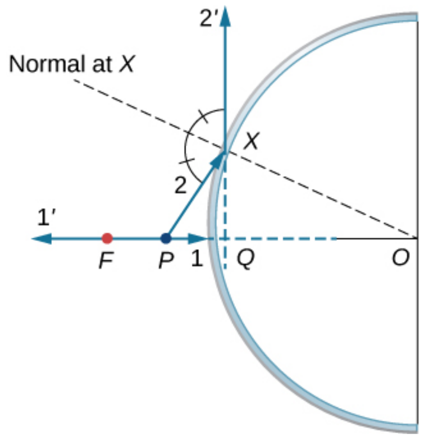
102. a.



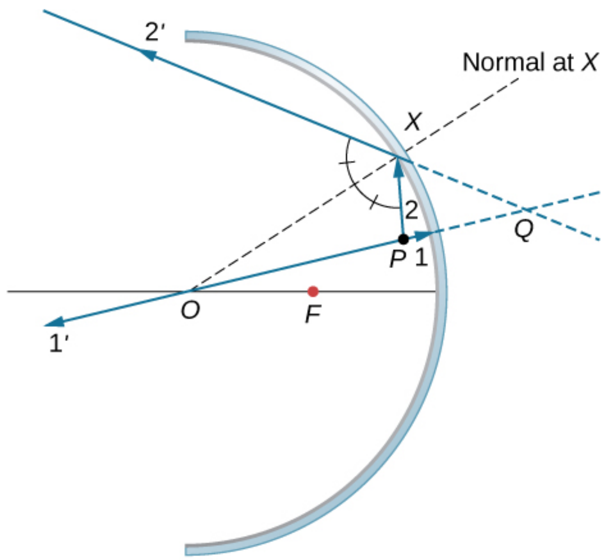
b.

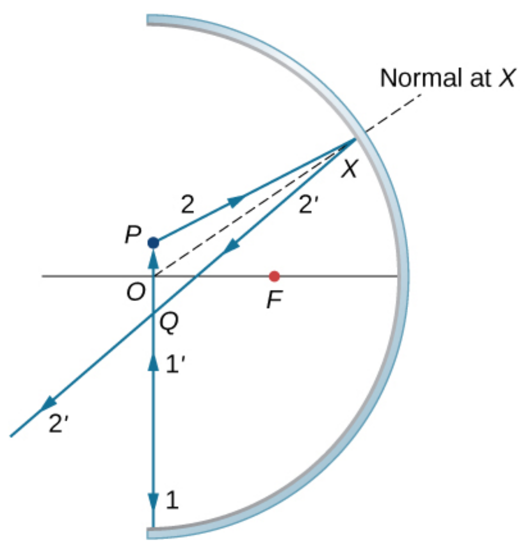


c.



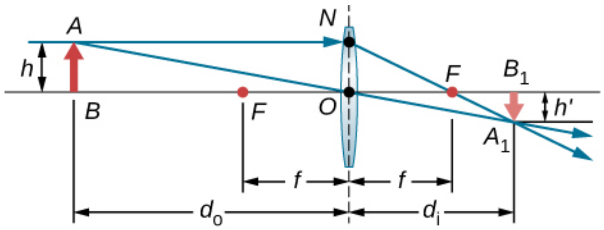
d. similar to the previous picture but with point *P* outside the focal length; e. Repeat (a)–(d) for a point object off the axis. For a point object placed off axis in front of a concave mirror corresponding to parts (a) and (b), the case for convex mirror left as exercises.





104. $d_i = -10/3\ \text{cm},\ h_i = 2\ \text{cm}$, upright

106. proof



108.

Triangles *BAO* and B_1A_1O are similar triangles. Thus, $\frac{A_1B_1}{AB} = \frac{d_i}{d_o}$.Triangles *NOF* and B_1A_1F are similar triangles. Thus, $\frac{NO}{f} = \frac{A_1B_1}{d_i - f}$. Noting that $NO = AB$ gives $\frac{AB}{f} = \frac{A_1B_1}{d_i - f}$ or $\frac{AB}{A_1B_1} = \frac{f}{d_i - f}$. Inverting this gives $\frac{A_1B_1}{AB} = \frac{d_i - f}{f}$. Equating the two expressions for the ratio $\frac{A_1B_1}{AB}$ gives $\frac{d_i}{d_o} = \frac{d_i - f}{f}$. Dividing through by d_i gives $\frac{1}{d_o} = \frac{1}{f} - \frac{1}{d_i}$ or $\frac{1}{d_o} + \frac{1}{d_i} = \frac{1}{f}$.

110. 70 cm

112. The plane mirror has an infinite focal point, so that $d_i = -d_o$. The total apparent distance of the man in the mirror will be his actual distance, plus the apparent image distance, or $d_o + (-d_i) = 2d_o$. If this distance must be less than 20 cm, he should stand at $d_o = 10\ \text{cm}$.

114. Here we want $d_o = 25\ \text{cm} - 2.20\ \text{cm} = 0.228\ \text{m}$. If $x =$ near point, $d_i = -(x - 0.0220\ \text{m})$. Thus, $P = \frac{1}{d_o} + \frac{1}{d_i} = \frac{1}{0.228\ \text{m}} + \frac{1}{x - 0.0220\ \text{m}}$. Using $P = 0.75\ \text{D}$ gives $x = 0.253\ \text{m}$, so the near point is 25.3 cm.

116. Assuming a lens at 2.00 cm from the boy's eye, the image distance must be $d_i = -(500\ \text{cm} - 2.00\ \text{cm}) = -498\ \text{cm}$. For an infinite-distance object, the required power is $P = \frac{1}{d_i} = -0.200\ \text{D}$. Therefore, the $-4.00\ \text{D}$ lens will correct the nearsightedness.

118. $87\ \mu\text{m}$

120. Use, $M_{\text{net}} = -\frac{d_i^{\text{obj}}\left(f^{\text{eye}} + 25\ \text{cm}\right)}{f^{\text{obj}}f^{\text{eye}}}$. The image distance for the objective is $d_i^{\text{obj}} = -\frac{M_{\text{net}}f^{\text{obj}}f^{\text{eye}}}{f^{\text{eye}} + 25\ \text{cm}}$. Using $f^{\text{obj}} = 3.0\ \text{cm},\ f^{\text{eye}} = 10\ \text{cm, and}\ M = -10$ gives $d_i^{\text{obj}} = 8.6\ \text{cm}$. We want this image to be at the focal point of the eyepiece so that the eyepiece forms an image at infinity for comfortable viewing. Thus, the distance *d* between the lenses should be $d = f^{\text{eye}} + d_i^{\text{obj}} = 10\ \text{cm} + 8.6\ \text{cm} = 19\ \text{cm}$.

122. a. focal length of the corrective lens $f_c = -80\ \text{cm}$; b. −1.25 D

124. 2×10^{16} km

126. 10^5 m

CHAPTER 32

CHECK YOUR UNDERSTANDING

32.1. $3.63°$ and $7.27°$, respectively

32.2. a. 853 nm, 1097 nm; b. 731 nm, 975 nm

32.3. a. too small; b. up to 8×10^{-5}

CONCEPTUAL QUESTIONS

1. No. Two independent light sources do not have coherent phase.

3. Because both the sodium lamps are not coherent pairs of light sources. Two lasers operating independently are also not coherent so no interference pattern results.

5. Monochromatic sources produce fringes at angles according to $d \sin \theta = m\lambda$. With white light, each constituent wavelength will produce fringes at its own set of angles, blending into the fringes of adjacent wavelengths. This results in rainbow patterns.

7. Differing path lengths result in different phases at destination resulting in constructive or destructive interference accordingly. Reflection can cause a $180°$ phase change, which also affects how waves interfere. Refraction into another medium changes the wavelength inside that medium such that a wave can emerge from the medium with a different phase compared to another wave that travelled the same distance in a different medium.

9. Phase changes occur upon reflection at the top of glass cover and the top of glass slide only.

11. The surface of the ham being moist means there is a thin layer of fluid, resulting in thin-film interference. Because the exact thickness of the film varies across the piece of ham, which is illuminated by white light, different wavelengths produce bright fringes at different locations, resulting in rainbow colors.

13. Other wavelengths will not generally satisfy $t = \frac{\lambda/n}{4}$ for the same value of *t* so reflections will result in completely destructive interference. For an incidence angle θ, the path length inside the coating will be increased by a factor $1/\cos\theta$ so the new condition for destructive interference becomes $\frac{t}{\cos\theta} = \frac{\lambda/n}{4}$.

15. In one arm, place a transparent chamber to be filled with the gas. See **Example 32.6**.

PROBLEMS

17. $0.997°$

19. $0.290\ \mu\text{m}$

21. $5.77 \times 10^{-7}\ \text{m} = 577\ \text{nm}$

23. 62.5; since *m* must be an integer, the highest order is then $m = 62$.

25. $1.44\ \mu\text{m}$

27. a. $20.3°$; b. $4.98°$; c. 5.76, the highest order is $m = 5$.

29. a. 2.37 cm; b. 1.78 cm

31. 560 nm

33. 1.2 mm

35. a. $0.40°,\ 0.53°;$ b. $4.6 \times 10^{-3}\ \text{m}$

37. 1:9

39. 532 nm (green)

41. $8.39 \times 10^{-8}\ \text{m} = 83.9\ \text{nm}$

43. 620 nm (orange)

45. 380 nm

47. a. Assuming *n* for the plane is greater than 1.20, then there are two phase changes: 0.833 cm. b. It is too thick, and the plane would be too heavy. c. It is unreasonable to think the layer of material could be any thickness when used on a real aircraft.

49. $4.55 \times 10^{-4}\ \text{m}$

51. $D = 2.53 \times 10^{-6}\ \text{m}$

ADDITIONAL PROBLEMS

53. $0.29°$ and $0.86°$

55. a. 4.26 cm; b. 2.84 cm

57. 6

59. 0.20 m

61. 0.0839 mm

63. a. 9.8, 10.4, 11.7, and 15.7 cm; b. 3.9 cm

65. $0.0575°$

67. 700 nm

69. 189 nm

71. a. green (504 nm); b. magenta (white minus green)

73. 1.29

75. $52.7\ \mu\text{m}$ and $53.0\ \mu\text{m}$

77. 160 nm

79. 413 nm and 689 nm

81. $73.9\ \mu\text{m}$

83. 47

85. $8.5\ \mu\text{m}$

87. $0.013°\text{C}$

CHALLENGE PROBLEMS

89. Bright and dark fringes switch places.
91. The path length must be less than one-fourth of the shortest visible wavelength in oil. The thickness of the oil is half the path length, so it must be less than one-eighth of the shortest visible wavelength in oil. If we take 380 nm to be the shortest visible wavelength in air, 33.9 nm.

93. $4.42 \times 10^{-5}\ \text{m}$

95. for one phase change: 950 nm (infrared); for three phase changes: 317 nm (ultraviolet); Therefore, the oil film will appear black, since the reflected light is not in the visible part of the spectrum.

CHAPTER 33

CHECK YOUR UNDERSTANDING

33.1. $17.8°$, $37.7°$, $66.4°$; no

33.2. $74.3°$, $0.0083 I_0$

33.3. From $d \sin \theta = m\lambda$, the interference maximum occurs at $2.87°$ for $m = 20.$ From **Equation 33.1**, this is also the angle for the second diffraction minimum. (*Note:* Both equations use the index *m* but they refer to separate phenomena.)

33.4. $3.332 \times 10^{-6}\ \text{m}$ or 300 lines per millimeter

33.5. $8.4 \times 10^{-4}\ \text{rad}$, 3000 times broader than the Hubble Telescope

33.6. also 90.0%

33.7. There will be only refraction but no reflection.

CONCEPTUAL QUESTIONS

1. The diffraction pattern becomes wider.
3. Walkie-talkies use radio waves whose wavelengths are comparable to the size of the hill and are thus able to diffract around the hill. Visible wavelengths of the flashlight travel as rays at this size scale.
5. The diffraction pattern becomes two-dimensional, with main fringes, which are now spots, running in perpendicular directions and fainter spots in intermediate directions.
8. blue; The shorter wavelength of blue light results in a smaller angle for diffraction limit.
10. No, these distances are three orders of magnitude smaller than the wavelength of visible light, so visible light makes a poor probe for atoms.
12. No. Sound waves are not transverse waves.
14. Energy is absorbed into the filters.
16. Sunsets are viewed with light traveling straight from the Sun toward us. When blue light is scattered out of this path, the remaining red light dominates the overall appearance of the setting Sun.
18. The axis of polarization for the sunglasses has been rotated $90°$.

PROBLEMS

19. a. $33.4°$; b. no

21. a. $1.35 \times 10^{-6}\ \text{m}$; b. $69.9°$

23. 750 nm
25. 2.4 mm, 4.7 mm
27. a. 1.00λ; b. 50.0λ; c. 1000λ

29. 1.92 m
31. $45.1°$

33. $I/I_0 = 2.2 \times 10^{-5}$

35. $0.63 I_0,\ 0.11 I_0,\ 0.0067 I_0,\ 0.0062 I_0,\ 0.00088 I_0$

37. 0.200
39. 3
41. 9
43. $5.97°$

45. 8.99×10^{3}

47. 707 nm

49. a. $11.8°$, $12.5°$, $14.1°$, $19.2°$; b. $24.2°$, $25.7°$, $29.1°$, $41.0°$; c. Decreasing the number of lines per centimeter by a factor of *x* means that the angle for the *x*-order maximum is the same as the original angle for the first-order maximum.
51. a. using $\lambda = 700\,\text{nm}, \theta = 5.0°$; b. using $\lambda = 460\,\text{nm}, \theta = 3.3°$
53. a. 26,300 lines/cm; b. yes; c. no
55. 1.13×10^{-2} m
57. 107 m
59. a. 7.72×10^{-4} rad; b. 23.2 m; c. 590 km
61. a. 2.24×10^{-4} rad; b. 5.81 km; c. 0.179 mm; d. can resolve details 0.2 mm apart at arm's length
63. $2.9\,\mu\text{m}$
65. 6.0 cm
67. 7.71 km
69. 1.0 m
71. 1.2 cm or closer
73. no
75. 0.500
77. 0.125 or 1/8
79. $84.3°$
81. $0.250\,I_0$
83. a. 0.500; b. 0.250; c. 0.187
85. $67.54°$
87. $53.1°$

ADDITIONAL PROBLEMS

89. 114 radian/s
91. 3.72 mm
93. $41.2°$
95. a. 1.92. The gem is not a diamond (it is zircon). b. $55.2°$
97. a. 0.898; b. We cannot have $n < 1.00$, since this would imply a speed greater than *c*. c. The refracted angle is too big relative to the angle of incidence.
99. $0.707\,B_1$
101. a. $1.69 \times 10^{-2}\,°\text{C/s}$; b. yes

CHALLENGE PROBLEMS

103. First part: $88.6°$. The remainder depends on the complexity of the solution the reader constructs.
105. proof; 1.33
107. a. 0.750; b. 0.563; c. 1.33

INDEX

Symbols

A

B

C

D

H

I

J

K

ATTRIBUTIONS

Collection: **The New Cosmic Universe**
Edited by: Steven Mellema
URL: **https://legacy.cnx.org/content/col31908/1.2/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: The Cosmic Universe <**http://legacy.cnx.org/content/col12169/1.81**> arranged by Steven Mellema.

Module: **New Cosmic Universe Preface**
Used here as: Preface
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m85849/1.3/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Cosmic Universe Preface <**http://legacy.cnx.org/content/m63936/1.18**> by Steven Mellema.

Module: **New Cosmic Universe Beginning Your Study of Physics**
Used here as: Introduction
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m86165/1.2/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Beginning Your Study of Physics <**http://legacy.cnx.org/content/m63978/1.14**> by Steven Mellema.

Module: **Cosmic Universe Introducing Astrophysics**
Used here as: Introducing Astrophysics
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m63947/1.12/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Introduction <**http://legacy.cnx.org/content/m58268/1.5**> by OpenStax.

Module: **Cosmic Universe The Scope and Scale of Physics**
Used here as: The Scope and Scale of Physics
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m64024/1.2/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: The Scope and Scale of Physics <**http://legacy.cnx.org/content/m58269/1.6**> by OpenStax.

Module: **Cosmic Universe Units and Standards**
Used here as: Units and Standards
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m64027/1.1/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Units and Standards <**http://legacy.cnx.org/content/m58270/1.5**> by OpenStax.

Module: **Cosmic Universe Unit Conversion**
Used here as: Unit Conversion
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m63948/1.4/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Unit Conversion <**http://legacy.cnx.org/content/m58271/1.5**> by OpenStax.

Module: **Cosmic Universe Dimensional Analysis**

Used here as: Dimensional Analysis
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77359/1.2/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Dimensional Analysis <**http://legacy.cnx.org/content/m58272/1.10**> by OpenStax.

Module: **Cosmic Universe Estimates and Fermi Calculations**
Used here as: Estimates and Fermi Calculations
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m64025/1.6/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Estimates and Fermi Calculations <**http://legacy.cnx.org/content/m58273/1.5**> by OpenStax.

Module: **Cosmic Universe Significant Figures**
Used here as: Uncertainties and Significant Figures
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m63995/1.6/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Significant Figures <**http://legacy.cnx.org/content/m58274/1.5**> by OpenStax.

Module: **Cosmic Universe Solving Problems in Physics**
Used here as: Solving Problems in Physics
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77360/1.1/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Solving Problems in Physics <**http://legacy.cnx.org/content/m58275/1.11**> by OpenStax.

Module: **Cosmic Universe Introduction to The Universe at Its Limits**
Used here as: Introduction
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m63977/1.9/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Numbers in Astronomy <**http://legacy.cnx.org/content/m59751/1.5**> by OpenStax Astronomy.

Module: **Cosmic Universe Consequences of Light Travel Time**
Used here as: Consequences of Light Travel Time
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m63983/1.2/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Consequences of Light Travel Time <**http://legacy.cnx.org/content/m59752/1.4**> by OpenStax Astronomy.

Module: **Cosmic Universe A Tour of the Universe**
Used here as: A Tour of the Universe
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m63984/1.2/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: A Tour of the Universe <**http://legacy.cnx.org/content/m59753/1.4**> by OpenStax Astronomy.

Module: **Cosmic Universe The Universe on the Large Scale**
Used here as: The Universe on the Large Scale

By: Steven Mellema
URL: **https://legacy.cnx.org/content/m63985/1.1/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: The Universe on the Large Scale <**http://legacy.cnx.org/content/m59754/1.4**> by OpenStax Astronomy.

Module: **Cosmic Universe The Universe of the Very Small**
Used here as: The Universe of the Very Small
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m63986/1.4/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: The Universe of the Very Small <**http://legacy.cnx.org/content/m59760/1.4**> by OpenStax Astronomy.

Module: **Cosmic Universe A Conclusion and a Beginning**
Used here as: A Conclusion and a Beginning
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m63987/1.3/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: A Conclusion and a Beginning <**http://legacy.cnx.org/content/m59762/1.4**> by OpenStax Astronomy.

Module: **Cosmic Universe Introduction to Motion Along a Straight Line**
Used here as: Introduction
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m63994/1.2/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Introduction <**http://legacy.cnx.org/content/m58281/1.4**> by OpenStax.

Module: **Cosmic Universe Position, Displacement, and Average Velocity**
Used here as: Position, Displacement and Average Velocity
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m64026/1.2/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Position, Displacement, and Average Velocity <**http://legacy.cnx.org/content/m58282/1.4**> by OpenStax.

Module: **Cosmic Universe Instantaneous Velocity and Speed**
Used here as: Instantaneous Velocity and Speed
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m63980/1.3/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Instantaneous Velocity and Speed <**http://legacy.cnx.org/content/m58283/1.4**> by OpenStax.

Module: **Cosmic Universe Average and Instantaneous Acceleration**
Used here as: Average and Instantaneous Acceleration
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m63981/1.3/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Average and Instantaneous Acceleration <**http://legacy.cnx.org/content/m58284/1.4**> by OpenStax.

Module: **Cosmic Universe Motion with Constant Acceleration**
Used here as: Motion with Constant Acceleration
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m71622/1.1/**

Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Motion with Constant Acceleration <**http://legacy.cnx.org/content/m58285/1.13**> by OpenStax.

Module: **Cosmic Universe Free Fall**
Used here as: Free Fall
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m64004/1.3/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Free Fall <**http://legacy.cnx.org/content/m58286/1.4**> by OpenStax.

Module: **Cosmic Universe Introduction to Rotation**
Used here as: Introduction
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m64031/1.11/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Introduction <**http://legacy.cnx.org/content/m58325/1.3**> by OpenStax.

Module: **Cosmic Universe Rotational Variables**
Used here as: Rotational Variables
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m64032/1.6/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Rotational Variables <**http://legacy.cnx.org/content/m58326/1.3**> by OpenStax.

Module: **Cosmic Universe Rotation with Constant Angular Acceleration**
Used here as: Rotation with Constant Angular Acceleration
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m64036/1.6/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Rotation with Constant Angular Acceleration <**http://legacy.cnx.org/content/m58327/1.4**> by OpenStax.

Module: **Cosmic Universe Relating Angular and Translational Quantities**
Used here as: Relating Angular and Translational Quantities
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m64037/1.8/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Relating Angular and Translational Quantities <**http://legacy.cnx.org/content/m58328/1.3**> by OpenStax.

Module: **Cosmic Universe Introduction to Vectors**
Used here as: Introduction
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77363/1.1/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Introduction <**http://legacy.cnx.org/content/m58276/1.8**> by OpenStax.

Module: **Cosmic Universe Scalars and Vectors**
Used here as: Scalars and Vectors
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77364/1.1/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**

Based on: Scalars and Vectors <**http://legacy.cnx.org/content/m58277/1.14**> by OpenStax.

Module: **Cosmic Universe Coordinate Systems and Components of a Vector**
Used here as: Coordinate Systems and Components of a Vector
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77365/1.1/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Coordinate Systems and Components of a Vector <**http://legacy.cnx.org/content/m58278/1.16**> by OpenStax.

Module: **Cosmic Universe Algebra of Vectors**
Used here as: Algebra of Vectors
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m64049/1.3/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Algebra of Vectors <**http://legacy.cnx.org/content/m58279/1.6**> by OpenStax.

Module: **Cosmic Universe Introduction to 2D&3D Kinematics**
Used here as: Introduction
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m64054/1.1/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Introduction <**http://legacy.cnx.org/content/m58288/1.6**> by OpenStax.

Module: **Cosmic Universe Displacement and Velocity Vectors**
Used here as: Displacement and Velocity Vectors
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m64162/1.2/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Displacement and Velocity Vectors <**http://legacy.cnx.org/content/m58289/1.4**> by OpenStax.

Module: **Cosmic Universe Acceleration Vector**
Used here as: Acceleration Vector
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m64064/1.4/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Acceleration Vector <**http://legacy.cnx.org/content/m58290/1.5**> by OpenStax.

Module: **Cosmic Universe Projectile Motion**
Used here as: Projectile Motion
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m64053/1.6/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Projectile Motion <**http://legacy.cnx.org/content/m58291/1.4**> by OpenStax.

Module: **Cosmic Universe Uniform Circular Motion**
Used here as: Uniform Circular Motion
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m64051/1.3/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Uniform Circular Motion <**http://legacy.cnx.org/content/m58292/1.4**> by OpenStax.

Module: **Cosmic Universe Relative Motion in One and Two Dimensions**
Used here as: Relative Motion in One and Two Dimensions
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m64050/1.6/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Relative Motion in One and Two Dimensions <**http://legacy.cnx.org/content/m58293/1.4**> by OpenStax.

Module: **Cosmic Universe Introduction to the Solar System**
Used here as: Introduction
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m64057/1.3/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Thinking Ahead <**http://legacy.cnx.org/content/m59814/1.3**> by OpenStax Astronomy.

Module: **Cosmic Universe Overview of Our Planetary System**
Used here as: Overview of Our Planetary System
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m64059/1.10/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Overview of Our Planetary System <**http://legacy.cnx.org/content/m59816/1.3**> by OpenStax Astronomy.

Module: **Cosmic Universe Composition and Structure of Planets**
Used here as: Composition and Structure of Planets
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m64060/1.5/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Composition and Structure of Planets <**http://legacy.cnx.org/content/m59819/1.3**> by OpenStax Astronomy.

Module: **Cosmic Universe Origin of the Solar System**
Used here as: Origin of the Solar System
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m64056/1.8/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Origin of the Solar System <**http://legacy.cnx.org/content/m59823/1.4**> by OpenStax Astronomy.

Module: **Cosmic Universe Kepler's Laws of Planetary Motion**
Used here as: Kepler's Laws of Planetary Motion
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m64058/1.14/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Kepler's Laws of Planetary Motion <**http://legacy.cnx.org/content/m58350/1.3**> by OpenStax.

Module: **Cosmic Universe Introduction to Newton's Laws**
Used here as: Introduction to Newton's Laws
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m64163/1.3/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Introduction <**http://legacy.cnx.org/content/m58294/1.4**> by OpenStax.

Module: **Cosmic Universe Forces**

Used here as: Forces
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m64169/1.5/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Forces <**http://legacy.cnx.org/content/m58295/1.4**> by OpenStax.

Module: **Cosmic Universe Newton's First Law**
Used here as: Newton's First Law
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m64170/1.3/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Newton's First Law <**http://legacy.cnx.org/content/m58296/1.4**> by OpenStax.

Module: **Cosmic Universe Newton's Second Law**
Used here as: Newton's Second Law
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m64171/1.7/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Newton's Second Law <**http://legacy.cnx.org/content/m58297/1.4**> by OpenStax.

Module: **Cosmic Universe Mass and Weight**
Used here as: Mass and Weight
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m64172/1.1/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Mass and Weight <**http://legacy.cnx.org/content/m58298/1.4**> by OpenStax.

Module: **Cosmic Universe Newton’s Third Law**
Used here as: Newton’s Third Law
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m64173/1.1/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Newton’s Third Law <**http://legacy.cnx.org/content/m58299/1.4**> by OpenStax.

Module: **Cosmic Universe Common Forces**
Used here as: Common Forces
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m64174/1.4/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Common Forces <**http://legacy.cnx.org/content/m58300/1.4**> by OpenStax.

Module: **Cosmic Universe Drawing Free-Body Diagrams**
Used here as: Drawing Free-Body Diagrams
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m64175/1.3/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Drawing Free-Body Diagrams <**http://legacy.cnx.org/content/m58301/1.4**> by OpenStax.

Module: **Cosmic Universe Friction**
Used here as: Friction
By: Steven Mellema

URL: **https://legacy.cnx.org/content/m64176/1.3/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Friction <**http://legacy.cnx.org/content/m58304/1.4**> by OpenStax.

Module: **Cosmic Universe Centripetal Force**
Used here as: Centripetal Force
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m64177/1.4/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Centripetal Force <**http://legacy.cnx.org/content/m58305/1.8**> by OpenStax.

Module: **Cosmic Universe Newton's Law of Universal Gravitation**
Used here as: Newton's Law of Universal Gravitation
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m64178/1.3/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Newton's Law of Universal Gravitation <**http://legacy.cnx.org/content/m58345/1.3**> by OpenStax.

Module: **Cosmic Universe The Newtonian Synthesis**
Used here as: The Newtonian Synthesis
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m64179/1.4/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Gravitation Near Earth's Surface <**http://legacy.cnx.org/content/m58346/1.4**> by OpenStax.

Module: **Cosmic Universe Introduction to Rotational Dynamics**
Used here as: Introduction
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m71578/1.1/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Introduction <**http://legacy.cnx.org/content/m58325/1.7**> by OpenStax.

Module: **Cosmic Universe Moment of Inertia**
Used here as: Moment of Inertia
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m71416/1.6/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Moment of Inertia and Rotational Kinetic Energy <**http://legacy.cnx.org/content/m58329/1.7**> by OpenStax.

Module: **Cosmic Universe Torque**
Used here as: Torque
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m71417/1.2/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Torque <**http://legacy.cnx.org/content/m58331/1.8**> by OpenStax.

Module: **Cosmic Universe Newton’s Second Law for Rotation**
Used here as: Newton's Second Law for Rotations
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m71422/1.3/**

Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Newton’s Second Law for Rotation <**http://legacy.cnx.org/content/m58332/1.8**> by OpenStax.

Module: **Cosmic Universe Introduction to Work and Energy**
Used here as: Introduction
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m71577/1.3/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Introduction <**http://legacy.cnx.org/content/m58307/1.8**> by OpenStax.

Module: **Cosmic Universe Work**
Used here as: Work
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m71418/1.3/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Work <**http://legacy.cnx.org/content/m58719/1.7**> by OpenStax University Physics.

Module: **Cosmic Universe Kinetic Energy**
Used here as: Kinetic Energy
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m71419/1.1/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Kinetic Energy <**http://legacy.cnx.org/content/m58308/1.7**> by OpenStax.

Module: **Cosmic Universe Work-Energy Theorem**
Used here as: Work-Energy Theorem
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m71570/1.3/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Work-Energy Theorem <**http://legacy.cnx.org/content/m58309/1.12**> by OpenStax.

Module: **Cosmic Universe Power**
Used here as: Power
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m71574/1.3/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Power <**http://legacy.cnx.org/content/m58310/1.7**> by OpenStax.

Module: **Cosmic Universe Potential Energy**
Used here as: Potential Energy
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m71571/1.2/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Potential Energy of a System <**http://legacy.cnx.org/content/m58312/1.15**> by OpenStax.

Module: **Cosmic Universe Conservation of Energy**
Used here as: Conservation of Energy
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m71572/1.7/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**

Based on: Conservation of Energy <**http://legacy.cnx.org/content/m58314/1.11**> by OpenStax.

Module: **Cosmic Universe Sources of Energy**
Used here as: Sources of Energy
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m71573/1.4/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Sources of Energy <**http://legacy.cnx.org/content/m58661/1.17**> by OpenStax.

Module: **Cosmic Universe Introduction to Momentum**
Used here as: Introduction
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m71576/1.1/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Introduction <**http://legacy.cnx.org/content/m58317/1.7**> by OpenStax.

Module: **Cosmic Universe Linear Momentum**
Used here as: Linear Momentum
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m71608/1.2/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Linear Momentum <**http://legacy.cnx.org/content/m58318/1.9**> by OpenStax.

Module: **Cosmic Universe Impulse and Collisions**
Used here as: Impulse and Collisions
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m71609/1.1/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Impulse and Collisions <**http://legacy.cnx.org/content/m58319/1.8**> by OpenStax.

Module: **Cosmic Universe Conservation of Linear Momentum**
Used here as: Conservation of Linear Momentum
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m71610/1.2/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Conservation of Linear Momentum <**http://legacy.cnx.org/content/m58320/1.9**> by OpenStax.

Module: **Cosmic Universe Types of Collisions**
Used here as: Types of Collisions
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m71611/1.1/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Types of Collisions <**http://legacy.cnx.org/content/m58321/1.10**> by OpenStax.

Module: **Cosmic Universe Center of Mass**
Used here as: Center of Mass
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m71612/1.3/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Center of Mass <**http://legacy.cnx.org/content/m58323/1.11**> by OpenStax.

Module: **Cosmic Universe Introduction to Angular Momentum**

Used here as: Introduction
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m71579/1.1/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Introduction <**http://legacy.cnx.org/content/m58334/1.6**> by OpenStax.

Module: **Cosmic Universe Angular Momentum**
Used here as: Angular Momentum
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m71620/1.6/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Angular Momentum <**http://legacy.cnx.org/content/m58336/1.13**> by OpenStax.

Module: **Cosmic Universe Conservation of Angular Momentum**
Used here as: Conservation of Angular Momentum
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m71619/1.4/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Conservation of Angular Momentum <**http://legacy.cnx.org/content/m58337/1.6**> by OpenStax.

Module: **Cosmic Universe Introduction to Waves**
Used here as: Introduction
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m71599/1.2/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Waves <**http://legacy.cnx.org/content/m42248/1.10**> by OpenStax.

Module: **Cosmic Universe The Wave Behavior of Light**
Used here as: The Wave Behavior of Light
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m76900/1.4/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: The Behavior of Light <**http://legacy.cnx.org/content/m59792/1.7**> by OpenStax Astronomy.

Module: **Cosmic Universe Huygens’s Principle**
Used here as: Huygens' Principle
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m76902/1.2/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Huygens’s Principle <**http://legacy.cnx.org/content/m58523/1.6**> by OpenStax.

Module: **Cosmic Universe The Electromagnetic Spectrum**
Used here as: The Electromagnetic Spectrum
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m76901/1.7/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: The Electromagnetic Spectrum <**http://legacy.cnx.org/content/m59798/1.13**> by OpenStax Astronomy.

Module: **Cosmic Universe The Brightness of Stars**
Used here as: The Brightness of Stars
By: Steven Mellema

URL: **https://legacy.cnx.org/content/m76888/1.10/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: The Brightness of Stars <**http://legacy.cnx.org/content/m59890/1.7**> by OpenStax Astronomy.

Module: **Cosmic Universe Introduction to Spectroscopy**
Used here as: Introduction
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m71596/1.6/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Thinking Ahead <**http://legacy.cnx.org/content/m59889/1.6**> by OpenStax Astronomy.

Module: **Cosmic Universe Spectroscopy in Astronomy**
Used here as: Spectroscopy
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m76920/1.4/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Spectroscopy in Astronomy <**http://legacy.cnx.org/content/m59793/1.7**> by OpenStax Astronomy.

Module: **Cosmic Universe The Doppler Effect**
Used here as: The Doppler Effect
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m76921/1.7/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: The Doppler Effect <**http://legacy.cnx.org/content/m59799/1.12**> by OpenStax Astronomy.

Module: **Cosmic Universe Introduction to Quantum Optics**
Used here as: Introduction
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m71601/1.1/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Introduction <**http://legacy.cnx.org/content/m58757/1.5**> by OpenStax.

Module: **Cosmic Universe Blackbody Radiation**
Used here as: Blackbody Radiation
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m76924/1.2/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Blackbody Radiation <**http://legacy.cnx.org/content/m58761/1.6**> by OpenStax.

Module: **Cosmic Universe Bohr’s Model of the Hydrogen Atom**
Used here as: Bohr’s Model of the Hydrogen Atom
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m76925/1.3/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Bohr’s Model of the Hydrogen Atom <**http://legacy.cnx.org/content/m58764/1.9**> by OpenStax.

Module: **Cosmic Universe Atomic Spectra and X-rays**
Used here as: Atomic Spectra and X-rays
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m76927/1.2/**
Copyright: Steven Mellema

License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Atomic Spectra and X-rays <**http://legacy.cnx.org/content/m58590/1.6**> by OpenStax.

Module: **Cosmic Universe Molecular Spectra**
Used here as: Molecular Spectra
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m76926/1.2/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Molecular Spectra <**http://legacy.cnx.org/content/m58593/1.5**> by OpenStax.

Module: **Cosmic Universe Introduction to Thermodynamics**
Used here as: Introduction
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m71603/1.1/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Introduction <**http://legacy.cnx.org/content/m58383/1.6**> by OpenStax.

Module: **Cosmic Universe Temperature and Thermal Equilibrium**
Used here as: Temperature and Thermal Equilibrium
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m76929/1.1/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Temperature and Thermal Equilibrium <**http://legacy.cnx.org/content/m58384/1.5**> by OpenStax.

Module: **Cosmic Universe Thermometers and Temperature Scales**
Used here as: Thermometers and Temperature Scales
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m76930/1.1/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Thermometers and Temperature Scales <**http://legacy.cnx.org/content/m58385/1.8**> by OpenStax.

Module: **Cosmic Universe Thermal Expansion**
Used here as: Thermal Expansion
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m76931/1.1/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Thermal Expansion <**http://legacy.cnx.org/content/m58386/1.9**> by OpenStax.

Module: **Cosmic Universe Heat Transfer and Specific Heat**
Used here as: Heat Transfer and Specific Heat
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m76932/1.3/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Heat Transfer, Specific Heat, and Calorimetry <**http://legacy.cnx.org/content/m58387/1.6**> by OpenStax.

Module: **Cosmic Universe Phase Changes**
Used here as: Phase Changes
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m76933/1.3/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Phase Changes <**http://legacy.cnx.org/content/m58388/1.7**> by OpenStax.

Module: **Cosmic Universe Mechanisms of Heat Transfer**
Used here as: Mechanisms of Heat Transfer
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m76934/1.2/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Mechanisms of Heat Transfer <**http://legacy.cnx.org/content/m58389/1.13**> by OpenStax.

Module: **Cosmic Universe Introduction to Kinetic Theory**
Used here as: Introduction
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m76935/1.1/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Introduction <**http://legacy.cnx.org/content/m58390/1.5**> by OpenStax.

Module: **Cosmic Universe Fluids, Density, and Pressure**
Used here as: Fluids, Density, and Pressure
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m76936/1.1/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Fluids, Density, and Pressure <**http://legacy.cnx.org/content/m58353/1.7**> by OpenStax.

Module: **Cosmic Universe Molecular Model of an Ideal Gas**
Used here as: Molecular Model of an Ideal Gas
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m76940/1.2/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Molecular Model of an Ideal Gas <**http://legacy.cnx.org/content/m58391/1.6**> by OpenStax.

Module: **Cosmic Universe Pressure, Temperature, and RMS Speed**
Used here as: Pressure, Temperature, and RMS Speed
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m76941/1.2/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Pressure, Temperature, and RMS Speed <**http://legacy.cnx.org/content/m58392/1.5**> by OpenStax.

Module: **Cosmic Universe Distribution of Molecular Speeds**
Used here as: Distribution of Molecular Speeds
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m76942/1.2/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Distribution of Molecular Speeds <**http://legacy.cnx.org/content/m58394/1.8**> by OpenStax.

Module: **Cosmic Universe Introduction to Energy and the Solar System**
Used here as: Introduction
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m71604/1.2/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Introduction <**http://legacy.cnx.org/content/m58606/1.6**> by OpenStax.

Module: **Cosmic Universe Formation of the Solar System**
Used here as: Formation of the Solar System

By: Steven Mellema
URL: **https://legacy.cnx.org/content/m76943/1.3/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Formation of the Solar System <**http://legacy.cnx.org/content/m59873/1.8**> by OpenStax Astronomy.

Module: **Cosmic Universe Properties of Nuclei**
Used here as: Properties of Nuclei
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77048/1.1/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Properties of Nuclei <**http://legacy.cnx.org/content/m58607/1.6**> by OpenStax.

Module: **Cosmic Universe Nuclear Binding Energy**
Used here as: Nuclear Binding Energy
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77049/1.4/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Nuclear Binding Energy <**http://legacy.cnx.org/content/m58608/1.6**> by OpenStax.

Module: **Cosmic Universe Radioactive Decay**
Used here as: Radioactive Decay
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77050/1.5/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Radioactive Decay <**http://legacy.cnx.org/content/m58609/1.6**> by OpenStax.

Module: **Cosmic Universe Introduction to Comparative Planetology**
Used here as: Introduction
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77149/1.5/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Thinking Ahead <**http://legacy.cnx.org/content/m59836/1.7**> by OpenStax Astronomy.

Module: **Cosmic Universe Composition and Structure of Planets**
Used here as: Composition and Structure of Planets
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77051/1.6/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Composition and Structure of Planets <**http://legacy.cnx.org/content/m59819/1.7**> by OpenStax Astronomy.

Module: **Cosmic Universe Dating Planetary Surfaces**
Used here as: Dating Planetary Surfaces
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77052/1.3/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Dating Planetary Surfaces <**http://legacy.cnx.org/content/m59820/1.7**> by OpenStax Astronomy.

Module: **Cosmic Universe Earth's Atmosphere**
Used here as: Earth's Atmosphere
By: Steven Mellema

URL: **https://legacy.cnx.org/content/m77053/1.4/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Earth’s Atmosphere <**http://legacy.cnx.org/content/m59827/1.7**> by OpenStax Astronomy.

Module: **Cosmic Universe The Massive Atmosphere of Venus**
Used here as: The Massive Atmosphere of Venus
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77054/1.2/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: The Massive Atmosphere of Venus <**http://legacy.cnx.org/content/m59840/1.7**> by OpenStax Astronomy.

Module: **Cosmic Universe Water and Life on Mars**
Used here as: Water and Life on Mars
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77055/1.3/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Water and Life on Mars <**http://legacy.cnx.org/content/m59856/1.12**> by OpenStax Astronomy.

Module: **Cosmic Universe Divergent Planetary Evolution**
Used here as: Divergent Planetary Evolution
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77056/1.11/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Divergent Planetary Evolution <**http://legacy.cnx.org/content/m59847/1.9**> by OpenStax Astronomy.

Module: **Cosmic Universe Introduction to Exoplanets**
Used here as: Introduction
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77137/1.7/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Comparison with Other Planetary Systems <**http://legacy.cnx.org/content/m59874/1.8**> by OpenStax Astronomy.

Module: **Cosmic Universe Planets beyond the Solar System: Search and Discovery**
Used here as: Planets Beyond the Solar System: Search and Discovery
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77138/1.8/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Planets beyond the Solar System: Search and Discovery <**http://legacy.cnx.org/content/m59921/1.8**> by OpenStax Astronomy.

Module: **Cosmic Universe Exoplanets Everywhere: What We Are Learning**
Used here as: Exoplanets Everywhere: What We Are Learning
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77139/1.3/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Exoplanets Everywhere: What We Are Learning <**http://legacy.cnx.org/content/m62183/1.8**> by OpenStax Astronomy.

Module: **Cosmic Universe New Perspectives on Planet Formation**

Used here as: New Perspectives on Planet Formation
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77140/1.4/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: New Perspectives on Planet Formation <**http://legacy.cnx.org/content/m62187/1.9**> by OpenStax Astronomy.

Module: **Cosmic Universe Introduction to The Sun**
Used here as: Introduction
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77141/1.3/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Thinking Ahead <**http://legacy.cnx.org/content/m59880/1.7**> by OpenStax Astronomy.

Module: **Cosmic Universe The Structure and Composition of the Sun**
Used here as: The Structure and Composition of the Sun
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77142/1.4/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: The Structure and Composition of the Sun <**http://legacy.cnx.org/content/m59877/1.9**> by OpenStax Astronomy.

Module: **Cosmic Universe Sources of Sunshine: Thermal and Gravitational Energy**
Used here as: Sources of Sunshine: Thermal and Gravitational Energy?
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77144/1.2/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Sources of Sunshine: Thermal and Gravitational Energy <**http://legacy.cnx.org/content/m59895/1.7**> by OpenStax Astronomy.

Module: **Cosmic Universe Source of Sunshine: Nuclear Fusion**
Used here as: Source of Sunshine: Nuclear Fusion!
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77145/1.5/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Nuclear Fusion <**http://legacy.cnx.org/content/m58613/1.8**> by OpenStax.

Module: **Cosmic Universe The Solar Interior: Theory**
Used here as: The Solar Interior: Theory
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77147/1.4/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: The Solar Interior: Theory <**http://legacy.cnx.org/content/m59887/1.8**> by OpenStax Astronomy.

Module: **Cosmic Universe The Solar Interior: Observations**
Used here as: The Solar Interior: Observations
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77148/1.4/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: The Solar Interior: Observations <**http://legacy.cnx.org/content/m59888/1.10**> by OpenStax Astronomy.

Module: **Cosmic Universe Introduction to Stars**
Used here as: Introduction
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m71605/1.4/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Cosmic Universe Introduction to Energy and the Solar System <**http://legacy.cnx.org/content/m71604/1.1**> by Steven Mellema.

Module: **Cosmic Universe Colors of Stars**
Used here as: Colors of Stars
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77152/1.4/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Colors of Stars <**http://legacy.cnx.org/content/m59891/1.7**> by OpenStax Astronomy.

Module: **Cosmic Universe The Spectra of Stars (and Brown Dwarfs)**
Used here as: The Spectra of Stars
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77153/1.4/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: The Spectra of Stars (and Brown Dwarfs) <**http://legacy.cnx.org/content/m59892/1.7**> by OpenStax Astronomy.

Module: **Cosmic Universe Using Spectra to Measure Stellar Radius, Composition, and Motion**
Used here as: Using Spectra to Measure Stellar Radius, Composition, and Motion
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77161/1.6/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Using Spectra to Measure Stellar Radius, Composition, and Motion <**http://legacy.cnx.org/content/m59896/1.11**> by OpenStax Astronomy.

Module: **Cosmic Universe A Stellar Census**
Used here as: A Stellar Census
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77157/1.2/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: A Stellar Census <**http://legacy.cnx.org/content/m59898/1.8**> by OpenStax Astronomy.

Module: **Cosmic Universe Measuring Stellar Masses**
Used here as: Measuring Stellar Masses
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77162/1.6/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Measuring Stellar Masses <**http://legacy.cnx.org/content/m59899/1.7**> by OpenStax Astronomy.

Module: **Cosmic Universe Diameters of Stars**
Used here as: Diameters of Stars
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77299/1.4/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Diameters of Stars <**http://legacy.cnx.org/content/m59900/1.8**> by OpenStax Astronomy.

Module: **Cosmic Universe The H-R Diagram**
Used here as: The H-R Diagram
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77300/1.5/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: The H–R Diagram <**http://legacy.cnx.org/content/m59901/1.13**> by OpenStax Astronomy.

Module: **Cosmic Universe Introduction to Celestial Distances**
Used here as: Introduction
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77301/1.3/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Thinking Ahead <**http://legacy.cnx.org/content/m59902/1.7**> by OpenStax Astronomy.

Module: **Cosmic Universe Fundamental Units of Distance**
Used here as: Fundamental Units of Distance
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77302/1.4/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Fundamental Units of Distance <**http://legacy.cnx.org/content/m59904/1.8**> by OpenStax Astronomy.

Module: **Cosmic Universe Surveying the Stars**
Used here as: Surveying the Stars
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77303/1.5/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Surveying the Stars <**http://legacy.cnx.org/content/m59905/1.9**> by OpenStax Astronomy.

Module: **Cosmic Universe Variable Stars: One Key to Cosmic Distances**
Used here as: Variable Stars: One Key to Cosmic Distances
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77304/1.4/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Variable Stars: One Key to Cosmic Distances <**http://legacy.cnx.org/content/m59906/1.8**> by OpenStax Astronomy.

Module: **Cosmic Universe The H–R Diagram and Cosmic Distances**
Used here as: The H–R Diagram and Cosmic Distances
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77305/1.7/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: The H–R Diagram and Cosmic Distances <**http://legacy.cnx.org/content/m59907/1.12**> by OpenStax Astronomy.

Module: **Cosmic Universe Introduction to Stellar Evolution**
Used here as: Introduction
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77151/1.3/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Thinking Ahead <**http://legacy.cnx.org/content/m59924/1.7**> by OpenStax Astronomy.

Module: **Cosmic Universe Star Formation**
Used here as: Star Formation
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77307/1.3/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Star Formation <**http://legacy.cnx.org/content/m59923/1.8**> by OpenStax Astronomy.

Module: **Cosmic Universe The H–R Diagram and the Study of Stellar Evolution**
Used here as: The H–R Diagram and the Study of Stellar Evolution
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77308/1.5/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: The H–R Diagram and the Study of Stellar Evolution <**http://legacy.cnx.org/content/m59918/1.8**> by OpenStax Astronomy.

Module: **Cosmic Universe Evolution from the Main Sequence to Red Giants**
Used here as: Evolution from the Main Sequence to Red Giants
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77309/1.3/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Evolution from the Main Sequence to Red Giants <**http://legacy.cnx.org/content/m59925/1.9**> by OpenStax Astronomy.

Module: **Cosmic Universe Star Clusters**
Used here as: Star Clusters
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77310/1.1/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Star Clusters <**http://legacy.cnx.org/content/m59926/1.7**> by OpenStax Astronomy.

Module: **Cosmic Universe Checking Out the Theory**
Used here as: Checking Out the Theory
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77311/1.7/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Checking Out the Theory <**http://legacy.cnx.org/content/m59927/1.8**> by OpenStax Astronomy.

Module: **Cosmic Universe Further Evolution of Stars**
Used here as: Further Evolution of Stars
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77312/1.3/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Further Evolution of Stars <**http://legacy.cnx.org/content/m59928/1.8**> by OpenStax Astronomy.

Module: **Cosmic Universe The Evolution of More Massive Stars**
Used here as: The Evolution of More Massive Stars
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77313/1.3/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: The Evolution of More Massive Stars <**http://legacy.cnx.org/content/m59930/1.8**> by OpenStax Astronomy.

Module: **Cosmic Universe Introduction to The Deaths of Stars**
Used here as: Introduction
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77314/1.1/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Thinking Ahead <**http://legacy.cnx.org/content/m59931/1.7**> by OpenStax Astronomy.

Module: **Cosmic Universe The Death of Low-Mass Stars**
Used here as: The Death of Low-Mass Stars
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77315/1.4/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: The Death of Low-Mass Stars <**http://legacy.cnx.org/content/m59932/1.7**> by OpenStax Astronomy.

Module: **Cosmic Universe Evolution of Massive Stars: An Explosive Finish**
Used here as: Evolution of Massive Stars: An Explosive Finish
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77316/1.3/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Evolution of Massive Stars: An Explosive Finish <**http://legacy.cnx.org/content/m59933/1.8**> by OpenStax Astronomy.

Module: **Cosmic Universe Supernova Observations**
Used here as: Supernova Observations
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77318/1.3/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Supernova Observations <**http://legacy.cnx.org/content/m59934/1.8**> by OpenStax Astronomy.

Module: **Cosmic Universe Pulsars and the Discovery of Neutron Stars**
Used here as: Pulsars and the Discovery of Neutron Stars
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77319/1.3/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Pulsars and the Discovery of Neutron Stars <**http://legacy.cnx.org/content/m59935/1.7**> by OpenStax Astronomy.

Module: **Cosmic Universe The Evolution of Binary Star Systems**
Used here as: The Evolution of Binary Star Systems
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77320/1.3/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: The Evolution of Binary Star Systems <**http://legacy.cnx.org/content/m59936/1.8**> by OpenStax Astronomy.

Module: **Cosmic Universe Introducing General Relativity**
Used here as: Introducing General Relativity
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77324/1.3/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Introducing General Relativity <**http://legacy.cnx.org/content/m59939/1.7**> by OpenStax Astronomy.

Module: **Cosmic Universe Black Holes**
Used here as: Black Holes
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77321/1.4/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Black Holes <**http://legacy.cnx.org/content/m59943/1.8**> by OpenStax Astronomy.

Module: **Cosmic Universe Evidence for Black Holes**
Used here as: Evidence for Black Holes
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77322/1.2/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Evidence for Black Holes <**http://legacy.cnx.org/content/m59944/1.7**> by OpenStax Astronomy.

Module: **Cosmic Universe Gravitational Wave Astronomy**
Used here as: Gravitational Wave Astronomy
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77323/1.3/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Gravitational Wave Astronomy <**http://legacy.cnx.org/content/m59945/1.11**> by OpenStax Astronomy.

Module: **Cosmic Universe Introduction to The Milky Way Galaxy**
Used here as: Introduction
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77325/1.2/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Thinking Ahead <**http://legacy.cnx.org/content/m59947/1.7**> by OpenStax Astronomy.

Module: **Cosmic Universe The Architecture of the Galaxy**
Used here as: The Architecture of the Galaxy
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77328/1.3/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: The Architecture of the Galaxy <**http://legacy.cnx.org/content/m59948/1.10**> by OpenStax Astronomy.

Module: **Cosmic Universe Spiral Structure**
Used here as: Spiral Structure
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77330/1.2/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Spiral Structure <**http://legacy.cnx.org/content/m59954/1.7**> by OpenStax Astronomy.

Module: **Cosmic Universe The Mass of the Galaxy**
Used here as: The Mass of the Galaxy
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77331/1.7/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: The Mass of the Galaxy <**http://legacy.cnx.org/content/m59950/1.8**> by OpenStax Astronomy.

Module: **Cosmic Universe The Center of the Galaxy**
Used here as: The Center of the Galaxy

By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77332/1.3/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: The Center of the Galaxy <**http://legacy.cnx.org/content/m59951/1.7**> by OpenStax Astronomy.

Module: **Cosmic Universe Stellar Populations in the Galaxy**
Used here as: Stellar Populations in the Galaxy
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77333/1.2/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Stellar Populations in the Galaxy <**http://legacy.cnx.org/content/m59952/1.7**> by OpenStax Astronomy.

Module: **Cosmic Universe The Formation of the Galaxy**
Used here as: The Formation of the Galaxy
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77334/1.4/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: The Formation of the Galaxy <**http://legacy.cnx.org/content/m59955/1.11**> by OpenStax Astronomy.

Module: **Cosmic Universe Introduction to Galaxies**
Used here as: Introduction
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m71606/1.3/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Cosmic Universe Introduction to Stars <**http://legacy.cnx.org/content/m71605/1.1**> by Steven Mellema.

Module: **Cosmic Universe The Discovery of Galaxies**
Used here as: The Discovery of Galaxies
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77335/1.3/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: The Discovery of Galaxies <**http://legacy.cnx.org/content/m59964/1.9**> by OpenStax Astronomy.

Module: **Cosmic Universe Types of Galaxies**
Used here as: Types of Galaxies
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77336/1.3/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Types of Galaxies <**http://legacy.cnx.org/content/m59958/1.7**> by OpenStax Astronomy.

Module: **Cosmic Universe Properties of Galaxies**
Used here as: Properties of Galaxies
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77337/1.2/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Properties of Galaxies <**http://legacy.cnx.org/content/m59959/1.7**> by OpenStax Astronomy.

Module: **Cosmic Universe The Extragalactic Distance Scale**
Used here as: The Extragalactic Distance Scale
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77338/1.5/**

Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: The Extragalactic Distance Scale <**http://legacy.cnx.org/content/m59960/1.7**> by OpenStax Astronomy.

Module: **Cosmic Universe The Expanding Universe**
Used here as: The Expanding Universe
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77339/1.4/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: The Expanding Universe <**http://legacy.cnx.org/content/m62195/1.9**> by OpenStax Astronomy.

Module: **Cosmic Universe Introduction to The Evolution and Distribution of Galaxies**
Used here as: Introduction
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77326/1.2/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Thinking Ahead <**http://legacy.cnx.org/content/m59961/1.7**> by OpenStax Astronomy.

Module: **Cosmic Universe Quasars**
Used here as: Quasars
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77341/1.4/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Quasars <**http://legacy.cnx.org/content/m59962/1.7**> by OpenStax Astronomy.

Module: **Cosmic Universe Supermassive Black Holes: What Quasars Really Are**
Used here as: Supermassive Black Holes
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77342/1.3/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Supermassive Black Holes: What Quasars Really Are <**http://legacy.cnx.org/content/m59965/1.7**> by OpenStax Astronomy.

Module: **Cosmic Universe Quasars as Probes of Evolution in the Universe**
Used here as: Quasars as Probes of Evolution in the Universe
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77343/1.3/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Quasars as Probes of Evolution in the Universe <**http://legacy.cnx.org/content/m59966/1.11**> by OpenStax Astronomy.

Module: **Cosmic Universe Observations of Distant Galaxies**
Used here as: Observations of Distant Galaxies
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77344/1.2/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Observations of Distant Galaxies <**http://legacy.cnx.org/content/m59970/1.11**> by OpenStax Astronomy.

Module: **Cosmic Universe Galaxy Mergers and Active Galactic Nuclei**
Used here as: Galaxy Mergers and Active Galactic Nuclei
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77345/1.3/**

Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Galaxy Mergers and Active Galactic Nuclei <**http://legacy.cnx.org/content/m59971/1.8**> by OpenStax Astronomy.

Module: **Cosmic Universe The Distribution of Galaxies in Space**
Used here as: The Distribution of Galaxies in Space
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77346/1.4/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: The Distribution of Galaxies in Space <**http://legacy.cnx.org/content/m59972/1.10**> by OpenStax Astronomy.

Module: **Cosmic Universe The Challenge of Dark Matter**
Used here as: The Challenge of Dark Matter
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77347/1.2/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: The Challenge of Dark Matter <**http://legacy.cnx.org/content/m59973/1.9**> by OpenStax Astronomy.

Module: **Cosmic Universe The Formation and Evolution of Galaxies and Structure in the Universe**
Used here as: The Formation and Evolution of Galaxies and Structure in the Universe
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77348/1.3/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: The Formation and Evolution of Galaxies and Structure in the Universe <**http://legacy.cnx.org/content/m59974/1.10**> by OpenStax Astronomy.

Module: **Cosmic Universe Introduction to Cosmology**
Used here as: Introduction
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m71607/1.3/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Cosmic Universe Introduction to Galaxies <**http://legacy.cnx.org/content/m71606/1.1**> by Steven Mellema.

Module: **Cosmic Universe The Age of the Universe**
Used here as: The Age of the Universe
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77349/1.2/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: The Age of the Universe <**http://legacy.cnx.org/content/m59976/1.7**> by OpenStax Astronomy.

Module: **Cosmic Universe A Model of the Universe**
Used here as: A Model of the Universe
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77352/1.5/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: A Model of the Universe <**http://legacy.cnx.org/content/m59977/1.7**> by OpenStax Astronomy.

Module: **Cosmic Universe The Beginning of the Universe**
Used here as: The Beginning of the Universe

By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77353/1.2/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: The Beginning of the Universe <**http://legacy.cnx.org/content/m59978/1.9**> by OpenStax Astronomy.

Module: **Cosmic Universe The Cosmic Microwave Background**
Used here as: The Cosmic Microwave Background
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77354/1.2/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: The Cosmic Microwave Background <**http://legacy.cnx.org/content/m59979/1.8**> by OpenStax Astronomy.

Module: **Cosmic Universe What Is the Universe Really Made Of?**
Used here as: What Is the Universe Really Made Of?
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77357/1.1/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: What Is the Universe Really Made Of? <**http://legacy.cnx.org/content/m59980/1.7**> by OpenStax Astronomy.

Module: **Cosmic Universe The Inflationary Universe**
Used here as: The Inflationary Universe
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m77358/1.3/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: The Inflationary Universe <**http://legacy.cnx.org/content/m59981/1.7**> by OpenStax Astronomy.

Module: **Introduction to Geometric Optics**
Used here as: Introduction
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m71580/1.4/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Introduction <**http://legacy.cnx.org/content/m58504/1.6**> by OpenStax.

Module: **Cosmic Universe The Propagation of Light**
Used here as: The Propagation of Light
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m76882/1.4/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: The Propagation of Light <**http://legacy.cnx.org/content/m58505/1.7**> by OpenStax.

Module: **Cosmic Universe The Law of Reflection**
Used here as: The Law of Reflection
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m76883/1.2/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: The Law of Reflection <**http://legacy.cnx.org/content/m58509/1.5**> by OpenStax.

Module: **Cosmic Universe Refraction**
Used here as: Refraction

By: Steven Mellema
URL: **https://legacy.cnx.org/content/m76884/1.2/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Refraction <**http://legacy.cnx.org/content/m58510/1.5**> by OpenStax.

Module: **Cosmic Universe Total Internal Reflection**
Used here as: Total Internal Reflection
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m76885/1.3/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Total Internal Reflection <**http://legacy.cnx.org/content/m58511/1.6**> by OpenStax.

Module: **Cosmic Universe Dispersion**
Used here as: Dispersion
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m76886/1.1/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Dispersion <**http://legacy.cnx.org/content/m58512/1.7**> by OpenStax.

Module: **Cosmic Universe Introduction to Image Formation**
Used here as: Introduction
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m71598/1.1/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Introduction <**http://legacy.cnx.org/content/m58526/1.5**> by OpenStax.

Module: **Cosmic Universe Images Formed by Plane Mirrors**
Used here as: Images Formed by Plane Mirrors
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m76889/1.1/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Images Formed by Plane Mirrors <**http://legacy.cnx.org/content/m58527/1.6**> by OpenStax.

Module: **Cosmic Universe Spherical Mirrors**
Used here as: Spherical Mirrors
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m76890/1.1/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Spherical Mirrors <**http://legacy.cnx.org/content/m58528/1.7**> by OpenStax.

Module: **Cosmic Universe Images Formed by Refraction**
Used here as: Images Formed by Refraction
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m76891/1.1/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Images Formed by Refraction <**http://legacy.cnx.org/content/m58529/1.6**> by OpenStax.

Module: **Cosmic Universe Thin Lenses**
Used here as: Thin Lenses
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m76893/1.2/**

Based on: Multiple-Slit Interference <**http://legacy.cnx.org/content/m58539/1.5**> by OpenStax.

Module: **Cosmic Universe Interference in Thin Films**
Used here as: Interference in Thin Films
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m76911/1.1/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Interference in Thin Films <**http://legacy.cnx.org/content/m58540/1.7**> by OpenStax.

Module: **Cosmic Universe The Michelson Interferometer**
Used here as: The Michelson Interferometer
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m76912/1.2/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: The Michelson Interferometer <**http://legacy.cnx.org/content/m58541/1.7**> by OpenStax.

Module: **Cosmic Universe Introduction to Diffraction**
Used here as: Introduction
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m76913/1.1/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Introduction <**http://legacy.cnx.org/content/m58542/1.5**> by OpenStax.

Module: **Cosmic Universe Single-Slit Diffraction**
Used here as: Single-Slit Diffraction
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m76914/1.2/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Single-Slit Diffraction <**http://legacy.cnx.org/content/m58543/1.5**> by OpenStax.

Module: **Cosmic Universe Intensity in Single-Slit Diffraction**
Used here as: Intensity in Single-Slit Diffraction
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m76915/1.2/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Intensity in Single-Slit Diffraction <**http://legacy.cnx.org/content/m58544/1.10**> by OpenStax.

Module: **Cosmic Universe Double-Slit Diffraction**
Used here as: Double-Slit Diffraction
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m76916/1.1/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Double-Slit Diffraction <**http://legacy.cnx.org/content/m58623/1.8**> by OpenStax.

Module: **Cosmic Universe Diffraction Gratings**
Used here as: Diffraction Gratings
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m76917/1.1/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Diffraction Gratings <**http://legacy.cnx.org/content/m62592/1.3**> by OpenStax.

Module: **Cosmic Universe Circular Apertures and Resolution**

Used here as: Circular Apertures and Resolution
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m76918/1.2/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Circular Apertures and Resolution <**http://legacy.cnx.org/content/m58548/1.7**> by OpenStax.

Module: **Cosmic Universe Polarization**
Used here as: Polarization
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m76919/1.2/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Polarization <**http://legacy.cnx.org/content/m58525/1.9**> by OpenStax.

Module: **Cosmic Universe Astronomical Instruments Thinking Ahead**
Used here as: Introduction
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m85896/1.1/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Thinking Ahead <**http://legacy.cnx.org/content/m59802/1.7**> by OpenStax Astronomy.

Module: **New Cosmic Universe Telescopes**
Used here as: Telescopes
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m85897/1.5/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Telescopes <**http://legacy.cnx.org/content/m59803/1.8**> by OpenStax Astronomy.

Module: **New Cosmic Universe Telescopes Today**
Used here as: Telescopes Today
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m85898/1.4/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Telescopes Today <**http://legacy.cnx.org/content/m59805/1.7**> by OpenStax Astronomy.

Module: **New Cosmic Universe Visible-Light Detectors and Instruments**
Used here as: Visible-Light Detectors and Instruments
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m85899/1.3/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Visible-Light Detectors and Instruments <**http://legacy.cnx.org/content/m59821/1.7**> by OpenStax Astronomy.

Module: **New Cosmic Universe Radio Telescopes**
Used here as: Radio Telescopes
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m85900/1.4/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Radio Telescopes <**http://legacy.cnx.org/content/m59809/1.10**> by OpenStax Astronomy.

Module: **New Cosmic Universe Observations outside Earth’s Atmosphere**
Used here as: Observations Outside Earth’s Atmosphere

By: Steven Mellema
URL: **https://legacy.cnx.org/content/m85928/1.4/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Observations outside Earth's Atmosphere <**http://legacy.cnx.org/content/m59811/1.8**> by OpenStax Astronomy.

Module: **New Cosmic Universe The Future of Large Telescopes**
Used here as: The Future of Large Telescopes
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m85977/1.3/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: The Future of Large Telescopes <**http://legacy.cnx.org/content/m59812/1.11**> by OpenStax Astronomy.

Module: **Cosmic Universe Units**
Used here as: Units
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m64018/1.1/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Units <**http://legacy.cnx.org/content/m58777/1.5**> by OpenStax.

Module: **Cosmic Universe Conversion Factors**
Used here as: Unit Conversions
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m64015/1.1/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Conversion Factors <**http://legacy.cnx.org/content/m58615/1.5**> by OpenStax.

Module: **Cosmic Universe Fundamental Constants**
Used here as: Fundamental Physics Constants
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m64016/1.1/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Fundamental Constants <**http://legacy.cnx.org/content/m58616/1.4**> by OpenStax.

Module: **Cosmic Universe Astronomical Data**
Used here as: Astronomical Data
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m64014/1.5/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Astronomical Data <**http://legacy.cnx.org/content/m58617/1.5**> by OpenStax.

Module: **Cosmic Universe Mathematical Formulas**
Used here as: Useful Mathematical Formulas
By: Steven Mellema
URL: **https://legacy.cnx.org/content/m64017/1.1/**
Copyright: Steven Mellema
License: **http://creativecommons.org/licenses/by/4.0/**
Based on: Mathematical Formulas <**http://legacy.cnx.org/content/m58618/1.5**> by OpenStax.

Module: **Cosmic Universe Chemistry**
Used here as: The Chemical Elements
By: Steven Mellema

ABOUT CONNEXIONS

Since 1999, Connexions has been pioneering a global system where anyone can create course materials and make them fully accessible and easily reusable free of charge. We are a Web-based authoring, teaching and learning environment open to anyone interested in education, including students, teachers, professors and lifelong learners. We connect ideas and facilitate educational communities. Connexions's modular, interactive courses are in use worldwide by universities, community colleges, K-12 schools, distance learners, and lifelong learners. Connexions materials are in many languages, including English, Spanish, Chinese, Japanese, Italian, Vietnamese, French, Portuguese, and Thai.