Fundamentals of Heat, Light & Sound

Fundamentals of Heat, Light & Sound

Lumen Learning and OpenStax

Fundamentals of Heat, Light & Sound by NSCC & Lumen/OpenStax is licensed under a Creative Commons Attribution 4.0 International License, except where otherwise noted.

Contents

3. Temperature, Kinetic Theory, and the Gas Laws

4. Heat and Heat Transfer Methods

5. Thermodynamics

6. Oscillatory Motion and Waves

7. Physics of Hearing

8. Electromagnetic Waves

9. Geometric Optics

10. Vision and Optical Instruments

11. Wave Optics

Appendix

How to Succeed in Physics Guide

Lumen Learning

Fundamentals of Heat, Light & Sound is an adapted and remixed version of Lumen learning Physics I and Lumen Learning Physics II. The Lumen textbooks are adapted from OpenStax College Physics.

A versioning history is located at the end of the book.

Download the “How to Succeed in Physics” guide from Veritas Learning to review many of the major math and physics concepts and techniques that will help you excel in this Physics course.

You may also download the Student Solution Manual and other helpful tools from the OpenStax website.

Instructor Resources

Lumen Learning

Visit the OpenStax College Supplemental Resources section to find more resources, including a Getting Started Guide, Instructor Solution Manual, and PowerPoint Slides.

1. The Nature of Science and Physics

Introduction to Science and the Realm of Physics, Physical Quantities, and Units

Figure 1. Galaxies are as immense as atoms are small. Yet the same laws of physics describe both, and all the rest of nature—an indication of the underlying unity in the universe. The laws of physics are surprisingly few in number, implying an underlying simplicity to nature's apparent complexity. (credit: NASA, JPL-Caltech, P. Barmby, Harvard-Smithsonian Center for Astrophysics)

What is your first reaction when you hear the word "physics"? Did you imagine working through difficult equations or memorizing formulas that seem to have no real use in life outside the physics classroom? Many people come to the subject of physics with a bit of fear. But as you begin your exploration of 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.

For example, take a look at the image above. This image is of the Andromeda Galaxy, which contains billions of individual stars, huge clouds of gas, and dust. Two smaller galaxies are also visible as bright blue spots in the background. At a staggering 2.5 million light years from the Earth, this galaxy is the nearest one to our own galaxy (which is called the Milky Way). The stars and planets that make up Andromeda might seem to be the furthest thing from most people's regular, everyday lives. But

Andromeda is a great starting point to think about the forces that hold together the universe. The forces that cause Andromeda to act as it does are the same forces we contend with here on Earth, whether we are planning to send a rocket into space or simply raise the walls for a new home. The same gravity that causes the stars of Andromeda to rotate and revolve also causes water to flow over hydroelectric dams here on Earth. Tonight, take a moment to look up at the stars. 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 of the technological devices that you use on a regular basis. Computers, smart phones, GPS systems, MP3 players, and satellite radio might come to mind. Next, think about the most exciting modern technologies that 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 of these groundbreaking advancements, 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 and 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, physics principles are propelling new, exciting technologies, and these principles are applied in a wide range of careers.

In this text, you will begin to explore the history of the formal study of physics, beginning with natural philosophy and the ancient Greeks, and leading up through a review of Sir Isaac Newton and the laws of physics that bear his name. You will also be introduced to the standards scientists use when they study physical quantities and the interrelated system of measurements most of the scientific community uses to communicate in a single mathematical language. Finally, you will study the limits of our ability to be accurate and precise, and the reasons scientists go to painstaking lengths to be as clear as possible regarding their own limitations.

Physics: An Introduction

Learning Objectives

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

- Explain the difference between a principle and a law.
- Explain the difference between a model and a theory.

Figure 1. The flight formations of migratory birds such as Canada geese are governed by the laws of physics. (credit: David Merrett)

The physical universe is enormously complex in its detail. Every day, each of us observes a great variety of objects and phenomena. Over the centuries, the curiosity of the human race has led us collectively to explore and catalog a tremendous wealth of information. From the flight of birds to the colors of flowers, from lightning to gravity, from quarks to clusters of galaxies, from the flow of time to the mystery of the creation of the universe, we have asked questions and assembled huge arrays of facts. In the face of all these details, we have discovered that a surprisingly small and unified set of physical laws can explain what we observe. As humans, we make generalizations and seek order. We have found that nature is remarkably cooperative—it exhibits the *underlying order and simplicity* we so value.

It is the underlying order of nature that makes science in general, and physics in particular, so 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.

The unifying aspect of physical laws and the basic simplicity of nature form the underlying themes of this text. In learning to apply these laws, you will, of course, study the most important topics in physics. More importantly, you will gain analytical abilities that will enable you to apply these laws far beyond the scope of what can be included in a single book. These analytical skills will help you to excel academically, and they will also help you to think critically in any professional career you choose to pursue. This module discusses the realm of physics (to define what physics is), some applications of

physics (to illustrate its relevance to other disciplines), and more precisely what constitutes a physical law (to illuminate the importance of experimentation to theory).

Science and the Realm of Physics

Science consists of the theories and laws that are the general truths of nature as well as the body of knowledge they encompass. Scientists are continually trying to expand this body of knowledge and to perfect the expression of the laws that describe it. *Physics* is concerned with describing the interactions of energy, matter, space, and time, and it is especially interested in what fundamental mechanisms underlie every phenomenon. The concern for describing the basic phenomena in nature essentially defines the *realm of physics*.

Physics aims to describe the function of everything around us, from the movement of tiny charged particles to the motion of people, cars, and spaceships. In fact, almost everything around you can be described quite accurately by the laws of physics. Consider a smart phone (Figure 2). 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 the smart phone. Next, consider a GPS system. 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 device in a vehicle, it utilizes these physics equations to determine the travel time from one location to another.

Applications of Physics

You need not be a scientist to use physics. On the contrary, 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. (See Figure 3.) Physics allows you to understand the hazards of radiation and rationally evaluate these hazards 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 through our body's nervous system are much easier to understand when you think about them in terms of basic physics.

Figure 2. The Apple "iPhone" is a common smart phone 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. GPS uses physics equations to determine the driving time between two locations on a map. (credit: @gletham GIS, Social, Mobile Tech Images)

Physics is the foundation of many important disciplines and contributes directly to others. Chemistry, for example—since it deals with the interactions of atoms and molecules—is rooted in atomic and molecular physics. Most branches of engineering are applied 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 in the 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 cell walls and cell membranes (Figure 4 and Figure 5). On the macroscopic level, it can explain the heat, work, and power associated with the human body. Physics is involved in medical diagnostics, such as x-rays, magnetic resonance imaging (MRI), and ultrasonic blood flow measurements. Medical therapy sometimes directly involves physics; for example, cancer radiotherapy uses ionizing radiation. Physics can also explain sensory phenomena, such as how musical instruments make sound, how the eye detects color, and how lasers can transmit information.

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

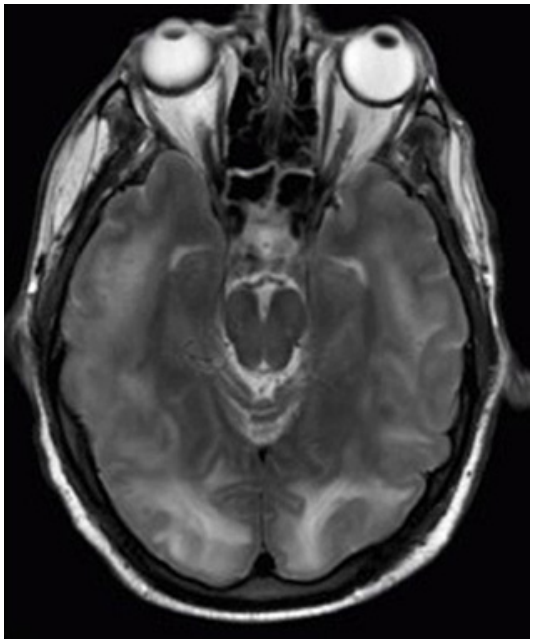


Figure 3. These two applications of physics have more in common than meets the eye. Microwave ovens use electromagnetic waves to heat food. Magnetic resonance imaging (MRI) also uses electromagnetic waves to yield an image of the brain, from which the exact location of tumors can be determined. (credit: Rashmi Chawla, Daniel Smith, and Paul E. Marik)

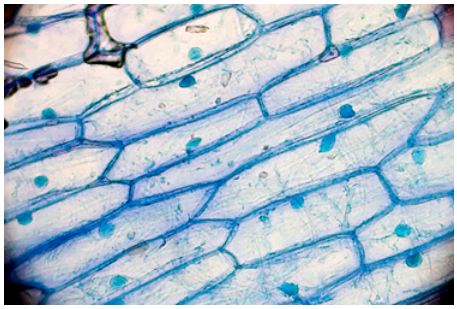


Figure 4. Physics, chemistry, and biology help describe the properties of cell walls in plant cells, such as the onion cells seen here. (credit: Umberto Salvagnin)

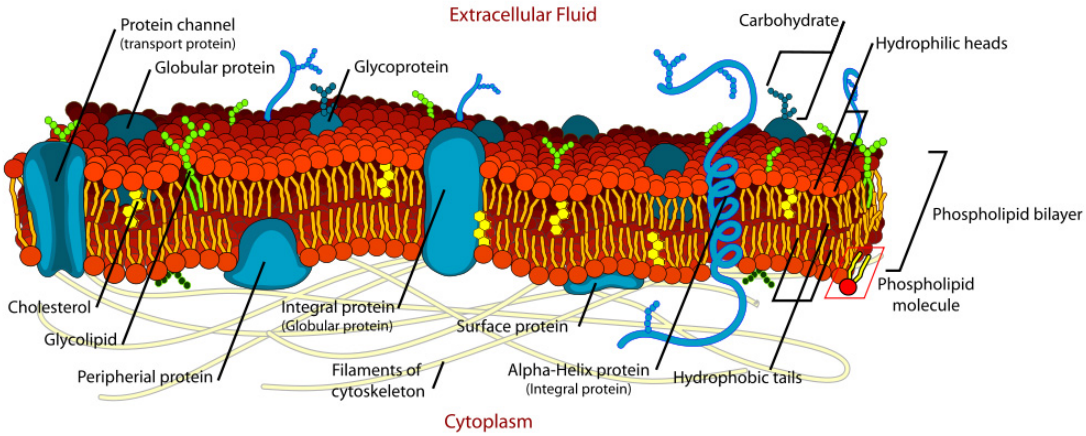


Figure 5. An artist's rendition of the the structure of a cell membrane. Membranes form the boundaries of animal cells and are complex in structure and function. Many of the most fundamental properties of life, such as the firing of nerve cells, are related to membranes. The disciplines of biology, chemistry, and physics all help us understand the membranes of animal cells. (credit: Mariana Ruiz)

Models, Theories, and Laws; The Role of Experimentation

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 so 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. (See Figure 6 and Figure 7.) The cornerstone of discovering natural laws is observation; science must describe the universe as it is, not as we may imagine it to be.

Figure 7. 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: Wikimedia Commons)

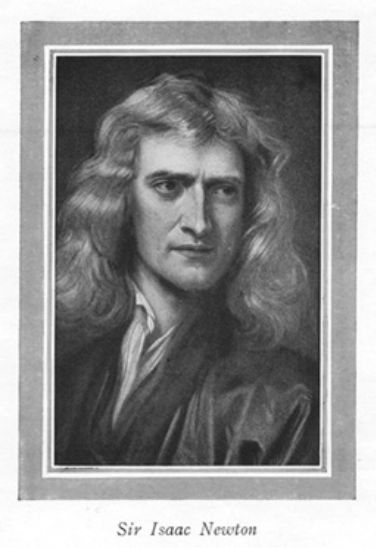


Figure 6. Isaac Newton (1642–1727) was very reluctant to publish his revolutionary work and had to be convinced to do so. In his later years, he stepped down from his academic post and became exchequer of the Royal Mint. He took this post seriously, inventing reeding (or creating ridges) on the edge of coins to prevent unscrupulous people from trimming the silver off of them before using them as currency. (credit: Arthur Shuster and Arthur E. Shipley: Britain's Heritage of Science. London, 1917.)

We all are curious to some extent. We look around, make generalizations, and try to understand what we see—for example, we look up and wonder whether one type of cloud signals an oncoming storm. As we become serious about exploring nature, we become more organized and formal in collecting and analyzing data. We attempt greater precision, perform controlled experiments (if we can), and write down ideas about how the data may be organized and unified. We then formulate models, theories, and laws based on the data we have collected and analyzed to generalize and communicate the results of these experiments.

A *model* is a representation of something that is often too difficult (or impossible) to display directly. While a model is justified with experimental proof, it is only accurate under limited situations. An example is the planetary model of the atom in which electrons are pictured as orbiting the nucleus, analogous to the way planets orbit the Sun. (See Figure 8.) We cannot observe electron orbits directly, but the mental image helps explain the observations we can make, such as the emission of light from hot gases (atomic spectra). Physicists use models for a variety of purposes. For example, models can help physicists analyze a scenario and perform a calculation, or they can be used to represent a situation in the form of a computer simulation. A *theory* is an explanation for patterns in nature that is 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 our instruments tell us about the behavior of gases.

A *law* uses concise language to describe a generalized pattern in nature that is 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 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, mass, and acceleration by the simple equation $\mathbf{F} = m\mathbf{a}$. A theory, in contrast, is a less concise statement of observed

phenomena. For example, the Theory of Evolution and the Theory of Relativity cannot be expressed concisely enough to be considered a law. 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. And, whereas a law is a postulate that forms the foundation of the scientific method, a theory is the end result of that process.

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 carefully made.

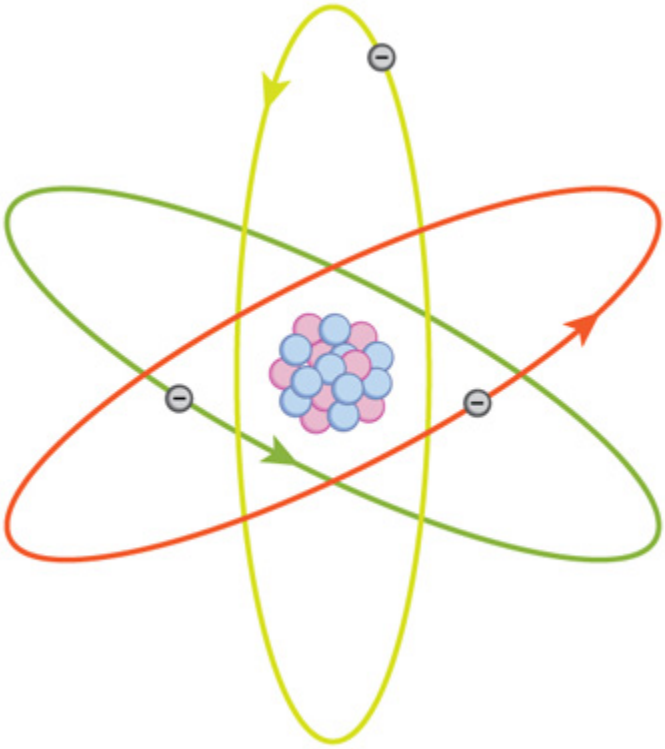


Figure 8. What is a model? This planetary model of the atom shows electrons orbiting the nucleus. It is a drawing that we use to form a mental image of the atom that we cannot see directly with our eyes because it is too small.

Models, Theories, and Laws

Models, theories, and laws are used to help scientists analyze the data they have already collected. However, often after a model, theory, or law has been developed, it points scientists toward new discoveries they would not otherwise have made.

The models, theories, and laws we devise sometimes *imply the existence of objects or phenomena 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 *experiment* 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 in order to confirm a law in 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, then the law must be modified or overthrown completely.

The study of science in general and physics in particular is an adventure much like the exploration of 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.

The Scientific Method

As scientists inquire and gather information about the world, they follow a process called the *scientific method*. This process typically begins with an observation and question that the scientist will research. Next, the scientist typically performs some research about the topic and then devises a hypothesis. Then, the scientist will test the hypothesis by performing an experiment. Finally, the scientist analyzes the results of the experiment and draws a conclusion. Note that the scientific method can be applied to many situations that are not limited to science, and this method can be modified to suit the situation.

Consider an example. Let us say that you try to turn on your car, but it will not start. You undoubtedly wonder: Why will the car not start? You can follow a scientific method to answer this question. First off, you may perform some research to determine a variety of reasons why the car will not start. Next, you will state a hypothesis. For example, you may believe that the car is not starting because it has no engine oil. To test this, you open the hood of the car and examine the oil level. You observe that the oil is at an acceptable level, and you thus conclude that the oil level is not contributing to your car issue. To troubleshoot the issue further, you may devise a new hypothesis to test and then repeat the process again.

The Evolution of Natural Philosophy into Modern Physics

Physics was not always a separate and distinct discipline. It remains connected to other sciences to this day. The word *physics* comes from Greek, meaning nature. The study of nature came to be called “natural philosophy.” From ancient times through the Renaissance, natural philosophy encompassed many fields, including astronomy, biology, chemistry, physics, mathematics, and medicine. Over the last few centuries, the growth of knowledge has resulted in ever-increasing specialization and branching of natural philosophy into separate fields, with physics retaining the most basic facets. (See Figure 9, Figure 10, and Figure 11.) Physics as it developed from the Renaissance to the end of the 19th century is called *classical physics*. It was transformed into modern physics by revolutionary discoveries made starting at the beginning of the 20th century.

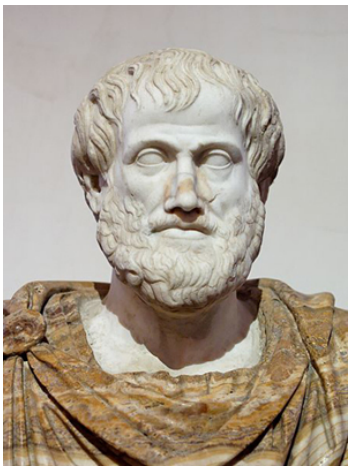


Figure 9. Over the centuries, natural philosophy has evolved into more specialized disciplines, as illustrated by the contributions of some of the greatest minds in history. The Greek philosopher Aristotle (384–322 B.C.) wrote on a broad range of topics including physics, animals, the soul, politics, and poetry. (credit: Jastrow (2006)/Ludovisi Collection)

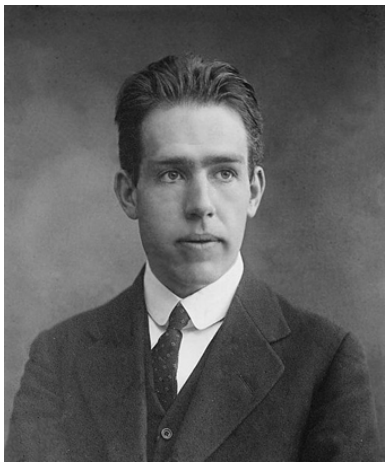


Figure 10. Niels Bohr (1885–1962) made fundamental contributions to the development of quantum mechanics, one part of modern physics. (credit: United States Library of Congress Prints and Photographs Division)

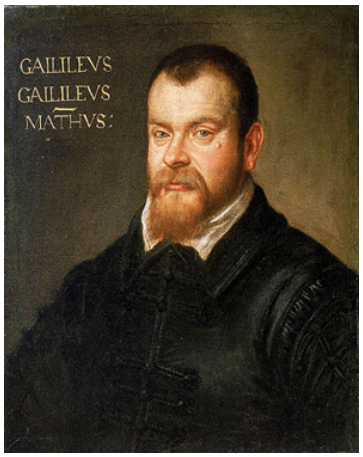


Figure 11. Galileo Galilei (1564–1642) laid the foundation of modern experimentation and made contributions in mathematics, physics, and astronomy. (credit: Domenico Tintoretto)

Classical physics is not an exact description of the universe, but it is an excellent approximation under the following conditions: Matter must be moving at speeds less than about 1% of the speed of light, the objects dealt with must be large enough to be seen with a microscope, and only weak gravitational fields, such as the field generated by the Earth, can be involved. Because humans live under such circumstances, classical physics seems intuitively reasonable, while many aspects of modern physics seem bizarre. This is why models are so useful in modern physics—they let us conceptualize phenomena we do not ordinarily experience. We can relate to models in human terms and visualize what happens when objects move at high speeds or imagine what objects too small to observe with our senses might be like. For example, we can understand an atom's properties because we can picture it in our minds, although we have never seen an atom with our eyes. New tools, of course, allow us to better picture phenomena we cannot see. In fact, new instrumentation has allowed us in recent years to actually "picture" the atom.

Limits on the Laws of Classical Physics

For the laws of classical physics to apply, the following criteria must be met: Matter must be moving at speeds less than about 1% of the speed of light, the objects dealt with must be large enough to be seen with a microscope, and only weak gravitational fields (such as the field generated by the Earth) can be involved.

Some of the most spectacular advances in science have been made in modern physics. Many of the laws of classical physics have been modified or rejected, and revolutionary changes in technology, society, and our view of the universe have resulted. Like science fiction, modern physics is filled with fascinating objects beyond our normal experiences, but it has the advantage over science fiction of being very real. Why, then, is the majority of this text devoted to topics of classical physics? There are two main reasons: Classical physics gives an extremely accurate description of the universe under a wide range of everyday circumstances, and knowledge of classical physics is necessary to understand modern physics.

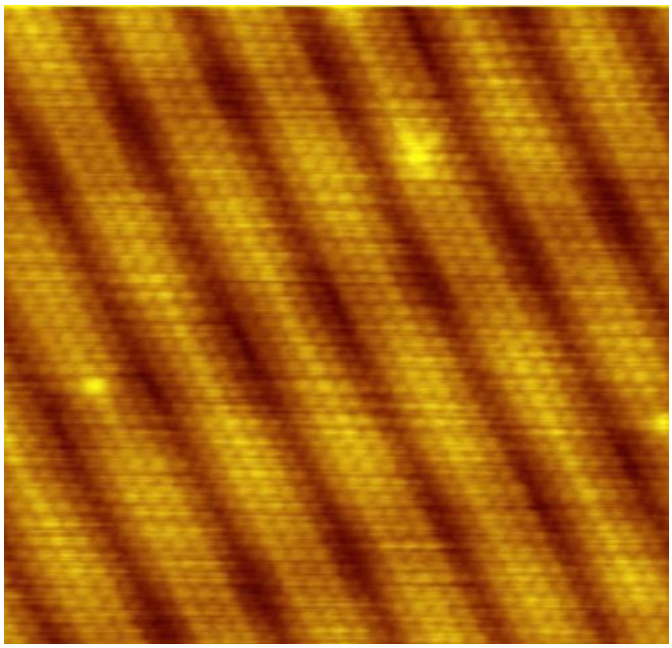


Figure 12. Using a scanning tunneling microscope (STM), scientists can see the individual atoms that compose this sheet of gold. (credit: Erwinrossen)

Modern physics itself consists of the two revolutionary theories, relativity and quantum mechanics. These theories deal with the very fast and the very small, respectively. *Relativity* must be used whenever an object is traveling at greater than about 1% of the speed of light or experiences a strong gravitational field such as that near the Sun. *Quantum mechanics* must be used for objects smaller than can be seen with a microscope. The combination of these two theories is *relativistic quantum mechanics*, and it describes the behavior of small objects traveling at high speeds or experiencing a strong gravitational field. Relativistic quantum mechanics is the best universally applicable theory we have. Because of its mathematical complexity, it is used only when necessary, and the other theories are used whenever they will produce sufficiently accurate results. We will find, however, that we can do a great deal of modern physics with the algebra and trigonometry used in this text.

Check Your Understanding

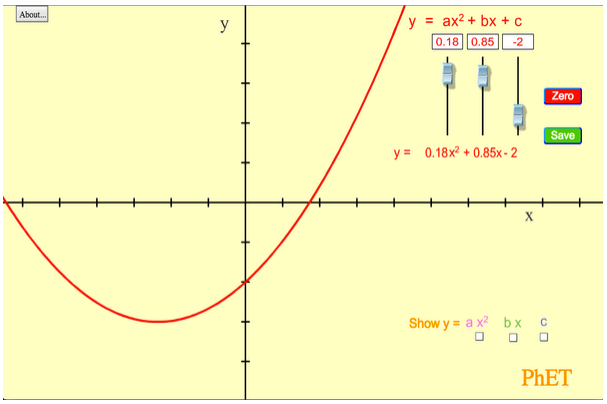
A friend tells you he has learned about a new law of nature. What can you know about the information even before your friend describes the law? How would the information be different if your friend told you he had learned about a scientific theory rather than a law?

Solution

Without knowing the details of the law, you can still infer that the information your friend has learned conforms to the requirements of all laws of nature: it will be a concise description of the universe around us; a statement of the underlying rules that all natural processes follow. If the information had been a theory, you would be able to infer that the information will be a large-scale, broadly applicable generalization.

PhET Explorations: Equation Grapher

Learn about graphing polynomials. The shape of the curve changes as the constants are adjusted. View the curves for the individual terms (e.g. $y = bx$) to see how they add to generate the polynomial curve.



Click to run the simulation.

Section Summary

- Science seeks to discover and describe the underlying order and simplicity in nature.
- Physics is the most basic of the sciences, concerning itself with energy, matter, space and time, and their interactions.
- Scientific laws and theories express the general truths of nature and the body of knowledge they encompass. These laws of nature are rules that all natural processes appear to follow.

Conceptual Questions

1. are particularly useful in relativity and quantum mechanics, where conditions are outside those normally encountered by humans. What is a model?

2. How does a model differ from a theory?

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 to the required validity of a theory or a law?

7. Classical physics is a good approximation to modern physics under certain circumstances. What are they?

8. When is it *necessary* to use relativistic quantum mechanics?

9. Can classical physics be used to accurately describe a satellite moving at a speed of 7500 m/s? Explain why or why not.

Glossary

classical physics:
physics that was developed from the Renaissance to the end of the 19th century

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

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

theory:
an explanation for patterns in nature that is supported by scientific evidence and verified multiple times by various groups of researchers

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

scientific method:
a method that typically begins with an observation and question that the scientist will research; next, the scientist typically performs some research about the topic and then devises a hypothesis; then, the scientist will test the hypothesis by performing an experiment; finally, the scientist analyzes the results of the experiment and draws a conclusion

modern physics:
the study of relativity, quantum mechanics, or both

relativity:
the study of objects moving at speeds greater than about 1% of the speed of light, or of objects being affected by a strong gravitational field

quantum mechanics:
the study of objects smaller than can be seen with a microscope

Physical Quantities and Units

Lumen Learning

Learning Objectives

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

- Perform unit conversions both in the SI and English units.
- Explain the most common prefixes in the SI units and be able to write them in scientific notation.

The range of objects and phenomena studied in physics is immense. From the incredibly short lifetime of a nucleus to the age of the Earth, from the tiny sizes of sub-nuclear 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 does qualitative description alone. To comprehend these vast ranges, we must also have accepted units in which to express them. And 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 four fundamental physical quantities: length, mass, time, and electric current.

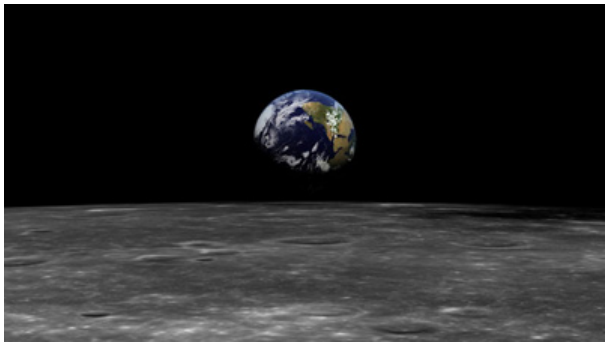


Figure 1. The distance from Earth to the Moon may seem immense, but it is just a tiny fraction of the distances from Earth to other celestial bodies. (credit: NASA)

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 define distance and time by specifying methods for measuring them, whereas we define *average speed* by stating that it is calculated as 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. (See Figure 2.)

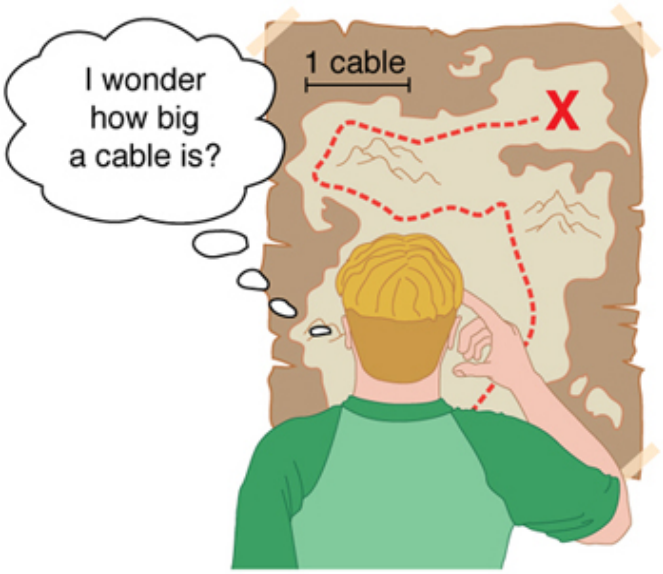


Figure 2. Distances given in unknown units are maddeningly useless.

There are two major systems of units used in the world: *SI units* (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. Virtually every other country in the world now uses SI units as the standard; the metric system is also the standard system agreed upon by scientists and mathematicians. The acronym "SI" is derived from the French *Système International*.

SI Units: Fundamental and Derived Units

Table 1 gives the fundamental SI units that are used throughout this textbook. This text uses non-SI units in a few applications where they are in very common use, such as the measurement of blood pressure in millimeters of mercury (mm Hg). Whenever non-SI units are discussed, they will be tied to SI units through conversions.

Table 1. Fundamental SI Units

Length	Mass	Time	Electric Current
meter (m)	kilogram (kg)	second (s)	ampere (A)

It is an intriguing fact that some physical quantities are more fundamental than others and that the most fundamental physical quantities can be defined *only* in terms of the procedure used to measure them. The units in which they are measured are thus called *fundamental units*. In this textbook, the fundamental physical quantities are taken to be length, mass, time, and electric current. (Note that electric current will not be introduced until much later in this text.) All other physical quantities, such as force and electric charge, can be expressed as algebraic combinations of length, mass, time, and current (for example, speed is length divided by time); these units are called *derived units*.

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

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 non-varying, or constant, physical phenomenon (because the solar day is getting longer due to very gradual slowing of the 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. (See Figure 3.)

Accuracy in the fundamental units is essential, because all measurements are ultimately expressed in terms of fundamental units and can be no more accurate than are the fundamental units themselves.

The Meter

The SI unit for length is the *meter* (abbreviated m); its definition has also changed over time to become more accurate and 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 present definition (partly for greater accuracy) as the distance light travels in a vacuum in 1/299,792,458 of a second. (See Figure 4.) This change defines the speed of light to be exactly 299,792,458 meters per second. The length of the meter will change if the speed of light is someday measured with greater accuracy.

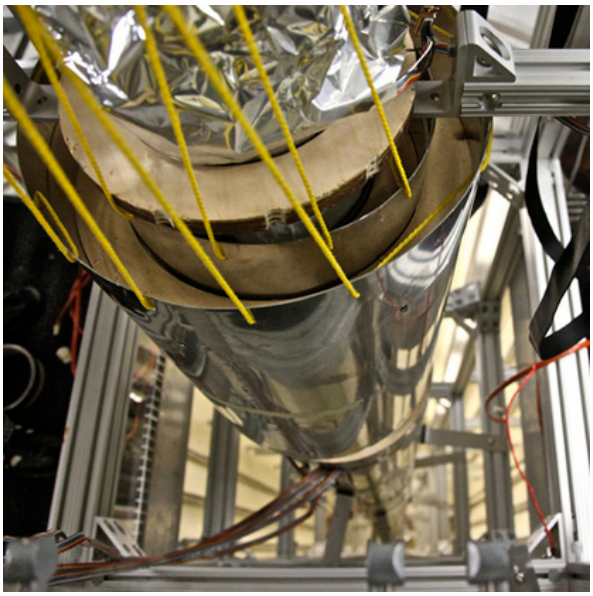


Figure 3. An atomic clock such as this one uses the vibrations of cesium atoms to keep time to a precision of better than a microsecond per year. The fundamental unit of time, the second, is based on such clocks. This image is looking down from the top of an atomic fountain nearly 30 feet tall! (credit: Steve Jurvetson/Flickr)

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 United States' National Institute of Standards and Technology, or NIST, located in Gaithersburg, Maryland outside of Washington D.C., and at other locations around the world. The determination of all other masses can be ultimately traced to a comparison with the standard mass.

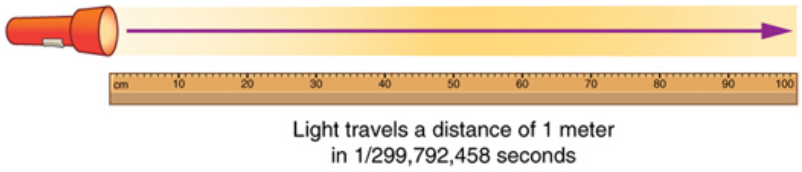


Figure 4. The meter is defined to be the distance light travels in 1/ 299,792,458 of a second in a vacuum. Distance traveled is speed multiplied by time.

Electric current and its accompanying unit, the ampere, will be introduced in Introduction to Electric Current, Resistance, and Ohm's Law when electricity and magnetism are covered. The initial modules in this textbook are concerned with mechanics, fluids, heat, and waves. In these subjects all pertinent physical quantities can be expressed in terms of the fundamental units of length, mass, and time.

Metric Prefixes

SI units are part of the *metric system*. The metric system is convenient for scientific and engineering calculations because the units are categorized by factors of 10. Table 2 gives metric prefixes and symbols used to denote various factors of 10.

Metric systems have the advantage that conversions of units involve only powers of 10. There are 100 centimeters in a meter, 1000 meters in a kilometer, and so on. In non-metric systems, such as the system of U.S. customary units, the relationships are not as simple—there are 12 inches in a foot, 5280 feet in a mile, and so on. Another advantage of the metric system is that the same unit can be used over extremely large ranges of values simply by using an appropriate metric prefix. For example, distances in meters are suitable in construction, while distances in kilometers are appropriate for air travel, and the tiny measure of nanometers are convenient in optical design. With the metric system there is no need to invent new units for particular applications.

The term order of magnitude refers to the scale of a value expressed in the metric system. Each power of 10 in the metric system represents a different order of magnitude. For example, 10^1, 10^2, 10^3, and so forth are all different orders of magnitude. All quantities that can be expressed as a product of a specific power of 10 are said to be of the *same* order of magnitude. For example, the number 800 can be written as 8×10^2, and the number 450 can be written as 4.5×10^2. Thus, the numbers 800 and 450 are of the same order of magnitude: 10^2. Order of magnitude can be thought of as a ballpark estimate for the scale of a value. The diameter of an atom is on the order of 10^{-9}m while the diameter of the Sun is on the order of 10^9m.

The Quest for Microscopic Standards for Basic Units

The fundamental units described in this chapter are those that produce the greatest accuracy and precision in measurement. There is a sense among physicists that, because there is an underlying microscopic substructure to matter, it would be most satisfying to base our standards of measurement on microscopic objects and fundamental physical phenomena such as the speed of light. A microscopic standard has been accomplished for the standard of time, which is based on the oscillations of the cesium atom.

The standard for length was once based on the wavelength of light (a small-scale length) emitted by a certain type of atom, but it has been supplanted by the more precise measurement of the speed of light. If it becomes possible to measure the mass of atoms or a particular arrangement of atoms such as a silicon sphere to greater precision than the kilogram standard, it may become possible to base mass measurements on the small scale. There are also possibilities that electrical phenomena on the small scale may someday allow us to base a unit of charge on the charge of electrons and protons, but at present current and charge are related to large-scale currents and forces between wires.

Table 2. Metric Prefixes for Powers of 10 and their Symbols

Prefix	Symbol	Value[1]	Example (some are approximate)			
exa	E	10^{18}	exameter	Em	10^{18} m	distance light travels in a century
peta	P	10^{15}	petasecond	Ps	10^{15} s	30 million years
tera	T	10^{12}	terawatt	TW	10^{12} W	powerful laser output
giga	G	10^{9}	gigahertz	GHz	10^{9} Hz	a microwave frequency
mega	M	10^{6}	megacurie	MCi	10^{6} Ci	high radioactivity
kilo	k	10^{3}	kilometer	km	10^{3} m	about 6/10 mile
hecto	h	10^{2}	hectoliter	hL	10^{2} L	26 gallons
deka	da	10^{1}	dekagram	dag	10^{1} g	teaspoon of butter
—	—	10^{0} (=1)				
deci	d	10^{-1}	deciliter	dL	10^{-1} L	less than half a soda
centi	c	10^{-2}	centimeter	cm	10^{-2} m	fingertip thickness
milli	m	10^{-3}	millimeter	mm	10^{-3} m	flea at its shoulders
micro	μ	10^{-6}	micrometer	μm	10^{-6} m	detail in microscope
nano	n	10^{-9}	nanogram	ng	10^{-9} g	small speck of dust
pico	p	10^{-12}	picofarad	pF	10^{-12} F	small capacitor in radio
femto	f	10^{-15}	femtometer	fm	10^{-15} m	size of a proton
atto	a	10^{-18}	attosecond	as	10^{-18} s	time light crosses an atom

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 in Table 1.3. Examination of this table will give you some feeling for the range of possible topics and numerical values. (See Figure 5 and Figure 6.)

1. See Appendix A for a discussion of powers of 10.

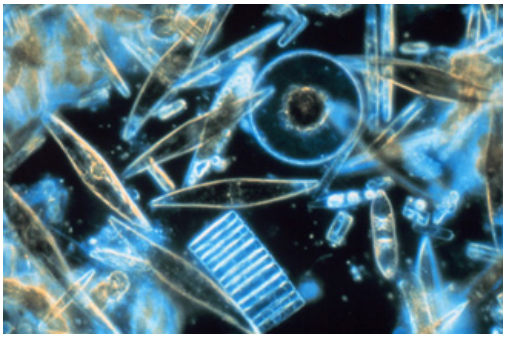


Figure 5. Tiny phytoplankton swims among crystals of ice in the Antarctic Sea. They range from a few micrometers to as much as 2 millimeters in length. (credit: Prof. Gordon T. Taylor, Stony Brook University; NOAA Corps Collections)

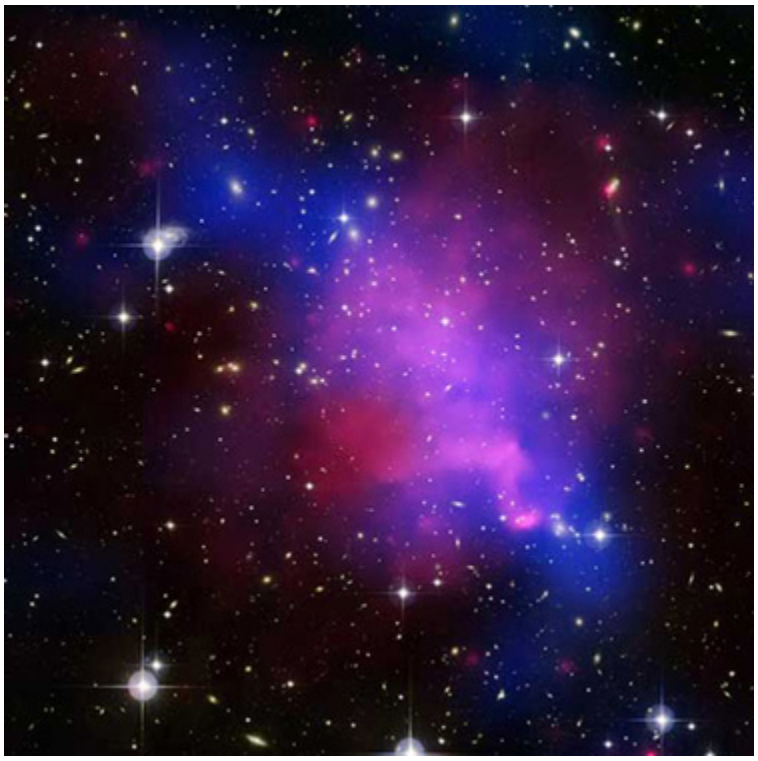


Figure 6. Galaxies collide 2.4 billion light years away from Earth. The tremendous range of observable phenomena in nature challenges the imagination. (credit: NASA/CXC/UVic./A. Mahdavi et al. Optical/lensing: CFHT/UVic./H. Hoekstra et al.)

Unit Conversion and Dimensional Analysis

It is often necessary to convert from one type of 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 will need to convert units of feet to miles.

Let us consider a simple example of how to convert units. Let us say that we want to convert 80 meters (m) to kilometers (km).

The first thing to do is to list the units that you have and the units that you want to convert to. 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 expressing how many of one unit are equal to another unit. For example, there are 12 inches in 1 foot, 100 centimeters in 1 meter, 60 seconds in 1 minute, and so on. In this case, we know that there are 1,000 meters in 1 kilometer.

Now we can set up our unit conversion. We will write the units that we have and then multiply them by the conversion factor so that the units cancel out, as shown:

$$80)\overline{\mathrm{m}} \times \frac{1\ \mathrm{km}}{1000)\overline{\mathrm{m}}} = 0.080\ \mathrm{km}.$$

Note that the unwanted m unit cancels, leaving only the desired km unit. You can use this method to convert between any types of unit.

Table 3. Approximate Values of Length, Mass, and Time

Lengths in meters		Masses in kilograms (more precise values in parentheses)		Times in seconds (more precise values in parentheses)	
10^{-18}	Present experimental limit to smallest observable detail	10^{-30}	Mass of an electron $(9.11 \times 10^{-31}\ \mathrm{kg})$	10^{-23}	Time for light to cross a proton
10^{-15}	Diameter of a proton	10^{-27}	Mass of a hydrogen atom $(1.67 \times 10^{-27}\ \mathrm{kg})$	10^{-22}	Mean life of an extremely unstable nucleus
10^{-14}	Diameter of a uranium nucleus	10^{-15}	Mass of a bacterium	10^{-15}	Time for one oscillation of visible light
10^{-10}	Diameter of a hydrogen atom	10^{-5}	Mass of a mosquito	10^{-13}	Time for one vibration of an atom in a solid
10^{-8}	Thickness of membranes in cells of living organisms	10^{-2}	Mass of a hummingbird	10^{-8}	Time for one oscillation of an FM radio wave
10^{-6}	Wavelength of visible light	1	Mass of a liter of water (about a quart)	10^{-3}	Duration of a nerve impulse
10^{-3}	Size of a grain of sand	10^{2}	Mass of a person	1	Time for one heartbeat
1	Height of a 4-year-old child	10^{3}	Mass of a car	10^{5}	One day $(8.64 \times 10^{4}\mathrm{s})$
10^{2}	Length of a football field	10^{8}	Mass of a large ship	10^{7}	One year (y) $(3.16 \times 10^{7}\mathrm{s})$
10^{4}	Greatest ocean depth	10^{12}	Mass of a large iceberg	10^{9}	About half the life expectancy of a human
10^{7}	Diameter of the Earth	10^{15}	Mass of the nucleus of a comet	10^{11}	Recorded history
10^{11}	Distance from the Earth to the Sun	10^{23}	Mass of the Moon $(7.35 \times 10^{22}\ \mathrm{kg})$	10^{17}	Age of the Earth
10^{16}	Distance traveled by light in 1 year (a light year)	10^{25}	Mass of the Earth $(5.97 \times 10^{24}\ \mathrm{kg})$	10^{18}	Age of the universe
10^{21}	Diameter of the Milky Way galaxy	10^{30}	Mass of the Sun $(1.99 \times 10^{30}\ \mathrm{kg})$		
10^{22}	Distance from the Earth to the nearest large galaxy (Andromeda)	10^{42}	Mass of the Milky Way galaxy (current upper limit)		
10^{26}	Distance from the Earth to the edges of the known universe	10^{53}	Mass of the known universe (current upper limit)		

Example 1. Unit Conversions – A Short Drive Home

Suppose that you drive the 10.0 km from your university to home in 20.0 min. Calculate your average speed (a) in kilometers per hour (km/h) and (b) 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 factor and multiplying by it. The correct conversion factor is the one that cancels the unwanted unit and leaves the desired unit in its place.

Solution for (a)

(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 will be covered in a later module.) In equation form,

$$\text{average speed} = \frac{\text{distance}}{\text{time}}.$$

(2) Substitute the given values for distance and time.

$$\text{average speed} = \frac{10.0\ \text{km}}{20.0\ \text{min}} = 0.500\ \frac{\text{km}}{\text{min}}.$$

(3) Convert km/min to km/h: multiply by the conversion factor that will cancel minutes and leave hours. That conversion factor is 60 min/hr. Thus,

$$\text{average speed} = 0.500\ \frac{\text{km}}{\text{min}} \times \frac{60\ \text{min}}{1\ \text{h}} = 30.0\ \frac{\text{km}}{\text{h}}.$$

Discussion for (a)

To check your answer, consider the following:

(1) Be sure that you have properly cancelled the units in the unit conversion. If you have written the unit conversion factor upside down, the units will not cancel properly in the equation. If you accidentally get the ratio upside down, then the units will not cancel; rather, they will give you the wrong units as follows:

$$\frac{\text{km}}{\text{min}} \times \frac{1\ \text{hr}}{60\ \text{min}} = \frac{1}{60}\ \frac{\text{km·hr}}{\text{min}^2},$$

which are obviously not the desired units of km/h.

(2) Check that the units of the final answer are the desired units. The problem asked us to solve for average speed in units of km/h and we have indeed obtained these units.

(3) Check the significant figures. Because each of the values given in the problem has three significant figures, the answer should also have three significant figures. The answer 30.0 km/hr does indeed have three significant figures, so this is appropriate. Note that the significant figures in the conversion factor are not relevant because an hour is *defined* to be 60 minutes, so the precision of the conversion factor is perfect.

(4) Next, check whether the answer is reasonable. Let us consider some information from the problem—if you travel 10 km in a third of an hour (20 min), you would travel three times that far in an hour. The answer does seem reasonable.

Solution for (b)

There are several ways to convert the average speed into meters per second.

(1) Start with the answer to (a) and convert km/h to m/s. Two conversion factors are needed—one to convert hours to seconds, and another to convert kilometers to meters.

(2) Multiplying by these yields

$$\text{Average speed} = 30.0\frac{\text{km}}{\text{h}} \times \frac{1\text{h}}{3{,}600\text{ s}} \times \frac{1{,}000\text{m}}{1\text{ km}},$$
$$\text{Average speed} = 8.33\frac{\text{m}}{\text{s}}$$
.

Discussion for (b)

If we had started with 0.500 km/min, we would have needed different conversion factors, but the answer would have been the same: 8.33 m/s.

You may have noted that the answers in the worked example just covered were given to three digits. Why? When do you need to be concerned about the number of digits in something you calculate? Why not write down all the digits your calculator produces? The module Accuracy, Precision, and Significant Figures will help you answer these questions.

Nonstandard Units

While there are numerous types of units that we are all familiar with, there are others that are much more obscure. For example, a **firkin** is a unit of volume that was once used to measure beer. One firkin equals about 34 liters. To learn more about nonstandard units, use a dictionary or encyclopedia to research different "weights and measures." Take note of any unusual units, such as a barleycorn, that are not listed in the text. Think about how the unit is defined and state its relationship to SI units.

Check Your Understanding

1. Some hummingbirds beat their wings more than 50 times per second. A scientist is measuring the time it takes for a hummingbird to beat its wings once. Which fundamental unit should the scientist use to describe the measurement? Which factor of 10 is the scientist likely to use to describe the motion precisely? Identify the metric prefix that corresponds to this factor of 10.

2. One cubic centimeter is equal to one milliliter. What does this tell you about the different units in the SI metric system?

Solutions

1. The scientist will measure the time between each movement using the fundamental unit of seconds.

Because the wings beat so fast, the scientist will probably need to measure in milliseconds, or 10^{-3} seconds. (50 beats per second corresponds to 20 milliseconds per beat.)

2. The fundamental unit of length (meter) is probably used to create the derived unit of volume (liter). The measure of a milliliter is dependent on the measure of a centimeter.

Section Summary

- Physical quantities are a characteristic or property of an object that can be measured or calculated from other measurements.
- Units are standards for expressing and comparing the measurement of physical quantities. All units can be expressed as combinations of four fundamental units.
- The four fundamental units we will use in this text are the meter (for length), the kilogram (for mass), the second (for time), and the ampere (for electric current). These units are part of the metric system, which uses powers of 10 to relate quantities over the vast ranges encountered in nature.
- The four fundamental units are abbreviated as follows: meter, m; kilogram, kg; second, s; and ampere, A. The metric system also uses a standard set of prefixes to denote each order of magnitude greater than or lesser than the fundamental unit itself.
- Unit conversions involve changing a value expressed in one type of unit to another type of unit. This is done by using conversion factors, which are ratios relating equal quantities of different units.

Conceptual Questions

1. Identify some advantages of metric units.

Problems & Exercises

1. 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?

2. 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?

3. Show that 1.0 m/s = 3.6 km/h. Hint: Show the explicit steps involved in converting 1.0 m/s = 3.6 km/h.

4. American football is played on a 100-yd-long field, excluding the end zones. How long is the field in meters? (Assume that 1 meter equals 3.281 feet.)

5. Soccer fields vary in size. A large soccer field is 115 m long and 85 m wide. What are its dimensions in feet and inches? (Assume that 1 meter equals 3.281 feet.)

6. What is the height in meters of a person who is 6 ft 1.0 in. tall? (Assume that 1 meter equals 39.37 in.)

7. Mount Everest, at 29,028 feet, is the tallest mountain on the Earth. What is its height in kilometers? (Assume that 1 kilometer equals 3,281 feet.)

8. The speed of sound is measured to be 342 m/s on a certain day. What is this in km/h?

9. Tectonic plates are large segments of the Earth's crust that move slowly. Suppose that one such plate has an average speed of 4.0 cm/year. (a) What distance does it move in 1 s at this speed? (b) What is its speed in kilometers per million years?

10. (a) Refer to *Table 2: Metric Prefixes for Powers of 10 and their Symbols* to determine the average distance between the Earth and the Sun. Then calculate the average speed of the Earth in its orbit in kilometers per second. (b) What is this in meters per second?

Glossary

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

units:
a standard used for expressing and comparing measurements

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

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

fundamental units:
units that can only be expressed relative to the procedure used to measure them

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

second:
the SI unit for time, abbreviated (s)

meter:
the SI unit for length, abbreviated (m)

kilogram:
the SI unit for mass, abbreviated (kg)

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

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

conversion factor:
a ratio expressing how many of one unit are equal to another unit

Selected Solutions to Problems & Exercises

1. (a) 27.8 m/s (b) 62.1 mph

3.

$$\frac{1.0\ \text{m}}{s} = \frac{1.0\ \text{m}}{s} \times \frac{3600\ \text{s}}{1\ \text{hr}} \times \frac{1km}{1000\ \text{m}}$$

5. length: 377 ft; 4.53×10^3 in. width: 280 ft; 3.3×10^3 in.

7. 8.847 kn

9. (a) 1.3×10^{-9} m (b) 40 km/My

Accuracy, Precision, and Significant Figures

Lumen Learning

Learning Objectives

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

- Determine the appropriate number of significant figures in both addition and subtraction, as well as multiplication and division calculations.
- Calculate the percent uncertainty of a measurement.

Figure 1. 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 gram, 10 grams, and 100 grams. (credit: Serge Melki)

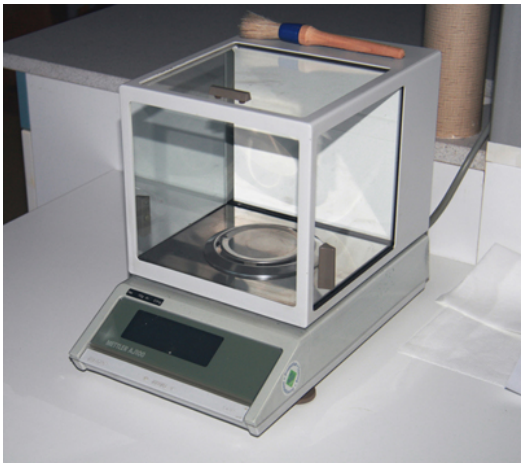


Figure 2. Many mechanical balances, such as double-pan balances, have been replaced by digital scales, which can typically measure the mass of an object more precisely. Whereas a mechanical balance may only read the mass of an object to the nearest tenth of a gram, many digital scales can measure the mass of an object up to the nearest thousandth of a gram. (credit: 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 correct value for that measurement. For example, let us say that you are measuring the length of standard computer paper. The packaging in which you purchased the paper states that it is 11.0 inches long. You 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 correct value of 11.0 inches. In contrast, if you had obtained a measurement of 12 inches, your measurement would not be very accurate.

The *precision* of a measurement system is refers to how close the agreement is between repeated 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 would be to determine the range, or difference, between the lowest and the highest measured values. In that 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, 11.1, and 11.9, then the measurements would not be very precise because there would be significant variation from one measurement to another.

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 us consider an example of a GPS system that is 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 3, you can see that 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 4, 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.

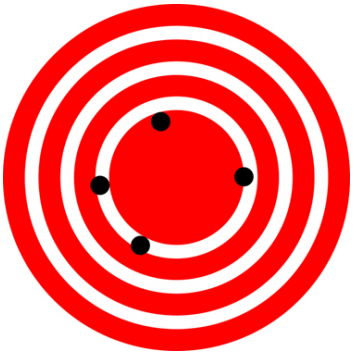


Figure 3. A GPS system attempts to locate a restaurant at the center of the bull's-eye. The black dots represent each attempt to pinpoint the location of the restaurant. The dots are spread out quite far apart from one another, indicating low precision, but they are each rather close to the actual location of the restaurant, indicating high accuracy. (credit: Dark Evil)

Accuracy, Precision, and Uncertainty

Figure 4. In this figure, 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: Dark Evil)

The degree of accuracy and precision of a measuring system are related to the *uncertainty* in the measurements. Uncertainty is a quantitative measure of how much your measured values deviate from a standard or expected value. If your measurements are not very accurate or precise, then the uncertainty of your values will be very high. In more general terms, uncertainty can be thought of as a disclaimer for your measured values. For example, if someone asked you to provide the mileage on your car, you might say that it is 45,000 miles, plus or minus 500 miles. The plus or minus amount is the uncertainty in your value. That is, you are indicating that the actual mileage of your car might be as low as 44,500 miles or as high as 45,500 miles, or anywhere in between. All measurements contain some amount of uncertainty. In our example of measuring the length of the paper, we might say that the length of the paper is 11 in., plus or minus 0.2 in. The uncertainty in a measurement, A, is often denoted as δA ("delta A"), so the measurement result would be recorded as $A \pm \delta A$. In our paper example, the length of the paper could be expressed as 11 in. $\pm$ 0.2.

The factors contributing to uncertainty in a measurement include:

1. Limitations of the measuring device,
2. The skill of the person making the measurement,
3. Irregularities in the object being measured,
4. Any other factors that affect the outcome (highly dependent on the situation).

In our example, such factors contributing to the uncertainty could be the following: the smallest division on the ruler is 0.1 in., the person using the ruler has bad eyesight, or one side of the paper is slightly longer than the other. At any rate, the uncertainty in a measurement must be based on a careful consideration of all the factors that might contribute and their possible effects.

Making Connections: Real-World Connections – Fevers or Chills?

Uncertainty is a critical piece of information, both in physics and in many other real-world applications. Imagine you are caring for a sick child. You suspect the child has a fever, so you check his or her temperature with a thermometer. What if the uncertainty of the thermometer were 3º? If the child's temperature reading was 37ºC(which is normal body temperature), the "true" temperature could be anywhere from a hypothermic 34º to a dangerously high 40º. A thermometer with an uncertainty of 3º would be useless.

Percent Uncertainty

One method of expressing uncertainty is as a percent of the measured value. If a measurement A is expressed with uncertainty, δA, the percent uncertainty (%unc) is defined to be

$$\%\ \text{unc} = \frac{\delta A}{A} \times 100\%$$

Example 1: Calculating Percent Uncertainty: A Bag of Apples

A grocery store sells a 5-pound bags of apples. You purchase four bags over the course of a month and weigh the apples each time. You 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

You determine that the weight of the 5-pound bag has an uncertainty of ±0.4 lb. What is the percent uncertainty of the bag's weight?

Strategy

First, observe that the expected value of the bag's weight, A, is 5 lb. The uncertainty in this value, δA, is 0.4 lb. We can use the following equation to determine the percent uncertainty of the weight:

$$\%\ \text{unc} = \frac{\delta A}{A} \times 100\%$$

Solution

Plug the known values into the equation:

$$\%\ \text{unc} = \frac{0.4\ \text{lb}}{5\ \text{lb}} \times 100\% = 8\%$$

Discussion

We can conclude that the weight of the apple bag is 5 lb ± 8%. Consider how this percent uncertainty would change if the bag of apples were half as heavy, but the uncertainty in the weight remained the same. Hint for future calculations: when calculating percent uncertainty, always remember that you must multiply the fraction by 100%. If you do not do this, you will have a decimal quantity, not a percent value.

Uncertainties in Calculations

There is an uncertainty 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 have uncertainties. How big is the uncertainty in something you calculate by multiplication or division? If the measurements going into the calculation have small uncertainties (a few percent or less), then the method of adding percents can be used for multiplication or division. This method says that *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*.For example, if a floor has a length of 4.00m and a width of 3.00m, with uncertainties of 2% and 1%, respectively, then the area of the floor is 12.0 m^2 and has an uncertainty of 3. (Expressed as an area this is 0.36 m^2, which we round to 0.4 m^2 since the area of the floor is given to a tenth of a square meter.)

Check Your Understanding

A high school track coach has just purchased a new stopwatch. The stopwatch manual states that 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?

No, the uncertainty in the stopwatch is too great to effectively differentiate between the sprint times.

Precision of Measuring Tools and Significant Figures

An important factor in the accuracy and 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, while a caliper can measure length to the nearest 0.01 millimeter. 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 and accurate the measurements can be.

When we express measured values, we can only list as many digits as we initially measured with our measuring tool. For example, if you use a standard ruler to measure the length of a stick, you may measure it to be 36.7 cm. You could not express this value as 36.71 cm because your measuring tool was 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 that 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*. In order 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 36.7cm has three digits, or significant figures. Significant figures indicate the precision of a measuring tool that was 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 only placekeepers that locate the decimal point. There are two significant figures in 0.053. The zeros in 10.053 are not placekeepers but 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 placekeepers. So 1300 could have two, three, or four significant figures. (To avoid this ambiguity, write 1300 in scientific notation.) *Zeros are significant except when they serve only as placekeepers*.

Check Your Understanding

Determine the number of significant figures in the following measurements:

1. 0.0009
2. 15,450.0
3. 6×10^3
4. 87.990
5. 30.42

(a) 1; the zeros in this number are placekeepers that indicate the decimal point

(b) 6; here, the zeros indicate that a measurement was made to the 0.1 decimal point, so the zeros are significant

(c) 1; the value 10^3 signifies the decimal place, not the number of measured values

(d) 5; the final zero indicates that a measurement was made to the 0.001 decimal point, so it is significant

(e) 4; any zeros located in between significant figures in a number are also significant

Significant Figures in Calculations

When combining measurements with different degrees of accuracy and 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, as discussed below.

1. For multiplication and division: *The result should have the same number of significant figures as the quantity having the least 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 us see how many significant figures the area has if the radius has only two—say, r = 1.2 m. Then,

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

is what you would get using a calculator that has an eight-digit output. But because the radius has only

two significant figures, it limits the calculated quantity to two significant figures or $A = 4.5\text{m}^2$, even though 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 that you buy 7.56-kg of potatoes in a grocery store as measured with a scale with precision 0.01 kg. Then you drop off 6.052-kg of potatoes at your laboratory as measured by a scale with precision 0.001 kg. Finally, you 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 you now have, and how many significant figures are appropriate in the answer? The mass is found by simple addition and subtraction:

$$7.56\text{ kg} - 6.052\text{ kg} + 13.7\text{ kg} = 15.208\text{ kg} = 15.2\text{kg}$$

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. You will note that 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 more than three significant figures will be used. 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.

Check Your Understanding

Perform the following calculations and express your answer using the correct number of significant digits.

(a) A woman has two bags weighing 13.5 pounds and one bag with a weight of 10.2 pounds. 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.)

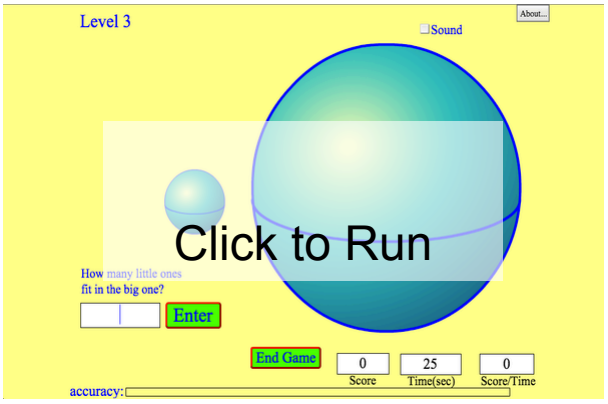
Solutions

(a) 37.2 pounds; 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.

PhET Explorations: Estimation

Explore size estimation in one, two, and three dimensions! Multiple levels of difficulty allow for progressive skill improvement.



Summary

- Accuracy of a measured value refers to how close a measurement is to the correct value. The uncertainty in a measurement is an estimate of the amount by which the measurement result may differ from this value.
- Precision of measured values refers to how close the agreement is between repeated measurements.
- 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.

Conceptual Questions

1. What is the relationship between the accuracy and uncertainty of a measurement?

2. Prescriptions for vision correction are given in units called *diopters* (D). Determine the meaning of that unit. Obtain information (perhaps by calling an optometrist or performing an internet search) on the minimum uncertainty with which corrections in diopters are determined and the accuracy with which corrective lenses can be produced. Discuss the sources of uncertainties in both the prescription and accuracy in the manufacture of lenses.

Problems & Exercises

Express your answers to problems in this section to the correct number of significant figures and proper units.

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

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

3. (a) A car speedometer has a 5.0% uncertainty. What is the range of possible speeds when it reads 90 km/h? Convert this range to miles per hour. (1 km = 0.6214 m)

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

5. (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 y? (b) In 2.00 y? (c) In 2.000 y?

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

7. 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) (1.60×10^{-19}) (3712).

8. (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?

9. (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?

10. (a) A person's blood pressure is measured to be 120 ± 2 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?

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

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

13. If a marathon runner averages 9.5 mi/h, how long does it take him or her to run a 26.22-mi marathon?

14. 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 1s in the elapsed time. (a) Calculate the percent uncertainty in the distance. (b) Calculate the uncertainty in the elapsed time. (c) What is the average speed in meters per second? (d) What is the uncertainty in the average speed?

15. The sides of a small rectangular box are measured to be 180 ± 0.01 cm long, 2.05 ± 0.02 cm, and 3.1 ± 0.1 cm long. Calculate its volume and uncertainty in cubic centimeters.

16. When non-metric units were used in the United Kingdom, a unit of mass called the *pound-mass* (lbm) was employed, where 11bm = 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?

17. 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.

18. A car engine moves a piston with a circular cross section of 7.500 ± 0.002 cm diameter in 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.

Glossary

accuracy:
the degree to which a measured value agrees with correct value for that measurement

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

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

precision:
the degree to which repeated measurements agree with each other

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

uncertainty:
a quantitative measure of how much your measured values deviate from a standard or expected value

Selected Solutions to Problems & Exercises

1. 2 kg

3. (a) 85.5 to 94.5 km/h (b)53.1 to 58.7 mi/h

5. (a) 7.6×10^7 beats (b) 7.57×10^7 beats (c) 7.57×10^7 beats

7. (a) 3 (b) 3 (c) 3

9. (a) 2.2% (b) 59 to 61 km/h

11. 80 ± 3 beats/min

13. 2.6 h

15. 11 ± 1 cm^3

17. 12.06 ± 0.04 m^2

Approximation

Lumen Learning

Learning Objectives

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

- Make reasonable approximations based on given data.

On many occasions, physicists, other scientists, and engineers need to make *approximations* or "guesstimates" for a particular quantity. What is the distance to a certain destination? What is the approximate density of a given item? About how large a current will there be in a circuit? Many approximate numbers are based on formulae in which the input quantities are known only to a limited accuracy. As you develop problem-solving skills (that can be applied to a variety of fields through a study of physics), you will also develop skills at approximating. You will develop these skills through thinking more quantitatively, and by being willing to take risks. As with any endeavor, experience helps, as well as familiarity with units. These approximations allow us to rule out certain scenarios or unrealistic numbers. Approximations also allow us to challenge others and guide us in our approaches to our scientific world. Let us do two examples to illustrate this concept.

Example 1. Approximating the Height of a Building

Can you approximate the height of one of the buildings on your campus, or in your neighborhood? Let us make an approximation based upon the height of a person. In this example, we will calculate the height of a 39-story building.

Strategy

Think about the average height of an adult male. We can approximate the height of the building by scaling up from the height of a person.

Solution

Based on information in the example, we know there are 39 stories in the building. If we use the fact that the height of one story is approximately equal to about the length of two adult humans (each human is about 2-m tall), then we can estimate the total height of the building to be

$$\frac{2\text{ m}}{1\text{ person}} \times \frac{2\text{ person}}{1\text{ story}} \times 39\text{ stories} = 156\text{ m}$$

.

Discussion

You can use known quantities to determine an approximate measurement of unknown quantities. If your hand measures 10 cm across, how many hand lengths equal the width of your desk? What other measurements can you approximate besides length?

Example 2. Approximating Vast Numbers: a Trillion Dollars

Figure 1. A bank stack contains one-hundred $100 bills, and is worth $10,000. How many bank stacks make up a trillion dollars? (credit: Andrew Magill)

The U.S. federal deficit in the 2008 fiscal year was a little greater than $10 trillion. Most of us do not have any concept of how much even one trillion actually is. Suppose that you were given a trillion dollars in $100 bills. If you made 100-bill stacks and used them to evenly cover a football field (between the end zones), make an approximation of how high the money pile would become. (We will use feet/inches rather than meters here because football fields are measured in yards.) One of your friends says 3 in., while another says 10 ft. What do you think?

Strategy

When you imagine the situation, you probably envision thousands of small stacks of 100 wrapped $100 bills, such as you might see in movies or at a bank. Since this is an easy-to-approximate quantity, let us start there. We can find the volume of a stack of 100 bills, find out how many stacks make up one trillion dollars, and then set this volume equal to the area of the football field multiplied by the unknown height.

Solution

(1) Calculate the volume of a stack of 100 bills. The dimensions of a single bill are approximately 3 in. by 6 in. A stack of 100 of these is about 0.5 in. thick. So the total volume of a stack of 100 bills is:

$$\text{volume of stack} = \text{length} \times \text{width} \times \text{height}$$

$$\text{volume of stack} = 6 \text{ in.} \times 3 \text{ in.} \times 0.5 \text{ in}$$

$$\text{volume of stack} = 9\ \text{in.}^3$$

(2) Calculate the number of stacks. Note that a trillion dollars is equal to $\$1 \times 10^{12}$, and a stack of one-hundred $100 bills is equal to $10,000, or $\$1 \times 10^4$. The number of stacks you will have is:

$$\$1 \times 10^{12}\ \text{(a trillion dollars)} / \$1 \times 10^4\ \text{per stack} = 1 \times 10^8\ \text{stacks.}$$

(3) Calculate the area of a football field in square inches. The area of a football field is 100 yd×50 yd, which gives 5,000 yd^2. Because we are working in inches, we need to convert square yards to square inches:

$$\begin{aligned} \text{Area} &= 5{,}000\ \text{yd}^2 \times \tfrac{3\text{ft}}{1\ \text{yd}} \times \tfrac{3\text{ft}}{1\ \text{yd}} \times \tfrac{12\text{in.}}{1\ \text{ft}} \times \tfrac{12\text{in.}}{1\ \text{ft}} = 6{,}480{,}000\ \text{in.}^2, \\ \text{Area} &\approx 6 \times 10^6\ \text{in.}^2. \end{aligned}$$

This conversion gives us 6×10^6 in^2 for the area of the field. (Note that we are using only one significant figure in these calculations.)

(4) Calculate the total volume of the bills. The volume of all the $100-bill stacks is

$$9\text{in.}^3/\text{stack} \times 10^8\ \text{stacks} = 9 \times 10^8\text{in.}^3$$

(5) Calculate the height. To determine the height of the bills, use the equation:

$$\begin{aligned} \text{volume of bills} &= \text{area of field} \times \text{height of money:} \\ \text{Height of money} &= \tfrac{\text{volume of bills}}{\text{area of field}}, \\ \text{Height of money} &= \tfrac{9\times10^8\text{in.}^3}{6\times10^6\text{in.}^2} = 1.33 \times 10^2\text{in.}, \\ \text{Height of money} &\approx 1 \times 10^2\text{in.} = 100\ \text{in.} \end{aligned}$$

The height of the money will be about 100 in. high. Converting this value to feet gives

$$100\ \text{in.} \times \tfrac{1\ \text{ft}}{12\ \text{in.}} = 8.33\ \text{ft} \approx 8\ \text{ft.}$$

Discussion

The final approximate value is much higher than the early estimate of 3 in., but the other early estimate of 10 ft (120 in.) was roughly correct. How did the approximation measure up to your first guess? What can this exercise tell you in terms of rough “guesstimates” versus carefully calculated approximations?

Check Your Understanding

Using mental math and your understanding of fundamental units, approximate the area of a regulation basketball court. Describe the process you used to arrive at your final approximation.

Solution

An average male is about two meters tall. It would take approximately 15 men laid out end to end to cover the length, and about 7 to cover the width. That gives an approximate area of 420 m^2.

Section Summary

Scientists often approximate the values of quantities to perform calculations and analyze systems.

Problems & Exercises

1. How many heartbeats are there in a lifetime?

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

3. How many times longer than the mean life of an extremely unstable atomic nucleus is the lifetime of a human? (Hint: The lifetime of an unstable atomic nucleus is on the order of 10^{-22}.)

4. Calculate the approximate number of atoms in a bacterium. Assume that the average mass of an atom in the bacterium is ten times the mass of a hydrogen atom. (Hint: The mass of a hydrogen atom is on the order of 10^{-27} and the mass of a bacterium is on the order of 10^{-15}).

Figure 2. This color-enhanced photo shows Salmonella typhimurium (red) attacking human cells. These bacteria are commonly known for causing foodborn illness. Can you estimate the number of atoms in each bacterium? (credit: Rocky Mountain Laboratories, NIAID, NIH)

6. (a) What fraction of Earth's diameter is the greatest ocean depth? (b) The greatest mountain height?

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

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

Glossary

approximation: an estimated value based on prior experience and reasoning

Selected Answers to Problems & Exercises

1. 2×10^9 heartbeats

3. 2×10^{31} if an average human lifetime is taken to be about 70 years.

5. 50 atoms

7. 10^{12} cells/hummingbird (b) 10^{16} cells/human

2. Fluid Statics

Introduction to Fluid Statics

Lumen Learning

Figure 1. The fluid essential to all life has a beauty of its own. It also helps support the weight of this swimmer. (credit: Terren, Wikimedia Commons)

Much of what we value in life is fluid: a breath of fresh winter air; the hot blue flame in our gas cooker; the water we drink, swim in, and bathe in; the blood in our veins. What exactly is a fluid? Can we understand fluids with the laws already presented, or will new laws emerge from their study? The physical characteristics of static or stationary fluids and some of the laws that govern their behavior are the topics of this chapter. Fluid Dynamics and Its Biological and Medical Applications explores aspects of fluid flow.

What Is a Fluid?

Lumen Learning

Learning Objectives

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

- State the common phases of matter.
- Explain the physical characteristics of solids, liquids, and gases.
- Describe the arrangement of atoms in solids, liquids, and gases.

Matter most commonly exists as a solid, liquid, or gas; these states are known as the three common *phases of matter*. Solids have a definite shape and a specific volume, liquids have a definite volume but their shape changes depending on the container in which they are held, and gases have neither a definite shape nor a specific volume as their molecules move to fill the container in which they are held. (See Figure 1.) Liquids and gases are considered to be fluids because they yield to shearing forces, whereas solids resist them. Note that the extent to which fluids yield to shearing forces (and hence flow easily and quickly) depends on a quantity called the viscosity which is discussed in detail in Viscosity and Laminar Flow; Poiseuille's Law. We can understand the phases of matter and what constitutes a fluid by considering the forces between atoms that make up matter in the three phases.

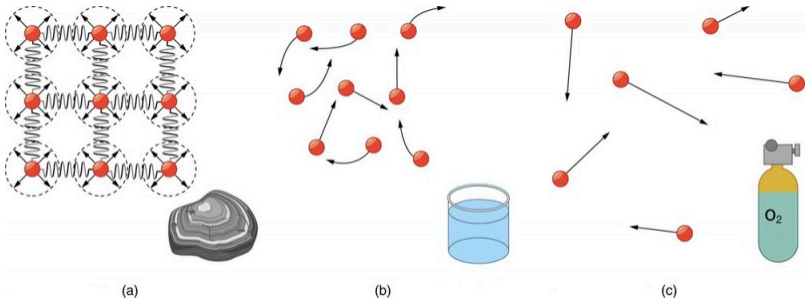


Figure 1. (a) Atoms in a solid always have the same neighbors, held near home by forces represented here by springs. These atoms are essentially in contact with one another. A rock is an example of a solid. This rock retains its shape because of the forces holding its atoms together. (b) Atoms in a liquid are also in close contact but can slide over one another. Forces between them strongly resist attempts to push them closer together and also hold them in close contact. Water is an example of a liquid. Water can flow, but it also remains in an open container because of the forces between its atoms. (c) Atoms in a gas are separated by distances that are considerably larger than the size of the atoms themselves, and they move about freely. A gas must be held in a closed container to prevent it from moving out freely.

Atoms in *solids* are in close contact, with forces between them that allow the atoms to vibrate but not to change positions with neighboring atoms. (These forces can be thought of as springs that can be stretched or compressed, but not easily broken.) Thus a solid *resists* all types of stress. A solid cannot be easily deformed because the atoms that make up the solid are not able to move about freely. Solids also resist compression, because their atoms form part of a lattice structure in which the atoms are a relatively fixed distance apart. Under compression, the atoms would be forced into one another. Most of the examples we have studied so far have involved solid objects which deform very little when stressed.

Connections: Submicroscopic Explanation of Solids and Liquids

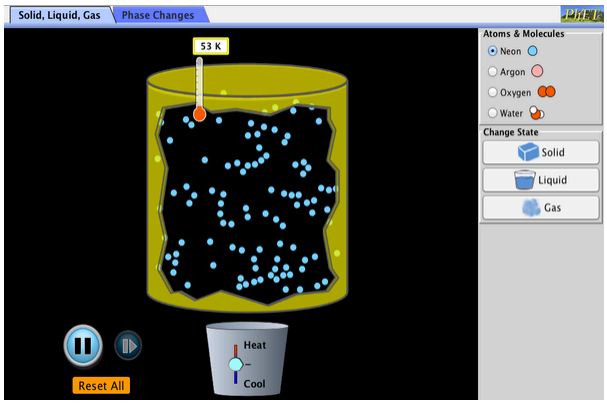
Atomic and molecular characteristics explain and underlie the macroscopic characteristics of solids and fluids. This submicroscopic explanation is one theme of this text and is highlighted in the Things Great and Small features in Conservation of Momentum. See, for example, microscopic description of collisions and momentum or microscopic description of pressure in a gas. This present section is devoted entirely to the submicroscopic explanation of solids and liquids.

In contrast, *liquids* deform easily when stressed and do not spring back to their original shape once the force is removed because the atoms are free to slide about and change neighbors—that is, they *flow* (so they are a type of fluid), with the molecules held together by their mutual attraction. When a liquid is placed in a container with no lid on, it remains in the container (providing the container has no holes below the surface of the liquid!). Because the atoms are closely packed, liquids, like solids, resist compression.

Atoms in *gases* are separated by distances that are large compared with the size of the atoms. The forces between gas atoms are therefore very weak, except when the atoms collide with one another. Gases thus not only flow (and are therefore considered to be fluids) but they are relatively easy to compress because there is much space and little force between atoms. When placed in an open container gases, unlike liquids, will escape. The major distinction is that gases are easily compressed, whereas liquids are not. We shall generally refer to both gases and liquids simply as *fluids*, and make a distinction between them only when they behave differently.

PhET Explorations: States of Matter—Basics

Heat, cool, and compress atoms and molecules and watch as they change between solid, liquid, and gas phases.



Click to download the simulation. Run using Java.

Section Summary

- A fluid is a state of matter that yields to sideways or shearing forces. Liquids and gases are both fluids. Fluid statics is the physics of stationary fluids.

Conceptual Questions

1. What physical characteristic distinguishes a fluid from a solid?
2. Which of the following substances are fluids at room temperature: air, mercury, water, glass?
3. Why are gases easier to compress than liquids and solids?
4. How do gases differ from liquids?

Glossary

fluids:
liquids and gases; a fluid is a state of matter that yields to shearing forces

Density

Lumen Learning

Learning Objectives

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

- Define density.
- Calculate the mass of a reservoir from its density.
- Compare and contrast the densities of various substances.

Which weighs more, a ton of feathers or a ton of bricks? This old riddle plays with the distinction between mass and density. A ton is a ton, of course; but bricks have much greater density than feathers, and so we are tempted to think of them as heavier. (See Figure 1.)

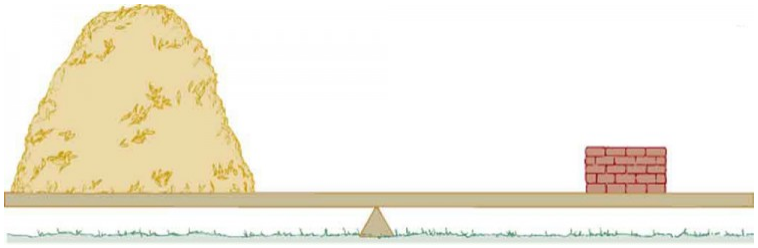


Figure 1. A ton of feathers and a ton of bricks have the same mass, but the feathers make a much bigger pile because they have a much lower density.

Density, as you will see, is an important characteristic of substances. It is crucial, for example, in determining whether an object sinks or floats in a fluid. Density is the mass per unit volume of a substance or object. In equation form, density is defined as

$$\rho = \frac{m}{V}$$

,

where the Greek letter ρ (rho) is the symbol for density, m is the mass, and V is the volume occupied by the substance.

Density

Density is mass per unit volume.

$$\rho = \frac{m}{V}$$

,

where ρ is the symbol for density, m is the mass, and V is the volume occupied by the substance.

In the riddle regarding the feathers and bricks, the masses are the same, but the volume occupied by the feathers is much greater, since their density is much lower. The SI unit of density is kg/m^3, representative values are given in Table 1. The metric system was originally devised so that water would have a density of 1 g/cm^3, equivalent to 10^3 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 cm^3.

Table 1. Densities of Various Substances

Substance	$\rho\left(10^3\mathrm{kg/m^3 org/mL}\right)$	Substance	$\rho\left(10^3\mathrm{kg/m^3 org/mL}\right)$	Substance	$\rho\left(10^3\mathrm{kg/m^3 org/m}$
Solids	**Liquids**	**Gases**			
Aluminum	2.7	Water (4ºC)	1.000	Air	1.29×10^{-3}
Brass	8.44	Blood	1.05	Carbon dioxide	1.98×10^{-3}
Copper (average)	8.8	Sea water	1.025	Carbon monoxide	1.25×10^{-3}
Gold	19.32	Mercury	13.6	Hydrogen	0.090×10^{-3}
Iron or steel	7.8	Ethyl alcohol	0.79	Helium	0.18×10^{-3}
Lead	11.3	Petrol	0.68	Methane	0.72×10^{-3}
Polystyrene	0.10	Glycerin	1.26	Nitrogen	1.25×10^{-3}
Tungsten	19.30	Olive oil	0.92	Nitrous oxide	1.98×10^{-3}
Uranium	18.70	Oxygen	1.43×10^{-3}		
Concrete	2.30–3.0	Steam (100º C)	0.60×10^{-3}		
Cork	0.24				
Glass, common (average)	2.6				
Granite	2.7				
Earth’s crust	3.3				
Wood	0.3–0.9				
Ice (0°C)	0.917				
Bone	1.7–2.0				

As you can see by examining Table 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.

Take-Home Experiment Sugar and Salt

A pile of sugar and a pile of salt look pretty similar, but which weighs more? If the volumes of both piles are the same, any difference in mass is due to their different densities (including the air space between crystals). Which do you think has the greater density? What values did you find? What method did you use to determine these values?

Example 1. Calculating the Mass of a Reservoir From Its Volume

A reservoir has a surface area of 50.0 km^2 and an average depth of 40.0 m. What mass of water is held behind the dam? (See Figure 2 for a view of a large reservoir—the Three Gorges Dam site on the Yangtze River in central China.)

Strategy

We can calculate the volume *V* of the reservoir from its dimensions, and find the density of water ρ in Table 1. Then the mass *m* can be found from the definition of density

$$\rho = \frac{m}{V}$$

.

Solution

Solving equation $\rho = m/V$ for *m* gives $m = \rho V$. The volume *V* of the reservoir is its surface area *A* times its average depth *h*:

$$\begin{aligned} V &= Ah = \left(50.0\text{km}^2\right)\left(40.0\text{m}\right) \\ &= \left[\left(50.0\ \text{km}^2\right)\left(\frac{10^3\text{m}}{1\text{km}}\right)^2\right]\left(40.0\ \text{m}\right) = 2.00 \times 10^9\text{m}^3 \end{aligned}$$

The density of water ρ from Table 1 is 1.000×10^3. Substituting *V* and ρ into the expression for mass gives

$$\begin{aligned} m &= \left(1.00 \times 10^3\ \text{kg/m}^3\right)\left(2.00 \times 10^9\text{m}^3\right) \\ &= 2.00 \times 10^{12}\ \text{kg} \end{aligned}$$

.

Discussion

A large reservoir contains a very large mass of water. In this example, the weight of the water in the reservoir is $mg = 1.96 \times 10^{13}$ N, where *g* is the acceleration due to the Earth's gravity (about 9.80 m/s^2). It is reasonable to ask whether the dam must supply a force equal to this tremendous weight. The answer is no. As we shall see in the following sections, the force the dam must supply can be much smaller than the weight of the water it holds back.

Figure 2. Three Gorges Dam in central China. When completed in 2008, this became the world's largest hydroelectric plant, generating power equivalent to that generated by 22 average-sized nuclear power plants. The concrete dam is 181 m high and 2.3 km across. The reservoir made by this dam is 660 km long. Over 1 million people were displaced by the creation of the reservoir. (credit: Le Grand Portage)

Section Summary

- Density is the mass per unit volume of a substance or object. In equation form, density is defined as

$$\rho = \frac{m}{V}$$

.

- The SI unit of density is kg/m^3.

Conceptual Questions

1. Approximately how does the density of air vary with altitude?

2. Give an example in which density is used to identify the substance composing an object. Would information in addition to average density be needed to identify the substances in an object composed of more than one material?

3. Figure 3 shows a glass of ice water filled to the brim. Will the water overflow when the ice melts? Explain your answer.

Figure 3.

Problems & Exercises

1. Gold is sold by the troy ounce (31.103 g). What is the volume of 1 troy ounce of pure gold?

2. Mercury is commonly supplied in flasks containing 34.5 kg (about 76 lb). What is the volume in liters of this much mercury?

3. (a) What is the mass of a deep breath of air having a volume of 2.00 L? (b) Discuss the effect taking such a breath has on your body's volume and density.

4, 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 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.)

5. 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.

6. (a) 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.

7. A trash compactor can reduce the volume of its contents to 0.350 their original value. Neglecting the mass of air expelled, by what factor is the density of the rubbish increased?

8. 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?

9. What is the density of 18.0-karat gold that is a mixture of 18 parts gold, 5 parts silver, and 1 part copper? (These values are parts by mass, not volume.) Assume that this is a simple mixture having an average density equal to the weighted densities of its constituents.

10. There is relatively little empty space between atoms in solids and liquids, so that the average density of an atom is about the same as matter on a macroscopic scale—approximately 10^3 kg/m^3. The nucleus of an atom

has a radius about 10^{-5} that of the atom and contains nearly all the mass of the entire atom. (a) What is the approximate density of a nucleus? (b) One remnant of a supernova, called a neutron star, can have the density of a nucleus. What would be the radius of a neutron star with a mass 10 times that of our Sun (the radius of the Sun is 7×10^8)?

Glossary

density:
the mass per unit volume of a substance or object

Selected Solutions to Problems & Exercises

1. 1.610 cm^3

3. (a) 2.58 g (b) 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 before you took the deep breath.

4. 2.70 g/cm^3

6. (a) 0.163 m (b) Equivalent to 19.4 gallons, which is reasonable

8. 7.9×10^2 kg/m^3

9. 15.6 g/cm^3

10. (a) 10^{18} kg/m^3 (b) 2×10^4 m

Pressure

Lumen Learning

Learning Objectives

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

- Define pressure.
- Explain the relationship between pressure and force.
- Calculate force given pressure and area.

You have no doubt heard the word *pressure* being used in relation to blood (high or low blood pressure) and in relation to the weather (high- and low-pressure weather systems). These are only two of many examples of pressures in fluids. Pressure *P* is defined as

$$P = \frac{F}{A}$$

where *F* is a force applied to an area *A* that is perpendicular to the force.

Pressure

Pressure is defined as the force divided by the area perpendicular to the force over which the force is applied, or

$$P = \frac{F}{A}$$

A given force can have a significantly different effect depending on the area over which the force is exerted, as shown in Figure 1. The SI unit for pressure is the *pascal*, where

$$1 \text{ Pa} = 1\text{N}/\text{m}^2$$

.

In addition to the pascal, there are many other units for pressure that are in common use. In meteorology, atmospheric pressure is often described in units of millibar (mb), where

$$100 \text{ mb} = 1 \times 10^5 \text{ Pa}$$

.

Pounds per square inch (lb/in^2 or psi) is still sometimes used as a measure of tire pressure, and millimeters of mercury (mm Hg) is still often used in the measurement of blood pressure. Pressure is defined for all states of matter but is particularly important when discussing fluids.

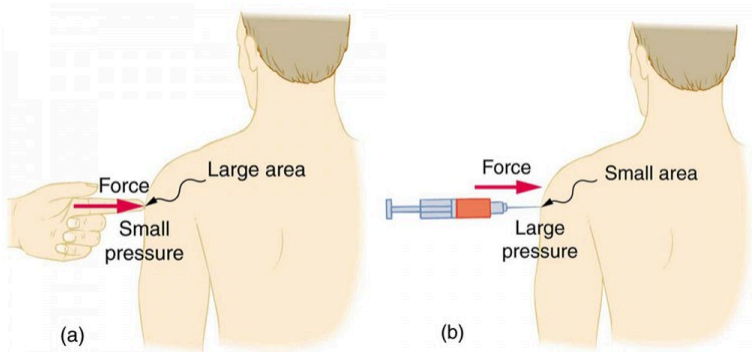


Figure 1. (a) While the person being poked with the finger might be irritated, the force has little lasting effect. (b) In contrast, the same force applied to an area the size of the sharp end of a needle is great enough to break the skin.

Example 1. What Force Does a Pressure Exert?

An astronaut is working outside the International Space Station where the atmospheric pressure is essentially zero. The pressure gauge on her air tank reads 6.90×10^6 Pa. What force does the air inside the tank exert on the flat end of the cylindrical tank, a disk 0.150 m in diameter?

Strategy

We can find the force exerted from the definition of pressure given in

$$P = \frac{F}{A}$$

, provided we can find the area *A* acted upon.

Solution

By rearranging the definition of pressure to solve for force, we see that

$$F = PA$$

Here, the pressure *P* is given, as is the area of the end of the cylinder *A*, given by $A = \pi r^2$. Thus,

$$\begin{aligned} F &= \left(6.90 \times 10^6 \text{N/m}^2\right)(3.14)(0.0750 \text{ m})^2 \\ &= 1.22 \times 10^5 \text{ N} \end{aligned}$$

.

Discussion

Wow! No wonder the tank must be strong. Since we found $F = PA$, we see that the force exerted by a pressure is directly proportional to the area acted upon as well as the pressure itself.

The force exerted on the end of the tank is perpendicular to its inside surface. This direction is because the force is exerted by a static or stationary fluid. We have already seen that fluids cannot *withstand* shearing (sideways) forces; they cannot *exert* shearing forces, either. Fluid pressure has no direction, being a scalar quantity. The forces due to pressure have well-defined directions: they are always exerted perpendicular to any surface. (See the tire in Figure 2, for example.) Finally, note that pressure is exerted on all surfaces. Swimmers, as well as the tire, feel pressure on all sides. (See Figure 3.)

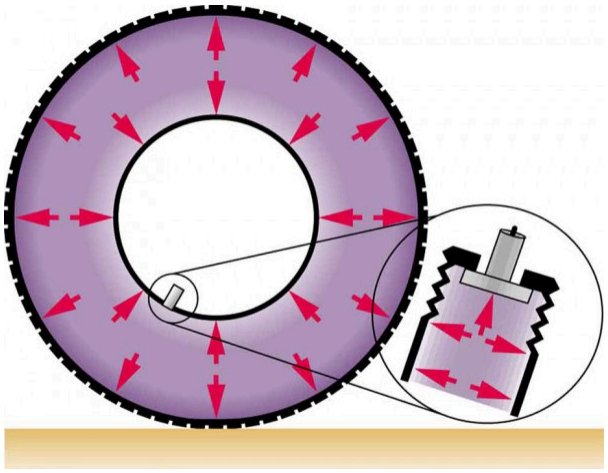


Figure 2. Pressure inside this tire exerts forces perpendicular to all surfaces it contacts. The arrows give representative directions and magnitudes of the forces exerted at various points. Note that static fluids do not exert shearing forces.

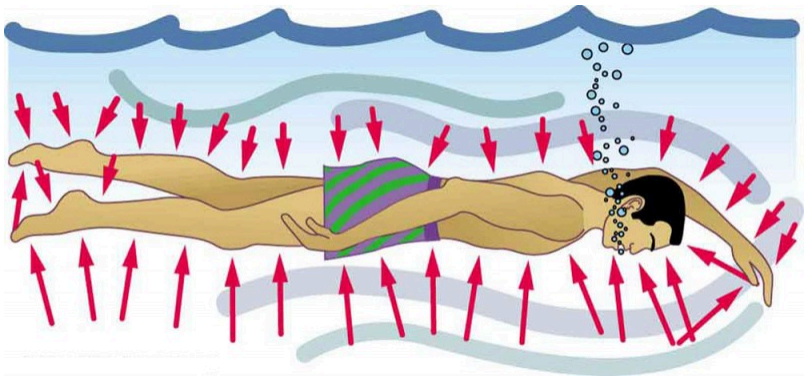
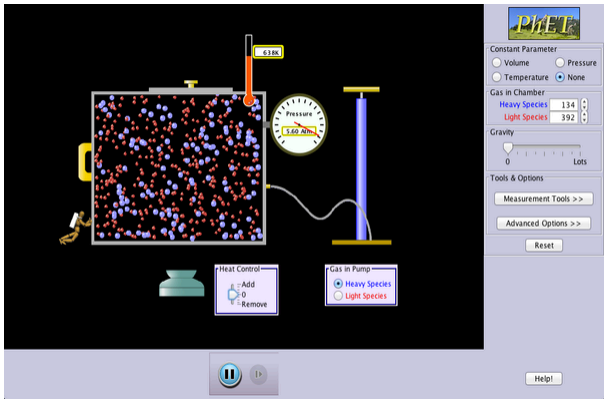


Figure 3. Pressure is exerted on all sides of this swimmer, since the water would flow into the space he occupies if he were not there. The arrows represent the directions and magnitudes of the forces exerted at various points on the swimmer. Note that the forces are larger underneath, due to greater depth, giving a net upward or buoyant force that is balanced by the weight of the swimmer.

PhET Explorations: Gas Properties

Pump gas molecules to a box and see what happens as you change the volume, add or remove heat, change gravity, and more. Measure the temperature and pressure, and discover how the properties of the gas vary in relation to each other.



Click to download the simulation. Run using Java.

Section Summary

- Pressure is the force per unit perpendicular area over which the force is applied. In equation form, pressure is defined as

$$P = \frac{F}{A}$$

.

- The SI unit of pressure is pascal and

$$1 \text{ Pa} = 1 \text{ N/m}^2$$

.

Conceptual Questions

1. How is pressure related to the sharpness of a knife and its ability to cut?

2. Why does a dull hypodermic needle hurt more than a sharp one?

3. The outward force on one end of an air tank was calculated in *Example 1: Calculating Force Exerted by the Air*. How is this force balanced? (The tank does not accelerate, so the force must be balanced.)

4. Why is force exerted by static fluids always perpendicular to a surface?

5. In a remote location near the North Pole, an iceberg floats in a lake. Next to the lake (assume it is not frozen) sits a comparably sized glacier sitting on land. If both chunks of ice should melt due to rising global

temperatures (and the melted ice all goes into the lake), which ice chunk would give the greatest increase in the level of the lake water, if any?

6. How do jogging on soft ground and wearing padded shoes reduce the pressures to which the feet and legs are subjected?

7. Toe dancing (as in ballet) is much harder on toes than normal dancing or walking. Explain in terms of pressure.

8. How do you convert pressure units like millimeters of mercury, centimeters of water, and inches of mercury into units like newtons per meter squared without resorting to a table of pressure conversion factors?

Problems & Exercises

1. As a woman walks, her entire weight is momentarily placed on one heel of her high-heeled shoes. Calculate the pressure exerted on the floor by the heel if it has an area of 1.50 cm^2 and the woman's mass is 55.0 kg. Express the pressure in Pa. (In the early days of commercial flight, women were not allowed to wear high-heeled shoes because aircraft floors were too thin to withstand such large pressures.)

2. The pressure exerted by a phonograph needle on a record is surprisingly large. If the equivalent of 1.00 g is supported by a needle, the tip of which is a circle 0.200 mm in radius, what pressure is exerted on the record in N/m^2?

3. Nail tips exert tremendous pressures when they are hit by hammers because they exert 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 3.00×10^9 N/m^2 (This high pressure is possible because the hammer striking the nail is brought to rest in such a short distance.)

Glossary

pressure:
the force per unit area perpendicular to the force, over which the force acts

Selected Solutions to Problems & Exercises

1. 3.59×10^6 Pa; or 521 lb/in^2

3. 2.36×10^3 N

Variation of Pressure with Depth in a Fluid

Lumen Learning

Learning Objectives

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

- Define pressure in terms of weight.
- Explain the variation of pressure with depth in a fluid.
- Calculate density given pressure and altitude.

If your ears have ever popped on a plane flight or ached during a deep dive in a swimming pool, you have experienced the effect of depth on pressure in a fluid. At the Earth's surface, the air pressure exerted on you is a result of the weight of air above you. This pressure is reduced as you climb up in altitude and the weight of air above you decreases. Under water, the pressure exerted on you increases with increasing depth. In this case, the pressure being exerted upon you is a result of both the weight of water above you *and* that of the atmosphere above you. You may notice an air pressure change on an elevator ride that transports you many stories, but you need only dive a meter or so below the surface of a pool to feel a pressure increase. The difference is that water is much denser than air, about 775 times as dense. Consider the container in Figure 1.

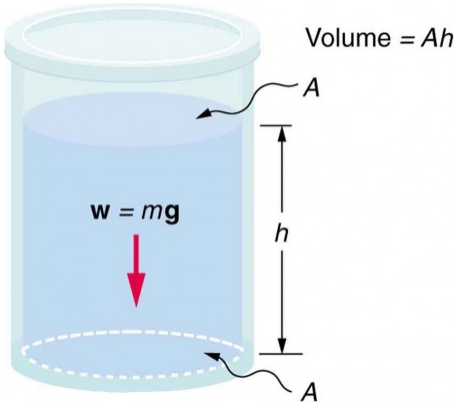


Figure 1. The bottom of this container supports the entire weight of the fluid in it. The vertical sides cannot exert an upward force on the fluid (since it cannot withstand a shearing force), and so the bottom must support it all.

Its bottom supports the weight of the fluid in it. Let us calculate the pressure exerted on the bottom by the weight of the fluid. That *pressure* is the weight of the fluid *mg* divided by the area *A* supporting it (the area of the bottom of the container):

$$P = \frac{mg}{A}$$

.

We can find the mass of the fluid from its volume and density:

$$m = \rho V.$$

The volume of the fluid V is related to the dimensions of the container. It is

$$V = Ah,$$

where A is the cross-sectional area and h is the depth. Combining the last two equations gives

$$m = \rho Ah$$

.

If we enter this into the expression for pressure, we obtain

$$P = \frac{(\rho Ah)g}{A}$$

.

The area cancels, and rearranging the variables yields

$$P = h\rho g.$$

This value is the *pressure due to the weight of a fluid*. The equation has general validity beyond the special conditions under which it is derived here. Even if the container were not there, the surrounding fluid would still exert this pressure, keeping the fluid static. Thus the equation $P = h\rho g$ represents the pressure due to the weight of any fluid of *average density* ρ at any depth h below its surface. For liquids, which are nearly incompressible, this equation holds to great depths. For gases, which are quite compressible, one can apply this equation as long as the density changes are small over the depth considered. *Example 2: Calculating Average Density: How Dense Is the Air?* illustrates this situation.

Example 1. Calculating the Average Pressure and Force Exerted: What Force Must a Dam Withstand?

In Example 1. Calculating the Mass of a Reservoir from Its Volume, we calculated the mass of water in a large reservoir. We will now consider the pressure and force acting on the dam retaining water. (See Figure 2.) The dam is 500 m wide, and the water is 80.0 m deep at the dam. (a) What is the average pressure on the dam due to the water? (b) Calculate the force exerted against the dam and compare it with the weight of water in the dam (previously found to be 1.96×10^{13} N).

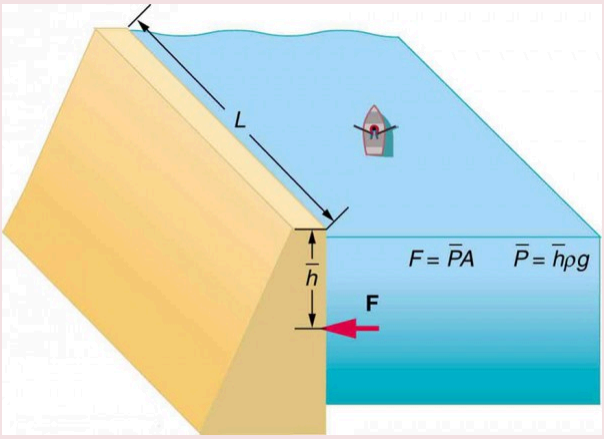


Figure 2. The dam must withstand the force exerted against it by the water it retains. This force is small compared with the weight of the water behind the dam.

Strategy for (a)

The average pressure

$$\overline{P}$$

due to the weight of the water is the pressure at the average depth

$$\overline{h}$$

of 40.0 m, since pressure increases linearly with depth.

Solution for (a)

The average pressure due to the weight of a fluid is

$$\overline{P} = \overline{h}rhog$$

.

Entering the density of water from Table 1 and taking

$$\overline{h}$$

to be the average depth of 40.0 m, we obtain

$$\begin{aligned} \overline{P} &= (40.0\ \text{m})\left(10^3 \tfrac{\text{kg}}{\text{m}^3}\right)\left(9.80 \tfrac{\text{m}}{\text{s}^2}\right) \\ &= 3.92 \times 10^5 \tfrac{\text{N}}{\text{m}^2} = 392\ \text{kPa}. \end{aligned}$$

Strategy for (b)

The force exerted on the dam by the water is the average pressure times the area of contact:

$$F = \overline{P}A$$

.

Solution for (b)

We have already found the value for

$$\overline{P}$$

. The area of the dam is A = 80.0 m x 500 m = 4.00 × 10^4 m^2, so that

$$\begin{aligned} F &= \left(3.92 \times 10^5 \mathrm{N/m^2}\right)\left(4.00 \times 10^4 \mathrm{m^2}\right) \\ &= 1.57 \times 10^{10} \mathrm{N}. \end{aligned}$$

Discussion

Although this force seems large, it is small compared with the 1.96 × 10^{13} N weight of the water in the reservoir—in fact, it is only 0.0800% of the weight. Note that the pressure found in part (a) is completely independent of the width and length of the lake—it depends only on its average depth at the dam. Thus the force depends only on the water's average depth and the dimensions of the dam, *not* on the horizontal extent of the reservoir. In the diagram, the thickness of the dam increases with depth to balance the increasing force due to the increasing pressure depth to balance the increasing force due to the increasing pressure.

Table 1. Densities of Various Substances

Substance	$\rho\left(10^3\mathrm{kg/m^3 org/mL}\right)$	Substance	$\rho\left(10^3\mathrm{kg/m^3 org/mL}\right)$	Substance	$\rho\left(10^3\mathrm{kg/m^3 org/m}$
Solids	**Liquids**	**Gases**			
Aluminum	2.7	Water (4ºC)	1.000	Air	1.29×10^{-3}
Brass	8.44	Blood	1.05	Carbon dioxide	1.98×10^{-3}
Copper (average)	8.8	Sea water	1.025	Carbon monoxide	1.25×10^{-3}
Gold	19.32	Mercury	13.6	Hydrogen	0.090×10^{-3}
Iron or steel	7.8	Ethyl alcohol	0.79	Helium	0.18×10^{-3}
Lead	11.3	Petrol	0.68	Methane	0.72×10^{-3}
Polystyrene	0.10	Glycerin	1.26	Nitrogen	1.25×10^{-3}
Tungsten	19.30	Olive oil	0.92	Nitrous oxide	1.98×10^{-3}
Uranium	18.70	Oxygen	1.43×10^{-3}		
Concrete	2.30–3.0	Steam (100º C)	0.60×10^{-3}		
Cork	0.24				
Glass, common (average)	2.6				
Granite	2.7				
Earth's crust	3.3				
Wood	0.3–0.9				
Ice (0°C)	0.917				
Bone	1.7–2.0				

Atmospheric pressure is another example of pressure due to the weight of a fluid, in this case due to the weight of *air* above a given height. The atmospheric pressure at the Earth's surface varies a little due to the large-scale flow of the atmosphere induced by the Earth's rotation (this creates weather "highs" and "lows"). However, the average pressure at sea level is given by the *standard atmospheric pressure* P_{atm}, measured to be

$$1\ \mathrm{atmosphere\ (atm)} = P_{\mathrm{atm}} = 1.01 \times 10^5 \mathrm{N/m^2} = 101\ \mathrm{kPa}$$

.

This relationship means that, on average, at sea level, a column of air above 1.00 m^2 of the Earth's surface has a weight of 1.01×10^5 N, equivalent to 1 atm. (See Figure 3.)

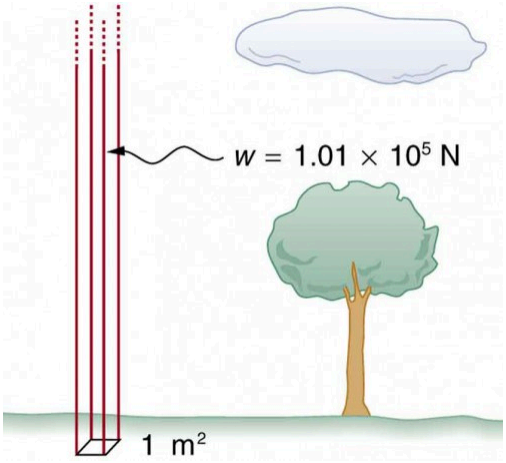


Figure 3. Atmospheric pressure at sea level averages 1.01×10^5 Pa (equivalent to 1 atm), since the column of air over this 1 m^2, extending to the top of the atmosphere, weighs 1.01×10^5.

Example 2. Calculating Average Density: How Dense Is the Air?

Calculate the average density of the atmosphere, given that it extends to an altitude of 120 km. Compare this density with that of air listed in Table 1.

Strategy

If we solve $P = h\rho g$ for density, we see that

$$\overline{\rho} = \frac{P}{\text{hg}}$$

.

We then take P to be atmospheric pressure, h is given, and g is known, and so we can use this to calculate

$$\overline{\rho}$$

.

Solution

Entering known values into the expression for

$$\overline{\rho}$$

yields

$$\overline{\rho} = \frac{1.01\times10^{5}\mathrm{N/m^{2}}}{\left(120\times10^{3}\mathrm{m}\right)\left(9.80\mathrm{m/s^{2}}\right)} = 8.59\times10^{-2}\mathrm{kg/m^{3}}$$

.

Discussion

This result is the average density of air between the Earth's surface and the top of the Earth's atmosphere, which essentially ends at 120 km. The density of air at sea level is given in Table 1 as 1.29 kg.m^3—about 15 times its average value. Because air is so compressible, its density has its highest value near the Earth's surface and declines rapidly with altitude.

Example 3. Calculating Depth Below the Surface of Water: What Depth of Water Creates the Same Pressure as the Entire Atmosphere?

Calculate the depth below the surface of water at which the pressure due to the weight of the water equals 1.00 atm.

Strategy

We begin by solving the equation $P = h\rho g$ for depth h:

$$h = \frac{P}{\rho g}$$

.

Then we take P to be 1.00 atm and ρ to be the density of the water that creates the pressure.

Solution

Entering the known values into the expression for h gives

$$h = \frac{1.01\times10^{5}\mathrm{N/m^{2}}}{\left(1.00\times10^{3}\mathrm{kg/m^{3}}\right)\left(9.80\mathrm{m/s^{2}}\right)} = 10.3\mathrm{m}$$

.

Discussion

Just 10.3 m of water creates the same pressure as 120 km of air. Since water is nearly incompressible, we can neglect any change in its density over this depth.

What do you suppose is the *total* pressure at a depth of 10.3 m in a swimming pool? Does the atmospheric pressure on the water's surface affect the pressure below? The answer is yes. This seems only logical, since both the water's weight and the atmosphere's weight must be supported. So the *total* pressure at a depth of 10.3 m is 2 atm—half from the water above and half from the air above. We shall see in Pascal's Principle that fluid pressures always add in this way.

Section Summary

- Pressure is the weight of the fluid *mg* divided by the area *A* supporting it (the area of the bottom of the container):

$$P = \frac{\mathrm{mg}}{A}$$

.

- Pressure due to the weight of a liquid is given by

$$P = h\rho g$$

,

where *P* is the pressure, *h* is the height of the liquid, *ρ* is the density of the liquid, and *g* is the acceleration due to gravity.

Conceptual Questions

1. 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?

2. Why does atmospheric pressure decrease more rapidly than linearly with altitude?

3. What are two reasons why mercury rather than water is used in barometers?

4. Figure 4 shows how sandbags placed around a leak outside a river levee can effectively stop the flow of water under the levee. Explain how the small amount of water inside the column formed by the sandbags is able to balance the much larger body of water behind the levee.

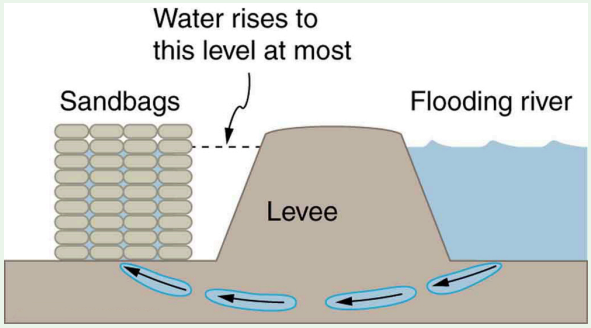


Figure 4. Because the river level is very high, it has started to leak under the levee. Sandbags are placed around the leak, and the water held by them rises until it is the same level as the river, at which point the water there stops rising.

5. Why is it difficult to swim under water in the Great Salt Lake?

6. Is there a net force on a dam due to atmospheric pressure? Explain your answer.

7. Does atmospheric pressure add to the gas pressure in a rigid tank? In a toy balloon? When, in general, does atmospheric pressure *not* affect the total pressure in a fluid?

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, and why it will not if there is air between the cork and liquid.

Problems & Exercises

1. What depth of mercury creates a pressure of 1.00 atm?

2. The greatest ocean depths on the Earth are found in the Marianas Trench near the Philippines. Calculate the pressure due to the ocean at the bottom of this trench, given its depth is 11.0 km and assuming the density of seawater is constant all the way down.

3. Verify that the SI unit of $h\rho g$ is N/m^2.

4. Water towers store water above the level of consumers for times of heavy use, eliminating the need for high-speed pumps. How high above a user must the water level be to create a gauge pressure of 3.00×10^5 N/m^2?

5. The aqueous humor in a person's eye is exerting a force of 0.300 N on the 1.10-cm^2 area of the cornea. (a) What pressure is this in mm Hg? (b) Is this value within the normal range for pressures in the eye?

6. How much force is exerted on one side of an 8.50 cm by 11.0 cm sheet of paper by the atmosphere? How can the paper withstand such a force?

7. What pressure is exerted on the bottom of a 0.500-m-wide by 0.900-m-long gas tank that can hold 50.0 kg of gasoline by the weight of the gasoline in it when it is full?

8. Calculate the average pressure exerted on the palm of a shot-putter's hand by the shot if the area of contact is 50.0 cm^2 and he exerts a force of 800 N on it. Express the pressure in N/m^2 and compare it with the 1.00×10^6 pressures sometimes encountered in the skeletal system.

9. The left side of the heart creates a pressure of 120 mm Hg by exerting a force directly on the blood over an effective area of 15.0 cm^2. What force does it exert to accomplish this?

10. Show that the total force on a rectangular dam due to the water behind it increases with the *square* of the water depth. In particular, show that this force is given by

$$F = \rho \mathrm{g} \mathrm{h}^2 L/2$$

, where

$$rho$$

is the density of water, h is its depth at the dam, and L is the length of the dam. You may assume the face of the dam is vertical. (Hint: Calculate the average pressure exerted and multiply this by the area in contact with the water. (See Figure 5.)

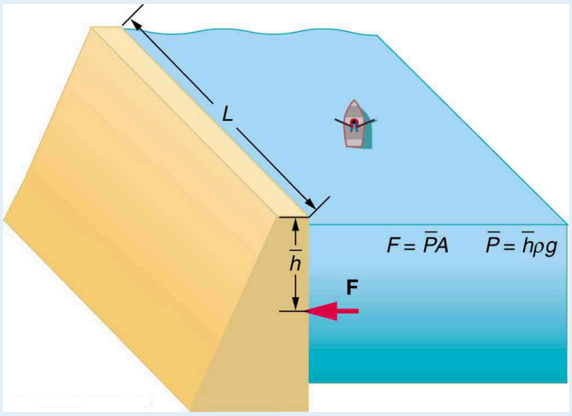


Figure 5.

Glossary

pressure:
the weight of the fluid divided by the area supporting it

Selected Solutions to Problems & Exercises

1. 0.760 m

3.

$$\begin{aligned}(h\rho g)_{\text{units}} &= (\text{m})\left(\text{kg/m}^3\right)\left(\text{m/s}^2\right) = \left(\text{kg}\cdot\text{m}^2\right)/\left(\text{m}^3\cdot\text{s}^2\right) \\ &= \left(\text{kg}\cdot\text{m/s}^2\right)\left(1/\text{m}^2\right) \\ &= \text{N/m}^2\end{aligned}$$

.

5. (a) 20.5 mm Hg (b) The range of pressures in the eye is 12–24 mm Hg, so the result in part (a) is within that range

7. 1.09×10^3 N/m^2

9. 24.0 N

Pascal's Principle

Lumen Learning

Learning Objectives

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

- Define pressure.
- State Pascal's principle.
- Understand applications of Pascal's principle.
- Derive relationships between forces in a hydraulic system.

Pressure is defined as force per unit area. Can pressure be increased in a fluid by pushing directly on the fluid? Yes, but it is much easier if the fluid is enclosed. The heart, for example, increases blood pressure by pushing directly on the blood in an enclosed system (valves closed in a chamber). If you try to push on a fluid in an open system, such as a river, the fluid flows away. An enclosed fluid cannot flow away, and so pressure is more easily increased by an applied force. What happens to a pressure in an enclosed fluid? Since atoms in a fluid are free to move about, they transmit the pressure to all parts of the fluid and to the walls of the container. Remarkably, the pressure is transmitted *undiminished*. This phenomenon is called *Pascal's principle*, because it was first clearly stated by the French philosopher and scientist Blaise Pascal (1623–1662): A change in pressure applied to an enclosed fluid is transmitted undiminished to all portions of the fluid and to the walls of its container.

Pascal's Principle

A change in pressure applied to an enclosed fluid is transmitted undiminished to all portions of the fluid and to the walls of its container.

Pascal's principle, an experimentally verified fact, is what makes pressure so important in fluids. Since a change in pressure is transmitted undiminished in an enclosed fluid, we often know more about pressure than other physical quantities in fluids. Moreover, Pascal's principle implies that *the total pressure in a fluid is the sum of the pressures from different sources*. We shall find this fact—that pressures add—very useful.

Blaise Pascal had an interesting life in that he was home-schooled by his father who removed all of the mathematics textbooks from his house and forbade him to study mathematics until the age of 15.

This, of course, raised the boy's curiosity, and by the age of 12, he started to teach himself geometry. Despite this early deprivation, Pascal went on to make major contributions in the mathematical fields of probability theory, number theory, and geometry. He is also well known for being the inventor of the first mechanical digital calculator, in addition to his contributions in the field of fluid statics.

Application of Pascal's Principle

One of the most important technological applications of Pascal's principle is found in a *hydraulic system*, which is an enclosed fluid system used to exert forces. The most common hydraulic systems are those that operate car brakes. Let us first consider the simple hydraulic system shown in Figure 1.

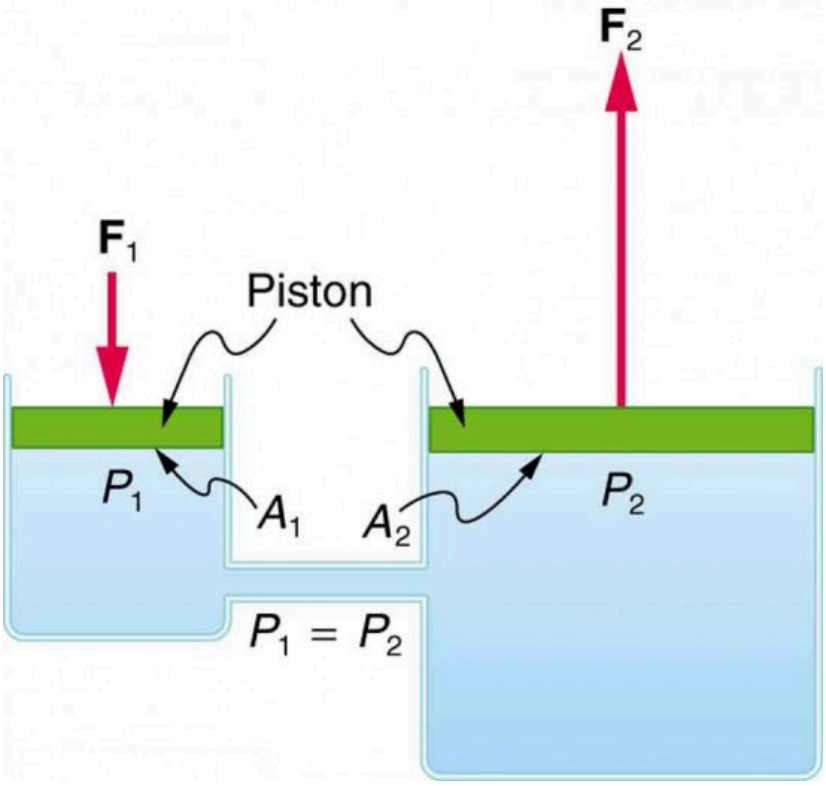


Figure 1. A typical hydraulic system with two fluid-filled cylinders, capped with pistons and connected by a tube called a hydraulic line. A downward force $\boldsymbol{F}_1$ *on the left piston creates a pressure that is transmitted undiminished to all parts of the enclosed fluid. This results in an upward force* $\boldsymbol{F}_2$ *on the right piston that is larger than* $\boldsymbol{F}_1$ *because the right piston has a larger area.*

Relationship Between Forces in a Hydraulic System

We can derive a relationship between the forces in the simple hydraulic system shown in Figure 1 by applying Pascal's principle. Note first that the two pistons in the system are at the same height, and so there will be no difference in pressure due to a difference in depth. Now the pressure due to F_1 acting on area A_1 is simply

$$P_1 = \frac{F_1}{A_1}$$

, as defined by

$$P = \frac{F}{A}$$

. According to Pascal's principle, this pressure is transmitted undiminished throughout the fluid and to all walls of the container. Thus, a pressure P_2 is felt at the other piston that is equal to P_1. That is

$$P_2 = \frac{F_2}{A_2} \qquad \frac{F_1}{A_1} = \frac{F_2}{A_2}$$

$P_1 = P_2$. But since , we see that . This equation relates the ratios of force to area in any hydraulic system, providing the pistons are at the same vertical height and that friction in the system is negligible. Hydraulic systems can increase or decrease the force applied to them. To make the force larger, the pressure is applied to a larger area. For example, if a 100-N force is applied to the left cylinder in Figure 1 and the right one has an area five times greater, then the force out is 500 N. Hydraulic systems are analogous to simple levers, but they have the advantage that pressure can be sent through tortuously curved lines to several places at once.

Example 1. Calculating Force of Slave Cylinders: Pascal Puts on the Brakes

Consider the automobile hydraulic system shown in Figure 2.

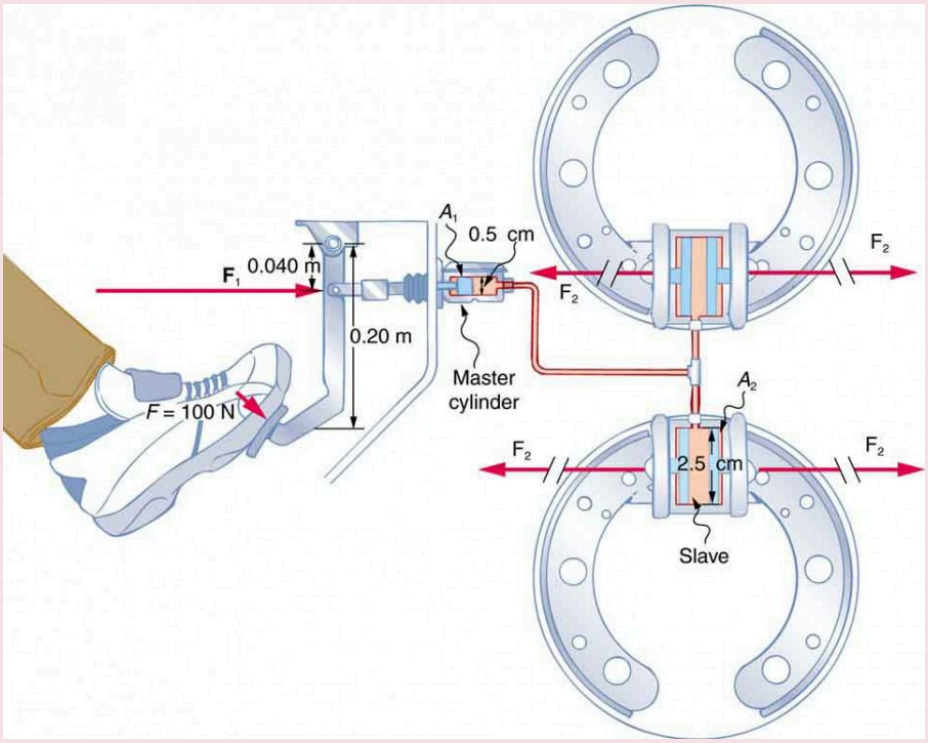


Figure 2. Hydraulic brakes use Pascal's principle. The driver exerts a force of 100 N on the brake pedal. This force is increased by the simple lever and again by the hydraulic system. Each of the identical slave cylinders receives the same pressure and, therefore, creates the same force output F_2. The circular cross-sectional areas of the master and slave cylinders are represented by A_1 and A_2, respectively

A force of 100 N is applied to the brake pedal, which acts on the cylinder—called the master—through a lever. A force of 500 N is exerted on the master cylinder. (The reader can verify that the force is 500 N using

techniques of statics from Applications of Statics, Including Problem-Solving Strategies.) Pressure created in the master cylinder is transmitted to four so-called slave cylinders. The master cylinder has a diameter of 0.500 cm, and each slave cylinder has a diameter of 2.50 cm. Calculate the force F_2 created at each of the slave cylinders.

Strategy

We are given the force F_1 that is applied to the master cylinder. The cross-sectional areas A_1 and A_2 can be calculated from their given diameters. Then

$$\frac{F_1}{A_1} = \frac{F_2}{A_2}$$

can be used to find the force F_2. Manipulate this algebraically to get F_2 on one side and substitute known values:

Solution

Pascal's principle applied to hydraulic systems is given by

$$\frac{F_1}{A_1} = \frac{F_2}{A_2}$$

:

$$F_2 = \frac{A_2}{A_1}F_1 = \frac{\pi {r_2}^2}{\pi {r_1}^2}F_1 = \frac{(1.25\text{ cm})^2}{(0.250\text{ cm})^2} \times 500\text{ N} = 1.25 \times 10^4\text{N}$$

.

Discussion

This value is the force exerted by each of the four slave cylinders. Note that we can add as many slave cylinders as we wish. If each has a 2.50-cm diameter, each will exert 1.25×10^4 N.

A simple hydraulic system, such as a simple machine, can increase force but cannot do more work than done on it. Work is force times distance moved, and the slave cylinder moves through a smaller distance than the master cylinder. Furthermore, the more slaves added, the smaller the distance each moves. Many hydraulic systems—such as power brakes and those in bulldozers—have a motorized pump that actually does most of the work in the system. The movement of the legs of a spider is achieved partly by hydraulics. Using hydraulics, a jumping spider can create a force that makes it capable of jumping 25 times its length!

Making Connections: Conservation of Energy

Conservation of energy applied to a hydraulic system tells us that the system cannot do more work than is done on it. Work transfers energy, and so the work output cannot exceed the work input. Power brakes and other similar hydraulic systems use pumps to supply extra energy when needed.

Section Summary

- Pressure is force per unit area.
- A change in pressure applied to an enclosed fluid is transmitted undiminished to all portions of the fluid and to the walls of its container.
- A hydraulic system is an enclosed fluid system used to exert forces.

Conceptual Questions

1. Suppose the master cylinder in a hydraulic system is at a greater height than the slave cylinder. Explain how this will affect the force produced at the slave cylinder.

Problems & Exercises

1. How much pressure is transmitted in the hydraulic system considered in Example 1? Express your answer in pascals and in atmospheres.

2. What force must be exerted on the master cylinder of a hydraulic lift to support the weight of a 2000-kg car (a large car) resting on the slave cylinder? The master cylinder has a 2.00-cm diameter and the slave has a 24.0-cm diameter.

3. A crass host pours the remnants of several bottles of wine into a jug after a party. He then inserts a cork with a 2.00-cm diameter into the bottle, placing it in direct contact with the wine. He is amazed when he pounds the cork into place and the bottom of the jug (with a 14.0-cm diameter) breaks away. Calculate the extra force exerted against the bottom if he pounded the cork with a 120-N force.

4. A certain hydraulic system is designed to exert a force 100 times as large as the one put into it. (a) What must be the ratio of the area of the slave cylinder to the area of the master cylinder? (b) What must be the ratio of their diameters? (c) By what factor is the distance through which the output force moves reduced relative to the distance through which the input force moves? Assume no losses to friction.

(5. a) Verify that work input equals work output for a hydraulic system assuming no losses to friction. Do this by showing that the distance the output force moves is reduced by the same factor that the output force is increased. Assume the volume of the fluid is constant. (b) What effect would friction within the fluid and between components in the system have on the output force? How would this depend on whether or not the fluid is moving?

Glossary

Pascal's Principle:
a change in pressure applied to an enclosed fluid is transmitted undiminished to all portions of the fluid and to the walls of its container

Selected Solutions to Problems & Exercises

1. 2.55×10^7 Pa; or 251 atm

3. 5.76×10^3 extra force

5. (a)

$$V = d_{\mathrm{i}}A_{\mathrm{i}} = d_{\mathrm{o}}A_{\mathrm{o}} \Rightarrow d_{\mathrm{o}} = d_{\mathrm{i}}\left(\frac{A_{\mathrm{i}}}{A_{\mathrm{o}}}\right)$$

.

Now, using equation:

$$\frac{F_1}{A_1} = \frac{F_2}{A_2} \Rightarrow F_{\mathrm{o}} = F_{\mathrm{i}}\left(\frac{A_{\mathrm{o}}}{A_{\mathrm{i}}}\right)$$

.

Finally,

$$W_{\mathrm{o}} = F_{\mathrm{o}}d_{\mathrm{o}} = \left(\frac{F_{\mathrm{i}}A_{\mathrm{o}}}{A_{\mathrm{i}}}\right)\left(\frac{d_{\mathrm{i}}A_{\mathrm{i}}}{A_{\mathrm{o}}}\right) = F_{\mathrm{i}}d_{\mathrm{i}} = W_{\mathrm{i}}$$

.

In other words, the work output equals the work input.

(b) If the system is not moving, friction would not play a role. With friction, we know there are losses, so that

$$W_{\mathrm{out}} = W_{\mathrm{in}} - W_{\mathrm{f}}$$

; therefore, the work output is less than the work input. In other words, with friction, you need to push harder on the input piston than was calculated for the nonfriction case.

Gauge Pressure, Absolute Pressure, and Pressure Measurement

Lumen Learning

Learning Objectives

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

- Define gauge pressure and absolute pressure.
- Understand the working of aneroid and open-tube barometers.

If you limp into a gas station with a nearly flat tire, you will notice the tire gauge on the airline reads nearly zero when you begin to fill it. In fact, if there were a gaping hole in your tire, the gauge would read zero, even though atmospheric pressure exists in the tire. Why does the gauge read zero? There is no mystery here. Tire gauges are simply designed to read zero at atmospheric pressure and positive when pressure is greater than atmospheric.

Similarly, atmospheric pressure adds to blood pressure in every part of the circulatory system. (As noted in Pascal's Principle, the total pressure in a fluid is the sum of the pressures from different sources—here, the heart and the atmosphere.) But atmospheric pressure has no net effect on blood flow since it adds to the pressure coming out of the heart and going back into it, too. What is important is how much *greater* blood pressure is than atmospheric pressure. Blood pressure measurements, like tire pressures, are thus made relative to atmospheric pressure.

In brief, it is very common for pressure gauges to ignore atmospheric pressure—that is, to read zero at atmospheric pressure. We therefore define *gauge pressure* to be the pressure relative to atmospheric pressure. Gauge pressure is positive for pressures above atmospheric pressure, and negative for pressures below it.

Gauge Pressure

Gauge pressure is the pressure relative to atmospheric pressure. Gauge pressure is positive for pressures above atmospheric pressure, and negative for pressures below it.

In fact, atmospheric pressure does add to the pressure in any fluid not enclosed in a rigid container. This happens because of Pascal's principle. The total pressure, or *absolute pressure*, is thus the sum of gauge pressure and atmospheric pressure: $P_{abs} = P_g + P_{atm}$ where P_{abs} is absolute pressure, P_g is gauge pressure, and P_{atm} is atmospheric pressure. For example, if your tire gauge reads 34 psi (pounds per

square inch), then the absolute pressure is 34 psi plus 14.7 psi (P_{atm} in psi), or 48.7 psi (equivalent to 336 kPa).

Absolute Pressure

Absolute pressure is the sum of gauge pressure and atmospheric pressure.

For reasons we will explore later, in most cases the absolute pressure in fluids cannot be negative. Fluids push rather than pull, so the smallest absolute pressure is zero. (A negative absolute pressure is a pull.) Thus the smallest possible gauge pressure is $P_g = -P_{atm}$ (this makes P_{abs} zero). There is no theoretical limit to how large a gauge pressure can be.

There are a host of devices for measuring pressure, ranging from tire gauges to blood pressure cuffs. Pascal's principle is of major importance in these devices. The undiminished transmission of pressure through a fluid allows precise remote sensing of pressures. Remote sensing is often more convenient than putting a measuring device into a system, such as a person's artery. Figure 1 shows one of the many types of mechanical pressure gauges in use today. In all mechanical pressure gauges, pressure results in a force that is converted (or transduced) into some type of readout.

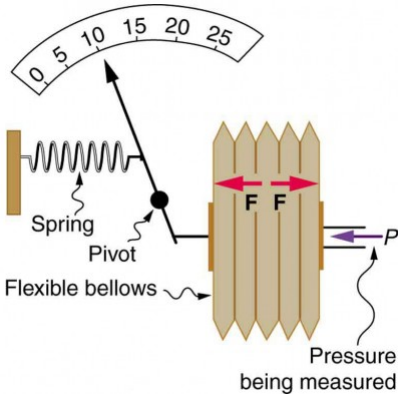


Figure 1. This aneroid gauge utilizes flexible bellows connected to a mechanical indicator to measure pressure.

An entire class of gauges uses the property that pressure due to the weight of a fluid is given by $P = h\rho g$. Consider the U-shaped tube shown in Figure 2, for example. This simple tube is called a *manometer*. In Figure 2(a), both sides of the tube are open to the atmosphere. Atmospheric pressure therefore pushes down on each side equally so its effect cancels. If the fluid is deeper on one side, there is a greater pressure on the deeper side, and the fluid flows away from that side until the depths are equal.

Let us examine how a manometer is used to measure pressure. Suppose one side of the U-tube is connected to some source of pressure P_{abs} such as the toy balloon in Figure 2(b) or the vacuum-packed peanut jar shown in Figure 2(c). Pressure is transmitted undiminished to the manometer, and the fluid levels are no longer equal. In Figure 2(b), P_{abs} is greater than atmospheric pressure, whereas in Figure 2(c), P_{abs} is less than atmospheric pressure. In both cases, P_{abs} differs from atmospheric pressure by an amount $h\rho g$, where ρ is the density of the fluid in the manometer. In Figure 2(b), P_{abs} can support a column of fluid of height h, and so it must exert a pressure $h\rho g$ greater than atmospheric pressure (the gauge pressure P_g is positive). In Figure 2(c), atmospheric pressure can support a column of fluid of height h, and so P_{abs} is less than atmospheric pressure by an amount $h\rho g$ (the gauge pressure P_g is negative). A manometer with one side open to the atmosphere is an ideal device for measuring gauge pressures. The gauge pressure is $P_g = h\rho g$ and is found by measuring h.

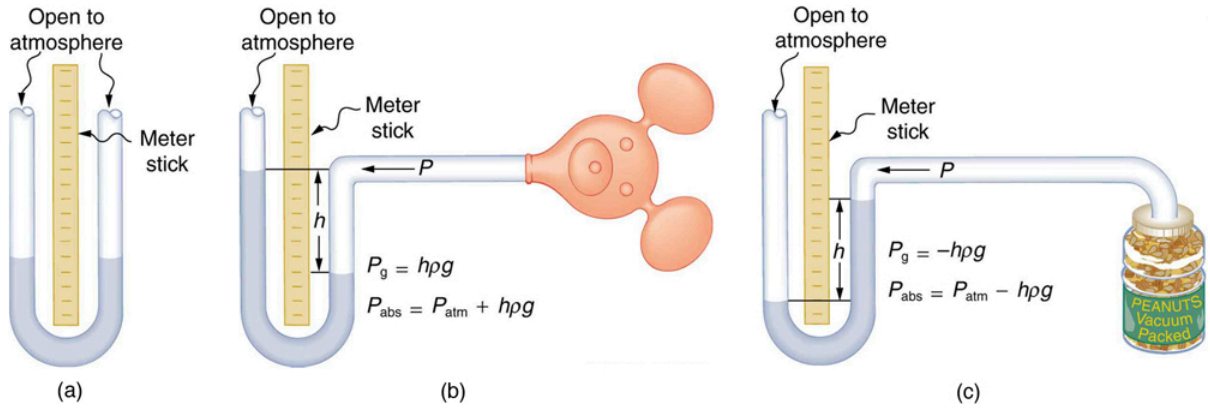


Figure 2. An open-tube manometer has one side open to the atmosphere. (a) Fluid depth must be the same on both sides, or the pressure each side exerts at the bottom will be unequal and there will be flow from the deeper side. (b) A positive gauge pressure $P_g = h\rho g$ transmitted to one side of the manometer can support a column of fluid of height h. (c) Similarly, atmospheric pressure is greater than a negative gauge pressure P_g by an amount $h\rho g$. The jar's rigidity prevents atmospheric pressure from being transmitted to the peanuts.

Mercury manometers are often used to measure arterial blood pressure. An inflatable cuff is placed on the upper arm as shown in Figure 3. By squeezing the bulb, the person making the measurement exerts pressure, which is transmitted undiminished to both the main artery in the arm and the manometer. When this applied pressure exceeds blood pressure, blood flow below the cuff is cut off. The person making the measurement then slowly lowers the applied pressure and listens for blood flow to resume. Blood pressure pulsates because of the pumping action of the heart, reaching a maximum, called *systolic pressure*, and a minimum, called *diastolic pressure*, with each heartbeat. Systolic pressure is measured by noting the value of *h* when blood flow first begins as cuff pressure is lowered. Diastolic pressure is measured by noting *h* when blood flows without interruption. The typical blood pressure of a young adult raises the mercury to a height of 120 mm at systolic and 80 mm at diastolic. This is commonly quoted as 120 over 80, or 120/80. The first pressure is representative of the maximum output of the heart; the second is due to the elasticity of the arteries in maintaining the pressure between beats. The density of the mercury fluid in the manometer is 13.6 times greater than water, so the height of the fluid will be 1/13.6 of that in a water manometer. This reduced height can make measurements difficult, so mercury manometers are used to measure larger pressures, such as blood pressure. The density of mercury is such that 1.0 mm Hg = 133 Pa.

Systolic Pressure

Systolic pressure is the maximum blood pressure.

Diastolic Pressure

Diastolic pressure is the minimum blood pressure.

Figure 3. In routine blood pressure measurements, an inflatable cuff is placed on the upper arm at the same level as the heart. Blood flow is detected just below the cuff, and corresponding pressures are transmitted to a mercury-filled manometer. (credit: U.S. Army photo by Spc. Micah E. Clare4TH BCT)

Example 1. Calculating Height of IV Bag: Blood Pressure and Intravenous Infusions

Intravenous infusions are usually made with the help of the gravitational force. Assuming that the density of the fluid being administered is 1.00 g/ml, at what height should the IV bag be placed above the entry point so that the fluid just enters the vein if the blood pressure in the vein is 18 mm Hg above atmospheric pressure? Assume that the IV bag is collapsible.

Strategy for (a)

For the fluid to just enter the vein, its pressure at entry must exceed the blood pressure in the vein (18 mm Hg above atmospheric pressure). We therefore need to find the height of fluid that corresponds to this gauge pressure.

Solution

We first need to convert the pressure into SI units. Since 1.0 mm Hg = 133 Pa,

$$P = 18 \text{ mm Hg} \times \frac{133 \text{ Pa}}{1.0 \text{ mm Hg}} = 2400 \text{ Pa}$$

Rearranging $P_g = h\rho g$ for h gives

$$h = \frac{P_g}{\rho g}$$

. Substituting known values into this equation gives

$$\begin{aligned} h &= \frac{2400 \text{ N/m}^2}{\left(1.0 \times 10^3 \text{kg/m}^3\right)\left(9.80 \text{m/s}^2\right)} \\ &= 0.24 \text{ m}. \end{aligned}$$

Discussion

The IV bag must be placed at 0.24 m above the entry point into the arm for the fluid to just enter the arm. Generally, IV bags are placed higher than this. You may have noticed that the bags used for blood collection are placed below the donor to allow blood to flow easily from the arm to the bag, which is the opposite direction of flow than required in the example presented here.

A *barometer* is a device that measures atmospheric pressure. A mercury barometer is shown in Figure 4. This device measures atmospheric pressure, rather than gauge pressure, because there is a nearly pure vacuum above the mercury in the tube. The height of the mercury is such that $h\rho g = P_{atm}$. When atmospheric pressure varies, the mercury rises or falls, giving important clues to weather forecasters. The barometer can also be used as an altimeter, since average atmospheric pressure varies with altitude. Mercury barometers and manometers are so common that units of mm Hg are often quoted for atmospheric pressure and blood pressures. Table 1 gives conversion factors for some of the more commonly used units of pressure.

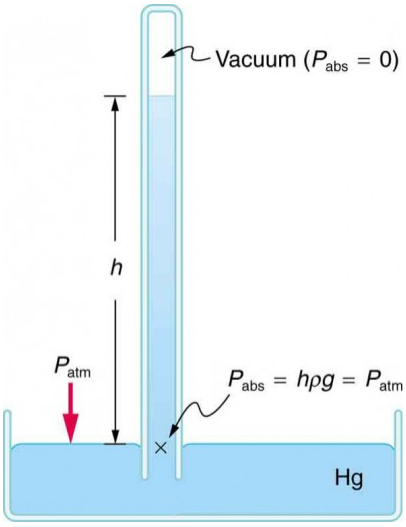


Figure 4. A mercury barometer measures atmospheric pressure. The pressure due to the mercury's weight, hρg, equals atmospheric pressure. The atmosphere is able to force mercury in the tube to a height h because the pressure above the mercury is zero.

Table 1. Conversion Factors for Various Pressure Units

Conversion to N/m^2 (Pa)	**Conversion from atm**
1.0 atm = 1.013×10^5 N/m^2	1.0 atm = 1.013×10^5 N/m^2
1.0 dyne/cm^2 = 0.10 N/m^2	1.0 atm = 1.013×10^6 dyne/cm^2
1.0 kg/cm^2 = 9.8×10^4 N/m^2	1.0 atm = 1.013 kg/cm^2
1.0 lb/$in.^2$ = 6.90×10^3 N/m^2	1.0 atm = 14.7 $lb/in.^2$
1.0 mm Hg = 133 N/m^2	1.0 atm = 760 mm Hg
1.0 cm Hg = 1.33×10^3 N/m^2	1.0 atm = 76.0 cm Hg
1.0 cm water = 98.1 N/m^2	1.0 atm = 1.03×10^3 cm water
1.0 bar = 1.000×10^5 N/m^2	1.0 atm = 1.013 bar
1.0 millibar = 1.000×10^2 N/m^2	1.0 atm = 1013 millibar

Section Summary

- Gauge pressure is the pressure relative to atmospheric pressure.
- Absolute pressure is the sum of gauge pressure and atmospheric pressure.

- Aneroid gauge measures pressure using a bellows-and-spring arrangement connected to the pointer of a calibrated scale.
- Open-tube manometers have U-shaped tubes and one end is always open. It is used to measure pressure.
- A mercury barometer is a device that measures atmospheric pressure.

Conceptual Questions

1. Explain why the fluid reaches equal levels on either side of a manometer if both sides are open to the atmosphere, even if the tubes are of different diameters.

2. Figure 3 shows how a common measurement of arterial blood pressure is made. Is there any effect on the measured pressure if the manometer is lowered? What is the effect of raising the arm above the shoulder? What is the effect of placing the cuff on the upper leg with the person standing? Explain your answers in terms of pressure created by the weight of a fluid.

3. Considering the magnitude of typical arterial blood pressures, why are mercury rather than water manometers used for these measurements?

Problems & Exercises

1. Find the gauge and absolute pressures in the balloon and peanut jar shown in Figure 2, assuming the manometer connected to the balloon uses water whereas the manometer connected to the jar contains mercury. Express in units of centimeters of water for the balloon and millimeters of mercury for the jar, taking h = 0.0500 m for each.

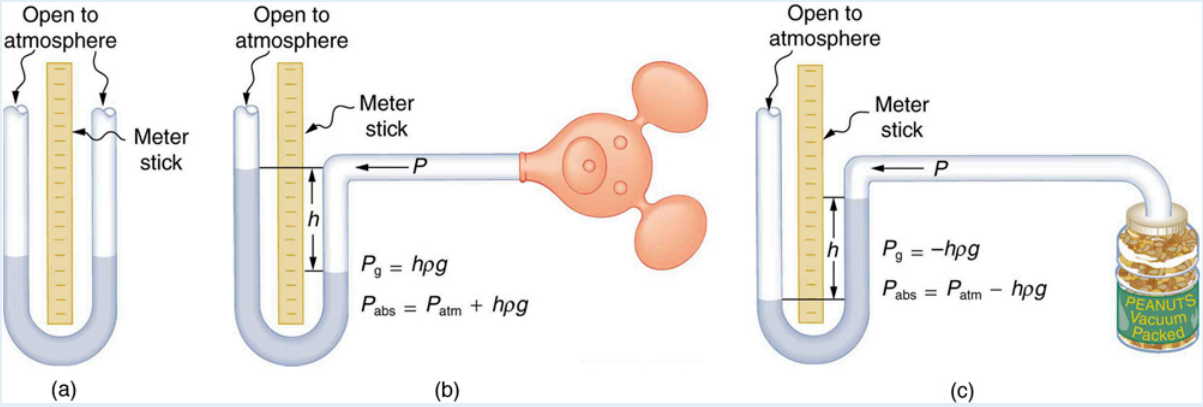


Figure 2. An open-tube manometer has one side open to the atmosphere. (a) Fluid depth must be the same on both sides, or the pressure each side exerts at the bottom will be unequal and there will be flow from the deeper side. (b) A positive gauge pressure $P_g = h\rho g$ transmitted to one side of the manometer can support a column of fluid of height h. (c) Similarly, atmospheric pressure is greater than a negative gauge pressure P_g by an amount $h\rho g$. The jar's rigidity prevents atmospheric pressure from being transmitted to the peanuts.

2. (a) Convert normal blood pressure readings of 120 over 80 mm Hg to newtons per meter squared using the relationship for pressure due to the weight of a fluid

$$(P = h\rho g)$$

rather than a conversion factor. (b) Discuss why blood pressures for an infant could be smaller than those for an adult. Specifically, consider the smaller height to which blood must be pumped.

3. How tall must a water-filled manometer be to measure blood pressures as high as 300 mm Hg?

4. Pressure cookers have been around for more than 300 years, although their use has strongly declined in recent years (early models had a nasty habit of exploding). How much force must the latches holding the lid onto a pressure cooker be able to withstand if the circular lid is 25.0 cm in diameter and the gauge pressure inside is 300 atm? Neglect the weight of the lid.

5. Suppose you measure a standing person's blood pressure by placing the cuff on his leg 0.500 m below the heart. Calculate the pressure you would observe (in units of mm Hg) if the pressure at the heart were 120 over 80 mm Hg. Assume that there is no loss of pressure due to resistance in the circulatory system (a reasonable assumption, since major arteries are large).

6. A submarine is stranded on the bottom of the ocean with its hatch 25.0 m below the surface. Calculate the force needed to open the hatch from the inside, given it is circular and 0.450 m in diameter. Air pressure inside the submarine is 1.00 atm.

7. Assuming bicycle tires are perfectly flexible and support the weight of bicycle and rider by pressure alone, calculate the total area of the tires in contact with the ground. The bicycle plus rider has a mass of 80.0 kg, and the gauge pressure in the tires is 3.50×10^5 Pa.

Glossary

absolute pressure:
the sum of gauge pressure and atmospheric pressure

diastolic pressure:
the minimum blood pressure in the artery

gauge pressure:
the pressure relative to atmospheric pressure

systolic pressure:
the maximum blood pressure in the artery

Selected Solutions to Problems & Exercises

1. Balloon:

P_g = 5.00 cm H_2O,

P_{abs} = 1.035×10^3 cm H_2O

Jar:

P_g = -50.0 mm Hg,

P_{abs} = 710 mm Hg.

3. 4.08 m

5.

```
*** QuickLaTeX cannot compile formula:
\begin{array}{}\Delta P=\text{38.7 mm Hg,}\\ \text{Leg blood pressure}=\frac

*** Error message:
Missing # inserted in alignment preamble.
leading text: $\begin{array}{}
Missing $ inserted.
leading text: $\begin{array}{}\Delta
Missing $ inserted.
leading text: ...array}{}\Delta P=\text{38.7 mm Hg,}\\ \text
Missing $ inserted.
leading text: ...ood pressure}=\frac{\text{159}}{\text{119}}
Extra }, or forgotten $.
leading text: ...ood pressure}=\frac{\text{159}}{\text{119}}
Missing } inserted.
leading text: ...e}=\frac{\text{159}}{\text{119}}\end{array}
Extra }, or forgotten $.
leading text: ...e}=\frac{\text{159}}{\text{119}}\end{array}
Missing } inserted.
leading text: ...e}=\frac{\text{159}}{\text{119}}\end{array}
Extra }, or forgotten $.
leading text: ...e}=\frac{\text{159}}{\text{119}}\end{array}
Missing } inserted.
leading text: ...e}=\frac{\text{159}}{\text{119}}\end{array}
```

7. 22.4 cm^2

Archimedes' Principle

Lumen Learning

Learning Objectives

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

- Define buoyant force.
- State Archimedes' principle.
- Understand why objects float or sink.
- Understand the relationship between density and Archimedes' principle.

When you rise from lounging in a warm bath, your arms feel strangely heavy. This is because you no longer have the buoyant support of the water. Where does this buoyant force come from? Why is it that some things float and others do not? Do objects that sink get any support at all from the fluid? Is your body buoyed by the atmosphere, or are only helium balloons affected? (See Figure 1.)

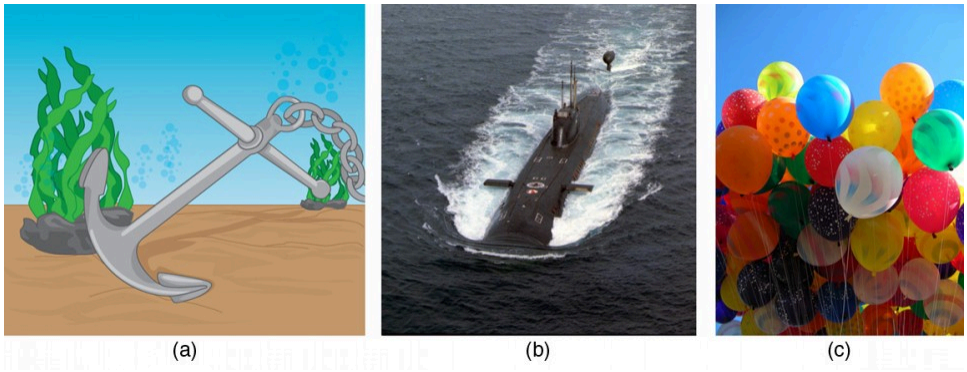


Figure 1. (a) Even objects that sink, like this anchor, are partly supported by water when submerged. (b) Submarines have adjustable density (ballast tanks) so that they may float or sink as desired. (credit: Allied Navy) (c) Helium-filled balloons tug upward on their strings, demonstrating air's buoyant effect. (credit: Crystl)

Answers to all these questions, and many others, are based on the fact that pressure increases with depth in a fluid. This means that the upward force on the bottom of an object in a fluid is greater than the downward force on the top of the object. There is a net upward, or *buoyant force* on any object in any fluid. (See Figure 2.) If the buoyant force is greater than the object's weight, the object will rise to the surface and float. If the buoyant force is less than the object's weight, the object will sink. If the buoyant force equals the object's weight, the object will remain suspended at that depth. The buoyant force is always present whether the object floats, sinks, or is suspended in a fluid.

Buoyant Force

The buoyant force is the net upward force on any object in any fluid.

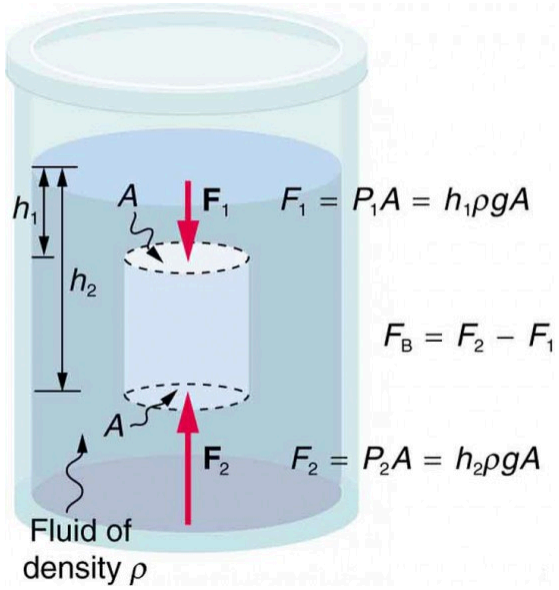


Figure 2. Pressure due to the weight of a fluid increases with depth since P = hpg. This pressure and associated upward force on the bottom of the cylinder are greater than the downward force on the top of the cylinder. Their difference is the buoyant force $\mathbf{F}_B$. *(Horizontal forces cancel.)*

Just how great is this buoyant force? To answer this question, think about what happens when a submerged object is removed from a fluid, as in Figure 3.

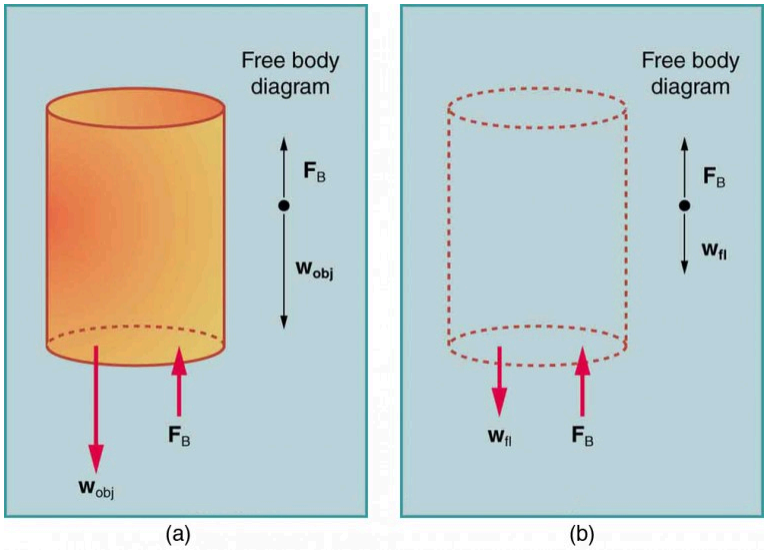


Figure 3. (a) An object submerged in a fluid experiences a buoyant force F_B. If F_B is greater than the weight of the object, the object will rise. If F_B is less than the weight of the object, the object will sink. (b) If the object is removed, it is replaced by fluid having weight w_{fl}. Since this weight is supported by surrounding fluid, the buoyant force must equal the weight of the fluid displaced. That is, $F_B = w_{fl}$,a statement of Archimedes' principle.

The space it occupied is filled by fluid having a weight w_{fl}. This weight is supported by the surrounding fluid, and so the buoyant force must equal w_{fl}, the weight of the fluid displaced by the object. It is a tribute to the genius of the Greek mathematician and inventor Archimedes (ca. 287–212 B.C.) that he stated this principle long before concepts of force were well established. Stated in words, *Archimedes' principle* is as follows: The buoyant force on an object equals the weight of the fluid it displaces. In equation form, Archimedes' principle is

$$F_B = w_{fl},$$

where F_B is the buoyant force and w_{fl} is the weight of the fluid displaced by the object. Archimedes' principle is valid in general, for any object in any fluid, whether partially or totally submerged.

Archimedes' Principle

According to this principle the buoyant force on an object equals the weight of the fluid it displaces. In equation form, Archimedes' principle is

$$F_B = w_{fl},$$

where F_B is the buoyant force and w_{fl} is the weight of the fluid displaced by the object.

Humm … High-tech body swimsuits were introduced in 2008 in preparation for the Beijing Olympics. One concern (and international rule) was that these suits should not provide any buoyancy advantage.

How do you think that this rule could be verified?

Making Connections: Take-Home Investigation

The density of aluminum foil is 2.7 times the density of water. Take a piece of foil, roll it up into a ball and drop it into water. Does it sink? Why or why not? Can you make it sink?

Floating and Sinking

Drop a lump of clay in water. It will sink. Then mold the lump of clay into the shape of a boat, and it will float. Because of its shape, the boat displaces more water than the lump and experiences a greater buoyant force. The same is true of steel ships.

Table 1. Densities of Various Substances

Substance	$\rho\left(10^3\text{kg/m}^3\text{org/mL}\right)$	Substance	$\rho\left(10^3\text{kg/m}^3\text{org/mL}\right)$	Substance	$\rho\left(10^3\text{kg/m}^3\text{org/ml}\right.$
Solids	**Liquids**	**Gases**			
Aluminum	2.7	Water (4ºC)	1.000	Air	1.29×10^{-3}
Brass	8.44	Blood	1.05	Carbon dioxide	1.98×10^{-3}
Copper (average)	8.8	Sea water	1.025	Carbon monoxide	1.25×10^{-3}
Gold	19.32	Mercury	13.6	Hydrogen	0.090×10^{-3}
Iron or steel	7.8	Ethyl alcohol	0.79	Helium	0.18×10^{-3}
Lead	11.3	Petrol	0.68	Methane	0.72×10^{-3}
Polystyrene	0.10	Glycerin	1.26	Nitrogen	1.25×10^{-3}
Tungsten	19.30	Olive oil	0.92	Nitrous oxide	1.98×10^{-3}
Uranium	18.70	Oxygen	1.43×10^{-3}		
Concrete	2.30–3.0	Steam (100º C)	0.60×10^{-3}		
Cork	0.24				
Glass, common (average)	2.6				
Granite	2.7				
Earth's crust	3.3				
Wood	0.3–0.9				
Ice (0°C)	0.917				
Bone	1.7–2.0				

Example 1. Calculating buoyant force: dependency on shape

(a) Calculate the buoyant force on 10,000 metric tons (1.00×10^7 kg) of solid steel completely submerged in water, and compare this with the steel's weight. (b) What is the maximum buoyant force that water could exert on this same steel if it were shaped into a boat that could displace 1.00×10^5 m^3 of water?

Strategy for (a)

To find the buoyant force, we must find the weight of water displaced. We can do this by using the densities of water and steel given in Table 1. We note that, since the steel is completely submerged, its volume and the water's volume are the same. Once we know the volume of water, we can find its mass and weight.

Solution for (a)

First, we use the definition of density

$$\rho = \frac{m}{V}$$

to find the steel's volume, and then we substitute values for mass and density. This gives

$$V_{\text{st}} = \frac{m_{\text{st}}}{\rho_{\text{st}}} = \frac{1.00 \times 10^7 \text{ kg}}{7.8 \times 10^3 \text{ kg/m}^3} = 1.28 \times 10^3 \text{m}^3$$

.

Because the steel is completely submerged, this is also the volume of water displaced, V_{w}. We can now find the mass of water displaced from the relationship between its volume and density, both of which are known. This gives

$$\begin{aligned} m_w &= \rho_w V_w = \left(1.000 \times 10^3 \text{kg/m}^3\right)\left(1.28 \times 10^3 m^3\right) \\ &= 1.28 \times 10^6 \text{kg}. \end{aligned}$$

By Archimedes' principle, the weight of water displaced is $m_{\text{w}}g$, so the buoyant force is

$$\begin{aligned} F_B &= w_w = m_w g = \left(1.28 \times 10^6 \text{kg}\right)\left(9.80 \text{m/s}^2\right) \\ &= 1.3 \times 10^7 \text{N} \end{aligned}$$

.

The steel's weight is

$$m_{\text{w}}g = 9.80 \times 10^7 \text{N}$$

, which is much greater than the buoyant force, so the steel will remain submerged. Note that the buoyant force is rounded to two digits because the density of steel is given to only two digits.

Strategy for (b)

Here we are given the maximum volume of water the steel boat can displace. The buoyant force is the weight of this volume of water.

Solution for (b)

The mass of water displaced is found from its relationship to density and volume, both of which are known. That is,

$$\begin{aligned} m_w &= \rho_w V_w = \left(1.000 \times 10^3 \text{kg/m}^3\right)\left(1.00 \times 10^5 m^3\right) \\ &= 1.00 \times 10^8 \text{kg} \end{aligned}$$

.

The maximum buoyant force is the weight of this much water, or

$$\begin{aligned} F_B &= w_w = m_w g = \left(1.00 \times 10^8 \text{kg}\right)\left(9.80 \text{m/s}^2\right) \\ &= 9.80 \times 10^8 \text{N} \end{aligned}$$

.

Discussion

The maximum buoyant force is ten times the weight of the steel, meaning the ship can carry a load nine times its own weight without sinking.

Making Connections: Take-Home Investigation

A piece of household aluminum foil is 0.016 mm thick. Use a piece of foil that measures 10 cm by 15 cm. (a) What is the mass of this amount of foil? (b) If the foil is folded to give it four sides, and paper clips or washers are added to this "boat," what shape of the boat would allow it to hold the most "cargo" when placed in water? Test your prediction.

Density and Archimedes' Principle

Density plays a crucial role in Archimedes' principle. The average density of an object is what ultimately determines whether it floats. If its average density is less than that of the surrounding fluid, it will float. This is because the fluid, having a higher density, contains more mass and hence more weight in the same volume. The buoyant force, which equals the weight of the fluid displaced, is thus greater than the weight of the object. Likewise, an object denser than the fluid will sink. The extent to which a floating object is submerged depends on how the object's density is related to that of the fluid. In Figure 4, for example, the unloaded ship has a lower density and less of it is submerged compared with the same ship loaded. We can derive a quantitative expression for the fraction submerged by considering density. The fraction submerged is the ratio of the volume submerged to the volume of the object, or

$$\text{fraction submerged} = \frac{V_{\text{sub}}}{V_{\text{obj}}} = \frac{V_{\text{fl}}}{V_{\text{obj}}}$$

.

The volume submerged equals the volume of fluid displaced, which we call V_{fl}. Now we can obtain the relationship between the densities by substituting

$$\rho = \frac{m}{V}$$

into the expression. This gives

$$\frac{V_{\text{fl}}}{V_{\text{obj}}} = \frac{m_{\text{fl}}/\rho_{\text{fl}}}{m_{\text{obj}}/\overline{\rho}_{\text{obj}}},$$

where

$$\overline{\rho}_{\text{obj}}$$

is the average density of the object and ρ_{fl} is the density of the fluid. Since the object floats, its mass and that of the displaced fluid are equal, and so they cancel from the equation, leaving

$$\text{fraction submerged} = \frac{\overline{\rho}_{\text{obj}}}{\rho_{\text{fl}}}$$

.

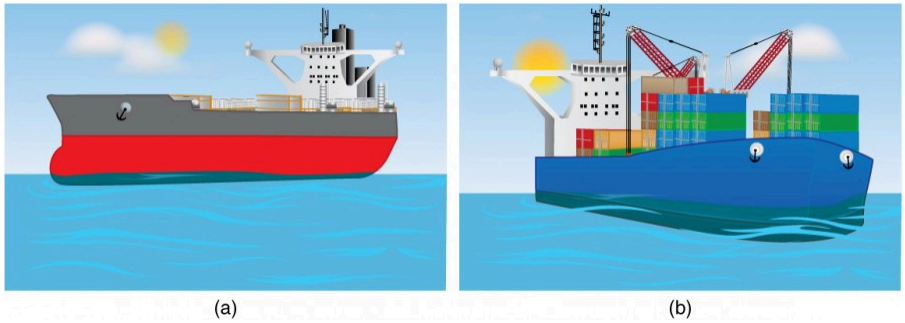


Figure 4. An unloaded ship (a) floats higher in the water than a loaded ship (b).

We use this last relationship to measure densities. This is done by measuring the fraction of a floating object that is submerged—for example, with a hydrometer. It is useful to define the ratio of the density of an object to a fluid (usually water) as *specific gravity*:

$$\text{specific gravity} = \frac{\overline{\rho}}{\rho_{\text{w}}}$$

,

where

$$\overline{\rho}$$

is the average density of the object or substance and ρ_{w} is the density of water at 4.00°C. Specific gravity is dimensionless, independent of whatever units are used for ρ. If an object floats, its specific gravity is less than one. If it sinks, its specific gravity is greater than one. Moreover, the fraction of a floating object that is submerged equals its specific gravity. If an object's specific gravity is exactly 1, then it will remain suspended in the fluid, neither sinking nor floating. Scuba divers try to obtain this state so that they can hover in the water. We measure the specific gravity of fluids, such as battery acid, radiator fluid, and urine, as an indicator of their condition. One device for measuring specific gravity is shown in Figure 5.

Specific Gravity

Specific gravity is the ratio of the density of an object to a fluid (usually water).

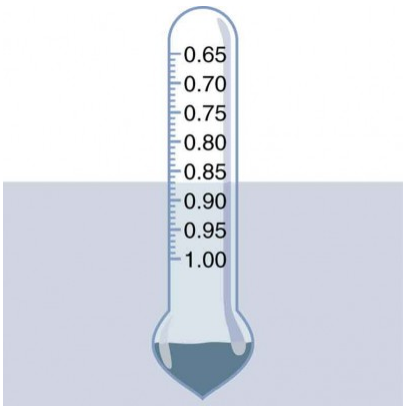


Figure 5. This hydrometer is floating in a fluid of specific gravity 0.87. The glass hydrometer is filled with air and weighted with lead at the bottom. It floats highest in the densest fluids and has been calibrated and labeled so that specific gravity can be read from it directly.

Example 2. Calculating Average Density: Floating Woman

Suppose a 60.0-kg woman floats in freshwater with 97.0% of her volume submerged when her lungs are full of air. What is her average density?

Strategy

We can find the woman's density by solving the equation

$$\text{fraction submerged} = \frac{\overline{\rho}_{\text{obj}}}{\rho_{\text{fl}}}$$

for the density of the object. This yields

$$\overline{\rho}_{\text{obj}} = \overline{\rho}_{\text{person}} = (\text{fraction submerged}) \cdot \rho_{\text{fl}}$$

.

We know both the fraction submerged and the density of water, and so we can calculate the woman's density.

Solution

Entering the known values into the expression for her density, we obtain

$$\overline{rho}_{\text{person}} = 0.970 \cdot \left(10^3 \frac{\text{kg}}{\text{m}^3}\right) = 970 \frac{\text{kg}}{\text{m}^3}$$

.

Discussion

Her density is less than the fluid density. We expect this because she floats. Body density is one indicator of a person's percent body fat, of interest in medical diagnostics and athletic training. (See Figure 6.)

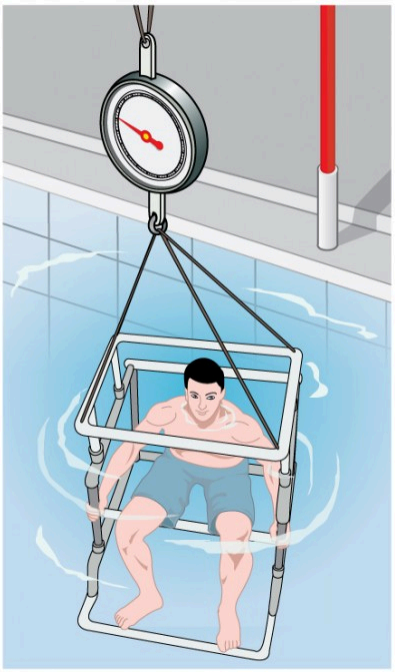


Figure 6. Subject in a "fat tank," where he is weighed while completely submerged as part of a body density determination. The subject must completely empty his lungs and hold a metal weight in order to sink. Corrections are made for the residual air in his lungs (measured separately) and the metal weight. His corrected submerged weight, his weight in air, and pinch tests of strategic fatty areas are used to calculate his percent body fat.

There are many obvious examples of lower-density objects or substances floating in higher-density fluids—oil on water, a hot-air balloon, a bit of cork in wine, an iceberg, and hot wax in a "lava lamp," to name a few. Less obvious examples include lava rising in a volcano and mountain ranges floating on the higher-density crust and mantle beneath them. Even seemingly solid Earth has fluid characteristics.

More Density Measurements

One of the most common techniques for determining density is shown in Figure 7.

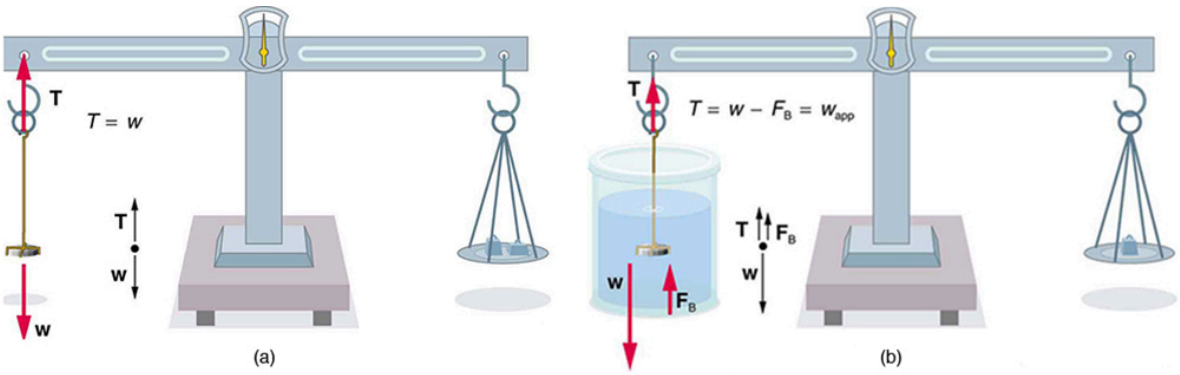


Figure 7. (a) A coin is weighed in air. (b) The apparent weight of the coin is determined while it is completely submerged in a fluid of known density. These two measurements are used to calculate the density of the coin.

An object, here a coin, is weighed in air and then weighed again while submerged in a liquid. The density of the coin, an indication of its authenticity, can be calculated if the fluid density is known. This same technique can also be used to determine the density of the fluid if the density of the coin is known. All of these calculations are based on Archimedes' principle. Archimedes' principle states that the buoyant force on the object equals the weight of the fluid displaced. This, in turn, means that the object *appears* to weigh less when submerged; we call this measurement the object's *apparent weight*. The object suffers an *apparent weight loss* equal to the weight of the fluid displaced. Alternatively, on balances that measure mass, the object suffers an *apparent mass loss* equal to the mass of fluid displaced. That is

apparent weight loss = weight of fluid displaced

or

apparent mass loss = mass of fluid displaced.

The next example illustrates the use of this technique.

Example 3. Calculating Density: Is the Coin Authentic?

The mass of an ancient Greek coin is determined in air to be 8.630 g. When the coin is submerged in water as shown in Figure 7, its apparent mass is 7.800 g. Calculate its density, given that water has a density of 1.000 g/cm^3 and that effects caused by the wire suspending the coin are negligible.

Strategy

To calculate the coin's density, we need its mass (which is given) and its volume. The volume of the coin equals the volume of water displaced. The volume of water displaced V_w can be found by solving the equation for density

$$\rho = \frac{m}{V}$$

for V.

Solution

The volume of water is

$$V_{\mathrm{w}} = \frac{m_{\mathrm{w}}}{\rho_{\mathrm{w}}}$$

where m_{w} is the mass of water displaced. As noted, the mass of the water displaced equals the apparent mass loss, which is m_{w} = 8.630 g − 7.800 g = 0.830 g. Thus the volume of water is

$$V_{\mathrm{w}} = \frac{0.830\ \mathrm{g}}{1.000\ \mathrm{g/cm^3}} = 0.830\mathrm{cm}^3$$

. This is also the volume of the coin, since it is completely submerged. We can now find the density of the coin using the definition of density:

$$\rho_{\mathrm{c}} = \frac{m_{\mathrm{c}}}{V_c} = \frac{8.630\ \mathrm{g}}{0.830\ \mathrm{cm^3}} = 10.4\ \mathrm{g/cm^3}$$

.

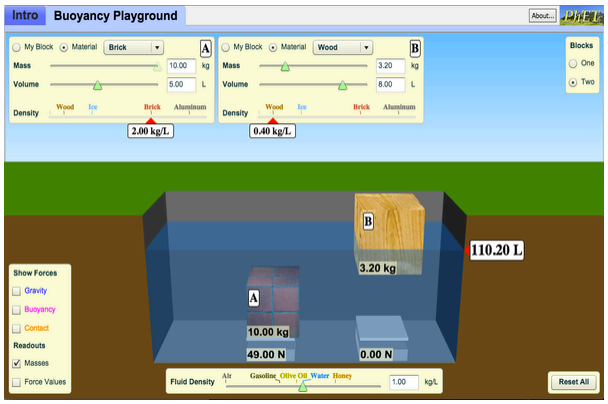
Discussion

You can see from Table 1 that this density is very close to that of pure silver, appropriate for this type of ancient coin. Most modern counterfeits are not pure silver.

This brings us back to Archimedes' principle and how it came into being. As the story goes, the king of Syracuse gave Archimedes the task of determining whether the royal crown maker was supplying a crown of pure gold. The purity of gold is difficult to determine by color (it can be diluted with other metals and still look as yellow as pure gold), and other analytical techniques had not yet been conceived. Even ancient peoples, however, realized that the density of gold was greater than that of any other then-known substance. Archimedes purportedly agonized over his task and had his inspiration one day while at the public baths, pondering the support the water gave his body. He came up with his now-famous principle, saw how to apply it to determine density, and ran naked down the streets of Syracuse crying "Eureka!" (Greek for "I have found it"). Similar behavior can be observed in contemporary physicists from time to time!

PhET Explorations: Buoyancy

When will objects float and when will they sink? Learn how buoyancy works with blocks. Arrows show the applied forces, and you can modify the properties of the blocks and the fluid.



Click to run the simulation.

Section Summary

- Buoyant force is the net upward force on any object in any fluid. If the buoyant force is greater than the object's weight, the object will rise to the surface and float. If the buoyant force is less than the object's weight, the object will sink. If the buoyant force equals the object's weight, the object will remain suspended at that depth. The buoyant force is always present whether the object floats, sinks, or is suspended in a fluid.
- Archimedes' principle states that the buoyant force on an object equals the weight of the fluid it displaces.
- Specific gravity is the ratio of the density of an object to a fluid (usually water).

Conceptual Questions

1. More force is required to pull the plug in a full bathtub than when it is empty. Does this contradict Archimedes' principle? Explain your answer.

2. Do fluids exert buoyant forces in a "weightless" environment, such as in the space shuttle? Explain your answer.

3. Will the same ship float higher in salt water than in freshwater? Explain your answer.

4. Marbles dropped into a partially filled bathtub sink to the bottom. Part of their weight is supported by buoyant force, yet the downward force on the bottom of the tub increases by exactly the weight of the marbles. Explain why.

Problem & Exercises

1. What fraction of ice is submerged when it floats in freshwater, given the density of water at 0°C is very close to 1000 kg/m^3?

2. Logs sometimes float vertically in a lake because one end has become water-logged and denser than the other. What is the average density of a uniform-diameter log that floats with 20.0% of its length above water?

3. Find the density of a fluid in which a hydrometer having a density of 0.750 g/mL floats with 92.0% of its volume submerged.

4. If your body has a density of 995 kg/m^3, what fraction of you will be submerged when floating gently in: (a) Freshwater? (b) Salt water, which has a density of 1027 kg/m^3?

5. Bird bones have air pockets in them to reduce their weight—this also gives them an average density significantly less than that of the bones of other animals. Suppose an ornithologist weighs a bird bone in air and in water and finds its mass is 45.0 g and its apparent mass when submerged is 3.60 g (the bone is watertight). (a) What mass of water is displaced? (b) What is the volume of the bone? (c) What is its average density?

6. A rock with a mass of 540 g in air is found to have an apparent mass of 342 g when submerged in water. (a) What mass of water is displaced? (b) What is the volume of the rock? (c) What is its average density? Is this consistent with the value for granite?

7. Archimedes' principle can be used to calculate the density of a fluid as well as that of a solid. Suppose a chunk of iron with a mass of 390.0 g in air is found to have an apparent mass of 350.5 g when completely submerged in an unknown liquid. (a) What mass of fluid does the iron displace? (b) What is the volume of iron, using its density as given in Table 1. (c) Calculate the fluid's density and identify it.

8. In an immersion measurement of a woman's density, she is found to have a mass of 62.0 kg in air and an apparent mass of 0.0850 kg when completely submerged with lungs empty. (a) What mass of water does she displace? (b) What is her volume? (c) Calculate her density. (d) If her lung capacity is 1.75 L, is she able to float without treading water with her lungs filled with air?

9. Some fish have a density slightly less than that of water and must exert a force (swim) to stay submerged. What force must an 85.0-kg grouper exert to stay submerged in salt water if its body density is 1015 kg/m^3?

10. (a) Calculate the buoyant force on a 2.00-L helium balloon. (b) Given the mass of the rubber in the balloon is 1.50 g, what is the net vertical force on the balloon if it is let go? You can neglect the volume of the rubber.

11. (a) What is the density of a woman who floats in freshwater with 4.00% of her volume above the surface? This could be measured by placing her in a tank with marks on the side to measure how much water she displaces when floating and when held under water (briefly). (b) What percent of her volume is above the surface when she floats in seawater?

12. A certain man has a mass of 80 kg and a density of 955 kg/m^3 (excluding the air in his lungs). (a) Calculate his volume. (b) Find the buoyant force air exerts on him. (c) What is the ratio of the buoyant force to his weight?

13. A simple compass can be made by placing a small bar magnet on a cork floating in water. (a) What fraction of a plain cork will be submerged when floating in water? (b) If the cork has a mass of 10.0 g and a 20.0-g magnet is placed on it, what fraction of the cork will be submerged? (c) Will the bar magnet and cork float in ethyl alcohol?

14. What fraction of an iron anchor's weight will be supported by buoyant force when submerged in saltwater?

15. Scurrilous con artists have been known to represent gold-plated tungsten ingots as pure gold and sell them to the greedy at prices much below gold value but deservedly far above the cost of tungsten. With what accuracy must you be able to measure the mass of such an ingot in and out of water to tell that it is almost pure tungsten rather than pure gold?

16. A twin-sized air mattress used for camping has dimensions of 100 cm by 200 cm by 15 cm when blown up. The weight of the mattress is 2 kg. How heavy a person could the air mattress hold if it is placed in freshwater?

17. Referring to Figure 3, prove that the buoyant force on the cylinder is equal to the weight of the fluid displaced (Archimedes' principle). You may assume that the buoyant force is $F_1 - F_2$ and that the ends of the cylinder have equal areas A. Note that the volume of the cylinder (and that of the fluid it displaces))A equals $(h_2 - h_1)A$.

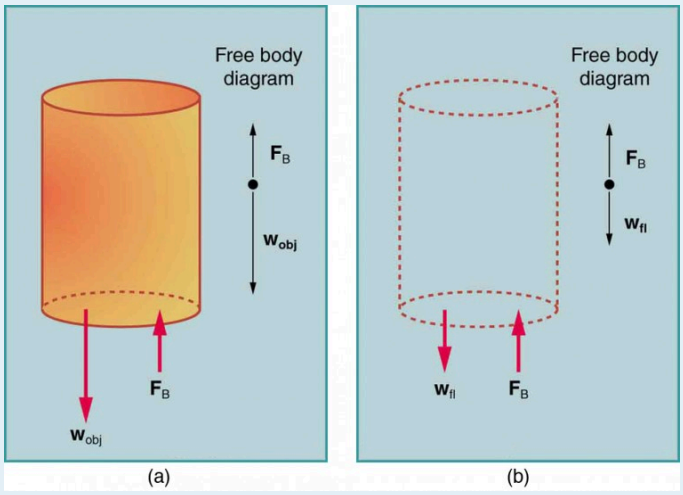


Figure 3. (a) An object submerged in a fluid experiences a buoyant force F_B. If F_B is greater than the weight of the object, the object will rise. If F_B is less than the weight of the object, the object will sink. (b) If the object is removed, it is replaced by fluid having weight w_{fl}. Since this weight is supported by surrounding fluid, the buoyant force must equal the weight of the fluid displaced. That is, $F_B = w_{fl}$,a statement of Archimedes' principle.

18. (a) A 75.0-kg man floats in freshwater with 3.00% of his volume above water when his lungs are empty, and 5.00% of his volume above water when his lungs are full. Calculate the volume of air he inhales—called his lung capacity—in liters. (b) Does this lung volume seem reasonable?

Glossary

Archimedes' principle:
the buoyant force on an object equals the weight of the fluid it displaces
buoyant force:
the net upward force on any object in any fluid
specific gravity:
the ratio of the density of an object to a fluid (usually water)

Selected Solutions to Problems & Exercises

1. 91.7%

3. 815 kg/m^3

5. (a) 41.4 g (b) 41.4cm^3 (c) 1.09 g/cm^3

7. (a) 39.5 g (b) 50cm^3 (c) 0.79g/cm^3

It is ethyl alcohol.

9. 8.21 N

11. (a) 960kg/m^3 (b) 6.34%

She indeed floats more in seawater.

13. (a) 0.24 (b) 0.68 (c) Yes, the cork will float because

$$\rho_{\text{obj}} < \rho_{\text{ethyl alcohol}} \left(0.678\text{g/cm}^3 < 0.79\text{g/cm}^3\right)$$

15. The difference is 0.006%.

17.

$$F_{\text{net}} = F_2 - F_1 = P_2A - P_1A = (P_2 - P_1)\,A$$

$$= (h_2\rho_{\text{fl}}g - h_1\rho_{\text{fl}}g)\,A$$

$$= (h_2 - h_1)\,\rho_{\text{fl}}\text{gA}$$

where

$$\rho_{\text{fl}}$$

= density of fluid. Therefore,

$$F_{\text{net}} = (h_2 - h_1)\,\text{A}\rho_{\text{fl}}g = V_{\text{fl}}\rho_{\text{fl}}g = m_{\text{fl}}g = w_{\text{fl}}$$

where is

$$w_{\text{fl}}$$

the weight of the fluid displaced.

Video: Buoyancy

Lumen Learning

Watch the following Physics Concept Trailer to see how sea creatures and scuba divers all change their buoyancy to dive to deeper depths.

A YouTube element has been excluded from this version of the text. You can view it online here: https://pressbooks.nscc.ca/heatlightsound/?p=55

Cohesion and Adhesion in Liquids: Surface Tension and Capillary Action

Lumen Learning

Learning Objectives

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

- Understand cohesive and adhesive forces.
- Define surface tension.
- Understand capillary action.

Cohesion and Adhesion in Liquids

Children blow soap bubbles and play in the spray of a sprinkler on a hot summer day. (See Figure 1.) An underwater spider keeps his air supply in a shiny bubble he carries wrapped around him. A technician draws blood into a small-diameter tube just by touching it to a drop on a pricked finger. A premature infant struggles to inflate her lungs. What is the common thread? All these activities are dominated by the attractive forces between atoms and molecules in liquids—both within a liquid and between the liquid and its surroundings.

Figure 1. The soap bubbles in this photograph are caused by cohesive forces among molecules in liquids. (credit: Steve Ford Elliott)

Attractive forces between molecules of the same type are called *cohesive forces*. Liquids can, for example, be held in open containers because cohesive forces hold the molecules together. Attractive forces between molecules of different types are called *adhesive forces*. Such forces cause liquid drops to cling to window panes, for example. In this section we examine effects directly attributable to cohesive and adhesive forces in liquids.

Cohesive Forces

Attractive forces between molecules of the same type are called cohesive forces.

Adhesive Forces

Attractive forces between molecules of different types are called adhesive forces.

Surface Tension

Cohesive forces between molecules cause the surface of a liquid to contract to the smallest possible surface area. This general effect is called *surface tension*. Molecules on the surface are pulled inward by cohesive forces, reducing the surface area. Molecules inside the liquid experience zero net force, since they have neighbors on all sides.

Surface Tension

Cohesive forces between molecules cause the surface of a liquid to contract to the smallest possible surface area. This general effect is called surface tension.

Making Connections: Surface Tension

Forces between atoms and molecules underlie the macroscopic effect called surface tension. These attractive forces pull the molecules closer together and tend to minimize the surface area. This is another example of a submicroscopic explanation for a macroscopic phenomenon.

The model of a liquid surface acting like a stretched elastic sheet can effectively explain surface tension effects. For example, some insects can walk on water (as opposed to floating in it) as we would walk on a trampoline—they dent the surface as shown in Figure 2(a). Figure 2(b) shows another example, where a needle rests on a water surface. The iron needle cannot, and does not, float, because its density is greater than that of water. Rather, its weight is supported by forces in the stretched surface that try to make the surface smaller or flatter. If the needle were placed point down on the surface, its weight acting on a smaller area would break the surface, and it would sink.

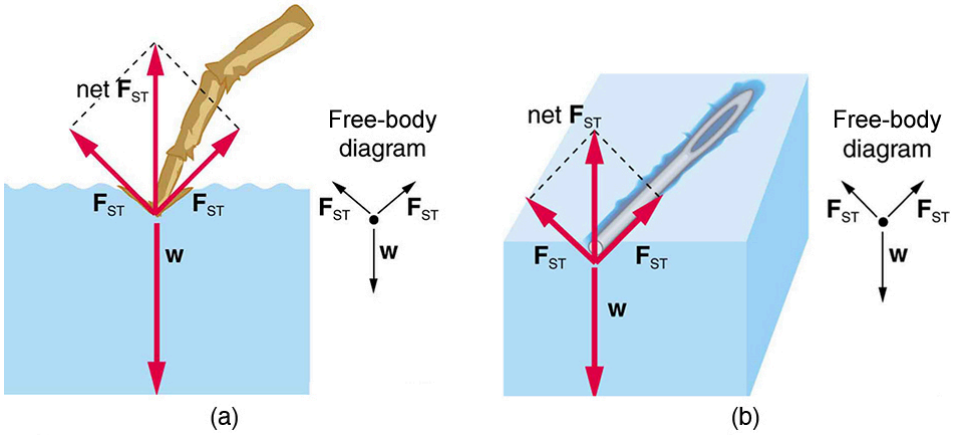


Figure 2. Surface tension supporting the weight of an insect and an iron needle, both of which rest on the surface without penetrating it. They are not floating; rather, they are supported by the surface of the liquid. (a) An insect leg dents the water surface. F_{ST} is a restoring force (surface tension) parallel to the surface. (b) An iron needle similarly dents a water surface until the restoring force (surface tension) grows to equal its weight.

Surface tension is proportional to the strength of the cohesive force, which varies with the type of liquid. Surface tension *γ* is defined to be the force F per unit length L exerted by a stretched liquid membrane:

$$\gamma = \frac{F}{L}$$

.

Table 1. Surface Tension of Some Liquids[1]

Liquid	Surface tension γ(N/m)
Water at 0ºC	0.0756
Water at 20ºC	0.0728
Water at 100ºC	0.0589
Soapy water (typical)	0.0370
Ethyl alcohol	0.0223
Glycerin	0.0631
Mercury	0.465
Olive oil	0.032
Tissue fluids (typical)	0.050
Blood, whole at 37ºC	0.058
Blood plasma at 37ºC	0.073
Gold at 1070ºC	1.000
Oxygen at −193ºC	0.0157
Helium at −269ºC	0.00012

Table 1 above lists values of γ for some liquids. For the insect of Figure 2(a), its weight w is supported by the upward components of the surface tension force: $w = \gamma L \sin\theta$, where L is the circumference of the insect's foot in contact with the water. Figure 3 shows one way to measure surface tension. The liquid film exerts a force on the movable wire in an attempt to reduce its surface area. The magnitude of this force depends on the surface tension of the liquid and can be measured accurately. Surface tension is the reason why liquids form bubbles and droplets. The inward surface tension force causes bubbles to be approximately spherical and raises the pressure of the gas trapped inside relative to atmospheric pressure outside. It can be shown that the gauge pressure P inside a spherical bubble is given by

$$P = \frac{4\gamma}{r}$$

,

where r is the radius of the bubble.

1. At 20ºC unless otherwise stated.

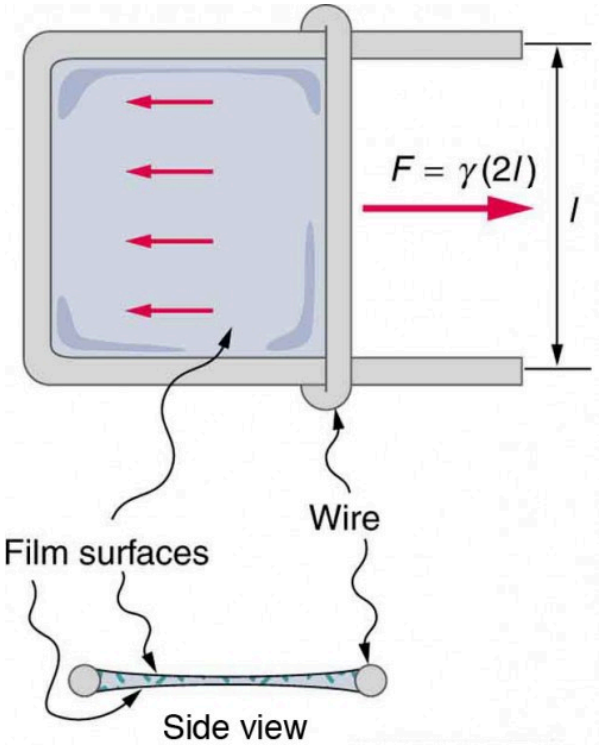


Figure 3. Sliding wire device used for measuring surface tension; the device exerts a force to reduce the film's surface area. The force needed to hold the wire in place is F = γL = γ(2l), since there are two liquid surfaces attached to the wire. This force remains nearly constant as the film is stretched, until the film approaches its breaking point.

Thus the pressure inside a bubble is greatest when the bubble is the smallest. Another bit of evidence for this is illustrated in Figure 4. When air is allowed to flow between two balloons of unequal size, the smaller balloon tends to collapse, filling the larger balloon.

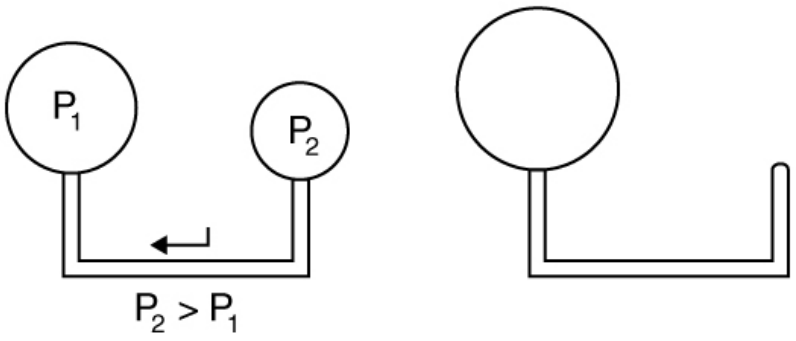


Figure 4. With the valve closed, two balloons of different sizes are attached to each end of a tube. Upon opening the valve, the smaller balloon decreases in size with the air moving to fill the larger balloon. The pressure in a spherical balloon is inversely proportional to its radius, so that the smaller balloon has a greater internal pressure than the larger balloon, resulting in this flow.

Example 1. Surface Tension: Pressure Inside a Bubble

Calculate the gauge pressure inside a soap bubble 2.00×10^{-4} m in radius using the surface tension for soapy water in Table 1. Convert this pressure to mm Hg.

Strategy

The radius is given and the surface tension can be found in Table 1, and so *P* can be found directly from the equation

$$P = \frac{4\gamma}{r}$$

.

Solution

Substituting *r* and *γ* into the equation

$$P = \frac{4\gamma}{r}$$

, we obtain

$$P = \frac{4\gamma}{r} = \frac{4(0.037\ \text{N/m})}{2.00 \times 10^{-4}\text{m}} = 740\ \text{N/m}^2 = 740\ \text{Pa}$$

.

We use a conversion factor to get this into units of mm Hg:

$$P = \left(740\ \text{N/m}^2\right) \frac{1.00\ \text{mm Hg}}{133\ \text{N/m}^2} = 5.56\ \text{mm Hg}$$

.

Discussion

Note that if a hole were to be made in the bubble, the air would be forced out, the bubble would decrease in radius, and the pressure inside would *increase* to atmospheric pressure (760 mm Hg).

Our lungs contain hundreds of millions of mucus-lined sacs called *alveoli*, which are very similar in size, and about 0.1 mm in diameter. (See Figure 5.) You can exhale without muscle action by allowing surface tension to contract these sacs. Medical patients whose breathing is aided by a positive pressure respirator have air blown into the lungs, but are generally allowed to exhale on their own. Even if there is paralysis, surface tension in the alveoli will expel air from the lungs. Since pressure increases as the radii of the alveoli decrease, an occasional deep cleansing breath is needed to fully reinflate the alveoli. Respirators are programmed to do this and we find it natural, as do our companion dogs and cats, to take a cleansing breath before settling into a nap.

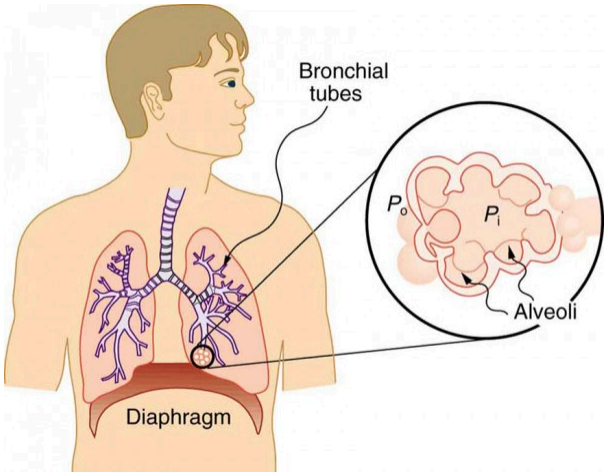


Figure 5. Bronchial tubes in the lungs branch into ever-smaller structures, finally ending in alveoli. The alveoli act like tiny bubbles. The surface tension of their mucous lining aids in exhalation and can prevent inhalation if too great.

The tension in the walls of the alveoli results from the membrane tissue and a liquid on the walls of the alveoli containing a long lipoprotein that acts as a surfactant (a surface-tension reducing substance). The need for the surfactant results from the tendency of small alveoli to collapse and the air to fill into the larger alveoli making them even larger (as demonstrated in Figure 4). During inhalation, the lipoprotein molecules are pulled apart and the wall tension increases as the radius increases (increased surface tension). During exhalation, the molecules slide back together and the surface tension decreases, helping to prevent a collapse of the alveoli. The surfactant therefore serves to change the wall tension so that small alveoli don't collapse and large alveoli are prevented from expanding too much. This tension change is a unique property of these surfactants, and is not shared by detergents (which simply lower surface tension). (See Figure 6.)

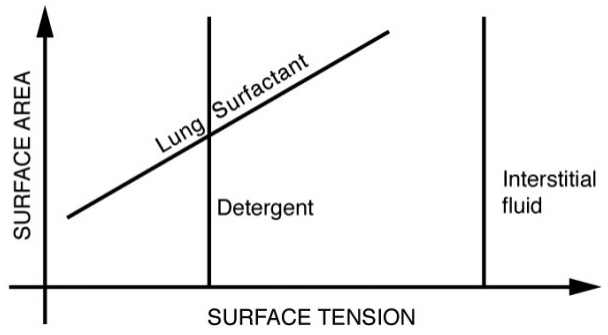


Figure 6. Surface tension as a function of surface area. The surface tension for lung surfactant decreases with decreasing area. This ensures that small alveoli don't collapse and large alveoli are not able to over expand.

If water gets into the lungs, the surface tension is too great and you cannot inhale. This is a severe problem in resuscitating drowning victims. A similar problem occurs in newborn infants who are born without this surfactant—their lungs are very difficult to inflate. This condition is known as *hyaline membrane disease* and is a leading cause of death for infants, particularly in premature births. Some

success has been achieved in treating hyaline membrane disease by spraying a surfactant into the infant's breathing passages. Emphysema produces the opposite problem with alveoli. Alveolar walls of emphysema victims deteriorate, and the sacs combine to form larger sacs. Because pressure produced by surface tension decreases with increasing radius, these larger sacs produce smaller pressure, reducing the ability of emphysema victims to exhale. A common test for emphysema is to measure the pressure and volume of air that can be exhaled.

Making Connections: Take-Home Investigation

(1) Try floating a sewing needle on water. In order for this activity to work, the needle needs to be very clean as even the oil from your fingers can be sufficient to affect the surface properties of the needle. (2) Place the bristles of a paint brush into water. Pull the brush out and notice that for a short while, the bristles will stick together. The surface tension of the water surrounding the bristles is sufficient to hold the bristles together. As the bristles dry out, the surface tension effect dissipates. (3) Place a loop of thread on the surface of still water in such a way that all of the thread is in contact with the water. Note the shape of the loop. Now place a drop of detergent into the middle of the loop. What happens to the shape of the loop? Why? (4) Sprinkle pepper onto the surface of water. Add a drop of detergent. What happens? Why? (5) Float two matches parallel to each other and add a drop of detergent between them. What happens? Note: For each new experiment, the water needs to be replaced and the bowl washed to free it of any residual detergent.

Adhesion and Capillary Action

Why is it that water beads up on a waxed car but does not on bare paint? The answer is that the adhesive forces between water and wax are much smaller than those between water and paint. Competition between the forces of adhesion and cohesion are important in the macroscopic behavior of liquids. An important factor in studying the roles of these two forces is the angle θ between the tangent to the liquid surface and the surface. (See Figure 7.) The *contact angle* θ is directly related to the relative strength of the cohesive and adhesive forces. The larger the strength of the cohesive force relative to the adhesive force, the larger θ is, and the more the liquid tends to form a droplet. The smaller θ is, the smaller the relative strength, so that the adhesive force is able to flatten the drop. Table 2 lists contact angles for several combinations of liquids and solids.

Contact Angle

The angle θ between the tangent to the liquid surface and the surface is called the contact angle.

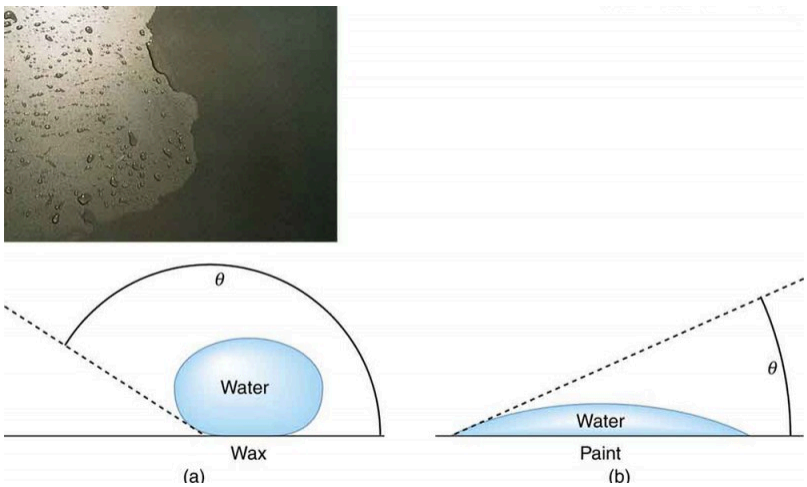


Figure 7. In the photograph, water beads on the waxed car paint and flattens on the unwaxed paint. (a) Water forms beads on the waxed surface because the cohesive forces responsible for surface tension are larger than the adhesive forces, which tend to flatten the drop. (b) Water beads on bare paint are flattened considerably because the adhesive forces between water and paint are strong, overcoming surface tension. The contact angle θ is directly related to the relative strengths of the cohesive and adhesive forces. The larger θ is, the larger the ratio of cohesive to adhesive forces. (credit: P. P. Urone)

One important phenomenon related to the relative strength of cohesive and adhesive forces is *capillary action*—the tendency of a fluid to be raised or suppressed in a narrow tube, or *capillary tube*. This action causes blood to be drawn into a small-diameter tube when the tube touches a drop.

Capillary Action

The tendency of a fluid to be raised or suppressed in a narrow tube, or capillary tube, is called capillary action.

If a capillary tube is placed vertically into a liquid, as shown in Figure 8, capillary action will raise or suppress the liquid inside the tube depending on the combination of substances. The actual effect depends on the relative strength of the cohesive and adhesive forces and, thus, the contact angle θ given in the table. If θ is less than 90º, then the fluid will be raised; if θ is greater than 90º, it will be suppressed. Mercury, for example, has a very large surface tension and a large contact angle with glass. When placed in a tube, the surface of a column of mercury curves downward, somewhat like a drop. The curved surface of a fluid in a tube is called a **meniscus**. The tendency of surface tension is always to reduce the surface area. Surface tension thus flattens the curved liquid surface in a capillary tube. This results in a downward force in mercury and an upward force in water, as seen in Figure 8.

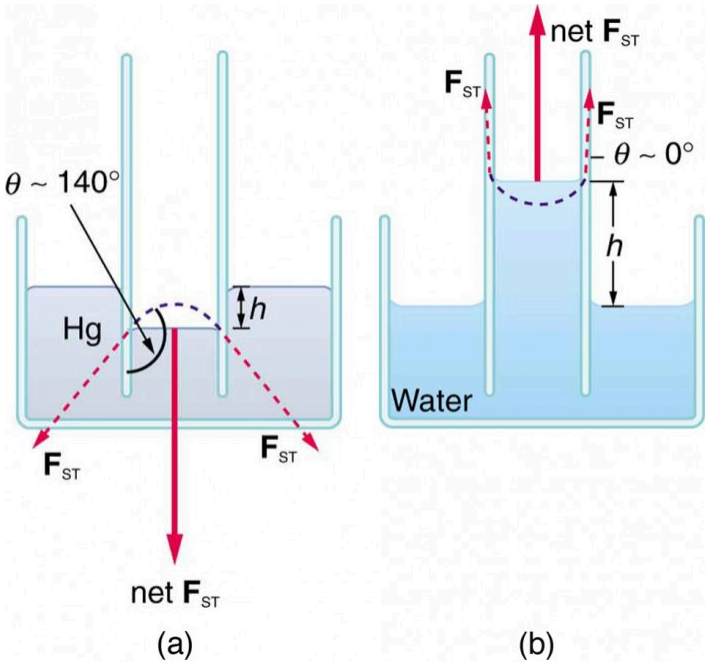


Figure 8. (a) Mercury is suppressed in a glass tube because its contact angle is greater than 90º. Surface tension exerts a downward force as it flattens the mercury, suppressing it in the tube. The dashed line shows the shape the mercury surface would have without the flattening effect of surface tension. (b) Water is raised in a glass tube because its contact angle is nearly 0º. Surface tension therefore exerts an upward force when it flattens the surface to reduce its area.

Table 2. Contact Angles of Some Substances

Interface	**Contact angle Θ**
Mercury–glass	140º
Water–glass	0º
Water–paraffin	107º
Water–silver	90º
Organic liquids (most)–glass	0º
Ethyl alcohol–glass	0º
Kerosene–glass	26º

Capillary action can move liquids horizontally over very large distances, but the height to which it can raise or suppress a liquid in a tube is limited by its weight. It can be shown that this height *h* is given by

$$h = \frac{2\gamma \cos\theta}{\rho g r}$$

.

If we look at the different factors in this expression, we might see how it makes good sense. The height is directly proportional to the surface tension γ, which is its direct cause. Furthermore, the height is inversely proportional to tube radius—the smaller the radius r, the higher the fluid can be raised, since a smaller tube holds less mass. The height is also inversely proportional to fluid density ρ, since a larger density means a greater mass in the same volume. (See Figure 9.)

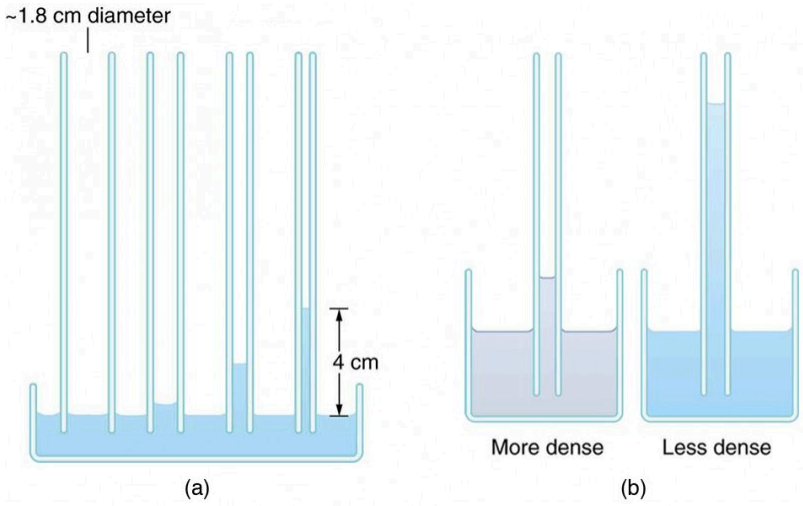


Figure 9. (a) Capillary action depends on the radius of a tube. The smaller the tube, the greater the height reached. The height is negligible for large-radius tubes. (b) A denser fluid in the same tube rises to a smaller height, all other factors being the same.

Example 2. Calculating Radius of a Capillary Tube: Capillary Action: Tree Sap

Can capillary action be solely responsible for sap rising in trees? To answer this question, calculate the radius of a capillary tube that would raise sap 100 m to the top of a giant redwood, assuming that sap's density is 1050 kg/m^3, its contact angle is zero, and its surface tension is the same as that of water at 20.0º C.

Strategy

The height to which a liquid will rise as a result of capillary action is given by

$$h = \frac{2\gamma \cos\theta}{\rho g r}$$

, and every quantity is known except for r.

Solution

Solving for r and substituting known values produces

$$\begin{aligned} r &= \frac{2\gamma\cos\theta}{\rho g h} = \frac{2(0.0728\ \text{N/m})\cos(0^{\text{O}})}{\left(1050\ \text{kg/m}^3\right)\left(9.80\ \text{m/s}^2\right)(100\ \text{m})} \\ &= 1.41 \times 10^{-7}\text{m}. \end{aligned}$$

Discussion

This result is unreasonable. Sap in trees moves through the *xylem*, which forms tubes with radii as small as

2.5×10^{-5} m. This value is about 180 times as large as the radius found necessary here to raise sap 100 m. This means that capillary action alone cannot be solely responsible for sap getting to the tops of trees.

How *does* sap get to the tops of tall trees? (Recall that a column of water can only rise to a height of 10 m when there is a vacuum at the top—see Example 3 from Variation of Pressure with Depth in a Fluid.) The question has not been completely resolved, but it appears that it is pulled up like a chain held together by cohesive forces. As each molecule of sap enters a leaf and evaporates (a process called transpiration), the entire chain is pulled up a notch. So a negative pressure created by water evaporation must be present to pull the sap up through the xylem vessels. In most situations, *fluids can push but can exert only negligible pull*, because the cohesive forces seem to be too small to hold the molecules tightly together. But in this case, the cohesive force of water molecules provides a very strong pull. Figure 10 shows one device for studying negative pressure. Some experiments have demonstrated that negative pressures sufficient to pull sap to the tops of the tallest trees *can* be achieved.

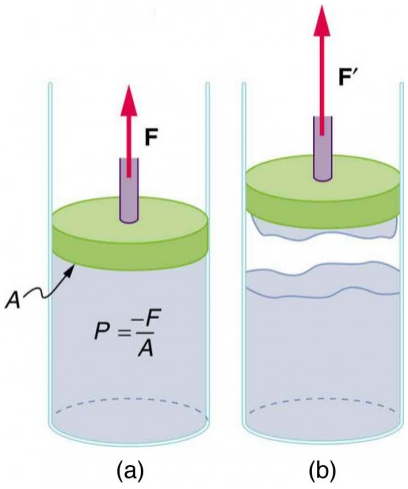


Figure 10. (a) When the piston is raised, it stretches the liquid slightly, putting it under tension and creating a negative absolute pressure $P = -F/A$. (b) The liquid eventually separates, giving an experimental limit to negative pressure in this liquid.

Section Summary

- Attractive forces between molecules of the same type are called cohesive forces.
- Attractive forces between molecules of different types are called adhesive forces.
- Cohesive forces between molecules cause the surface of a liquid to contract to the smallest possible surface area. This general effect is called surface tension.
- Capillary action is the tendency of a fluid to be raised or suppressed in a narrow tube, or capillary tube which is due to the relative strength of cohesive and adhesive forces.

Conceptual Questions

1. The density of oil is less than that of water, yet a loaded oil tanker sits lower in the water than an empty one. Why?

2. Is surface tension due to cohesive or adhesive forces, or both?

3. Is capillary action due to cohesive or adhesive forces, or both?

4. Birds such as ducks, geese, and swans have greater densities than water, yet they are able to sit on its surface. Explain this ability, noting that water does not wet their feathers and that they cannot sit on soapy water.

5. Water beads up on an oily sunbather, but not on her neighbor, whose skin is not oiled. Explain in terms of cohesive and adhesive forces.

6. Could capillary action be used to move fluids in a “weightless” environment, such as in an orbiting space probe?

7. What effect does capillary action have on the reading of a manometer with uniform diameter? Explain your answer.

8. Pressure between the inside chest wall and the outside of the lungs normally remains negative. Explain how pressure inside the lungs can become positive (to cause exhalation) without muscle action.

Problems & Exercises

1. What is the pressure inside an alveolus having a radius of 2.50×10^{-4} if the surface tension of the fluid-lined wall is the same as for soapy water? You may assume the pressure is the same as that created by a spherical bubble.

2. (a) The pressure inside an alveolus with a 2.00×10^{-4} -m radius is 1.40×10^{3}, due to its fluid-lined walls. Assuming the alveolus acts like a spherical bubble, what is the surface tension of the fluid? (b) Identify the likely fluid. (You may need to extrapolate between values in Table 1.)

3. What is the gauge pressure in millimeters of mercury inside a soap bubble 0.100 m in diameter?

4. Calculate the force on the slide wire in Figure 3 (shown again below) if it is 3.50 cm long and the fluid is ethyl alcohol.

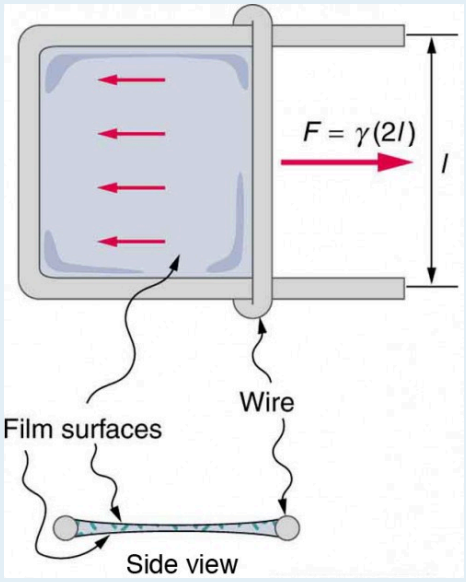


Figure 3. Sliding wire device used for measuring surface tension; the device exerts a force to reduce the film's surface area. The force needed to hold the wire in place is F = γL = γ(2l), since there are two liquid surfaces attached to the wire. This force remains nearly constant as the film is stretched, until the film approaches its breaking point.

5. Figure 9(a) (shown again below) shows the effect of tube radius on the height to which capillary action can raise a fluid. (a) Calculate the height *h* for water in a glass tube with a radius of 0.900 cm—a rather large tube like the one on the left. (b) What is the radius of the glass tube on the right if it raises water to 4.00 cm?

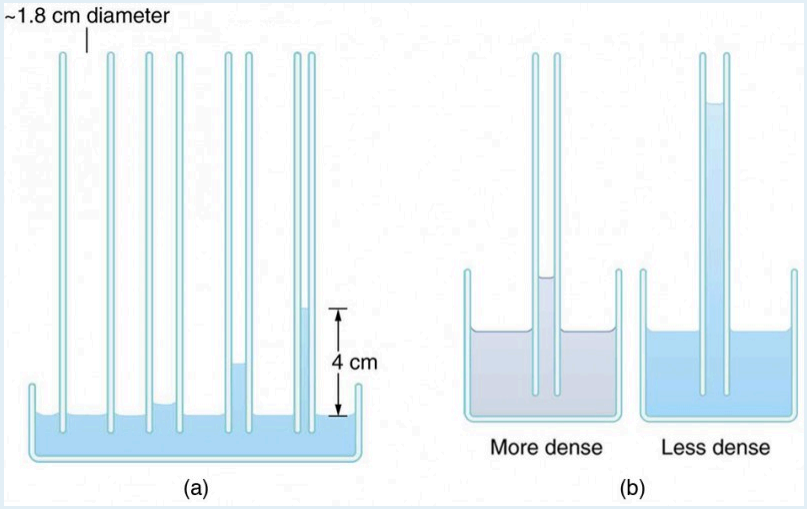


Figure 9. (a) Capillary action depends on the radius of a tube. The smaller the tube, the greater the height reached. The height is negligible for large-radius tubes. (b) A denser fluid in the same tube rises to a smaller height, all other factors being the same.

6. We stated in Example 2 above that a xylem tube is of radius 2.50×10^{-5} m. Verify that such a tube raises

sap less than a meter by finding *h* for it, making the same assumptions that sap's density is 1050 kg/m^3, its contact angle is zero, and its surface tension is the same as that of water at 20.0º C.

7. What fluid is in the device shown in Figure 3 (shown again below) if the force is 3.16×10^{-3} and the length of the wire is 2.50 cm? Calculate the surface tension γ and find a likely match from Table 1 (above).

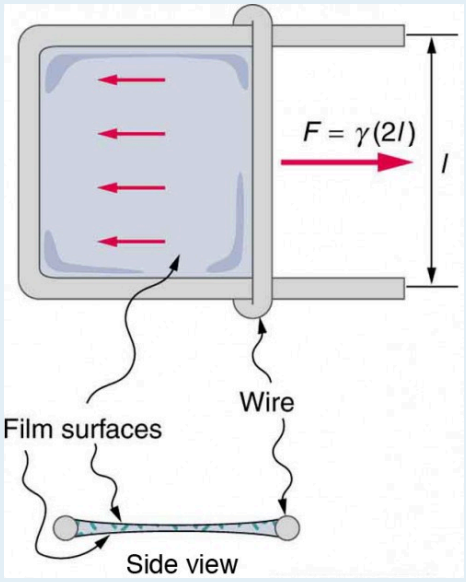


Figure 3. Sliding wire device used for measuring surface tension; the device exerts a force to reduce the film's surface area. The force needed to hold the wire in place is F = γL = γ(2l), since there are two liquid surfaces attached to the wire. This force remains nearly constant as the film is stretched, until the film approaches its breaking point.

8. If the gauge pressure inside a rubber balloon with a 10.0-cm radius is 1.50 cm of water, what is the effective surface tension of the balloon?

9. Calculate the gauge pressures inside 2.00-cm-radius bubbles of water, alcohol, and soapy water. Which liquid forms the most stable bubbles, neglecting any effects of evaporation?

10. Suppose water is raised by capillary action to a height of 5.00 cm in a glass tube. (a) To what height will it be raised in a paraffin tube of the same radius? (b) In a silver tube of the same radius?

11. Calculate the contact angle θ for olive oil if capillary action raises it to a height of 7.07 cm in a glass tube with a radius of 0.100 mm. Is this value consistent with that for most organic liquids?

12. When two soap bubbles touch, the larger is inflated by the smaller until they form a single bubble. (a) What is the gauge pressure inside a soap bubble with a 1.50-cm radius? (b) Inside a 4.00-cm-radius soap bubble? (c) Inside the single bubble they form if no air is lost when they touch?

13. Calculate the ratio of the heights to which water and mercury are raised by capillary action in the same glass tube.

14. What is the ratio of heights to which ethyl alcohol and water are raised by capillary action in the same glass tube?

Glossary

adhesive forces:
the attractive forces between molecules of different types

capillary action:
the tendency of a fluid to be raised or lowered in a narrow tube

cohesive forces:
the attractive forces between molecules of the same type

contact angle:
the angle θ between the tangent to the liquid surface and the surface

surface tension:
the cohesive forces between molecules which cause the surface of a liquid to contract to the smallest possible surface area

Selected Solutions to Problems & Exercises

1. 592 N/m^2

3. 2.23×10^{-2} mm Hg

5. (a) 1.65×10^{-3} m (b) 3.71×10^{-4} m

7. 6.32×10^{-2} N/m. Based on the values in table, the fluid is probably glycerin.

9.

$$\begin{aligned} P_{\text{w}} &= 14.6\ \text{N/m}^2, \\ P_{\text{a}} &= 4.46\ \text{N/m}^2, \\ P_{\text{sw}} &= 7.40\ \text{N/m}^2. \end{aligned}$$

Alcohol forms the most stable bubble, since the absolute pressure inside is closest to atmospheric pressure.

11. 5.1º. This is near the value of θ=0º for most organic liquids.

13. -2.78. The ratio is negative because water is raised whereas mercury is lowered.

Pressures in the Body

Lumen Learning

Learning Objectives

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

- Explain the concept of pressure the in human body.
- Explain systolic and diastolic blood pressures.
- Describe pressures in the eye, lungs, spinal column, bladder, and skeletal system.

Pressure in the Body

Next to taking a person’s temperature and weight, measuring blood pressure is the most common of all medical examinations. Control of high blood pressure is largely responsible for the significant decreases in heart attack and stroke fatalities achieved in the last three decades. The pressures in various parts of the body can be measured and often provide valuable medical indicators. In this section, we consider a few examples together with some of the physics that accompanies them. Table 1 lists some of the measured pressures in mm Hg, the units most commonly quoted.

Table 1. Typical Pressures in Humans

Body system	Gauge pressure in mm Hg
Blood pressures in large arteries (resting)	
Maximum (systolic)	100–140
Minimum (diastolic)	60–90
Blood pressure in large veins	4–15
Eye	12–24
Brain and spinal fluid (lying down)	5–12
Bladder	
While filling	0–25
When full	100–150
Chest cavity between lungs and ribs	−8 to −4
Inside lungs	−2 to +3
Digestive tract	
Esophagus	−2
Stomach	0–20
Intestines	10–20
Middle ear	<1

Blood Pressure

Common arterial blood pressure measurements typically produce values of 120 mm Hg and 80 mm Hg, respectively, for systolic and diastolic pressures. Both pressures have health implications. When systolic pressure is chronically high, the risk of stroke and heart attack is increased. If, however, it is too low, fainting is a problem. *Systolic pressure* increases dramatically during exercise to increase blood flow and returns to normal afterward. This change produces no ill effects and, in fact, may be beneficial to the tone of the circulatory system. *Diastolic pressure* can be an indicator of fluid balance. When low, it may indicate that a person is hemorrhaging internally and needs a transfusion. Conversely, high diastolic pressure indicates a ballooning of the blood vessels, which may be due to the transfusion of too much fluid into the circulatory system. High diastolic pressure is also an indication that blood vessels are not dilating properly to pass blood through. This can seriously strain the heart in its attempt to pump blood.

Blood leaves the heart at about 120 mm Hg but its pressure continues to decrease (to almost 0) as it goes from the aorta to smaller arteries to small veins (see Figure 1). The pressure differences in the circulation system are caused by blood flow through the system as well as the position of the person. For a person standing up, the pressure in the feet will be larger than at the heart due to the weight of the blood ($P = h\rho g$). If we assume that the distance between the heart and the feet of a person in an upright

position is 1.4 m, then the increase in pressure in the feet relative to that in the heart (for a static column of blood) is given by

$$\Delta P = \Delta h \rho g = (1.4 \text{ m}) \left(1050 \text{ kg/m}^3\right) \left(9.80 \text{ m/s}^2\right) = 1.4 \times 10^4 \text{ Pa} = 108 \text{ mm Hg}$$

.

Increase in Pressure in the Feet of a Person

$$\Delta P = \Delta h \rho g = (1.4 \text{ m}) \left(1050 \text{ kg/m}^3\right) \left(9.80 \text{ m/s}^2\right) = 1.4 \times 10^4 \text{ Pa} = 108 \text{ mm Hg}$$

.

Standing a long time can lead to an accumulation of blood in the legs and swelling. This is the reason why soldiers who are required to stand still for long periods of time have been known to faint. Elastic bandages around the calf can help prevent this accumulation and can also help provide increased pressure to enable the veins to send blood back up to the heart. For similar reasons, doctors recommend tight stockings for long-haul flights.

Blood pressure may also be measured in the major veins, the heart chambers, arteries to the brain, and the lungs. But these pressures are usually only monitored during surgery or for patients in intensive care since the measurements are invasive. To obtain these pressure measurements, qualified health care workers thread thin tubes, called catheters, into appropriate locations to transmit pressures to external measuring devices. The heart consists of two pumps—the right side forcing blood through the lungs and the left causing blood to flow through the rest of the body (Figure 1). Right-heart failure, for example, results in a rise in the pressure in the vena cavae and a drop in pressure in the arteries to the lungs. Left-heart failure results in a rise in the pressure entering the left side of the heart and a drop in aortal pressure. Implications of these and other pressures on flow in the circulatory system will be discussed in more detail in Fluid Dynamics and Its Biological and Medical Applications.

Two Pumps of the Heart

The heart consists of two pumps—the right side forcing blood through the lungs and the left causing blood to flow through the rest of the body.

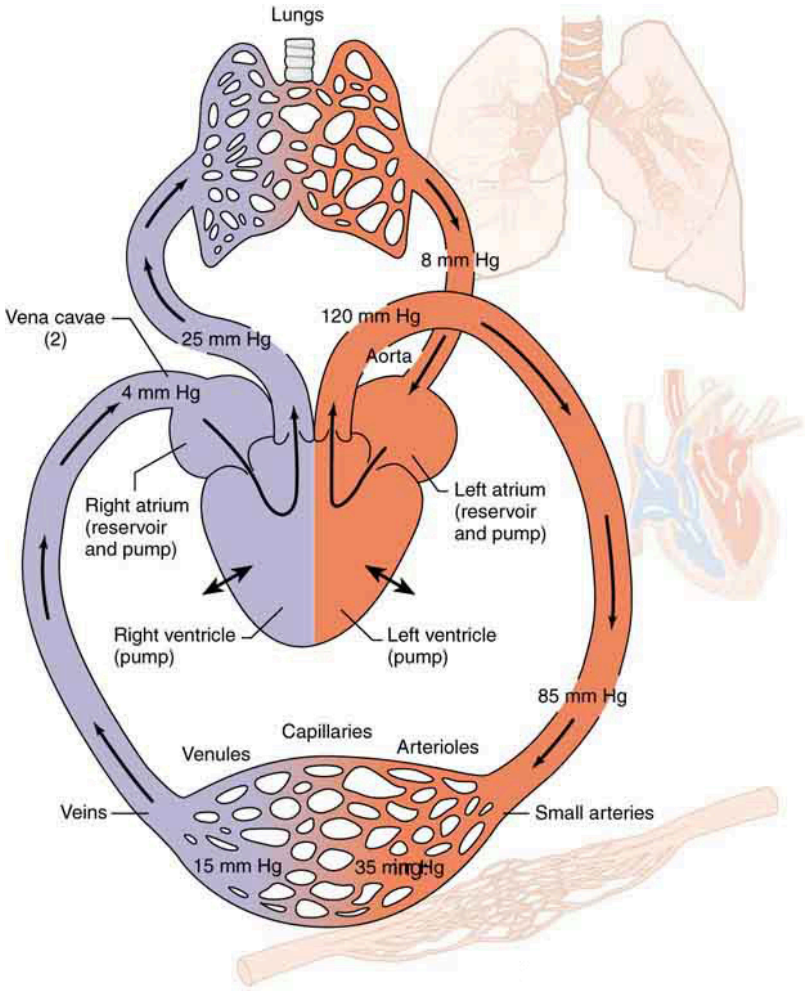


Figure 1. Schematic of the circulatory system showing typical pressures. The two pumps in the heart increase pressure and that pressure is reduced as the blood flows through the body. Long-term deviations from these pressures have medical implications discussed in some detail in the Fluid Dynamics and Its Biological and Medical Applications. Only aortal or arterial blood pressure can be measured non-invasively.

Pressure in the Eye

The shape of the eye is maintained by fluid pressure, called *intraocular pressure*, which is normally in the range of 12.0 to 24.0 mm Hg. When the circulation of fluid in the eye is blocked, it can lead to a buildup in pressure, a condition called *glaucoma*. The net pressure can become as great as 85.0 mm Hg, an abnormally large pressure that can permanently damage the optic nerve. To get an idea of the force involved, suppose the back of the eye has an area of 6.0 cm^2, and the net pressure is 85.0 mm Hg. Force is given by $F = PA$. To get F in newtons, we convert the area to m^2 (1 m^2 = 10^4 cm^2). Then we calculate as follows:

$$F = h\rho gA = \left(85.0 \times 10^{-3}\text{m}\right)\left(13.6 \times 10^3\ \text{kg/m}^3\right)\left(9.80\ \text{m/s}^2\right)\left(6.0 \times 10^{-4}\text{m}^2\right) = 6.8\ \text{N}$$

.

Eye Pressure

The shape of the eye is maintained by fluid pressure, called intraocular pressure. When the circulation of fluid in the eye is blocked, it can lead to a buildup in pressure, a condition called glaucoma. The force is calculated as

$$F = h\rho gA = \left(85.0 \times 10^{-3}\text{m}\right)\left(13.6 \times 10^{3}\text{kg/m}^{3}\right)\left(9.80\ \text{m/s}^{2}\right)\left(6.0 \times 10^{-4}\text{m}^{2}\right) = 6.8\ \text{N}$$

.

This force is the weight of about a 680-g mass. A mass of 680 g resting on the eye (imagine 1.5 lb resting on your eye) would be sufficient to cause it damage. (A normal force here would be the weight of about 120 g, less than one-quarter of our initial value.)

People over 40 years of age are at greatest risk of developing glaucoma and should have their intraocular pressure tested routinely. Most measurements involve exerting a force on the (anesthetized) eye over some area (a pressure) and observing the eye's response. A noncontact approach uses a puff of air and a measurement is made of the force needed to indent the eye (Figure 2). If the intraocular pressure is high, the eye will deform less and rebound more vigorously than normal. Excessive intraocular pressures can be detected reliably and sometimes controlled effectively.

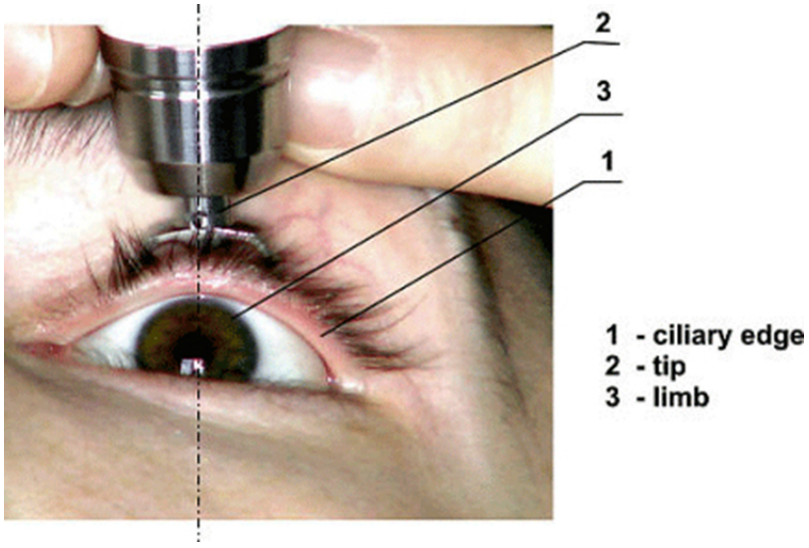


Figure 12. The intraocular eye pressure can be read with a tonometer. (credit: DevelopAll at the Wikipedia Project.)

Example 1. Calculating Gauge Pressure and Depth: Damage to the Eardrum

Suppose a 3.00-N force can rupture an eardrum. (a) If the eardrum has an area of 1.00 cm^2, calculate the maximum tolerable gauge pressure on the eardrum in newtons per meter squared and convert it to millimeters of mercury. (b) At what depth in freshwater would this person's eardrum rupture, assuming the gauge pressure in the middle ear is zero?

Strategy for (a)

The pressure can be found directly from its definition since we know the force and area. We are looking for the gauge pressure.

Solution for (a)

$$P_{\text{g}} = F/A = 3.00\ \text{N}/\left(1.00 \times 10^{-4}\text{m}^2\right) = 3.00 \times 10^4\ \text{N/m}^2.$$

We now need to convert this to units of mm Hg:

$$P_{\text{g}} = 3.0 \times 10^4\ \text{N/m}^2 \left(\frac{1.0\ \text{mm Hg}}{133\ \text{N/m}^2}\right) = 226\ \text{mm Hg}.$$

Strategy for (b)

Here we will use the fact that the water pressure varies linearly with depth *h* below the surface.

Solution for (b)

$P = h\rho g$ and therefore $h = P/\rho g$. Using the value above for P, we have

$$h = \frac{3.0 \times 10^4\ \text{N/m}^2}{\left(1.00 \times 10^3\ \text{kg/m}^3\right)\left(9.80\ \text{m/s}^2\right)} = 3.06\ \text{m}.$$

Discussion

Similarly, increased pressure exerted upon the eardrum from the middle ear can arise when an infection causes a fluid buildup.

Pressure Associated with the Lungs

The pressure inside the lungs increases and decreases with each breath. The pressure drops to below atmospheric pressure (negative gauge pressure) when you inhale, causing air to flow into the lungs. It increases above atmospheric pressure (positive gauge pressure) when you exhale, forcing air out. Lung pressure is controlled by several mechanisms. Muscle action in the diaphragm and rib cage is necessary for inhalation; this muscle action increases the volume of the lungs thereby reducing the pressure within them Figure 3. Surface tension in the alveoli creates a positive pressure opposing inhalation. (See Cohesion and Adhesion in Liquids: Surface Tension and Capillary Action.) You can exhale without muscle action by letting surface tension in the alveoli create its own positive pressure. Muscle action can add to this positive pressure to produce forced exhalation, such as when you blow up a balloon, blow out a candle, or cough. The lungs, in fact, would collapse due to the surface tension in the alveoli, if they

were not attached to the inside of the chest wall by liquid adhesion. The gauge pressure in the liquid attaching the lungs to the inside of the chest wall is thus negative, ranging from −4 to −8 mm Hg during exhalation and inhalation, respectively. If air is allowed to enter the chest cavity, it breaks the attachment, and one or both lungs may collapse. Suction is applied to the chest cavity of surgery patients and trauma victims to reestablish negative pressure and inflate the lungs.

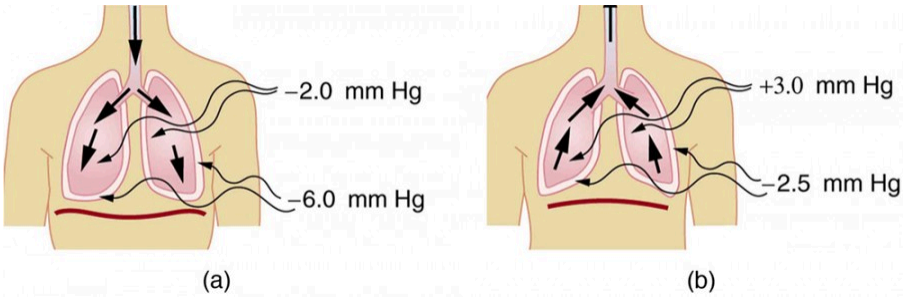


Figure 3. (a) During inhalation, muscles expand the chest, and the diaphragm moves downward, reducing pressure inside the lungs to less than atmospheric (negative gauge pressure). Pressure between the lungs and chest wall is even lower to overcome the positive pressure created by surface tension in the lungs. (b) During gentle exhalation, the muscles simply relax and surface tension in the alveoli creates a positive pressure inside the lungs, forcing air out. Pressure between the chest wall and lungs remains negative to keep them attached to the chest wall, but it is less negative than during inhalation.

Other Pressures in the Body

Spinal Column and Skull

Normally, there is a 5- to 12-mm Hg pressure in the fluid surrounding the brain and filling the spinal column. This cerebrospinal fluid serves many purposes, one of which is to supply flotation to the brain. The buoyant force supplied by the fluid nearly equals the weight of the brain, since their densities are nearly equal. If there is a loss of fluid, the brain rests on the inside of the skull, causing severe headaches, constricted blood flow, and serious damage. Spinal fluid pressure is measured by means of a needle inserted between vertebrae that transmits the pressure to a suitable measuring device.

Bladder Pressure

This bodily pressure is one of which we are often aware. In fact, there is a relationship between our awareness of this pressure and a subsequent increase in it. Bladder pressure climbs steadily from zero to about 25 mm Hg as the bladder fills to its normal capacity of 500 cm^3. This pressure triggers the *micturition reflex*, which stimulates the feeling of needing to urinate. What is more, it also causes muscles around the bladder to contract, raising the pressure to over 100 mm Hg, accentuating the sensation. Coughing, straining, tensing in cold weather, wearing tight clothes, and experiencing simple nervous tension all can increase bladder pressure and trigger this reflex. So can the weight of a pregnant woman’s fetus, especially if it is kicking vigorously or pushing down with its head! Bladder pressure can be measured by a catheter or by inserting a needle through the bladder wall and transmitting the pressure to an appropriate measuring device. One hazard of high bladder pressure (sometimes created by an obstruction), is that such pressure can force urine back into the kidneys, causing potentially severe damage.

Pressures in the Skeletal System

These pressures are the largest in the body, due both to the high values of initial force, and the small areas to which this force is applied, such as in the joints.. For example, when a person lifts an object improperly, a force of 5000 N may be created between vertebrae in the spine, and this may be applied to an area as small as 10 cm^2. The pressure created is $P = F/A$ = (5000 N)/($10^{-3}m^2$) = 5.0×10^6 N/m^2 or about 50 atm! This pressure can damage both the spinal discs (the cartilage between vertebrae), as well as the bony vertebrae themselves. Even under normal circumstances, forces between vertebrae in the spine are large enough to create pressures of several atmospheres. Most causes of excessive pressure in the skeletal system can be avoided by lifting properly and avoiding extreme physical activity. (See Forces and Torques in Muscles and Joints.)

There are many other interesting and medically significant pressures in the body. For example, pressure caused by various muscle actions drives food and waste through the digestive system. Stomach pressure behaves much like bladder pressure and is tied to the sensation of hunger. Pressure in the relaxed esophagus is normally negative because pressure in the chest cavity is normally negative. Positive pressure in the stomach may thus force acid into the esophagus, causing “heartburn.” Pressure in the middle ear can result in significant force on the eardrum if it differs greatly from atmospheric pressure, such as while scuba diving. The decrease in external pressure is also noticeable during plane flights (due to a decrease in the weight of air above relative to that at the Earth’s surface). The Eustachian tubes connect the middle ear to the throat and allow us to equalize pressure in the middle ear to avoid an imbalance of force on the eardrum.

Many pressures in the human body are associated with the flow of fluids. Fluid flow will be discussed in detail in the Fluid Dynamics and Its Biological and Medical Applications.

Section Summary

- Measuring blood pressure is among the most common of all medical examinations.
- The pressures in various parts of the body can be measured and often provide valuable medical indicators.
- The shape of the eye is maintained by fluid pressure, called intraocular pressure.
- When the circulation of fluid in the eye is blocked, it can lead to a buildup in pressure, a condition called glaucoma.
- Some of the other pressures in the body are spinal and skull pressures, bladder pressure, pressures in the skeletal system.

Problems & Exercises

1. During forced exhalation, such as when blowing up a balloon, the diaphragm and chest muscles create a pressure of 60.0 mm Hg between the lungs and chest wall. What force in newtons does this pressure create on the 600 cm^2 surface area of the diaphragm?

2. You can chew through very tough objects with your incisors because they exert a large force on the small area of a pointed tooth. What pressure in pascals can you create by exerting a force of 500 Nwith your tooth on an area of 1.00 mm^2?

3. One way to force air into an unconscious person's lungs is to squeeze on a balloon appropriately connected to the subject. What force must you exert on the balloon with your hands to create a gauge pressure of 4.00 cm water, assuming you squeeze on an effective area of 50.0 cm^2 ?

4. Heroes in movies hide beneath water and breathe through a hollow reed (villains never catch on to this trick). In practice, you cannot inhale in this manner if your lungs are more than 60.0 cm below the surface. What is the maximum negative gauge pressure you can create in your lungs on dry land, assuming you can achieve -3.00 cm water pressure with your lungs 60.0 cm below the surface?

5. Gauge pressure in the fluid surrounding an infant's brain may rise as high as 85.0 mm Hg (5 to 12 mm Hg is normal), creating an outward force large enough to make the skull grow abnormally large. (a) Calculate this outward force in newtons on each side of an infant's skull if the effective area of each side is 70.0 cm^2. (b) What is the net force acting on the skull?

6. A full-term fetus typically has a mass of 3.50 kg. (a) What pressure does the weight of such a fetus create if it rests on the mother's bladder, supported on an area of 90.0cm^2? (b) Convert this pressure to millimeters of mercury and determine if it alone is great enough to trigger the micturition reflex (it will add to any pressure already existing in the bladder).

7. If the pressure in the esophagus is -2.00 mm Hgwhile that in the stomach is +20.0 mm Hg, to what height could stomach fluid rise in the esophagus, assuming a density of 1.10 g/mL? (This movement will not occur if the muscle closing the lower end of the esophagus is working properly.)

8. Pressure in the spinal fluid is measured as shown in Figure 4. If the pressure in the spinal fluid is 10.0 mm Hg: (a) What is the reading of the water manometer in cm water? (b) What is the reading if the person sits up, placing the top of the fluid 60 cm above the tap? The fluid density is 1.05 g/mL.

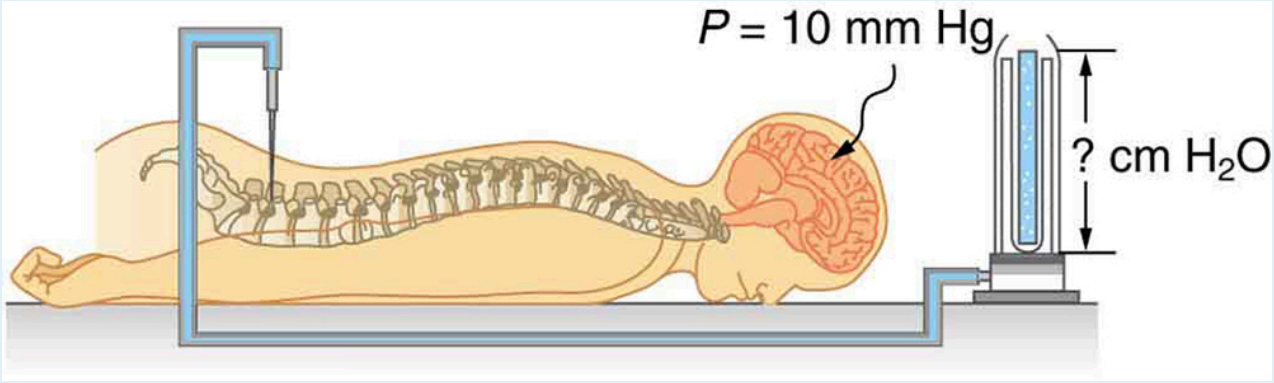


Figure 4. A water manometer used to measure pressure in the spinal fluid. The height of the fluid in the manometer is measured relative to the spinal column, and the manometer is open to the atmosphere. The measured pressure will be considerably greater if the person sits up.

9. Calculate the maximum force in newtons exerted by the blood on an aneurysm, or ballooning, in a major artery, given the maximum blood pressure for this person is 150 mm Hg and the effective area of the aneurysm is 20.0 cm^2. Note that this force is great enough to cause further enlargement and subsequently greater force on the ever-thinner vessel wall.

10. During heavy lifting, a disk between spinal vertebrae is subjected to a 5000-N compressional force. (a) What pressure is created, assuming that the disk has a uniform circular cross section 2.00 cm in radius? (b) What deformation is produced if the disk is 0.800 cm thick and has a Young's modulus of 1.5×10^9 N/m^2?

11. When a person sits erect, increasing the vertical position of their brain by 36.0 cm, the heart must continue to pump blood to the brain at the same rate. (a) What is the gain in gravitational potential energy for 100 mL of blood raised 36.0 cm? (b) What is the drop in pressure, neglecting any losses due to friction? (c) Discuss how the gain in gravitational potential energy and the decrease in pressure are related.

12. (a) How high will water rise in a glass capillary tube with a 0.500-mm radius? (b) How much gravitational potential energy does the water gain? (c) Discuss possible sources of this energy.

13. A negative pressure of 25.0 atm can sometimes be achieved with the device in Figure 5 before the water separates. (a) To what height could such a negative gauge pressure raise water? (b) How much would a steel wire of the same diameter and length as this capillary stretch if suspended from above?

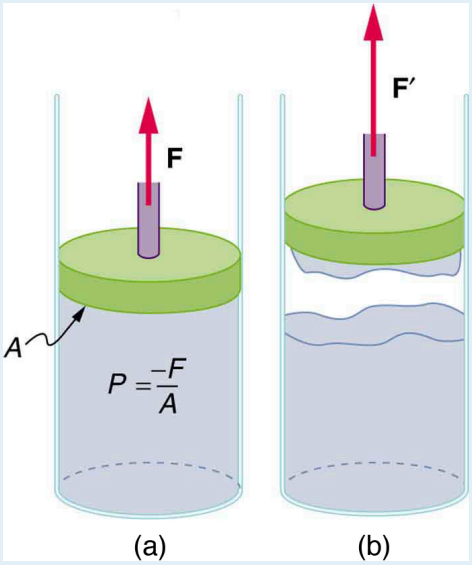


Figure 5. (a) When the piston is raised, it stretches the liquid slightly, putting it under tension and creating a negative absolute pressure $P = -F/A$ *(b) The liquid eventually separates, giving an experimental limit to negative pressure in this liquid.*

14. Suppose you hit a steel nail with a 0.500-kg hammer, initially moving at 15.0 m/s and brought to rest in 2.80 mm. (a) What average force is exerted on the nail? (b) How much is the nail compressed if it is 2.50 mm in diameter and 6.00-cm long? (c) What pressure is created on the 1.00-mm-diameter tip of the nail?

15. Calculate the pressure due to the ocean at the bottom of the Marianas Trench near the Philippines, given its depth is 11.0 km and assuming the density of sea water is constant all the way down. (b) Calculate the percent decrease in volume of sea water due to such a pressure, assuming its bulk modulus is the same as water and is constant. (c) What would be the percent increase in its density? Is the assumption of constant density valid? Will the actual pressure be greater or smaller than that calculated under this assumption?

16. The hydraulic system of a backhoe is used to lift a load as shown in Figure 6. (a) Calculate the force *F* the slave cylinder must exert to support the 400-kg load and the 150-kg brace and shovel. (b) What is the pressure in the hydraulic fluid if the slave cylinder is 2.50 cm in diameter? (c) What force would you have to exert on a lever with a mechanical advantage of 5.00 acting on a master cylinder 0.800 cm in diameter to create this pressure?

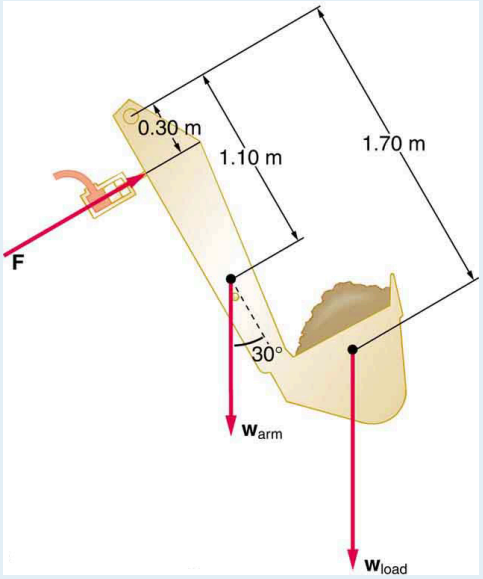


Figure 6. Hydraulic and mechanical lever systems are used in heavy machinery such as this back hoe.

17. Some miners wish to remove water from a mine shaft. A pipe is lowered to the water 90 m below, and a negative pressure is applied to raise the water. (a) Calculate the pressure needed to raise the water. (b) What is unreasonable about this pressure? (c) What is unreasonable about the premise?

18. You are pumping up a bicycle tire with a hand pump, the piston of which has a 2.00-cm radius. (a) What force in newtons must you exert to create a pressure of 6.90×10^5 Pa (b) What is unreasonable about this (a) result? (c) Which premises are unreasonable or inconsistent?

19. Consider a group of people trying to stay afloat after their boat strikes a log in a lake. Construct a problem in which you calculate the number of people that can cling to the log and keep their heads out of the water. Among the variables to be considered are the size and density of the log, and what is needed to keep a person's head and arms above water without swimming or treading water.

20. The alveoli in emphysema victims are damaged and effectively form larger sacs. Construct a problem in which you calculate the loss of pressure due to surface tension in the alveoli because of their larger average diameters. (Part of the lung's ability to expel air results from pressure created by surface tension in the alveoli.) Among the things to consider are the normal surface tension of the fluid lining the alveoli, the average alveolar radius in normal individuals and its average in emphysema sufferers.

Glossary

diastolic pressure:
minimum arterial blood pressure; indicator for the fluid balance

glaucoma:
condition caused by the buildup of fluid pressure in the eye

intraocular pressure:
fluid pressure in the eye

micturition reflex:
stimulates the feeling of needing to urinate, triggered by bladder pressure

systolic pressure:
maximum arterial blood pressure; indicator for the blood flow

Selected Solutions to Problems & Exercises

1. 479 N

3. 1.96 N

4. -63.0 cm H_2O

6. (a) 3.81 x 10^3 N/m (b) 28.7 mm Hg, which is sufficient to trigger micturition reflex

8. (a) 13.6 m water (b) 76.5 cm water

10. (a) 3.98×10^6 Pa (b) 2.1×10^{-3} cm

12. (a) 2.97 cm (b) 3.39×10^{-6} J (c) Work is done by the surface tension force through an effective distance $h/2$ to raise the column of water.

14. (a) 2.01×10^4 N (b) 1.17×10^{-3} m (c) 2.56×10^{10} N/m^2

16. (a) 1.38×10^4 N (b) 2.81×10^7 N/m^2 (c) 283 N

18. (a) 867 N (b) This is too much force to exert with a hand pump. (c) The assumed radius of the pump is too large; it would be nearly two inches in diameter—too large for a pump or even a master cylinder. The pressure is reasonable for bicycle tires.

3. Temperature, Kinetic Theory, and the Gas Laws

Introduction to Temperature, Kinetic Theory, and the Gas Laws

Lumen Learning

Figure 1. The welder's gloves and helmet protect him from the electric arc that transfers enough thermal energy to melt the rod, spray sparks, and burn the retina of an unprotected eye. The thermal energy can be felt on exposed skin a few meters away, and its light can be seen for kilometers. (credit: Kevin S. O'Brien/U.S. Navy)

Heat is something familiar to each of us. We feel the warmth of the summer Sun, the chill of a clear summer night, the heat of coffee after a winter stroll, and the cooling effect of our sweat. Heat transfer is maintained by temperature differences. Manifestations of *heat transfer*—the movement of heat energy from one place or material to another—are apparent throughout the universe. Heat from beneath Earth's surface is brought to the surface in flows of incandescent lava. The Sun warms Earth's surface and is the source of much of the energy we find on it. Rising levels of atmospheric carbon dioxide threaten to trap more of the Sun's energy, perhaps fundamentally altering the ecosphere. In space, supernovas explode, briefly radiating more heat than an entire galaxy does.

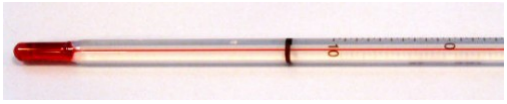


Figure 2. In a typical thermometer like this one, the alcohol, with a red dye, expands more rapidly than the glass containing it. When the thermometer's temperature increases, the liquid from the bulb is forced into the narrow tube, producing a large change in the length of the column for a small change in temperature. (credit: Chemical Engineer, Wikimedia Commons)

What is heat? How do we define it? How is it related to temperature? What are heat's effects? How is it related to other forms of energy and to work? We will find that, in spite of the richness of the phenomena, there is a small set of underlying physical principles that unite the subjects and tie them to other fields.

Temperature

Lumen Learning

Learning Objectives

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

- Define temperature.
- Convert temperatures between the Celsius, Fahrenheit, and Kelvin scales.
- Define thermal equilibrium.
- State the zeroth law of thermodynamics.

The concept of temperature has evolved from the common concepts of hot and cold. Human perception of what feels hot or cold is a relative one. For example, if you place one hand in hot water and the other in cold water, and then place both hands in tepid water, the tepid water will feel cool to the hand that was in hot water, and warm to the one that was in cold water. The scientific definition of temperature is less ambiguous than your senses of hot and cold. *Temperature* is operationally defined to be what we measure with a thermometer. (Many physical quantities are defined solely in terms of how they are measured. We shall see later how temperature is related to the kinetic energies of atoms and molecules, a more physical explanation.) Two accurate thermometers, one placed in hot water and the other in cold water, will show the hot water to have a higher temperature. If they are then placed in the tepid water, both will give identical readings (within measurement uncertainties). In this section, we discuss temperature, its measurement by thermometers, and its relationship to thermal equilibrium. Again, temperature is the quantity measured by a thermometer.

Misconception Alert: Human Perception vs. Reality

On a cold winter morning, the wood on a porch feels warmer than the metal of your bike. The wood and bicycle are in thermal equilibrium with the outside air, and are thus the same temperature. They *feel* different because of the difference in the way that they conduct heat away from your skin. The metal conducts heat away from your body faster than the wood does (see more about conductivity in Conduction). This is just one example demonstrating that the human sense of hot and cold is not determined by temperature alone.

Another factor that affects our perception of temperature is humidity. Most people feel much hotter on hot, humid days than on hot, dry days. This is because on humid days, sweat does not evaporate from the skin as efficiently as it does on dry days. It is the evaporation of sweat (or water from a sprinkler or pool) that cools us off.

Any physical property that depends on temperature, and whose response to temperature is reproducible, can be used as the basis of a thermometer. Because many physical properties depend on temperature, the variety of thermometers is remarkable. For example, volume increases with temperature for most substances. This property is the basis for the common alcohol thermometer, the old mercury thermometer, and the bimetallic strip (Figure 1).

Other properties used to measure temperature include electrical resistance and color and the emission of infrared radiation.

One example of electrical resistance and color is found in a plastic thermometer. Each of the six squares on the plastic (liquid crystal) thermometer in Figure 2 contains a film of a different heat-sensitive liquid crystal material Below 95ºF, all six squares are black. When the plastic thermometer is exposed to temperature that increases to 95ºF, the first liquid crystal square changes color. When the temperature increases above 96.8ºF the second liquid crystal square also changes color, and so forth.

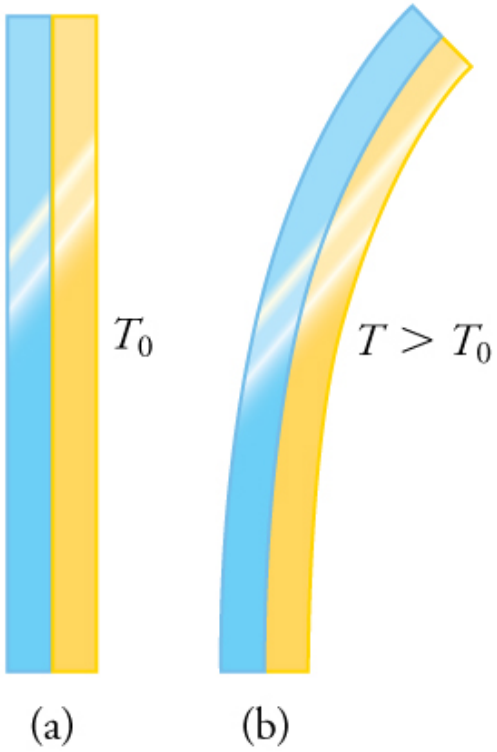


Figure 1. The curvature of a bimetallic strip depends on temperature. (a) The strip is straight at the starting temperature, where its two components have the same length. (b) At a higher temperature, this strip bends to the right, because the metal on the left has expanded more than the metal on the right.

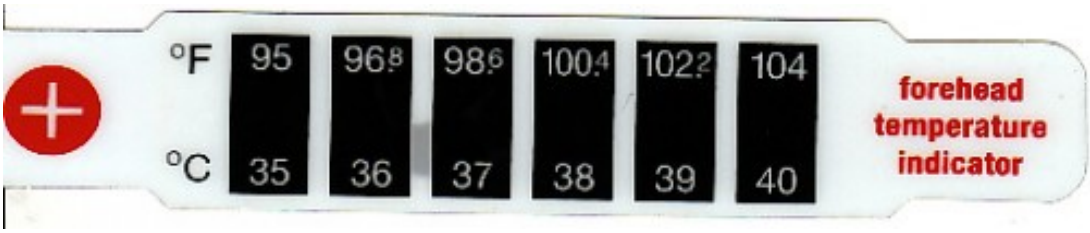


Figure 2. A plastic (liquid crystal) thermometer. (credit: Arkrishna, Wikimedia Commons)

An example of emission of radiation is shown in the use of a pyrometer (Figure 3). Infrared radiation (whose emission varies with temperature) from the vent in Figure 3 is measured and a temperature readout is quickly produced. Infrared measurements are also frequently used as a measure of body temperature. These modern thermometers, placed in the ear canal, are more accurate than alcohol thermometers placed under the tongue or in the armpit.

Figure 3. Fireman Jason Ormand uses a pyrometer to check the temperature of an aircraft carrier's ventilation system. (credit: Lamel J. Hinton/U.S. Navy)

Temperature Scales

Thermometers are used to measure temperature according to well-defined scales of measurement, which use pre-defined reference points to help compare quantities. The three most common temperature scales are the Fahrenheit, Celsius, and Kelvin scales. A temperature scale can be created by identifying two easily reproducible temperatures. The freezing and boiling temperatures of water at standard atmospheric pressure are commonly used.

The *Celsius* scale (which replaced the slightly different *centigrade* scale) has the freezing point of water at 0ºC and the boiling point at 100ºC. Its unit is the *degree Celsius*(ºC). On the *Fahrenheit* scale (still the most frequently used in the United States), the freezing point of water is at 32ºF and the boiling point is at 212ºF. The unit of temperature on this scale is the *degree Fahrenheit*(ºF). Note that a temperature difference of one degree Celsius is greater than a temperature difference of one degree Fahrenheit. Only 100 Celsius degrees span the same range as 180 Fahrenheit degrees, thus one degree on the Celsius scale is 1.8 times larger than one degree on the Fahrenheit scale 180/100=9/5.

The *Kelvin* scale is the temperature scale that is commonly used in science. It is an *absolute temperature* scale defined to have 0 K at the lowest possible temperature, called *absolute zero*. The official temperature unit on this scale is the *kelvin*, which is abbreviated K, and is not accompanied by a degree sign. The freezing and boiling points of water are 273.15 K and 373.15 K, respectively. Thus, the magnitude of temperature differences is the same in units of kelvins and degrees Celsius. Unlike other temperature scales, the Kelvin scale is an absolute scale. It is used extensively in scientific work because a number of physical quantities, such as the volume of an ideal gas, are directly related to absolute temperature. The kelvin is the SI unit used in scientific work.

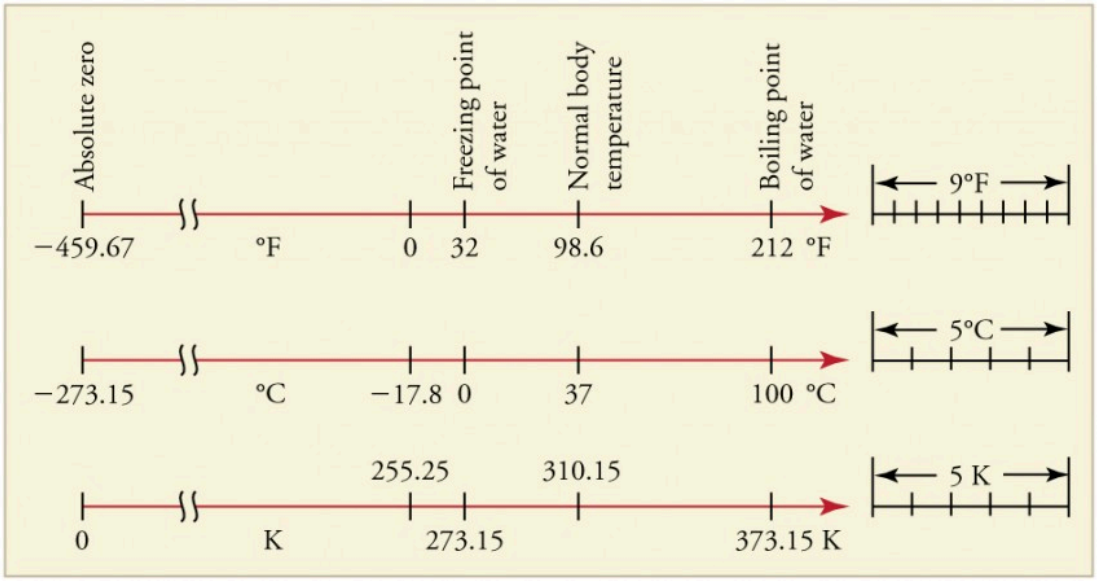


Figure 4. Relationships between the Fahrenheit, Celsius, and Kelvin temperature scales, rounded to the nearest degree. The relative sizes of the scales are also shown.

The relationships between the three common temperature scales is shown in Figure 4. Temperatures on these scales can be converted using the equations in Table 1.

Table 1. Temperature Conversions

To convert from . . .	Use this equation . . .	Also written as . . .
Celsius to Fahrenheit	$T\,(^\circ\mathrm{F}) = \frac{9}{5}T\,(^\circ\mathrm{C}) + 32$	$T_{^\circ\mathrm{F}} = \frac{9}{5}T_{^\circ\mathrm{C}} + 32$
Fahrenheit to Celsius	$T\,(^\circ\mathrm{C}) = \frac{5}{9}\,(T\,(^\circ\mathrm{F}) - 32)$	$T_{^\circ\mathrm{C}} = \frac{5}{9}\,(T_{^\circ\mathrm{F}} - 32)$
Celsius to Kelvin	$T(\mathrm{K}) = T(^\circ\mathrm{C}) + 273.15$	$T_\mathrm{K} = T_{^\circ\mathrm{C}} + 273.15$
Kelvin to Celsius	$T(^\circ\mathrm{C}) = T(\mathrm{K}) - 273.15$	$T_{^\circ\mathrm{C}} = T_\mathrm{K} - 273.15$
Fahrenheit to Kelvin	$T\,(K) = \frac{5}{9}\,(T\,(^\circ\mathrm{F}) - 32) + 273.15$	$T_K = \frac{5}{9}\,(T_{^\circ\mathrm{F}} - 32) + 273.15$
Kelvin to Fahrenheit	$T\,(^\circ\mathrm{F}) = \frac{9}{5}\,(T\,(K) - 273.15) + 32$	$T_{^\circ\mathrm{F}} = \frac{9}{5}\,(T_K - 273.15) + 32$

Notice that the conversions between Fahrenheit and Kelvin look quite complicated. In fact, they are simple combinations of the conversions between Fahrenheit and Celsius, and the conversions between Celsius and Kelvin.

Example 1. Converting between Temperature Scales: Room Temperature

"Room temperature" is generally defined to be 25ºC.

1. What is room temperature in ºF?
2. What is it in K?

Strategy

To answer these questions, all we need to do is choose the correct conversion equations and plug in the known values.

Solution for Part 1

$$T_{\circ\mathrm{F}} = \frac{9}{5}T_{\circ\mathrm{C}} + 32$$

1. Choose the right equation. To convert from ºC to ºF, use the equation .
$$T_{\circ\mathrm{F}} = \frac{9}{5}25^{\circ}\mathrm{C} + 32 = 77^{\circ}\mathrm{F}$$
2. Plug the known value into the equation and solve:

Solution for Part 2

1. Choose the right equation. To convert from ºC to K, use the equation $T_K = T_{ºC} + 273.15$
2. Plug the known value into the equation and solve: T_K = 25ºC + 273.15 = 298 K.

Example 2. Converting between Temperature Scales: the Reaumur Scale

The Reaumur scale is a temperature scale that was used widely in Europe in the eighteenth and nineteenth centuries. On the Reaumur temperature scale, the freezing point of water is 0ºR and the boiling temperature is 80ºR. If "room temperature" is 25ºC on the Celsius scale, what is it on the Reaumur scale?

Strategy

To answer this question, we must compare the Reaumur scale to the Celsius scale. The difference between the freezing point and boiling point of water on the Reaumur scale is 80ºR. On the Celsius scale it is 100ºC. Therefore 100º C=80ºR. Both scales start at 0 º for freezing, so we can derive a simple formula to convert between temperatures on the two scales.

Solution

$$T_{\circ\mathrm{R}} = \frac{0.8^{\circ}\mathrm{R}}{^{\circ}\mathrm{C}} \times T_{\circ\mathrm{C}}$$

1. Derive a formula to convert from one scale to the other:
$$T_{\circ\mathrm{R}} = \frac{0.8^{\circ}\mathrm{R}}{^{\circ}\mathrm{C}} \times 25^{\circ}\mathrm{C} = 20^{\circ}\mathrm{R}$$
2. Plug the known value into the equation and solve:

Temperature Ranges in the Universe

Figure 6 shows the wide range of temperatures found in the universe. Human beings have been known to survive with body temperatures within a small range, from 24ºC to 44ºC (75ºF to 111ºF). The average normal body temperature is usually given as 37.0ºC (98.6ºF), and variations in this temperature can indicate a medical condition: a fever, an infection, a tumor, or circulatory problems (see Figure 5).

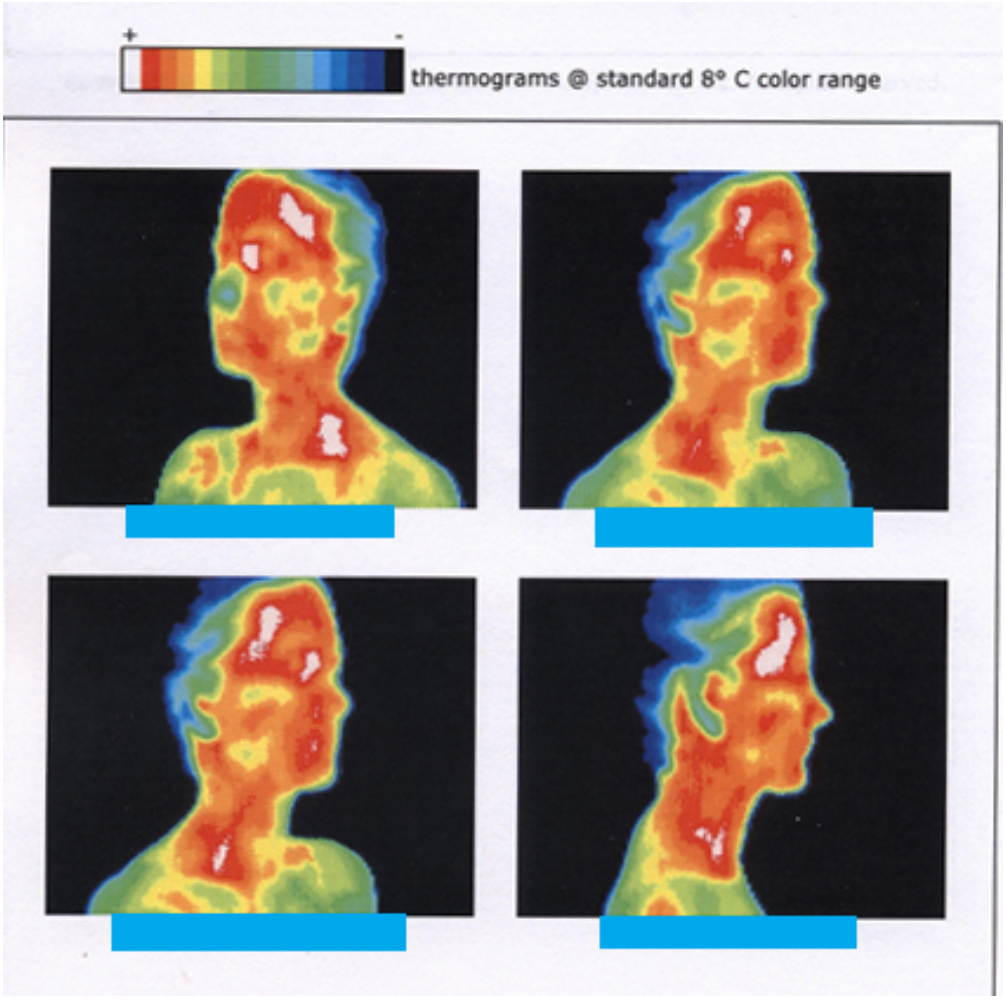


Figure 5. This image of radiation from a person's body (an infrared thermograph) shows the location of temperature abnormalities in the upper body. Dark blue corresponds to cold areas and red to white corresponds to hot areas. An elevated temperature might be an indication of malignant tissue (a cancerous tumor in the breast, for example), while a depressed temperature might be due to a decline in blood flow from a clot. In this case, the abnormalities are caused by a condition called hyperhidrosis. (credit: Porcelina81, Wikimedia Commons)

The lowest temperatures ever recorded have been measured during laboratory experiments: 4.5×10^{-10} K at the Massachusetts Institute of Technology (USA), and 1.0×10^{-10} K at Helsinki University of Technology (Finland). In comparison, the coldest recorded place on Earth's surface is Vostok, Antarctica at 183 K (–89ºC), and the coldest place (outside the lab) known in the universe is the Boomerang Nebula, with a temperature of 1 K.

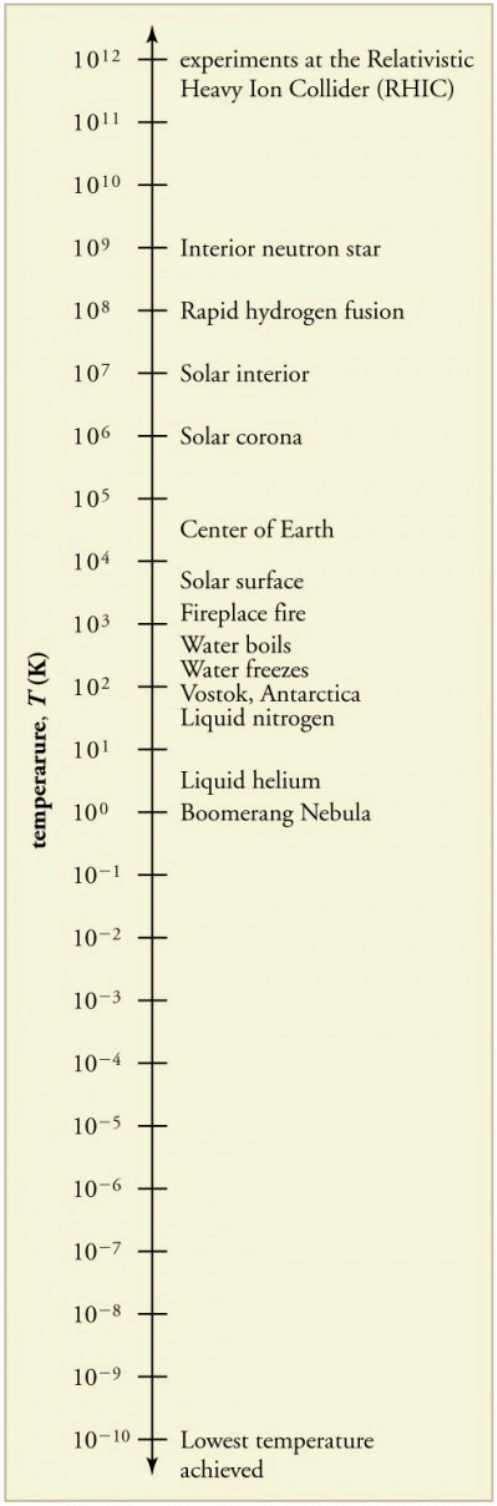


Figure 6. Each increment on this logarithmic scale indicates an increase by a factor of ten, and thus illustrates the tremendous range of temperatures in nature. Note that zero on a logarithmic scale would occur off the bottom of the page at infinity.

Making Connections: Absolute Zero

What is absolute zero? Absolute zero is the temperature at which all molecular motion has ceased. The concept of absolute zero arises from the behavior of gases. Figure 7 shows how the pressure of gases at a constant volume decreases as temperature decreases. Various scientists have noted that the pressures of gases extrapolate to zero at the same temperature, –273.15ºC. This extrapolation implies that there is a lowest temperature. This temperature is called *absolute zero*. Today we know that most gases first liquefy and then freeze, and it is not actually possible to reach absolute zero. The numerical value of absolute zero temperature is –273.15ºC or 0 K.

Thermal Equilibrium and the Zeroth Law of Thermodynamics

Thermometers actually take their *own* temperature, not the temperature of the object they are measuring. This raises the question of how we can be certain that a thermometer measures the temperature of the object with which it is in contact. It is based on the fact that any two systems placed in *thermal contact* (meaning heat transfer can occur between them) will reach the same temperature. That is, heat will flow from the hotter object to the cooler one until they have exactly the same temperature. The objects are then in *thermal equilibrium*, and no further changes will occur. The systems interact and change because their temperatures differ, and the changes stop once their temperatures are the same. Thus, if enough time is allowed for this transfer of heat to run its course, the temperature a thermometer registers *does* represent the system with which it is in thermal equilibrium. Thermal equilibrium is established when two bodies are in contact with each other and can freely exchange energy.

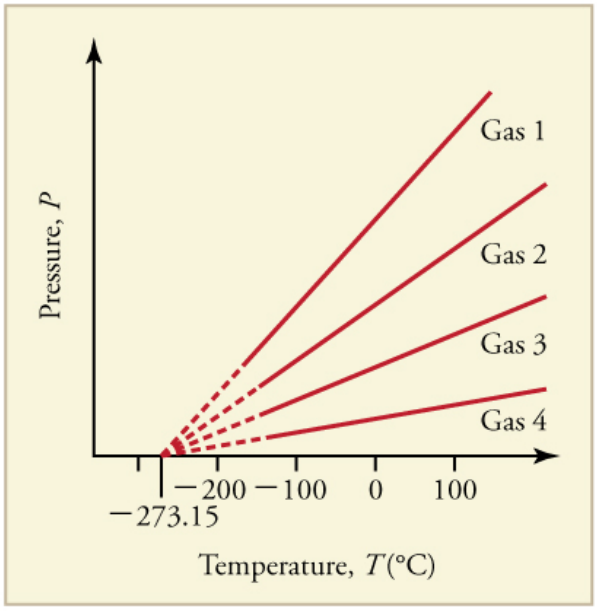


Figure 7. Graph of pressure versus temperature for various gases kept at a constant volume. Note that all of the graphs extrapolate to zero pressure at the same temperature.

Furthermore, experimentation has shown that if two systems, A and B, are in thermal equilibrium with each another, and B is in thermal equilibrium with a third system C, then A is also in thermal equilibrium with C. This conclusion may seem obvious, because all three have the same temperature, but it is basic to thermodynamics. It is called the *zeroth law of thermodynamics*.

The Zeroth Law of Thermodynamics

If 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.

This law was postulated in the 1930s, after the first and second laws of thermodynamics had been developed and named. It is called the *zeroth law* because it comes logically before the first and second laws (discussed in Thermodynamics). An example of this law in action is seen in babies in incubators: babies in incubators normally have very few clothes on, so to an observer they look as if they may not be warm enough. However, the temperature of the air, the cot, and the baby is the same, because they are in thermal equilibrium, which is accomplished by maintaining air temperature to keep the baby comfortable.

Check Your Understanding

Does the temperature of a body depend on its size?

Solution

No, the system can be divided into smaller parts each of which is at the same temperature. We say that the temperature is an *intensive* quantity. Intensive quantities are independent of size.

Section Summary

- Temperature is the quantity measured by a thermometer.
- Temperature is related to the average kinetic energy of atoms and molecules in a system.
- Absolute zero is the temperature at which there is no molecular motion.
- There are three main temperature scales: Celsius, Fahrenheit, and Kelvin.
- Temperatures on one scale can be converted to temperatures on another scale using the following equations:
  - $T_{^\circ\mathrm{F}} = \frac{9}{5}T_{^\circ\mathrm{C}} + 32$
  - $T_{^\circ\mathrm{C}} = \frac{5}{9}\left(T_{^\circ\mathrm{F}} - 32\right)$
  - $T_{\mathrm{K}} = T_{^\circ\mathrm{C}} + 273.15$
  - $T_{^\circ\mathrm{C}} = T_{\mathrm{K}} - 273.15$
- Systems are in thermal equilibrium when they have the same temperature.

Thermal equilibrium occurs when two bodies are in contact with each other and can freely exchange energy.
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.

Conceptual Questions

1. What does it mean to say that two systems are in thermal equilibrium?
2. Give an example of a physical property that varies with temperature and describe how it is used to measure temperature.
3. When a cold alcohol thermometer is placed in a hot liquid, the column of alcohol goes down slightly before going up. Explain why.
4. If you add boiling water to a cup at room temperature, what would you expect the final equilibrium temperature of the unit to be? You will need to include the surroundings as part of the system. Consider the zeroth law of thermodynamics.

Problems & Exercises

1. What is the Fahrenheit temperature of a person with a 39.0ºC fever?
2. Frost damage to most plants occurs at temperatures of 28.0ºF or lower. What is this temperature on the Kelvin scale?
3. To conserve energy, room temperatures are kept at 68.0ºF in the winter and 78.0ºF in the summer. What are these temperatures on the Celsius scale?
4. A tungsten light bulb filament may operate at 2900 K. What is its Fahrenheit temperature? What is this on the Celsius scale?
5. The surface temperature of the Sun is about 5750 K. What is this temperature on the Fahrenheit scale?
6. One of the hottest temperatures ever recorded on the surface of Earth was 134ºF in Death Valley, CA. What is this temperature in Celsius degrees? What is this temperature in Kelvin?
7. (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 there is a 40.0ºF decrease in temperature? (b) Show that any change in temperature in Fahrenheit degrees is nine-fifths the change in Celsius degrees.
8. (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?

Glossary

temperature: the quantity measured by a thermometer

Celsius scale: temperature scale in which the freezing point of water is 0ºC and the boiling point of water is 100ºC

degree Celsius: unit on the Celsius temperature scale

Fahrenheit scale: temperature scale in which the freezing point of water is 32ºF and the boiling point of water is 212ºF

degree Fahrenheit: unit on the Fahrenheit temperature scale

Kelvin scale: temperature scale in which 0 K is the lowest possible temperature, representing absolute zero

absolute zero: the lowest possible temperature; the temperature at which all molecular motion ceases

thermal equilibrium: the condition in which heat no longer flows between two objects that are in contact; the two objects have the same temperature

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

Selected Solutions to Problems & Exercises

1. 102ºF

3. 20.0ºC and 25.6ºC

5. 9890ºF

7. (a) 22.2ºC; (b)

$$\begin{aligned} \Delta T\,(^\circ\mathrm{F}) &= T_2\,(^\circ\mathrm{F}) - T_1\,(^\circ\mathrm{F}) \\ &= \tfrac{9}{5}T_2\,(^\circ\mathrm{C}) + 32.0^\circ - \left(\tfrac{9}{5}T_1\,(^\circ\mathrm{C}) + 32.0^\circ\right) \\ &= \tfrac{9}{5}\left(T_2\,(^\circ\mathrm{C}) - T_1\,(^\circ\mathrm{C})\right) = \tfrac{9}{5}\Delta T\,(^\circ\mathrm{C}) \end{aligned}$$

Thermal Expansion of Solids and Liquids

Lumen Learning

Learning Objectives

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

- Define and describe thermal expansion.
- Calculate the linear expansion of an object given its initial length, change in temperature, and coefficient of linear expansion.
- Calculate the volume expansion of an object given its initial volume, change in temperature, and coefficient of volume expansion.
- Calculate thermal stress on an object given its original volume, temperature change, volume change, and bulk modulus.

The expansion of alcohol in a thermometer is one of many commonly encountered examples of *thermal expansion*, the change in size or volume of a given mass with temperature. Hot air rises because its volume increases, which causes the hot air's density to be smaller than the density of surrounding air, causing a buoyant (upward) force on the hot air. The same happens in all liquids and gases, driving natural heat transfer upwards in homes, oceans, and weather systems. 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.

Figure 1. Thermal expansion joints like these in the Auckland Harbour Bridge in New Zealand allow bridges to change length without buckling. (credit: Ingolfson, Wikimedia Commons)

What are the basic properties of thermal expansion? First, thermal expansion is clearly related to temperature change. The greater the temperature change, the more a bimetallic strip will bend. Second, it depends on the material. In a thermometer, for example, the expansion of alcohol is much greater than the expansion of the glass containing it.

What is the underlying cause of thermal expansion? As is discussed in Kinetic Theory: Atomic and Molecular Explanation of Pressure and Temperature, an increase in temperature implies an increase in the kinetic energy of the individual atoms. In a solid, unlike in a gas, the atoms or molecules are closely packed together, but their kinetic energy (in the form of small, rapid vibrations) pushes neighboring atoms or molecules apart from each other. This neighbor-to-neighbor pushing results in a slightly greater distance, on average, between neighbors, and adds up to a larger size for the whole body. For most

substances under ordinary conditions, there is no preferred direction, and an increase in temperature will increase the solid's size by a certain fraction in each dimension.

Linear Thermal Expansion—Thermal Expansion in One Dimension

The change in length ΔL is proportional to length L. The dependence of thermal expansion on temperature, substance, and length is summarized in the equation $\Delta L = \alpha L \Delta T$,where ΔL is the change in length L, ΔT is the change in temperature, and α is the *coefficient of linear expansion*, which varies slightly with temperature.

Table 1 lists representative values of the coefficient of linear expansion, which may have units of 1/ºC or 1/K. Because the size of a kelvin and a degree Celsius are the same, both α and ΔT can be expressed in units of kelvins or degrees Celsius. The equation $\Delta L = \alpha L \Delta T$ is accurate for small changes in temperature and can be used for large changes in temperature if an average value of α is used.

Table 1. Thermal Expansion Coefficients at 20ºC[1]

Material	Coefficient of linear expansion α(1/ºC)	Coefficient of volume expansion β(1/ºC)
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}
Petrol		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}

Example 1. Calculating Linear Thermal Expansion: The Golden Gate Bridge

The main span of San Francisco's Golden Gate Bridge is 1275 m long at its coldest. The bridge is exposed to

1. Values for liquids and gases are approximate.

temperatures ranging from –15ºC to 40ºC. 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 1, and note that the change in temperature, ΔT, is 55ºC.

Solution

Plug all of the known values into the equation to solve for ΔL.

$$\Delta L = \alpha L \Delta L = \left(\frac{12 \times 10^{-6}}{^\circ\text{C}}\right)(1275\ \text{m})(55^\circ\text{C}) = 0.84\ \text{m}$$

Discussion

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

Objects expand in all dimensions, as illustrated in Figure 2. That is, their areas and volumes, as well as their lengths, increase with temperature. Holes 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 plug was still in place. The plug would get bigger, and so the hole must get bigger too. (Think of the ring of neighboring atoms or molecules on the wall of the hole as pushing each other farther apart as temperature increases. Obviously, the ring of neighbors must get slightly larger, so the hole gets slightly larger).

Thermal Expansion in Two Dimensions

For small temperature changes, the change in area ΔA is given by $\Delta A = 2\alpha A \Delta T$, 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.

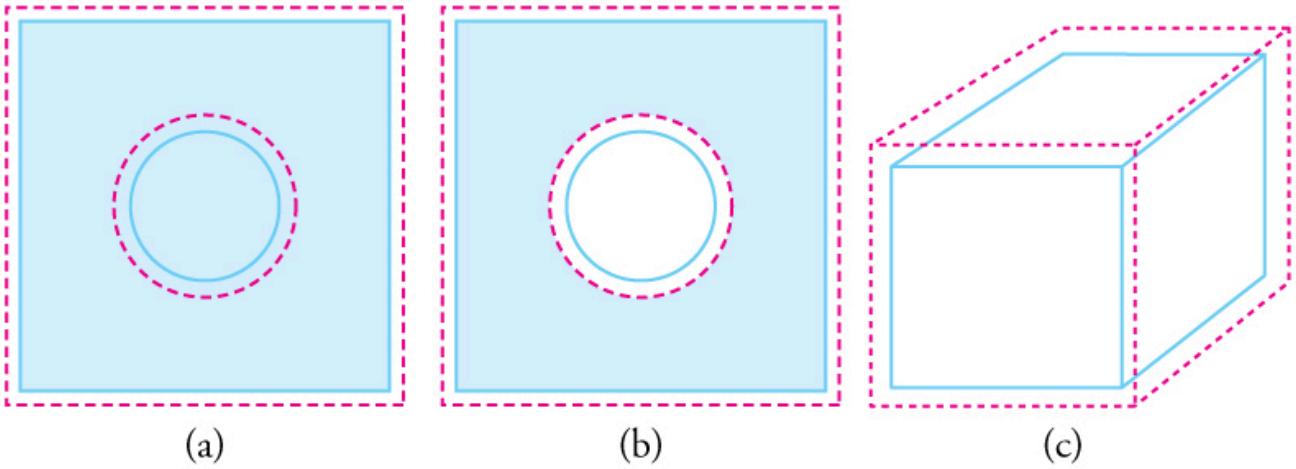


Figure 2. In general, objects expand in all directions as temperature increases. In these drawings, the original boundaries of the objects are shown with solid lines, and the expanded boundaries with dashed lines. (a) Area increases because both length and width increase. The area of a circular plug 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 change in volume ΔV is very nearly $\Delta V = 3\alpha V \Delta T$. This equation is usually written as $\Delta V = \beta V \Delta T$, where β is the *coefficient of volume expansion* and $\beta \approx 3\alpha$. Note that the values of β in Table 1 are almost exactly equal to 3α.

In general, objects will expand with increasing temperature. Water is the most important exception to this rule. Water expands with increasing temperature (its density *decreases*) when it is at temperatures greater than 4ºC (40ºF). However, it expands with *decreasing* temperature when it is between +4ºC and 0ºC (40ºF to 32ºF). Water is densest at +4ºC. (See Figure 3.) Perhaps the most striking effect of this phenomenon is the freezing of water in a pond. When water near the surface cools down to 4ºC it is denser than the remaining water and thus will sink to the bottom. This "turnover" results in a layer of warmer water near the surface, which is then cooled. Eventually the pond has a uniform temperature of 4ºC. If the temperature in the surface layer drops below 4ºC, the water is less dense than the water below, and thus stays near the top. As a result, the pond surface can completely freeze over. The ice on top of liquid water provides an insulating layer from winter's harsh exterior air temperatures. Fish and other aquatic life can survive in 4ºC water beneath ice, due to this unusual characteristic of water. It also produces circulation of water in the pond that is necessary for a healthy ecosystem of the body of water.

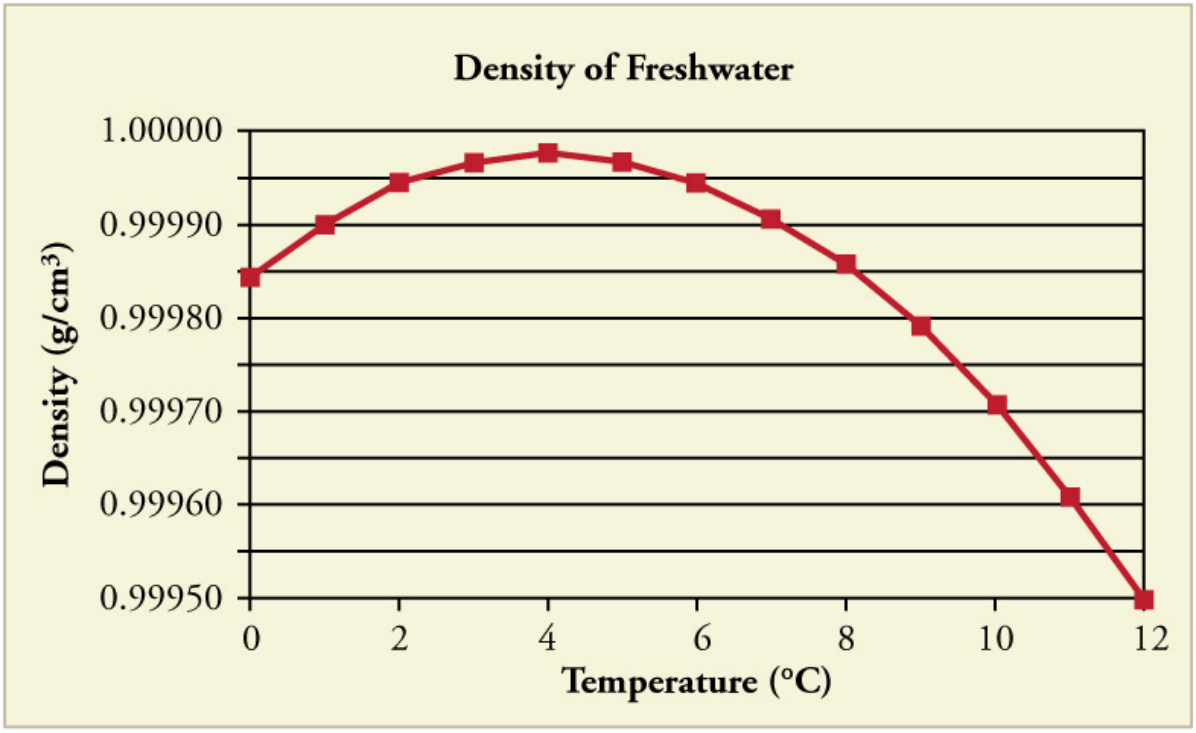


Figure 3. The density of water as a function of temperature. Note that the thermal expansion is actually very small. The maximum density at +4ºC is only 0.0075% greater than the density at 2ºC, and 0.012% greater than that at 0ºC.

Making Connections: Real-World Connections—Filling the Tank

Differences in the thermal expansion of materials can lead to interesting effects at the gas station. One example is the dripping of gasoline from a freshly filled tank on a hot day. Gasoline starts out at the temperature of the ground under the gas station, which is cooler than the air temperature above. The gasoline cools the steel tank when it is filled. Both gasoline and steel tank expand as they warm to air temperature, but gasoline expands much more than steel, and so it may overflow.

This difference in expansion can also cause problems when interpreting the gasoline gauge. The actual amount (mass) of gasoline left in the tank when the gauge hits "empty" is a lot 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. If you are used to getting another 40 miles on "empty" in the winter, beware—you will probably run out much more quickly in the summer.

Figure 4. Because the gas expands more than the gas tank with increasing temperature, you can't drive as many miles on "empty" in the summer as you can in the winter. (credit: Hector Alejandro, Flickr)

Example 2. Calculating Thermal Expansion: Gas vs. Gas Tank

Suppose your 60.0-L (15.9-gal) steel gasoline tank is full of gas, so both the tank and the gasoline have a temperature of 15.0ºC. How much gasoline has spilled by the time they warm to 35.0ºC?

Strategy

The tank and gasoline increase in volume, but the gasoline increases more, so the amount spilled is the difference in their volume changes. (The gasoline tank can be treated as solid steel.) We can use the equation for volume expansion to calculate the change in volume of the gasoline and of the tank.

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.)

$$\begin{aligned} V_{\text{spill}} &= (\beta_{\text{gas}} - \beta_{\text{s}})\, V \Delta T \\ &= \left[(950 - 35) \times 10^{-6}/^{\circ}\text{C}\right] (60.0\text{L})\, (20.0^{\circ}\text{C}) \\ &= 1.10\text{L} \end{aligned}$$

Discussion

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 in the chapter Heat and Heat Transfer Methods.

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 being compressed with extremely large forces. To avoid rupturing rigid containers, these containers have air gaps, which allow them to expand and contract without stressing them.

Thermal Stress

Thermal stress is created by thermal expansion or contraction (see Elasticity: Stress and Strain for a discussion of stress and strain). Thermal stress can be destructive, such as when expanding gasoline ruptures a tank. It can also be useful, for example, when two parts are joined together by heating one in manufacturing, then slipping it over the other and allowing the combination to cool. Thermal stress can explain many phenomena, such as the weathering of rocks and pavement by the expansion of ice when it freezes.

Example 3. Calculating Thermal Stress: Gas Pressure

What pressure would be created in the gasoline tank considered in Example 2, if the gasoline increases in temperature from 15.0ºC to 35.0ºC without being allowed to expand? Assume that the bulk modulus *B* for gasoline is 1.00×10^9 N/m^2.

Strategy

To solve this problem, we must use the following equation, which relates a change in volume ΔV to pressure:

$$\Delta V = \frac{1}{B}\frac{F}{A}V_0$$

where

$$\frac{F}{A}$$

is pressure, V_0 is the original volume, and *B* is the bulk modulus of the material involved. We will use the amount spilled in Example 2 as the change in volume, ΔV.

Solution

$$P = \frac{F}{A} = \frac{\Delta V}{V_0}B$$

1. Rearrange the equation for calculating pressure: .
2. Insert the known values. The bulk modulus for gasoline is $B = 1.00 \times 10^9$ N/m^2. In the previous example, the change in volume $\Delta V = 1.10$ L is the amount that would spill. Here, $V_0 = 60.0$ L is the original volume of the gasoline. Substituting these values into the equation, we obtain
$$P = \frac{1.10\ \text{L}}{60.0\ \text{L}}\left(1.00 \times 10^9\ \text{Pa}\right) = 1.83 \times 10^7\ \text{Pa}$$
.

Discussion

This pressure is about 2500 lb/in^2, *much* more than a gasoline tank can handle.

Forces and pressures created by thermal stress are typically as great as that in the example above. Railroad tracks and roadways can buckle on hot days if they lack sufficient expansion joints. (See Figure 5.) Power lines sag more in the summer than in the winter, and will snap in cold weather if there is insufficient slack. Cracks open and close in plaster walls as a house warms and cools. Glass cooking pans will crack if cooled rapidly or unevenly, because of differential contraction and the stresses it creates. (Pyrex® is less susceptible because of its small coefficient of thermal expansion.) Nuclear reactor pressure vessels are threatened by overly rapid cooling, and although none have failed, several have been cooled faster than considered desirable. Biological cells are ruptured when foods are frozen, detracting from their taste. Repeated thawing and freezing accentuate the damage. Even the oceans can be affected. A significant portion of the rise in sea level that is resulting from global warming is due to the thermal expansion of sea water.

Metal is regularly used in the human body for hip and knee implants. Most implants need to be replaced over time because, among other things, metal does not bond with bone. Researchers are trying to find better metal coatings that would allow metal-to-bone bonding. One challenge is to find a coating that has an expansion coefficient similar to that of metal. If the expansion coefficients are too different, the thermal stresses during the manufacturing process lead to cracks at the coating-metal interface.

Figure 5. Thermal stress contributes to the formation of potholes. (credit: Editor5807, Wikimedia Commons)

Another example of thermal stress is found in the mouth. Dental fillings can expand differently from tooth enamel. It can give pain when eating ice cream or having a hot drink. Cracks might occur in the filling. Metal fillings (gold, silver, etc.) are being replaced by composite fillings (porcelain), which have smaller coefficients of expansion, and are closer to those of teeth.

Check Your Understanding

Two blocks, A and B, are made of the same material. Block A has dimensions $l \times w \times h = L \times 2L \times L$ and Block B has dimensions $2L \times 2L \times 2L$. If the temperature changes, what is

1. the change in the volume of the two blocks,
2. the change in the cross-sectional area $l \times w$, and
3. the change in the height h of the two blocks?

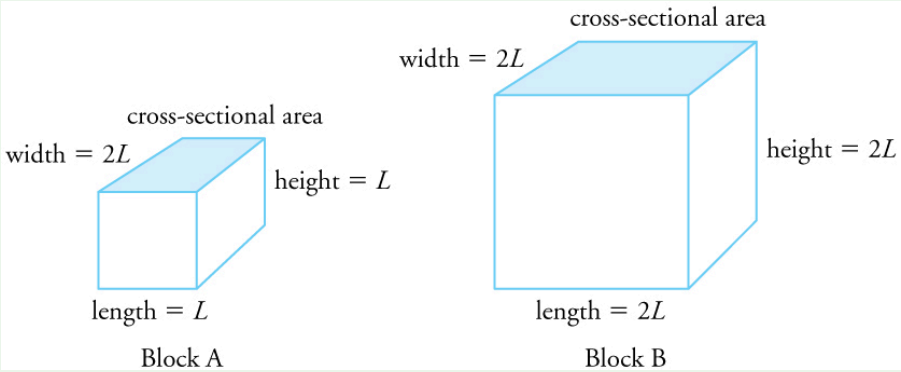


Figure 6.

Solution

1. The change in volume is proportional to the original volume. Block A has a volume of $L \times 2L \times L = 2L^3$. Block B has a volume of $2L \times 2L \times 2L = 8L^3$, which is 4 times that of Block A. Thus the change in volume of Block B should be 4 times the change in volume of Block A.
2. The change in area is proportional to the area. The cross-sectional area of Block A is $L \times 2L =$

$2L^2$, while that of Block B is $2L \times 2L = 4L^2$. Because cross-sectional area of Block B is twice that of Block A, the change in the cross-sectional area of Block B is twice that of Block A.

3. The change in height is proportional to the original height. Because the original height of Block B is twice that of A, the change in the height of Block B is twice that of Block A.

Section Summary

- Thermal expansion is the increase, or decrease, of the size (length, area, or volume) of a body due to a change in temperature.
- Thermal expansion is large for gases, and relatively small, but not negligible, for liquids and solids.
- Linear thermal expansion is $\Delta L = \alpha L \Delta T$, where ΔL is the change in length L, ΔT is the change in temperature, and α is the coefficient of linear expansion, which varies slightly with temperature.
- The change in area due to thermal expansion is $\Delta A = 2\alpha A \Delta T$, where ΔA is the change in area.
- The change in volume due to thermal expansion is $\Delta V = \beta V \Delta T$, where β is the coefficient of volume expansion and $\beta \approx 3\alpha$. Thermal stress is created when thermal expansion is constrained.

Conceptual Questions

1. Thermal stresses caused by uneven cooling can easily break glass cookware. Explain why Pyrex®, a glass with a small coefficient of linear expansion, is less susceptible.
2. Water expands significantly when it freezes: a volume increase of about 9% occurs. As a result of this expansion and because of the formation and growth of crystals as water freezes, anywhere from 10% to 30% of biological cells are burst when animal or plant material is frozen. Discuss the implications of this cell damage for the prospect of preserving human bodies by freezing so that they can be thawed at some future date when it is hoped that all diseases are curable.
3. 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.
4. 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.
5. Liquids and solids expand with increasing temperature, because the kinetic energy of a body's atoms and molecules increases. Explain why some materials shrink with increasing temperature.

Problems & Exercises

1. The height of the Washington Monument is measured to be 170 m on a day when the temperature is 35.0ºC. What will its height be on a day when the temperature falls to –10.0ºC? Although the monument is made of limestone, assume that its thermal coefficient of expansion is the same as marble's.
2. How much taller does the Eiffel Tower become at the end of a day when the temperature has increased by 15ºC? Its original height is 321 m and you can assume it is made of steel.
3. What is the change in length of a 3.00-cm-long column of mercury if its temperature changes from 37.0ºC to 40.0ºC, assuming the mercury is unconstrained?
4. How large an expansion gap should be left between steel railroad rails if they may reach a maximum temperature 35.0ºC greater than when they were laid? Their original length is 10.0 m.
5. You are looking to purchase 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 m × 30 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 20ºC above normal?
6. Global warming will produce rising sea levels partly due to melting ice caps but also 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 1.00ºC. Note that this calculation is only approximate because ocean warming is not uniform with depth.
7. Show that 60.0 L of gasoline originally at 15.0ºC will expand to 61.1 L when it warms to 35.0ºC, as claimed in Example 2.
8. (a) Suppose a meter stick made of steel and one made of invar (an alloy of iron and nickel) are the same length at 0ºC. What is their difference in length at 22.0ºC? (b) Repeat the calculation for two 30.0-m-long surveyor's tapes.
9. (a) If a 500-mL glass beaker is filled to the brim with ethyl alcohol at a temperature of 5.00ºC, how much will overflow when its temperature reaches 22.0ºC? (b) How much less water would overflow under the same conditions?
10. Most automobiles 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 10.0ºC. What volume of radiator fluid will overflow when the radiator and fluid reach their 95.0ºC operating temperature, given that the fluid's volume coefficient of expansion is $\beta = 400 \times 10^{-6}/^{\circ}\text{C}$? Note that this coefficient is approximate, because most car radiators have operating temperatures of greater than 95.0ºC.
11. 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 350cm3 of coffee is in a 7.00-cm-diameter cup and decreases in temperature from 95.0ºC to 45.0ºC. (Most of the drop in level is actually due to escaping bubbles of air.)
12. (a) The density of water at 0ºC is very nearly 1000kg/m3 (it is actually 999.84 kg/m^3), whereas the density of ice at 0ºC is 917 kg/m^3. Calculate the pressure necessary to keep ice from expanding when it freezes, neglecting the effect such a large pressure would have on the freezing temperature. (This problem gives you only an indication of how large the forces associated with

freezing water might be.) (b) What are the implications of this result for biological cells that are frozen?

13. Show that $\beta \approx 3\alpha$, by calculating the change in volume ΔV of a cube with sides of length L.

Glossary

thermal expansion: the change in size or volume of an object with change in temperature

coefficient of linear expansion: α, the change in length, per unit length, per 1ºC change in temperature; a constant used in the calculation of linear expansion; the coefficient of linear expansion depends on the material and to some degree on the temperature of the material

coefficient of volume expansion: β, the change in volume, per unit volume, per 1ºC change in temperature

thermal stress: stress caused by thermal expansion or contraction

Selected Answers to Problems & Exercises

1. 169.98 m

3. 5.4×10^{-6} m

5. Because the area gets smaller, the price of the land DECREASES by ~$17,000.

7.

$$\begin{aligned} V &= V_0 + \Delta V = V_0 \left(1 + \beta \Delta T\right) \\ &= (60.00\ \text{L}) \left[1 + \left(950 \times 10^{-6}/^\circ\text{C}\right) (35.0^\circ\text{C} - 15.0^\circ\text{C})\right] \\ &= 61.1\text{L} \end{aligned}$$

9. (a) 9.35 mL; (b) 7.56 mL

11. 0.832 mm

13. We know how the length changes with temperature: $\Delta L = \alpha L_0 \Delta T$. Also we know that the volume of a cube is related to its length by $V = L^3$, so the final volume is then $V = V_0 + \Delta V = (L_0 + \Delta L)^3$. Substituting for ΔL gives $V = (L_0 + \alpha L_0 \Delta T)^3 = L_0^3(1 + \alpha \Delta T)^3$.

Now, because $\alpha \Delta T$ is small, we can use the binomial expansion: $V \approx L_0^3(1 + 3\alpha \Delta T) = L_0^3 + 3\alpha L_0^3 \Delta T$.

So writing the length terms in terms of volumes gives $V = V_0 + \Delta V \approx V_0 + 3\alpha V_0 \Delta T$, and so $\Delta V = \beta V_0 \Delta T \approx 3\alpha V_0 \Delta T$, or $\beta \approx 3\alpha$.

The Ideal Gas Law

Lumen Learning

Learning Objectives

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

- State the ideal gas law in terms of molecules and in terms of moles.
- Use the ideal gas law to calculate pressure change, temperature change, volume change, or the number of molecules or moles in a given volume.
- Use Avogadro's number to convert between number of molecules and number of moles.

In this section, we continue to explore the thermal behavior of gases. In particular, we examine the characteristics of atoms and molecules that compose gases. (Most gases, for example nitrogen, N_2, and oxygen, O_2, are composed of two or more atoms. We will primarily use the term "molecule" in discussing a gas because the term can also be applied to monatomic gases, such as helium.)

Figure 1. The air inside this hot air balloon flying over Putrajaya, Malaysia, is hotter than the ambient air. As a result, the balloon experiences a buoyant force pushing it upward. (credit: Kevin Poh, Flickr)

Gases are easily compressed. We can see evidence of this in Table 1 in Thermal Expansion of Solids and Liquids, where you will note that gases have the *largest* coefficients of volume expansion. The large coefficients mean that gases expand and contract very rapidly with temperature changes. In addition, you will note that most gases expand at the *same* rate, or have the same β. This raises the question as to why gases should all act in nearly the same way, when liquids and solids have widely varying expansion rates.

The answer lies in the large separation of atoms and molecules in gases, compared to their sizes, as illustrated in Figure 2. Because atoms and molecules have large separations, forces between them can be ignored, except when they collide with each other during collisions. The motion of atoms and molecules (at temperatures well above the boiling temperature) is fast, such that the gas occupies all of the accessible volume and the expansion of gases is rapid. In contrast, in liquids and solids, atoms and molecules are closer together and are quite sensitive to the forces between them.

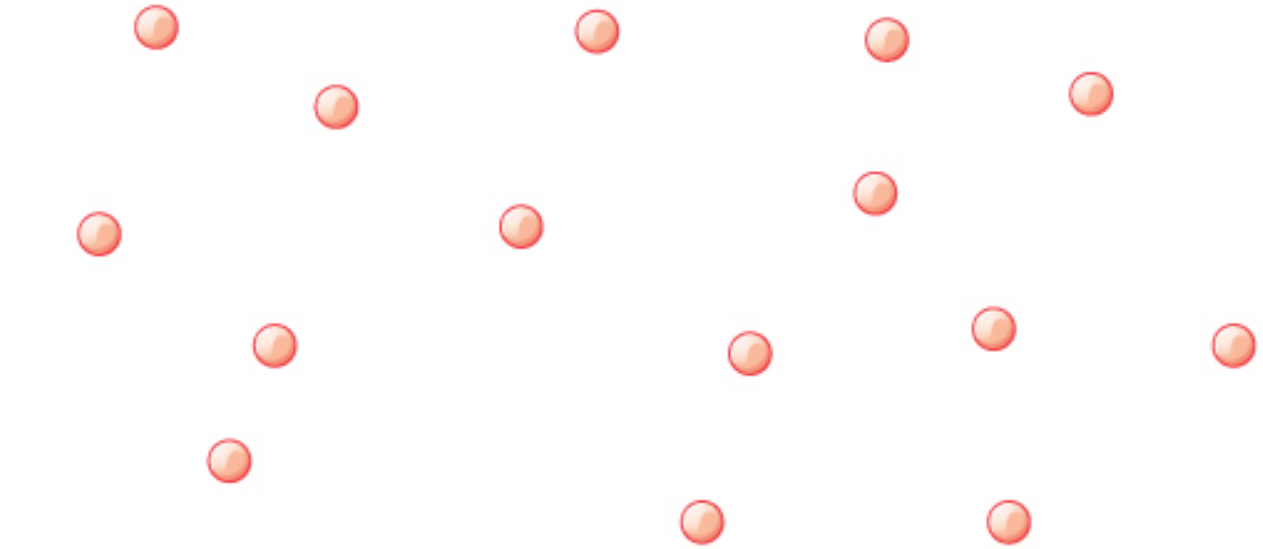


Figure 2. Atoms and molecules in a gas are typically widely separated, as shown. Because the forces between them are quite weak at these distances, the properties of a gas depend more on the number of atoms per unit volume and on temperature than on the type of atom.

To get some idea of how pressure, temperature, and volume of a gas are related to one another, consider what happens when you pump air into an initially deflated 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 walls limit volume expansion. If we continue to pump air into it, the pressure increases. The pressure will further increase when the car is driven and the tires move. Most manufacturers specify optimal tire pressure for cold tires. (See Figure 3.)

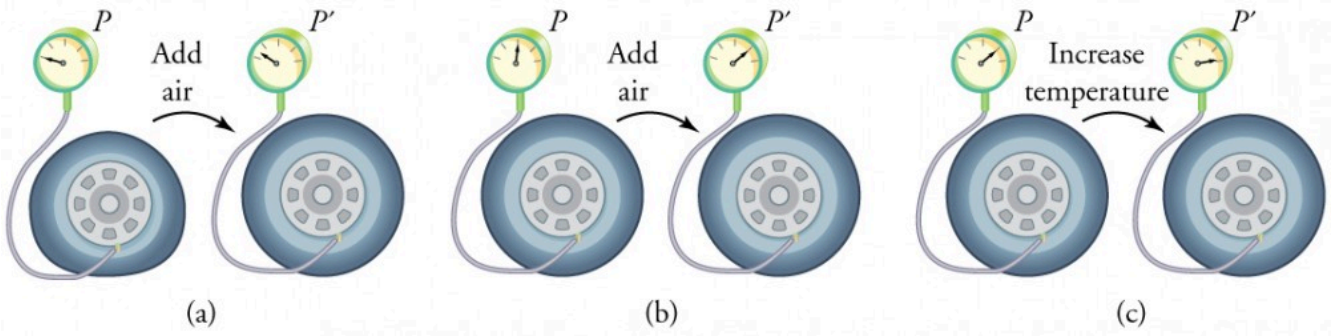


Figure 3. (a) When air is pumped into a deflated tire, its volume first increases without much increase in pressure. (b) When the tire is filled to a certain point, the tire walls resist further expansion and the pressure increases with more air. (c) Once the tire is inflated, its pressure increases with temperature.

At room temperatures, collisions between atoms and molecules can be ignored. In this case, the gas is called an ideal gas, in which case the relationship between the pressure, volume, and temperature is given by the equation of state called the ideal gas law.

Ideal Gas Law

The *ideal gas law* states that $PV = NkT$, where P is the absolute pressure of a gas, V is the volume it occupies, N is the number of atoms and molecules in the gas, and T is its absolute temperature. The constant k is called the *Boltzmann constant* in honor of Austrian physicist Ludwig Boltzmann (1844–1906) and has the value $k = 1.38 \times 10^{-23}$ J/K.

The ideal gas law can be derived from basic principles, but was originally deduced from experimental measurements of Charles' law (that volume occupied by a gas is proportional to temperature at a fixed pressure) and from Boyle's law (that for a fixed temperature, the product PV is a constant). In the ideal gas model, the volume occupied by its atoms and molecules is a negligible fraction of V. The ideal gas law describes the behavior of real gases under most conditions. (Note, for example, that N is the total number of atoms and molecules, independent of the type of gas.)

Let us see how the ideal gas law is consistent with the behavior of filling the tire when it is pumped slowly and the temperature is constant. At first, the pressure P is essentially equal to atmospheric pressure, and the volume V increases in direct proportion to the number of atoms and molecules N put into the tire. Once the volume of the tire is constant, the equation $PV = NkT$ predicts that the pressure should increase in proportion to *the number N of atoms and molecules*.

Example 1. Calculating Pressure Changes Due to Temperature Changes: Tire Pressure

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 lb/in^2) at a temperature of 18.0ºC. What is the pressure after its temperature has risen to 35.0ºC? Assume that there are no appreciable leaks or changes in volume.

Strategy

The pressure in the tire is changing only because of changes in temperature. First we need to identify what we know and what we want to know, and then identify an equation to solve for the unknown.

We know the initial pressure $P_0 = 7.00 \times 10^5$ Pa, the initial temperature $T_0 = 18.0$ºC, and the final temperature $T_f = 35.0$ºC. We must find the final pressure P_f. How can we use the equation $PV = NkT$? At first, it may seem that not enough information is given, because the volume V and number of atoms N are not specified. What we can do is use the equation twice: $P_0V_0 = NkT_0$ and $P_fV_f = NkT_f$. If we divide P_fV_f by P_0V_0 we can come up with an equation that allows us to solve for P_f.

$$\frac{P_fV_f}{P_0V_0} = \frac{N_fkT_f}{N_0kT_0}$$

Since the volume is constant, V_f and V_0 are the same and they cancel out. The same is true for N_f and N_0, and k, which is a constant. Therefore,

$$\frac{P_f}{P_0} = \frac{T_f}{T_0}$$

We can then rearrange this to solve for P_f:

$$P_f = P_0 \frac{T_f}{T_0}$$

, where the temperature must be in units of kelvins, because T_0 and T_f are absolute temperatures.

Solution

Convert temperatures from Celsius to Kelvin:

$$T_0 = (18.0 + 273)\text{K} = 291 \text{ K}$$

$$T_f = (35.0 + 273)\text{K} = 308 \text{ K}$$

Substitute the known values into the equation.

$$P_f = P_0 \frac{T_f}{T_0} = 7.00 \times 10^5 \text{ Pa} \left(\frac{308 \text{ K}}{291 \text{ K}} \right) = 7.41 \times 10^5 \text{ Pa}$$

Discussion

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 must be used in the ideal gas law.

Making Connections: Take-Home Experiment—Refrigerating a Balloon

Inflate a balloon at room temperature. Leave the inflated balloon in the refrigerator overnight. What happens to the balloon, and why?

Example 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 drink? We can use the ideal gas law to give us an idea of how large N typically is.

Calculate the number of molecules in a cubic meter of gas at standard temperature and pressure (STP), which is defined to be 0ºC and atmospheric pressure.

Strategy

Because pressure, volume, and temperature are all specified, we can use the ideal gas law $PV = NkT$, to find N.

Solution

Identify the knowns:

$$\begin{aligned} T &= 0^\circ\text{C} = 273\ \text{K} \\ P &= 1.01 \times 10^5\ \text{Pa} \\ V &= 1.00\ \text{m}^3 \\ k &= 1.38 \times 10^{-23}\ \text{J/K} \end{aligned}$$

Identify the unknown: number of molecules, *N*.

Rearrange the ideal gas law to solve for *N*:

$$\begin{aligned} PV &= NkT \\ N &= \frac{PV}{kT} \end{aligned}$$

Substitute the known values into the equation and solve for *N*:

$$N = \frac{PV}{kT} = \frac{\left(1.01 \times 10^5\ \text{Pa}\right)\left(1.00\ \text{m}^3\right)}{\left(1.38 \times 10^{-23}\ \text{J/K}\right)\left(273\ \text{K}\right)} = 2.68 \times 10^{25}\ \text{molecules}$$

Discussion

This number is undeniably large, considering that a gas is mostly empty space. *N* is huge, even in small volumes. For example, 1 cm^3 of a gas at STP has 2.68×10^{19} molecules in it. Once again, note that *N* is the same for all types or mixtures of gases.

Moles and Avogadro's Number

It is sometimes convenient to work with a unit other than molecules when measuring the amount of substance. A *mole* (abbreviated mol) is defined to be the amount of a substance that contains as many atoms or molecules as there are atoms in exactly 12 grams (0.012 kg) of carbon-12. The actual number of atoms or molecules in one mole is called *Avogadro's number* (N_A), in recognition of Italian scientist Amedeo Avogadro (1776–1856). He developed the concept of the mole, based on the hypothesis that equal volumes of gas, at the same pressure and temperature, contain equal numbers of molecules. That is, the number is independent of the type of gas. This hypothesis has been confirmed, and the value of Avogadro's number is $N_A = 6.02 \times 10^{23}\ \text{mol}^{-1}$.

Avogadro's Number

One mole always contains 6.02×10^{23} particles (atoms or molecules), independent of the element or substance. A mole of any substance has a mass in grams equal to its molecular mass, which can be calculated from the atomic masses given in the periodic table of elements.

$$N_A = 6.02 \times 10^{23}\ \text{mol}^{-1}$$

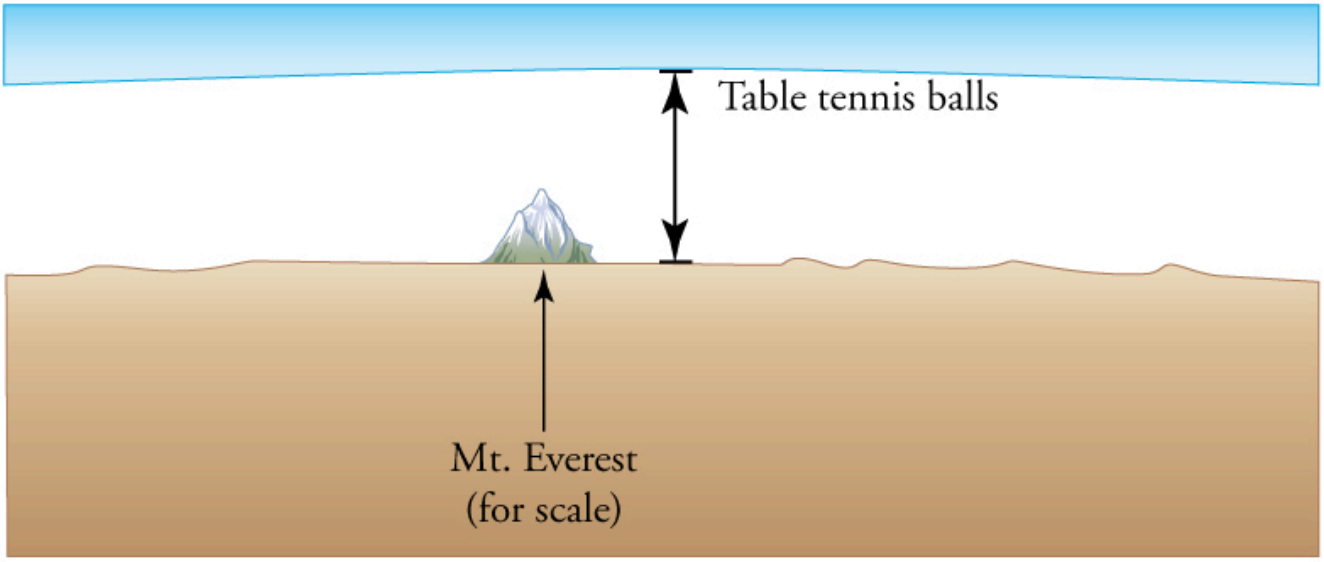


Figure 4. How big is a mole? On a macroscopic level, one mole of table tennis balls would cover the Earth to a depth of about 40 km.

Check Your Understanding

The active ingredient in a Tylenol pill is 325 mg of acetaminophen ($C_8H_9NO_2$). Find the number of active molecules of acetaminophen in a single pill.

Solution

We first need to calculate the molar mass (the mass of one mole) of acetaminophen. To do this, we need to multiply the number of atoms of each element by the element's atomic mass.

(8 moles of carbon)(12 grams/mole) + (9 moles hydrogen)(1 gram/mole) + (1 mole nitrogen)(14 grams/mole) + (2 moles oxygen)(16 grams/mole) = 151 g

Then we need to calculate the number of moles in 325 mg.

$$\left(\frac{325\text{ mg}}{151\text{ grams/mole}}\right)\left(\frac{1\text{ gram}}{1000\text{ mg}}\right) = 2.15 \times 10^{-3}\text{ moles}$$

Then use Avogadro's number to calculate the number of molecules.

$$N = (2.15 \times 10^{-3}\text{ moles})(6.02 \times 10^{23}\text{ molecules/mole}) = 1.30 \times 10^{21}\text{ molecules}$$

Example 3. Calculating Moles per Cubic Meter and Liters per Mole

Calculate the following:

1. The number of moles in 1.00 m^3 of gas at STP
2. The number of liters of gas per mole.

Strategy and Solution

1. We are asked to find the number of moles per cubic meter, and we know from Example 2 that the number of molecules per cubic meter at STP is 2.68×10^{25}. The number of moles can be found by dividing the number of molecules by Avogadro's number. We let *n* stand for the number of moles,
$$m \text{ mol/m}^3 = \frac{N \text{ molecules/m}^3}{6.02\times10^{23} \text{ molecules/mol}} = \frac{2.68\times10^{25} \text{ molecules/m}^3}{6.02\times10^{23} \text{ molecules/mol}} = 44.5 \text{ mol/m}^3$$
2. Using the value obtained for the number of moles in a cubic meter, and converting cubic meters to liters, we obtain
$$\frac{\left(10^3 \text{ L/m}^3\right)}{44.5 \text{ mol/m}^3} = 22.5 \text{ L/mol}$$

Discussion

This value is very close to the accepted value of 22.4 L/mol. The slight difference is due to rounding errors caused by using three-digit input. Again this number is the same for all gases. In other words, it is independent of the gas.

The (average) molar weight of air (approximately 80% N_2 and 20% O_2 is $M = 28.8$ g. Thus the mass of one cubic meter of air is 1.28 kg. If a living room has dimensions 5 m × 5 m × 3 m, the mass of air inside the room is 96 kg, which is the typical mass of a human.

Check Your Understanding

The density of air at standard conditions (P = 1 atm and T = 20ºC) is 1.28 kg/m^3. At what pressure is the density 0.64 kg/m^3 if the temperature and number of molecules are kept constant?

Solution

The best way to approach this question is to think about what is happening. If the density drops to half its original value and no molecules are lost, then the volume must double. If we look at the equation $PV = NkT$, we see that when the temperature is constant, the pressure is inversely proportional to volume. Therefore, if the volume doubles, the pressure must drop to half its original value, and $P_f = 0.50$ atm.

The Ideal Gas Law Restated Using Moles

A very common expression of the ideal gas law uses the number of moles, *n*, rather than the number of atoms and molecules, *N*. We start from the ideal gas law, $PV = NkT$, and multiply and divide the equation by Avogadro's number N_A. This gives
$$PV = \frac{N}{N_A} N_A kT$$

.

Note that

$$n = \frac{N}{N_A}$$

is the number of moles. We define the universal gas constant $R=N_Ak$, and obtain the ideal gas law in terms of moles.

Ideal Gas Law (in terms of moles)

The ideal gas law (in terms of moles) is $PV = nRT$.

The numerical value of R in SI units is $R = N_Ak = (6.02 \times 10^{23}\ \text{mol}^{-1})(1.38 \times 10^{-23}\ \text{J/K}) = 8.31\ \text{J/mol} \cdot \text{K}$.

In other units,

$$R = 1.99\ \text{cal/mol} \cdot \text{K}$$

$$R = 0.0821\ \text{L} \cdot \text{atm/mol} \cdot \text{K}$$

You can use whichever value of R is most convenient for a particular problem.

Example 4. Calculating Number of Moles: Gas in a Bike Tire

How many moles of gas are in a bike tire with a volume of $2.00 \times 10^{-3}\ \text{m}^3$(2.00 L), a pressure of 7.00×10^5 Pa (a gauge pressure of just under 90.0 lb/in^2), and at a temperature of 18.0ºC?

Strategy

Identify the knowns and unknowns, and choose an equation to solve for the unknown. In this case, we solve the ideal gas law, $PV = nRT$, for the number of moles n.

Solution

Identify the knowns:

$$\begin{aligned} P &= 7.00 \times 10^5\ \text{Pa} \\ V &= 2.00 \times 10^{-3}\ \text{m}^3 \\ T &= 18.0^\circ\text{C} = 291\ \text{K} \\ R &= 8.31\ \text{J/mol} \cdot\ \text{K} \end{aligned}$$

Rearrange the equation to solve for n and substitute known values.

$$\begin{aligned} n &= \frac{PV}{RT} = \frac{(7.00\times10^5\ \text{Pa})(2.00\times10^{-3}\ \text{m}^3)}{(8.31\ \text{J/mol}\cdot\ \text{K})(291\ \text{K})} \\ &= 0.579\ \text{mol} \end{aligned}$$

Discussion

The most convenient choice for R in this case is 8.31 J/mol · K, because our known quantities are in SI units. The pressure and temperature are obtained from the initial conditions in Example 1, but we would get the same answer if we used the final values.

The ideal gas law can be considered to be another manifestation of the law of conservation of energy (see Conservation of Energy). Work done on a gas results in an increase in its energy, increasing pressure and/or temperature, or decreasing volume. This increased energy can also be viewed as increased internal kinetic energy, given the gas's atoms and molecules.

The Ideal Gas Law and Energy

Let us now examine the role of energy in the behavior of gases. When you inflate a bike tire by hand, you do work by repeatedly exerting a force through a distance. This energy goes into increasing the pressure of air inside the tire and increasing the temperature of the pump and the air.

The ideal gas law is closely related to energy: the units on both sides are joules. The right-hand side of the ideal gas law in $PV = NkT$ is NkT. This term is roughly the amount of translational kinetic energy of N atoms or molecules at an absolute temperature T, as we shall see formally in Kinetic Theory: Atomic and Molecular Explanation of Pressure and Temperature. The left-hand side of the ideal gas law is PV, which also has the units of joules. We know from our study of fluids that pressure is one type of potential energy per unit volume, so pressure multiplied by volume is energy. The important point is that there is energy in a gas related to both its pressure and its volume. The energy can be changed when the gas is doing work as it expands—something we explore in Heat and Heat Transfer Methods—similar to what occurs in gasoline or steam engines and turbines.

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.

Step 2. Make a list of what quantities are given, or can be inferred from the problem as stated (identify the known quantities). Convert known values into proper SI units (K for temperature, Pa for pressure, m^3 for volume, molecules for N, and moles for n).

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, in order to decide which form of the ideal gas law to use. The first form is $PV = NkT$ and involves N, the number of atoms or molecules. The second form is $PV = nRT$ and involves n, the number of moles.

Step 5. 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 6. Substitute the known quantities, along with their units, into the appropriate equation, and obtain numerical solutions complete with units. Be certain to use absolute temperature and absolute pressure.

Step 7. Check the answer to see if it is reasonable: Does it make sense?

Check Your Understanding

Liquids and solids have densities about 1000 times greater than gases. Explain how this implies that the distances between atoms and molecules in gases are about 10 times greater than the size of their atoms and molecules.

Solution

Atoms and molecules are close together in solids and liquids. In gases they are separated by empty space. Thus gases have lower densities than liquids and solids. Density is mass per unit volume, and volume is related to the size of a body (such as a sphere) cubed. So if the distance between atoms and 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.

Section Summary

- The ideal gas law relates the pressure and volume of a gas to the number of gas molecules and the temperature of the gas.
- The ideal gas law can be written in terms of the number of molecules of gas: $PV = NkT$, where P is pressure, V is volume, T is temperature, N is number of molecules, and k is the Boltzmann constant $k = 1.38 \times 10^{-23}$ J/K.
- A mole is the number of atoms in a 12-g sample of carbon-12.
- The number of molecules in a mole is called Avogadro's number NA, $NA = 6.02 \times 10^{23}$ mol^{-1}.
- A mole of any substance has a mass in grams equal to its molecular weight, 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 number of moles and R is the universal gas constant, $R = 8.31$ J/mol · K.
- The ideal gas law is generally valid at temperatures well above the boiling temperature.

Conceptual Questions

Find out the human population of Earth. Is there a mole of people inhabiting Earth? If the average mass of a person is 60 kg, calculate the mass of a mole of people. How does the mass of a mole of people compare with the mass of Earth?

Under what circumstances would you expect a gas to behave significantly differently than predicted by the ideal gas law?

A constant-volume gas thermometer contains a fixed amount of gas. What property of the gas is measured to indicate its temperature?

Problems & Exercises

1. The gauge pressure in your car tires is 2.50×10^5 N/m^2 at a temperature of 35.0ºC when you drive it onto a ferry boat to Alaska. What is their gauge pressure later, when their temperature has dropped to –40.0ºC?
2. Convert an absolute pressure of 7.00×10^5 N/m^2 to gauge pressure in lb/in^2. (This value was stated to be just less than 90.0 lb/in^2 in Example 4. Is it?)
3. 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 20.0ºC. (a) Find the gauge pressure inside such a bulb when it is hot, assuming its average temperature is 60.0ºC (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. What will the actual final pressure be, taking this into account? Is this a negligible difference?
4. Large helium-filled balloons are used to lift scientific equipment to high altitudes. (a) What is the pressure inside such a balloon if it starts out at sea level with a temperature of 10.0ºC and rises to an altitude where its volume is twenty times the original volume and its temperature is –50.0ºC? (b) What is the gauge pressure? (Assume atmospheric pressure is constant.)
5. Confirm that the units of nRT are those of energy for each value of R: (a) 8.31 J/mol · K, (b) 1.99 cal/mol · K, and (c) 0.0821 L · atm/mol · K.
6. In the text, it was shown that $N/V = 2.68 \times 10^{25}$ m^{-3} for gas at STP. (a) Show that this quantity is equivalent to $N/V = 2.68 \times 10^{19}$ cm^{-3}, as stated. (b) About how many atoms are there in one μm^3 (a cubic micrometer) at STP? (c) What does your answer to part (b) imply about the separation of atoms and molecules?
7. 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 37.0ºC (body temperature).
8. An airplane passenger has 100 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 7.50×10^4 N/m^2?
9. (a) What is the volume (in km^3) of Avogadro's number of sand grains if each grain is a cube and has sides that are 1.0 mm long? (b) How many kilometers of beaches in length would this cover if the beach averages 100 m in width and 10.0 m in depth? Neglect air spaces between grains.
10. An expensive vacuum system can achieve a pressure as low as 1.00×10^{-7} N/m^2 at 20ºC. How many atoms are there in a cubic centimeter at this pressure and temperature?
11. The number density of gas atoms at a certain location in the space above our planet is about 1.00×10^{11} m^{-3}, and the pressure is 2.75×10^{-10} N/m^2 in this space. What is the temperature there?
12. A bicycle tire has a pressure of 7.00×10^5 N/m^2 at a temperature of 18.0ºC and contains 2.00 L of gas. What will its pressure be if you let out an amount of air that has a volume of 100cm3 at atmospheric pressure? Assume tire temperature and volume remain constant.
13. A high-pressure gas cylinder contains 50.0 L of toxic gas at a pressure of 1.40×10^7 N/m^2 and a temperature of 25.0ºC. Its valve leaks after the cylinder is dropped. The cylinder is cooled to dry ice temperature (–78.5ºC) 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 appear to be a practical solution?

14. Find the number of moles in 2.00 L of gas at 35.0ºC and under 7.41×10^7 N/m^2 of pressure.
15. 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.
16. (a) What is the gauge pressure in a 25.0ºC 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 25.0ºC? Assume the temperature returns to 25.0ºC and the volume remains constant.
17. (a) In the deep space between galaxies, the density of atoms is 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?

Glossary

ideal gas law: the physical law that relates the pressure and volume of a gas to the number of gas molecules or number of moles of gas and the temperature of the gas

Boltzmann constant: *k*, a physical constant that relates energy to temperature; $k = 1.38 \times 10^{-23}$ J/K

Avogadro's number: *NA*, the number of molecules or atoms in one mole of a substance; $NA = 6.02 \times 10^{23}$ particles/mole

mole: the quantity of a substance whose mass (in grams) is equal to its molecular mass

Selected Solutions to Problems & Exercises

1. 1.62 atm

3. (a) 0.136 atm; (b) 0.135 atm. The difference between this value and the value from part (a) is negligible.

5. (a)

$$\mathrm{nRT} = (\mathrm{mol})\,(\mathrm{J/mol \cdot K})\,(\mathrm{K}) = \mathrm{J}$$

;

(b)

$$\mathrm{nRT} = (\mathrm{mol})\,(\mathrm{cal/mol \cdot K})\,(\mathrm{K}) = \mathrm{cal}$$

;

(c)

$$\begin{aligned} \mathrm{nRT} &= (\mathrm{mol})\,(\mathrm{L}\cdot\mathrm{atm}/\mathrm{mol}\cdot\mathrm{K})\,(\mathrm{K}) \\ &= \mathrm{L}\cdot\mathrm{atm} = \left(\mathrm{m}^3\right)\left(\mathrm{N}/\mathrm{m}^2\right) \\ &= \mathrm{N}\cdot\mathrm{m} = \mathrm{J} \end{aligned}$$

7. 7.86×10^{-2} mol

9. (a) 6.02×10^{5} km^3; (b) 6.02×10^{8} km

11. −73.9ºC

13. (a) 9.14×10^{6} N/m^2; (b) 8.23×10^{6} N/m^2; (c) 2.16 K; (d) No. The final temperature needed is much too low to be easily achieved for a large object.

15. 41 km

17. (a) 3.7×10^{-17} Pa; (b) 6.0×10^{17} m^3; (c) 8.4×10^{2} km

Kinetic Theory: Atomic and Molecular Explanation of Pressure and Temperature

Lumen Learning

Learning Objectives

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

- Express the ideal gas law in terms of molecular mass and velocity.
- Define thermal energy.
- Calculate the kinetic energy of a gas molecule, given its temperature.
- Describe the relationship between the temperature of a gas and the kinetic energy of atoms and molecules.
- Describe the distribution of speeds of molecules in a gas.

We have developed macroscopic definitions of pressure and temperature. Pressure is the force divided by the area on which the force is exerted, and temperature is measured with a thermometer. We gain a better understanding of pressure and temperature from the kinetic theory of gases, which assumes that atoms and molecules are in continuous random motion.

Figure 1 shows an elastic 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). Because a huge number of molecules will collide with the wall in a short time, we observe an average force per unit area. These collisions are the source of pressure in a gas. As the number of molecules increases, the number of collisions and thus the pressure increase. Similarly, the gas pressure is higher if the average velocity of molecules is higher. The actual relationship is derived in the Making Connections feature below. The following relationship is found:

$$PV = \frac{1}{3}Nm\overline{v^2}$$

, where *P* is the pressure (average force per unit area), *V* is the volume of gas in the container, *N* is the number of molecules in the container, *m* is the mass of a molecule, and $\overline{v^2}$ is the average of the molecular speed squared.

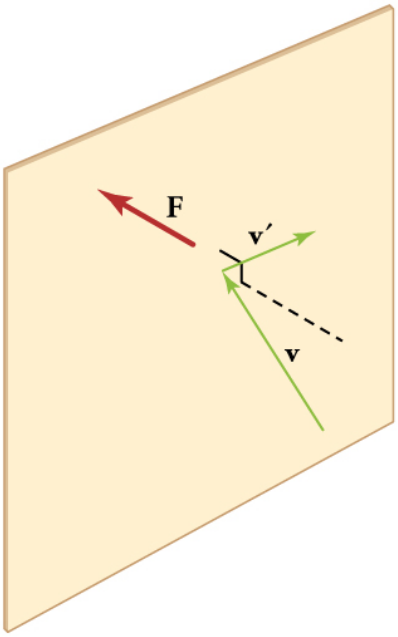


Figure 1. When a molecule collides with a rigid wall, the component of its momentum perpendicular to the wall is reversed. A force is thus exerted on the wall, creating pressure.

What can we learn from this atomic and molecular version of the ideal gas law? We can derive a relationship between temperature and the average translational kinetic energy of molecules in a gas. Recall the previous expression of the ideal gas law: *PV = NkT*.

Equating the right-hand side of this equation with the right-hand side of

$$PV = \frac{1}{3}Nm\overline{v^2}$$

gives

$$\frac{1}{3}Nm\overline{v^2} = NkT$$

.

Making Connections: Things Great and Small—Atomic and Molecular Origin of Pressure in a Gas

Figure 2 shows a box filled with a gas. We know from our previous discussions that putting more gas into the box produces greater pressure, and that increasing the temperature of the gas also produces a greater pressure. But why should increasing the temperature of the gas increase the pressure in the box? A look at the atomic and molecular scale gives us some answers, and an alternative expression for the ideal gas law.

The figure shows an expanded view of an elastic collision of a gas molecule with the wall of a container. Calculating the average force exerted by such molecules will lead us to the ideal gas law, and to the connection between temperature and molecular kinetic energy. We assume that a molecule is small compared with the separation of molecules in the gas, and that its interaction with other molecules can be ignored. We also assume the wall is rigid and that the molecule's direction changes, but that its speed remains constant (and hence its kinetic energy and the magnitude of its momentum remain constant as well). This assumption is not always valid, but the same result is obtained with a more detailed description of the molecule's exchange of energy and momentum with the wall.

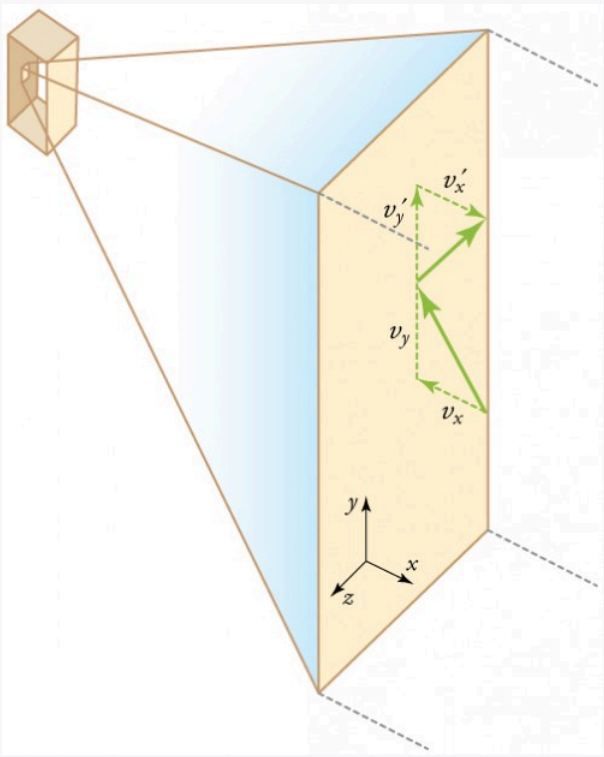


Figure 2. Gas in a box exerts an outward pressure on its walls. A molecule colliding with a rigid wall has the direction of its velocity and momentum in the x-direction reversed. This direction is perpendicular to the wall. The components of its velocity momentum in the y- and z-directions are not changed, 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 $\Delta mv = +mv_x - (-mv_x) = 2mv_x$. The force exerted on the molecule is given by

$$F = \frac{\Delta p}{\Delta t} = \frac{2mv_x}{\Delta t}$$

.

There is no force between the wall and the molecule until the molecule hits the wall. During the short time of the collision, the force between the molecule and wall is relatively large. We are looking for an average force; we take Δt to be the average time between collisions of the molecule with this wall. It is the time it would take the molecule to go across the box and back (a distance $2l$) at a speed of v_x. Thus

$$\Delta t = \frac{2l}{v_x}$$

, and the expression for the force becomes

$$F = \frac{2mv_x}{\frac{2l}{v_x}} = \frac{mv_x^2}{l}$$

This force is due to *one* molecule. We multiply by the number of molecules N and use their average squared velocity to find the force

$$F = N\frac{m\overline{{v_x}^2}}{l}$$

,

where the bar over a quantity means its average value. We would like to have the force in terms of the speed *v*, rather than the *x*-component of the velocity. We note that the total velocity squared is the sum of the squares of its components, so that

$$\overline{v^2} = \overline{{v_x}^2} + \overline{{v_y}^2} + \overline{{v_z}^2}$$

.

Because the velocities are random, their average components in all directions are the same:

$$\overline{{v_x}^2} = \overline{{v_y}^2} = \overline{{v_z}^2}$$

.

Thus,

$$\overline{v^2} = 3\overline{{v_x}^2}$$

or

$$\overline{{v_x}^2} = \frac{1}{3}\overline{v^2}$$

.

Substituting

$$\frac{1}{3}\overline{v^2}$$

into the expression for *F* gives

$$F = N\frac{m\overline{v^2}}{3l}$$

.

The pressure is

$$\frac{F}{A}$$

, so that we obtain

$$P = \frac{F}{A} = N\frac{m\overline{v^2}}{3Al} = \frac{1}{3}\frac{Nm\overline{v^2}}{V}$$

, where we used *V* = *Al* for the volume. This gives the important result.

$$PV = \frac{1}{3}Nm\overline{v^2}$$

This equation is another expression of the ideal gas law.

We can get the average kinetic energy of a molecule,

$$\frac{1}{2}mv^2$$

, from the left-hand side of the equation by canceling *N* and multiplying by 3/2. This calculation produces the result that the average kinetic energy of a molecule is directly related to absolute temperature.

$$\overline{\text{KE}} = \frac{1}{2}m\overline{v^2} = \frac{3}{2}kT$$

The average translational kinetic energy of a molecule,

$$\overline{\text{KE}}$$

, is called *thermal energy*. The equation

$$\overline{\text{KE}} = \frac{1}{2}m\overline{v^2} = \frac{3}{2}kT$$

is a molecular interpretation of temperature, and it has been found to be valid for gases and reasonably accurate in liquids and solids. It is another definition of temperature based on an expression of the molecular energy.

It is sometimes useful to rearrange

$$\overline{\text{KE}} = \frac{1}{2}m\overline{v^2} = \frac{3}{2}kT$$

, and solve for the average speed of molecules in a gas in terms of temperature,

$$\sqrt{\overline{v^2}} = v_{\text{rms}} = \sqrt{\frac{3kT}{m}}$$

where v_{rms} stands for root-mean-square (rms) speed.

Example 1. Calculating Kinetic Energy and Speed of a Gas Molecule

1. What is the average kinetic energy of a gas molecule at 20.0ºC (room temperature)?
2. Find the rms speed of a nitrogen molecule (N_2) at this temperature.

Strategy for Part 1

The known in the equation for the average kinetic energy is the temperature.

$$\overline{\text{KE}} = \frac{1}{2}m\overline{v^2} = \frac{3}{2}kT$$

Before substituting values into this equation, we must convert the given temperature to kelvins. This conversion gives T = (20.0 + 273)K = 293K.

Solution for Part 1

The temperature alone is sufficient to find the average translational kinetic energy. Substituting the temperature into the translational kinetic energy equation gives

$$\overline{\text{KE}} = \tfrac{3}{2}kT = \tfrac{3}{2}\left(1.38 \times 10^{-23}\ \text{J/K}\right)(293\ \text{K}) = 6.07 \times 10^{-21}\ \text{J}$$

Strategy for Part 2

Finding the rms speed of a nitrogen molecule involves a straightforward calculation using the equation

$$\sqrt{\overline{v^2}} = v_{\text{rms}} = \sqrt{\frac{3kT}{m}}$$

but we must first find the mass of a nitrogen molecule. Using the molecular mass of nitrogen N_2 from the periodic table,

$$m = \frac{2(14.0067) \times 10^{-3}\ \text{kg/mol}}{6.02 \times 10^{23}\ \text{mol}^{-1}} = 4.65 \times 10^{-26}\ \text{kg}$$

Solution for Part 2

Substituting this mass and the value for *k* into the equation for v_{rms} yields

$$v_{\text{rms}} = \sqrt{\frac{3kT}{m}} = \sqrt{\frac{3\left(1.38 \times 10^{-23}\ \text{J/K}\right)(293\ \text{K})}{4.65 \times 10^{-26}\ \text{kg}}} = 511\ \text{m/s}$$

Discussion

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. The rms velocity 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 can move on average between collisions) of molecules in air is very small, and 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, however, which is about 340 m/s at room temperature. The faster the rms speed of air molecules, the faster that 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 3.)

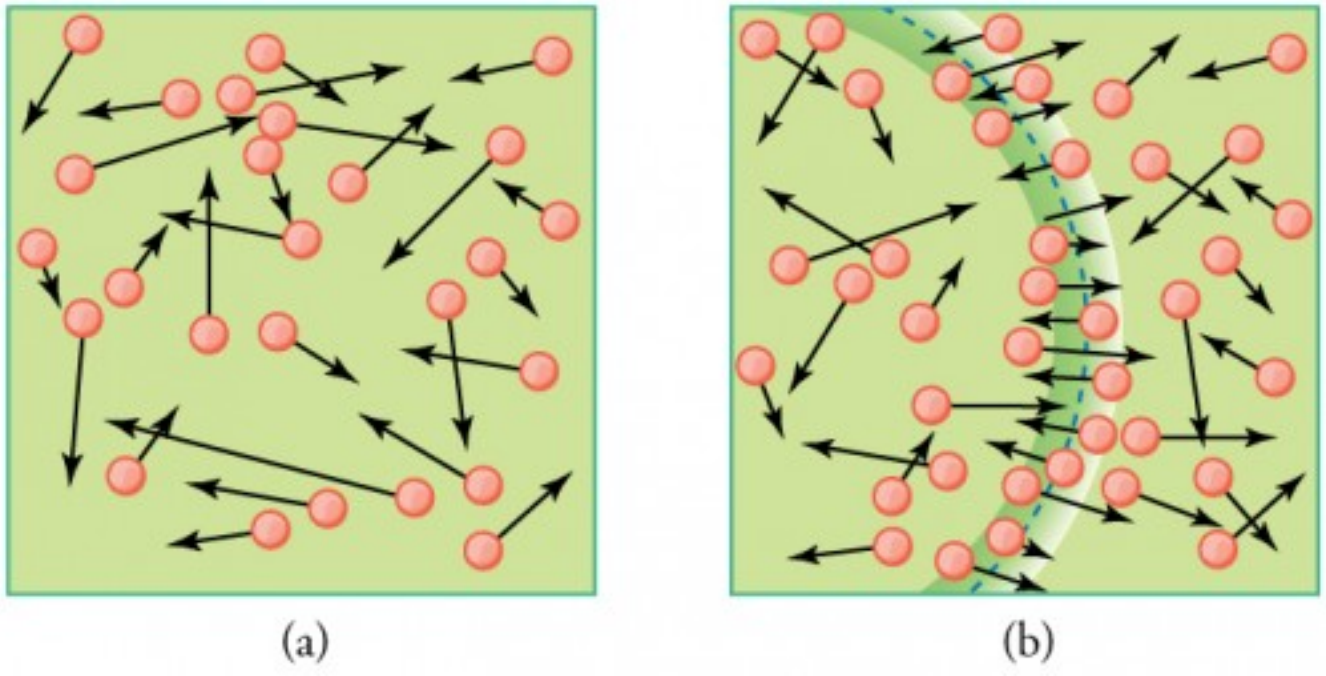


Figure 3. (a) There are many molecules moving so fast in an ordinary gas that they collide a billion times every second. (b) Individual molecules do not move very far in a small amount of time, but disturbances like sound waves are transmitted at speeds related to the molecular speeds.

Making Connections: Historical Note—Kinetic Theory of Gases

The kinetic theory of gases was developed by Daniel Bernoulli (1700–1782), who is best known in physics for his work on fluid flow (hydrodynamics). Bernoulli's work predates the atomistic view of matter established by Dalton.

Distribution of Molecular Speeds

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 distribution is called the *Maxwell-Boltzmann distribution*, after its originators, who calculated it based on kinetic theory, and has since been confirmed experimentally. (See Figure 4.) The distribution has a long tail, because a few molecules may go several times the rms speed. The most probable speed v_p is less than the rms speed v_{rms}. Figure 5 shows that the curve is shifted to higher speeds at higher temperatures, with a broader range of speeds.

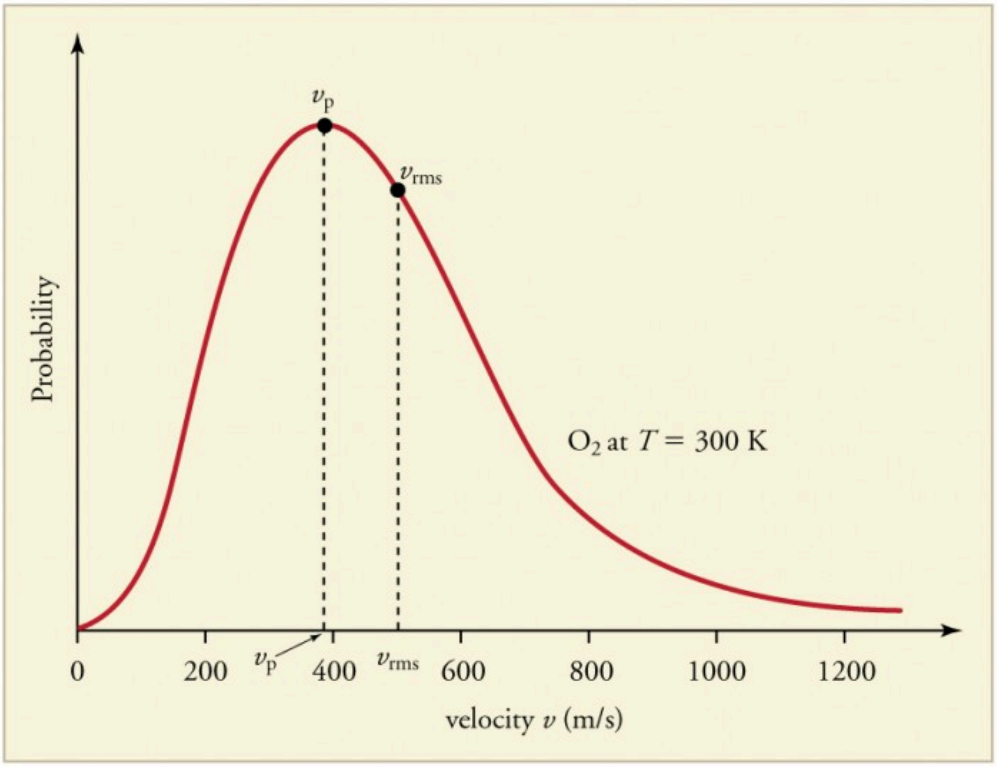


Figure 4. The Maxwell-Boltzmann distribution of molecular speeds in an ideal gas. The most likely speed vp is less than the rms speed vrms. Although very high speeds are possible, only a tiny fraction of the molecules have speeds that are an order of magnitude greater than vrms.

The distribution of thermal speeds depends strongly on temperature. As temperature increases, the speeds are shifted to higher values and the distribution is broadened.

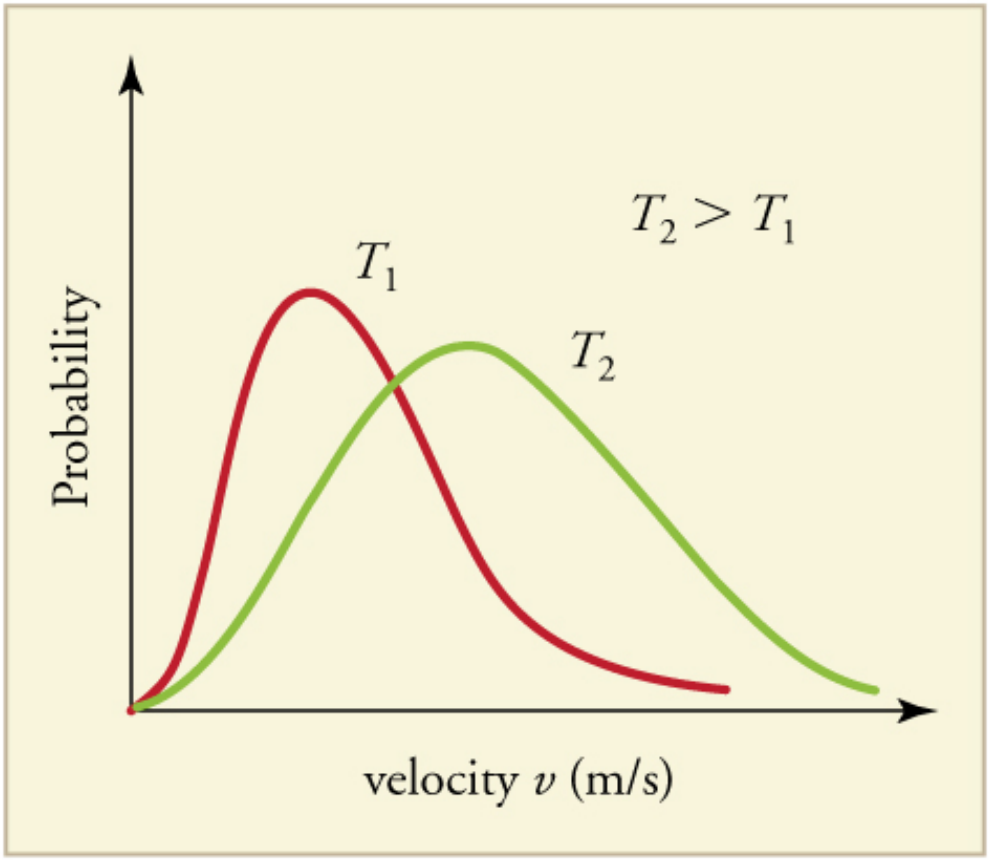


Figure 5. The Maxwell-Boltzmann distribution is shifted to higher speeds and is broadened at higher temperatures.

What is the implication of the change in distribution with temperature shown in Figure 5 for humans? All other things being equal, if a person has a fever, he or she is likely to lose more water molecules, particularly from linings along moist cavities such as the lungs and mouth, creating a dry sensation in the mouth.

Example 2. Calculating Temperature: Escape Velocity of Helium Atoms

In order 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

Identify the knowns: v is the escape velocity, 11.1 km/s.

Identify the unknowns: We need to solve for temperature, T. We also need to solve for the mass m of the helium atom.

Determine which equations are needed. To solve for mass *m* of the helium atom, we can use information from the periodic table:

$$m = \frac{\text{molar mass}}{\text{number of atoms per mole}}$$

To solve for temperature *T*, we can rearrange either

$$\overline{KE} = \tfrac{1}{2}m\overline{v^2} = \tfrac{3}{2}kT$$

or

$$\sqrt{\overline{v^2}} = v_{\text{rms}} = \sqrt{\frac{3kT}{m}}$$

$T = \frac{m\overline{v^2}}{3k}$

to yield , where *k* is the Boltzmann constant and *m* is the mass of a helium atom.

Plug the known values into the equations and solve for the unknowns.

$$m = \frac{\text{molar mass}}{\text{number of atoms per mole}} = \frac{4.0026\times10^{-3}\ \text{kg/mol}}{6.02\times10^{23}\ \text{mol}} = 6.65\times10^{-27}\ \text{kg}$$

$$T = \frac{\left(6.65\times10^{-27}\ \text{kg}\right)\left(11.1\times10^{3}\ \text{m/s}\right)^2}{3(1.38\times10^{-23}\ \text{J/K})} = 1.98\times10^{4}\ \text{K}$$

Discussion

This temperature is much higher than atmospheric temperature, which is approximately 250 K (–25ºC or –10ºF) at high altitude. Very few helium atoms are left in the atmosphere, but there were many when the atmosphere was formed. The reason for the loss of helium atoms is that there are a small number of helium atoms with speeds higher than Earth's escape velocity even at normal temperatures. The speed of a helium atom changes from one instant to the next, so that at any instant, there is a small, but nonzero chance that the speed is greater than the escape speed and the molecule escapes from Earth's gravitational pull. Heavier molecules, such as oxygen, nitrogen, and water (very little of which reach a very high altitude), 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, so few have speeds above the escape velocity that billions of years are required to lose significant amounts of the atmosphere. Figure 6 shows the impact 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 comparison between Earth and the Moon is discussed in this chapter's Problems and Exercises.

Figure 6. 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)

Check Your Understanding

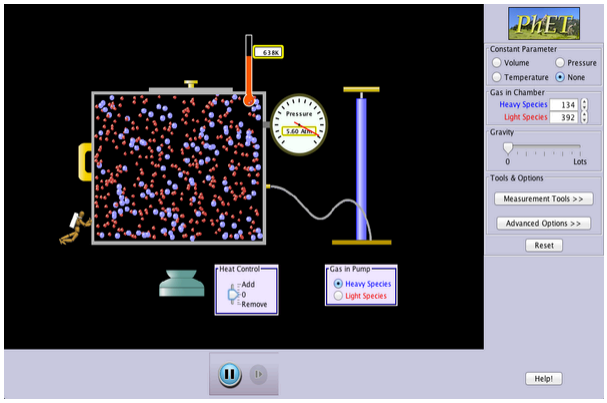
If you consider a very small object such as a grain of pollen, in a gas, then the number of atoms and molecules striking its surface would also be relatively small. Would the grain of pollen experience any fluctuations in pressure due to statistical fluctuations in the number of gas atoms and molecules striking it in a given amount of time?

Solution

Yes. Such fluctuations actually occur for a body of any size in a gas, but since the numbers of atoms and 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 proportional to the inverse square root of the number of collisions, so for small bodies they can become significant. This was actually observed in the 19th century for pollen grains in water, and is known as the Brownian effect.

PhET Explorations: Gas Properties

Pump gas molecules into a box and see what happens as you change the volume, add or remove heat, change gravity, and more. Measure the temperature and pressure, and discover how the properties of the gas vary in relation to each other.



Click to download the simulation. Run using Java.

Section Summary

- Kinetic theory is the atomistic description of gases as well as liquids and solids.
- Kinetic theory models the properties of matter in terms of continuous random motion of atoms and molecules.

$$\mathrm{PV} = \frac{1}{3}\mathrm{Nm}\overline{v^2}$$

- The ideal gas law can also be expressed as , where *P* is the pressure (average force per unit area), *V* is the volume of gas in the container, *N* is the number of molecules in the container, *m* is the mass of a molecule, and is the average of the molecular speed squared.

$\overline{v^2}$

$\overline{\mathrm{KE}}$

- Thermal energy is defined to be the average translational kinetic energy of an atom or molecule.
- The temperature of gases is proportional to the average translational kinetic energy of atoms and molecules: or .

$$\overline{\mathrm{KE}} = \frac{1}{2}m\overline{v^2} = \frac{3}{2}\mathrm{kT} \qquad \sqrt{\overline{v^2}} = v_{\mathrm{rms}} = \sqrt{\frac{3\mathrm{kT}}{m}}$$

- 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.

Conceptual Questions

1. How is momentum related to the pressure exerted by a gas? Explain on the atomic and molecular level, considering the behavior of atoms and molecules.

Problems & Exercises

1. 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?
2. Average atomic and molecular speeds (v_{rms}) are large, even at low temperatures. What is v_{rms} for helium atoms at 5.00 K, just one degree above helium's liquefaction temperature?
3. (a) What is the average kinetic energy in joules of hydrogen atoms on the 5500ºC 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 6.00×10^5 K?
4. The escape velocity of any object from Earth is 11.2 km/s. (a) Express this speed in m/s and km/h. (b) At what temperature would oxygen molecules (molecular mass is equal to 32.0 g/mol) have an average velocity v_{rms} equal to Earth's escape velocity of 11.1 km/s?
5. The escape velocity from the Moon is much smaller than from Earth and is only 2.38 km/s. At what temperature would hydrogen molecules (molecular mass is equal to 2.016 g/mol) have an average velocity v_{rms} equal to the Moon's escape velocity?
6. 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 6.40×10^{-14} J. What temperature is needed?
7. Suppose that the average velocity (v_{rms}) of carbon dioxide molecules (molecular mass is equal to 44.0 g/mol) in a flame is found to be 1.05×10^5 m/s. What temperature does this represent?
8. Hydrogen molecules (molecular mass is equal to 2.016 g/mol) have an average velocity v_{rms} equal to 193 m/s. What is the temperature?
9. Much of the gas near the Sun is atomic hydrogen. Its temperature would have to be 1.5×10^7 K for the average velocity v_{rms} to equal the escape velocity from the Sun. What is that velocity?
10. 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. One of the techniques for separating them (gas diffusion) is based on the different average velocities v_{rms} of uranium hexafluoride gas, UF_6. (a) The molecular masses for $^{235}U\ UF_6$ and $^{238}U\ UF_6$ are 349.0 g/mol and 352.0 g/mol, respectively. What is the ratio of their average velocities? (b) At what temperature would their average velocities differ by 1.00 m/s? (c) Do your answers in this problem imply that this technique may be difficult?

Glossary

thermal energy: $\overline{\text{KE}}$, the average translational kinetic energy of a molecule

Selected Solutions to Problems & Exercises

1. 1.25×10^3 m/s
3. (a) 1.20×10^{-19} J; (b) 1.24×10^{-17} J
5. 458 K
7. 1.95×10^7 K
9. 6.09×10^5 m/s

Phase Changes

Lumen Learning

Learning Objectives

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

- Interpret a phase diagram.
- State Dalton’s law.
- Identify and describe the triple point of a gas from its phase diagram.
- Describe the state of equilibrium between a liquid and a gas, a liquid and a solid, and a gas and a solid.

Up to now, we have considered the behavior of ideal gases. Real gases are like ideal gases at high temperatures. At lower temperatures, however, the interactions between the molecules and their volumes cannot be ignored. The molecules are very close (condensation occurs) and there is a dramatic decrease in volume, as seen in Figure 1. The substance changes from a gas to a liquid. When a liquid is cooled to even lower temperatures, it becomes a solid. The volume never reaches zero because of the finite volume of the molecules.

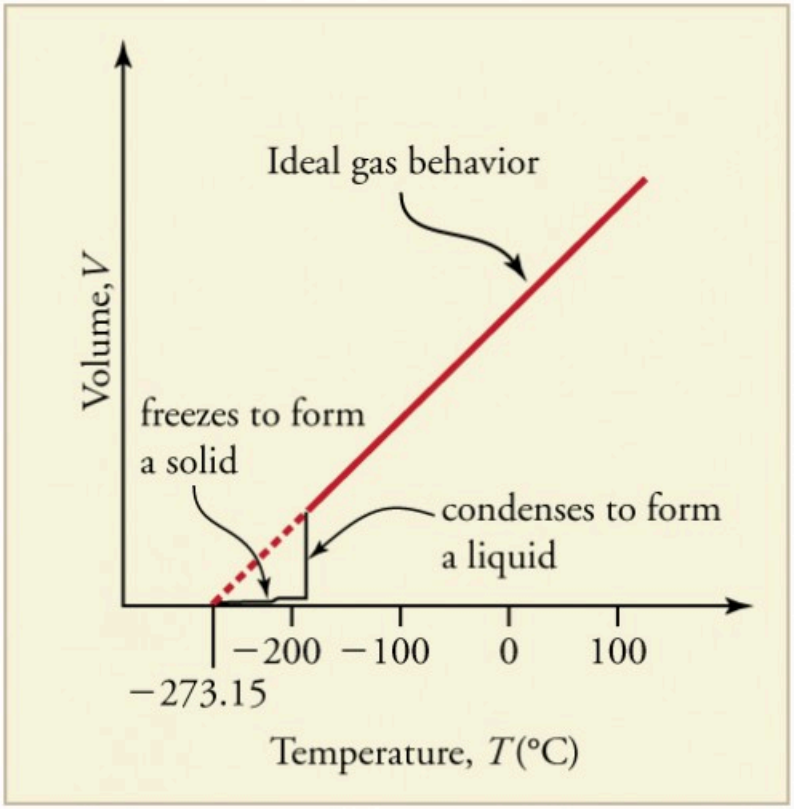


Figure 1. A sketch of volume versus temperature for a real gas at constant pressure. The linear (straight line) part of the graph represents ideal gas behavior—volume and temperature are directly and positively related and the line extrapolates to zero volume at –273.15ºC, or absolute zero. When the gas becomes a liquid, however, the volume actually decreases precipitously at the liquefaction point. The volume decreases slightly once the substance is solid, but it never becomes zero.

High pressure may also cause a gas to change phase to a liquid. Carbon dioxide, for example, is a gas at room temperature and atmospheric pressure, but becomes a liquid under sufficiently high pressure. If the pressure is reduced, the temperature drops and the liquid carbon dioxide solidifies into a snow-like substance at the temperature –78ºC. Solid CO_2 is called "dry ice." Another example of a gas that can be in a liquid phase is liquid nitrogen (LN_2). LN_2 is made by liquefaction of atmospheric air (through compression and cooling). It boils at 77 K (–196ºC) at atmospheric pressure. LN_2 is useful as a refrigerant and allows for the preservation of blood, sperm, and other biological materials. It is also used to reduce noise in electronic sensors and equipment, and to help cool down their current-carrying wires. In dermatology, LN_2 is used to freeze and painlessly remove warts and other growths from the skin.

PV Diagrams

We can examine aspects of the behavior of a substance by plotting a graph of pressure versus volume, called a *PV diagram*. When the substance behaves like an ideal gas, the ideal gas law describes the relationship between its pressure and volume. That is, $PV = NkT$ (ideal gas).

Now, assuming the number of molecules and the temperature are fixed, PV = constant (ideal gas, constant temperature).

For example, the volume of the gas will decrease as the pressure increases. If you plot the relationship PV = constant on a PV diagram, you find a hyperbola. Figure 2 shows a graph of pressure versus volume. The hyperbolas represent ideal-gas behavior at various fixed temperatures, and are called *isotherms*. At lower temperatures, the curves begin to look less like hyperbolas—the gas is not behaving ideally and may even contain liquid. There is a *critical point*—that is, a *critical temperature*—above which liquid cannot exist. At sufficiently high pressure above the critical point, the gas will have the density of a liquid but will not condense. Carbon dioxide, for example, cannot be liquefied at a temperature above 31.0ºC. *Critical pressure* is the minimum pressure needed for liquid to exist at the critical temperature. Table 1 lists representative critical temperatures and pressures.

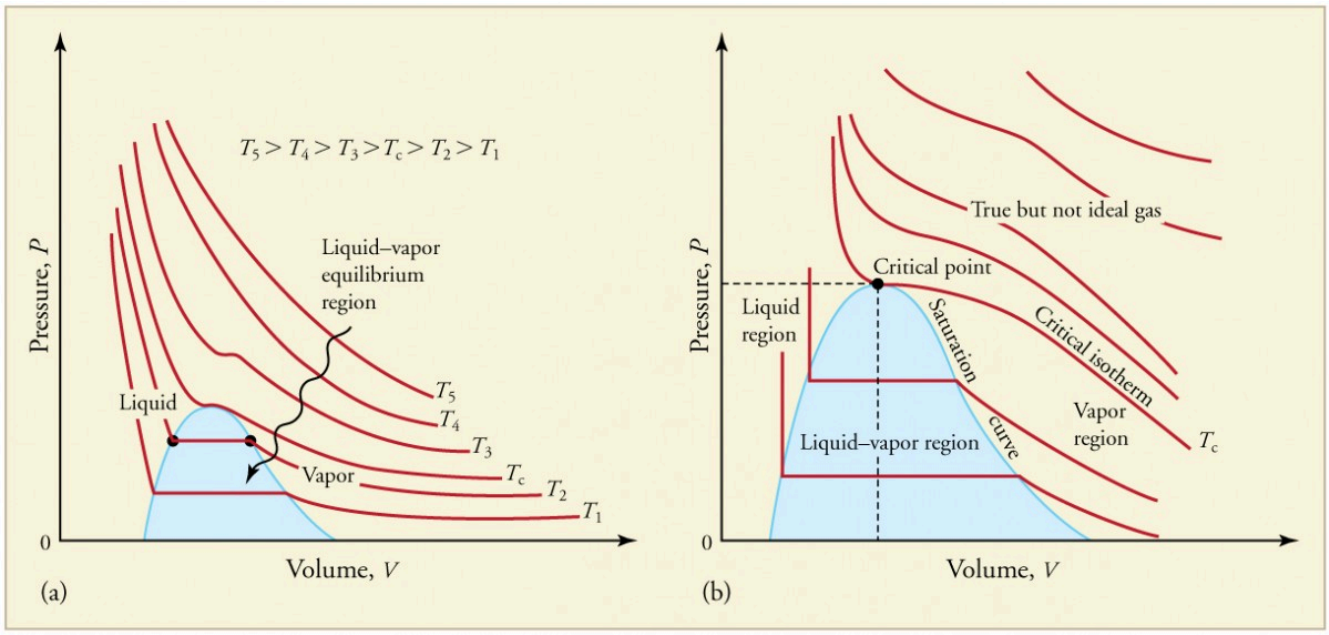


Figure 2. PV diagrams. (a) Each curve (isotherm) represents the relationship between P and V at a fixed temperature; the upper curves are at higher temperatures. The lower curves are not hyperbolas, because the gas is no longer an ideal gas. (b) An expanded portion of the diagram for low temperatures, where the phase can change from a gas to a liquid. The term "vapor" refers to the gas phase when it exists at a temperature below the boiling temperature.

Table 1. Critical Temperatures and Pressures

Substance	**Critical temperature**		**Critical pressure**	
	K	**° C**	**Pa**	**atm**
Water	647.4	374.3	22.12×10^6	219.0
Sulfur dioxide	430.7	157.6	7.88×10^6	78.0
Ammonia	405.5	132.4	11.28×10^6	111.7
Carbon dioxide	304.2	31.1	7.39×10^6	73.2
Oxygen	154.8	−118.4	5.08×10^6	50.3
Nitrogen	126.2	−146.9	3.39×10^6	33.6
Hydrogen	33.3	−239.9	1.30×10^6	12.9
Helium	5.3	−267.9	0.229×10^6	2.27

Phase Diagrams

The plots of pressure versus temperatures provide considerable insight into thermal properties of substances. There are well-defined regions on these graphs that correspond to various phases of matter, so *PT* graphs are called *phase diagrams*. Figure 3 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 lines—boundaries between phases—indicate temperatures and pressures at which the phases coexist (that is, they exist together in ratios, depending on pressure and temperature). For example, the boiling point of water is 100ºC at 1.00 atm. As the pressure increases, the boiling temperature rises steadily to 374ºC at a pressure of 218 atm. A pressure cooker (or even a covered pot) will cook food faster because the water can exist as a liquid at temperatures greater than 100ºC without all boiling away. The curve ends at a point called the *critical point*, because at higher temperatures the liquid phase does not exist at any pressure. The critical point occurs at the critical temperature, as you can see for water from Table 1. The critical temperature for oxygen is –118ºC, so oxygen cannot be liquefied above this temperature.

Similarly, the curve between the solid and liquid regions in Figure 3 gives the melting temperature at various pressures. For example, the melting point is 0ºC at 1.00 atm, as expected. Note that, at a fixed temperature, you can change the phase from solid (ice) to liquid (water) by increasing the pressure. Ice melts from pressure in the hands of a snowball maker. From the phase diagram, we can also say that the melting temperature of ice rises with increased pressure. 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.

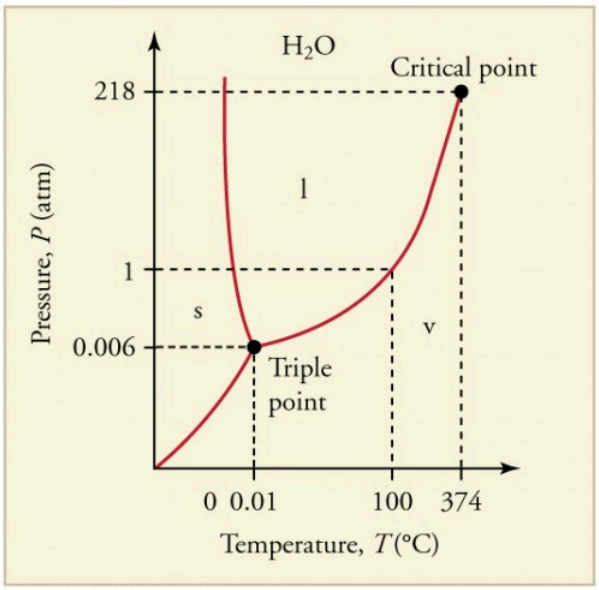


Figure 3. The phase diagram (PT graph) for water. Note that the axes are nonlinear and the graph is not to scale. This graph is simplified—there are several other exotic phases of ice at higher pressures.

At sufficiently low pressures there is no liquid phase, but 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*. It accounts for large losses of snow pack that never make it into a river, the routine automatic defrosting of a freezer, and the freeze-drying process applied to many foods. Carbon dioxide, on the other hand, sublimates at standard atmospheric pressure of 1 atm. (The solid form of CO_2 is known as dry ice because it does not melt. Instead, it moves directly from the solid to the gas state.)

All three curves on the phase diagram meet at a single point, the *triple point*, where all three phases exist in equilibrium. For water, the triple point occurs at 273.16 K (0.01ºC), and is a more accurate calibration temperature than the melting point of water at 1.00 atm, or 273.15 K (0.0ºC). See Table 2 for the triple point values of other substances.

Equilibrium

Liquid and gas phases are in equilibrium at the boiling temperature. (See Figure 4.) 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 relative amount. Molecules in the liquid escape as a gas at the same rate at which gas molecules stick to the liquid, or form droplets and become part of the liquid phase. The combination of temperature and pressure has to be “just right”; if the temperature and pressure are increased, equilibrium is maintained by the same increase of boiling and condensation rates.

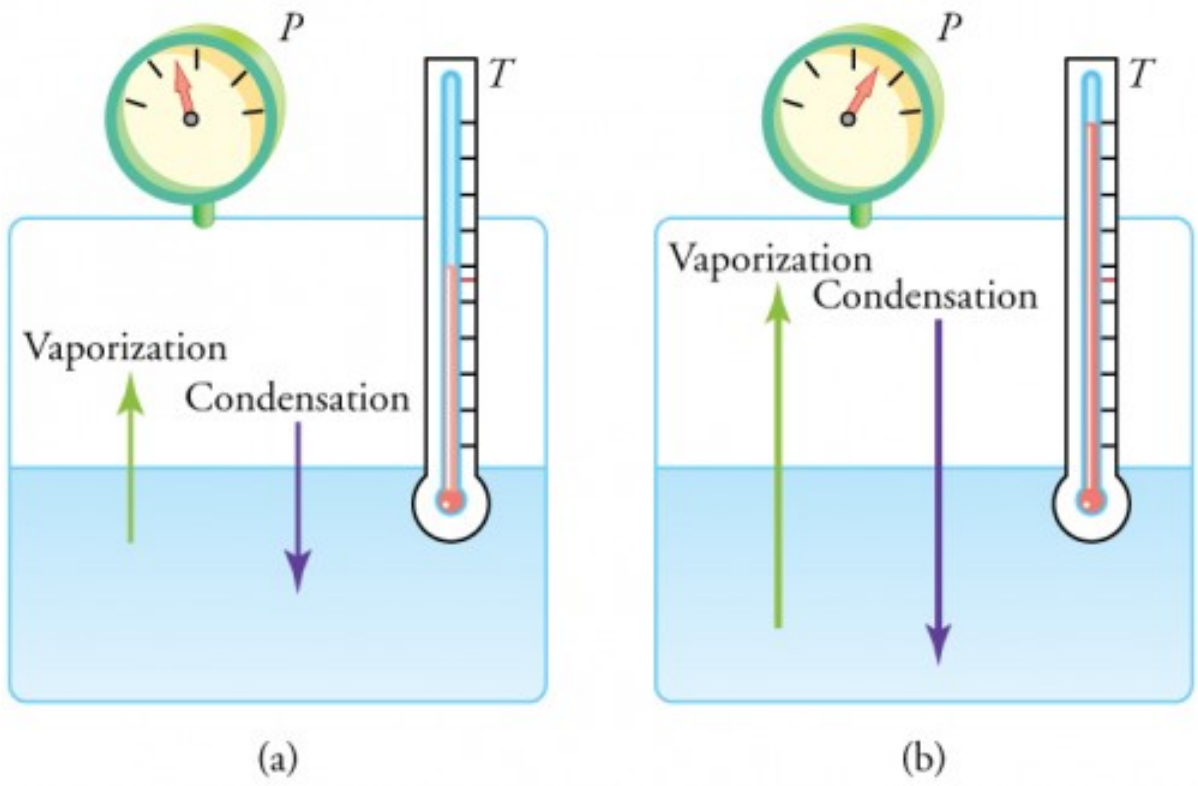


Figure 4. Equilibrium between liquid and gas at two different boiling points inside a closed container. (a) The rates of boiling and condensation are equal at this combination of temperature and pressure, so the liquid and gas phases are in equilibrium. (b) At a higher temperature, the boiling rate is faster and the rates at which molecules leave the liquid and enter the gas are also faster. Because there are more molecules in the gas, the gas pressure is higher and the rate at which gas molecules condense and enter the liquid is faster. As a result the gas and liquid are in equilibrium at this higher temperature.

Table 2. Triple Point Temperatures and Pressures

Substance	**Temperature**		**Pressure**	
	K	**° C**	**Pa**	**atm**
Water	273.16	0.01	6.10×10^2	0.00600
Carbon dioxide	216.55	−56.60	5.16×10^5	5.11
Sulfur dioxide	197.68	−75.47	1.67×10^3	0.0167
Ammonia	195.40	−77.75	6.06×10^3	0.0600
Nitrogen	63.18	−210.0	1.25×10^4	0.124
Oxygen	54.36	−218.8	1.52×10^2	0.00151
Hydrogen	13.84	−259.3	7.04×10^3	0.0697

One example of equilibrium between liquid and gas is that of water and steam at 100ºC and 1.00 atm. This temperature is the boiling point at that pressure, so they should exist in equilibrium. Why does an open pot of water at 100ºC 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 100ºC and 1.00 atm, they would coexist—but with air over the pot, there are fewer water molecules to condense, and water boils. What about water at 20.0ºC and 1.00 atm? This temperature and pressure correspond to the liquid region, yet an open glass of water at this temperature will completely evaporate. Again, the gas around it is air and not pure water vapor, so that the reduced evaporation rate is greater than the condensation rate of water from dry air. If the glass is sealed, then the liquid phase remains. We call the gas phase a *vapor* when it exists, as it does for water at 20.0ºC, at a temperature below the boiling temperature.

Check Your Understanding

Explain why a cup of water (or soda) with ice cubes stays at 0ºC, even on a hot summer day.

Solution

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.)

Vapor Pressure, Partial Pressure, and Dalton's Law

Vapor pressure is defined as the pressure at which a gas coexists with its solid or liquid phase. Vapor pressure is created by faster molecules that break away from the liquid or solid and enter the gas phase. The vapor pressure of a substance depends on both the substance and its temperature—an increase in temperature increases the vapor pressure.

Partial pressure is defined as the pressure a gas would create if it occupied the total volume available. In a mixture of gases, *the total pressure is the sum of partial pressures of the component gases*, assuming ideal gas behavior and no chemical reactions between the components. This law is known as *Dalton's law of partial pressures*, after the English scientist John Dalton (1766–1844), who proposed it. Dalton's law is based on kinetic theory, where each gas creates its pressure by molecular collisions, independent of other gases present. It is consistent with the fact that pressures add according to Pascal's Principle. Thus water evaporates and ice sublimates when their vapor pressures exceed the partial pressure of water vapor in the surrounding mixture of gases. If their vapor pressures are less than the partial pressure of water vapor in the surrounding gas, liquid droplets or ice crystals (frost) form.

Check Your Understanding

Is energy transfer involved in a phase change? If so, will energy have to be supplied to change phase from solid to liquid and liquid to gas? What about gas to liquid and liquid to solid? Why do they spray the orange trees with water in Florida when the temperatures are near or just below freezing?

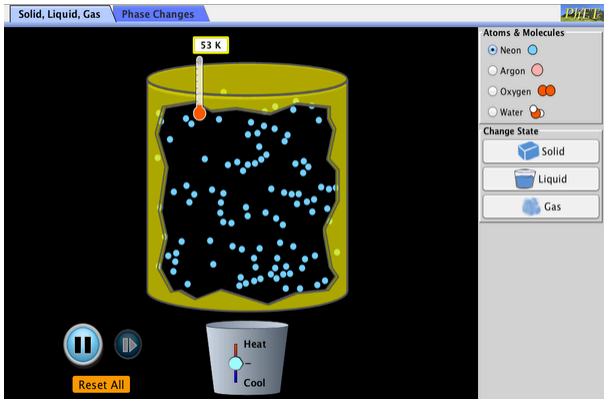
Solution

Yes, energy transfer is involved in a phase change. We know that atoms and molecules in solids and liquids are bound to each other because we know that force is required to separate them. So in a phase change from solid to liquid and liquid to gas, a force must be exerted, perhaps by collision, to separate atoms and

molecules. Force exerted through a distance is work, and energy is needed to do work to go from solid to liquid and liquid to gas. This is intuitively consistent with the need for energy to melt ice or boil water. The converse is also true. Going from gas to liquid or liquid to solid involves atoms and molecules pushing together, doing work and releasing energy.

PhET Explorations: States of Matter—Basics

Heat, cool, and compress atoms and molecules and watch as they change between solid, liquid, and gas phases.



Click to download the simulation. Run using Java.

Section Summary

- Most substances have three distinct phases: gas, liquid, and solid.
- Phase changes among the various phases of matter depend on temperature and pressure.
- The existence of the three phases with respect to pressure and temperature can be described in a phase diagram.
- Two phases coexist (i.e., they are in thermal equilibrium) at a set of pressures and temperatures. These are described as a line on a phase diagram.
- The three phases coexist at a single pressure and temperature. This is known as the triple point and is described by a single point on a phase diagram.
- A gas at a temperature below its boiling point is called a vapor.
- Vapor pressure is the pressure at which a gas coexists with its solid or liquid phase.
- Partial pressure is the pressure a gas would create if it existed alone.
- Dalton’s law states that the total pressure is the sum of the partial pressures of all of the gases

present.

Conceptual Questions

1. A pressure cooker contains water and steam in equilibrium at a pressure greater than atmospheric pressure. How does this greater pressure increase cooking speed?
2. Why does condensation form most rapidly on the coldest object in a room—for example, on a glass of ice water?
3. What is the vapor pressure of solid carbon dioxide (dry ice) at –78.5ºC?

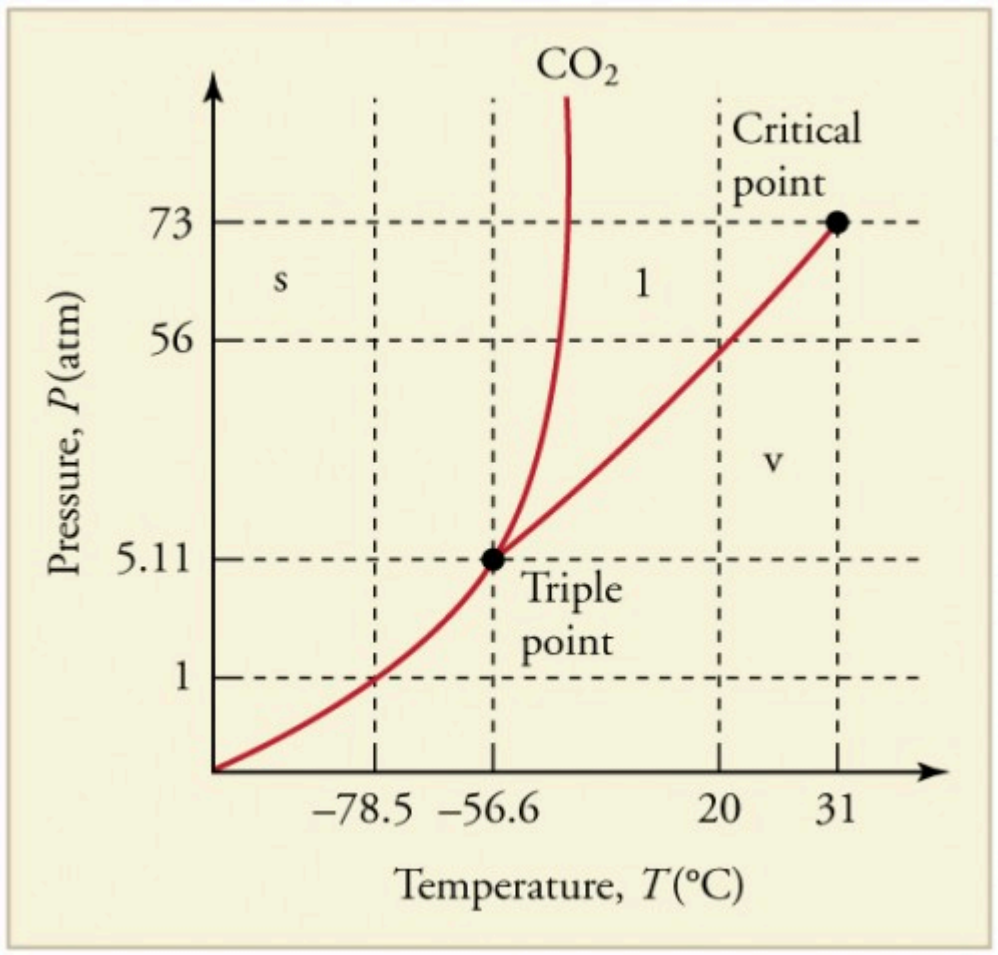


Figure 5. The phase diagram for carbon dioxide. The axes are nonlinear, and the graph is not to scale. Dry ice is solid carbon dioxide and has a sublimation temperature of –78.5ºC.

4. Can carbon dioxide be liquefied at room temperature (20ºC)? If so, how? If not, why not? (See Figure 5)
5. Oxygen cannot be liquefied at room temperature by placing it under a large enough pressure to force its molecules together. Explain why this is.
6. What is the distinction between gas and vapor?

Glossary

PV diagram: a graph of pressure vs. volume

critical point: the temperature above which a liquid cannot exist

critical temperature: the temperature above which a liquid cannot exist

critical pressure: the minimum pressure needed for a liquid to exist at the critical temperature

vapor: a gas at a temperature below the boiling temperature

vapor pressure: the pressure at which a gas coexists with its solid or liquid phase

phase diagram: a graph of pressure vs. temperature of a particular substance, showing at which pressures and temperatures the three phases of the substance occur

triple point: the pressure and temperature at which a substance exists in equilibrium as a solid, liquid, and gas

sublimation: the phase change from solid to gas

partial pressure: the pressure a gas would create if it occupied the total volume of space available

Dalton's law of partial pressures: the physical law that states that the total pressure of a gas is the sum of partial pressures of the component gases

Video: Phase Changes

Lumen Learning

Watch the following Physics Concept Trailer to examine the energy requied to melt ice and how the rising global temperatures contribute to melting ice caps.

A YouTube element has been excluded from this version of the text. You can view it online here: https://pressbooks.nscc.ca/heatlightsound/?p=75

Humidity, Evaporation, and Boiling

Lumen Learning

Learning Objectives

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

- Explain the relationship between vapor pressure of water and the capacity of air to hold water vapor.
- Explain the relationship between relative humidity and partial pressure of water vapor in the air.
- Calculate vapor density using vapor pressure.
- Calculate humidity and dew point.

The expression “it’s not the heat, it’s the humidity” makes a valid point. We keep cool in hot weather by evaporating sweat from our skin and water from our breathing passages. Because evaporation is inhibited by high humidity, we feel hotter at a given temperature when the humidity is high. Low humidity, on the other hand, can cause discomfort from excessive drying of mucous membranes and can lead to an increased risk of respiratory infections.

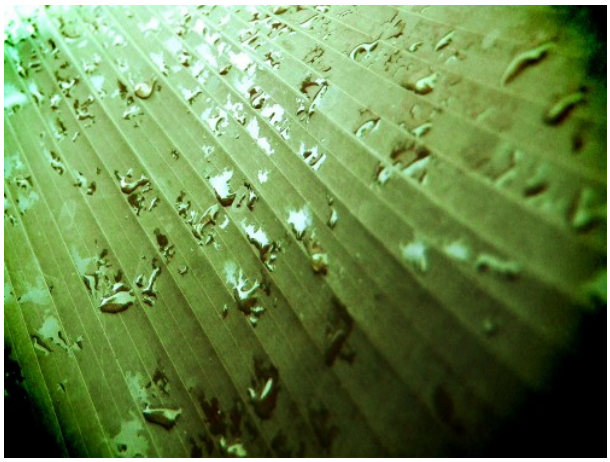


Figure 1. Dew drops like these, on a banana leaf photographed just after sunrise, form when the air temperature drops to or below the dew point. At the dew point, the air can no longer hold all of the water vapor it held at higher temperatures, and some of the water condenses to form droplets. (credit: Aaron Escobar, Flickr)

When we say humidity, we really mean *relative humidity*. Relative humidity tells us how much water vapor is in the air compared with the maximum possible. At its maximum, denoted as *saturation*, the relative humidity is 100%, and evaporation is inhibited. The amount of water vapor the air can hold depends on its temperature. For example, relative humidity rises in the evening, as air temperature declines, sometimes reaching the *dew point*. At the dew point temperature, relative humidity is 100%, and fog may result from the condensation of water droplets if they are small enough to stay in suspension. Conversely, if you wish to dry something (perhaps your hair), it is more effective to blow hot air over it rather than cold air, because, among other things, hot air can hold more water vapor.

The capacity of air to hold water vapor is based on vapor pressure of water. The liquid and solid phases are continuously giving off vapor because some of the molecules have high enough speeds to enter the gas phase; see Figure 2a. If a lid is placed over the container, as in Figure 2b, evaporation continues, increasing the pressure, until sufficient vapor has built up for condensation to balance evaporation.

Then equilibrium has been achieved, and the vapor pressure is equal to the partial pressure of water in the container. Vapor pressure increases with temperature because molecular speeds are higher as temperature increases. Table 1 gives representative values of water vapor pressure over a range of temperatures.

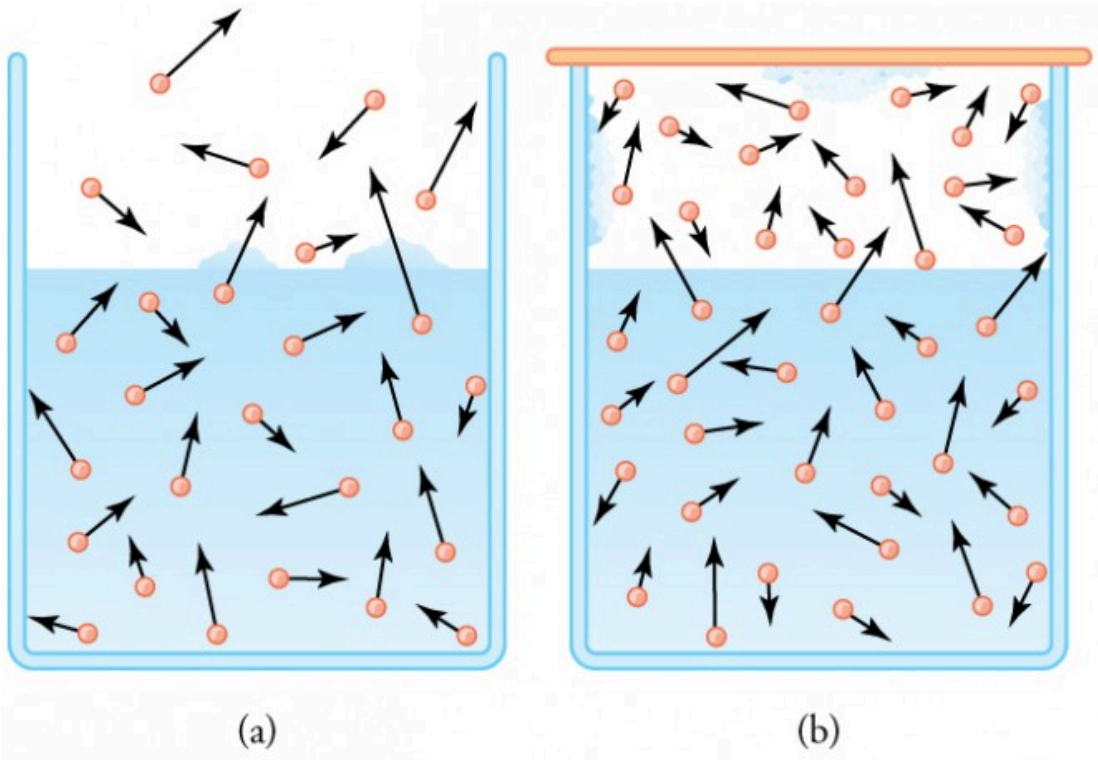


Figure 2. (a) Because of the distribution of speeds and kinetic energies, some water molecules can break away to the vapor phase even at temperatures below the ordinary boiling point. (b) If the container is sealed, evaporation will continue until there is enough vapor density for the condensation rate to equal the evaporation rate. This vapor density and the partial pressure it creates are the saturation values. They increase with temperature and are independent of the presence of other gases, such as air. They depend only on the vapor pressure of water.

Relative humidity is related to the partial pressure of water vapor in the air. At 100% humidity, the partial pressure is equal to the vapor pressure, and no more water can enter the vapor phase. If the partial pressure is less than the vapor pressure, then evaporation will take place, as humidity is less than 100%. If the partial pressure is greater than the vapor pressure, condensation takes place. The capacity of air to “hold” water vapor is determined by the vapor pressure of water and has nothing to do with the properties of air.

Table 1. Saturation Vapor Density of Water

Temperature (°C)	Vapor pressure (Pa)	Saturation vapor density (g/m^3)
−50	4.0	0.039
−20	1.04×10^2	0.89
−10	2.60×10^2	2.36
0	6.10×10^2	4.84
5	8.68×10^2	6.80
10	1.19×10^3	9.40
15	1.69×10^3	12.8
20	2.33×10^3	17.2
25	3.17×10^3	23.0
30	4.24×10^3	30.4
37	6.31×10^3	44.0
40	7.34×10^3	51.1
50	1.23×10^4	82.4
60	1.99×10^4	130
70	3.12×10^4	197
80	4.73×10^4	294
90	7.01×10^4	418
95	8.59×10^4	505
100	**1.01×10^5**	**598**
120	1.99×10^5	1095
150	4.76×10^5	2430
200	1.55×10^6	7090
220	2.32×10^6	10,200

Example 1. Calculating Density Using Vapor Pressure

Table 1 gives the vapor pressure of water at 20.0°C as 2.33×10^3 Pa Use the ideal gas law to calculate the density of water vapor in g/m^3 that would create a partial pressure equal to this vapor pressure. Compare the result with the saturation vapor density given in the table.

Strategy

To solve this problem, we need to break it down into a two steps. The partial pressure follows the ideal gas law, $PV = nRT$, where n is the number of moles. If we solve this equation for n/V to calculate the number of moles per cubic meter, we can then convert this quantity to grams per cubic meter as requested. To do this, we need to use the molecular mass of water, which is given in the periodic table.

Solution

1. Identify the knowns and convert them to the proper units:

   temperature T = 20ºC = 293 K

   vapor pressure P of water at 20ºC is 2.33×10^3 Pa

   molecular mass of water is 18.0 g/mol

2. Solve the ideal gas law for

$$\frac{n}{V}$$

$$:$$

$$\frac{n}{V} = \frac{P}{RT}$$

3. Substitute known values into the equation and solve for n/V.

$$\frac{n}{V} = \frac{P}{RT} = \frac{2.33 \times 10^3 \text{ Pa}}{(8.31 \text{ J/mol} \cdot \text{ K})(293 \text{ K})} = 0.957 \text{ mol/m}^3$$

4. Convert the density in moles per cubic meter to grams per cubic meter.

$$\rho = \left(0.957 \frac{\text{mol}}{\text{m}^3}\right)\left(\frac{18.0 \text{ g}}{\text{mol}}\right) = 17.2 \text{ g/m}^3$$

Discussion

The density is obtained by assuming a pressure equal to the vapor pressure of water at 20.0ºC. The density found is identical to the value in Table 1, which means that a vapor density of 17.2 g/m^3 at 20.0ºC creates a partial pressure of 2.33×10^3 Pa, equal to the vapor pressure of water at that temperature. If the partial pressure is equal to the vapor pressure, then the liquid and vapor phases are in equilibrium, and the relative humidity is 100%. Thus, there can be no more than 17.2 g of water vapor per m^3 at 20.0ºC, so that this value is the saturation vapor density at that temperature. This example illustrates how water vapor behaves like an ideal gas: the pressure and density are consistent with the ideal gas law (assuming the density in the table is correct). The saturation vapor densities listed in Table 1 are the maximum amounts of water vapor that air can hold at various temperatures.

Percent Relative Humidity

We define *percent relative humidity* as the ratio of vapor density to saturation vapor density, or

$$\text{percent relative humidity} = \frac{\text{vapor density}}{\text{saturation vapor density}} \times 100$$

We can use this and the data in Table 1 to do a variety of interesting calculations, keeping in mind that relative humidity is based on the comparison of the partial pressure of water vapor in air and ice.

Example 2. Calculating Humidity and Dew Point

1. Calculate the percent relative humidity on a day when the temperature is 25.0ºC and the air contains 9.40 g of water vapor per m^3.

2. At what temperature will this air reach 100% relative humidity (the saturation density)? This temperature is the dew point.

3. What is the humidity when the air temperature is 25.0ºC and the dew point is –10.0ºC?

Strategy and Solution

1. Percent relative humidity is defined as the ratio of vapor density to saturation vapor density

$$\text{percent relative humidity} = \frac{\text{vapor density}}{\text{saturation vapor density}} \times 100$$

2. The first is given to be 9.40 g/m^3, and the second is found in Table 1 to be 23.0 g/m^3. Thus,

$$\text{percent relative humidity} = \frac{9.40\ \text{g/m}^3}{23.0\ \text{g/m}^3} \times 100 = 40.9\%$$

3. The air contains 9.40 g/m^3 of water vapor. The relative humidity will be 100% at a temperature where 9.40 g/m^3 is the saturation density. Inspection of Table 1 reveals this to be the case at 10.0ºC, where the relative humidity will be 100%. That temperature is called the dew point for air with this concentration of water vapor.

Here, the dew point temperature is given to be –10.0ºC. Using Table 1, we see that the vapor density is 2.36 g/m^3, because this value is the saturation vapor density at –10.0ºC. The saturation vapor density at 25.0ºC is seen to be 23.0 g/m^3. Thus, the relative humidity at 25.0ºC is

$$\text{percent relative humidity} = \frac{2.36\ \text{g/m}^3}{23.0\ \text{g/m}^3} \times 100 = 10.3\%$$

Discussion

The importance of dew point is that air temperature cannot drop below 10.0ºC in part (b), or –10.0ºC in part

(c), without water vapor condensing out of the air. If condensation occurs, considerable transfer of heat occurs (discussed in Heat and Heat Transfer Methods), which prevents the temperature from further dropping. When dew points are below 0ºC, freezing temperatures are a greater possibility, which explains why farmers keep track of the dew point. Low humidity in deserts means low dew-point temperatures. Thus condensation is unlikely. If the temperature drops, vapor does not condense in liquid drops. Because no heat is released into the air, the air temperature drops more rapidly compared to air with higher humidity. Likewise, at high temperatures, liquid droplets do not evaporate, so that no heat is removed from the gas to the liquid phase. This explains the large range of temperature in arid regions.

Why does water boil at 100ºC? You will note from Table 1 that the vapor pressure of water at 100ºC is 1.01×10^5 Pa, or 1.00 atm. Thus, it can evaporate without limit at this temperature and pressure. But why does it form bubbles when it boils? This is because water ordinarily contains significant amounts of dissolved air and other impurities, which are observed as small bubbles of air in a glass of water. If a bubble starts out at the bottom of the container at 20ºC, it contains water vapor (about 2.30%). The pressure inside the bubble is fixed at 1.00 atm (we ignore the slight pressure exerted by the water around it). As the temperature rises, the amount of air in the bubble stays the same, but the water vapor increases; the bubble expands to keep the pressure at 1.00 atm. At 100ºC, water vapor enters the bubble continuously since the partial pressure of water is equal to 1.00 atm in equilibrium. It cannot reach this pressure, however, since the bubble also contains air and total pressure is 1.00 atm. The bubble grows in size and thereby increases the buoyant force. The bubble breaks away and rises rapidly to the surface—we call this boiling! (See Figure 3.)

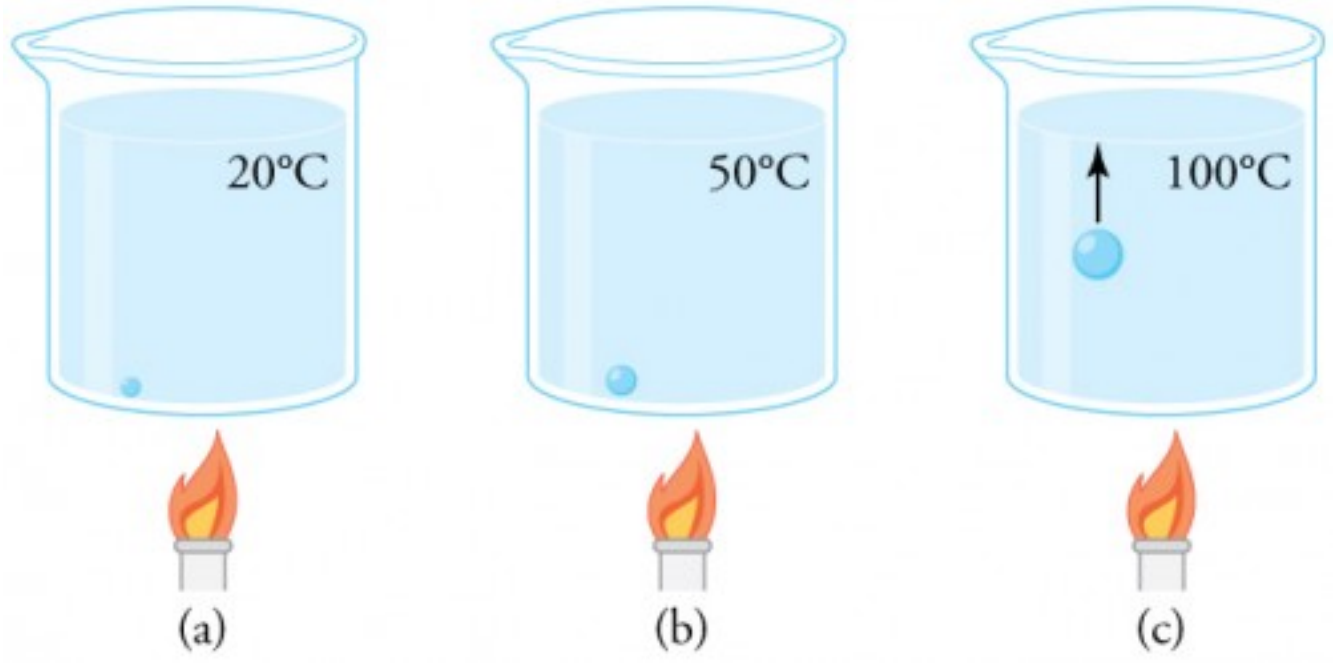


Figure 3. (a) An air bubble in water starts out saturated with water vapor at 20ºC. (b) As the temperature rises, water vapor enters the bubble because its vapor pressure increases. The bubble expands to keep its pressure at 1.00 atm. (c) At 100ºC, water vapor enters the bubble continuously because water's vapor pressure exceeds its partial pressure in the bubble, which must be less than 1.00 atm. The bubble grows and rises to the surface.

Check Your Understanding

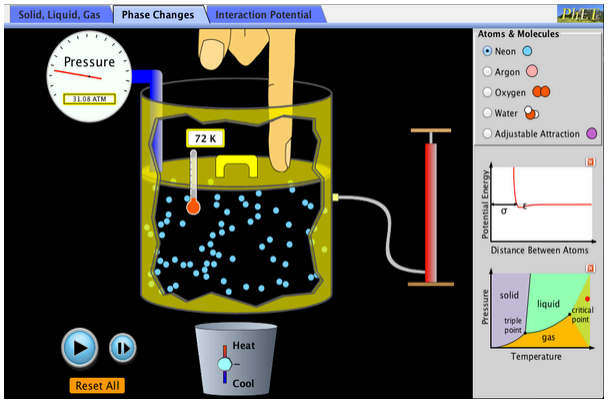
Freeze drying is a process in which substances, such as foods, are dried by placing them in a vacuum chamber and lowering the atmospheric pressure around them. How does the lowered atmospheric pressure speed the drying process, and why does it cause the temperature of the food to drop?

Solution

Decreased the atmospheric pressure results in decreased partial pressure of water, hence a lower humidity. So evaporation of water from food, for example, will be enhanced. The molecules of water most likely to break away from the food will be those with the greatest velocities. Those remaining thus have a lower average velocity and a lower temperature. This can (and does) result in the freezing and drying of the food; hence the process is aptly named freeze drying.

PhET Explorations: States of Matter

Watch different types of molecules form a solid, liquid, or gas. Add or remove heat and watch the phase change. Change the temperature or volume of a container and see a pressure-temperature diagram respond in real time. Relate the interaction potential to the forces between molecules.



Click to download the simulation. Run using Java.

Section Summary

- Relative humidity is the fraction of water vapor in a gas compared to the saturation value.
- The saturation vapor density can be determined from the vapor pressure for a given temperature.
- Percent relative humidity is defined to be

$$\text{percent relative humidity} = \frac{\text{vapor density}}{\text{saturation vapor density}} \times 100$$

.

- The dew point is the temperature at which air reaches 100% relative humidity.

Conceptual Questions

1. Because humidity depends only on water's vapor pressure and temperature, are the saturation vapor densities listed in Table 1 valid in an atmosphere of helium at a pressure of 1.01×10^5 N/m^2, rather than air? Are those values affected by altitude on Earth?
2. Why does a beaker of 40.0ºC water placed in a vacuum chamber start to boil as the chamber is evacuated (air is pumped out of the chamber)? At what pressure does the boiling begin? Would food cook any faster in such a beaker?
3. Why does rubbing alcohol evaporate much more rapidly than water at STP (standard temperature and pressure)?

Problems & Exercises

1. Dry air is 78.1% nitrogen. What is the partial pressure of nitrogen when the atmospheric pressure is 1.01×10^5 N/m^2?
2. (a) What is the vapor pressure of water at 20.0ºC? (b) What percentage of atmospheric pressure does this correspond to? (c) What percent of 20.0ºC air is water vapor if it has 100% relative humidity? (The density of dry air at 20.0ºC is 1.20kg/m3.)
3. Pressure cookers increase cooking speed by raising the boiling temperature of water above its value at atmospheric pressure. (a) What pressure is necessary to raise the boiling point to 120.0ºC? (b) What gauge pressure does this correspond to?
4. (a) At what temperature does water boil at an altitude of 1500 m (about 5000 ft) on a day when atmospheric pressure is 8.59×10^4 N/m^2? (b) What about at an altitude of 3000 m (about 10,000 ft) when atmospheric pressure is 7.00×10^4 N/m^2?
5. What is the atmospheric pressure on top of Mt. Everest on a day when water boils there at a temperature of 70.0ºC?
6. At a spot in the high Andes, water boils at 80.0ºC, greatly reducing the cooking speed of potatoes, for example. What is atmospheric pressure at this location?
7. What is the relative humidity on a 25.0ºC day when the air contains 18.0 g/m^3 of water vapor?
8. What is the density of water vapor in g/m^3 on a hot dry day in the desert when the temperature is 40.0ºC and the relative humidity is 6.00%?
9. A deep-sea diver should breathe a gas mixture that has the same oxygen partial pressure as at sea level, where dry air contains 20.9% oxygen and has a total pressure of 1.01×10^5 N/m^2. (a) What is the partial pressure of oxygen at sea level? (b) If the diver breathes a gas mixture at a pressure of 2.00×10^6 N/m^2, what percent oxygen should it be to have the same oxygen partial pressure as at sea level?

10. The vapor pressure of water at 40.0ºC is 7.34×10^3 N/m^2. Using the ideal gas law, calculate the density of water vapor in g/m^3 that creates a partial pressure equal to this vapor pressure. The result should be the same as the saturation vapor density at that temperature 51.1 g/m^3.
11. Air in human lungs has a temperature of 37.0ºC and a saturation vapor density of 44.0 g/m^3. (a) If 2.00 L of air is exhaled and very dry air inhaled, what is the maximum loss of water vapor by the person? (b) Calculate the partial pressure of water vapor having this density, and compare it with the vapor pressure of 6.31×10^3 N/m^2.
12. If the relative humidity is 90.0% on a muggy summer morning when the temperature is 20.0ºC, what will it be later in the day when the temperature is 30.0ºC, assuming the water vapor density remains constant?
13. Late on an autumn day, the relative humidity is 45.0% and the temperature is 20.0ºC. What will the relative humidity be that evening when the temperature has dropped to 10.0ºC, assuming constant water vapor density?
14. Atmospheric pressure atop Mt. Everest is 3.30×10^4 N/m^2. (a) What is the partial pressure of oxygen there if it is 20.9% of the air? (b) What percent oxygen should a mountain climber breathe so that its partial pressure is the same as at sea level, where atmospheric pressure is 1.01×10^5 N/m^2? (c) One of the most severe problems for those climbing very high mountains is the extreme drying of breathing passages. Why does this drying occur?
15. What is the dew point (the temperature at which 100% relative humidity would occur) on a day when relative humidity is 39.0% at a temperature of 20.0ºC?
16. On a certain day, the temperature is 25.0ºC and the relative humidity is 90.0%. How many grams of water must condense out of each cubic meter of air if the temperature falls to 15.0ºC? Such a drop in temperature can, thus, produce heavy dew or fog.
17. **Integrated Concepts.** The boiling point of water increases with depth because pressure increases with depth. At what depth will fresh water have a boiling point of 150ºC, if the surface of the water is at sea level?
18. **Integrated Concepts.** (a) At what depth in fresh water is the critical pressure of water reached, given that the surface is at sea level? (b) At what temperature will this water boil? (c) Is a significantly higher temperature needed to boil water at a greater depth?
19. **Integrated Concepts.** To get an idea of the small effect that temperature has on Archimedes' principle, calculate the fraction of a copper block's weight that is supported by the buoyant force in 0ºC water and compare this fraction with the fraction supported in 95.ºC water.
20. **Integrated Concepts.** If you want to cook in water at 150ºC, you need a pressure cooker that can withstand the necessary pressure. (a) What pressure is required for the boiling point of water to be this high? (b) If the lid of the pressure cooker is a disk 25.0 cm in diameter, what force must it be able to withstand at this pressure?
21. **Unreasonable Results.** (a) How many moles per cubic meter of an ideal gas are there at a pressure of 1.00×10^{14} N/m^2 and at 0ºC? (b) What is unreasonable about this result? (c) Which premise or assumption is responsible?
22. **Unreasonable Results.** (a) An automobile mechanic claims that an aluminum rod fits loosely into its hole on an aluminum engine block because the engine is hot and the rod is cold. If the hole is 10.0% bigger in diameter than the 22.0ºC rod, at what temperature will the rod be the same size as the hole? (b) What is unreasonable about this temperature? (c) Which premise is responsible?

23. **Unreasonable Results.** The temperature inside a supernova explosion is said to be 2.00×10^{13} K. (a) What would the average velocity v_{rms} of hydrogen atoms be? (b) What is unreasonable about this velocity? (c) Which premise or assumption is responsible?
24. **Unreasonable Results.** Suppose the relative humidity is 80% on a day when the temperature is 30.0ºC. (a) What will the relative humidity be if the air cools to 25.0ºC and the vapor density remains constant? (b) What is unreasonable about this result? (c) Which premise is responsible?

Glossary

dew point: the temperature at which relative humidity is 100%; the temperature at which water starts to condense out of the air

saturation: the condition of 100% relative humidity

percent relative humidity: the ratio of vapor density to saturation vapor density

relative humidity: the amount of water in the air relative to the maximum amount the air can hold

Selected Solutions to Problems & Exercises

1. 7.89×10^4 Pa

3. (a) 1.99×10^5 Pa; (b) 0.97 atm

5. 3.12×10^4 Pa

7. 78.3%

9. (a) 2.12×10^4 Pa; (b) 1.06%

11. (a) 8.80×10^{-2} g; (b) 6.30×10^3 Pa; the two values are nearly identical.

13. 82.3%

15. 4.77ºC

17. 38.3 m

19.

$$\frac{\frac{F_{\text{B}}}{w_{\text{Cu}}}}{\frac{F_{\text{B}}}{w_{\text{Cu}'}}} = 1.02$$

. The buoyant force supports nearly the exact same amount of force on the copper block in both circumstances.

21. (a) 4.41×10^{10} mol/m^3; (b) It's unreasonably large; (c) At high pressures such as these, the ideal gas law can no longer be applied. As a result, unreasonable answers come up when it is used.

23. (a) 7.03×10^8 m/s; (b) The velocity is too high—it's greater than the speed of light; (c) The assumption that hydrogen inside a supernova behaves as an idea gas is responsible, because of the great temperature and

density in the core of a star. Furthermore, when a velocity greater than the speed of light is obtained, classical physics must be replaced by relativity, a subject not yet covered.

4. Heat and Heat Transfer Methods

Introduction to Heat and Heat Transfer Methods

Lumen Learning

Energy can exist in many forms and heat is one of the most intriguing. Heat is often hidden, as it only exists when in transit, and is transferred by a number of distinctly different methods. Heat transfer touches every aspect of our lives and helps us understand how the universe functions. It explains the chill we feel on a clear breezy night, or why Earth's core has yet to cool. This module defines and explores heat transfer, its effects, and the methods by which heat is transferred. These topics are fundamental, as well as practical, and will often be referred to in the modules ahead.

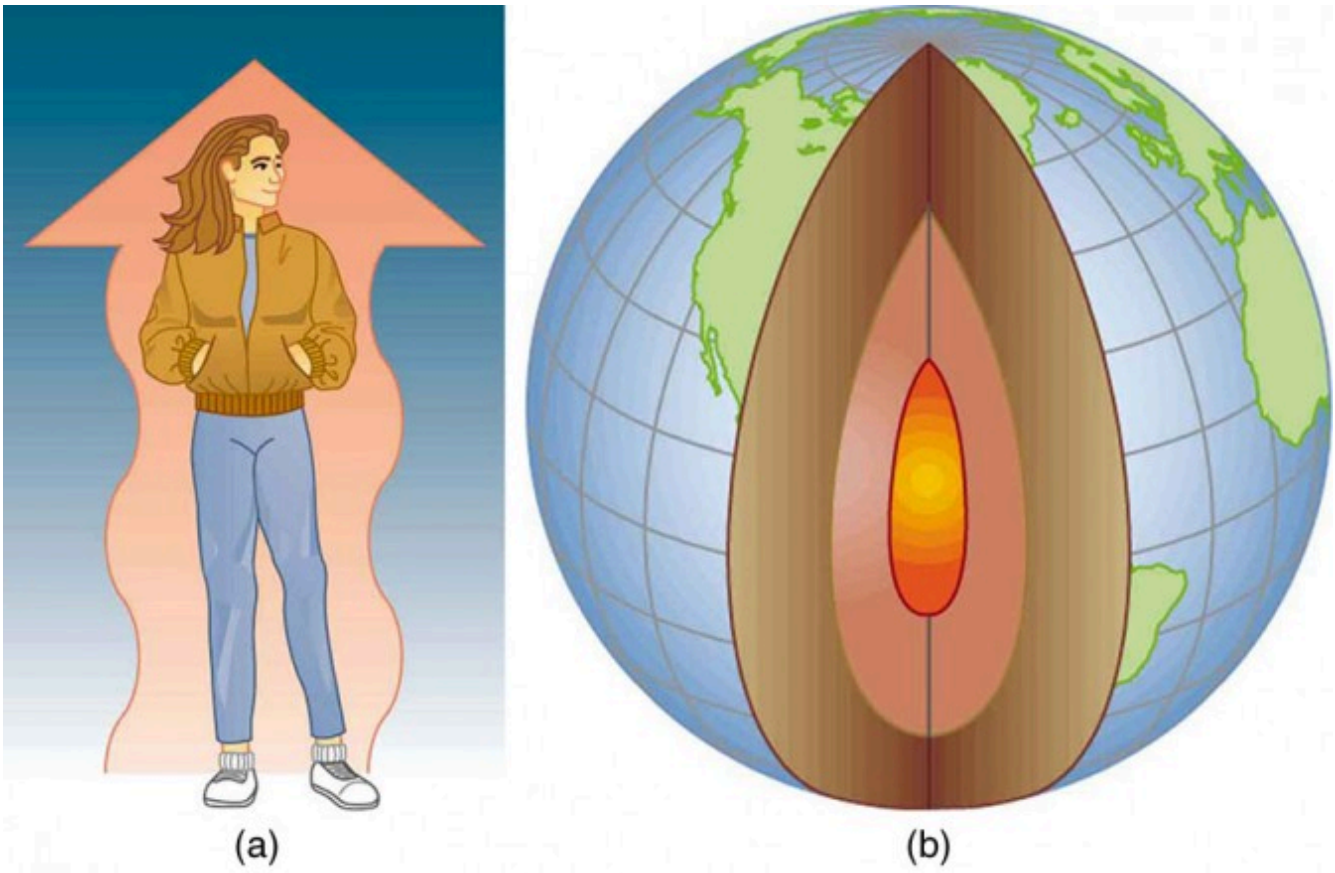


Figure 1. (a) The chilling effect of a clear breezy night is produced by the wind and by radiative heat transfer to cold outer space. (b) There was once great controversy about the Earth's age, but it is now generally accepted to be about 4.5 billion years old. Much of the debate is centered on the Earth's molten interior. According to our understanding of heat transfer, if the Earth is really that old, its center should have cooled off long ago. The discovery of radioactivity in rocks revealed the source of energy that keeps the Earth's interior molten, despite heat transfer to the surface, and from there to cold outer space.

Heat

Lumen Learning

Learning Objectives

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

- Define heat as transfer of energy.

In the chapter Work, Energy, and Energy Resources, we defined work as force times distance and learned that work done on an object changes its kinetic energy. We also saw in Temperature, Kinetic Theory, and the Gas Laws that temperature is proportional to the (average) kinetic energy of atoms and molecules. We say that a thermal system has a certain internal energy: its internal energy is higher if the temperature is higher. If two objects at different temperatures are brought in contact with each other, energy is transferred from the hotter to the colder object until equilibrium is reached and the bodies reach thermal equilibrium (i.e., they are at the same temperature). No work is done by either object, because no force acts through a distance. The transfer of energy is caused by the temperature difference, and ceases once the temperatures are equal. These observations lead to the following definition of *heat*: Heat is the spontaneous transfer of energy due to a temperature difference.

As noted in the chapter Temperature, Kinetic Theory, and the Gas Laws, heat is often confused with temperature. For example, we may say the heat was unbearable, when we actually mean that the temperature was high. Heat is a form of energy, whereas temperature is not. The misconception arises because we are sensitive to the flow of heat, rather than the temperature.

Owing to the fact that heat is a form of energy, it has the SI unit of *joule* (J). The *calorie* (cal) is a common unit of energy, defined as the energy needed to change the temperature of 1.00 g of water by 1.00ºC—specifically, between 14.5ºC and 15.5ºC, since there is a slight temperature dependence. Perhaps the most common unit of heat is the *kilocalorie* (kcal), which is the energy needed to change the temperature of 1.00 kg of water by 1.00ºC. Since mass is most often specified in kilograms, kilocalorie is commonly used. Food calories (given the notation Cal, and sometimes called “big calorie”) are actually kilocalories (1 kilocalorie = 1000 calories), a fact not easily determined from package labeling.

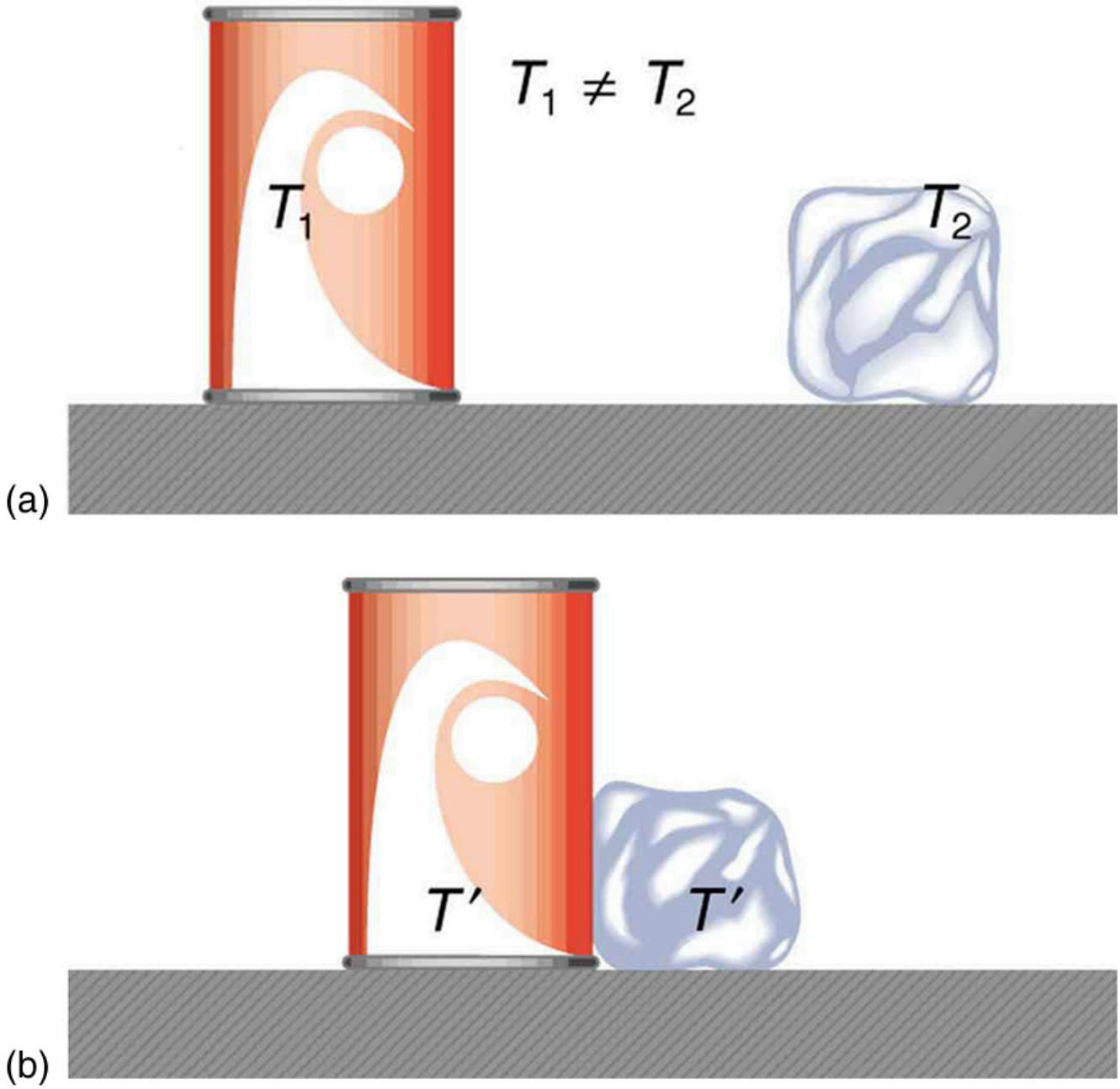


Figure 1. In figure (a) the soft drink and the ice have different temperatures, T_1 and T_2, and are not in thermal equilibrium. In figure (b), when the soft drink and ice are allowed to interact, energy is transferred until they reach the same temperature T', achieving equilibrium. Heat transfer occurs due to the difference in temperatures. In fact, since the soft drink and ice are both in contact with the surrounding air and bench, the equilibrium temperature will be the same for both.

Mechanical Equivalent of Heat

It is also possible to change the temperature of a substance by doing work. Work can transfer energy into or out of a system. This realization helped establish the fact 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 terms of the units used for these two terms, the best modern value for this equivalence is 1.000 kcal = 4186 J.

We consider this equation as the conversion between two different units of energy.

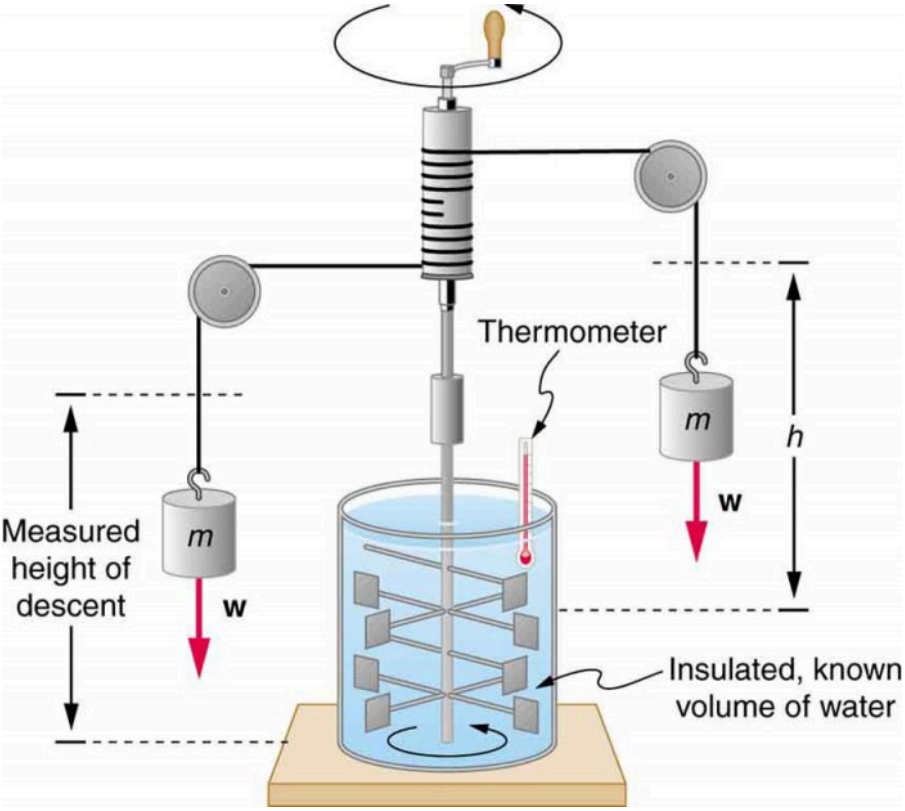


Figure 2. Schematic depiction of Joule's experiment that established the equivalence of heat and work.

Figure 2 above shows one of Joule's most famous experimental setups for demonstrating the mechanical equivalent of heat. It demonstrated that work and heat can produce the same effects, and helped establish the principle of conservation of energy. Gravitational potential energy (PE) (work done by the gravitational force) is converted into kinetic energy (KE), and then randomized by viscosity and turbulence into increased average kinetic energy of atoms and molecules in the system, producing a temperature increase. His contributions to the field of thermodynamics were so significant that the SI unit of energy was named after him.

Heat added or removed from a system changes its internal energy and thus its temperature. Such a temperature increase is observed while cooking. However, adding heat does not necessarily increase the temperature. An example is melting of ice; that is, when a substance changes from one phase to another. Work done on the system or by the system can also change the internal energy of the system. Joule demonstrated that the temperature of a system can be increased by stirring. If an ice cube is rubbed against a rough surface, work is done by the frictional force. A system has a well-defined internal energy, but we cannot say that it has a certain "heat content" or "work content." We use the phrase "heat transfer" to emphasize its nature.

Check Your Understanding

Two samples (A and B) of the same substance are kept in a lab. Someone adds 10 kilojoules (kJ) of heat to one sample, while 10 kJ of work is done on the other sample. How can you tell to which sample the heat was added?

Solution

Heat and work both change the internal energy of the substance. However, the properties of the sample only depend on the internal energy so that it is impossible to tell whether heat was added to sample A or B.

Section Summary

- Heat and work are the two distinct methods of energy transfer.
- Heat is energy transferred solely due to a temperature difference.
- Any energy unit can be used for heat transfer, and the most common are kilocalorie (kcal) and joule (J).
- Kilocalorie is defined to be the energy needed to change the temperature of 1.00 kg of water between 14.5ºC and 15.5ºC.
- The mechanical equivalent of this heat transfer is 1.00 kcal=4186 J.

Conceptual Questions

1. How is heat transfer related to temperature?
2. Describe a situation in which heat transfer occurs. What are the resulting forms of energy?
3. When heat transfers into a system, is the energy stored as heat? Explain briefly.

Glossary

heat: the spontaneous transfer of energy due to a temperature difference

kilocalorie: 1 kilocalorie = 1000 calories

mechanical equivalent of heat: the work needed to produce the same effects as heat transfer

Temperature Change and Heat Capacity

Lumen Learning

Learning Objectives

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

- Observe heat transfer and change in temperature and mass.
- Calculate final temperature after heat transfer between two objects.

One of the major effects of heat transfer is temperature change: heating increases the temperature while cooling decreases it. We assume that there is no phase change and that no work is done on or by the system. Experiments show that the transferred heat depends on three factors—the change in temperature, the mass of the system, and the substance and phase of the substance.

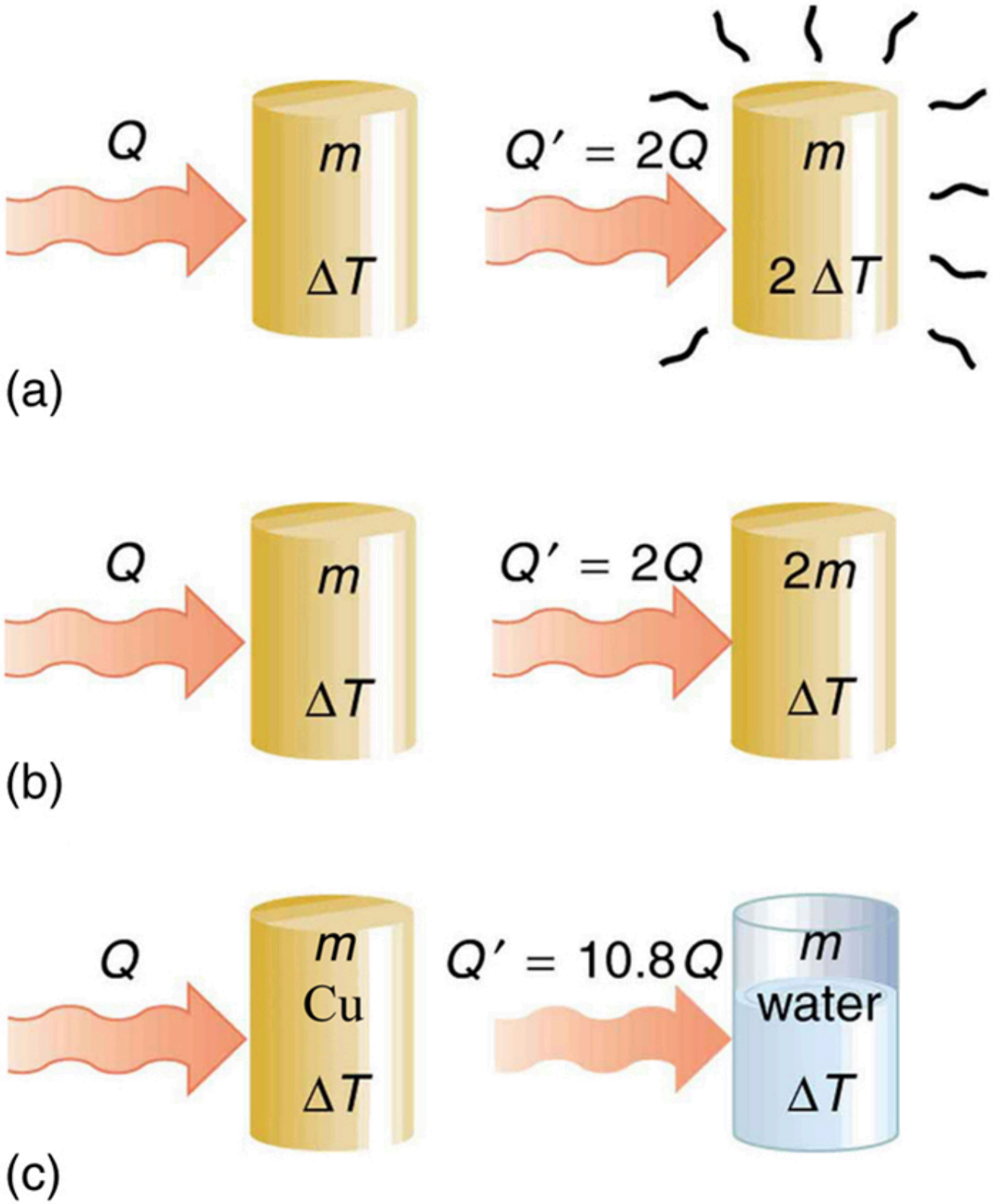


Figure 1. The heat Q transferred to cause a temperature change depends on the magnitude of the temperature change, the mass of the system, and the substance and phase involved. (a) The amount of heat transferred is directly proportional to the temperature change. To double the temperature change of a mass m, you need to add twice the heat. (b) The amount of heat transferred is also directly proportional to the mass. To cause an equivalent temperature change in a doubled mass, you need to add twice the heat. (c) The amount of heat transferred depends on the

substance and its phase. If it takes an amount Q of heat to cause a temperature change ΔT in a given mass of copper, it will take 10.8 times that amount of heat to cause the equivalent temperature change in the same mass of water assuming no phase change in either substance.

The dependence on temperature change and mass are easily understood. Owing to the fact that the (average) kinetic energy of an atom or molecule is proportional to the absolute temperature, the internal energy of a system is proportional to the absolute temperature and the number of atoms or molecules. Owing to the fact that the transferred heat is equal to the change in the internal energy, the heat is proportional to the mass of the substance and the temperature change. The transferred heat also depends on the substance so that, for example, the heat necessary to raise the temperature is less for alcohol than for water. For the same substance, the transferred heat also depends on the phase (gas, liquid, or solid).

Heat Transfer and Temperature Change

The quantitative relationship between heat transfer and temperature change contains all three factors: $Q = mc\Delta T$, where Q is the symbol for heat transfer, m is the mass of the substance, and ΔT is the change in temperature. The symbol c stands for *specific heat* and depends on the material and phase. The specific heat is the amount of heat necessary to change the temperature of 1.00 kg of mass by 1.00ºC. The specific heat c is a property of the substance; its SI unit is J/(kg · K) or J/(kg · ºC). Recall that the temperature change (ΔT) is the same in units of kelvin and degrees Celsius. If heat transfer is measured in kilocalories, then *the unit of specific heat* is kcal/(kg · ºC).

Values of specific heat must generally be looked up in tables, because there is no simple way to calculate them. In general, the specific heat also depends on the temperature. Table 1 lists representative values of specific heat for various substances. Except for gases, the temperature and volume dependence of the specific heat of most substances is weak. We see from this table that the specific heat of water is five times that of glass and ten times that of iron, which means that it takes five times as much heat to raise the temperature of water the same amount as for glass and ten times as much heat to raise the temperature of water as for iron. In fact, water has one of the largest specific heats of any material, which is important for sustaining life on Earth.

Example 1. Calculating the Required Heat: Heating Water in an Aluminum Pan

A 0.500 kg aluminum pan on a stove is used to heat 0.250 liters of water from 20.0ºC to 80.0ºC. (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

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 the pan is 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 1.

Solution

Because water is in thermal contact with the aluminum, the pan and the water are at the same temperature.

Calculate the temperature difference:

$$\Delta T = T_f - T_i = 60.0^oC.$$

Calculate the mass of water. Because the density of water is 1000 kg/m^3, one liter of water has a mass of 1 kg, and the mass of 0.250 liters of water is $m_w = 0.250$ kg.

Calculate the heat transferred to the water. Use the specific heat of water in Table 1:

$$Q_w = m_w c_w \Delta T = (0.250 \text{ kg})(4186 \text{ J/kg}^oC)(60.0^oC) = 62.8 \text{ kJ}.$$

Calculate the heat transferred to the aluminum. Use the specific heat for aluminum in Table 1:

$$Q_{Al} = m_{Al} c_{Al} \Delta T = (0.500 \text{ kg})(900 \text{ J/kg}^oC)(60.0^oC) = 27.0 \times 10^4 \text{ J} = 27.0 \text{ kJ}.<$$

Compare the percentage of heat going into the pan versus that going into the water. First, find the total transferred heat:

$$Q_{Total} = Q_w + Q_{Al} = 62.8 \text{ kJ} + 27.0 \text{ kJ} = 89.8 \text{ kJ}.$$

Thus, the amount of heat going into heating the pan is

$$\frac{27.0 \text{ kJ}}{89.8 \text{ kJ}} \times 100\% = 30.1\%$$

and the amount going into heating the water is

$$\frac{62.8 \text{ kJ}}{89.8 \text{ kJ}} \times 100\% = 69.9\%$$

.

Discussion

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 greater than that of aluminum. Therefore, it takes a bit more than twice the heat to achieve the given temperature change for the water as compared to the aluminum pan.

Example 2. Calculating the Temperature Increase from the Work Done on a Substance: Truck Brakes Overheat on Downhill Runs

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. This conversion prevents the gravitational potential energy from being converted into kinetic energy of the truck. The problem is that the mass of the truck is large compared with that of the brake material absorbing the energy, and the temperature increase may occur too fast for sufficient heat to transfer from the brakes to the environment.

Calculate the temperature increase of 100 kg of brake material with an average specific heat of 800 J/kg · ºC 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.

Figure 2. The smoking brakes on this truck are a visible evidence of the mechanical equivalent of heat.

Strategy

If the brakes are not applied, gravitational potential energy is converted into kinetic energy. When brakes are applied, gravitational potential energy is converted into internal energy of the brake material. We first calculate the gravitational potential energy (*Mgh*) that the entire truck loses in its descent and then find the temperature increase produced in the brake material alone.

Solution

1. Calculate the change in gravitational potential energy as the truck goes downhill *Mgh* = (10,000 kg)(9.80 m/s^2)(75.0 m) = 7.35 × 10^6 J.

$$\Delta T = \frac{Q}{mc}$$

2. Calculate the temperature from the heat transferred using $Q = Mgh$ and , where m is the mass of the brake material. Insert the values m = 100 kg and c = 800 J/kg · ºC to find $\Delta T = \frac{\left(7.35 \times 10^6 \text{ J}\right)}{(100 \text{ kg})(800 \text{ J/kg}^\circ\text{C})} = 92^\circ C$.

Discussion

This temperature is close to the boiling point of water. If the truck had been traveling for some time, then just before the descent, the brake temperature would likely be higher than the ambient temperature. The temperature increase in the descent would likely raise the temperature of the brake material above the boiling point of water, so this technique is not practical. However, the same idea underlies the recent hybrid technology of cars, where mechanical energy (gravitational potential energy) is converted by the brakes into electrical energy (battery).

Table 1. Specific Heats[1] of Various Substances

Substances	**Specific heat (*c*)**	
Solids	J/kg · ºC	kcal/kg · ºC[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	0.4
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)

1. The values for solids and liquids are at constant volume and at 25ºC, except as noted.
2. These values are identical in units of cal/g · ºC.
3. c_v at constant volume and at 20.0ºC, except as noted, and at 1.00 atm average pressure. Values in parentheses are c_p at a constant pressure of 1.00 atm.

Note that Example 2 is an illustration of the mechanical equivalent of heat. Alternatively, the temperature increase could be produced by a blow torch instead of mechanically.

Example 3. Calculating the Final Temperature When Heat Is Transferred Between Two Bodies: Pouring Cold Water in a Hot Pan

Suppose you pour 0.250 kg of 20.0ºC water (about a cup) into a 0.500-kg aluminum pan off the stove with a temperature of 150ºC. Assume that the pan is placed on an insulated pad and that a negligible amount of water boils off. What is the temperature when the water and pan reach thermal equilibrium a short time later?

Strategy

The pan is placed on an insulated pad so that little heat transfer occurs with the surroundings. Originally the pan and water are not in thermal equilibrium: the pan is at a higher temperature than the water. Heat transfer then restores thermal equilibrium once the water and pan are in contact. Because heat transfer between the pan and water takes place rapidly, the mass of evaporated water is negligible and the magnitude of the heat lost by the pan is equal to the heat gained by the water. The exchange of heat stops once a thermal equilibrium between the pan and the water is achieved. The heat exchange can be written as $|Q_{\text{hot}}|=Q_{\text{cold}}$.

Solution

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: $Q_{\text{hot}} = m_{\text{Al}}c_{\text{Al}}(T_{\text{f}} - 150^{\text{o}}\text{C})$.

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: $Q_{\text{cold}} = m_{\text{W}}c_{\text{W}}(T_{\text{f}} - 20.0^{\text{o}}\text{C})$.

Note that $Q_{\text{hot}}<0$ and $Q_{\text{cold}}>0$ and that they must sum to zero because the heat lost by the hot pan must be the same as the heat gained by the cold water:

$$\begin{aligned} Q_{\text{cold}} + Q_{\text{hot}} &= 0 \\ Q_{\text{cold}} &= -Q_{\text{hot}} \\ m_{\text{W}}c_{\text{W}}\left(T_{\text{f}} - 20.0^{\circ}\text{C}\right) &= -m_{\text{Al}}c_{\text{Al}}\left(T_{\text{f}} - 150^{\circ}\text{C}\right) \end{aligned}$$

This an equation for the unknown final temperature, T_{f}.

Bring all terms involving T_{f} on the left hand side and all other terms on the right hand side. Solve for T_{f},

$$T_{\text{f}} = \frac{m_{\text{Al}}c_{\text{Al}}\left(T_{\text{f}} - 150^{\circ}\text{C}\right) + m_{\text{W}}c_{\text{W}}\left(T_{\text{f}} - 20.0^{\circ}\text{C}\right)}{m_{\text{Al}}c_{\text{Al}} + m_{\text{W}}c_{\text{W}}}$$

,

and insert the numerical values:

$$\begin{aligned} T_{\text{f}} &= \frac{(0.500\text{ kg})(900\text{ J/kg}^{\circ}\text{C})(150^{\circ}\text{C}) + (0.250\text{ kg})(4186\text{ J/kg}^{\circ}\text{C})(20.0^{\circ}\text{C})}{(0.500\text{ kg})(900\text{ J/kg}^{\circ}\text{C}) + (0.250\text{ kg})(4186\text{ J/kg}^{\circ}\text{C})} \\ &= \frac{88430\text{ J}}{1496.5\text{ J/}^{\circ}\text{C}} \\ &= 59.1^{\circ}\text{C} \end{aligned}$$

Discussion

This is a typical *calorimetry* problem—two bodies at different temperatures are brought in contact with each other and exchange heat until a common temperature is reached. Why is the final temperature so much closer

to 20.0ºC than 150ºC? The reason is that water has a greater specific heat than most common substances and thus undergoes a small 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 a 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).

Take-Home Experiment: Temperature Change of Land and Water

What heats faster, land or water?

To study differences in heat capacity:

- Place equal masses of dry sand (or soil) and water at the same temperature into two small jars. (The average density of soil or sand is about 1.6 times that of water, so you can achieve approximately equal masses by using 50% more water by volume.)
- Heat both (using an oven or a heat lamp) for the same amount of time.
- Record the final temperature of the two masses.
- Now bring both jars to the same temperature by heating for a longer period of time.
- Remove the jars from the heat source and measure their temperature every 5 minutes for about 30 minutes.

Which sample cools off the fastest? This activity replicates the phenomena responsible for land breezes and sea breezes.

Check Your Understanding

If 25 kJ is necessary to raise the temperature of a block from 25ºC to 30ºC, how much heat is necessary to heat the block from 45ºC to 50ºC?

Solution

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.

Section Summary

- The transfer of heat Q that leads to a change ΔT in the temperature of a body with mass m is $Q = mc\Delta T$, where c is the specific heat of the material. This relationship can also be considered as the definition of specific heat.

Conceptual Questions

1. What three factors affect the heat transfer that is necessary to change an object's temperature?
2. 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 to stop sufficiently fast so that no heat transfers out of the brakes.

Problems & Exercises

1. On a hot day, the temperature of an 80,000-L swimming pool increases by 1.50ºC. What is the net heat transfer during this heating? Ignore any complications, such as loss of water by evaporation.
2. Show that 1 cal/g · ºC =1 kcal/kg · ºC.
3. To sterilize a 50.0-g glass baby bottle, we must raise its temperature from 22.0ºC to 95.0ºC. How much heat transfer is required?
4. 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 20.0ºC: (a) water; (b) concrete; (c) steel; and (d) mercury.
5. 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.
6. A 0.250-kg block of a pure material is heated from 20.0ºC to 65.0ºC by the addition of 4.35 kJ of energy. Calculate its specific heat and identify the substance of which it is most likely composed.
7. 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?
8. (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 54.9ºC temperature increase? (b) Compare your answer to labeling information found on a package of peanuts and comment on whether the values are consistent.
9. Following vigorous exercise, the body temperature of an 80.0-kg person is 40.0ºC. At what rate in watts must the person transfer thermal energy to reduce the the body temperature to 37.0ºC in 30.0 min, assuming the body continues to produce energy at the rate of 150 W? 1 watt = 1 joule/second or 1 W = 1 J/s.
10. 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 watt = 1 joule/second or 1 W = 1 J/s and 1 MW = 1 megawatt). (a) Calculate the rate of temperature increase in degrees Celsius per second (ºC/s) if the mass of the reactor core is 1.60×10^5 kg and it has an average specific heat of 0.3349 kJ/kg · ºC. (b) How long would it take to obtain a temperature increase of 2000ºC, 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 5×10^5-kg steel containment vessel would also begin to heat up.)

Figure 3. Radioactive spent-fuel pool at a nuclear power plant. Spent fuel stays hot for a long time. (credit: U.S. Department of Energy)

Glossary

specific heat: the amount of heat necessary to change the temperature of 1.00 kg of a substance by 1.00 ºC

Selected Solutions to Problems & Exercises

1. 5.02×10^8 J
3. 3.07×10^3 J
5. 0.171ºC

7. 10.8

9. 617 W

Phase Change and Latent Heat

Lumen Learning

Learning Objectives

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

- Examine heat transfer.
- Calculate final temperature from heat transfer.

So far we have discussed temperature change due to heat transfer. No temperature change occurs from heat transfer if ice melts and becomes liquid water (i.e., during a phase change). For example, consider water dripping from icicles melting on a roof warmed by the Sun. Conversely, water freezes in an ice tray cooled by lower-temperature surroundings.

Energy is required to melt a solid because the cohesive bonds between the molecules in the solid must be broken apart such that, in the liquid, the molecules can move around at comparable kinetic energies; thus, there is no rise in temperature. Similarly, energy is needed to vaporize a liquid, because molecules in a liquid interact with each other via attractive forces. There is no temperature change until a phase change is complete. The temperature of a cup of soda initially at 0ºC stays at 0ºC until all the ice has melted. Conversely, energy is released during freezing and condensation, usually in the form of thermal energy. Work is done by cohesive forces when molecules are brought together. The corresponding energy must be given off (dissipated) to allow them to stay together Figure 2.

Figure 1. Heat from the air transfers to the ice causing it to melt. (credit: Mike Brand)

The energy involved in a phase change depends on two major factors: the number and strength of bonds or force pairs. The number of bonds is proportional to the number of molecules and thus to the mass of the sample. The strength of forces depends on the type of molecules. The heat Q required to change the phase of a sample of mass m is given by

$$Q = mL_{\text{f}} \text{ (melting/freezing,}$$

$$Q = mL_{\text{v}} \text{ (vaporization/condensation),}$$

where the latent heat of fusion, L_{f}, and latent heat of vaporization, L_{v}, are material constants that are determined experimentally. See (Table 1).

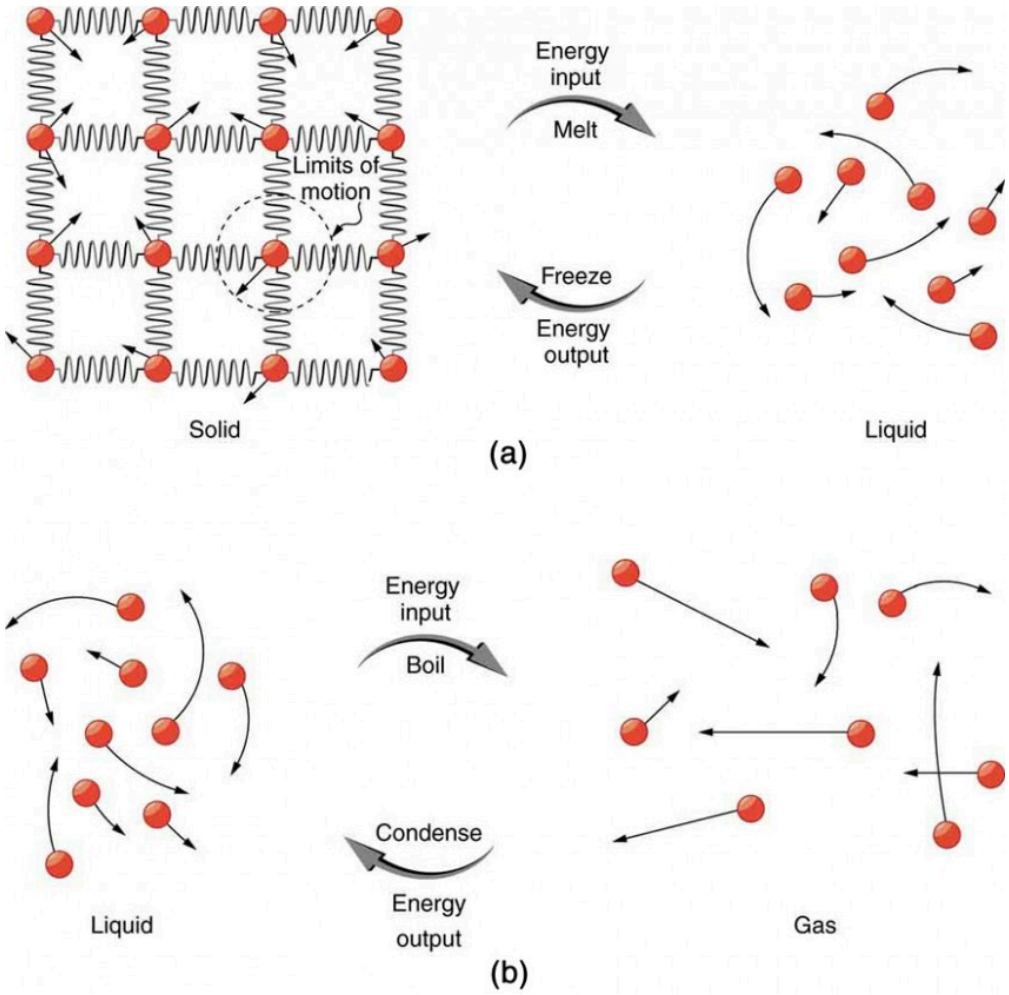


Figure 2. (a) Energy is required to partially overcome the attractive forces between molecules in a solid to form a liquid. That same energy must be removed for freezing to take place. (b) Molecules are separated by large distances when going from liquid to vapor, requiring significant energy to overcome molecular attraction. The same energy must be removed for condensation to take place. There is no temperature change until a phase change is complete.

Latent heat is measured in units of J/kg. Both L_f and L_v depend on the substance, particularly on the strength of its molecular forces as noted earlier. L_f and L_v are collectively called *latent heat coefficients*. They 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. Table 1 lists representative values of L_f and L_v, together with melting and boiling points.

The table shows that significant amounts of energy are involved in phase changes. Let us look, for example, at how much energy is needed to melt a kilogram of ice at 0ºC to produce a kilogram of water at 0°C. Using the equation for a change in temperature and the value for water from Table 1, we find that $Q = mL_f$ = (1.0 kg)(334 kJ/kg) = 334 kJ is the energy to melt a kilogram of ice. This is a lot of energy as it represents the same amount of energy needed to raise the temperature of 1 kg of liquid water from 0ºC to 79.8ºC. Even more energy is required to vaporize water; it would take 2256 kJ to change 1 kg of liquid water at the normal boiling point (100ºC at atmospheric pressure) to steam (water vapor). This example

shows that the energy for a phase change is enormous compared to energy associated with temperature changes without a phase change.

Table 1. Heats of Fusion and Vaporization[1]

		L_f			L_v	
Substance	**Melting point (ºC)**	**kJ/kg**	**kcal/kg**	**Boiling point (°C)**	**kJ/kg**	**kcal/kg**
Helium	−269.7	5.23	1.25	−268.9	20.9	4.99
Hydrogen	−259.3	58.6	14.0	−252.9	452	108
Nitrogen	−210.0	25.5	6.09	−195.8	201	48.0
Oxygen	−218.8	13.8	3.30	−183.0	213	50.9
Ethanol	−114	104	24.9	78.3	854	204
Ammonia	−75		108	−33.4	1370	327
Mercury	−38.9	11.8	2.82	357	272	65.0
Water	0.00	334	79.8	100.0	2256[2]	539[3]
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
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

Phase changes can have a tremendous stabilizing effect even on temperatures that are not near the melting and boiling points, because evaporation and condensation (conversion of a gas into a liquid state) occur even at temperatures below the boiling point. Take, for example, the fact that air temperatures in humid climates rarely go above 35.0ºC, which is because most heat transfer goes into evaporating water into the air. Similarly, temperatures in humid weather rarely fall below the dew point because enormous heat is released when water vapor condenses.

We examine the effects of phase change more precisely by considering adding heat into a sample of ice at −20ºC (Figure 3). The temperature of the ice rises linearly, absorbing heat at a constant rate of 0.50 cal/g·ºC until it reaches 0ºC. Once at this temperature, the ice begins to melt until all the ice has melted,

1. Values quoted at the normal melting and boiling temperatures at standard atmospheric pressure (1 atm).
2. At 37.0ºC (body temperature), the heat of vaporization Lv for water is 2430 kJ/kg or 580 kcal/kg
3. At 37.0ºC (body temperature), the heat of vaporization Lv for water is 2430 kJ/kg or 580 kcal/kg

absorbing 79.8 cal/g of heat. The temperature remains constant at 0ºC 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 1.00 cal/g·ºC. At 100ºC, the water begins to boil and the temperature again remains constant while the water absorbs 539 cal/g of heat during this phase change. When all the liquid has become steam vapor, the temperature rises again, absorbing heat at a rate of 0.482 cal/g·ºC.

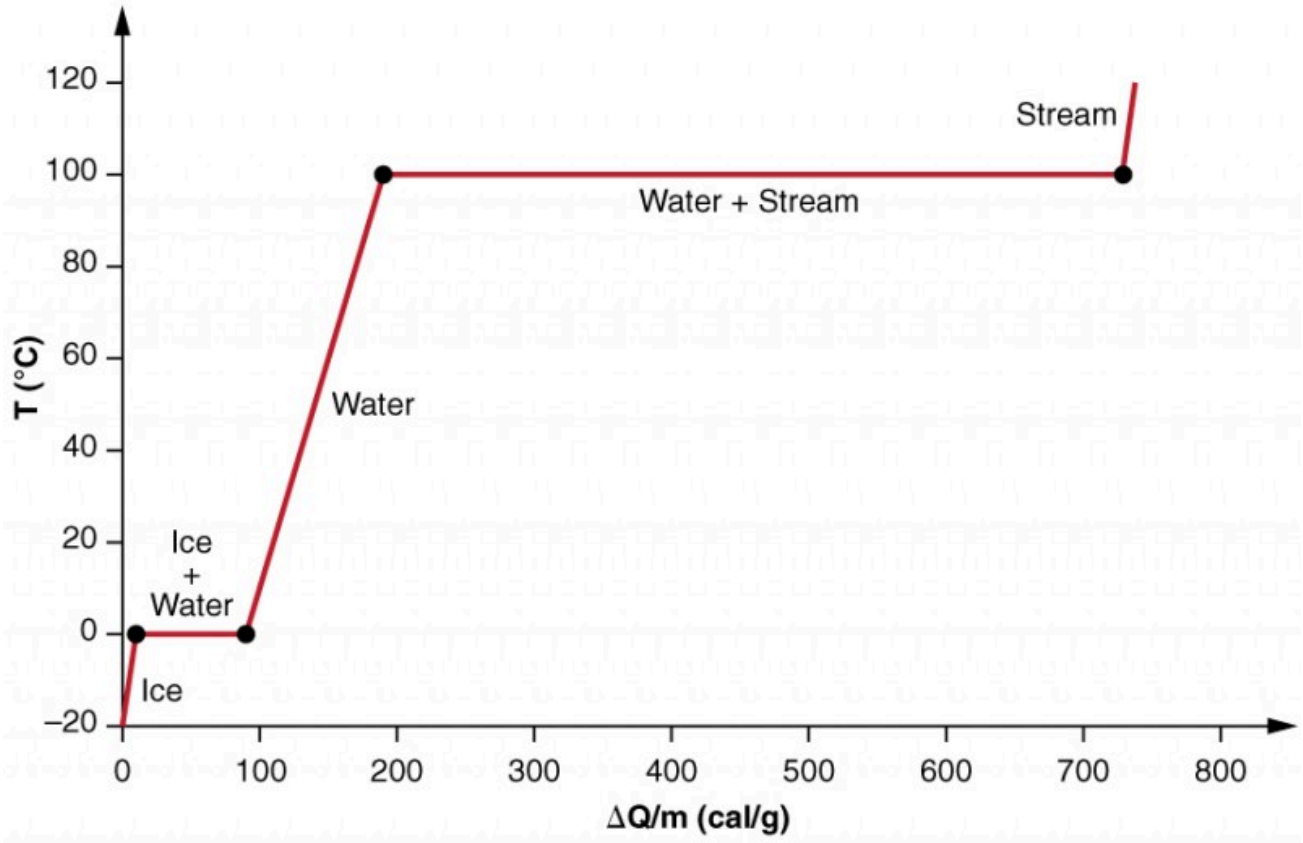


Figure 3. A graph of temperature versus energy added. The system is constructed so that no vapor evaporates while ice warms to become liquid water, and so that, when vaporization occurs, the vapor remains in of the system. The long stretches of constant temperature values at 0ºC and 100ºC reflect the large latent heat of melting and vaporization, respectively.

Water can evaporate at temperatures below the boiling point. More energy is required than at the boiling point, because the kinetic energy of water molecules at temperatures below 100ºC is less than that at 100ºC, hence less energy is available from random thermal motions. Take, for example, the fact that, at body temperature, perspiration from the skin requires a heat input of 2428 kJ/kg, which is about 10 percent higher than the latent heat of vaporization at 100ºC. This heat comes from the skin, and thus provides an effective cooling mechanism in hot weather. High humidity inhibits evaporation, so that body temperature might rise, leaving unevaporated sweat on your brow.

Example 1. Calculate Final Temperature from Phase Change: Cooling Soda with Ice Cubes

Three ice cubes are used to chill a soda at 20ºC with mass m_{soda} = 0.25 kg. The ice is at 0ºC 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. Assume the soda has the same heat capacity as water. Find the final temperature when all ice has melted.

Strategy

The ice cubes are at the melting temperature of 0ºC. Heat is transferred from the soda to the ice for melting. Melting of ice occurs in two steps: first the phase change occurs and solid (ice) transforms into liquid water at the melting temperature, then the temperature of this water rises. Melting yields water at 0ºC, so more heat is transferred from the soda to this water until the water plus soda system reaches thermal equilibrium, $Q_{ice} = -Q_{soda}$.

The heat transferred to the ice is

$$Q_{ice} = m_{ice} L_f + m_{ice} c_W (T_f - 0^ºC).$$

The heat given off by the soda is $Q_{soda} = m_{soda} c_W (T_f - 20^ºC)$. Since no heat is lost, $Q_{ice} = -Q_{soda}$, so that

$$m_{ice} L_f + m_{ice} c_W (T_f - 0^ºC) = -m_{soda} c_W (T_f - 20^ºC).$$

Bring all terms involving T_f on the left-hand-side and all other terms on the right-hand-side. Solve for the unknown quantity T_f:

$$T_f = \frac{m_{soda} c_W (20^\circ C) - m_{ice} L_f}{(m_{soda} + m_{ice}) c_W}$$

Solution

1. Identify the known quantities. The mass of ice is $m_{ice} = 3 \times 6.0$ g = 0.018 kg and the mass of soda is $m_{soda} = 0.25$ kg.
2. Calculate the terms in the numerator: $m_{soda} c_W (20^ºC) = (0.25 \text{ kg})(4186 \text{ J/kg} \cdot ^ºC)(20^ºC) = 20{,}930$ J and $m_{ice} L_f = (0.018 \text{ kg})(334{,}000 \text{ J/kg}) = 6012$ J.
3. Calculate the denominator: $(m_{soda} + m_{ice}) c_W = (0.25 \text{ kg} + 0.018 \text{ kg})(4186 \text{ K/(kg}\cdot^ºC) = 1122 \text{ J/}^ºC$.
4. Calculate the final temperature:

$$T_f = \frac{20{,}930 \text{ J} - 6012 \text{ J}}{1122 \text{ J}/^\circ C} = 13^\circ C$$

Discussion

This example illustrates the enormous energies involved during a phase change. The mass of ice is about 7 percent the mass of water but leads to a noticeable change in the temperature of soda. Although we assumed that the ice was at the freezing temperature, this is incorrect: the typical temperature is −6ºC. However, this correction gives a final temperature that is essentially identical to the result we found. Can you explain why?

We have seen that vaporization requires heat transfer to a liquid from the surroundings, so that energy is released by the surroundings. Condensation is the reverse process, increasing the temperature of the surroundings. This increase may seem surprising, since we associate condensation with cold objects—the glass in the figure, for example. However, energy must be removed from the condensing molecules to make a vapor condense. The energy is exactly the same as that required to make the phase change in the other direction, from liquid to vapor, and so it can be calculated from $Q = mL_v$.

Figure 4. Condensation on a glass of iced tea. (credit: Jenny Downing)

Condensation forms in Figure 4 because the temperature of the nearby air is reduced to below the dew point. The air cannot hold as much water as it did at room temperature, and so water condenses. Energy is released when the water condenses, speeding the melting of the ice in the glass.

Real-World Application

Energy is also released when a liquid freezes. This phenomenon is used by fruit growers in Florida to protect oranges when the temperature is close to the freezing point (0ºC). Growers spray water on the plants in orchards so that the water freezes and heat is released to the growing oranges on the trees. This prevents the temperature inside the orange from dropping below freezing, which would damage the fruit.

Figure 14.11. The ice on these trees released large amounts of energy when it froze, helping to prevent the temperature of the trees from dropping below 0ºC. Water is intentionally sprayed on orchards to help prevent hard frosts. (credit: Hermann Hammer)

Sublimation is the transition from solid to vapor phase. You may have noticed that snow can disappear

into thin air without a trace of liquid water, or the disappearance of ice cubes in a freezer. The reverse is also true: Frost can form on very cold windows without going through the liquid stage. A popular effect is the making of “smoke” from dry ice, which is solid carbon dioxide. Sublimation occurs because the equilibrium vapor pressure of solids is not zero. Certain air fresheners use the sublimation of a solid to inject a perfume into the room. Moth balls are a slightly toxic example of a phenol (an organic compound) that sublimates, while 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)

Figure 5. Direct transitions between solid and vapor are common, sometimes useful, and even beautiful. (a) Dry ice sublimates directly to carbon dioxide gas. The visible vapor is made of water droplets. (credit: Windell Oskay) (b) Frost forms patterns on a very cold window, an example of a solid formed directly from a vapor. (credit: Liz West)

All phase transitions involve heat. In the case of direct solid-vapor transitions, the energy required is given by the equation $Q = mL_s$, where L_s is the *heat of sublimation*, which is the energy required to change 1.00 kg of a substance from the solid phase to the vapor phase. L_s is analogous to L_f and L_v, and its value depends on the substance. Sublimation requires energy input, so that dry ice is an effective coolant, whereas the reverse process (i.e., frosting) releases energy. The amount of energy required for sublimation is of the same order of magnitude as that for other phase transitions.

The material presented in this section and the preceding section allows us to 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 and then to apply the appropriate equation. Keep in mind that heat transfer and work can cause both temperature and phase changes.

Problem-Solving Strategies for 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 the presence or absence of a phase change is not obvious, 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 and 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).*
5. *Solve the appropriate equation for the quantity to be determined (the unknown).* If there is a temperature change, the transferred heat depends on the specific heat (see Table 1 in Temperature Change and Heat Capacity) whereas, for a phase change, the transferred heat depends on the latent heat. See Table 1.
6. *Substitute the knowns along with their units into the appropriate equation and obtain numerical solutions complete with units.* You will need to do this in steps if there is more than one stage to the process (such as a temperature change followed by a phase change).
7. *Check the answer to see if it is reasonable: Does it make sense?* As an example, be certain that the temperature change does not also cause a phase change that you have not taken into account.

Check Your Understanding

Why does snow remain on mountain slopes even when daytime temperatures are higher than the freezing temperature?

Solution

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 accumulated from the air, even if the air is above 0ºC. The warmer the air is, the faster this heat exchange occurs and the faster the snow melts.

Section Summary

- Most substances can exist either in solid, liquid, and gas forms, which are referred to as "phases."
- Phase changes occur at fixed temperatures for a given substance at a given pressure, and these temperatures are called boiling and freezing (or melting) points.
- During phase changes, heat absorbed or released is given by: $Q = mL$ where L is the latent heat coefficient.

Conceptual Questions

1. Heat transfer can cause temperature and phase changes. What else can cause these changes?
2. How does the latent heat of fusion of water help slow the decrease of air temperatures, perhaps preventing temperatures from falling significantly below ºC, in the vicinity of large bodies of water?
3. What is the temperature of ice right after it is formed by freezing water?
4. If you place ºC ice into ºC water in an insulated container, what will happen? Will some ice melt, will more water freeze, or will neither take place?
5. 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?
6. In very humid climates where there are numerous bodies of water, such as in Florida, it is unusual for temperatures to rise above about 35ºC (95ºF). In deserts, however, temperatures can rise far above this. Explain how the evaporation of water helps limit high temperatures in humid climates.
7. In winters, it is often warmer in San Francisco than in nearby Sacramento, 150 km inland. In summers, it is nearly always hotter in Sacramento. Explain how the bodies of water surrounding San Francisco moderate its extreme temperatures.
8. Putting a lid on a boiling pot greatly reduces the heat transfer necessary to keep it boiling. Explain why.
9. 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.

10. When still air cools by radiating at night, it is unusual for temperatures to fall below the dew point. Explain why.
11. 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 sublimate, with some crystals lingering for awhile and then producing a liquid. Identify the blue liquid and the two solids in the cold balloon. Justify your identifications using data from Table 1.

Problems & Exercises

1. How much heat transfer (in kilocalories) is required to thaw a 0.450-kg package of frozen vegetables originally at 0ºC if their heat of fusion is the same as that of water?
2. A bag containing 0ºC ice is much more effective in absorbing energy than one containing the same amount of 0ºC water. (a) How much heat transfer is necessary to raise the temperature of 0.800 kg of water from 0ºC to 30.0ºC? (b) How much heat transfer is required to first melt 0.800 kg of 0ºC ice and then raise its temperature? (c) Explain how your answer supports the contention that the ice is more effective.
3. (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 30.0ºC 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 1 watt = 1 joule/second (1 W = 1 J/s)?
4. The formation of condensation on a glass of ice water causes the ice to melt faster than it would otherwise. If 8.00 g of condensation forms 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.
5. 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 0ºC and completely melts to 0ºC water in exactly one day 1 watt = 1 joule/second (1 W = 1 J/s)?
6. On a certain dry sunny day, a swimming pool's temperature would rise by 1.50ºC if not for evaporation. What fraction of the water must evaporate to carry away precisely enough energy to keep the temperature constant?
7. (a) How much heat transfer is necessary to raise the temperature of a 0.200-kg piece of ice from −20.0ºC to 130ºC, 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.
8. In 1986, a gargantuan iceberg broke away from the Ross Ice Shelf in Antarctica. It was approximately a rectangle 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?
9. How many grams of coffee must evaporate from 350 g of coffee in a 100-g glass cup to cool the

coffee from 95.0ºC to 45.0ºC? You may assume the coffee has the same thermal properties as water and that the average heat of vaporization is 2340 kJ/kg (560 cal/g). (You may neglect the change in mass of the coffee as it cools, which will give you an answer that is slightly larger than correct.)

10. (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 has its temperature raised from 20.0ºC to 100ºC, it boils, and the resulting steam is raised to 300ºC. (b) Discuss additional complications caused by the fact that crude oil has a smaller density than water.
11. 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.
12. To help prevent frost damage, 4.00 kg of 0ºC water 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 kJ/kg · ºC, and assume that no phase change occurs.
13. A 0.250-kg aluminum bowl holding 0.800 kg of soup at 25.0ºC is placed in a freezer. What is the final temperature if 377 kJ of energy is transferred from the bowl and soup, assuming the soup's thermal properties are the same as that of water?
14. A 0.0500-kg ice cube at −30.0ºC is placed in 0.400 kg of 35.0ºC water in a very well-insulated container. What is the final temperature?
15. If you pour 0.0100 kg of 20.0ºC water onto a 1.20-kg block of ice (which is initially at −15.0ºC), what is the final temperature? You may assume that the water cools so rapidly that effects of the surroundings are negligible.
16. Indigenous people sometimes cook in watertight baskets by placing hot rocks into water to bring it to a boil. What mass of 500ºC rock must be placed in 4.00 kg of 15.0ºC water to bring its temperature to 100ºC, if 0.0250 kg of water escapes as vapor from the initial sizzle? You may neglect the effects of the surroundings and take the average specific heat of the rocks to be that of granite.
17. What would be the final temperature of the pan and water in Calculating the Final Temperature When Heat Is Transferred Between Two Bodies: Pouring Cold Water in a Hot Pan if 0.260 kg of water was 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?
18. 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 kg/m^3. (a) Calculate the heat transfer necessary to evaporate this amount of liquid nitrogen and raise its temperature to 3.00ºC. (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 0ºC ice with that from evaporating the liquid nitrogen.
19. Some gun fanciers make their own bullets, which involves melting and casting the lead slugs. How much heat transfer is needed to raise the temperature and melt 0.500 kg of lead, starting from 25.0ºC?

Glossary

heat of sublimation: the energy required to change a substance from the solid phase to the vapor phase

latent heat coefficient: a physical constant equal to the amount of heat transferred for every 1 kg of a substance during the change in phase of the substance

sublimation: the transition from the solid phase to the vapor phase

Selected Solutions to Problems & Exercises

1. 35.9 kcal

3. (a) 591 kcal; (b) 4.94×10^3 s

5. 13.5 W

7. (a) 148 kcal; (b) 0.418 s, 3.34 s, 4.19 s, 22.6 s, 0.456 s

9. 33.0 g

10. (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 efficient in absorbing the heat generated by the oil.

12. (a) 319 kcal; (b) 2.00ºC

14. 20.6ºC

16. 4.38 kg

18. (a) 1.57×10^4 kcal; (b) 18.3 kW · h; (c) 1.29×10^4 kcal

Heat Transfer Methods

Lumen Learning

Learning Objective

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

- Discuss the different methods of heat transfer.

Equally as interesting as the effects of heat transfer on a system are the methods by which this occurs. Whenever there is a temperature difference, heat transfer occurs. Heat transfer may occur rapidly, such as through a cooking pan, or slowly, such as through the walls of a picnic ice chest. We can control rates of heat transfer by choosing materials (such as thick wool clothing for the winter), controlling air movement (such as the use of weather stripping around doors), or by choice of color (such as a white roof to reflect summer sunlight). So many processes involve heat transfer, so that it is hard to imagine a situation where no heat transfer occurs. Yet every process involving 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 there is thermal motion of the atoms and molecules at any temperature above absolute zero.) Heat transferred between the electric burner of a stove and the bottom of a 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 the Earth by the Sun. A less obvious example is thermal radiation from the human body.

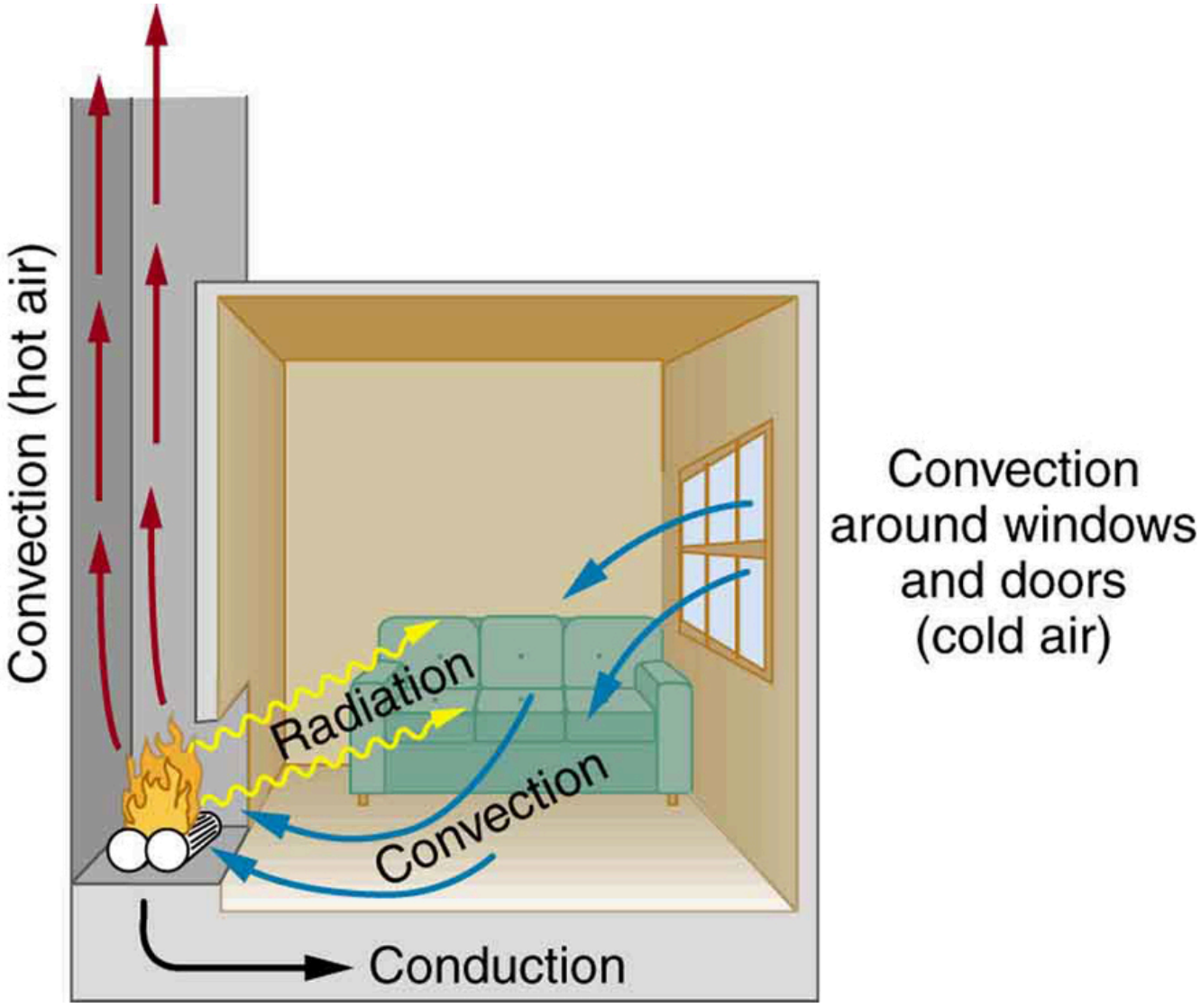


Figure 1. In a fireplace, heat transfer occurs by all three methods: conduction, convection, and radiation. Radiation is responsible for most of the heat transferred into the room. Heat transfer also occurs through conduction into the room, but at a much slower rate. Heat transfer by convection also occurs through cold air entering the room around windows and hot air leaving the room by rising up the chimney.

We examine these methods in some detail in the three following modules. Each method has unique and interesting characteristics, but all three do have one thing in common: they transfer heat solely because of a temperature difference Figure 1.

Check Your Understanding

Name an example from daily life (different from the text) for each mechanism of heat transfer.

Solution

- 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: Reheating a cold cup of coffee in a microwave oven.

Section Summary

- Heat is transferred by three different methods: conduction, convection, and radiation.

Conceptual Questions

1. What are the main methods of heat transfer from the hot core of Earth to its surface? From Earth’s surface to outer space?
2. 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 this have on a person in a 40.0ºC hot tub?
3. Figure 2 shows 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.

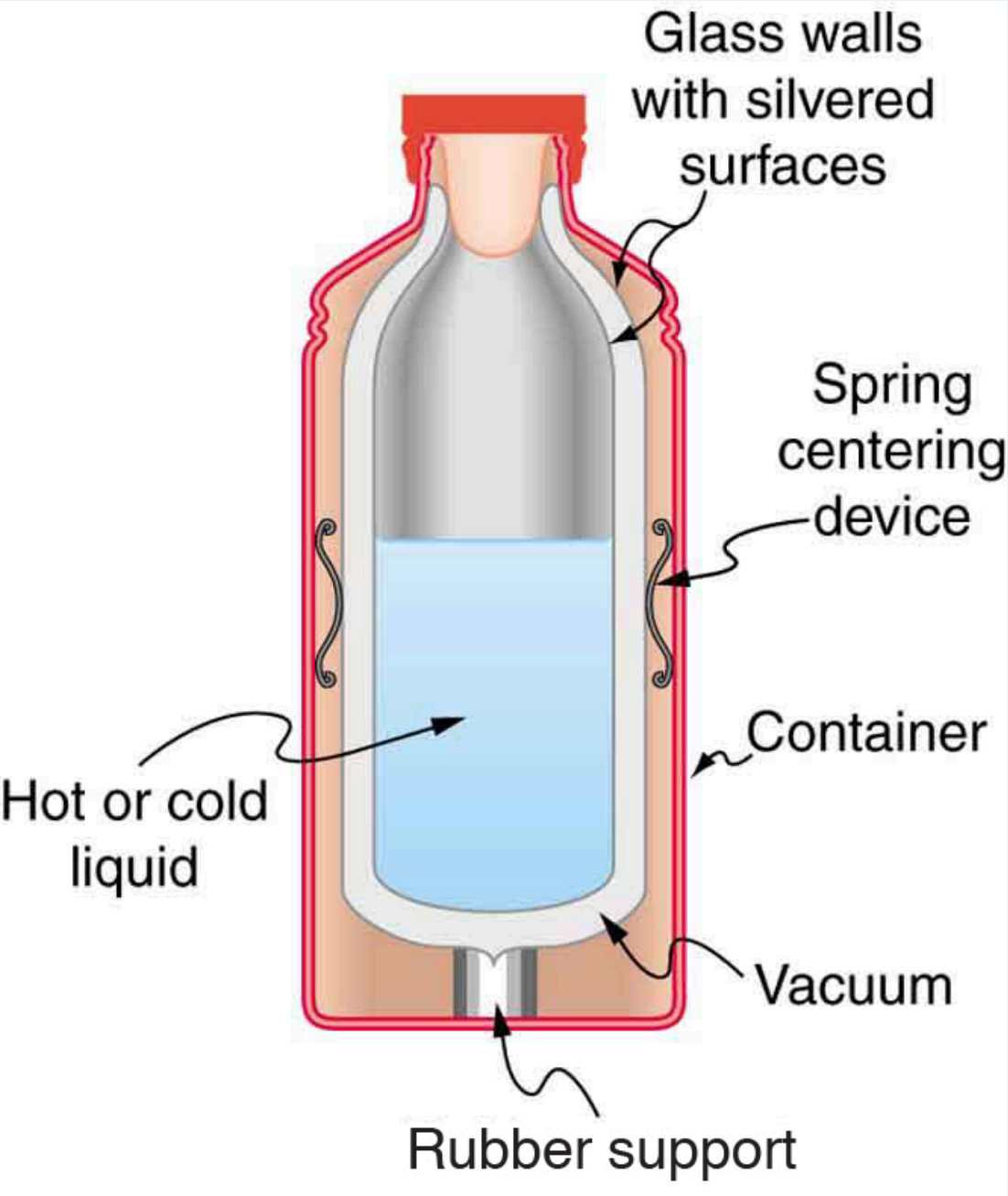


Figure 2. The construction of a thermos bottle is designed to inhibit all methods of heat transfer.

4. The construction of a thermos bottle is designed to inhibit all methods of heat transfer.
5. The figure shows a cutaway drawing of a thermos bottle, with various parts labeled.

Glossary

conduction: heat transfer through stationary matter by physical contact

convection: heat transfer by the macroscopic movement of fluid

radiation: heat transfer which occurs when microwaves, infrared radiation, visible light, or other electromagnetic radiation is emitted or absorbed

Video: Heat Transfer

Lumen Learning

Watch the following Physics Concept Trailer to see how conduction, convection, and radiation are all important processes in preparing the perfect steak.

A YouTube element has been excluded from this version of the text. You can view it online here: https://pressbooks.nscc.ca/heatlightsound/?p=91

Conduction

Lumen Learning

Learning Objectives

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

- Calculate thermal conductivity.
- Observe conduction of heat in collisions.
- Study thermal conductivities of common substances.

Your feet feel cold as you walk barefoot across the living room carpet in your cold house and then step onto the kitchen tile floor. This result is intriguing, since the carpet and tile floor are both at the same temperature. The different sensation you feel is explained by the different rates of heat transfer: the heat loss during the same time interval is greater for skin in contact with the tiles than with the carpet, so the temperature drop is greater on the tiles.

Figure 1. Insulation is used to limit the conduction of heat from the inside to the outside (in winters) and from the outside to the inside (in summers). (credit: Giles Douglas)

Some materials conduct thermal energy faster than others. In general, good conductors of electricity (metals like copper, aluminum, gold, and silver) are also good heat conductors, whereas insulators of electricity (wood, plastic, and rubber) are poor heat conductors. Figure 2 shows molecules in two bodies at different temperatures. The (average) kinetic energy of a molecule in the hot body is higher than in the colder body. If two molecules collide, an energy transfer from the hot to the cold molecule occurs. The cumulative effect from all collisions results in a net flux of heat from the hot body to the colder body. The heat flux thus depends on the temperature difference $\Delta T = T_{\text{hot}} - T_{\text{cold}}$. Therefore, you will get a more severe burn from boiling water than from hot tap water. Conversely, if the temperatures are the same, the net heat transfer rate falls to zero, and equilibrium is achieved. Owing to the fact that the number of collisions increases with increasing area, heat conduction depends on the cross-sectional area. If you touch a cold wall with your palm, your hand cools faster than if you just touch it with your fingertip.

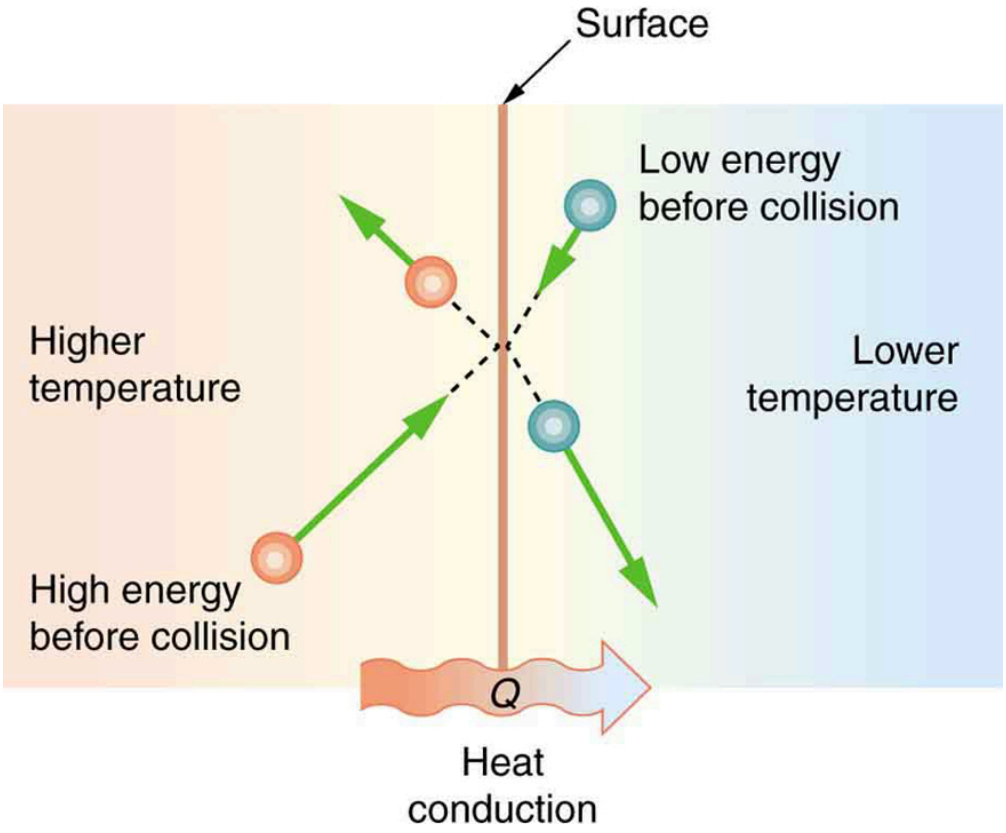


Figure 2. The molecules in two bodies at different temperatures have different average kinetic energies. Collisions occurring at the contact surface tend to transfer energy from high-temperature regions to low-temperature regions. In this illustration, a molecule in the lower temperature region (right side) has low energy before collision, but its energy increases after colliding with 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 the contact surface.

A third factor in the mechanism of conduction is the thickness of the material through which heat transfers. The figure below shows a slab of material with different temperatures on either side. Suppose that T_2 is greater than T_1, so that heat is transferred from left to right. Heat transfer from the left side to the right side is accomplished by a series of molecular collisions. The thicker the material, the more time it takes to transfer the same amount of heat. This model explains why thick clothing is warmer than thin clothing in winters, and why Arctic mammals protect themselves with thick blubber.

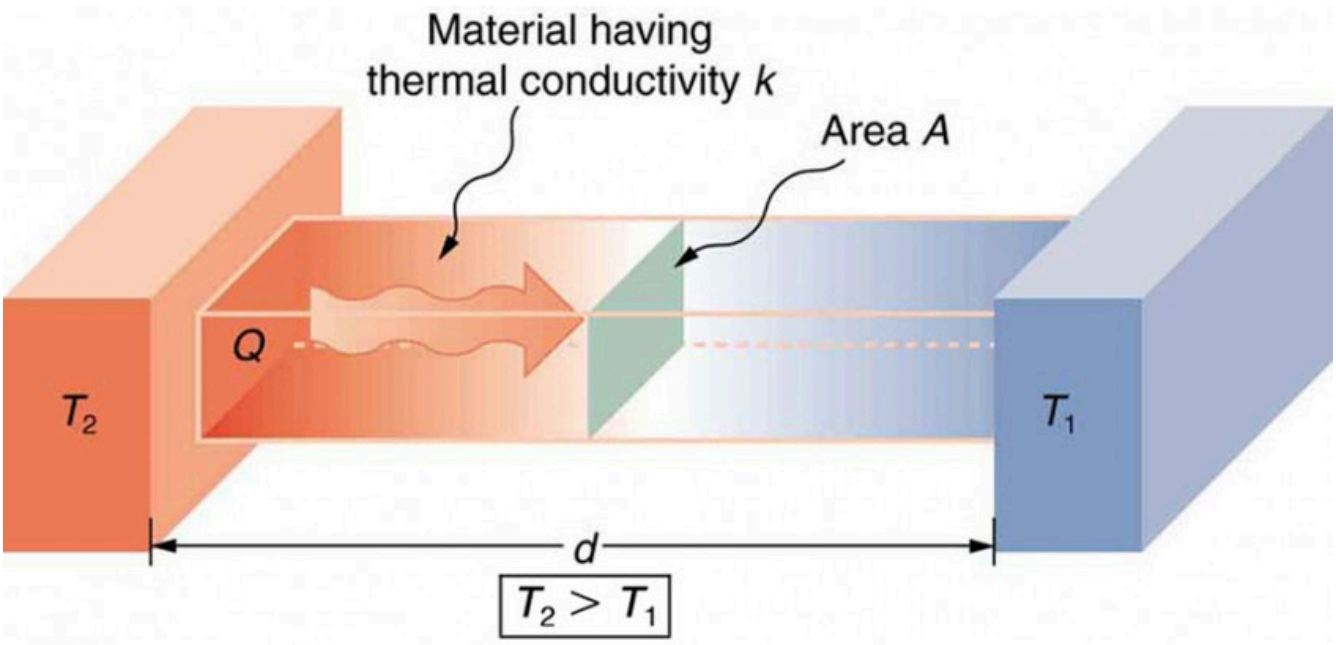


Figure 3. Heat conduction occurs through any material, represented here by a rectangular bar, whether window glass or walrus blubber. The temperature of the material is T_2 on the left and T_1 on the right, where T_2 is greater than T_1. The rate of heat transfer by conduction is directly proportional to the surface area A, the temperature difference $T_2 - T_1$, and the substance's conductivity k. The rate of heat transfer is inversely proportional to the thickness d.

Lastly, the heat transfer rate depends on the material properties described by the coefficient of thermal conductivity. All four factors are included in a simple equation that was deduced from and is confirmed by experiments. The *rate of conductive heat transfer* through a slab of material, such as the one in Figure 3, is given by

$$\frac{Q}{t} = \frac{kA\,(T_2 - T_1)}{d}$$

,

where

$$\frac{Q}{t}$$

is the rate of heat transfer in watts or kilocalories per second, *k* is the *thermal conductivity* of the material, *A* and *d* are its surface area and thickness, as shown in Figure 3, and $(T_2 - T_1)$ is the temperature difference across the slab. Table 1 gives representative values of thermal conductivity.

Example 1. Calculating Heat Transfer Through Conduction: Conduction Rate Through an Ice Box

A Styrofoam ice box has a total area of 0.950 m^2 and walls with an average thickness of 2.50 cm. The box contains ice, water, and canned beverages at 0ºC. The inside of the box is kept cold by melting ice. How much ice melts in one day if the ice box is kept in the trunk of a car at 35.0ºC?

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

Identify the knowns.

$$\begin{aligned} A &= 0.950\ \text{m}^2; \\ d &= 2.50\ \text{cm} = 0.0250\ \text{m}; \\ T_1 &= 0^\circ\text{C}; \\ T_2 &= 35.0^\circ\text{C}; \\ t &= 1\ \text{day} = 24\text{hours} = 86,400\ \text{s}. \end{aligned}$$

Identify the unknowns. We need to solve for the mass of the ice, *m*. We will also need to solve for the net heat transferred to melt the ice, *Q*. Determine which equations to use. The rate of heat transfer by conduction is given by

$$\frac{Q}{t} = \frac{kA\left(T_2 - T_1\right)}{d}$$

The heat is used to melt the ice: $Q = mL_\text{f}$.

Insert the known values:

$$\frac{Q}{t} = \frac{(0.010\ \text{J/s} \cdot \text{m} \cdot^\circ \text{C})\left(0.950\ \text{m}^2\right)(35.0^\circ\text{C} - 0^\circ\text{C})}{0.0250\ \text{m}} = 13.3\ \text{J/s}$$

Multiply the rate of heat transfer by the time (1 day = 86,400 s): $Q =$

$$\left(\frac{Q}{t}\right)t$$

$= (13.3\ \text{J/s})(86{,}400\ \text{s}) = 1.15 \times 10^6\ \text{J}$.

Set this equal to the heat transferred to melt the ice: $Q = mL_\text{f}$. Solve for the mass *m*:

$$m = \frac{Q}{L_\text{f}} = \frac{1.15 \times 10^6\ \text{J}}{334 \times 10^3\ \text{J/kg}} = 3.44\ \text{kg}$$

Discussion

The result of 3.44 kg, or about 7.6 lbs, 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.

Inspecting the conductivities in Table 1 shows that Styrofoam is a very poor conductor and thus a good insulator. Other good insulators include fiberglass, wool, and goose-down feathers. Like Styrofoam, these all incorporate many small pockets of air, taking advantage of air's poor thermal conductivity.

Table 1. Thermal Conductivities of Common Substances[1]

Substance	Thermal conductivity k (J/s·m·ºC)
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
Styrofoam	0.010

1. At temperatures near 0ºC.

A combination of material and thickness is often manipulated to develop good insulators—the smaller the conductivity k and the larger the thickness d, the better. The ratio of

$$\frac{d}{k}$$

will thus be large for a good insulator. The ratio

$$\frac{d}{k}$$

is called the *R factor*. The rate of conductive heat transfer is inversely proportional to *R*. The larger the value of *R*, the better the insulation. *R* factors are most commonly quoted for household insulation, refrigerators, and the like—unfortunately, it is still in non-metric units of $ft^2 \cdot °F \cdot h/Btu$, 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°F). A couple of representative values are an *R* factor of 11 for 3.5-in-thick fiberglass batts (pieces) of insulation and an *R* factor of 19 for 6.5-in-thick fiberglass batts. Walls are usually insulated with 3.5-in batts, while ceilings are usually insulated with 6.5-in batts. In cold climates, thicker batts may be used in ceilings and walls.

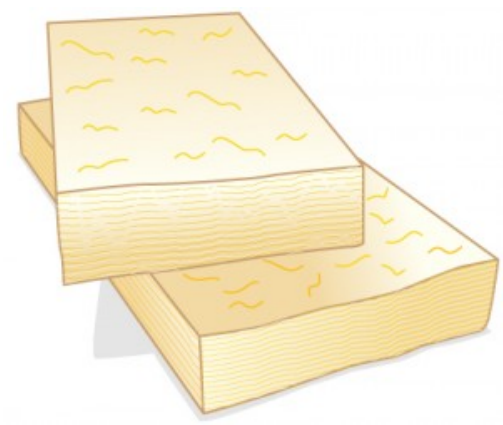


Figure 4. The fiberglass batt is used for insulation of walls and ceilings to prevent heat transfer between the inside of the building and the outside environment.

Note that in Table 1, the best thermal conductors—silver, copper, gold, and aluminum—are also the best electrical conductors, again related to the density of free electrons in them. Cooking utensils are typically made from good conductors.

Example 2. Calculating the Temperature Difference Maintained by a Heat Transfer: Conduction Through an Aluminum Pan

Water is boiling in an aluminum pan placed on an electrical element on a stovetop. The sauce pan has a bottom that is 0.800 cm thick and 14.0 cm in diameter. The boiling water is evaporating at the rate of 1.00 g/s. What is the temperature difference across (through) the bottom of the pan?

Strategy

Conduction through the aluminum is the primary method of heat transfer here, and so we use the equation for the rate of heat transfer and solve for the temperature difference.

$$T_2 - T_1 = \frac{Q}{t}\left(\frac{d}{kA}\right)$$

Solution

Identify the knowns and convert them to the SI units.The thickness of the pan, $d = 0.900$ cm $= 8.0 \times 10^{-3}$ m the area of the pan, $A = \pi(0.14/2)^2$ m^2 $= 1.54 \times 10^{-2}$ m^2, and the thermal conductivity, $k = 220$ J/s · m · °C.

Calculate the necessary heat of vaporization of 1 g of water: $Q = mL_v = (1.00 \times 10^{-3}$ kg$)(2256 \times 10^3$ J/kg$) = 2256$ J.

Calculate the rate of heat transfer given that 1 g of water melts in one second:

$$\frac{Q}{t}$$

= 2256 J/s or 2.26 kW.

Insert the knowns into the equation and solve for the temperature difference:

$$T_2 - T_1 = \frac{Q}{t}\left(\frac{d}{kA}\right) = (2256\ \text{J/s})\,\frac{8.00 \times 10^{-3}\ \text{m}}{(220\text{J/s} \cdot \text{m} \cdot^{\circ}\text{C})\,(1.54 \times 10^{-2}\text{m}^2)} = 5.33^{\circ}\text{C}$$

Discussion

The value for the heat transfer

$$\frac{Q}{t}$$

= 2.26 kW or 2256 J/s is typical for an electric stove. This value gives a remarkably small temperature difference between the stove and the pan. Consider that the stove burner is red hot while the inside of the pan is nearly 100ºC because of its contact with boiling water. This contact effectively cools the bottom of the pan in spite of its proximity to the very hot stove burner. Aluminum is such a good conductor that it only takes this small temperature difference to produce a heat transfer of 2.26 kW into the pan.

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 time distances. Take, for example, the temperature on the Earth, which would be unbearably cold during the night and extremely hot during the day if heat transport in the atmosphere was to be only through conduction. In another example, car engines would overheat unless there was a more efficient way to remove excess heat from the pistons.

Check Your Understanding

How does the rate of heat transfer by conduction change when all spatial dimensions are doubled?

Solution

Because area is the product of two spatial dimensions, it increases by a factor of four when each dimension is doubled ($A_{\text{final}} = (2d)^2 = 4d^2 = 4A_{\text{initial}}$). 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:

$$\left(\frac{Q}{t}\right)_{\text{final}} = \frac{kA_{\text{final}}(T_2 - T_1)}{d_{\text{final}}} = \frac{k(4A_{\text{initial}})(T_2 - T_1)}{2d_{\text{initial}}} = 2\frac{kA_{\text{initial}}(T_2 - T_1)}{d_{\text{initial}}} = 2\left(\frac{Q}{t}\right)_{\text{initial}}$$

Section Summary

- Heat conduction is the transfer of heat between two objects in direct contact with each other.

- The rate of heat transfer $\frac{Q}{t}$ (energy per unit time) is proportional to the temperature difference $T_2 - T_1$ and the contact area A and inversely proportional to the distance d between the objects: $\frac{Q}{t} = \frac{\mathrm{kA}(T_2 - T_1)}{d}$.

Conceptual Questions

1. Some electric stoves have a flat ceramic surface with heating elements hidden beneath. A pot placed over a heating element will be heated, while it is safe to touch the surface only a few centimeters away. 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?
2. Loose-fitting white clothing covering most of the body 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.

Figure 5. A jellabiya is worn by many men in Egypt. (credit: Zerida)

Problems & Exercises

1. (a) Calculate the rate of heat conduction through house walls that are 13.0 cm thick and that have an average thermal conductivity twice that of glass wool. Assume there are no windows or doors. The surface area of the walls is 120 m^2 and their inside surface is at 18.0ºC, while their outside surface is at 5.00ºC. (b) How many 1-kW room heaters would be needed to balance the heat transfer due to conduction?
2. 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 0.635 cm thick (1/4 in) if the temperatures of the inner and outer surfaces are 5.00ºC and −10.0ºC, respectively. This rapid rate will not be maintained—the inner surface will cool, and even result in frost formation.
3. Calculate the rate of heat conduction out of the human body, assuming that the core internal temperature is 37.0ºC, the skin temperature is 34.0ºC, the thickness of the tissues between averages 1.00 cm, and the surface area is 1.40 m^2.
4. 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 10.0ºC on their bottom sides. At what rate must heat transfer occur from each foot to keep the top of the ceramic and carpet at 33.0ºC?
5. A man consumes 3000 kcal of food in one day, converting most of it to maintain body temperature. If he loses half this energy by evaporating water (through breathing and sweating), how many kilograms of water evaporate?
6. (a) 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 700ºC, and the time in contact is 1.00 s. (b) What temperature increase is produced in the 25.0 cm^3 of tissue affected? (c) What effect do you think this will have on the tissue, keeping in mind that a callus is made of dead cells?
7. (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 32.0ºC, that the air temperature is −5.00ºC, 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?
8. A walrus transfers energy by conduction through its blubber at the rate of 150 W when immersed in −1.00ºC water. The walrus's internal core temperature is 37.0ºC, and it has a surface area of 2.00 m^2. What is the average thickness of its blubber, which has the conductivity of fatty tissues without blood?

Figure 6. Walrus on ice. (credit: Captain Budd Christman, NOAA Corps)

9. Compare the rate of heat conduction through a 13.0-cm-thick wall that has an area of 10.0 m^2 and a thermal conductivity twice that of glass wool with the rate of heat conduction through a window that is 0.750 cm thick and that has an area of 2.00 m^2, assuming the same temperature difference across each.
10. 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 m^2?
11. 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 2, what is the temperature difference across it? Ceramic has the same thermal conductivity as glass and brick.
12. One easy way to reduce heating (and cooling) costs is to add extra insulation in the attic of a house. Suppose the 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, then by what percentage would the heating cost of the house drop? Take the single story house to be of dimensions 10 m by 15 m by 3.0 m. Ignore air infiltration and heat loss through windows and doors.
13. (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 15.0ºC, while that on the outside is −10.0ºC. (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).
14. 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 two years.) Suppose you wish to install the extra insulation in question 12. 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 15.0ºC.

15. 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 change in temperature is 2.00ºC, and the skin area is 1.50 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.)

Glossary

R factor: the ratio of thickness to the conductivity of a material

rate of conductive heat transfer: rate of heat transfer from one material to another

thermal conductivity: the property of a material's ability to conduct heat

Selected Solutions to Problems & Exercises

1. (a) 1.01×10^3 W; (b) One

3. 84.0 W

5. 2.59 kg

7. (a) 39.7 W; (b) 820 kcal

9. 35 to 1, window to wall

11. 1.05×10^3 K

13. (a) 83 W; (b) 24 times that of a double pane window.

15. 20.0 W, 17.2% of 2400 kcal per day

Convection

Lumen Learning

Learning Objectives

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

- Discuss the method of heat transfer by convection.

Convection is driven by large-scale flow of matter. In the case of Earth, the atmospheric circulation is caused by the flow of hot air from the tropics to the poles, and the flow of cold air from the poles toward the tropics. (Note that Earth's rotation causes the observed easterly flow of air in the northern hemisphere). Car engines are kept cool by the flow of water in the cooling system, with the water pump maintaining a flow of cool water to the pistons. The circulatory system is used the body: when the body overheats, the blood vessels in the skin expand (dilate), which increases the blood flow to the skin where it can be cooled by sweating. These vessels become smaller when it is cold outside and larger when it is hot (so more fluid flows, and more energy is transferred).

The body also loses a significant fraction of its heat through the breathing process.

While convection is usually more complicated than conduction, we can describe convection and do some straightforward, realistic calculations of its effects. Natural convection is driven by buoyant forces: hot air rises because density decreases as temperature increases. The house in Figure 1 is kept warm in this manner, as is the pot of water on the stove in Figure 2. Ocean currents and large-scale atmospheric circulation transfer energy from one part of the globe to another. Both are examples of natural convection.

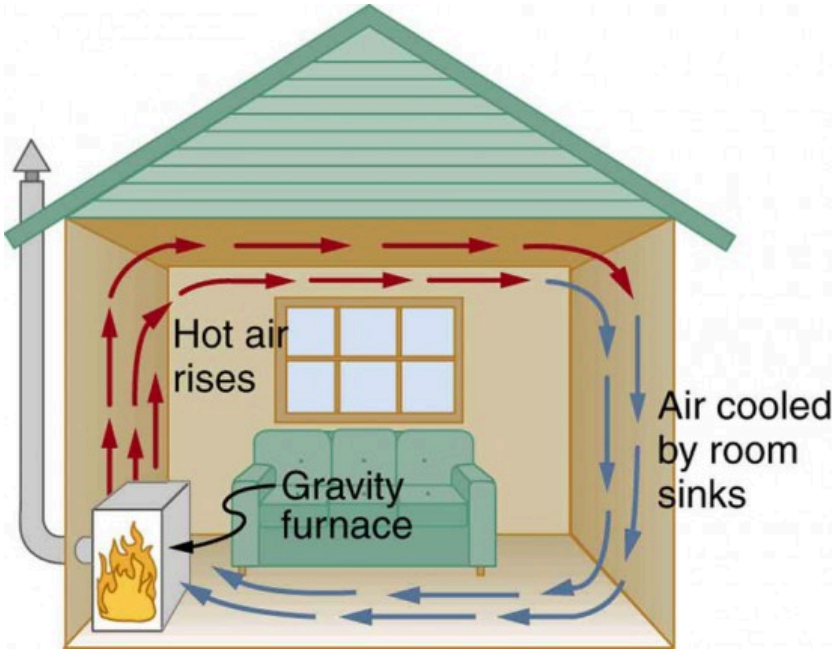


Figure 1. Air heated by the so-called gravity furnace expands and rises, forming a convective loop that transfers energy to other parts of the room. As the air is cooled at the ceiling and outside walls, it contracts, eventually becoming denser than room air and sinking to the floor. A properly designed heating system using natural convection, like this one, can be quite efficient in uniformly heating a home.

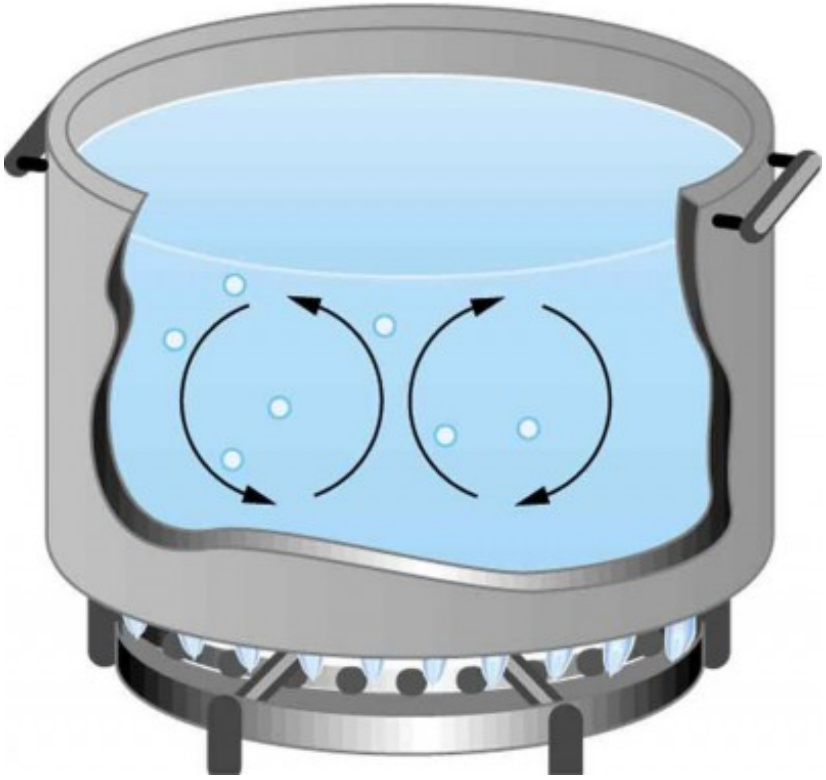


Figure 2. Convection plays an important role in heat transfer inside this pot of water. Once conducted to the inside, heat transfer to other parts of the pot is mostly by convection. The hotter water expands, decreases in density, and rises to transfer heat to other regions of the water, while colder water sinks to the bottom. This process keeps repeating.

Take-Home Experiment: Convection Rolls in a Heated Pan

Take two small pots of water and use an eye dropper to place a drop of food coloring near the bottom of each. Leave one on a bench top and heat the other over a stovetop. Watch how the color spreads and how long it takes the color to reach the top. Watch how convective loops form.

Example 1. Calculating Heat Transfer by Convection: Convection of Air Through the Walls of a House

Most houses are not airtight: air goes in and out around doors and windows, through cracks and crevices, following wiring to switches and outlets, and so on. The air in a typical house is completely replaced in less than an hour. Suppose that a moderately-sized house has inside dimensions 12.0 m × 18.0 m × 3.00 m high, and that all air is replaced in 30.0 min. Calculate the heat transfer per unit time in watts needed to warm the incoming cold air by 10.0ºC, thus replacing the heat transferred by convection alone.

Strategy

Heat is used to raise the temperature of air so that $Q = mc\Delta T$. The rate of heat transfer is then

$$\frac{Q}{t}$$

, where t is the time for air turnover. We are given that ΔT is 10.0ºC, but we must still find values for the mass of air and its specific heat before we can calculate Q. The specific heat of air is a weighted average of the specific heats of nitrogen and oxygen, which gives $c = c_p \cong 1000$ J/kg · ºC from Table 1 (note that the specific heat at constant pressure must be used for this process).

Solution

1. Determine the mass of air from its density and the given volume of the house. The density is given from the density ρ and the volume $m = \rho V = (1.29 \text{ kg/m}^3)(12.0 \text{ m} \times 18.0 \text{ m} \times 3.00 \text{ m}) = 836$ kg.
2. Calculate the heat transferred from the change in air temperature: $Q = mc\Delta T$ so that $Q = (836 \text{ kg})(1000 \text{ J/kg} \cdot ^\circ\text{C})(10.0^\circ\text{C}) = 8.36 \times 10^6$ J.
3. Calculate the heat transfer from the heat Q and the turnover time t. Since air is turned over in $t = 0.500$ h = 1800 s, the heat transferred per unit time is $\frac{Q}{t} = \frac{8.36 \times 10^6 \text{ J}}{1800 \text{ s}} = 4.64 \text{ kW}$.

Discussion

This rate of heat transfer is equal to the power consumed by about forty-six 100-W light bulbs. Newly constructed homes are designed for a turnover time of 2 hours or more, rather than 30 minutes for the house of this example. Weather stripping, caulking, and improved window seals are commonly employed. More extreme measures are sometimes taken in very cold (or hot) climates to achieve a tight standard of more than 6 hours for one air turnover. Still longer turnover times are unhealthy, because a minimum amount of fresh air is necessary to supply oxygen for breathing and to dilute household pollutants. The term used for the process by which outside air leaks into the house from cracks around windows, doors, and the foundation is called “air infiltration.”

A cold wind is much more chilling than still cold air, because convection combines with conduction in the body to increase the rate at which energy is transferred away from the body. The table below gives approximate wind-chill factors, which are the temperatures of still air that produce the same rate of cooling as air of a given temperature and speed. Wind-chill factors are a dramatic reminder of convection’s ability to transfer heat faster than conduction. For example, a 15.0 m/s wind at 0ºC has the chilling equivalent of still air at about −18ºC.

Table 1. Wind-Chill Factors

Moving air temperature	**Wind speed (m/s)**				
(ºC)	2	5	10	15	0
5	3	−1	−8	−10	−12
2	0	−7	−12	−16	−18
0	−2	−9	−15	−18	−20
−5	−7	−15	−22	−26	−29
−10	−12	−21	−29	−34	−36
−20	−23	−34	−44	−50	−52
−10	−12	−21	−29	−34	−36
−20	−23	−34	−44	−50	−52
−40	−44	−59	−73	−82	−84

Although air can transfer heat rapidly by convection, it is a poor conductor and thus a good insulator. The amount of available space for airflow determines whether air acts as an insulator or conductor. The space between the inside and outside walls of a 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. Similarly, the gap between the two panes of a double-paned window is about 1 cm, which prevents convection and takes advantage of air’s low conductivity to prevent greater loss. Fur, fiber, and fiberglass also take advantage of the low conductivity of air by trapping it in spaces too small to support convection, as shown in the figure. Fur and feathers are lightweight and thus ideal for the protection of animals.

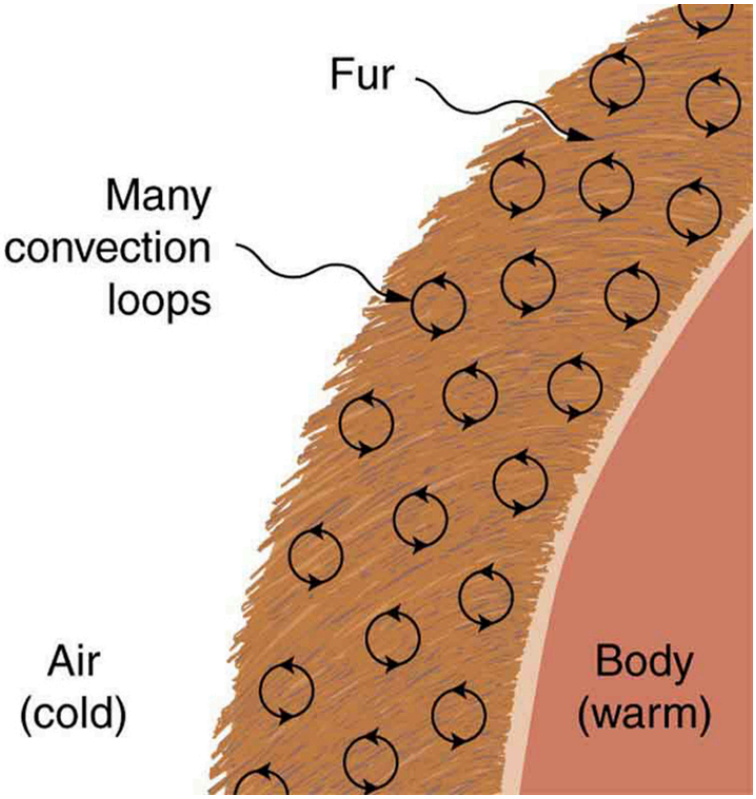


Figure 3. Fur is filled with air, breaking it up into many small pockets. Convection is very slow here, because the loops are so small. The low conductivity of air makes fur a very good lightweight insulator.

Some interesting phenomena happen *when convection is accompanied by a phase change*. It 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 2. Calculate the Flow of Mass during Convection: Sweat-Heat Transfer away from the Body

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? (This evaporation might occur 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 a phase change ($Q = mL_v$). Thus, the energy loss per unit time is

$$\frac{Q}{t} = \frac{mL_v}{t} = 120 \text{ W} = 120 \text{ J/s}$$

.

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 1 in Phase Change and Latent Heat, L_v = 2430 kJ/kg = 2430 J/g. This yields

$$\frac{m}{t} = \frac{120\ \text{J/s}}{2430\ \text{J/g}} = 0.0494\ \text{g/s} = 2.96\ \text{g/min}$$

Discussion

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.

Figure 4. Cumulus clouds are caused by water vapor that rises because of convection. The rise of clouds is driven by a positive feedback mechanism. (credit: Mike Love)

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, 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. Water vapor carried in by convection condenses, releasing tremendous amounts of energy. This energy causes the air to expand and rise, where it is colder. More condensation occurs in these colder regions, which in turn drives the cloud even higher. Such a mechanism is called positive feedback, since the process reinforces and accelerates itself.

Figure 5. Convection accompanied by a phase change releases the energy needed to drive this thunderhead into the stratosphere. (credit: Gerardo García Moretti)

These systems sometimes produce violent storms, with lightning and hail, and constitute the mechanism driving hurricanes (Figure 5).

The movement of icebergs (Figure 6) is another example of convection accompanied by a phase change. Suppose an iceberg drifts from Greenland into warmer Atlantic waters. Heat is removed from the warm ocean water when the ice melts and heat is released to the land mass when the iceberg forms on Greenland.

Figure 6. The phase change that occurs when this iceberg melts involves tremendous heat transfer. (credit: Dominic Alves)

Check Your Understanding

Explain why using a fan in the summer feels refreshing!

Solution

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.

Section Summary

Convection is heat transfer by the macroscopic movement of mass. Convection can be natural or forced and generally transfers thermal energy faster than conduction. Table 1 gives wind-chill factors, indicating that moving air has the same chilling effect of much colder stationary air. Convection that occurs along with a phase change can transfer energy from cold regions to warm ones.

Conceptual Questions

1. One way to make a fireplace more energy efficient is to have an external air supply for the combustion of its fuel. Another 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 in each.
2. On cold, clear nights horses will sleep under the cover of large trees. How does this help them keep warm?

Problems & Exercises

1. At what wind speed does −10ºC air cause the same chill factor as still air at −29ºC?
2. At what temperature does still air cause the same chill factor as −5ºC air moving at 15 m/s?
3. The “steam” above a freshly made cup of instant coffee is really water vapor droplets condensing after evaporating from the hot coffee. What is the final temperature of 250 g of hot coffee initially at 90.0ºC if 2.00 g evaporates from it? The coffee is in a Styrofoam cup, so other methods of heat transfer can be neglected.
4. (a) How many kilograms of water must evaporate from a 60.0-kg woman to lower her body temperature by 0.750ºC? (b) Is this a reasonable amount of water to evaporate in the form of perspiration, assuming the relative humidity of the surrounding air is low?
5. On a hot dry day, evaporation from a lake has just enough heat transfer to balance the 1.00 kW/m^2 of incoming heat from the Sun. What mass of water evaporates in 1.00 h from each square meter?
6. One winter day, the climate control system of a large university classroom building malfunctions. As a result, 500 m^3 of excess cold air is brought in each minute. At what rate in kilowatts must heat transfer occur to warm this air by 10.0ºC (that is, to bring the air to room temperature)?
7. The Kilauea volcano in Hawaii is the world’s most active, disgorging about 5×10^5 m^3 of 1200ºC lava per day. What is the rate of heat transfer out of Earth by convection if this lava has a density of 2700 kg/m^3 and eventually cools to 30ºC? Assume that the specific heat of lava is the same as that of granite.

Figure 7. Lava flow on Kilauea volcano in Hawaii. (credit: J. P. Eaton, U.S. Geological Survey)

8. During heavy exercise, the body pumps 2.00 L of blood per minute to the surface, where it is cooled by 2.00ºC. What is the rate of heat transfer from this forced convection alone, assuming blood has the same specific heat as water and its density is 1050 kg/m^3?
9. A person inhales and exhales 2.00 L of 37.0ºC air, evaporating 4.00×10^{-2} g of water from the lungs and breathing passages with each breath. (a) How much heat transfer occurs due to evaporation in each breath? (b) What is the rate of heat transfer in watts if the person is breathing at a moderate rate of 18.0 breaths per minute? (c) If the inhaled air had a temperature of 20.0ºC, what is the rate of heat transfer for warming the air? (d) Discuss the total rate of heat transfer as it relates to typical metabolic rates. Will this breathing be a major form of heat transfer for this person?
10. A glass coffee pot has a circular bottom with a 9.00-cm diameter in contact with a heating element that keeps the coffee warm with a continuous heat transfer rate of 50.0 W (a) What is the temperature of the bottom of the pot, if it is 3.00 mm thick and the inside temperature is 60.0ºC? (b) If the temperature of the coffee remains constant and all of the heat transfer is removed by evaporation, how many grams per minute evaporate? Take the heat of vaporization to be 2340 kJ/kg.

Selected Solutions to Problems & Exercises

1. 10 m/s

3. 85.7ºC

5. 1.48 kg

7. 2×10^4 MW

9. (a) 97.2 J; (b) 29.2 W; (c) 9.49 W; (d) The total rate of heat loss would be 29.2 W + 9.49 W = 38.7 W.

While sleeping, our body consumes 83 W of power, while sitting it consumes 120 to 210 W. Therefore, the total rate of heat loss from breathing will not be a major form of heat loss for this person.

Radiation

Lumen Learning

Learning Objectives

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

- Discuss heat transfer by radiation.
- Explain the power of different materials.

You can feel the heat transfer from a fire and from the Sun. 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. The space between the Earth and the Sun is largely empty, without any possibility of heat transfer by convection or conduction. In these examples, heat is transferred by radiation. That is, the hot body emits electromagnetic waves that are absorbed by our 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.

The energy of electromagnetic radiation depends on the wavelength (color) and varies over a wide range: a smaller wavelength (or higher frequency) corresponds to a higher energy. Because more heat is radiated at higher temperatures, a temperature change is accompanied by a color change. Take, for example, an electrical element on a stove, which glows from red to orange, while the higher-temperature steel in a blast furnace glows from yellow to white. The radiation you feel is mostly infrared, which corresponds to a lower temperature than that of the electrical element and the steel. The radiated energy depends on its intensity, which is represented in Figure 2 by the height of the distribution.

Electromagnetic Waves explains more about the electromagnetic spectrum and Introduction to Quantum Physics discusses how the decrease in wavelength corresponds to an increase in energy.

Figure 1. Most of the heat transfer from this fire to the observers is 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 that you can sense the presence of a fire without looking at it directly. (credit: Daniel X. O'Neil)

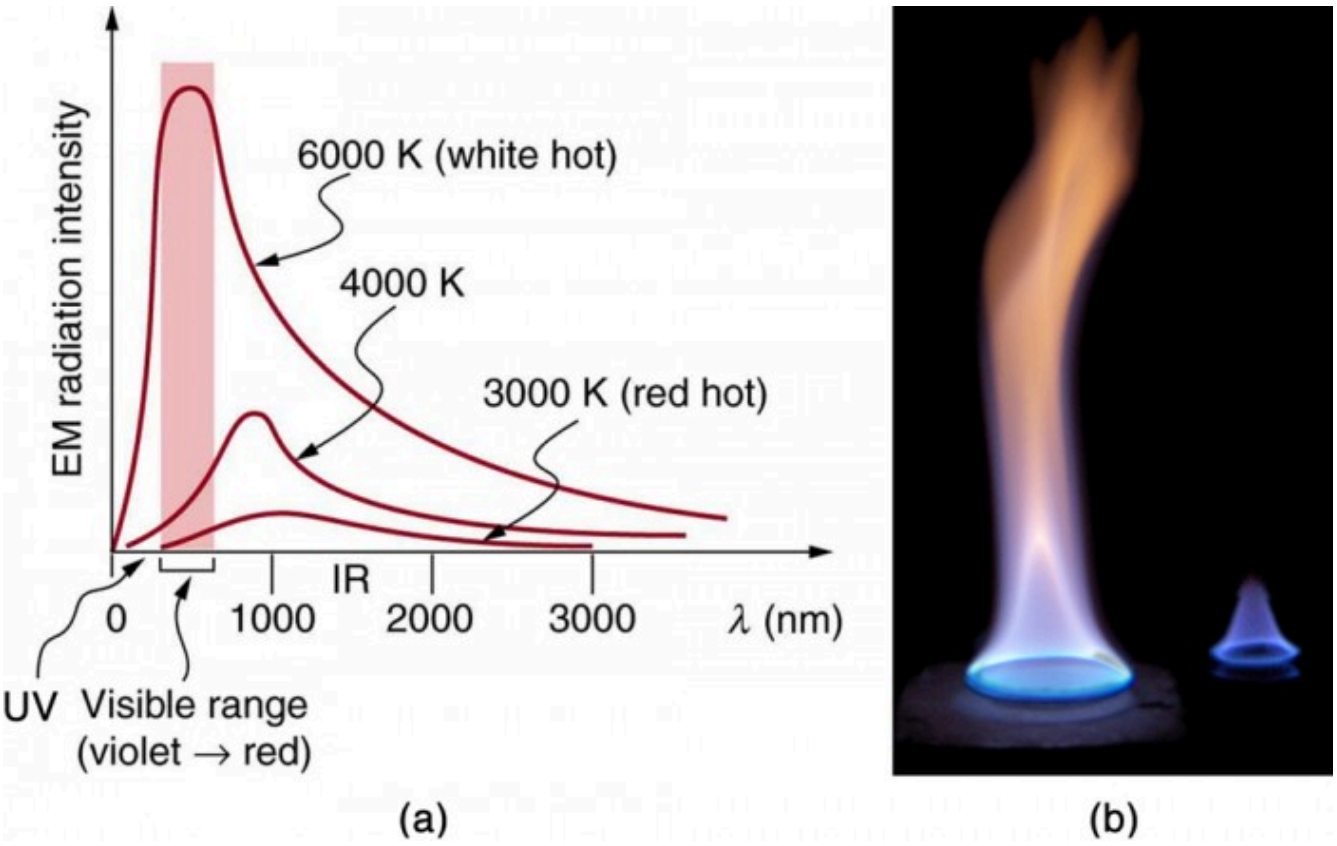


Figure 2. (a) A graph of the spectra of electromagnetic waves emitted from an ideal radiator at three different temperatures. The intensity or rate of radiation emission increases dramatically with temperature, and the spectrum shifts toward the visible and ultraviolet parts of the spectrum. The shaded 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. (credit: Tuohirulla)

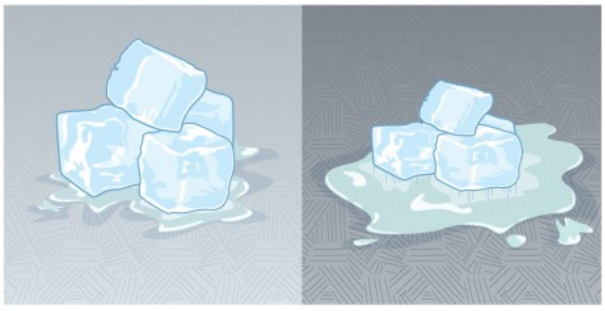


Figure 3. This illustration shows that the darker pavement is hotter than the lighter pavement (much more of the ice on the right has melted), although both have been in the sunlight for the same time. The thermal conductivities of the pavements are the same.

All objects absorb and emit electromagnetic radiation. The rate of heat transfer by radiation is largely determined by the color of the object. Black is the most effective, and white is the least effective. People living in hot climates generally avoid wearing black clothing, for instance (see Take-Home Experiment: Temperature in the Sun). Similarly, black asphalt in a parking lot will be hotter than adjacent gray sidewalk on a summer day, because black absorbs better than gray. The reverse is also true—black radiates better than gray. Thus, on a clear summer night, the asphalt will be colder than the gray sidewalk, because black radiates the energy more rapidly than gray. An *ideal radiator* is the same color as an *ideal absorber*, and captures all the radiation that falls on it. In contrast, white is a poor absorber and is also a poor radiator. A white object reflects all radiation, like a mirror. (A perfect, polished white surface is mirror-like in appearance, and a crushed mirror looks white.)

Gray objects have a uniform ability to absorb all parts of the electromagnetic spectrum. Colored objects behave in similar but more complex ways, which gives them a particular color in the visible range and may make them special in other ranges of the nonvisible spectrum. Take, for example, the strong absorption of infrared radiation by the skin, which allows us to be very sensitive to it.

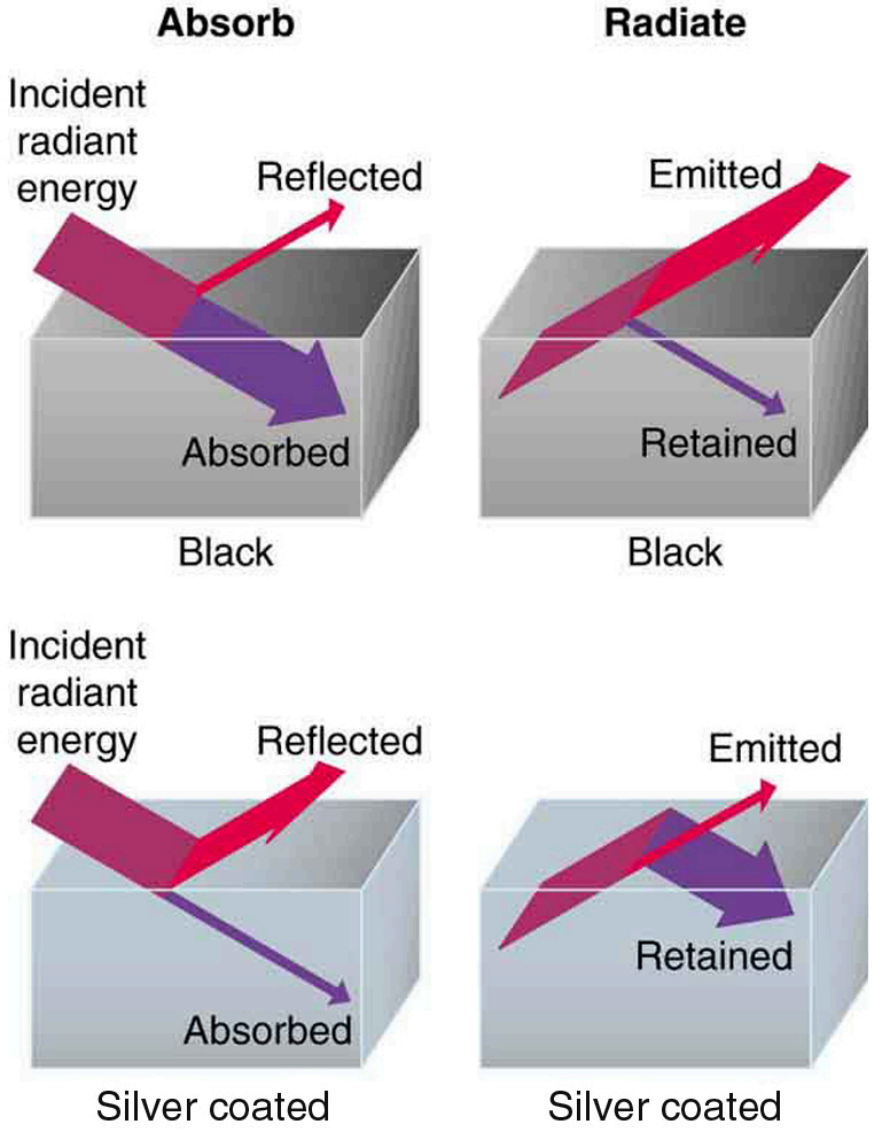


Figure 4. A black object is a good absorber and a good radiator, while a white (or silver) object is a poor absorber and a poor radiator. It is as if radiation from the inside is reflected back into the silver object, whereas radiation from the inside of the black object is "absorbed" when it hits the surface and finds itself on the outside and is strongly emitted.

The rate of heat transfer by emitted radiation is determined by the *Stefan-Boltzmann law of radiation*:

$$\frac{Q}{t} = \sigma e A T^4$$

,

where $\sigma = 5.67 \times 10^{-8}$ J/s · m^2 · K^4 is the Stefan-Boltzmann constant, *A* is the surface area of the object, and *T* is its absolute temperature in kelvin. The symbol *e* stands for the *emissivity* of the object, which is a measure of how well it radiates. An ideal jet-black (or black body) radiator has $e = 1$, whereas a perfect reflector has $e = 0$. Real objects fall between these two values. Take, for example, tungsten light bulb filaments which have an *e* of about 0.5, and carbon black (a material used in printer toner), which has the (greatest known) emissivity of about 0.99.

The radiation rate is directly proportional to the *fourth power* of the absolute temperature—a remarkably strong temperature dependence. Furthermore, the radiated heat is proportional to the surface area of the object. If you knock apart the coals of a fire, there is a noticeable increase in radiation due to an increase in radiating surface area.

Skin is a remarkably good absorber and emitter of infrared radiation, having an emissivity of 0.97 in the infrared spectrum. Thus, we are all nearly (jet) black in the infrared, in spite of the obvious variations in skin color. This high infrared emissivity is why we can so easily feel radiation on our skin. It is also the basis for the use of night scopes used by law enforcement and the military to detect human beings. Even small temperature variations can be detected because of the T^4 dependence. 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 5, optimize performance of blast furnaces, improve comfort levels in work environments, and even remotely map the Earth's temperature profile.

Figure 5. 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: U.S. Army)

All objects emit and absorb radiation. The *net* rate of heat transfer by radiation (absorption minus emission) is related to both the temperature of the object and the temperature of 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

$$\frac{Q_{\text{net}}}{t} = \sigma e\left(T_2^4 - T_1^4\right)$$

,

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

$$\frac{Q_{\text{net}}}{t}$$

is positive; that is, the net heat transfer is from hot to cold.

Take-Home Experiment: Temperature in the Sun

Place a thermometer out in the sunshine and shield it from direct sunlight using an aluminum foil. What is the reading? Now remove the shield, and note what the thermometer reads. Take a handkerchief soaked in nail polish remover, wrap it around the thermometer and place it in the sunshine. What does the thermometer read?

Example 1. Calculate the Net Heat Transfer of a Person: Heat Transfer by Radiation

What is the rate of heat transfer by radiation, with an unclothed person standing in a dark room whose ambient temperature is 22.0ºC. The person has a normal skin temperature of 33.0ºC and a surface area of 1.50 m^2. The emissivity of skin is 0.97 in the infrared, where the radiation takes place.

Strategy

We can solve this by using the equation for the rate of radiative heat transfer.

Solution

Insert the temperatures values T_2 = 295 K and T_1 = 306 K, so that

$$\begin{aligned}\frac{Q}{t} &= \sigma e A\left(T_2^4 - T_1^4\right)\\ &= \left(5.67 \times 10^{-8}\text{J/s} \cdot \text{m}^2 \cdot \text{K}^4\right)(0.97)\left(1.50\ \text{m}^2\right)\left[\left(295\ \text{K}^4\right) - \left(306\ \text{K}^4\right)\right]\\ &= -99\ \text{J/s} = -99\ \text{W}\end{aligned}$$

Discussion

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 will also be transferring energy to the environment. Indeed, we would probably expect this person to feel cold. Clothing significantly reduces heat transfer to the environment by many methods, because clothing slows down both conduction and convection, and has a lower emissivity (especially if it is white) than skin.

The Earth receives almost all its energy from radiation of the Sun and reflects some of it back into outer space. Because the Sun is hotter than the Earth, the net energy flux is from the Sun to the Earth. However, the rate of energy transfer is less than the equation for the radiative heat transfer would predict because the Sun does not fill the sky. The average emissivity (*e*) of the Earth is about 0.65, but the calculation of this value is complicated by the fact that the highly reflective cloud coverage varies greatly from day to day. There is a negative feedback (one in which a change produces an effect that opposes that change) between clouds and heat transfer; greater 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 the Earth's emissivity with radiation type (see Figure 6). The greenhouse effect is a natural phenomenon responsible for providing temperatures suitable for life on Earth. The Earth's relatively constant temperature is a result of the energy balance between the incoming solar radiation and the energy radiated from the Earth. Most of the infrared radiation emitted from the Earth is absorbed by carbon dioxide (CO_2) and water (H_2O) in the atmosphere and then re-radiated back

to the Earth or into outer space. Re-radiation back to the Earth maintains its surface temperature about 40ºC higher than it would be if there was no atmosphere, similar to the way glass increases temperatures in a greenhouse.

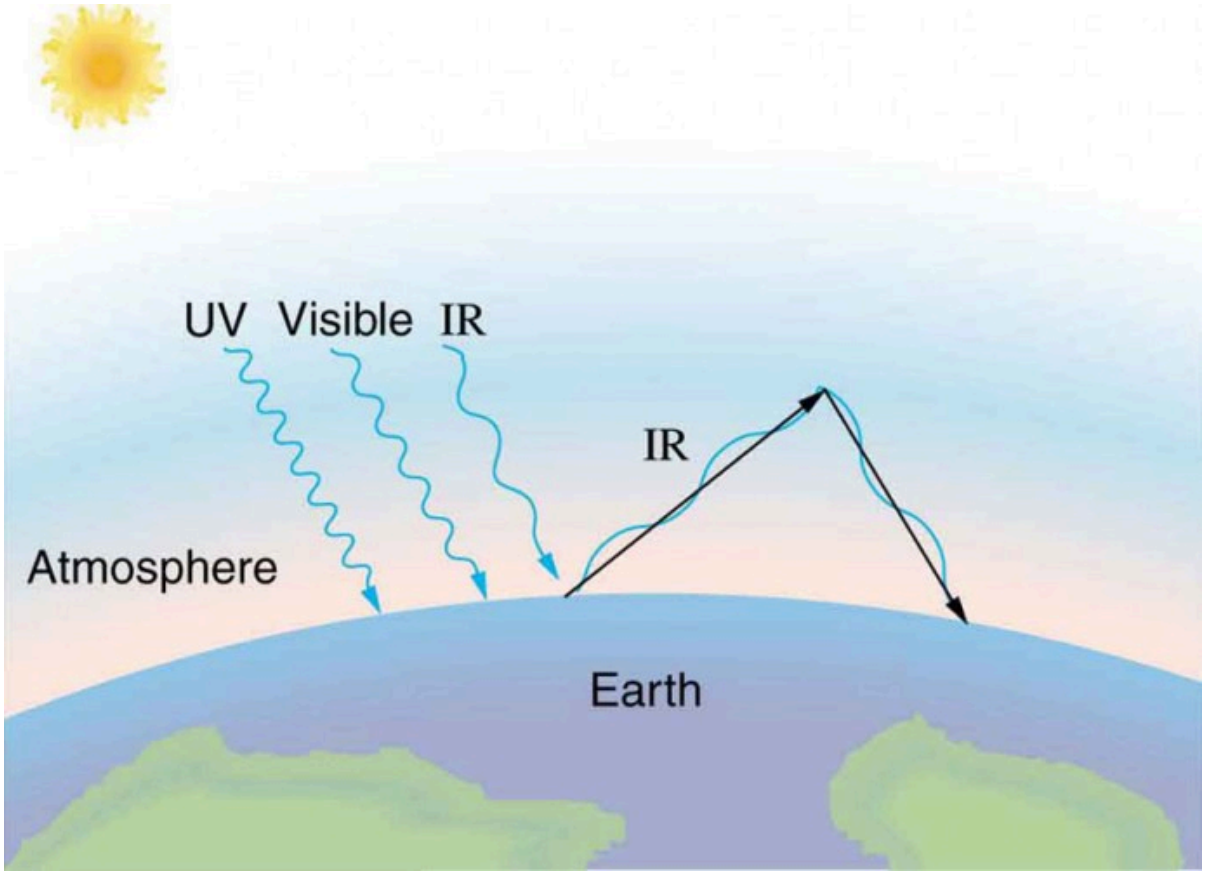


Figure 6. The greenhouse effect is a name given to the trapping of energy in the Earth's atmosphere by a process similar to that used in greenhouses. The atmosphere, like window glass, is transparent to incoming visible radiation and most of the Sun's infrared. These wavelengths are absorbed by the Earth and re-emitted as infrared. Since Earth's temperature is much lower than that of the Sun, the infrared radiated by the Earth has a much longer wavelength. The atmosphere, like glass, traps these longer infrared rays, keeping the Earth warmer than it would otherwise be. The amount of trapping depends on concentrations of trace gases like carbon dioxide, and a change in the concentration of these gases is believed to affect the Earth's surface temperature.

The greenhouse effect is also central to the discussion of global warming due to emission of carbon dioxide and methane (and other so-called greenhouse gases) into the Earth's atmosphere from industrial production 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.

Heating and cooling are often significant contributors to energy use in individual homes. Current research efforts into developing environmentally friendly homes quite often focus on reducing conventional heating and cooling through better building materials, strategically positioning windows to optimize radiation gain from the Sun, and opening spaces to allow convection. It is possible to build a zero-energy house that allows for comfortable living in most parts of the United States with hot and humid summers and cold winters.

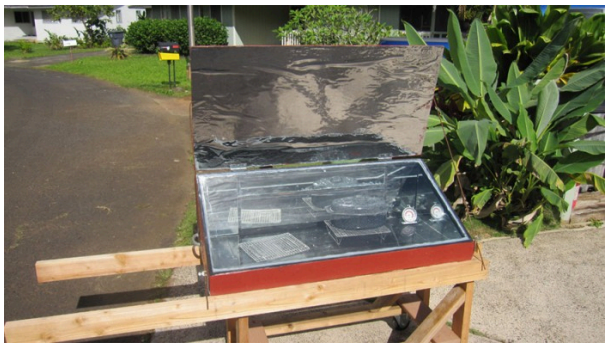


Figure 7. This simple but effective solar cooker uses the greenhouse effect and reflective material to trap and retain solar energy. Made of inexpensive, durable materials, it saves money and labor, and is of particular economic value in energy-poor developing countries. (credit: E.B. Kauai)

Conversely, dark space is very cold, about 3K(−454ºF), so that the Earth radiates energy into the dark sky. Owing to the fact that 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. 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.

Check Your Understanding

What is the change in the rate of the radiated heat by a body at the temperature T_1 = 20ºC compared to when the body is at the temperature T_2 = 40ºC?

Solution

The radiated heat is proportional to the fourth power of the *absolute temperature*. Because T_1 = 293 K and T_2 = 313 K, the rate of heat transfer increases by about 30 percent of the original rate.

Career Connection: Energy Conservation Consultation

The cost of energy is generally believed to remain very high for the foreseeable future. Thus, passive control of heat loss in both commercial and domestic housing will become increasingly important. Energy consultants measure and analyze the flow of energy into and out of houses and ensure that a healthy exchange of air is maintained inside the house. The job prospects for an energy consultant are strong.

Problem-Solving Strategies for the Methods 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 very useful.
4. Make a list of what is given or can be inferred from the problem as stated (identify the knowns).
5. Solve the appropriate equation for the quantity to be determined (the unknown).
$$\frac{Q}{t} = \frac{kA\,(T_2 - T_1)}{d}$$
6. For conduction, equation is appropriate. Table 1 in Conduction lists thermal conductivities. For convection, determine the amount of matter moved and use equation Q = mcΔT, to calculate the heat transfer involved in the temperature change of the fluid. If a phase change accompanies convection, equation $Q = mL_f$ or $Q = mL_v$ is appropriate to find the heat transfer involved in the phase change. Table 1 in Phase Change and Latent Heat lists information relevant to phase change. For radiation, equation $\frac{Q_{\text{net}}}{t} = \sigma e A\left(T_2^4 - T_1^4\right)$ gives the net heat transfer rate.
7. Insert 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?

Section Summary

- Radiation is the rate of heat transfer through the emission or absorption of electromagnetic waves.
- The rate of heat transfer depends on the surface area and the fourth power of the absolute temperature: $\frac{Q}{t} = \sigma e A T^4$, where $\sigma = 5.67 \times 10^{-8}$ J/s · m^2 · K^4 is the Stefan-Boltzmann constant and e is the emissivity of the body. For a black body, $e = 1$ whereas a shiny white or perfect reflector has $e = 0$, with real objects having values of e between 1 and 0. The net rate of heat transfer 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 an object surrounded by an environment with uniform temperature T_2 and e is the emissivity of the object.

Conceptual Questions

1. When watching a daytime circus in a large, dark-colored tent, you sense significant heat transfer from the tent. Explain why this occurs.
2. Satellites designed to observe the radiation from cold (3 K) dark space have sensors that are shaded from the Sun, Earth, and Moon and that are cooled to very low temperatures. Why must the sensors be at low temperature?

3. Why are cloudy nights generally warmer than clear ones?
4. 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 and also if it is not?
5. On average, would Earth be warmer or cooler without the atmosphere? Explain your answer.

Problems & Exercises

1. At what net rate does heat radiate from a 275-m^2 black roof on a night when the roof's temperature is 30.0ºC and the surrounding temperature is 15.0ºC? The emissivity of the roof is 0.900.
2. (a) Cherry-red embers in a fireplace are at 850ºC and have an exposed area of 0.200 m^2 and an emissivity of 0.980. The surrounding room has a temperature of 18.0ºC. If 50% of the radiant energy enters the room, what is the net rate of radiant heat transfer in kilowatts? (b) Does your answer support the contention that most of the heat transfer into a room by a fireplace comes from infrared radiation?
3. Radiation makes it impossible to stand close to a hot lava flow. Calculate the rate of heat transfer by radiation from 1.00 m^2 of 1200ºC fresh lava into 30.0ºC surroundings, assuming lava's emissivity is 1.00.
4. (a) Calculate the rate of heat transfer by radiation from a car radiator at 110ºC into a 50.0ºC environment, if the radiator has an emissivity of 0.750 and a 1.20-m^2 surface area. (b) Is this a significant fraction of the heat transfer by an automobile engine? To answer this, assume a horsepower of 200 hp(1.5 kW) and the efficiency of automobile engines as 25%.
5. Find the net rate of heat transfer by radiation from a skier standing in the shade, given the following. She is completely clothed in white (head to foot, including a ski mask), the clothes have an emissivity of 0.200 and a surface temperature of 10.0ºC, the surroundings are at−15.0ºC, and her surface area is 1.60 m^2.
6. Suppose you walk into a sauna that has an ambient temperature of 50.0ºC. (a) Calculate the rate of heat transfer to you by radiation given your skin temperature is 37.0ºC, the emissivity of skin is 0.98, and the surface area of your body is 1.50 m^2. (b) If all other forms of heat transfer are balanced (the net heat transfer is zero), at what rate will your body temperature increase if your mass is 75.0 kg?
7. Thermography is a technique for measuring radiant heat and detecting variations in surface temperatures that may be medically, environmentally, or militarily meaningful.(a) What is the percent increase in the rate of heat transfer by radiation from a given area at a temperature of 34.0ºC compared with that at 33.0ºC, such as on a person's skin? (b) What is the percent increase in the rate of heat transfer by radiation from a given area at a temperature of 34.0ºC compared with that at 20.0ºC, such as for warm and cool automobile hoods?

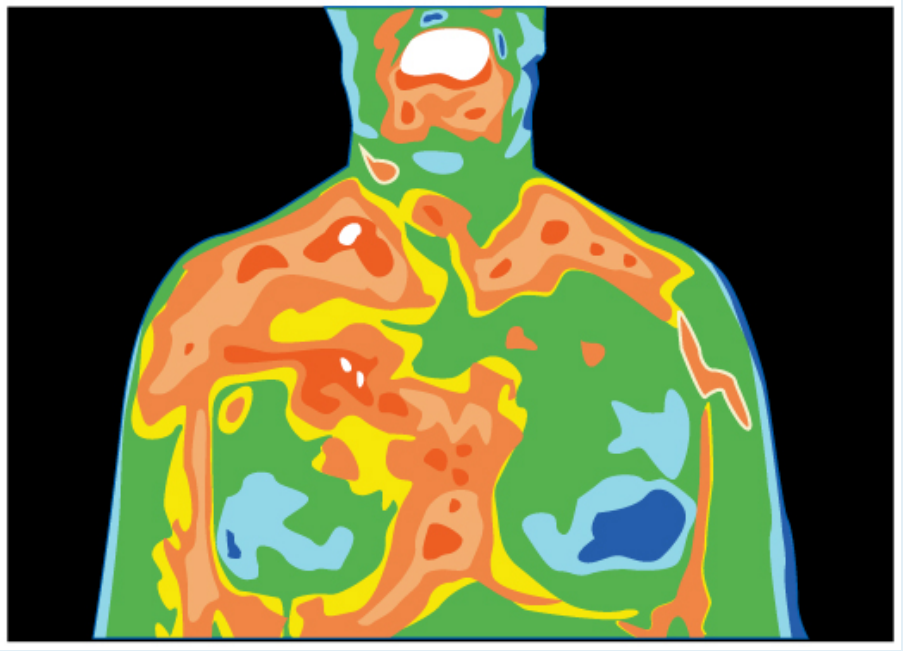


Figure 8. Artist's rendition of a thermograph of a patient's upper body, showing the distribution of heat represented by different colors.

8. The Sun radiates like a perfect black body with an emissivity of exactly 1. (a) Calculate the surface temperature of the Sun, given that it is a sphere with a 7.00×10^8-m radius that radiates 3.80×10^{26} W into 3-K space. (b) How much power does the Sun radiate per square meter of its surface? (c) How much power in watts per square meter is that value at the distance of Earth, 1.50×10^{11} m away? (This number is called the solar constant.)
9. A large body of lava from a volcano has stopped flowing and is slowly cooling. The interior of the lava is at 1200ºC, its surface is at 450ºC, and the surroundings are at 27.0ºC. (a) Calculate the rate at which energy is transferred by radiation from 1.00 m2 of surface lava into the surroundings, assuming the emissivity is 1.00. (b) Suppose heat conduction to the surface occurs at the same rate. What is the thickness of the lava between the 450ºC surface and the 1200ºC interior, assuming that the lava's conductivity is the same as that of brick?
10. Calculate the temperature the entire sky would have to be in order to transfer energy by radiation at 1000 W/m^2—about the rate at which the Sun radiates when it is directly overhead on a clear day. This value is the effective temperature of the sky, a kind of average that takes account of the fact that the Sun occupies only a small part of the sky but is much hotter than the rest. Assume that the body receiving the energy has a temperature of 27.0ºC.
11. (a) A shirtless rider under a circus tent feels the heat radiating from the sunlit portion of the tent. Calculate the temperature of the tent canvas based on the following information: The shirtless rider's skin temperature is 34.0ºC and has an emissivity of 0.970. The exposed area of skin is 0.400 m2. He receives radiation at the rate of 20.0 W—half what you would calculate if the entire region behind him was hot. The rest of the surroundings are at 34.0ºC. (b) Discuss how this situation would change if the sunlit side of the tent was nearly pure white and if the rider was covered by a white tunic.
12. **Integrated Concepts.** One 30.0ºC day the relative humidity is 75.0%, and that evening the temperature drops to 20.0ºC, well below the dew point. (a) How many grams of water condense

from each cubic meter of air? (b) How much heat transfer occurs by this condensation? (c) What temperature increase could this cause in dry air?

13. **Integrated Concepts.** Large meteors sometimes strike the Earth, converting most of their kinetic energy into thermal energy. (a) What is the kinetic energy of a 109 kg meteor moving at 25.0 km/s? (b) If this meteor lands in a deep ocean and 80% of its kinetic energy goes into heating water, how many kilograms of water could it raise by 5.0ºC? (c) Discuss how the energy of the meteor is more likely to be deposited in the ocean and the likely effects of that energy.
14. **Integrated Concepts.** Frozen waste from airplane toilets has sometimes been accidentally ejected at high altitude. Ordinarily it breaks up and disperses over a large area, but sometimes it holds together and strikes the ground. Calculate the mass of 0ºC ice that can be melted by the conversion of kinetic and gravitational potential energy when a 20.0 kg piece of frozen waste is released at 12.0 km altitude while moving at 250 m/s and strikes the ground at 100 m/s (since less than 20.0 kg melts, a significant mess results).
15. **Integrated Concepts.** (a) A large electrical power facility produces 1600 MW of "waste heat," which is dissipated to the environment in cooling towers by warming air flowing through the towers by 5.00ºC. What is the necessary flow rate of air in m^3/s? (b) Is your result consistent with the large cooling towers used by many large electrical power plants?
16. **Integrated Concepts.** (a) Suppose you start a workout on a Stairmaster, producing power at the same rate as climbing 116 stairs per minute. Assuming your mass is 76.0 kg and your efficiency is 20.0%, how long will it take for your body temperature to rise 1.00ºC if all other forms of heat transfer in and out of your body are balanced? (b) Is this consistent with your experience in getting warm while exercising?
17. **Integrated Concepts.** A 76.0-kg person suffering from hypothermia comes indoors and shivers vigorously. How long does it take the heat transfer to increase the person's body temperature by 2.00ºC if all other forms of heat transfer are balanced?
18. **Integrated Concepts.** In certain large geographic regions, the underlying rock is hot. Wells can be drilled and water circulated through the rock for heat transfer for the generation of electricity. (a) Calculate the heat transfer that can be extracted by cooling 1.00 km^3 of granite by 100ºC. (b) How long will it take for heat transfer at the rate of 300 MW, assuming no heat transfers back into the 1.00km3 of rock by its surroundings?
19. **Integrated Concepts.** Heat transfers from your lungs and breathing passages by evaporating water. (a) Calculate the maximum number of grams of water that can be evaporated when you inhale 1.50 L of 37ºC air with an original relative humidity of 40.0%. (Assume that body temperature is also 37ºC.) (b) How many joules of energy are required to evaporate this amount? (c) What is the rate of heat transfer in watts from this method, if you breathe at a normal resting rate of 10.0 breaths per minute?
20. **Integrated Concepts.** (a) What is the temperature increase of water falling 55.0 m over Niagara Falls? (b) What fraction must evaporate to keep the temperature constant?
21. **Integrated Concepts.** Hot air rises because it has expanded. It then displaces a greater volume of cold air, which increases the buoyant force on it. (a) Calculate the ratio of the buoyant force to the weight of 50.0ºC air surrounded by 20.0ºC air. (b) What energy is needed to cause 1.00m3 of air to go from 20.0ºC to 50.0ºC? (c) What gravitational potential energy is gained by this volume of air if it rises 1.00 m? Will this cause a significant cooling of the air?
22. **Unreasonable Results.** (a) What is the temperature increase of an 80.0 kg person who consumes

2500 kcal of food in one day with 95.0% of the energy transferred as heat to the body? (b) What is unreasonable about this result? (c) Which premise or assumption is responsible?

23. **Unreasonable Results.** A slightly deranged Arctic inventor surrounded by ice thinks it would be much less mechanically complex to cool a car engine by melting ice on it than by having a water-cooled system with a radiator, water pump, antifreeze, and so on. (a) If 80.0% of the energy in 1.00 gal of gasoline is converted into "waste heat" in a car engine, how many kilograms of 0ºC ice could it melt? (b) Is this a reasonable amount of ice to carry around to cool the engine for 1.00 gal of gasoline consumption? (c) What premises or assumptions are unreasonable?
24. **Unreasonable Results.** (a) Calculate the rate of heat transfer by conduction through a window with an area of 1.00 m2 that is 0.750 cm thick, if its inner surface is at 22.0ºC and its outer surface is at 35.0ºC. (b) What is unreasonable about this result? (c) Which premise or assumption is responsible?
25. **Unreasonable Results.** A meteorite 1.20 cm in diameter is so hot immediately after penetrating the atmosphere that it radiates 20.0 kW of power. (a) What is its temperature, if the surroundings are at 20.0ºC and it has an emissivity of 0.800? (b) What is unreasonable about this result? (c) Which premise or assumption is responsible?
26. **Construct Your Own Problem.** Consider a new model of commercial airplane having its brakes tested as a part of the initial flight permission procedure. The airplane is brought to takeoff speed and then stopped with the brakes alone. Construct a problem in which you calculate the temperature increase of the brakes during this process. You may assume most of the kinetic energy of the airplane is converted to thermal energy in the brakes and surrounding materials, and that little escapes. Note that the brakes are expected to become so hot in this procedure that they ignite and, in order to pass the test, the airplane must be able to withstand the fire for some time without a general conflagration.
27. **Construct Your Own Problem.** Consider a person outdoors on a cold night. Construct a problem in which you calculate the rate of heat transfer from the person by all three heat transfer methods. Make the initial circumstances such that at rest the person will have a net heat transfer and then decide how much physical activity of a chosen type is necessary to balance the rate of heat transfer. Among the things to consider are the size of the person, type of clothing, initial metabolic rate, sky conditions, amount of water evaporated, and volume of air breathed. Of course, there are many other factors to consider and your instructor may wish to guide you in the assumptions made as well as the detail of analysis and method of presenting your results.

Glossary

emissivity: measure of how well an object radiates

greenhouse effect: warming of the Earth that is due to gases such as carbon dioxide and methane that absorb infrared radiation from the Earth's surface and reradiate it in all directions, thus sending a fraction of it back toward the surface of the Earth

net rate of heat transfer by radiation: is

$$\frac{Q_{\text{net}}}{t} = \sigma e A \left(T_2^4 - T_1^4\right)$$

radiation: energy transferred by electromagnetic waves directly as a result of a temperature difference

Stefan-Boltzmann law of radiation:

$$\frac{Q}{t} = \sigma e A T^4$$

, where σ is the Stefan-Boltzmann constant, A is the surface area of the object, T is the absolute temperature, and e is the emissivity

Selected Solutions to Problems & Exercises

1. −21.7 kW; note that the negative answer implies heat loss to the surroundings.

3. −266 kW

5. −36.0 W

7. (a) 1.31%; (b) 20.5%

9. (a) −15.0 kW; (b) 4.2 cm

11. (a) 48.5ºC; (b) A pure white object reflects more of the radiant energy that hits it, so a white tent would prevent more of the sunlight from heating up the inside of the tent, and the white tunic would prevent that heat which entered the tent from heating the rider. Therefore, with a white tent, the temperature would be lower than 48.5ºC, and the rate of radiant heat transferred to the rider would be less than 20.0 W.

13. (a) 3×10^{17} J; (b) 1×10^{13} kg; (c) When a large meteor hits the ocean, it causes great tidal waves, dissipating large amount of its energy in the form of kinetic energy of the water.

15. (a) 3.44×10^5 m^3/s; (b) This is equivalent to 12 million cubic feet of air per second. That is tremendous. This is too large to be dissipated by heating the air by only 5ºC. Many of these cooling towers use the circulation of cooler air over warmer water to increase the rate of evaporation. This would allow much smaller amounts of air necessary to remove such a large amount of heat because evaporation removes larger quantities of heat than was considered in part (a).

17. 20.9 min

19. (a) 3.96×10^{-2} g; (b) 96.2 J; (c) 16.0 W

21. (a) 1.102; (b) 2.79×10^4 J; (c) 12.6 J. This will not cause a significant cooling of the air because it is much less than the energy found in part (b), which is the energy required to warm the air from 20.0ºC to 50.0ºC.

22. (a) 36ºC; (b) Any temperature increase greater than about 3ºC would be unreasonably large. In this case the final temperature of the person would rise to 73ºC(163ºF); (c) The assumption of 95% heat retention is unreasonable.

24. (a) 1.46 kW; (b) Very high power loss through a window. An electric heater of this power can keep an entire room warm; (c) The surface temperatures of the window do not differ by as great an amount as assumed. The inner surface will be warmer, and the outer surface will be cooler.

5. Thermodynamics

Introduction to Thermodynamics

Lumen Learning

Heat transfer is energy in transit, and it can be used to do work. It can also be converted to any other form of energy. A car engine, for example, burns fuel for heat transfer into a gas. Work is done by the gas as it exerts a force through a distance, converting its energy into a variety of other forms—into the car's kinetic or gravitational potential energy; into electrical energy to run the spark plugs, radio, and lights; and back into stored energy in the car's battery. But most of the heat transfer produced from burning fuel in the engine does not do work on the gas. Rather, the energy is released into the environment, implying that the engine is quite inefficient.

Figure 1. A steam engine uses heat transfer to do work. Tourists regularly ride this narrow-gauge steam engine train near the San Juan Skyway in Durango, Colorado, part of the National Scenic Byways Program. (credit: Dennis Adams)

It is often said that modern gasoline engines cannot be made to be significantly more efficient. We hear the same about heat transfer to electrical energy in large power stations, whether they are coal, oil, natural gas, or nuclear powered. Why is that the case? Is the inefficiency caused by design problems that could be solved with better engineering and superior materials? Is it part of some money-making conspiracy by those who sell energy? Actually, the truth is more interesting, and reveals much about the nature of heat transfer.

Basic physical laws govern how heat transfer for doing work takes place and place insurmountable limits onto its efficiency. This chapter will explore these laws as well as many applications and concepts associated with them. These topics are part of *thermodynamics*—the study of heat transfer and its relationship to doing work.

The First Law of Thermodynamics

Lumen Learning

Learning Objectives

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

- Define the first law of thermodynamics.
- Describe how conservation of energy relates to the first law of thermodynamics.
- Identify instances of the first law of thermodynamics working in everyday situations, including biological metabolism.
- Calculate changes in the internal energy of a system, after accounting for heat transfer and work done.

If we are interested in how heat transfer is converted into doing work, then the conservation of energy principle is important. The first law of thermodynamics applies the conservation of energy principle to systems where heat transfer and doing work are the methods of transferring energy into and out of the system. The *first law of thermodynamics* states that the change in internal energy of a system equals the net heat transfer *into* the system minus the net work done *by* the system. In equation form, the first law of thermodynamics is $\Delta U = Q - W$.

Figure 1. This boiling tea kettle represents energy in motion. The water in the kettle is turning to water vapor because heat is being transferred from the stove to the kettle. As the entire system gets hotter, work is done—from the evaporation of the water to the whistling of the kettle. (credit: Gina Hamilton)

Here ΔU is the *change in internal energy* U of the system. Q is the *net heat transferred into the system*—that is, Q is the sum of all heat transfer into and out of the system. W is the *net work done by the system*—that is, W is the sum of all work done on or by the system. We use the following sign conventions: if Q is positive, then there is a net heat transfer into the system; if W is positive, then there is net work done by the system. So positive Q adds energy to the system and positive W takes energy from the system. Thus $\Delta U = Q - W$. Note also that if more heat transfer into the system occurs than work done, the difference is stored as internal energy. Heat engines are a good example of this—heat transfer into them takes place so that they can do work. (See Figure 2.) We will now examine Q, W, and ΔU further.

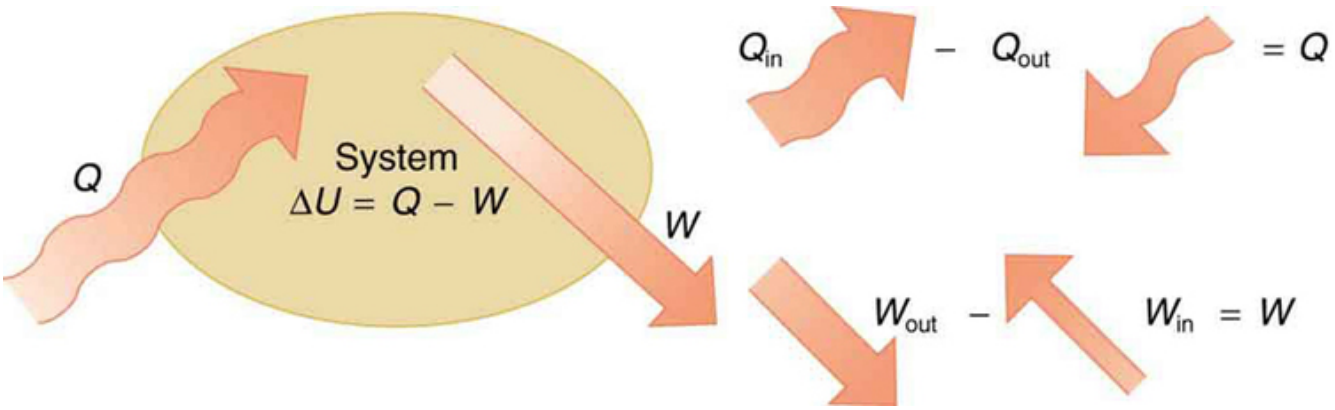


Figure 2. The first law of thermodynamics is the conservation-of-energy principle stated for a system where heat and work are the methods of transferring energy for a system in thermal equilibrium. Q represents the net heat transfer—it is the sum of all heat transfers into and out of the system. Q is positive for net heat transfer into the system. W is the total work done on and by the system. W is positive when more work is done by the system than on it. The change in the internal energy of the system, ΔU, is related to heat and work by the first law of thermodynamics, ΔU = Q − W.

Making Connections: Law of Thermodynamics and Law of Conservation of Energy

The first law of thermodynamics is actually the law of conservation of energy stated in a form most useful in thermodynamics. The first law gives the relationship between heat transfer, work done, and the change in internal energy of a system.

Heat *Q* and Work *W*

Heat transfer (*Q*) and doing work (*W*) are the two everyday means of bringing energy into or taking energy out of a system. The processes are quite different. Heat transfer, a less organized process, is driven by temperature differences. Work, a quite organized process, involves a macroscopic force exerted through a distance. Nevertheless, heat and work can produce identical results.For example, both can cause a temperature increase. Heat transfer into a system, such as when the Sun warms the air in a bicycle tire, can increase its temperature, and so can work done on the system, as when the bicyclist pumps air into the tire. Once the temperature increase has occurred, it is impossible to tell whether it was caused by heat transfer or by doing work. This uncertainty is an important point. Heat transfer and work are both energy in transit—neither is stored as such in a system. However, both can change the internal energy *U* of a system. Internal energy is a form of energy completely different from either heat or work.

Internal Energy *U*

We can think about the internal energy of a system in two different but consistent ways. The first is the atomic and molecular view, which examines the system on the atomic and molecular scale. The *internal energy U* of a system is the sum of the kinetic and potential energies of its atoms and molecules. Recall that kinetic plus potential energy is called mechanical energy. Thus internal energy is the sum of atomic and molecular mechanical energy. Because it is impossible to keep track of all individual atoms and molecules, we must deal with averages and distributions. A second way to view the internal energy of

a system is in terms of its macroscopic characteristics, which are very similar to atomic and molecular average values.

Macroscopically, we define the change in internal energy ΔU to be that given by the first law of thermodynamics: $\Delta U = Q - W$.

Many detailed experiments have verified that $\Delta U = Q - W$, where ΔU is the change in total kinetic and potential energy of all atoms and molecules in a system. It has also been determined experimentally that the internal energy U of a system depends only on the state of the system and *not how it reached that state.* More specifically, U is found to be a function of a few macroscopic quantities (pressure, volume, and temperature, for example), independent of past history such as whether there has been heat transfer or work done. This independence means that if we know the state of a system, we can calculate changes in its internal energy U from a few macroscopic variables.

Making Connections: Macroscopic and Microscopic

In thermodynamics, we often use the macroscopic picture when making calculations of how a system behaves, while the atomic and molecular picture gives underlying explanations in terms of averages and distributions. We shall see this again in later sections of this chapter. For example, in the topic of entropy, calculations will be made using the atomic and molecular view.

To get a better idea of how to think about the internal energy of a system, let us examine a system going from State 1 to State 2. The system has internal energy U_1 in State 1, and it has internal energy U_2 in State 2, no matter how it got to either state. So the change in internal energy $\Delta U = U_2 - U_1$ is independent of what caused the change. In other words, ΔU *is independent of path*. By path, we mean the method of getting from the starting point to the ending point. Why is this independence important? Note that $\Delta U = Q - W$. Both Q and *Wdepend on path*, but ΔU does not. This path independence means that internal energy U is easier to consider than either heat transfer or work done.

Example 1. Calculating Change in Internal Energy: The Same Change in *U* is Produced by Two Different Processes

1. Suppose there is heat transfer of 40.00 J to a system, while the system does 10.00 J of work. Later, there is heat transfer of 25.00 J out of the system while 4.00 J of work is done on the system. What is the net change in internal energy of the system?
2. What is the change in internal energy of a system when a total of 150.00 J of heat transfer occurs out of (from) the system and 159.00 J of work is done on the system? (See Figure 3).

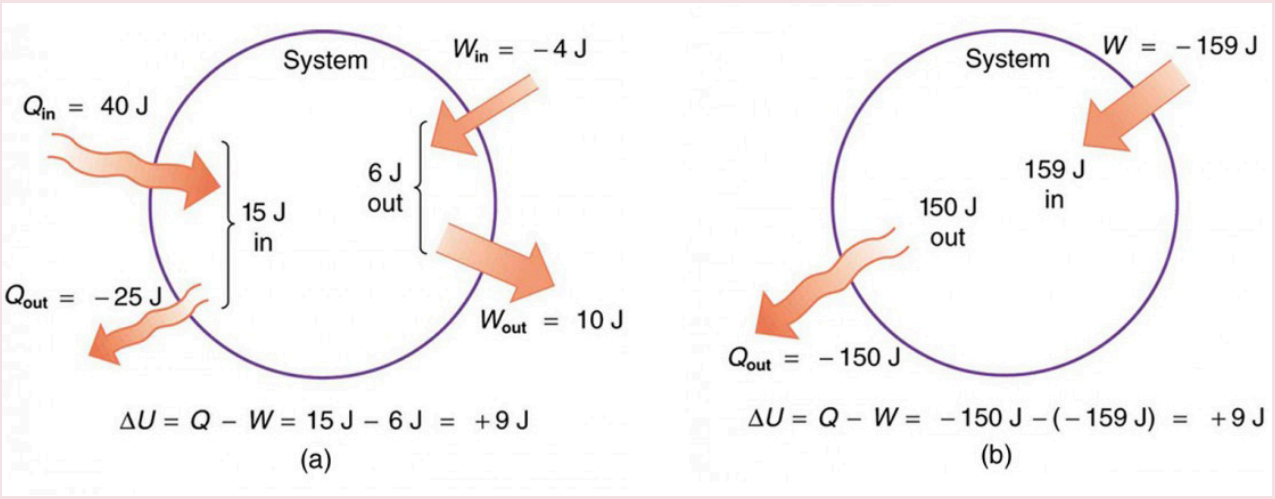


Figure 3. Two different processes produce the same change in a system. (a) A total of 15.00 J of heat transfer occurs into the system, while work takes out a total of 6.00 J. The change in internal energy is ΔU=Q−W=9.00 J. (b) Heat transfer removes 150.00 J from the system while work puts 159.00 J into it, producing an increase of 9.00 J in internal energy. If the system starts out in the same state in (a) and (b), it will end up in the same final state in either case—its final state is related to internal energy, not how that energy was acquired.

Strategy

In part 1, we must first find the net heat transfer and net work done from the given information. Then the first law of thermodynamics ($\Delta U = Q - W$) can be used to find the change in internal energy. In part (b), the net heat transfer and work done are given, so the equation can be used directly.

Solution for Part 1

The net heat transfer is the heat transfer into the system minus the heat transfer out of the system, or

$$Q = 40.00\ \text{J} - 25.00\ \text{J} = 15.00\ \text{J}.$$

Similarly, the total work is the work done by the system minus the work done on the system, or

$$W = 10.00\ \text{J} - 4.00\ \text{J} = 6.00\ \text{J}.$$

Thus the change in internal energy is given by the first law of thermodynamics:

$$\Delta U = Q - W = 15.00\ \text{J} - 6.00\ \text{J} = 9.00\ \text{J}.$$

We can also find the change in internal energy for each of the two steps. First, consider 40.00 J of heat transfer in and 10.00 J of work out, or $\Delta U_1 = Q_1 - W_1 = 40.00\ \text{J} - 10.00\ \text{J} = 30.00\ \text{J}$.

Now consider 25.00 J of heat transfer out and 4.00 J of work in, or

$$\Delta U_2 = Q_2 - W_2 = -25.00\ \text{J} - (-4.00\ \text{J}) = -21.00\ \text{J}.$$

The total change is the sum of these two steps, or $\Delta U = \Delta U_1 + \Delta U_2 = 30.00\ \text{J} + (-21.00\ \text{J}) = 9.00\ \text{J}$.

Discussion on Part 1

No matter whether you look at the overall process or break it into steps, the change in internal energy is the same.

Solution for Part 2

Here the net heat transfer and total work are given directly to be Q=–150.00 J and W=–159.00 J, so that

$$\Delta U = Q - W = -150.00 \text{ J} - (-159.00 \text{ J}) = 9.00 \text{ J}.$$

Discussion on Part 2

A very different process in part 2 produces the same 9.00-J change in internal energy as in part 1. Note that the change in the system in both parts is related to ΔU and not to the individual Qs or Ws involved. The system ends up in the *same* state in both parts. Parts 1 and 2 present two different paths for the system to follow between the same starting and ending points, and the change in internal energy for each is the same—it is independent of path.

Human Metabolism and the First Law of Thermodynamics

Human metabolism is the conversion of food into heat transfer, work, and stored fat. Metabolism is an interesting example of the first law of thermodynamics in action. We now take another look at these topics via the first law of thermodynamics. Considering the body as the system of interest, we can use the first law to examine heat transfer, doing work, and internal energy in activities ranging from sleep to heavy exercise. What are some of the major characteristics of heat transfer, doing work, and energy in the body? For one, body temperature is normally kept constant by heat transfer to the surroundings. This means Q is negative. Another fact is that the body usually does work on the outside world. This means W is positive. In such situations, then, the body loses internal energy, since $\Delta U = Q - W$ is negative.

Now consider the effects of eating. Eating increases the internal energy of the body by adding chemical potential energy (this is an unromantic view of a good steak). The body *metabolizes* all the food we consume. Basically, metabolism is an oxidation process in which the chemical potential energy of food is released. This implies that food input is in the form of work. Food energy is reported in a special unit, known as the Calorie. This energy is measured by burning food in a calorimeter, which is how the units are determined.

In chemistry and biochemistry, one calorie (spelled with a *lowercase* c) is defined as the energy (or heat transfer) required to raise the temperature of one gram of pure water by one degree Celsius. Nutritionists and weight-watchers tend to use the *dietary* calorie, which is frequently called a Calorie (spelled with a *capital* C). One food Calorie is the energy needed to raise the temperature of one *kilogram* of water by one degree Celsius. This means that one dietary Calorie is equal to one kilocalorie for the chemist, and one must be careful to avoid confusion between the two.

Again, consider the internal energy the body has lost. There are three places this internal energy can go—to heat transfer, to doing work, and to stored fat (a tiny fraction also goes to cell repair and growth). Heat transfer and doing work take internal energy out of the body, and food puts it back. If you eat just the right amount of food, then your average internal energy remains constant. Whatever you lose to heat transfer and doing work is replaced by food, so that, in the long run, ΔU=0. If you overeat repeatedly, then ΔU is always positive, and your body stores this extra internal energy as fat. The reverse is true if you eat too little. If ΔU is negative for a few days, then the body metabolizes its own fat to maintain

body temperature and do work that takes energy from the body. This process is how dieting produces weight loss.

Life is not always this simple, as any dieter knows. The body stores fat or metabolizes it only if energy intake changes for a period of several days. Once you have been on a major diet, the next one is less successful because your body alters the way it responds to low energy intake. Your basal metabolic rate (BMR) is the rate at which food is converted into heat transfer and work done while the body is at complete rest. The body adjusts its basal metabolic rate to partially compensate for over-eating or under-eating. The body will decrease the metabolic rate rather than eliminate its own fat to replace lost food intake. You will chill more easily and feel less energetic as a result of the lower metabolic rate, and you will not lose weight as fast as before. Exercise helps to lose weight, because it produces both heat transfer from your body and work, and raises your metabolic rate even when you are at rest. Weight loss is also aided by the quite low efficiency of the body in converting internal energy to work, so that the loss of internal energy resulting from doing work is much greater than the work done.It should be noted, however, that living systems are not in thermalequilibrium.

The body provides us with an excellent indication that many thermodynamic processes are *irreversible*. An irreversible process can go in one direction but not the reverse, under a given set of conditions. For example, although body fat can be converted to do work and produce heat transfer, work done on the body and heat transfer into it cannot be converted to body fat. Otherwise, we could skip lunch by sunning ourselves or by walking down stairs. Another example of an irreversible thermodynamic process is photosynthesis. This process is the intake of one form of energy—light—by plants and its conversion to chemical potential energy. Both applications of the first law of thermodynamics are illustrated in Figure 4. One great advantage of conservation laws such as the first law of thermodynamics is that they accurately describe the beginning and ending points of complex processes, such as metabolism and photosynthesis, without regard to the complications in between. Table 1 presents a summary of terms relevant to the first law of thermodynamics.

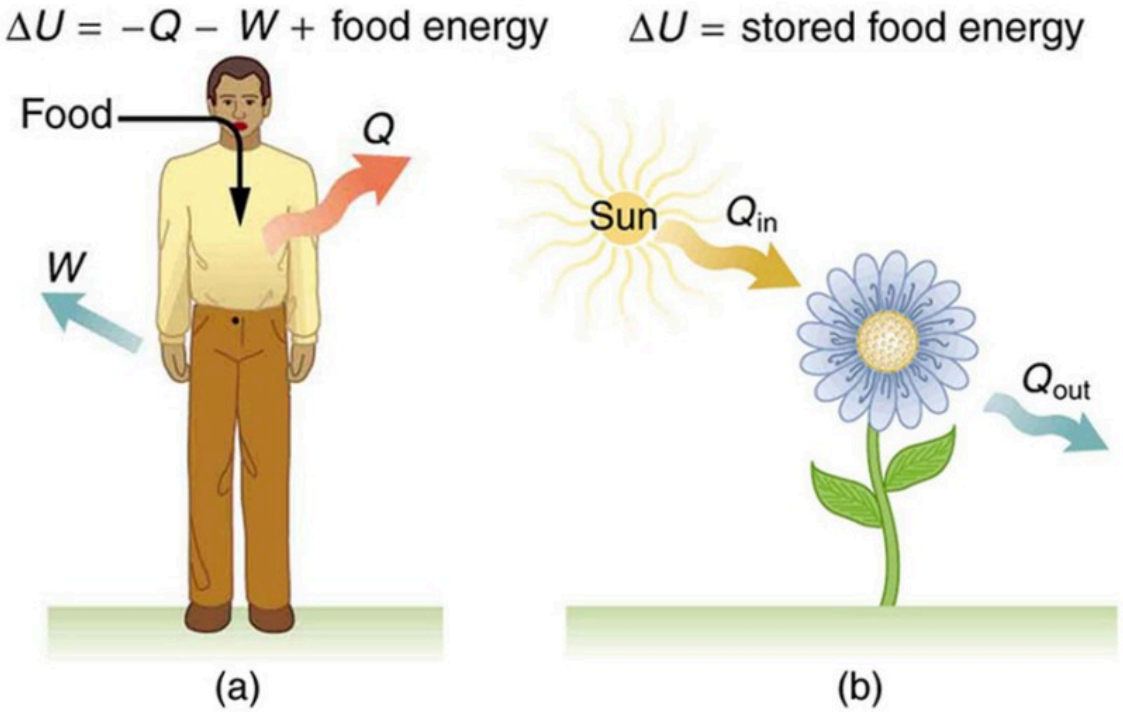


Figure 4. (a) The first law of thermodynamics applied to metabolism. Heat transferred out of the body (Q) and work done by the body (W) remove internal energy, while food intake replaces it. (Food intake may be considered as work done on the body.) (b) Plants convert part of the radiant heat transfer in sunlight to stored chemical energy, a process called photosynthesis.

Table 1. Summary of Terms for the First Law of Thermodynamics, $\Delta U = Q - W$

Term	Definition
U	Internal energy—the sum of the kinetic and potential energies of a system's atoms and molecules. Can be divided into many subcategories, such as thermal and chemical energy. Depends only on the state of a system (such as its P, V, and T), not on how the energy entered the system. Change in internal energy is path independent.
Q	Heat—energy transferred because of a temperature difference. Characterized by random molecular motion. Highly dependent on path. Q entering a system is positive.
W	Work—energy transferred by a force moving through a distance. An organized, orderly process. Path dependent. W done by a system (either against an external force or to increase the volume of the system) is positive.

Section Summary

- The first law of thermodynamics is given as $\Delta U = Q - W$, where ΔU is the change in internal energy of a system, Q is the net heat transfer (the sum of all heat transfer into and out of the system), and W is the net work done (the sum of all work done on or by the system).
- Both Q and W are energy in transit; only ΔU represents an independent quantity capable of being stored.
- The internal energy U of a system depends only on the state of the system and not how it

reached that state.

- Metabolism of living organisms, and photosynthesis of plants, are specialized types of heat transfer, doing work, and internal energy of systems.

Conceptual Questions

1. Describe the photo of the tea kettle at the beginning of this section in terms of heat transfer, work done, and internal energy. How is heat being transferred? What is the work done and what is doing it? How does the kettle maintain its internal energy?
2. The first law of thermodynamics and the conservation of energy, as discussed in Conservation of Energy, are clearly related. How do they differ in the types of energy considered?
3. Heat transfer Q and work done W are always energy in transit, whereas internal energy U is energy stored in a system. Give an example of each type of energy, and state specifically how it is either in transit or resides in a system.
4. How do heat transfer and internal energy differ? In particular, which can be stored as such in a system and which cannot?
5. If you run down some stairs and stop, what happens to your kinetic energy and your initial gravitational potential energy?
6. Give an explanation of how food energy (calories) can be viewed as molecular potential energy (consistent with the atomic and molecular definition of internal energy).
7. Identify the type of energy transferred to your body in each of the following as either internal energy, heat transfer, or doing work: (a) basking in sunlight; (b) eating food; (c) riding an elevator to a higher floor.

Problems & Exercises

1. What is the change in internal energy of a car if you put 12.0 gal of gasoline into its tank? The energy content of gasoline is 1.3×10^8 J/gal. All other factors, such as the car's temperature, are constant.
2. How much heat transfer occurs from a system, if its internal energy decreased by 150 J while it was doing 30.0 J of work?
3. A system does 1.80×10^8 J of work while 7.50×10^8 J of heat transfer occurs to the environment. What is the change in internal energy of the system assuming no other changes (such as in temperature or by the addition of fuel)?
4. What is the change in internal energy of a system which does 4.50×10^5 J of work while 3.00×10^6 J of heat transfer occurs into the system, and 8.00×10^6 J of heat transfer occurs to the environment?
5. Suppose a woman does 500 J of work and 9500 J of heat transfer occurs into the environment in the process. (a) What is the decrease in her internal energy, assuming no change in temperature or consumption of food? (That is, there is no other energy transfer.) (b) What is her efficiency?
6. (a) How much food energy will a man metabolize in the process of doing 35.0 kJ of work with an

efficiency of 5.00%? (b) How much heat transfer occurs to the environment to keep his temperature constant?

7. (a) What is the average metabolic rate in watts of a man who metabolizes 10,500 kJ of food energy in one day? (b) What is the maximum amount of work in joules he can do without breaking down fat, assuming a maximum efficiency of 20.0%? (c) Compare his work output with the daily output of a 187-W (0.250-horsepower) motor.
8. (a) How long will the energy in a 1470-kJ (350-kcal) cup of yogurt last in a woman doing work at the rate of 150 W with an efficiency of 20.0% (such as in leisurely climbing stairs)? (b) Does the time found in part (a) imply that it is easy to consume more food energy than you can reasonably expect to work off with exercise?
9. (a) A woman climbing the Washington Monument metabolizes 6.00×10^2 kJ of food energy. If her efficiency is 18.0%, how much heat transfer occurs to the environment to keep her temperature constant? (b) Discuss the amount of heat transfer found in (a). Is it consistent with the fact that you quickly warm up when exercising?

Glossary

first law of thermodynamics: states that the change in internal energy of a system equals the net heat transfer into the system minus the net work done by the system

internal energy: the sum of the kinetic and potential energies of a system's atoms and molecules

human metabolism: conversion of food into heat transfer, work, and stored fat

Selected Solutions to Problems & Exercises

1. 1.6×10^9 J

3. -9.30×10^8 J

5. (a) -1.0×10^4 J , or −2.39 kcal; (b) 5.00%

7. (a) 122 W; (b) 2.10×10^6 J; (c) Work done by the motor is 1.61×10^7 J; thus the motor produces 7.67 times the work done by the man

9. (a) 492 kJ; (b) This amount of heat is consistent with the fact that you warm quickly when exercising. Since the body is inefficient, the excess heat produced must be dissipated through sweating, breathing, etc.

The First Law of Thermodynamics and Some Simple Processes

Lumen Learning

Learning Objectives

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

- Describe the processes of a simple heat engine.
- Explain the differences among the simple thermodynamic processes—isobaric, isochoric, isothermal, and adiabatic.
- Calculate total work done in a cyclical thermodynamic process.

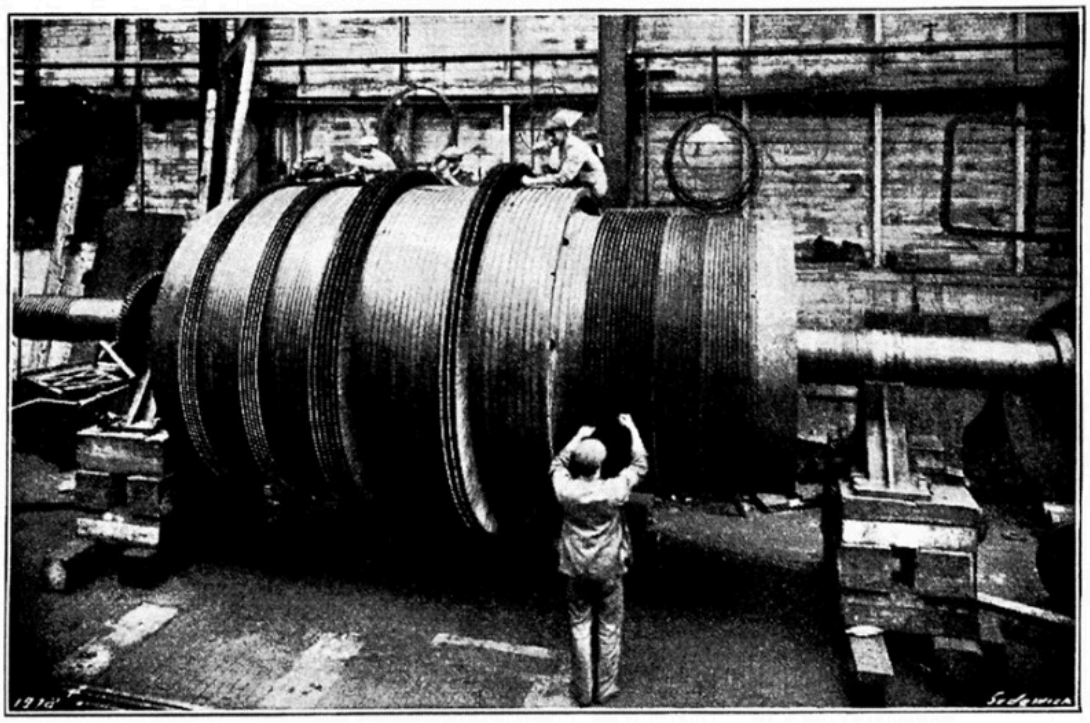


Figure 1. Beginning with the Industrial Revolution, humans have harnessed power through the use of the first law of thermodynamics, before we even understood it completely. This photo, of a steam engine at the Turbinia Works, dates from 1911, a mere 61 years after the first explicit statement of the first law of thermodynamics by Rudolph Clausius. (credit: public domain; author unknown)

One of the most important things we can do with heat transfer is to use it to do work for us. Such a device is called a *heat engine*. Car engines and steam turbines that generate electricity are examples of heat engines. Figure 2 shows schematically how the first law of thermodynamics applies to the typical heat engine.

It is impossible to devise a system where $Q_{out} = 0$, that is, in which no heat transfer occurs to the environment.

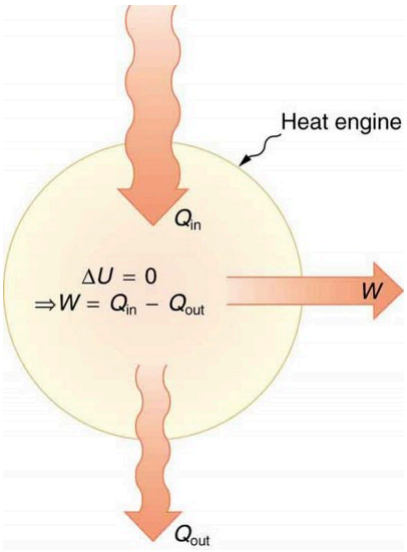
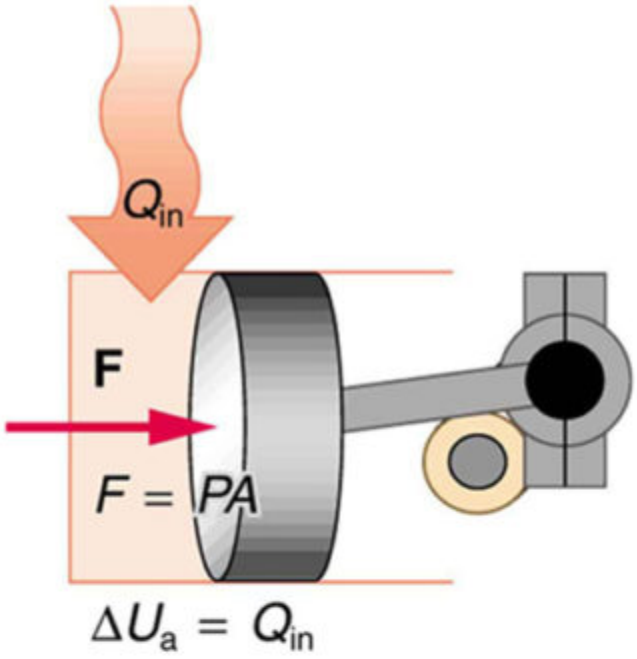
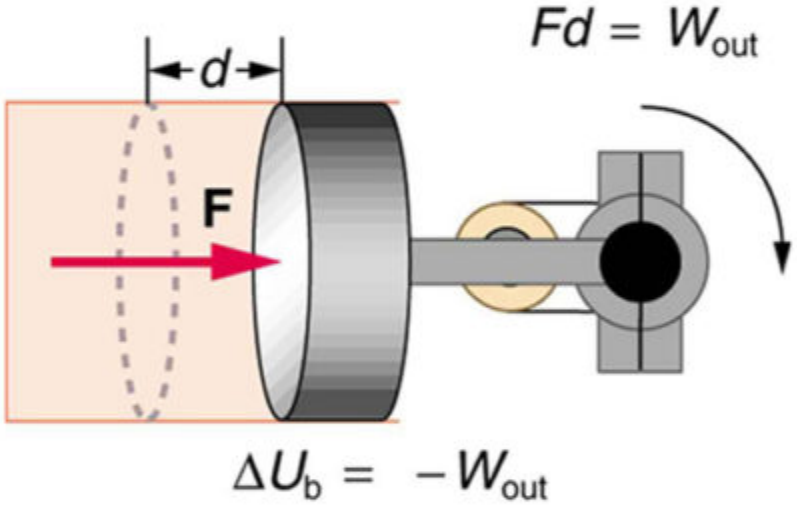


Figure 2. Schematic representation of a heat engine, governed, of course, by the first law of thermodynamics.



(a)



(b)

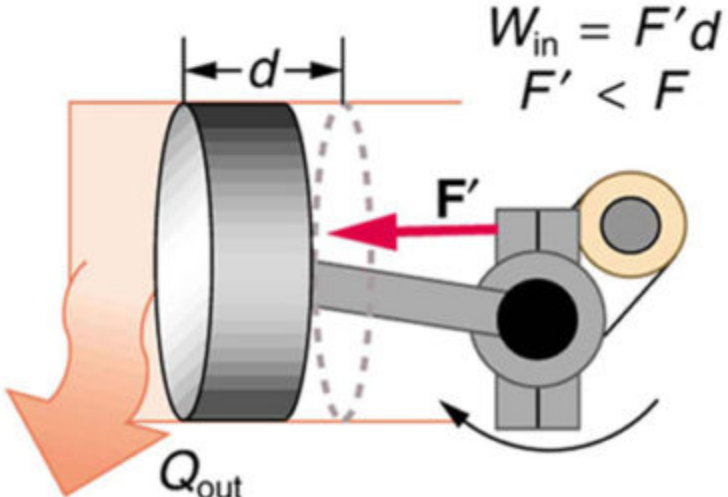


Figure 3. (a) Heat transfer to the gas in a cylinder increases the internal energy of the gas, creating higher pressure and temperature. (b) The force exerted on the movable cylinder does work as the gas expands. Gas pressure and temperature decrease when it expands, indicating that the gas's internal energy has been decreased by doing work. (c) Heat transfer to the environment further reduces pressure in the gas so that the piston can be more easily returned to its starting position.

The illustrations above show one of the ways in which heat transfer does work. Fuel combustion produces heat transfer to a gas in a cylinder, increasing the pressure of the gas and thereby the force it exerts on a movable piston. The gas does work on the outside world, as this force moves the piston through some distance. Heat transfer to the gas cylinder results in work being done. To repeat this process, the piston needs to be returned to its starting point. Heat transfer now occurs from the gas to the surroundings so that its pressure decreases, and a force is exerted by the surroundings to push the piston back through some distance. Variations of this process are employed daily in hundreds of millions of heat engines. We will examine heat engines in detail in the next section. In this section, we consider some of the simpler underlying processes on which heat engines are based.

PV Diagrams and their Relationship to Work Done on or by a Gas

A process by which a gas does work on a piston at constant pressure is called an *isobaric process*. Since the pressure is constant, the force exerted is constant and the work done is given as $P\Delta V$.

$W = Fd$. See the symbols as shown in Figure 4. Now $F = PA$, and so $W = PAd$.

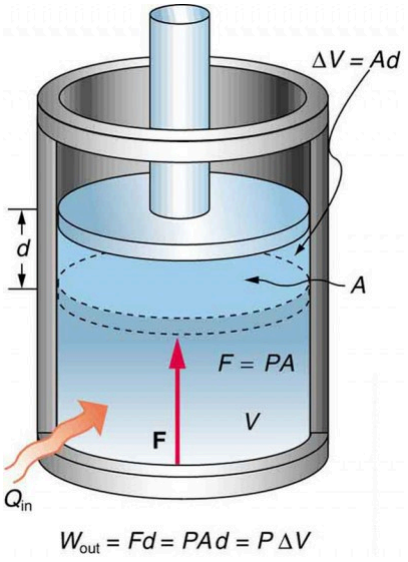


Figure 4. An isobaric expansion of a gas requires heat transfer to keep the pressure constant. Since pressure is constant, the work done is PΔV.

Because the volume of a cylinder is its cross-sectional area A times its length d, we see that $Ad = \Delta V$, the change in volume; thus, $W = P\Delta V$ (isobaric process).

Note that if ΔV is positive, then W is positive, meaning that work is done *by* the gas on the outside world.

(Note that the pressure involved in this work that we've called P is the pressure of the gas *inside* the tank. If we call the pressure outside the tank P_{ext}, an expanding gas would be working *against* the external pressure; the work done would therefore be $W = -P_{ext}\Delta V$ (isobaric process). Many texts use this definition of work, and not the definition based on internal pressure, as the basis of the First Law of Thermodynamics. This definition reverses the sign conventions for work, and results in a statement of the first law that becomes $\Delta U = Q + W$.)

It is not surprising that $W = P\Delta V$, since we have already noted in our treatment of fluids that pressure is a type of potential energy per unit volume and that pressure in fact has units of energy divided by volume. We also noted in our discussion of the ideal gas law that PV has units of energy. In this case, some of the energy associated with pressure becomes work.

Figure 5 shows a graph of pressure versus volume (that is, a *PV* diagram for an isobaric process. You can see in the figure that the work done is the area under the graph. This property of *PV* diagrams is very useful and broadly applicable: *the work done on or by a system in going from one state to another equals the area under the curve on a* PV *diagram*.

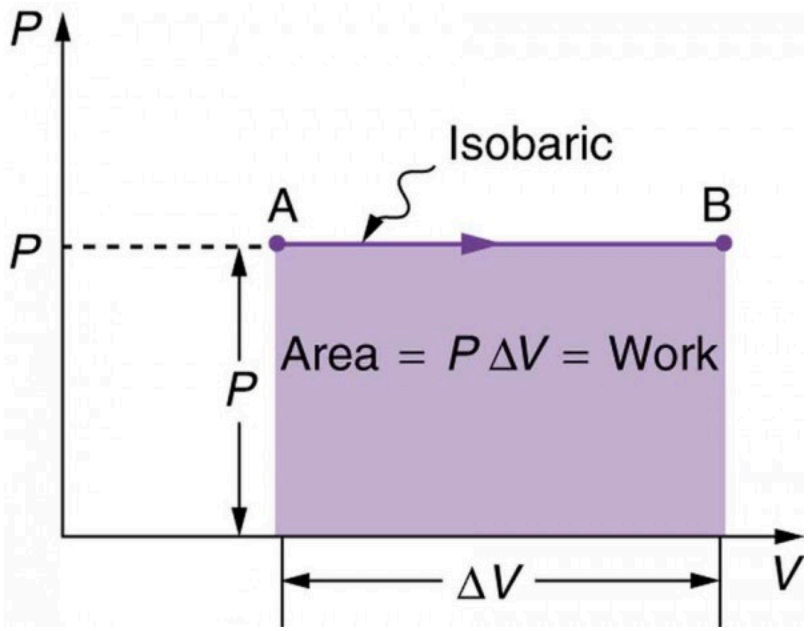


Figure 5. A graph of pressure versus volume for a constant-pressure, or isobaric, process, such as the one shown in Figure 4. The area under the curve equals the work done by the gas, since W = PΔV.

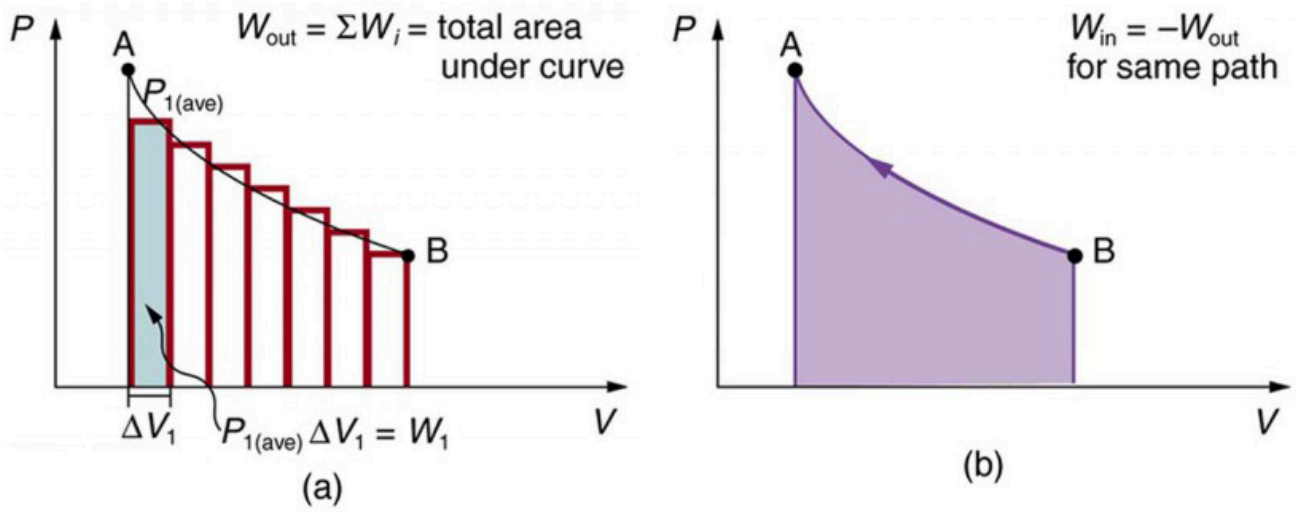


Figure 6. (a) A PV diagram in which pressure varies as well as volume. The work done for each interval is its average pressure times the change in volume, or the area under the curve over that interval. Thus the total area under the curve equals the total work done. (b) Work must be done on the system to follow the reverse path. This is interpreted as a negative area under the curve.

We can see where this leads by considering Figure 6a, which shows a more general process in which both pressure and volume change. The area under the curve is closely approximated by dividing it into strips, each having an average constant pressure $P_{i(ave)}$. The work done is $W_i = P_{i(ave)}\Delta V_i$ for each strip,

and the total work done is the sum of the W_i. Thus the total work done is the total area under the curve. If the path is reversed, as in Figure 6b, then work is done on the system. The area under the curve in that case is negative, because ΔV is negative.

PV diagrams clearly illustrate that *the work done depends on the path taken and not just the endpoints*. This path dependence is seen in Figure 7a, where more work is done in going from A to C by the path via point B than by the path via point D. The vertical paths, where volume is constant, are called *isochoric* processes. Since volume is constant, $\Delta V = 0$, and no work is done in an isochoric process. Now, if the system follows the cyclical path ABCDA, as in Figure 7b, then the total work done is the area inside the loop. The negative area below path CD subtracts, leaving only the area inside the rectangle. In fact, the work done in any cyclical process (one that returns to its starting point) is the area inside the loop it forms on a *PV* diagram, as Figure 7c illustrates for a general cyclical process. Note that the loop must be traversed in the clockwise direction for work to be positive—that is, for there to be a net work output.

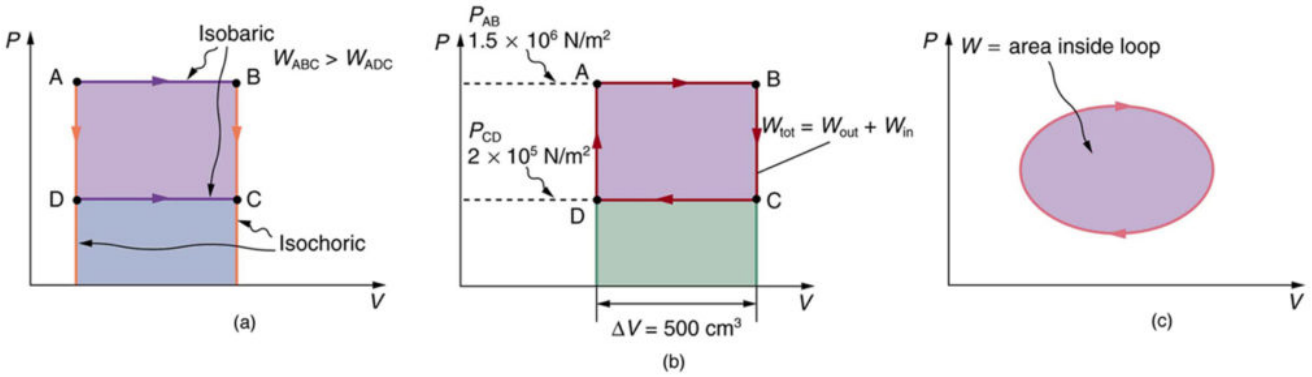


Figure 7. (a) The work done in going from A to C depends on path. The work is greater for the path ABC than for the path ADC, because the former is at higher pressure. In both cases, the work done is the area under the path. This area is greater for path ABC. (b) The total work done in the cyclical process ABCDA is the area inside the loop, since the negative area below CD subtracts out, leaving just the area inside the rectangle. (The values given for the pressures and the change in volume are intended for use in the example below.) (c) The area inside any closed loop is the work done in the cyclical process. If the loop is traversed in a clockwise direction, W is positive—it is work done on the outside environment. If the loop is traveled in a counter-clockwise direction, W is negative—it is work that is done to the system.

Example 1. Total Work Done in a Cyclical Process Equals the Area Inside the Closed Loop on a *PV* Diagram

Calculate the total work done in the cyclical process ABCDA shown in Figure 7b by the following two methods to verify that work equals the area inside the closed loop on the *PV* diagram. (Take the data in the figure to be precise to three significant figures.)

1. Calculate the work done along each segment of the path and add these values to get the total work.
2. Calculate the area inside the rectangle ABCDA.

Strategy

To find the work along any path on a *PV* diagram, you use the fact that work is pressure times change in volume, or $W = P\Delta V$. So in part 1, this value is calculated for each leg of the path around the closed loop.

Solution for Part 1

The work along path AB is

$$\begin{aligned} W_{\text{AB}} &= P_{\text{AB}}\Delta V_{\text{AB}} \\ &= \left(1.50 \times 10^6 \text{ N/m}^2\right)\left(5.00 \times 10^{-4} \text{ m}^3\right) = 750 \text{ J} \end{aligned}$$

Since the path BC is isochoric, $\Delta V_{\text{BC}}=0$, and so $W_{\text{BC}}=0$. The work along path CD is negative, since ΔV_{CD} is negative (the volume decreases). The work is

$$\begin{aligned} W_{\text{CD}} &= P_{\text{CD}}\Delta V_{\text{CD}} \\ &= \left(2.00 \times 10^5 \text{ N/m}^2\right)\left(-5.00 \times 10^{-4} \text{ m}^3\right) = -100 \text{ J} \end{aligned}$$

Again, since the path DA is isochoric, $\Delta V_{\text{DA}}=0$, and so $W_{\text{DA}}=0$. Now the total work is

$$\begin{aligned} W &= W_{\text{AB}} + W_{\text{BC}} + W_{\text{CD}} + W_{\text{DA}} \\ &= 750 \text{ J} + 0 + (-100 \text{ J}) + 0 = 650 \text{ J} \end{aligned}$$

Solution for Part 2

The area inside the rectangle is its height times its width, or

$$\begin{aligned} \text{area} &= \left(P_{\text{AB}} - P_{\text{CD}}\right)\Delta V \\ &= \left[\left(1.50 \times 10^6 \text{ N/m}^2\right) - \left(2.00 \times 10^5 \text{ N/m}^2\right)\right]\left(5.00 \times 10^{-4} \text{ m}^3\right) \\ &= 650 \text{ J} \end{aligned}$$

Thus, area = 650 J = *W*.

Discussion

The result, as anticipated, is that the area inside the closed loop equals the work done. The area is often easier to calculate than is the work done along each path. It is also convenient to visualize the area inside different curves on *PV* diagrams in order to see which processes might produce the most work. Recall that work can be done to the system, or by the system, depending on the sign of *W*. A positive *W* is work that is done by the system on the outside environment; a negative *W* represents work done by the environment on the system.

Figure 8a shows two other important processes on a *PV* diagram. For comparison, both are shown starting from the same point A. The upper curve ending at point B is an *isothermal* process—that is, one in which temperature is kept constant. If the gas behaves like an ideal gas, as is often the case, and if no phase change occurs, then $PV = nRT$. Since *T* is constant, *PV* is a constant for an isothermal process. We ordinarily expect the temperature of a gas to decrease as it expands, and so we correctly suspect that heat transfer must occur from the surroundings to the gas to keep the temperature constant during an isothermal expansion. To show this more rigorously for the special case of a monatomic ideal gas, we note that the average kinetic energy of an atom in such a gas is given by

$$\tfrac{1}{2}m\overline{v}^2 = \tfrac{3}{2}kT$$

.

The kinetic energy of the atoms in a monatomic ideal gas is its only form of internal energy, and so its total internal energy *U* is

$$U = N\tfrac{1}{2}m\overline{v}^2 = \tfrac{3}{2}NkT$$

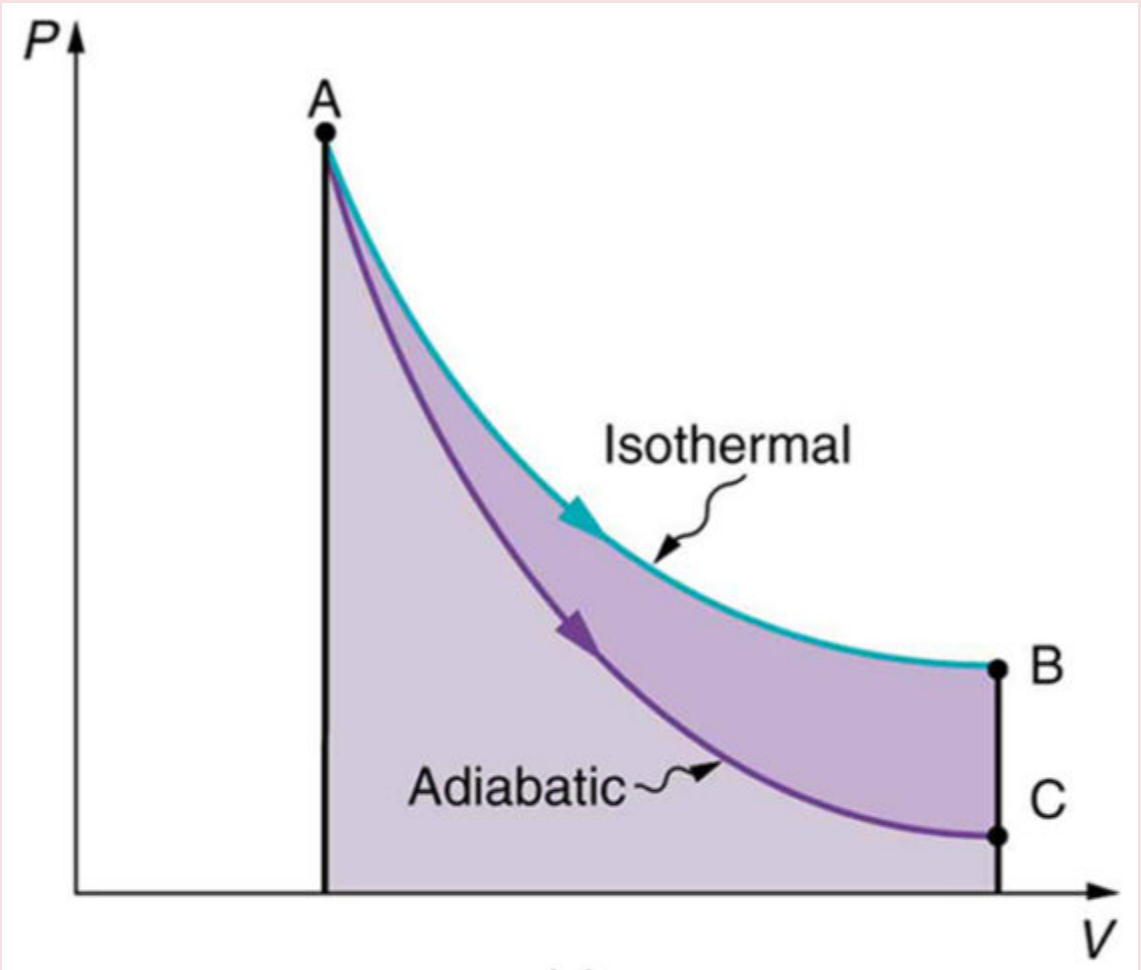
, (monatomic ideal gas), where *N* is the number of atoms in the gas. This relationship means that the internal energy of an ideal monatomic gas is constant during an isothermal process—that is, ΔU=0. If the internal energy does not change, then the net heat transfer into the gas must equal the net work done by the gas. That is, because $\Delta U = Q - W = 0$ here, $Q = W$. We must have just enough heat transfer to replace the work done. An isothermal process is inherently slow, because heat transfer occurs continuously to keep the gas temperature constant at all times and must be allowed to spread through the gas so that there are no hot or cold regions.

Also shown in Figure 8a is a curve AC for an *adiabatic* process, defined to be one in which there is no heat transfer—that is, $Q = 0$. Processes that are nearly adiabatic can be achieved either by using very effective insulation or by performing the process so fast that there is little time for heat transfer. Temperature must decrease during an adiabatic process, since work is done at the expense of internal energy:

$$U = \tfrac{3}{2}NkT$$

.

(You might have noted that a gas released into atmospheric pressure from a pressurized cylinder is substantially colder than the gas in the cylinder.) In fact, because $Q = 0$, $\Delta U = -W$ for an adiabatic process. Lower temperature results in lower pressure along the way, so that curve AC is lower than curve AB, and less work is done. If the path ABCA could be followed by cooling the gas from B to C at constant volume (isochorically), Figure 8b, there would be a net work output.


P
A
Isothermal
B
Adiabatic
C
V

(a)

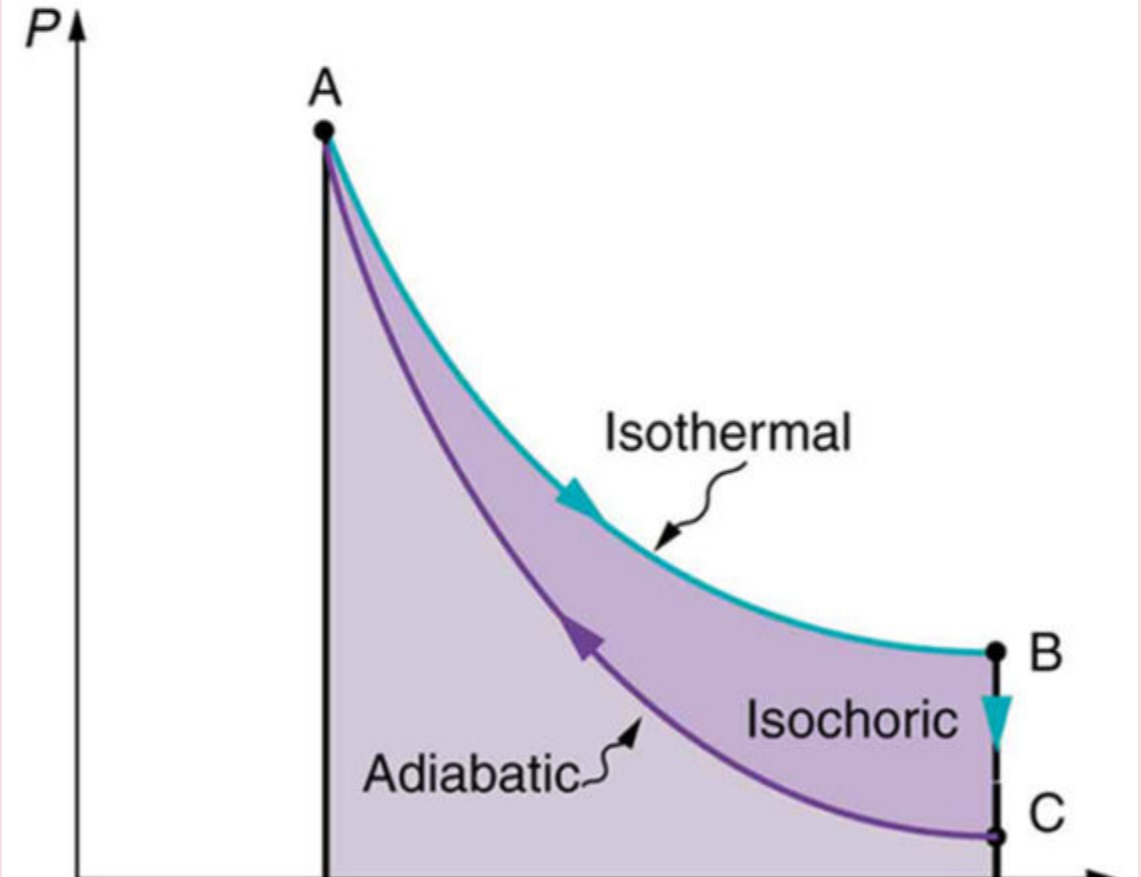

P
A
Isothermal
B
Isochoric
Adiabatic
C

Figure 8. (a) The upper curve is an isothermal process ($\Delta T = 0$), whereas the lower curve is an adiabatic process ($Q = 0$). Both start from the same point A, but the isothermal process does more work than the adiabatic because heat transfer into the gas takes place to keep its temperature constant. This keeps the pressure higher all along the isothermal path than along the adiabatic path, producing more work. The adiabatic path thus ends up with a lower pressure and temperature at point C, even though the final volume is the same as for the isothermal process. (b) The cycle ABCA produces a net work output.

Reversible Processes

Both isothermal and adiabatic processes such as shown in Figure 8 are reversible in principle. A *reversible process* is one in which both the system and its environment can return to exactly the states they were in by following the reverse path. The reverse isothermal and adiabatic paths are BA and CA, respectively. Real macroscopic processes are never exactly reversible. In the previous examples, our system is a gas (like that in Figure 4), and its environment is the piston, cylinder, and the rest of the universe. If there are any energy-dissipating mechanisms, such as friction or turbulence, then heat transfer to the environment occurs for either direction of the piston. So, for example, if the path BA is followed and there is friction, then the gas will be returned to its original state but the environment will not—it will have been heated in both directions. Reversibility requires the direction of heat transfer to reverse for the reverse path. Since dissipative mechanisms cannot be completely eliminated, real processes cannot be reversible.

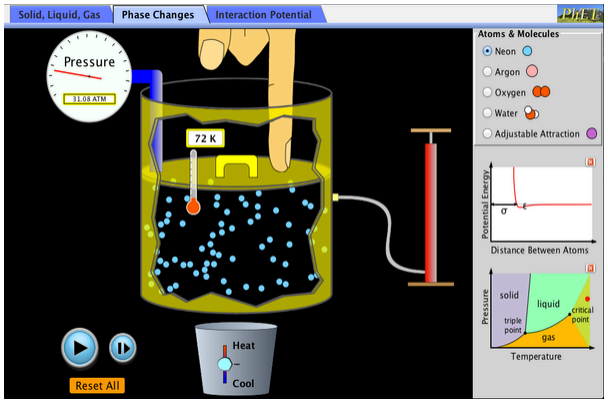
There must be reasons that real macroscopic processes cannot be reversible. We can imagine them going in reverse. For example, heat transfer occurs spontaneously from hot to cold and never spontaneously the reverse. Yet it would not violate the first law of thermodynamics for this to happen. In fact, all spontaneous processes, such as bubbles bursting, never go in reverse. There is a second thermodynamic law that forbids them from going in reverse. When we study this law, we will learn something about nature and also find that such a law limits the efficiency of heat engines. We will find that heat engines with the greatest possible theoretical efficiency would have to use reversible processes, and even they cannot convert all heat transfer into doing work. Table 1 summarizes the simpler thermodynamic processes and their definitions.

Table 1. Summary of Simple Thermodynamic Processes

Isobaric	Constant pressure	$W = P\Delta V$
Isochoric	Constant volume	$W = 0$
Isothermal	Constant temperature	$Q = W$
Adiabatic	No heat transfer	$Q = 0$

PhET Explorations: States of Matter

Watch different types of molecules form a solid, liquid, or gas. Add or remove heat and watch the phase change. Change the temperature or volume of a container and see a pressure-temperature diagram respond in real time. Relate the interaction potential to the forces between molecules.



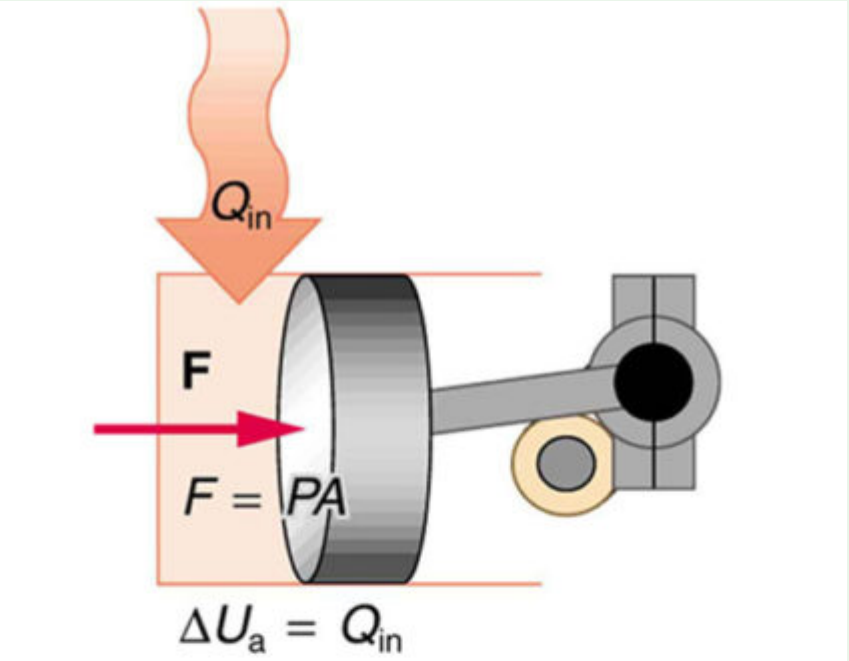
Click to run the simulation.

Section Summary

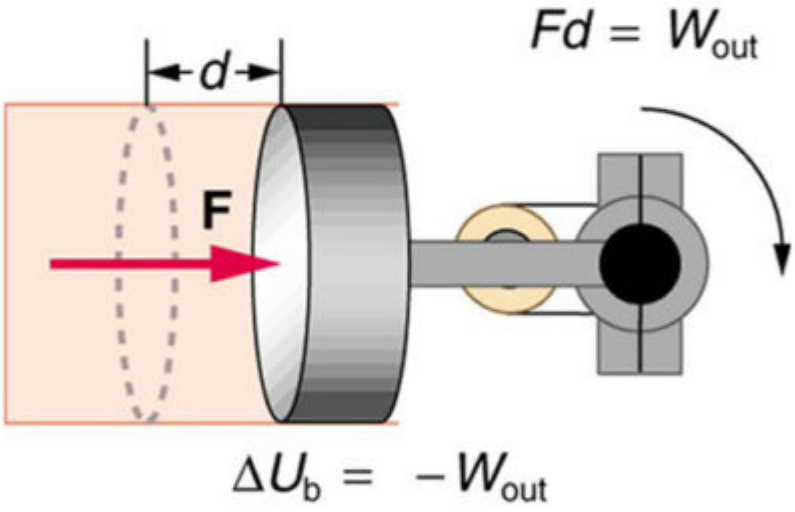
- One of the important implications of the first law of thermodynamics is that machines can be harnessed to do work that humans previously did by hand or by external energy supplies such as running water or the heat of the Sun. A machine that uses heat transfer to do work is known as a heat engine.
- There are several simple processes, used by heat engines, that flow from the first law of thermodynamics. Among them are the isobaric, isochoric, isothermal and adiabatic processes.
- These processes differ from one another based on how they affect pressure, volume, temperature, and heat transfer.
- If the work done is performed on the outside environment, work (W) will be a positive value. If the work done is done to the heat engine system, work (W) will be a negative value.
- Some thermodynamic processes, including isothermal and adiabatic processes, are reversible in theory; that is, both the thermodynamic system and the environment can be returned to their initial states. However, because of loss of energy owing to the second law of thermodynamics, complete reversibility does not work in practice.

Conceptual Questions

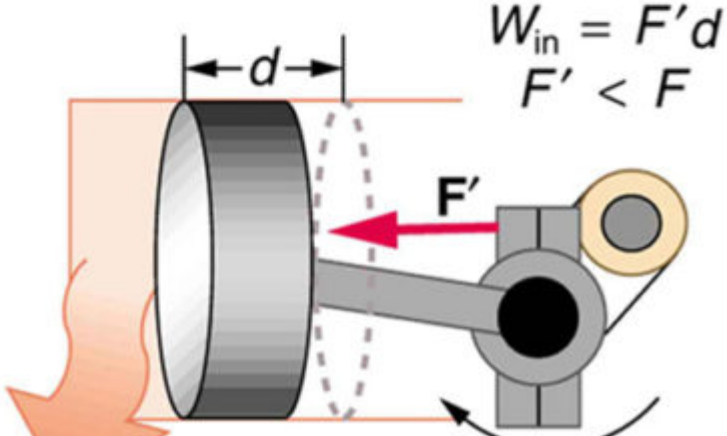
1. A great deal of effort, time, and money has been spent in the quest for the so-called perpetual-motion machine, which is defined as a hypothetical machine that operates or produces useful work indefinitely and/or a hypothetical machine that produces more work or energy than it consumes. Explain, in terms of heat engines and the first law of thermodynamics, why or why not such a machine is likely to be constructed.
2. One method of converting heat transfer into doing work is for heat transfer into a gas to take place, which expands, doing work on a piston, as shown in the figure below. (a) Is the heat transfer converted directly to work in an isobaric process, or does it go through another form first? Explain your answer. (b) What about in an isothermal process? (c) What about in an adiabatic process (where heat transfer occurred prior to the adiabatic process)?



(a)



(b)



3. Would the previous question make any sense for an isochoric process? Explain your answer.
4. We ordinarily say that $\Delta U = 0$ for an isothermal process. Does this assume no phase change takes place? Explain your answer.
5. The temperature of a rapidly expanding gas decreases. Explain why in terms of the first law of thermodynamics. (Hint: Consider whether the gas does work and whether heat transfer occurs rapidly into the gas through conduction.)
6. Which cyclical process represented by the two closed loops, ABCFA and ABDEA, on the *PV* diagram in the figure below produces the greatest net work? Is that process also the one with the smallest work input required to return it to point A? Explain your responses.

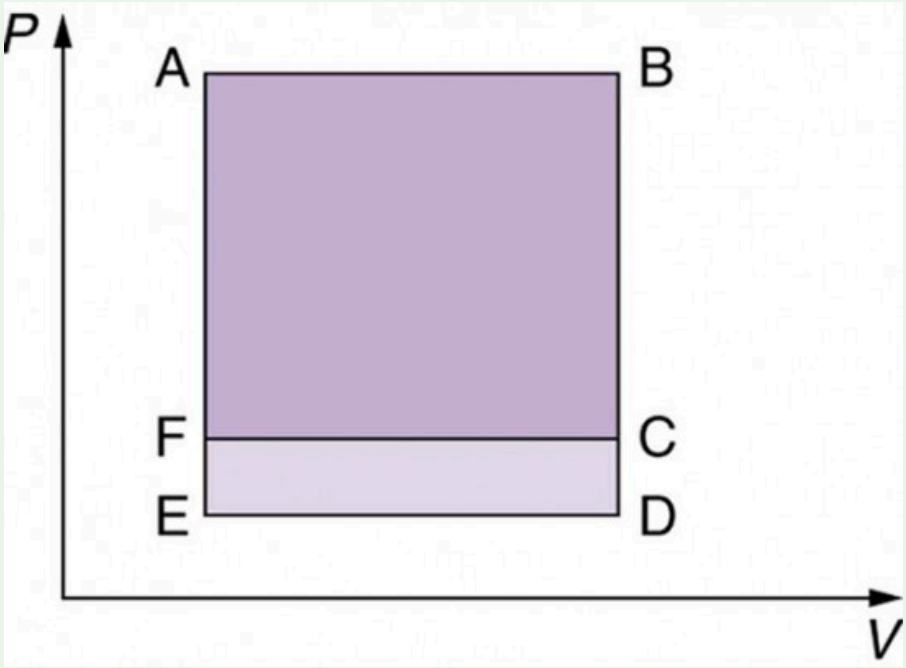


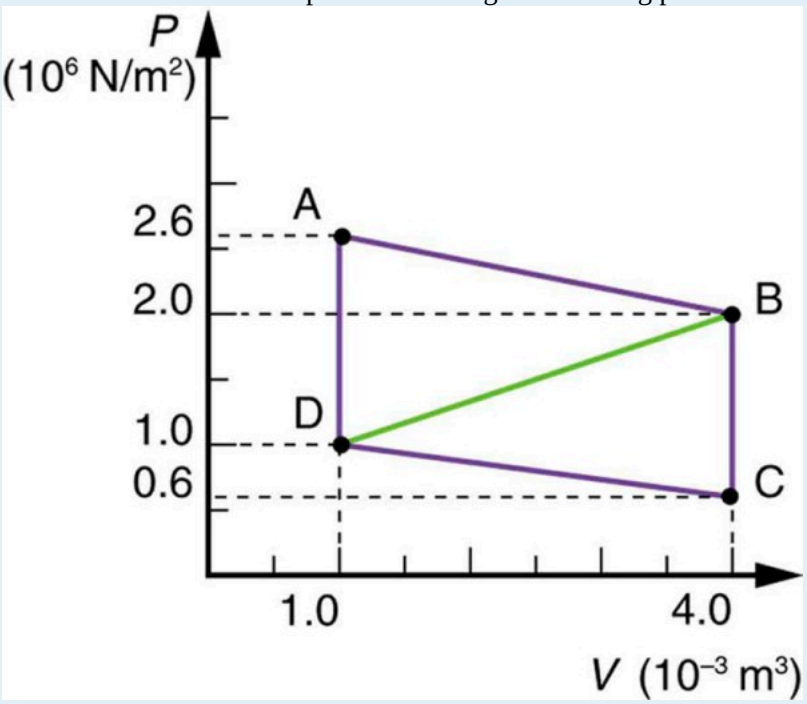
Figure 11. The two cyclical processes shown on this PV diagram start with and return the system to the conditions at point A, but they follow different paths and produce different amounts of work.

7. A real process may be nearly adiabatic if it occurs over a very short time. How does the short time span help the process to be adiabatic?
8. It is unlikely that a process can be isothermal unless it is a very slow process. Explain why. Is the same true for isobaric and isochoric processes? Explain your answer.

Problems & Exercises

1. A car tire contains 0.0380 m^3 of air at a pressure of 2.20×10^5 N/m^2 (about 32 psi). How much more internal energy does this gas have than the same volume has at zero gauge pressure (which is equivalent to normal atmospheric pressure)?
2. A helium-filled toy balloon has a gauge pressure of 0.200 atm and a volume of 10.0 L. How much greater is the internal energy of the helium in the balloon than it would be at zero gauge pressure?

3. Steam to drive an old-fashioned steam locomotive is supplied at a constant gauge pressure of 1.75×10^6 N/m^2 (about 250 psi) to a piston with a 0.200-m radius. (a) By calculating $P\Delta V$, find the work done by the steam when the piston moves 0.800 m. Note that this is the net work output, since gauge pressure is used. (b) Now find the amount of work by calculating the force exerted times the distance traveled. Is the answer the same as in part (a)?
4. A hand-driven tire pump has a piston with a 2.50-cm diameter and a maximum stroke of 30.0 cm. (a) How much work do you do in one stroke if the average gauge pressure is 2.40×10^5 N/m^2 (about 35 psi)? (b) What average force do you exert on the piston, neglecting friction and gravitational force?
5. Calculate the net work output of a heat engine following path ABCDA in the figure below.



6. What is the net work output of a heat engine that follows path ABDA in the figure above, with a straight line from B to D? Why is the work output less than for path ABCDA? Explicitly show how you follow the steps in the Problem-Solving Strategies for Thermodynamics.
7. **Unreasonable Results.** What is wrong with the claim that a cyclical heat engine does 4.00 kJ of work on an input of 24.0 kJ of heat transfer while 16.0 kJ of heat transfers to the environment?
8. (a) A cyclical heat engine, operating between temperatures of 450ºC and 150ºC produces 4.00 MJ of work on a heat transfer of 5.00 MJ into the engine. How much heat transfer occurs to the environment? (b) What is unreasonable about the engine? (c) Which premise is unreasonable?
9. **Construct Your Own Problem.** Consider a car's gasoline engine. Construct a problem in which you calculate the maximum efficiency this engine can have. Among the things to consider are the effective hot and cold reservoir temperatures. Compare your calculated efficiency with the actual efficiency of car engines.
10. **Construct Your Own Problem.** Consider a car trip into the mountains. Construct a problem in which you calculate the overall efficiency of the car for the trip as a ratio of kinetic and potential energy gained to fuel consumed. Compare this efficiency to the thermodynamic efficiency quoted

for gasoline engines and discuss why the thermodynamic efficiency is so much greater. Among the factors to be considered are the gain in altitude and speed, the mass of the car, the distance traveled, and typical fuel economy.

Glossary

heat engine: a machine that uses heat transfer to do work

isobaric process: constant-pressure process in which a gas does work

isochoric process: a constant-volume process

isothermal process: a constant-temperature process

adiabatic process: a process in which no heat transfer takes place

reversible process: a process in which both the heat engine system and the external environment theoretically can be returned to their original states

Selected Solutions to Problems & Exercises

1. 6.77×10^3 J

3. (a) $W = P\Delta V = 1.76 \times 10^5$ J; (b) $W = Fd = 1.76 \times 10^5$ J. Yes, the answer is the same.

5. $W = 4.5 \times 10^3$ J

7. W is not equal to the difference between the heat input and the heat output.

Introduction to the Second Law of Thermodynamics: Heat Engines and Their Efficiency

Lumen Learning

Learning Objectives

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

- State the expressions of the second law of thermodynamics.
- Calculate the efficiency and carbon dioxide emission of a coal-fired electricity plant, using second law characteristics.
- Describe and define the Otto cycle.

The second law of thermodynamics deals with the direction taken by spontaneous processes. Many processes occur spontaneously in one direction only—that is, they are irreversible, under a given set of conditions. Although irreversibility is seen in day-to-day life—a broken glass does not resume its original state, for instance—complete irreversibility is a statistical statement that cannot be seen during the lifetime of the universe. More precisely, an *irreversible process* is one that depends on path. If the process can go in only one direction, then the reverse path differs fundamentally and the process cannot be reversible. For example, as noted in the previous section, heat involves the transfer of energy from higher to lower temperature. A cold object in contact with a hot one never gets colder, transferring heat to the hot object and making it hotter. Furthermore, mechanical energy, such as kinetic energy, can be completely converted to thermal energy by friction, but the reverse is impossible. A hot stationary object never spontaneously cools off and starts moving. Yet another example is the expansion of a puff of gas introduced into one corner of a vacuum chamber. The gas expands to fill the chamber, but it never regroups in the corner. The random motion of the gas molecules could take them all back to the corner, but this is never observed to happen. (See Figure 2.)

Figure 1. These ice floes melt during the Arctic summer. Some of them refreeze in the winter, but the second law of thermodynamics predicts that it would be extremely unlikely for the water molecules contained in these particular floes to reform the distinctive alligator-like shape they formed when the picture was taken in the summer of 2009. (credit: Patrick Kelley, U.S. Coast Guard, U.S. Geological Survey)

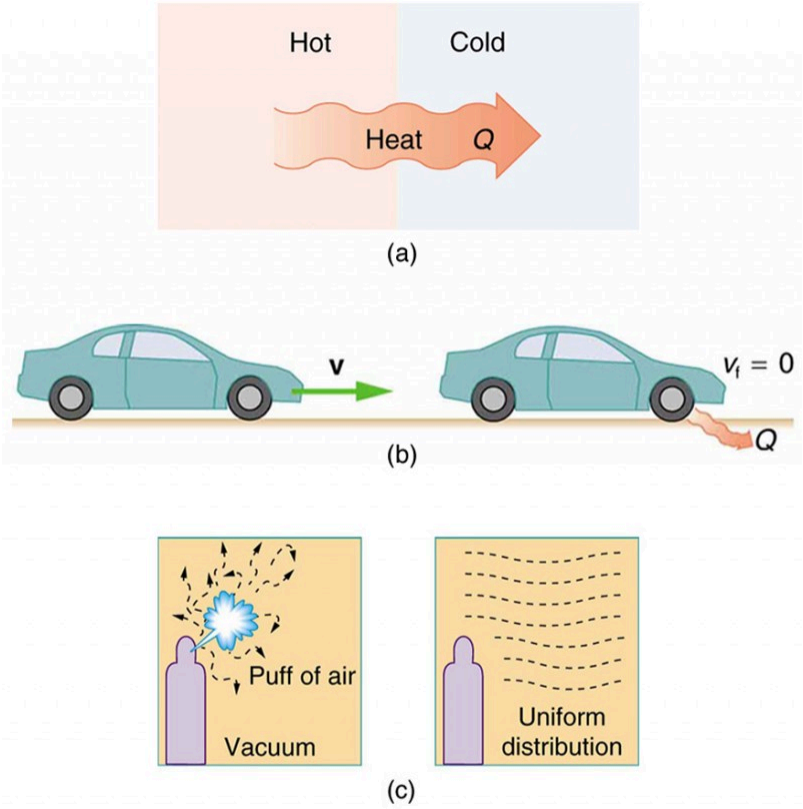


Figure 2. Examples of one-way processes in nature. (a) Heat transfer occurs spontaneously from hot to cold and not from cold to hot. (b) The brakes of this car convert its kinetic energy to heat transfer to the environment. The reverse process is impossible. (c) The burst of gas let into this vacuum chamber quickly expands to uniformly fill every part of the chamber. The random motions of the gas molecules will never return them to the corner.

The fact that certain processes never occur suggests that there is a law forbidding them to occur. The first law of thermodynamics would allow them to occur—none of those processes violate conservation of energy. The law that forbids these processes is called the second law of thermodynamics. We shall see that the second law can be stated in many ways that may seem different, but which in fact are equivalent. Like all natural laws, the second law of thermodynamics gives insights into nature, and its several statements imply that it is broadly applicable, fundamentally affecting many apparently disparate processes.

The already familiar direction of heat transfer from hot to cold is the basis of our first version of the *second law of thermodynamics*

The Second Law of Thermodynamics (first expression)

Heat transfer occurs spontaneously from higher- to lower-temperature bodies but never spontaneously in the reverse direction.

Another way of stating this: It is impossible for any process to have as its sole result heat transfer from a cooler to a hotter object.

Heat Engines

Now let us consider a device that uses heat transfer to do work. As noted in the previous section, such a device is called a heat engine, and one is shown schematically in Figure 3b. Gasoline and diesel engines, jet engines, and steam turbines are all heat engines that do work by using part of the heat transfer from some source. Heat transfer from the hot object (or hot reservoir) is denoted as Q_h, while heat transfer into the cold object (or cold reservoir) is Q_c, and the work done by the engine is W. The temperatures of the hot and cold reservoirs are T_h and T_c, respectively.

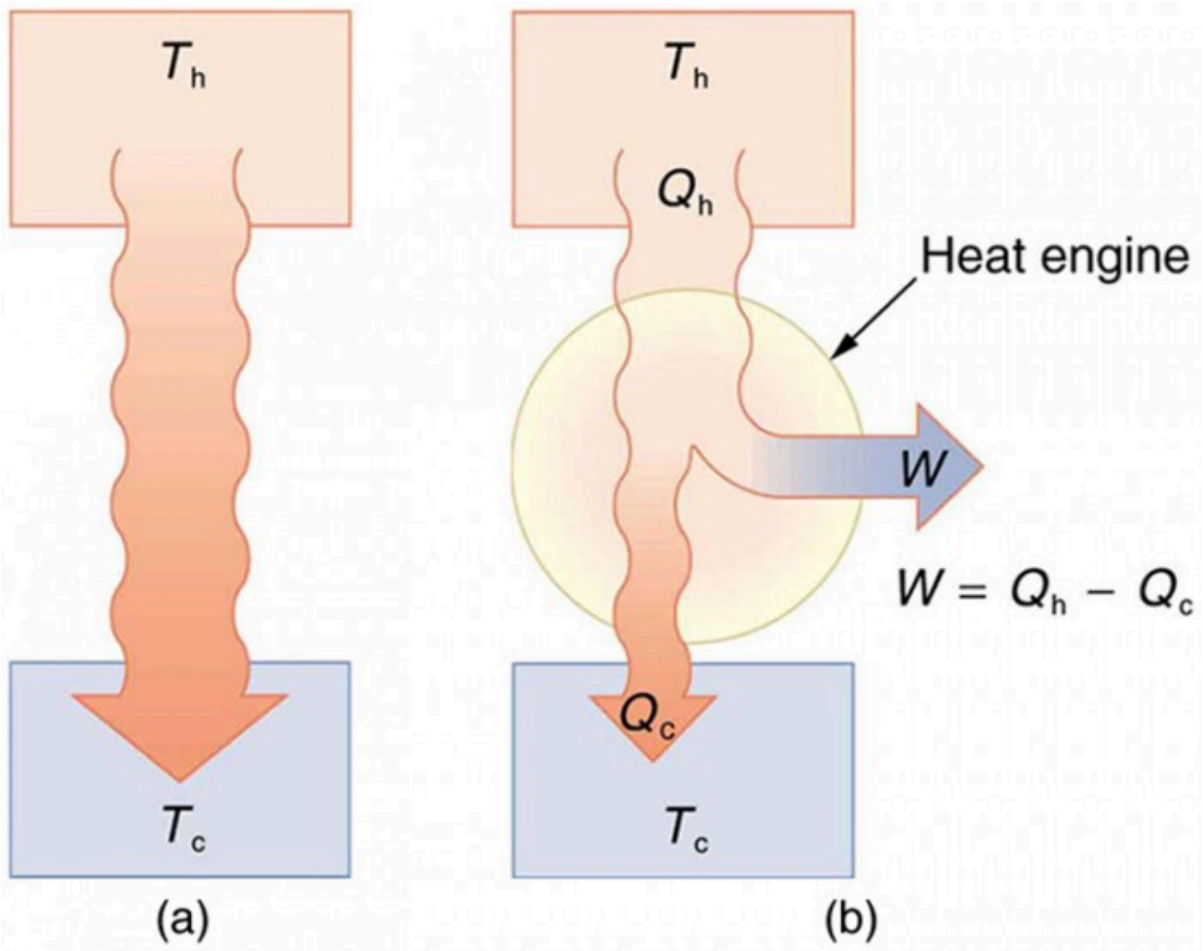


Figure 3. (a) Heat transfer occurs spontaneously from a hot object to a cold one, consistent with the second law of thermodynamics. (b) A heat engine, represented here by a circle, uses part of the heat transfer to do work. The hot and cold objects are called the hot and cold reservoirs. Qh is the heat transfer out of the hot reservoir, W is the work output, and Qc is the heat transfer into the cold reservoir.

Because the hot reservoir is heated externally, which is energy intensive, it is important that the work is done as efficiently as possible. In fact, we would like W to equal Q_h, and for there to be no heat transfer to the environment (Q_c=0). Unfortunately, this is impossible. The *second law of thermodynamics* also states, with regard to using heat transfer to do work (the second expression of the second law):

The Second Law of Thermodynamics (second expression)

It is impossible in any system for heat transfer from a reservoir to completely convert to work in a cyclical process in which the system returns to its initial state.

A *cyclical process* brings a system, such as the gas in a cylinder, back to its original state at the end of every cycle. Most heat engines, such as reciprocating piston engines and rotating turbines, use cyclical processes. The second law, just stated in its second form, clearly states that such engines cannot have

perfect conversion of heat transfer into work done. Before going into the underlying reasons for the limits on converting heat transfer into work, we need to explore the relationships among W, Q_h, and Q_c, and to define the efficiency of a cyclical heat engine. As noted, a cyclical process brings the system back to its original condition at the end of every cycle. Such a system's internal energy U is the same at the beginning and end of every cycle—that is, $\Delta U = 0$. The first law of thermodynamics states that $\Delta U = Q - W$, where Q is the *net* heat transfer during the cycle ($Q = Q_h - Q_c$) and W is the net work done by the system. Since ΔU=0 for a complete cycle, we have $0 = Q - W$, so that $W = Q$.

Thus the net work done by the system equals the net heat transfer into the system, or $W = Q_h - Q_c$ (cyclical process), just as shown schematically in Figure 3b. The problem is that in all processes, there is some heat transfer Q_c to the environment—and usually a very significant amount at that.

In the conversion of energy to work, we are always faced with the problem of getting less out than we put in. We define *conversion efficiency Eff* to be the ratio of useful work output to the energy input (or, in other words, the ratio of what we get to what we spend). In that spirit, we define the efficiency of a heat engine to be its net work output W divided by heat transfer to the engine Q_h; that is,

$$Eff = \frac{W}{Q_h}$$

Since $W = Q_h - Q_c$ in a cyclical process, we can also express this as

$$Eff = \frac{Q_h - Q_c}{Q_h} = 1 - \frac{Q_c}{Q_h}$$

(cyclical process),

making it clear that an efficiency of 1, or 100%, is possible only if there is no heat transfer to the environment ($Q_c = 0$). Note that all Qs are positive. The direction of heat transfer is indicated by a plus or minus sign. For example, Q_c is out of the system and so is preceded by a minus sign.

Example 1. Daily Work Done by a Coal-Fired Power Station, Its Efficiency and Carbon Dioxide Emissions

A coal-fired power station is a huge heat engine. It uses heat transfer from burning coal to do work to turn turbines, which are used to generate electricity. In a single day, a large coal power station has 2.50×10^{14} J of heat transfer from coal and 1.48×10^{14} J of heat transfer into the environment.

1. What is the work done by the power station?
2. What is the efficiency of the power station?
3. In the combustion process, the following chemical reaction occurs: $C + O_2 \rightarrow CO_2$. This implies that every 12 kg of coal puts 12 kg + 16 kg + 16 kg = 44 kg of carbon dioxide into the atmosphere. Assuming that 1 kg of coal can provide 2.5×10^6 J of heat transfer upon combustion, how much CO_2 is emitted per day by this power plant?

Strategy for Part 1

We can use $W = Q_h - Q_c$ to find the work output W, assuming a cyclical process is used in the power station.

In this process, water is boiled under pressure to form high-temperature steam, which is used to run steam turbine-generators, and then condensed back to water to start the cycle again.

Solution for Part 1

Work output is given by: $W = Q_h - Q_c$.

Substituting the given values:

$$\begin{aligned} W &= 2.50 \times 10^{14}\ \text{J} - 1.48 \times 10^{14}\ \text{J} \\ &= 1.02 \times 10^{14}\ \text{J} \end{aligned}$$

Strategy for Part 2

The efficiency can be calculated with

$$Eff = \frac{W}{Q_h}$$

since Q_h is given and work W was found in the first part of this example.

Solution for Part 2

Efficiency is given by:

$$Eff = \frac{W}{Q_h}$$

. The work W was just found to be 1.02×10^{14} J, and Q_h is given, so the efficiency is

$$\begin{aligned} Eff &= \frac{1.02 \times 10^{14}\ \text{J}}{2.50 \times 10^{14}\ \text{J}} \\ &= 0.408, \text{ or } 40.8\% \end{aligned}$$

Strategy for Part 3

The daily consumption of coal is calculated using the information that each day there is 2.50×10^{14} J of heat transfer from coal. In the combustion process, we have $C + O_2 \rightarrow CO_2$. So every 12 kg of coal puts 12 kg + 16 kg + 16 kg = 44 kg of CO_2 into the atmosphere.

Solution for Part 3

The daily coal consumption is

$$\frac{2.50 \times 10^{14}\ \text{J}}{2.50 \times 10^{6}\ \text{J/kg}} = 1.0 \times 10^{8}\ \text{kg}$$

.

Assuming that the coal is pure and that all the coal goes toward producing carbon dioxide, the carbon dioxide produced per day is

$$1.0 \times 10^{8}\ \text{kg coal} \times \frac{44\ \text{kg CO}_2}{12\ \text{kg coal}} = 3.7 \times 10^{8}\ \text{kg CO}_2$$

This is 370,000 metric tons of CO_2 produced every day.

Discussion

If all the work output is converted to electricity in a period of one day, the average power output is 1180 MW (this is left to you as an end-of-chapter problem). This value is about the size of a large-scale conventional power plant. The efficiency found is acceptably close to the value of 42% given for coal power stations. It means that fully 59.2% of the energy is heat transfer to the environment, which usually results in warming lakes, rivers, or the ocean near the power station, and is implicated in a warming planet generally. While the laws of thermodynamics limit the efficiency of such plants—including plants fired by nuclear fuel, oil, and natural gas—the heat transfer to the environment could be, and sometimes is, used for heating homes or for industrial processes. The generally low cost of energy has not made it economical to make better use of the waste heat transfer from most heat engines. Coal-fired power plants produce the greatest amount of CO_2 per unit energy output (compared to natural gas or oil), making coal the least efficient fossil fuel.

With the information given in Example 1, we can find characteristics such as the efficiency of a heat engine without any knowledge of how the heat engine operates, but looking further into the mechanism of the engine will give us greater insight. Figure 4 illustrates the operation of the common four-stroke gasoline engine. The four steps shown complete this heat engine's cycle, bringing the gasoline-air mixture back to its original condition.

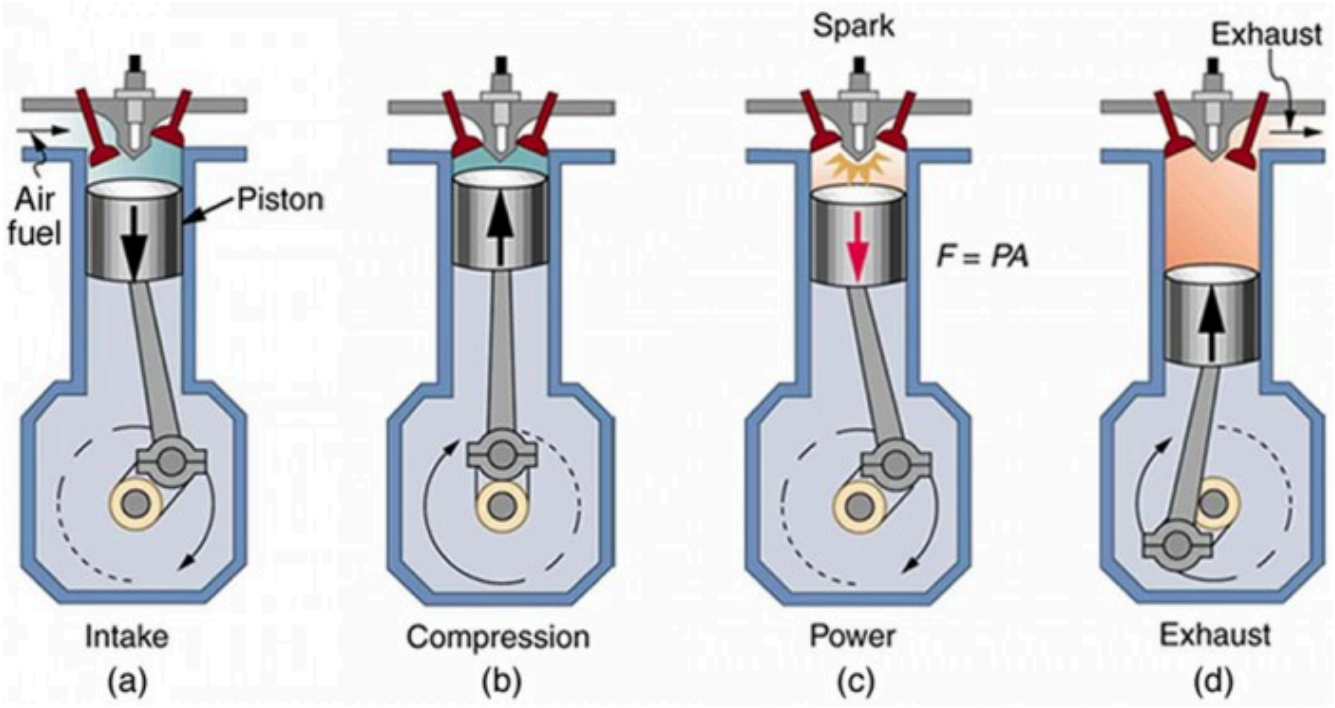


Figure 4. In the four-stroke internal combustion gasoline engine, heat transfer into work takes place in the cyclical process shown here. The piston is connected to a rotating crankshaft, which both takes work out of and does work on the gas in the cylinder. (a) Air is mixed with fuel during the intake stroke. (b) During the compression stroke, the air-fuel mixture is rapidly compressed in a nearly adiabatic process, as the piston rises with the valves closed. Work is done on the gas. (c) The power stroke has two distinct parts. First, the air-fuel mixture is ignited, converting chemical potential energy into thermal energy almost instantaneously, which leads to a great increase in pressure. Then the piston descends, and the gas does work by exerting a force through a distance in a nearly adiabatic process. (d) The exhaust stroke expels the hot gas to prepare the engine for another cycle, starting again with the intake stroke.

The *Otto cycle* shown in Figure 5a is used in four-stroke internal combustion engines, although in fact the true Otto cycle paths do not correspond exactly to the strokes of the engine.

The adiabatic process AB corresponds to the nearly adiabatic compression stroke of the gasoline engine. In both cases, work is done on the system (the gas mixture in the cylinder), increasing its temperature and pressure. Along path BC of the Otto cycle, heat transfer Q_h into the gas occurs at constant volume, causing a further increase in pressure and temperature. This process corresponds to burning fuel in an internal combustion engine, and takes place so rapidly that the volume is nearly constant. Path CD in the Otto cycle is an adiabatic expansion that does work on the outside world, just as the power stroke of an internal combustion engine does in its nearly adiabatic expansion. The work done by the system along path CD is greater than the work done on the system along path AB, because the pressure is greater, and so there is a net work output. Along path DA in the Otto cycle, heat transfer Q_c from the gas at constant volume reduces its temperature and pressure, returning it to its original state. In an internal combustion engine, this process corresponds to the exhaust of hot gases and the intake of an air-gasoline mixture at a considerably lower temperature. In both cases, heat transfer into the environment occurs along this final path.

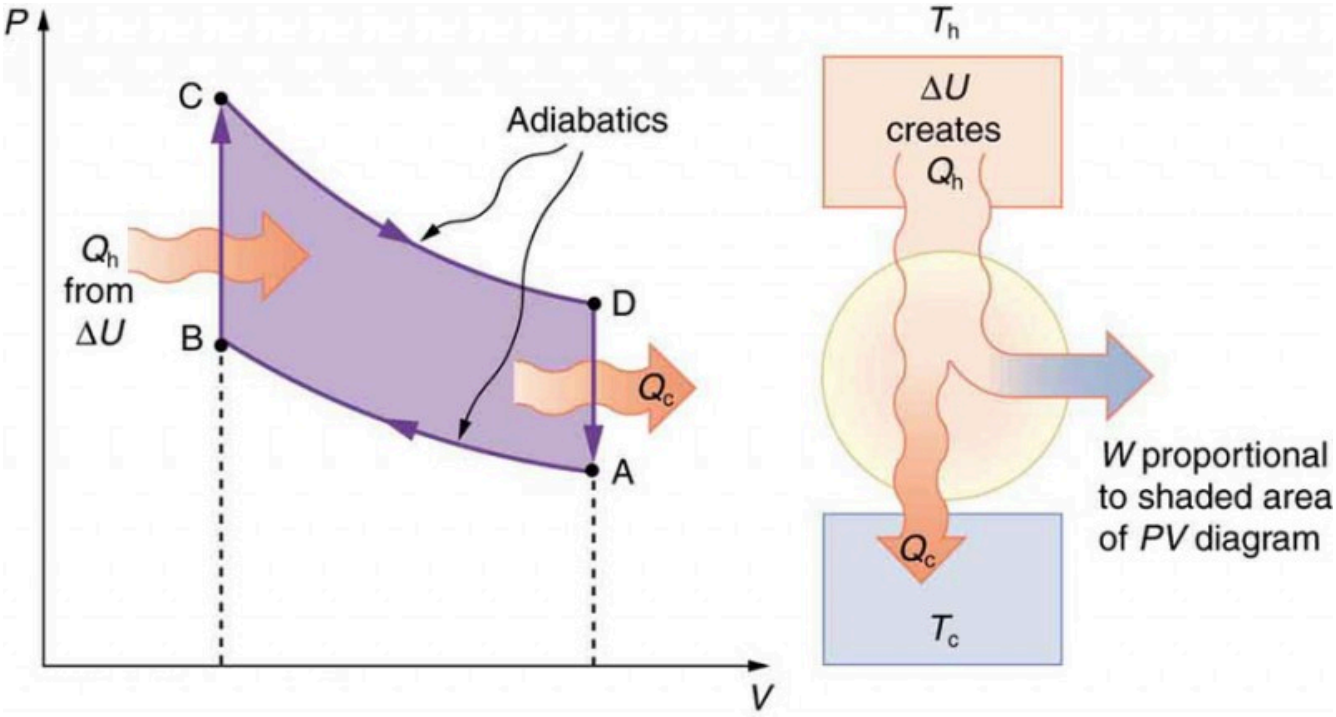


Figure 5. diagram for a simplified Otto cycle, analogous to that employed in an internal combustion engine. Point A corresponds to the start of the compression stroke of an internal combustion engine. Paths AB and CD are adiabatic and correspond to the compression and power strokes of an internal combustion engine, respectively. Paths BC and DA are isochoric and accomplish similar results to the ignition and exhaust-intake portions, respectively, of the internal combustion engine's cycle. Work is done on the gas along path AB, but more work is done by the gas along path CD, so that there is a net work output.

The net work done by a cyclical process is the area inside the closed path on a *PV* diagram, such as that inside path ABCDA in Figure 5. Note that in every imaginable cyclical process, it is absolutely necessary for heat transfer from the system to occur in order to get a net work output. In the Otto cycle, heat transfer occurs along path DA. If no heat transfer occurs, then the return path is the same, and the net work output is zero. The lower the temperature on the path AB, the less work has to be done to

compress the gas. The area inside the closed path is then greater, and so the engine does more work and is thus more efficient. Similarly, the higher the temperature along path CD, the more work output there is. (See Figure 6.) So efficiency is related to the temperatures of the hot and cold reservoirs. In the next section, we shall see what the absolute limit to the efficiency of a heat engine is, and how it is related to temperature.

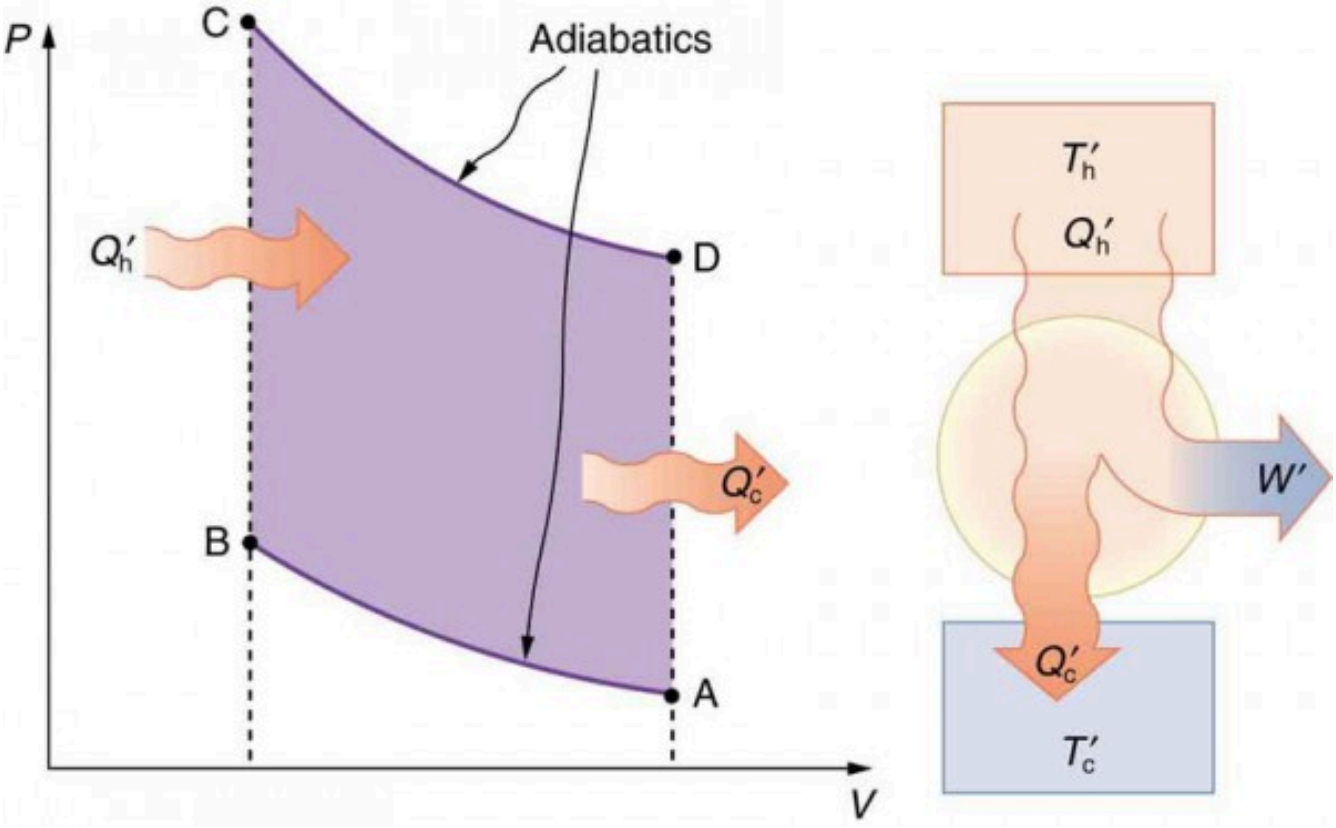


Figure 6. This Otto cycle produces a greater work output than the one in Figure 5, because the starting temperature of path CD is higher and the starting temperature of path AB is lower. The area inside the loop is greater, corresponding to greater net work output.

Section Summary

- The two expressions of the second law of thermodynamics are: (i) Heat transfer occurs spontaneously from higher- to lower-temperature bodies but never spontaneously in the reverse direction; and (ii) It is impossible in any system for heat transfer from a reservoir to completely convert to work in a cyclical process in which the system returns to its initial state.
- Irreversible processes depend on path and do not return to their original state. Cyclical processes are processes that return to their original state at the end of every cycle.
- In a cyclical process, such as a heat engine, the net work done by the system equals the net heat transfer into the system, or $W = Q_h - Q_c$, where Q_h is the heat transfer from the hot object (hot reservoir), and Q_c is the heat transfer into the cold object (cold reservoir).

$$Eff = \frac{W}{Q_h}$$

- Efficiency can be expressed as , the ratio of work output divided by the amount

of energy input.

- The four-stroke gasoline engine is often explained in terms of the Otto cycle, which is a repeating sequence of processes that convert heat into work.

Conceptual Questions

1. Imagine you are driving a car up Pike's Peak in Colorado. To raise a car weighing 1000 kilograms a distance of 100 meters would require about a million joules. You could raise a car 12.5 kilometers with the energy in a gallon of gas. Driving up Pike's Peak (a mere 3000-meter climb) should consume a little less than a quart of gas. But other considerations have to be taken into account. Explain, in terms of efficiency, what factors may keep you from realizing your ideal energy use on this trip.
2. Is a temperature difference necessary to operate a heat engine? State why or why not.
3. Definitions of efficiency vary depending on how energy is being converted. Compare the definitions of efficiency for the human body and heat engines. How does the definition of efficiency in each relate to the type of energy being converted into doing work?
4. Why—other than the fact that the second law of thermodynamics says reversible engines are the most efficient—should heat engines employing reversible processes be more efficient than those employing irreversible processes? Consider that dissipative mechanisms are one cause of irreversibility.

Problems & Exercises

1. A certain heat engine does 10.0 kJ of work and 8.50 kJ of heat transfer occurs to the environment in a cyclical process. (a) What was the heat transfer into this engine? (b) What was the engine's efficiency?
2. With 2.56×10^6 J of heat transfer into this engine, a given cyclical heat engine can do only 1.50×10^5 J of work. (a) What is the engine's efficiency? (b) How much heat transfer to the environment takes place?
3. (a) What is the work output of a cyclical heat engine having a 22.0% efficiency and 6.00×10^9 J of heat transfer into the engine? (b) How much heat transfer occurs to the environment?
4. (a) What is the efficiency of a cyclical heat engine in which 75.0 kJ of heat transfer occurs to the environment for every 95.0 kJ of heat transfer into the engine? (b) How much work does it produce for 100 kJ of heat transfer into the engine?
5. The engine of a large ship does 2.00×10^8 J of work with an efficiency of 5.00%. (a) How much heat transfer occurs to the environment? (b) How many barrels of fuel are consumed, if each barrel produces 6.00×10^9 J of heat transfer when burned?
6. (a) How much heat transfer occurs to the environment by an electrical power station that uses 1.25×10^{14} J of heat transfer into the engine with an efficiency of 42.0%? (b) What is the ratio of heat transfer to the environment to work output? (c) How much work is done?
7. Assume that the turbines at a coal-powered power plant were upgraded, resulting in an

improvement in efficiency of 3.32%. Assume that prior to the upgrade the power station had an efficiency of 36% and that the heat transfer into the engine in one day is still the same at 2.50×1014J. (a) How much more electrical energy is produced due to the upgrade? (b) How much less heat transfer occurs to the environment due to the upgrade?

8. This problem compares the energy output and heat transfer to the environment by two different types of nuclear power stations—one with the normal efficiency of 34.0%, and another with an improved efficiency of 40.0%. Suppose both have the same heat transfer into the engine in one day, 2.50×10^{14} J. (a) How much more electrical energy is produced by the more efficient power station? (b) How much less heat transfer occurs to the environment by the more efficient power station? (One type of more efficient nuclear power station, the gas-cooled reactor, has not been reliable enough to be economically feasible in spite of its greater efficiency.)

Glossary

irreversible process: any process that depends on path direction

second law of thermodynamics: heat transfer flows from a hotter to a cooler object, never the reverse, and some heat energy in any process is lost to available work in a cyclical process

cyclical process: a process in which the path returns to its original state at the end of every cycle

Otto cycle: a thermodynamic cycle, consisting of a pair of adiabatic processes and a pair of isochoric processes, that converts heat into work, e.g., the four-stroke engine cycle of intake, compression, ignition, and exhaust

Selected Solutions to Problems & Exercises

1. (a) 18.5 kJ; (b) 54.1%

3. (a) 1.32×10^9 J; (b) 4.68×10^9 J

5. (a) 3.80×10^9 J; (b) 0.667 barrels

7. (a) 8.30×10^{12} J, which is 3.32% of 2.50×10^{14} J; (b) -8.30×10^{12} J, where the negative sign indicates a reduction in heat transfer to the environment.

Carnot's Perfect Heat Engine: The Second Law of Thermodynamics Restated

Lumen Learning

Learning Objectives

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

- Identify a Carnot cycle.
- Calculate maximum theoretical efficiency of a nuclear reactor.
- Explain how dissipative processes affect the ideal Carnot engine.

The novelty toy known as the drinking bird (seen in Figure 1) is an example of Carnot's engine. It contains methylene chloride (mixed with a dye) in the abdomen, which boils at a very low temperature—about 100ºF . To operate, one gets the bird's head wet. As the water evaporates, fluid moves up into the head, causing the bird to become top-heavy and dip forward back into the water. This cools down the methylene chloride in the head, and it moves back into the abdomen, causing the bird to become bottom heavy and tip up. Except for a very small input of energy—the original head-wetting—the bird becomes a perpetual motion machine of sorts.

Figure 1. A drinking bird (credit: Arabesk.nl, Wikimedia Commons)

We know from the second law of thermodynamics that a heat engine cannot be 100% efficient, since there must always be some heat transfer Q_c to the environment, which is often called waste heat. How efficient, then, can a heat engine be? This question was answered at a theoretical level in 1824 by a young French engineer, Sadi Carnot (1796–1832), in his study of the then-emerging heat engine technology crucial to the Industrial Revolution. He devised a theoretical cycle, now called the *Carnot cycle*, which is the most efficient cyclical process possible. The second law of thermodynamics can be restated in terms of the Carnot cycle, and so what Carnot actually discovered was this fundamental law. Any heat engine employing the Carnot cycle is called a *Carnot engine*.

What is crucial to the Carnot cycle—and, in fact, defines it—is that only reversible processes are used. Irreversible processes involve dissipative factors, such as friction and turbulence. This increases heat transfer Q_c to the environment and reduces the efficiency of the engine. Obviously, then, reversible processes are superior.

Carnot Engine

Stated in terms of reversible processes, the *second law of thermodynamics* has a third form:

> A Carnot engine operating between two given temperatures has the greatest possible efficiency of any heat engine operating between these two temperatures. Furthermore, all engines employing only reversible processes have this same maximum efficiency when operating between the same given temperatures.

Figure 2 shows the *PV* diagram for a Carnot cycle. The cycle comprises two isothermal and two adiabatic processes. Recall that both isothermal and adiabatic processes are, in principle, reversible.

Carnot also determined the efficiency of a perfect heat engine—that is, a Carnot engine. It is always true that the efficiency of a cyclical heat engine is given by:

$$Eff = \frac{Q_{\mathrm{h}} - Q_{\mathrm{c}}}{Q_{\mathrm{h}}} = 1 - \frac{Q_{\mathrm{c}}}{Q_{\mathrm{h}}}$$

What Carnot found was that for a perfect heat engine, the ratio

$$\frac{Q_{\mathrm{c}}}{Q_{\mathrm{h}}}$$

equals the ratio of the absolute temperatures of the heat reservoirs. That is,

$$\frac{Q_{\mathrm{c}}}{Q_{\mathrm{h}}} = \frac{T_{\mathrm{c}}}{T_{\mathrm{h}}}$$

for a Carnot engine, so that the maximum or *Carnot efficiency* Eff_{C} is given by

$$Eff_{\mathrm{C}} = 1 - \frac{T_{\mathrm{c}}}{T_{\mathrm{h}}}$$

where T_{h} and T_{c} are in kelvins (or any other absolute temperature scale). No real heat engine can do as well as the Carnot efficiency—an actual efficiency of about 0.7 of this maximum is usually the best that can be accomplished. But the ideal Carnot engine, like the drinking bird above, while a fascinating novelty, has zero power. This makes it unrealistic for any applications.

Carnot's interesting result implies that 100% efficiency would be possible only if $T_{\mathrm{c}} = 0$ K—that is, only if the cold reservoir were at absolute zero, a practical and theoretical impossibility. But the physical implication is this—the only way to have all heat transfer go into doing work is to remove *all* thermal energy, and this requires a cold reservoir at absolute zero.

It is also apparent that the greatest efficiencies are obtained when the ratio

$$\frac{T_{\mathrm{c}}}{T_{\mathrm{h}}}$$

is as small as possible. Just as discussed for the Otto cycle in the previous section, this means that efficiency is greatest for the highest possible temperature of the hot reservoir and lowest possible temperature of the cold reservoir. (This setup increases the area inside the closed loop on the *PV* diagram; also, it seems reasonable that the greater the temperature difference, the easier it is to divert the heat transfer to work.) The actual reservoir temperatures of a heat engine are usually related to the type of heat source and the temperature of the environment into which heat transfer occurs. Consider the following example.

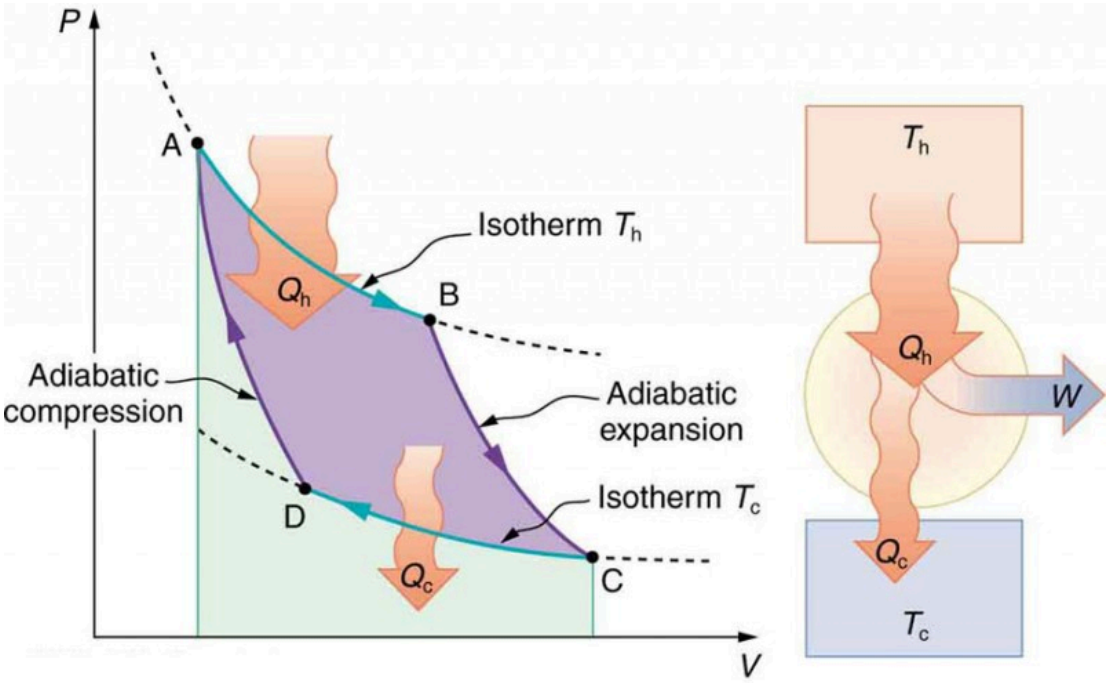


Figure 2. PV diagram for a Carnot cycle, employing only reversible isothermal and adiabatic processes. Heat transfer Q_h occurs into the working substance during the isothermal path AB, which takes place at constant temperature T_h. Heat transfer Q_c occurs out of the working substance during the isothermal path CD, which takes place at constant temperature T_c. The net work output W equals the area inside the path ABCDA. Also shown is a schematic of a Carnot engine operating between hot and cold reservoirs at temperatures T_h and T_c. Any heat engine using reversible processes and operating between these two temperatures will have the same maximum efficiency as the Carnot engine.

Example 1. Maximum Theoretical Efficiency for a Nuclear Reactor

A nuclear power reactor has pressurized water at 300ºC. (Higher temperatures are theoretically possible but practically not, due to limitations with materials used in the reactor.) Heat transfer from this water is a complex process (see Figure 3). Steam, produced in the steam generator, is used to drive the turbine-generators. Eventually the steam is condensed to water at 27ºC and then heated again to start the cycle over. Calculate the maximum theoretical efficiency for a heat engine operating between these two temperatures.

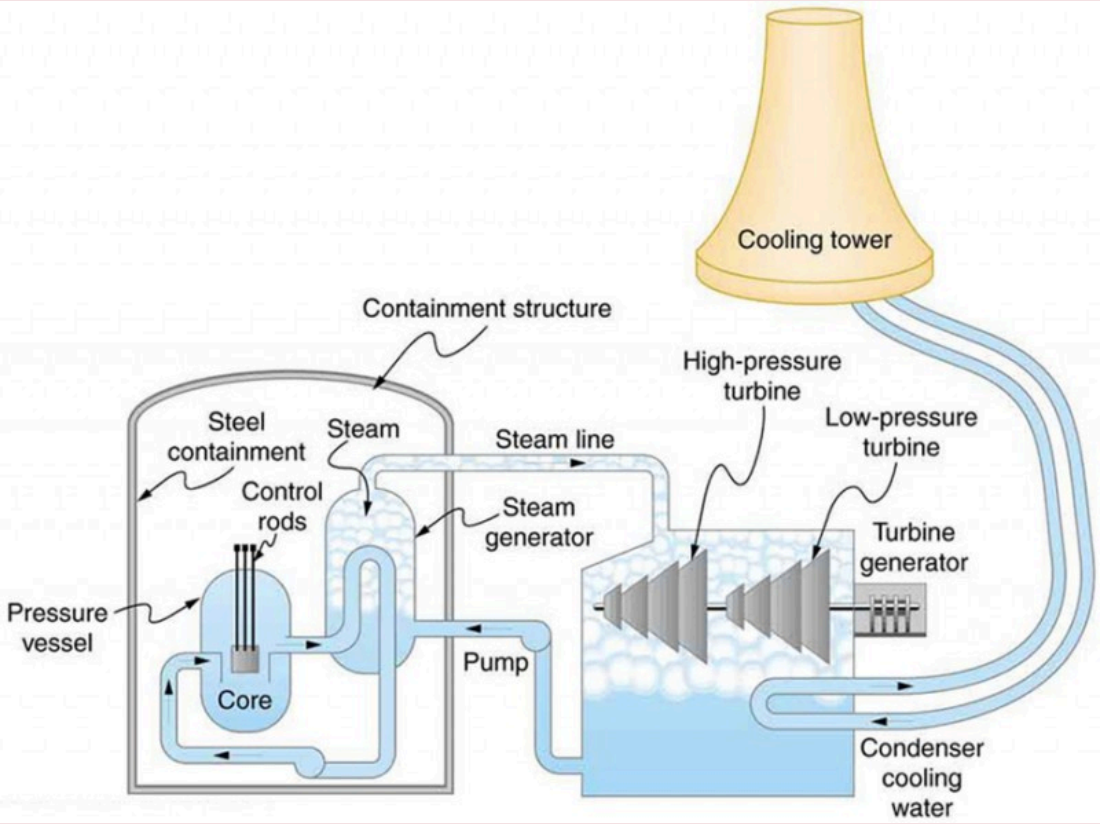


Figure 3. Schematic diagram of a pressurized water nuclear reactor and the steam turbines that convert work into electrical energy. Heat exchange is used to generate steam, in part to avoid contamination of the generators with radioactivity. Two turbines are used because this is less expensive than operating a single generator that produces the same amount of electrical energy. The steam is condensed to liquid before being returned to the heat exchanger, to keep exit steam pressure low and aid the flow of steam through the turbines (equivalent to using a lower-temperature cold reservoir). The considerable energy associated with condensation must be dissipated into the local environment; in this example, a cooling tower is used so there is no direct heat transfer to an aquatic environment. (Note that the water going to the cooling tower does not come into contact with the steam flowing over the turbines.)

Strategy

Since temperatures are given for the hot and cold reservoirs of this heat engine,

$$Eff_{\mathrm{C}} = 1 - \frac{T_{\mathrm{c}}}{T_{\mathrm{h}}}$$

can be used to calculate the Carnot (maximum theoretical) efficiency. Those temperatures must first be converted to kelvins.

Solution

The hot and cold reservoir temperatures are given as 300ºC and 27.0ºC, respectively. In kelvins, then, T_{h} = 573 K and T_{c} = 300 K, so that the maximum efficiency is

$$Eff_{\mathrm{C}} = 1 - \frac{T_{\mathrm{c}}}{T_{\mathrm{h}}}$$

.

Thus,

$$\begin{aligned} Eff_{\mathrm{C}} &= 1 - \frac{300\ \mathrm{K}}{573\ \mathrm{K}} \\ &= 0.476, \text{ or } 47.6\% \end{aligned}$$

Discussion

A typical nuclear power station's actual efficiency is about 35%, a little better than 0.7 times the maximum possible value, a tribute to superior engineering. Electrical power stations fired by coal, oil, and natural gas have greater actual efficiencies (about 42%), because their boilers can reach higher temperatures and pressures. The cold reservoir temperature in any of these power stations is limited by the local environment. Figure 4 shows (a) the exterior of a nuclear power station and (b) the exterior of a coal-fired power station. Both have cooling towers into which water from the condenser enters the tower near the top and is sprayed downward, cooled by evaporation.

(a)

(b)

Figure 4. (a) A nuclear power station (credit: BlatantWorld.com) and (b) a coal-fired power station. Both have cooling towers in which water evaporates into the environment, representing Q_c. The nuclear reactor, which supplies Q_h, is housed inside the dome-shaped containment buildings. (credit: Robert & Mihaela Vicol, publicphoto.org)

Since all real processes are irreversible, the actual efficiency of a heat engine can never be as great as that of a Carnot engine, as illustrated in Figure 5a. Even with the best heat engine possible, there are always dissipative processes in peripheral equipment, such as electrical transformers or car transmissions. These further reduce the overall efficiency by converting some of the engine's work output back into heat transfer, as shown in Figure 5b.

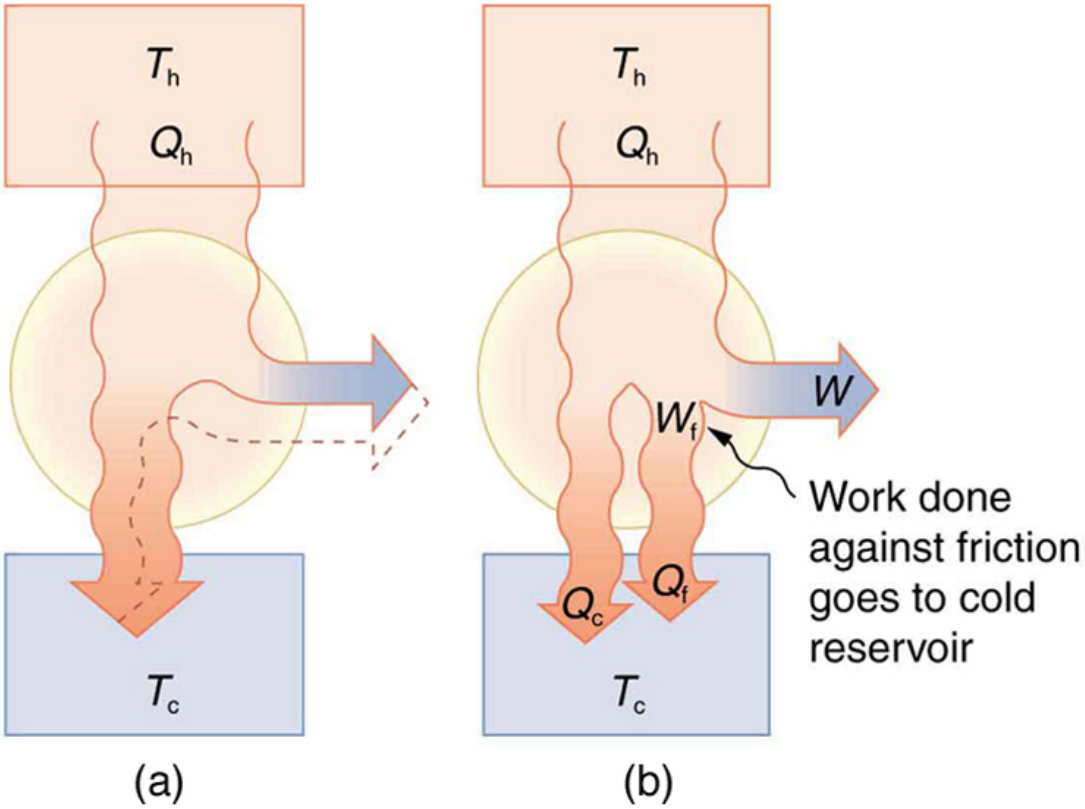


Figure 5. Real heat engines are less efficient than Carnot engines. (a) Real engines use irreversible processes, reducing the heat transfer to work. Solid lines represent the actual process; the dashed lines are what a Carnot engine would do between the same two reservoirs. (b) Friction and other dissipative processes in the output mechanisms of a heat engine convert some of its work output into heat transfer to the environment.

Section Summary

- The Carnot cycle is a theoretical cycle that is the most efficient cyclical process possible. Any engine using the Carnot cycle, which uses only reversible processes (adiabatic and isothermal), is known as a Carnot engine.
- Any engine that uses the Carnot cycle enjoys the maximum theoretical efficiency.
- While Carnot engines are ideal engines, in reality, no engine achieves Carnot's theoretical

maximum efficiency, since dissipative processes, such as friction, play a role. Carnot cycles without heat loss may be possible at absolute zero, but this has never been seen in nature.

Conceptual Questions

1. Think about the drinking bird at the beginning of this section (Figure 1). Although the bird enjoys the theoretical maximum efficiency possible, if left to its own devices over time, the bird will cease "drinking." What are some of the dissipative processes that might cause the bird's motion to cease?
2. Can improved engineering and materials be employed in heat engines to reduce heat transfer into the environment? Can they eliminate heat transfer into the environment entirely?
3. Does the second law of thermodynamics alter the conservation of energy principle?

Problems & Exercises

1. A certain gasoline engine has an efficiency of 30.0%. What would the hot reservoir temperature be for a Carnot engine having that efficiency, if it operates with a cold reservoir temperature of 200ºC?

2. A gas-cooled nuclear reactor operates between hot and cold reservoir temperatures of 700ºC and 27.0ºC. (a) What is the maximum efficiency of a heat engine operating between these temperatures? (b) Find the ratio of this efficiency to the Carnot efficiency of a standard nuclear reactor (found in Example 1).

3. (a) What is the hot reservoir temperature of a Carnot engine that has an efficiency of 42.0% and a cold reservoir temperature of 27.0ºC? (b) What must the hot reservoir temperature be for a real heat engine that achieves 0.700 of the maximum efficiency, but still has an efficiency of 42.0% (and a cold reservoir at 27.0ºC)? (c) Does your answer imply practical limits to the efficiency of car gasoline engines?

4. Steam locomotives have an efficiency of 17.0% and operate with a hot steam temperature of 425ºC. (a) What would the cold reservoir temperature be if this were a Carnot engine? (b) What would the maximum efficiency of this steam engine be if its cold reservoir temperature were 150ºC?

5. Practical steam engines utilize 450ºC steam, which is later exhausted at 270ºC. (a) What is the maximum efficiency that such a heat engine can have? (b) Since 270ºC steam is still quite hot, a second steam engine is sometimes operated using the exhaust of the first. What is the maximum efficiency of the second engine if its exhaust has a temperature of 150ºC? (c) What is the overall efficiency of the two engines? (d) Show that this is the same efficiency as a single Carnot engine operating between 450ºC and 150ºC.

6. A coal-fired electrical power station has an efficiency of 38%. The temperature of the steam leaving the boiler is

$$550^{\circ}\mathrm{C}$$

. What percentage of the maximum efficiency does this station obtain? (Assume the temperature of the environment is

$$20^{\circ}\mathrm{C}$$

.)

7. Would you be willing to financially back an inventor who is marketing a device that she claims has 25 kJ of heat transfer at 600 K, has heat transfer to the environment at 300 K, and does 12 kJ of work? Explain your answer.

8. **Unreasonable Results** (a) Suppose you want to design a steam engine that has heat transfer to the environment at 270ºC and has a Carnot efficiency of 0.800. What temperature of hot steam must you use? (b) What is unreasonable about the temperature? (c) Which premise is unreasonable?

9. **Unreasonable Results** Calculate the cold reservoir temperature of a steam engine that uses hot steam at 450ºC and has a Carnot efficiency of 0.700. (b) What is unreasonable about the temperature? (c) Which premise is unreasonable?

Glossary

Carnot cycle: a cyclical process that uses only reversible processes, the adiabatic and isothermal processes

Carnot engine: a heat engine that uses a Carnot cycle

Carnot efficiency: the maximum theoretical efficiency for a heat engine

Selected Solutions to Problems & Exercises

1. 403ºC

3. (a) 244ºC; (b) 477ºC; (c)Yes, since automobiles engines cannot get too hot without overheating, their efficiency is limited.

5. (a)

$$\text{Eff}_1 = 1 - \frac{T_{\text{c},1}}{T_{\text{h},1}} = 1 - \frac{543\text{ K}}{723\text{ K}} = 0.249 \text{ or } 24.9\%$$

(b)

$$\text{Eff}_2 = 1 - \frac{423\text{ K}}{543\text{ K}} = 0.221 \text{ or } 22.1\%$$

(c)

$$\text{Eff}_1 = 1 - \frac{T_{\text{c},1}}{T_{\text{h},1}} \Rightarrow T_{\text{c},1} = T_{\text{h},1}\left(1, -, \text{eff}_1\right) \text{similarly, } T_{\text{c},2} = T_{\text{h},2}\left(1 - \text{Eff}_2\right)$$

using $T_{\text{h},2} = T_{\text{c},1}$ in above equation gives

$$T_{\text{c},2} = T_{\text{h},1}\left(1 - Eff_1\right)\left(1 - Eff_2\right) \equiv T_{\text{h},1}\left(1 - Eff_{\text{overall}}\right)$$
$$\therefore \left(1 - Eff_{\text{overall}}\right) = \left(1 - \text{Eff}_1\right)\left(1 - Eff_2\right)$$
$$Eff_{\text{overall}} = 1 - (1 - 0.249)\,(1 - 0.221) = 41.5\%$$

(d)

$$\text{Eff}_{\text{overall}} = 1 - \frac{423\text{ K}}{723\text{ K}} = 0.415 \text{ or } 41.5$$

7. The heat transfer to the cold reservoir is

$$Q_{\text{c}} = Q_{\text{h}} - W = 25\text{kJ} - 12\text{kJ} = 13\text{kJ}$$

, so the efficiency is

$$Eff = 1 - \frac{Q_{\mathrm{c}}}{Q_{\mathrm{h}}} = 1 - \frac{13\mathrm{kJ}}{25\mathrm{kJ}} = 0.48$$

$$\mathrm{Eff_C} = 1 - \frac{T_{\mathrm{c}}}{T_{\mathrm{h}}} = 1 - \frac{300\mathrm{K}}{600\mathrm{K}} = 0.50$$

. The Carnot efficiency is . The actual efficiency is 96% of the Carnot efficiency, which is much higher than the best-ever achieved of about 70%, so her scheme is likely to be fraudulent.

9. (a) -56.3ºC (b) The temperature is too cold for the output of a steam engine (the local environment). It is below the freezing point of water. (c) The assumed efficiency is too high.

Applications of Thermodynamics: Heat Pumps and Refrigerators

Lumen Learning

Learning Objectives

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

- Describe the use of heat engines in heat pumps and refrigerators.
- Demonstrate how a heat pump works to warm an interior space.
- Explain the differences between heat pumps and refrigerators.
- Calculate a heat pump's coefficient of performance.

Heat pumps, air conditioners, and refrigerators utilize heat transfer from cold to hot. They are heat engines run backward. We say backward, rather than reverse, because except for Carnot engines, all heat engines, though they can be run backward, cannot truly be reversed. Heat transfer occurs from a cold reservoir Q_c and into a hot one. This requires work input W, which is also converted to heat transfer. Thus the heat transfer to the hot reservoir is $Q_h = Q_c + W$. (Note that Q_h, Q_c, and W are positive, with their directions indicated on schematics rather than by sign.) A heat pump's mission is for heat transfer Q_h to occur into a warm environment, such as a home in the winter. The mission of air conditioners and refrigerators is for heat transfer Q_c to occur from a cool environment, such as chilling a room or keeping food at lower temperatures than the environment. (Actually, a heat pump can be used both to heat and cool a space. It is essentially an air conditioner and a heating unit all in one. In this section we will concentrate on its heating mode.)

Figure 1. Almost every home contains a refrigerator. Most people don't realize they are also sharing their homes with a heat pump. (credit: Id1337x, Wikimedia Commons)

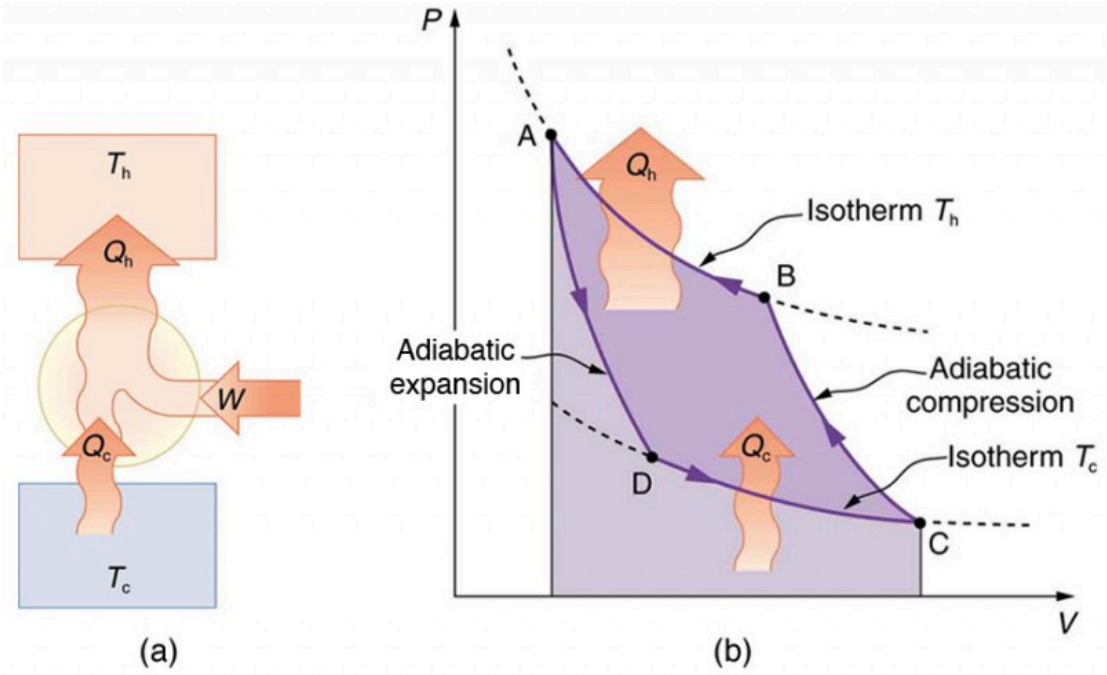


Figure 2. Heat pumps, air conditioners, and refrigerators are heat engines operated backward. The one shown here is based on a Carnot (reversible) engine. (a) Schematic diagram showing heat transfer from a cold reservoir to a warm reservoir with a heat pump. The directions of W, Q_h, and Q_c are opposite what they would be in a heat engine. (b) diagram for a Carnot cycle similar to that in Figure 3 but reversed, following path ADCBA. The area inside the loop is negative, meaning there is a net work input. There is heat transfer Q_c into the system from a cold reservoir along path DC, and heat transfer Q_h out of the system into a hot reservoir along path BA.

Heat Pumps

The great advantage of using a heat pump to keep your home warm, rather than just burning fuel, is that a heat pump supplies $Q_h = Q_c + W$. Heat transfer is from the outside air, even at a temperature below freezing, to the indoor space. You only pay for W, and you get an additional heat transfer of Q_c from the outside at no cost; in many cases, at least twice as much energy is transferred to the heated space as is used to run the heat pump. When you burn fuel to keep warm, you pay for all of it. The disadvantage is that the work input (required by the second law of thermodynamics) is sometimes more expensive than simply burning fuel, especially if the work is done by electrical energy.

The basic components of a heat pump in its heating mode are shown in Figure 3. A working fluid such as a non-CFC refrigerant is used. In the outdoor coils (the evaporator), heat transfer Q_c occurs to the working fluid from the cold outdoor air, turning it into a gas.

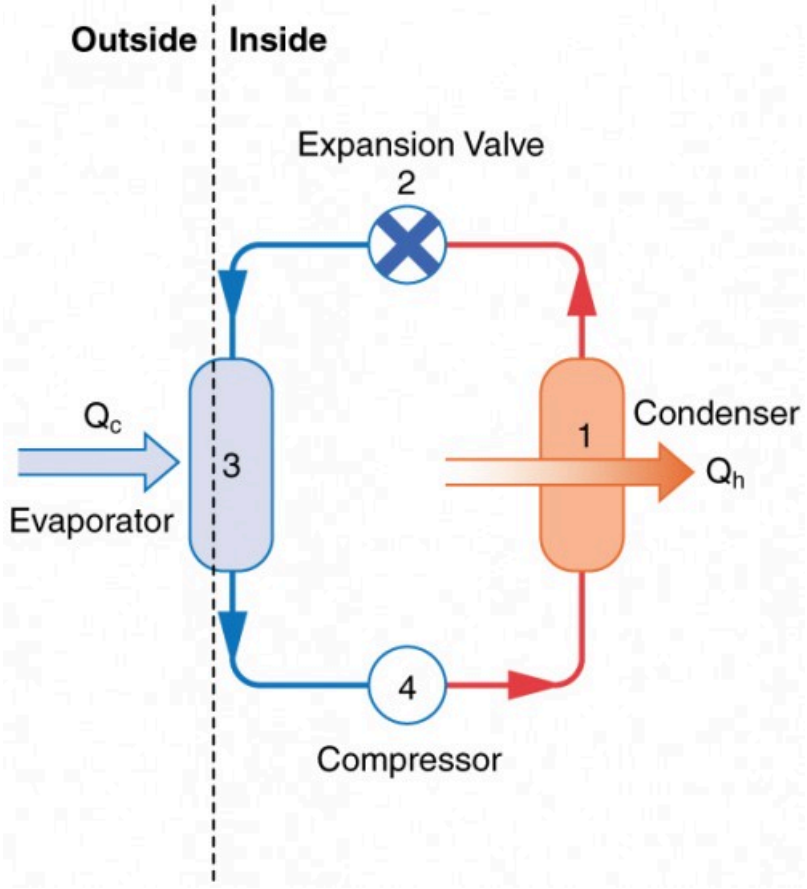


Figure 3. A simple heat pump has four basic components: (1) condenser, (2) expansion valve, (3) evaporator, and (4) compressor. In the heating mode, heat transfer Q_c occurs to the working fluid in the evaporator (3) from the colder outdoor air, turning it into a gas. The electrically driven compressor (4) increases the temperature and pressure of the gas and forces it into the condenser coils (1) inside the heated space. Because the temperature of the gas is higher than the temperature in the room, heat transfer from the gas to the room occurs as the gas condenses to a liquid. The working fluid is then cooled as it flows back through an expansion valve (2) to the outdoor evaporator coils.

The electrically driven compressor (work input *W*) raises the temperature and pressure of the gas and forces it into the condenser coils that are inside the heated space. Because the temperature of the gas is higher than the temperature inside the room, heat transfer to the room occurs and the gas condenses to a liquid. The liquid then flows back through a pressure-reducing valve to the outdoor evaporator coils, being cooled through expansion. (In a cooling cycle, the evaporator and condenser coils exchange roles and the flow direction of the fluid is reversed.)

The quality of a heat pump is judged by how much heat transfer Q_h occurs into the warm space compared with how much work input *W* is required. In the spirit of taking the ratio of what you get to what you spend, we define a *heat pump's coefficient of performance* (COP_{hp}) to be

$$COP_{hp} = \frac{Q_h}{W}$$

.

Since the efficiency of a heat engine is

$$Eff = \frac{W}{Q_{\mathrm{h}}}$$

, we see that

$$COP_{\mathrm{hp}} = \frac{1}{Eff}$$

, an important and interesting fact. First, since the efficiency of any heat engine is less than 1, it means that COP_{hp} is always greater than 1—that is, a heat pump always has more heat transfer Q_{h} than work put into it. Second, it means that heat pumps work best when temperature differences are small. The efficiency of a perfect, or Carnot, engine is $Eff_{\mathrm{C}} = 1 - \left(\frac{T_{\mathrm{c}}}{T_{\mathrm{h}}}\right)$; thus, the smaller the temperature difference, the smaller the efficiency and the greater the COP_{hp} (because $COP_{\mathrm{hp}} = \frac{1}{Eff}$). In other words, heat pumps do not work as well in very cold climates as they do in more moderate climates.

Friction and other irreversible processes reduce heat engine efficiency, but they do *not* benefit the operation of a heat pump—instead, they reduce the work input by converting part of it to heat transfer back into the cold reservoir before it gets into the heat pump.

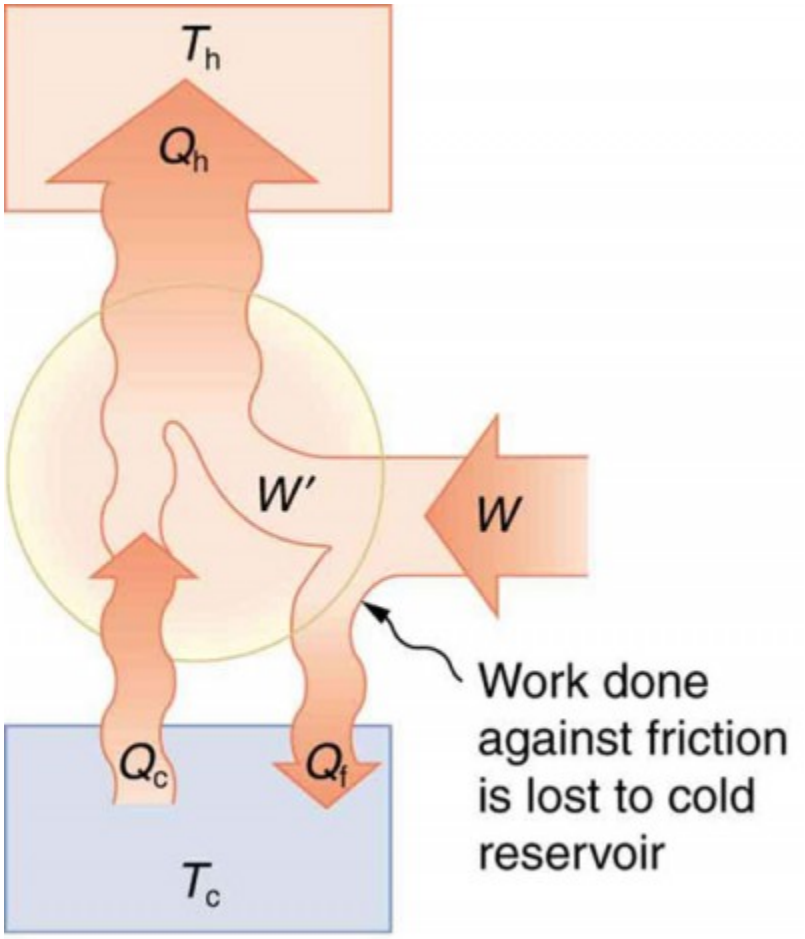


Figure 4. When a real heat engine is run backward, some of the intended work input (W) goes into heat transfer before it gets into the heat engine, thereby reducing its coefficient of performance . In this figure, W' represents the portion of W that goes into the heat pump, while the remainder of W is lost in the form of frictional heat (Q_f) to the cold reservoir. If all of W had gone into the heat pump, then Q_h would have been greater. The best heat pump uses adiabatic and isothermal processes, since, in theory, there would be no dissipative processes to reduce the heat transfer to the hot reservoir.

Example 1. The Best COP_{hp} of a Heat Pump for Home Use

A heat pump used to warm a home must employ a cycle that produces a working fluid at temperatures greater than typical indoor temperature so that heat transfer to the inside can take place. Similarly, it must produce a working fluid at temperatures that are colder than the outdoor temperature so that heat transfer occurs from outside. Its hot and cold reservoir temperatures therefore cannot be too close, placing a limit on its COP_{hp}. (See Figure 5.) What is the best coefficient of performance possible for such a heat pump, if it has a hot reservoir temperature of 45.0ºC and a cold reservoir temperature of −15.0ºC?

Strategy

A Carnot engine reversed will give the best possible performance as a heat pump. As noted above,

$$COP_{\text{hp}} = \frac{1}{Eff}$$

, so that we need to first calculate the Carnot efficiency to solve this problem.

Solution

Carnot efficiency in terms of absolute temperature is given by:

$$Eff_{\text{C}} = 1 - \frac{T_{\text{c}}}{T_{\text{h}}}$$

.

The temperatures in kelvins are T_{h} = 318 K and T_{c} = 258 K, so that

$$Eff_{\text{C}} = 1 - \frac{258\ \text{K}}{318\ \text{K}} = 0.1887$$

.

Thus, from the discussion above,

$$COP_{\text{hp}} = \frac{1}{Eff} = \frac{1}{0.1887} = 5.30$$

, or

$$COP_{\text{hp}} = \frac{Q_{\text{h}}}{W} = \frac{1}{0.1887} = 5.30$$

so that Q_{h} = 5.30 W.

Discussion

This result means that the heat transfer by the heat pump is 5.30 times as much as the work put into it. It would cost 5.30 times as much for the same heat transfer by an electric room heater as it does for that produced by this heat pump. This is not a violation of conservation of energy. Cold ambient air provides 4.3 J per 1 J of work from the electrical outlet.

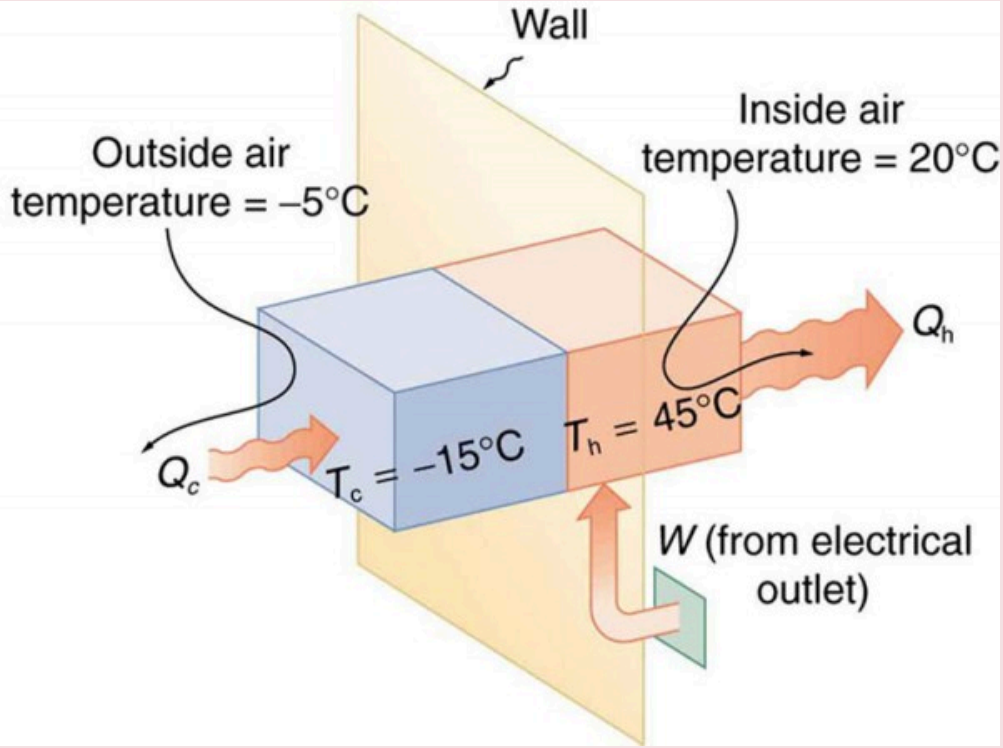


Figure 5. Heat transfer from the outside to the inside, along with work done to run the pump, takes place in the heat pump of the example above. Note that the cold temperature produced by the heat pump is lower than the outside temperature, so that heat transfer into the working fluid occurs. The pump's compressor produces a temperature greater than the indoor temperature in order for heat transfer into the house to occur.

Real heat pumps do not perform quite as well as the ideal one in the previous example; their values of COP_{hp} range from about 2 to 4. This range means that the heat transfer Q_h from the heat pumps is 2 to 4 times as great as the work W put into them. Their economical feasibility is still limited, however, since W is usually supplied by electrical energy that costs more per joule than heat transfer by burning fuels like natural gas. Furthermore, the initial cost of a heat pump is greater than that of many furnaces, so that a heat pump must last longer for its cost to be recovered. Heat pumps are most likely to be economically superior where winter temperatures are mild, electricity is relatively cheap, and other fuels are relatively expensive. Also, since they can cool as well as heat a space, they have advantages where cooling in summer months is also desired. Thus some of the best locations for heat pumps are in warm summer climates with cool winters. Figure 6 shows a heat pump, called a *"reverse cycle"* or *"split-system cooler"* in some countries.

Figure 6. In hot weather, heat transfer occurs from air inside the room to air outside, cooling the room. In cool weather, heat transfer occurs from air outside to air inside, warming the room. This switching is achieved by reversing the direction of flow of the working fluid.

Air Conditioners and Refrigerators

Air conditioners and refrigerators are designed to cool something down in a warm environment. As with heat pumps, work input is required for heat transfer from cold to hot, and this is expensive. The quality of air conditioners and refrigerators is judged by how much heat transfer Q_c occurs from a cold environment compared with how much work input W is required. What is considered the benefit in a heat pump is considered waste heat in a refrigerator. We thus define the *coefficient of performance* (COP_{ref}) of an air conditioner or refrigerator to be

$$COP_{\text{ref}} = \frac{Q_c}{W}$$

.

Noting again that $Q_h = Q_c + W$, we can see that an air conditioner will have a lower coefficient of performance than a heat pump, because

$$COP_{\text{hp}} = \frac{Q_h}{W}$$

and Q_h is greater than Q_c. In this module's Problems and Exercises, you will show that $COP_{ref} = COP_{hp} - 1$ for a heat engine used as either an air conditioner or a heat pump operating between the same two temperatures. Real air conditioners and refrigerators typically do remarkably well, having values of COP_{ref} ranging from 2 to 6. These numbers are better than the COP_{hp} values for the heat pumps mentioned above, because the temperature differences are smaller, but they are less than those for Carnot engines operating between the same two temperatures.

A type of *COP* rating system called the "energy efficiency rating" (*EER*) has been developed. This rating is an example where non-SI units are still used and relevant to consumers. To make it easier for the

consumer, Australia, Canada, New Zealand, and the U.S. use an Energy Star Rating out of 5 stars—the more stars, the more energy efficient the appliance. *EER*s are expressed in mixed units of British thermal units (Btu) per hour of heating or cooling divided by the power input in watts. Room air conditioners are readily available with *EER*s ranging from 6 to 12. Although not the same as the *COP*s just described, these *EER*s are good for comparison purposes—the greater the *EER*, the cheaper an air conditioner is to operate (but the higher its purchase price is likely to be).

The *EER* of an air conditioner or refrigerator can be expressed as

$$EER = \frac{\frac{Q_c}{t_1}}{\frac{W}{t_2}}$$

,

where Q_c is the amount of heat transfer from a cold environment in British thermal units, t_1 is time in hours, W is the work input in joules, and t_2 is time in seconds.

Problem-Solving Strategies for Thermodynamics

1. *Examine the situation to determine whether heat, work, or internal energy are involved.* Look for any system where the primary methods of transferring energy are heat and work. Heat engines, heat pumps, refrigerators, and air conditioners are examples of such systems.
2. *Identify the system of interest and draw a labeled diagram of the system showing energy flow.*
3. *Identify exactly what needs to be determined in the problem (identify the unknowns).* A written list is useful. Maximum efficiency means a Carnot engine is involved. Efficiency is not the same as the coefficient of performance.
4. *Make a list of what is given or can be inferred from the problem as stated (identify the knowns).* Be sure to distinguish heat transfer into a system from heat transfer out of the system, as well as work input from work output. In many situations, it is useful to determine the type of process, such as isothermal or adiabatic.
5. *Solve the appropriate equation for the quantity to be determined (the unknown).*
6. *Substitute the known quantities along with their units into the appropriate equation and obtain numerical solutions complete with units.*
7. *Check the answer to see if it is reasonable: Does it make sense?* For example, efficiency is always less than 1, whereas coefficients of performance are greater than 1.

Section Summary

- An artifact of the second law of thermodynamics is the ability to heat an interior space using a heat pump. Heat pumps compress cold ambient air and, in so doing, heat it to room temperature without violation of conservation principles.

$$\mathrm{COP_{hp}} = \frac{Q_h}{W}$$

- To calculate the heat pump's coefficient of performance, use the equation .

- A refrigerator is a heat pump; it takes warm ambient air and expands it to chill it.

Conceptual Questions

1. Explain why heat pumps do not work as well in very cold climates as they do in milder ones. Is the same true of refrigerators?
2. In some Northern European nations, homes are being built without heating systems of any type. They are very well insulated and are kept warm by the body heat of the residents. However, when the residents are not at home, it is still warm in these houses. What is a possible explanation?
3. Why do refrigerators, air conditioners, and heat pumps operate most cost-effectively for cycles with a small difference between T_h and T_c? (Note that the temperatures of the cycle employed are crucial to its *COP*.)
4. Grocery store managers contend that there is less total energy consumption in the summer if the store is kept at a low temperature. Make arguments to support or refute this claim, taking into account that there are numerous refrigerators and freezers in the store.
5. Can you cool a kitchen by leaving the refrigerator door open?

Problems & Exercises

1. What is the coefficient of performance of an ideal heat pump that has heat transfer from a cold temperature of −25.0ºC to a hot temperature of 40.0ºC?
2. Suppose you have an ideal refrigerator that cools an environment at−20.0ºC and has heat transfer to another environment at 50.0ºC. What is its coefficient of performance?
3. What is the best coefficient of performance possible for a hypothetical refrigerator that could make liquid nitrogen at −200ºC and has heat transfer to the environment at 35.0ºC?
4. In a very mild winter climate, a heat pump has heat transfer from an environment at 5.00ºC to one at 35.0ºC. What is the best possible coefficient of performance for these temperatures? Explicitly show how you follow the steps in the Problem-Solving Strategies for Thermodynamics.
5. (a) What is the best coefficient of performance for a heat pump that has a hot reservoir temperature of 50.0ºC and a cold reservoir temperature of −20.0ºC? (b) How much heat transfer occurs into the warm environment if 3.60×10^7 J of work (10.0 kW · h) is put into it? (c) If the cost of this work input is 10.0 cents/kW · h, how does its cost compare with the direct heat transfer achieved by burning natural gas at a cost of 85.0 cents per therm. (A therm is a common unit of energy for natural gas and equals 1.055×10^8 J.)
6. (a) What is the best coefficient of performance for a refrigerator that cools an environment at −30.0ºC and has heat transfer to another environment at 45.0ºC? (b) How much work in joules must be done for a heat transfer of 4186 kJ from the cold environment? (c) What is the cost of doing this if the work costs 10.0 cents per 3.60×10^6 J (a kilowatt-hour)? (d) How many kJ of heat transfer occurs into the warm environment? (e) Discuss what type of refrigerator might operate between these temperatures.
7. Suppose you want to operate an ideal refrigerator with a cold temperature of −10.0ºC, and you would like it to have a coefficient of performance of 7.00. What is the hot reservoir temperature

for such a refrigerator?

8. An ideal heat pump is being considered for use in heating an environment with a temperature of 22.0ºC. What is the cold reservoir temperature if the pump is to have a coefficient of performance of 12.0?
9. A 4-ton air conditioner removes 5.06×10^7 J (48,000 British thermal units) from a cold environment in 1.00 h. (a) What energy input in joules is necessary to do this if the air conditioner has an energy efficiency rating (*EER*) of 12.0? (b) What is the cost of doing this if the work costs 10.0 cents per 3.60×10^6 J (one kilowatt-hour)? (c) Discuss whether this cost seems realistic. Note that the energy efficiency rating (*EER*) of an air conditioner or refrigerator is defined to be the number of British thermal units of heat transfer from a cold environment per hour divided by the watts of power input.
10. Show that the coefficients of performance of refrigerators and heat pumps are related by $COP_{\text{ref}} = COP_{\text{hp}} - 1$. Start with the definitions of the *COP*s and the conservation of energy relationship between Q_{h}, Q_{c}, and W.

Glossary

heat pump: a machine that generates heat transfer from cold to hot

coefficient of performance: for a heat pump, it is the ratio of heat transfer at the output (the hot reservoir) to the work supplied; for a refrigerator or air conditioner, it is the ratio of heat transfer from the cold reservoir to the work supplied

Selected Solutions to Problems & Exercises

1. 4.82

3. 0.311

5. (a) 4.61; (b) 1.66×10^8 J or 3.97×10^4 kcal; (c) To transfer 1.66×10^8 J, heat pump costs $1.00, natural gas costs $1.34.

7. 27.6ºC

9. (a) 1.44×10^7 J; (b) 40 cents; (c) This cost seems quite realistic; it says that running an air conditioner all day would cost $9.59 (if it ran continuously).

Entropy and the Second Law of Thermodynamics: Disorder and the Unavailability of Energy

Lumen Learning

Learning Objectives

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

- Define entropy.
- Calculate the increase of entropy in a system with reversible and irreversible processes.
- Explain the expected fate of the universe in entropic terms.
- Calculate the increasing disorder of a system.

Figure 1. The ice in this drink is slowly melting. Eventually the liquid will reach thermal equilibrium, as predicted by the second law of thermodynamics. (credit: Jon Sullivan, PDPhoto.org)

There is yet another way of expressing the second law of thermodynamics. This version relates to a

concept called *entropy*. By examining it, we shall see that the directions associated with the second law—heat transfer from hot to cold, for example—are related to the tendency in nature for systems to become disordered and for less energy to be available for use as work. The entropy of a system can in fact be shown to be a measure of its disorder and of the unavailability of energy to do work.

Making Connections: Entropy, Energy, and Work

Recall that the simple definition of energy is the ability to do work. Entropy is a measure of how much energy is not available to do work. Although all forms of energy are interconvertible, and all can be used to do work, it is not always possible, even in principle, to convert the entire available energy into work. That unavailable energy is of interest in thermodynamics, because the field of thermodynamics arose from efforts to convert heat to work.

We can see how entropy is defined by recalling our discussion of the Carnot engine. We noted that for a Carnot cycle, and hence for any reversible processes,

$$\frac{Q_{\mathrm{c}}}{Q_{\mathrm{h}}} = \frac{T_{\mathrm{c}}}{T_{\mathrm{h}}}$$

.

Rearranging terms yields

$$\frac{Q_{\mathrm{c}}}{T_{\mathrm{c}}} = \frac{Q_{\mathrm{h}}}{T_{\mathrm{h}}}$$

for any reversible process. Q_{c} and Q_{h} are absolute values of the heat transfer at temperatures T_{c} and T_{h}, respectively. This ratio of

$$\frac{Q}{T}$$

is defined to be the *change in entropy* ΔS for a reversible process,

$$\Delta S = \left(\frac{Q}{T}\right)_{\mathrm{rev}}$$

, where Q is the heat transfer, which is positive for heat transfer into and negative for heat transfer out of, and T is the absolute temperature at which the reversible process takes place. The SI unit for entropy is joules per kelvin (J/K). If temperature changes during the process, then it is usually a good approximation (for small changes in temperature) to take T to be the average temperature, avoiding the need to use integral calculus to find ΔS.

The definition of ΔS is strictly valid only for reversible processes, such as used in a Carnot engine. However, we can find ΔS precisely even for real, irreversible processes. The reason is that the entropy S of a system, like internal energy U, depends only on the state of the system and not how it reached that condition. Entropy is a property of state. Thus the change in entropy ΔS of a system between state 1 and state 2 is the same no matter how the change occurs. We just need to find or imagine a reversible process

that takes us from state 1 to state 2 and calculate Δ*S* for that process. That will be the change in entropy for any process going from state 1 to state 2. (See Figure 2.)

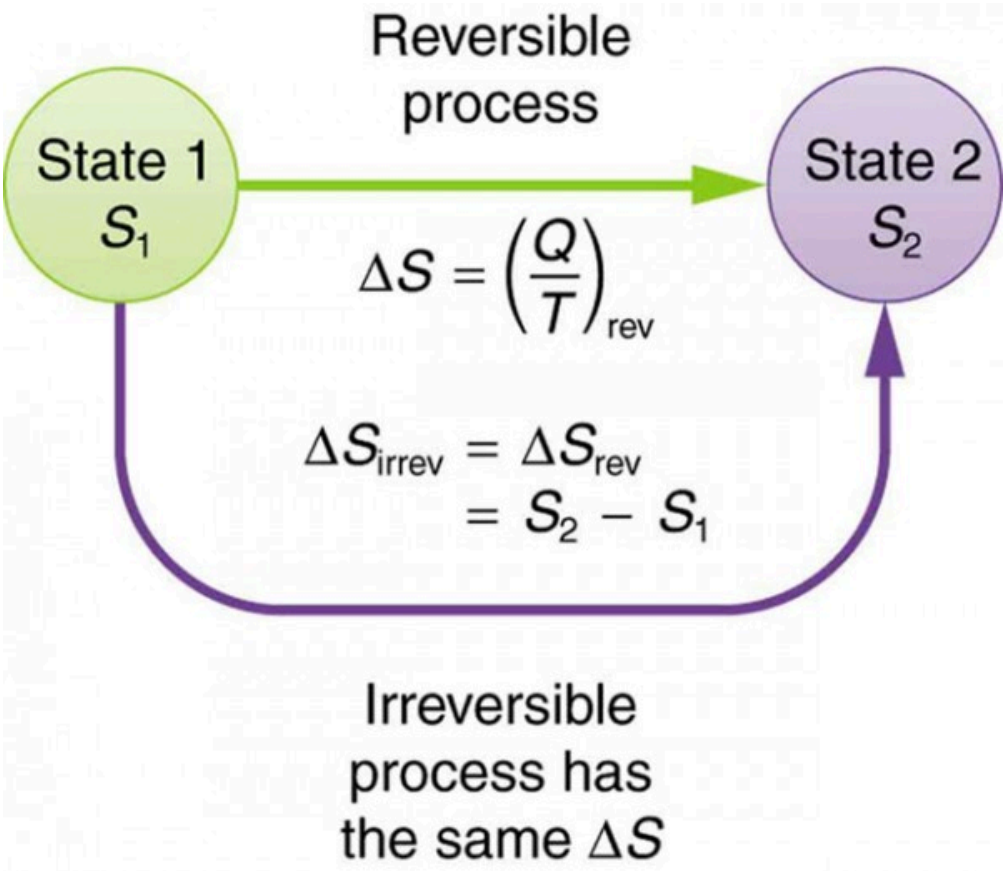


*Figure 2. When a system goes from state 1 to state 2, its entropy changes by the same amount Δ*S*, whether a hypothetical reversible path is followed or a real irreversible path is taken.*

Now let us take a look at the change in entropy of a Carnot engine and its heat reservoirs for one full cycle. The hot reservoir has a loss of entropy

$$\Delta S_{\mathrm{h}} = \frac{-Q_{\mathrm{h}}}{T_{\mathrm{h}}}$$

, because heat transfer occurs out of it (remember that when heat transfers out, then *Q* has a negative sign). The cold reservoir has a gain of entropy

$$\Delta S_{\mathrm{c}} = \frac{Q_{\mathrm{c}}}{T_{\mathrm{c}}}$$

, because heat transfer occurs into it. (We assume the reservoirs are sufficiently large that their temperatures are constant.) So the total change in entropy is $\Delta S_{\mathrm{tot}} = \Delta S_{\mathrm{h}} + \Delta S_{\mathrm{c}}$.

Thus, since we know that

$$\frac{Q_{\mathrm{h}}}{T_{\mathrm{h}}} = \frac{Q_{\mathrm{c}}}{T_{\mathrm{c}}}$$

for a Carnot engine,

$$\Delta S_{\text{tot}} = \frac{Q_{\text{h}}}{T_{\text{h}}} = \frac{Q_{\text{c}}}{T_{\text{c}}} = 0$$

.

This result, which has general validity, means that *the total change in entropy for a system in any reversible process is zero.*

The entropy of various parts of the system may change, but the total change is zero. Furthermore, the system does not affect the entropy of its surroundings, since heat transfer between them does not occur. Thus the reversible process changes neither the total entropy of the system nor the entropy of its surroundings. Sometimes this is stated as follows: *Reversible processes do not affect the total entropy of the universe.* Real processes are not reversible, though, and they do change total entropy. We can, however, use hypothetical reversible processes to determine the value of entropy in real, irreversible processes. Example 1 illustrates this point.

Example 1. Entropy Increases in an Irreversible (Real) Process

Spontaneous heat transfer from hot to cold is an irreversible process. Calculate the total change in entropy if 4000 J of heat transfer occurs from a hot reservoir at T_{h} = 600 K(327ºC) to a cold reservoir at T_{c} = 250 K(−23ºC), assuming there is no temperature change in either reservoir. (See Figure 3.)

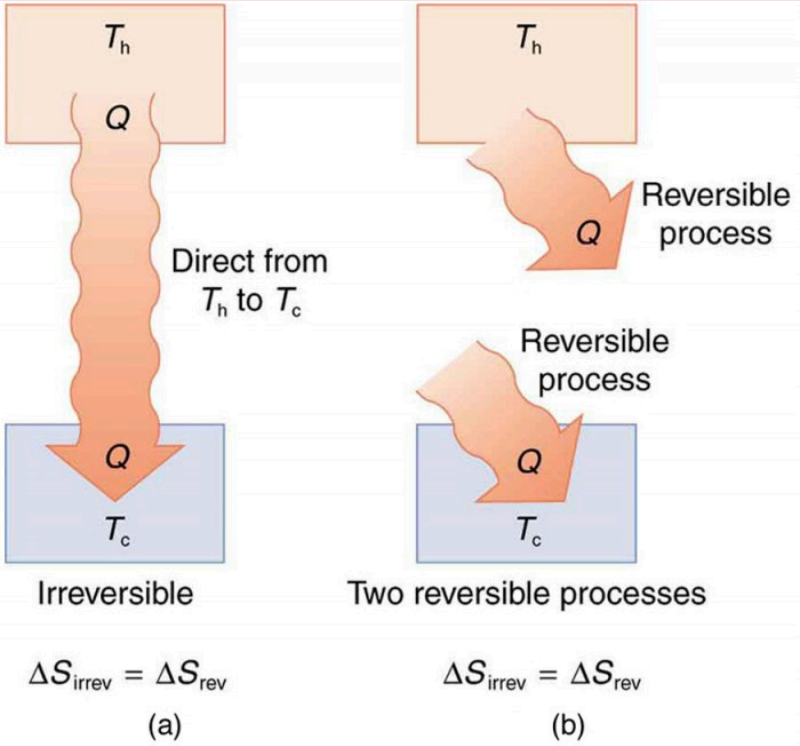


Figure 3. (a) Heat transfer from a hot object to a cold one is an irreversible process that produces an overall increase in entropy. (b) The same final state and, thus, the same change in entropy is achieved for the objects if reversible heat transfer processes occur between the two objects whose temperatures are the same as the temperatures of the corresponding objects in the irreversible process.

Strategy

How can we calculate the change in entropy for an irreversible process when $\Delta S_{tot} = \Delta S_h + \Delta S_c$ is valid only for reversible processes? Remember that the total change in entropy of the hot and cold reservoirs will be the same whether a reversible or irreversible process is involved in heat transfer from hot to cold. So we can calculate the change in entropy of the hot reservoir for a hypothetical reversible process in which 4000 J of heat transfer occurs from it; then we do the same for a hypothetical reversible process in which 4000 J of heat transfer occurs to the cold reservoir. This produces the same changes in the hot and cold reservoirs that would occur if the heat transfer were allowed to occur irreversibly between them, and so it also produces the same changes in entropy.

Solution

We now calculate the two changes in entropy using $\Delta S_{tot} = \Delta S_h + \Delta S_c$. First, for the heat transfer from the hot reservoir,

$$\Delta S_{\mathrm{h}} = \frac{-Q_{\mathrm{h}}}{T_{\mathrm{h}}} = \frac{-4000\ \mathrm{J}}{600\ \mathrm{K}} = -6.67\ \mathrm{J/K}$$

And for the cold reservoir,

$$\Delta S_{\mathrm{c}} = \frac{-Q_{\mathrm{c}}}{T_{\mathrm{c}}} = \frac{4000\ \mathrm{J}}{250\ \mathrm{K}} = 16.0\ \mathrm{J/K}$$

Thus the total is

$$\begin{aligned} \Delta S_{\text{tot}} &= \Delta S_{\text{h}} + \Delta S_{\text{c}} \\ &= (-6.67 + 16.0)\ \text{J/K} \\ &= 9.33\ \text{J/K} \end{aligned}$$

Discussion

There is an *increase* in entropy for the system of two heat reservoirs undergoing this irreversible heat transfer. We will see that this means there is a loss of ability to do work with this transferred energy. Entropy has increased, and energy has become unavailable to do work.

It is reasonable that entropy increases for heat transfer from hot to cold. Since the change in entropy is $\frac{Q}{T}$, there is a larger change at lower temperatures. The decrease in entropy of the hot object is therefore less than the increase in entropy of the cold object, producing an overall increase, just as in the previous example. This result is very general:

There is an increase in entropy for any system undergoing an irreversible process.

With respect to entropy, there are only two possibilities: entropy is constant for a reversible process, and it increases for an irreversible process. There is a fourth version of *the second law of thermodynamics stated in terms of entropy*:

The total entropy of a system either increases or remains constant in any process; it never decreases.

For example, heat transfer cannot occur spontaneously from cold to hot, because entropy would decrease.

Entropy is very different from energy. Entropy is *not* conserved but increases in all real processes. Reversible processes (such as in Carnot engines) are the processes in which the most heat transfer to work takes place and are also the ones that keep entropy constant. Thus we are led to make a connection between entropy and the availability of energy to do work.

Entropy and the Unavailability of Energy to Do Work

What does a change in entropy mean, and why should we be interested in it? One reason is that entropy is directly related to the fact that not all heat transfer can be converted into work. Example 2 gives some indication of how an increase in entropy results in less heat transfer into work.

Example 2. Less Work is Produced by a Given Heat Transfer When Entropy Change is Greater

1. Calculate the work output of a Carnot engine operating between temperatures of 600 K and 100 K for 4000 J of heat transfer to the engine.
2. Now suppose that the 4000 J of heat transfer occurs first from the 600 K reservoir to a 250 K reservoir (without doing any work, and this produces the increase in entropy calculated above) before transferring into a Carnot engine operating between 250 K and 100 K. What work output is produced? (See Figure 4.)

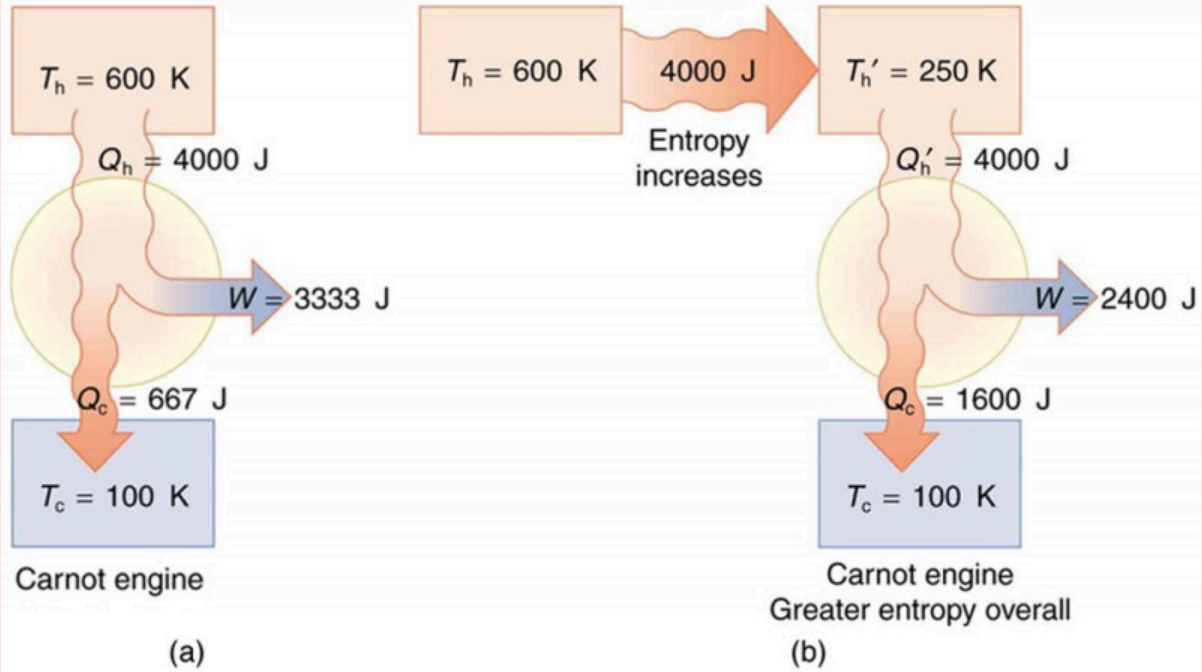


Figure 4. (a) A Carnot engine working at between 600 K and 100 K has 4000 J of heat transfer and performs 3333 J of work. (b) The 4000 J of heat transfer occurs first irreversibly to a 250 K reservoir and then goes into a Carnot engine. The increase in entropy caused by the heat transfer to a colder reservoir results in a smaller work output of 2400 J. There is a permanent loss of 933 J of energy for the purpose of doing work.

Strategy

In both parts, we must first calculate the Carnot efficiency and then the work output.

Solution to Part 1

The Carnot efficiency is given by

$$Eff_{\mathrm{C}} = 1 - \frac{T_{\mathrm{c}}}{T_{\mathrm{h}}}$$

.

Substituting the given temperatures yields

$$Eff_{\mathrm{C}} = 1 - \frac{100\ \mathrm{K}}{600\ \mathrm{K}} = 0.833$$

.

Now the work output can be calculated using the definition of efficiency for any heat engine as given by

$$Eff = \frac{W}{Q_{\mathrm{h}}}$$

.

Solving for *W* and substituting known terms gives

$$\begin{aligned} W &= Eff_{\mathrm{C}} Q_{\mathrm{h}} \\ &= (0.833)\,(4000\ \mathrm{J}) = 3333\ \mathrm{J} \end{aligned}$$

Solution to Part 2

Similarly,

$$Eff'_{\mathrm{C}} = 1 - \frac{T_{\mathrm{c}}}{T'_{\mathrm{c}}} = \frac{100\ \mathrm{K}}{250\ \mathrm{K}} = 0.600$$

so that

$$\begin{aligned} W &= Eff'_{\mathrm{C}} Q_{\mathrm{h}} \\ &= (0.600)\,(4000\ \mathrm{J}) = 2400\ \mathrm{J} \end{aligned}$$

Discussion

There is 933 J less work from the same heat transfer in the second process. This result is important. The same heat transfer into two perfect engines produces different work outputs, because the entropy change differs in the two cases. In the second case, entropy is greater and less work is produced. Entropy is associated with the *un*availability of energy to do work.

When entropy increases, a certain amount of energy becomes *permanently* unavailable to do work. The energy is not lost, but its character is changed, so that some of it can never be converted to doing work—that is, to an organized force acting through a distance. For instance, in Example 2, 933 J less work was done after an increase in entropy of 9.33 J/K occurred in the 4000 J heat transfer from the 600 K reservoir to the 250 K reservoir. It can be shown that the amount of energy that becomes unavailable for work is $W_{\mathrm{unavail}} = \Delta S \cdot T_0$, where T_0 is the lowest temperature utilized. In Example 2, W_{unavail} = (9.33 J/K)(100 K) = 933 J as found.

Heat Death of the Universe: An Overdose of Entropy

In the early, energetic universe, all matter and energy were easily interchangeable and identical in nature. Gravity played a vital role in the young universe. Although it may have *seemed* disorderly, and therefore, superficially entropic, in fact, there was enormous potential energy available to do work—all the future energy in the universe.

As the universe matured, temperature differences arose, which created more opportunity for work. Stars are hotter than planets, for example, which are warmer than icy asteroids, which are warmer still than the vacuum of the space between them.

Most of these are cooling down from their usually violent births, at which time they were provided with energy of their own—nuclear energy in the case of stars, volcanic energy on Earth and other planets, and so on. Without additional energy input, however, their days are numbered.

As entropy increases, less and less energy in the universe is available to do work. On Earth, we still have great stores of energy such as fossil and nuclear fuels; large-scale temperature differences, which can provide wind energy; geothermal energies due to differences in temperature in Earth's layers; and tidal energies owing to our abundance of liquid water. As these are used, a certain fraction of the energy they contain can never be converted into doing work. Eventually, all fuels will be exhausted, all temperatures will equalize, and it will be impossible for heat engines to function, or for work to be done.

Entropy increases in a closed system, such as the universe. But in parts of the universe, for instance, in the Solar system, it is not a locally closed system. Energy flows from the Sun to the planets, replenishing Earth's stores of energy. The Sun will continue to supply us with energy for about another five billion years. We will enjoy direct solar energy, as well as side effects of solar energy, such as wind power and biomass energy from photosynthetic plants. The energy from the Sun will keep our water at the liquid state, and the Moon's gravitational pull will continue to provide tidal energy. But Earth's geothermal energy will slowly run down and won't be replenished.

But in terms of the universe, and the very long-term, very large-scale picture, the entropy of the universe is increasing, and so the availability of energy to do work is constantly decreasing. Eventually, when all stars have died, all forms of potential energy have been utilized, and all temperatures have equalized (depending on the mass of the universe, either at a very high temperature following a universal contraction, or a very low one, just before all activity ceases) there will be no possibility of doing work.

Either way, the universe is destined for thermodynamic equilibrium—maximum entropy. This is often called the *heat death of the universe*, and will mean the end of all activity. However, whether the universe contracts and heats up, or continues to expand and cools down, the end is not near. Calculations of black holes suggest that entropy can easily continue for at least 10^{100} years.

Order to Disorder

Entropy is related not only to the unavailability of energy to do work—it is also a measure of disorder. This notion was initially postulated by Ludwig Boltzmann in the 1800s. For example, melting a block of ice means taking a highly structured and orderly system of water molecules and converting it into a disorderly liquid in which molecules have no fixed positions. (See Figure 5.) There is a large increase in entropy in the process, as seen in the following example.

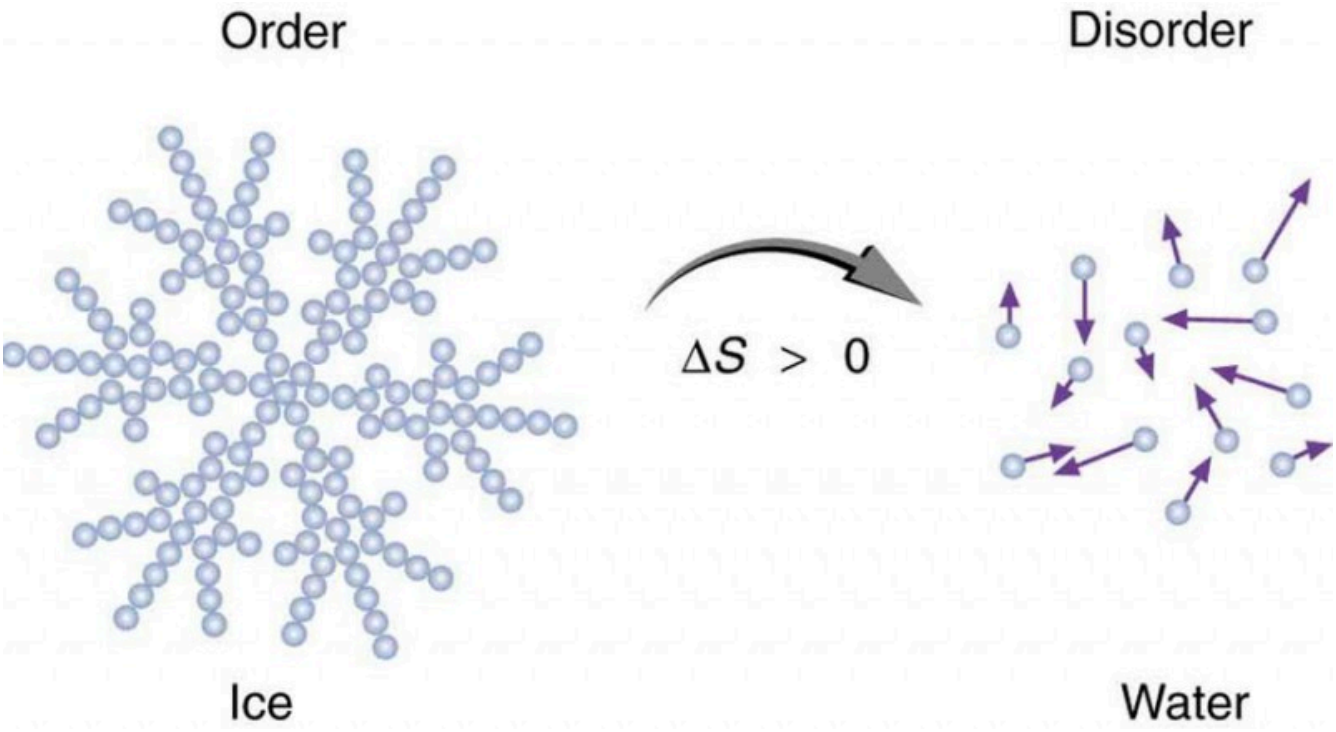


Figure 5. When ice melts, it becomes more disordered and less structured. The systematic arrangement of molecules in a crystal structure is replaced by a more random and less orderly movement of molecules without fixed locations or orientations. Its entropy increases because heat transfer occurs into it. Entropy is a measure of disorder.

Example 3. Entropy Associated with Disorder

Find the increase in entropy of 1.00 kg of ice originally at 0º C that is melted to form water at 0º C.

Strategy

As before, the change in entropy can be calculated from the definition of ΔS once we find the energy Q needed to melt the ice.

Solution

The change in entropy is defined as:

$$\Delta S = \frac{Q}{T}$$

.

Here Q is the heat transfer necessary to melt 1.00 kg of ice and is given by $Q = mL_f$, where m is the mass and L_f is the latent heat of fusion. L_f = 334 kJ/kg for water, so that Q = (1.00 kg)(334 kJ/kg) = 3.34×10^5 J.

Now the change in entropy is positive, since heat transfer occurs into the ice to cause the phase change; thus,

$$\Delta S = \frac{Q}{T} = \frac{3.34 \times 10^5 \text{ J}}{T}$$

T is the melting temperature of ice. That is, T = 0ºC = 273 K. So the change in entropy is

$$\begin{aligned} \Delta S &= \frac{3.34\times10^5\ \text{J}}{273\ \text{K}} \\ &= 1.22\times10^3\ \text{J/K} \end{aligned}$$

Discussion

This is a significant increase in entropy accompanying an increase in disorder.

In another easily imagined example, suppose we mix equal masses of water originally at two different temperatures, say 20.0ºC and 40.0ºC. The result is water at an intermediate temperature of 30.0ºC. Three outcomes have resulted: entropy has increased, some energy has become unavailable to do work, and the system has become less orderly. Let us think about each of these results.

First, entropy has increased for the same reason that it did in Example 3. Mixing the two bodies of water has the same effect as heat transfer from the hot one and the same heat transfer into the cold one. The mixing decreases the entropy of the hot water but increases the entropy of the cold water by a greater amount, producing an overall increase in entropy.

Second, once the two masses of water are mixed, there is only one temperature—you cannot run a heat engine with them. The energy that could have been used to run a heat engine is now unavailable to do work.

Third, the mixture is less orderly, or to use another term, less structured. Rather than having two masses at different temperatures and with different distributions of molecular speeds, we now have a single mass with a uniform temperature.

These three results—entropy, unavailability of energy, and disorder—are not only related but are in fact essentially equivalent.

Life, Evolution, and the Second Law of Thermodynamics

Some people misunderstand the second law of thermodynamics, stated in terms of entropy, to say that the process of the evolution of life violates this law. Over time, complex organisms evolved from much simpler ancestors, representing a large decrease in entropy of the Earth's biosphere. It is a fact that living organisms have evolved to be highly structured, and much lower in entropy than the substances from which they grow. But it is *always* possible for the entropy of one part of the universe to decrease, provided the total change in entropy of the universe increases. In equation form, we can write this as $\Delta S_{\text{tot}} = \Delta S_{\text{syst}} + \Delta S_{\text{envir}} > 0$.

Thus ΔS_{syst} can be negative as long as ΔS_{envir} is positive and greater in magnitude.

How is it possible for a system to decrease its entropy? Energy transfer is necessary. If I pick up marbles that are scattered about the room and put them into a cup, my work has decreased the entropy of that system. If I gather iron ore from the ground and convert it into steel and build a bridge, my work has decreased the entropy of that system. Energy coming from the Sun can decrease the entropy of local systems on Earth—that is, ΔS_{syst} is negative. But the overall entropy of the rest of the universe

increases by a greater amount—that is, ΔS_{envir} is positive and greater in magnitude. Thus, $\Delta S_{tot} = \Delta S_{syst} + \Delta S_{envir} > 0$, and the second law of thermodynamics is *not* violated.

Every time a plant stores some solar energy in the form of chemical potential energy, or an updraft of warm air lifts a soaring bird, the Earth can be viewed as a heat engine operating between a hot reservoir supplied by the Sun and a cold reservoir supplied by dark outer space—a heat engine of high complexity, causing local decreases in entropy as it uses part of the heat transfer from the Sun into deep space. There is a large total increase in entropy resulting from this massive heat transfer. A small part of this heat transfer is stored in structured systems on Earth, producing much smaller local decreases in entropy. (See Figure 6.)

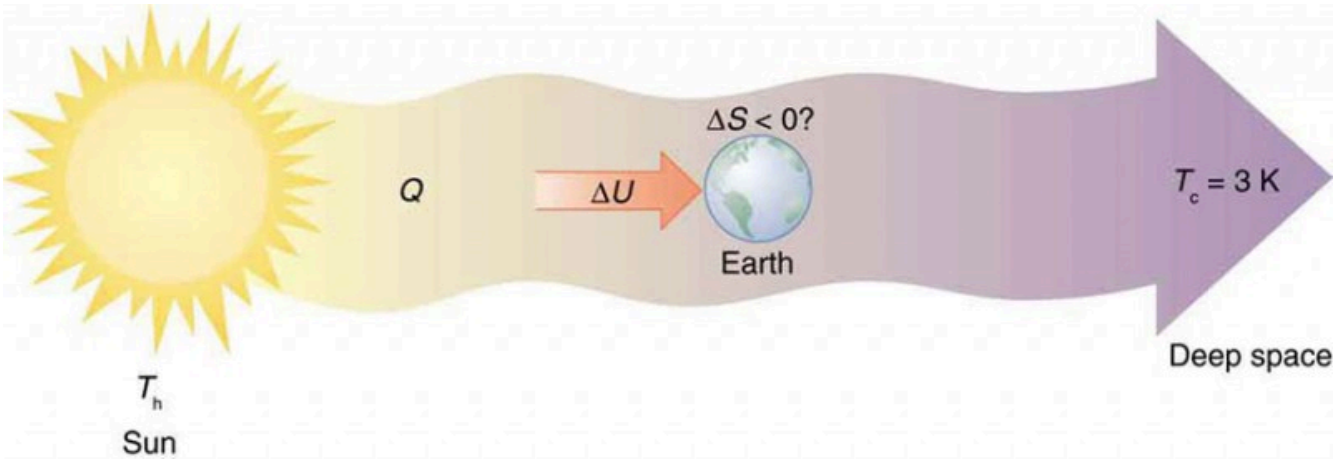
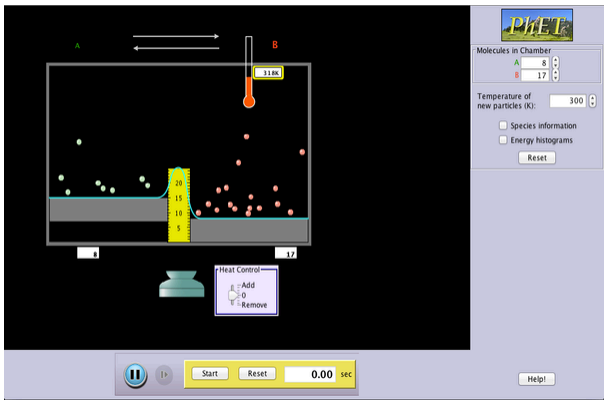


Figure 6. Earth's entropy may decrease in the process of intercepting a small part of the heat transfer from the Sun into deep space. Entropy for the entire process increases greatly while Earth becomes more structured with living systems and stored energy in various forms.

PhET Explorations: Reversible Reactions

Watch a reaction proceed over time. How does total energy affect a reaction rate? Vary temperature, barrier height, and potential energies. Record concentrations and time in order to extract rate coefficients. Do temperature dependent studies to extract Arrhenius parameters. This simulation is best used with teacher guidance because it presents an analogy of chemical reactions.



Click to download the simulation. Run using Java.

Section Summary

- Entropy is the loss of energy available to do work.
- Another form of the second law of thermodynamics states that the total entropy of a system either increases or remains constant; it never decreases.
- Entropy is zero in a reversible process; it increases in an irreversible process.
- The ultimate fate of the universe is likely to be thermodynamic equilibrium, where the universal temperature is constant and no energy is available to do work.
- Entropy is also associated with the tendency toward disorder in a closed system.

Conceptual Questions

1. A woman shuts her summer cottage up in September and returns in June. No one has entered the cottage in the meantime. Explain what she is likely to find, in terms of the second law of thermodynamics.
2. Consider a system with a certain energy content, from which we wish to extract as much work as possible. Should the system's entropy be high or low? Is this orderly or disorderly? Structured or uniform? Explain briefly.
3. Does a gas become more orderly when it liquefies? Does its entropy change? If so, does the entropy increase or decrease? Explain your answer.
4. Explain how water's entropy can decrease when it freezes without violating the second law of thermodynamics. Specifically, explain what happens to the entropy of its surroundings.
5. Is a uniform-temperature gas more or less orderly than one with several different temperatures? Which is more structured? In which can heat transfer result in work done without heat transfer from another system?

6. Give an example of a spontaneous process in which a system becomes less ordered and energy becomes less available to do work. What happens to the system's entropy in this process?
7. What is the change in entropy in an adiabatic process? Does this imply that adiabatic processes are reversible? Can a process be precisely adiabatic for a macroscopic system?
8. Does the entropy of a star increase or decrease as it radiates? Does the entropy of the space into which it radiates (which has a temperature of about 3 K) increase or decrease? What does this do to the entropy of the universe?
9. Explain why a building made of bricks has smaller entropy than the same bricks in a disorganized pile. Do this by considering the number of ways that each could be formed (the number of microstates in each macrostate).

Problems & Exercises

1. (a) On a winter day, a certain house loses 5.00×10^8 J of heat to the outside (about 500,000 Btu). What is the total change in entropy due to this heat transfer alone, assuming an average indoor temperature of 21.0ºC and an average outdoor temperature of 5.00ºC? (b) This large change in entropy implies a large amount of energy has become unavailable to do work. Where do we find more energy when such energy is lost to us?
2. On a hot summer day, 4.00×10^6 J of heat transfer into a parked car takes place, increasing its temperature from 35.0ºC to 45.0ºC. What is the increase in entropy of the car due to this heat transfer alone?
3. A hot rock ejected from a volcano's lava fountain cools from 1100ºC to 40.0ºC, and its entropy decreases by 950 J/K. How much heat transfer occurs from the rock?
4. When 1.60×10^5 J of heat transfer occurs into a meat pie initially at 20.0ºC, its entropy increases by 480 J/K. What is its final temperature?
5. The Sun radiates energy at the rate of 3.80×10^{26} W from its 5500ºC surface into dark empty space (a negligible fraction radiates onto Earth and the other planets). The effective temperature of deep space is −270ºC. (a) What is the increase in entropy in one day due to this heat transfer? (b) How much work is made unavailable?
6. (a) In reaching equilibrium, how much heat transfer occurs from 1.00 kg of water at 40.0ºC when it is placed in contact with 1.00 kg of 20.0ºC water in reaching equilibrium? (b) What is the change in entropy due to this heat transfer? (c) How much work is made unavailable, taking the lowest temperature to be 20.0ºC? Explicitly show how you follow the steps in the Problem-Solving Strategies for Entropy.
7. What is the decrease in entropy of 25.0 g of water that condenses on a bathroom mirror at a temperature of 35.0ºC, assuming no change in temperature and given the latent heat of vaporization to be 2450 kJ/kg?
8. Find the increase in entropy of 1.00 kg of liquid nitrogen that starts at its boiling temperature, boils, and warms to 20.0ºC at constant pressure.
9. A large electrical power station generates 1000 MW of electricity with an efficiency of 35.0%. (a) Calculate the heat transfer to the power station, Q_h, in one day. (b) How much heat transfer Q_c

occurs to the environment in one day? (c) If the heat transfer in the cooling towers is from 35.0ºC water into the local air mass, which increases in temperature from 18.0ºC to 20.0ºC, what is the total increase in entropy due to this heat transfer? (d) How much energy becomes unavailable to do work because of this increase in entropy, assuming an 18.0ºC lowest temperature? (Part of Q_c could be utilized to operate heat engines or for simply heating the surroundings, but it rarely is.)

10. (a) How much heat transfer occurs from 20.0 kg of 90.0ºC water placed in contact with 20.0 kg of 10.0ºC water, producing a final temperature of 50.0ºC? (b) How much work could a Carnot engine do with this heat transfer, assuming it operates between two reservoirs at constant temperatures of 90.0ºC and 10.0ºC? (c) What increase in entropy is produced by mixing 20.0 kg of 90.0ºC water with 20.0 kg of 10.0ºC water? (d) Calculate the amount of work made unavailable by this mixing using a low temperature of 10.0ºC, and compare it with the work done by the Carnot engine. Explicitly show how you follow the steps in the Problem-Solving Strategies for Entropy. (e) Discuss how everyday processes make increasingly more energy unavailable to do work, as implied by this problem.

Glossary

entropy: a measurement of a system's disorder and its inability to do work in a system

change in entropy: the ratio of heat transfer to temperature

$$\frac{Q}{T}$$

second law of thermodynamics stated in terms of entropy: the total entropy of a system either increases or remains constant; it never decreases

Selected Solutions to Problems & Exercises

1. (a) 9.78×10^4 J/K; (b) In order to gain more energy, we must generate it from things within the house, like a heat pump, human bodies, and other appliances. As you know, we use a lot of energy to keep our houses warm in the winter because of the loss of heat to the outside.

3. 8.01×10^5 J

5. (a) 1.04×10^{31} J/K; (b) 3.28×10^{31} J

7. 199 J/K

9. (a) 2.47×10^{14} J; (b) 1.60×10^{14} J; (c) 2.85×10^{10} J/K; (d) 8.29×10^{12} J

Statistical Interpretation of Entropy and the Second Law of Thermodynamics: The Underlying Explanation

Lumen Learning

Learning Objectives

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

- Identify probabilities in entropy.
- Analyze statistical probabilities in entropic systems.

The various ways of formulating the second law of thermodynamics tell what happens rather than why it happens. Why should heat transfer occur only from hot to cold? Why should energy become ever less available to do work? Why should the universe become increasingly disorderly? The answer is that it is a matter of overwhelming probability. Disorder is simply vastly more likely than order.

Figure 1. When you toss a coin a large number of times, heads and tails tend to come up in roughly equal numbers. Why doesn't heads come up 100, 90, or even 80% of the time? (credit: Jon Sullivan, PDPhoto.org)

When you watch an emerging rain storm begin to wet the ground, you will notice that the drops fall in a disorganized manner both in time and in space. Some fall close together, some far apart, but they never fall in straight, orderly rows. It is not impossible for rain to fall in an orderly pattern, just highly unlikely, because there are many more disorderly ways than orderly ones. To illustrate this fact, we will examine some random processes, starting with coin tosses.

Coin Tosses

What are the possible outcomes of tossing 5 coins? Each coin can land either heads or tails. On the large scale, we are concerned only with the total heads and tails and not with the order in which heads and tails appear. The following possibilities exist:

5 heads, 0 tails
4 heads, 1 tail
3 heads, 2 tails
2 heads, 3 tails
1 head, 4 tails
0 head, 5 tails

These are what we call macrostates. A *macrostate* is an overall property of a system. It does not specify the details of the system, such as the order in which heads and tails occur or which coins are heads or tails.

Using this nomenclature, a system of 5 coins has the 6 possible macrostates just listed. Some macrostates are more likely to occur than others. For instance, there is only one way to get 5 heads, but there are several ways to get 3 heads and 2 tails, making the latter macrostate more probable. Table 1 lists of all the ways in which 5 coins can be tossed, taking into account the order in which heads and tails occur. Each sequence is called a *microstate*—a detailed description of every element of a system.

Table 1. 5-Coin Toss

	Individual microstates	**Number of microstates**
5 heads, 0 tails	HHHHH	1
4 heads, 1 tail	HHHHT, HHHTH, HHTHH, HTHHH, THHHH	5
3 heads, 2 tails	HTHTH, THTHH, HTHHT, THHTH, THHHT HTHTH, THTHH, HTHHT, THHTH, THHHT	10
2 heads, 3 tails	TTTHH, TTHHT, THHTT, HHTTT, TTHTH, THTHT, HTHTT, THTTH, HTTHT, HTTTH	10
1 head, 4 tails	TTTTH, TTTHT, TTHTT, THTTT, HTTTT	5
0 heads, 5 tails	TTTTT	1
		Total: 32

The macrostate of 3 heads and 2 tails can be achieved in 10 ways and is thus 10 times more probable than the one having 5 heads. Not surprisingly, it is equally probable to have the reverse, 2 heads and 3 tails. Similarly, it is equally probable to get 5 tails as it is to get 5 heads. Note that all of these conclusions are based on the crucial assumption that each microstate is equally probable. With coin tosses, this requires that the coins not be asymmetric in a way that favors one side over the other, as with loaded dice. With any system, the assumption that all microstates are equally probable must be valid, or the analysis will be erroneous.

The two most orderly possibilities are 5 heads or 5 tails. (They are more structured than the others.) They are also the least likely, only 2 out of 32 possibilities. The most disorderly possibilities are 3 heads and 2 tails and its reverse. (They are the least structured.) The most disorderly possibilities are also the most likely, with 20 out of 32 possibilities for the 3 heads and 2 tails and its reverse. If we start with an orderly

array like 5 heads and toss the coins, it is very likely that we will get a less orderly array as a result, since 30 out of the 32 possibilities are less orderly. So even if you start with an orderly state, there is a strong tendency to go from order to disorder, from low entropy to high entropy. The reverse can happen, but it is unlikely.

Table 2. 100-Coin Toss

Macrostate		**Number of microstates**
Heads	Tails	(*W*)
100	0	1
99	1	1.0×10^{2}
95	5	7.5×10^{7}
90	10	1.7×10^{13}
75	25	2.4×10^{23}
60	40	1.4×10^{28}
55	45	6.1×10^{28}
51	49	9.9×10^{28}
50	50	1.0×10^{29}
49	51	9.9×10^{28}
45	55	6.1×10^{28}
40	60	1.4×10^{28}
25	75	2.4×10^{23}
10	90	1.7×10^{13}
5	95	7.5×10^{7}
1	99	1.0×10^{2}
0	100	1
		Total: 1.27×10^{30}

This result becomes dramatic for larger systems. Consider what happens if you have 100 coins instead of just 5. The most orderly arrangements (most structured) are 100 heads or 100 tails. The least orderly (least structured) is that of 50 heads and 50 tails. There is only 1 way (1 microstate) to get the most orderly arrangement of 100 heads. There are 100 ways (100 microstates) to get the next most orderly arrangement of 99 heads and 1 tail (also 100 to get its reverse). And there are 1.0×10^{29} ways to get 50 heads and 50 tails, the least orderly arrangement. Table 2 is an abbreviated list of the various macrostates and the number of microstates for each macrostate. The total number of microstates—the total number of different ways 100 coins can be tossed—is an impressively large 1.27×10^{30}. Now, if we start with an orderly macrostate like 100 heads and toss the coins, there is a virtual certainty that we will get a less

orderly macrostate. If we keep tossing the coins, it is possible, but exceedingly unlikely, that we will ever get back to the most orderly macrostate. If you tossed the coins once each second, you could expect to get either 100 heads or 100 tails once in 2×10^{22} years! This period is 1 trillion (10^{12}) times longer than the age of the universe, and so the chances are essentially zero. In contrast, there is an 8% chance of getting 50 heads, a 73% chance of getting from 45 to 55 heads, and a 96% chance of getting from 40 to 60 heads. Disorder is highly likely.

Disorder in a Gas

The fantastic growth in the odds favoring disorder that we see in going from 5 to 100 coins continues as the number of entities in the system increases. Let us now imagine applying this approach to perhaps a small sample of gas. Because counting microstates and macrostates involves statistics, this is called *statistical analysis*. The macrostates of a gas correspond to its macroscopic properties, such as volume, temperature, and pressure; and its microstates correspond to the detailed description of the positions and velocities of its atoms. Even a small amount of gas has a huge number of atoms: 1.0 cm^3 of an ideal gas at 1.0 atm and 0º C has 2.7×10^{19} atoms. So each macrostate has an immense number of microstates. In plain language, this means that there are an immense number of ways in which the atoms in a gas can be arranged, while still having the same pressure, temperature, and so on.

The most likely conditions (or macrostates) for a gas are those we see all the time—a random distribution of atoms in space with a Maxwell-Boltzmann distribution of speeds in random directions, as predicted by kinetic theory. This is the most disorderly and least structured condition we can imagine. In contrast, one type of very orderly and structured macrostate has all of the atoms in one corner of a container with identical velocities. There are very few ways to accomplish this (very few microstates corresponding to it), and so it is exceedingly unlikely ever to occur. (See Figure 2b.) Indeed, it is so unlikely that we have a law saying that it is impossible, which has never been observed to be violated—the second law of thermodynamics.

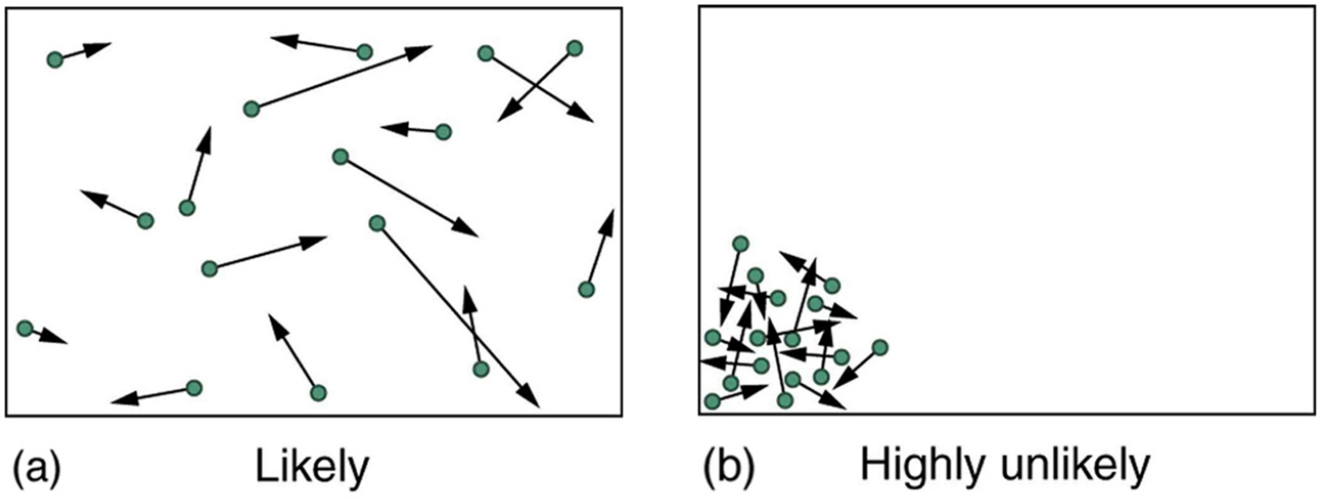


Figure 2. (a) The ordinary state of gas in a container is a disorderly, random distribution of atoms or molecules with a Maxwell-Boltzmann distribution of speeds. It is so unlikely that these atoms or molecules would ever end up in one corner of the container that it might as well be impossible. (b) With energy transfer, the gas can be forced into one corner and its entropy greatly reduced. But left alone, it will spontaneously increase its entropy and return to the normal conditions, because they are immensely more likely.

The disordered condition is one of high entropy, and the ordered one has low entropy. With a transfer of energy from another system, we could force all of the atoms into one corner and have a local decrease in entropy, but at the cost of an overall increase in entropy of the universe. If the atoms start out in one corner, they will quickly disperse and become uniformly distributed and will never return to the orderly original state (Figure 2b). Entropy will increase. With such a large sample of atoms, it is possible—but unimaginably unlikely—for entropy to decrease. Disorder is vastly more likely than order.

The arguments that disorder and high entropy are the most probable states are quite convincing. The great Austrian physicist Ludwig Boltzmann (1844–1906)—who, along with Maxwell, made so many contributions to kinetic theory—proved that the entropy of a system in a given state (a macrostate) can be written as $S = k\ \ln W$, where $k = 1.38 \times 10^{-23}$ J/K is Boltzmann's constant, and $\ln W$ is the natural logarithm of the number of microstates W corresponding to the given macrostate. W is proportional to the probability that the macrostate will occur. Thus entropy is directly related to the probability of a state—the more likely the state, the greater its entropy. Boltzmann proved that this expression for S is equivalent to the definition $\Delta S = \frac{Q}{T}$, which we have used extensively.

Thus the second law of thermodynamics is explained on a very basic level: entropy either remains the same or increases in every process. This phenomenon is due to the extraordinarily small probability of a decrease, based on the extraordinarily larger number of microstates in systems with greater entropy. Entropy *can* decrease, but for any macroscopic system, this outcome is so unlikely that it will never be observed.

Example 1. Entropy Increases in a Coin Toss

Suppose you toss 100 coins starting with 60 heads and 40 tails, and you get the most likely result, 50 heads and 50 tails. What is the change in entropy?

Strategy

Noting that the number of microstates is labeled W in Table 2 for the 100-coin toss, we can use $\Delta S = S_f - S_i = k\ \ln W_f - k \ln W_i$ to calculate the change in entropy.

Solution

The change in entropy is $\Delta S = S_f - S_i = k\ \ln W_f - k \ln W_i$,

where the subscript i stands for the initial 60 heads and 40 tails state, and the subscript f for the final 50 heads and 50 tails state. Substituting the values for W from Table 2 gives

$$\begin{aligned} \Delta S &= \left(1.38 \times 10^{-23}\ \text{J/K}\right)\left[\ln\left(1.0 \times 10^{29}\right) - \ln\left(1.4 \times 10^{29}\right)\right] \\ &= 2.7 \times 10^{-23}\ \text{J/K} \end{aligned}$$

Discussion

This increase in entropy means we have moved to a less orderly situation. It is not impossible for further tosses to produce the initial state of 60 heads and 40 tails, but it is less likely. There is about a 1 in 90 chance for that decrease in entropy (-2.7×10^{-23} J/K) to occur. If we calculate the decrease in entropy to move to the most orderly state, we get $\Delta S = -92 \times 10^{-23}$ J/K. There is about a 1 in 10^{30} chance of this change occurring.

So while very small decreases in entropy are unlikely, slightly greater decreases are impossibly unlikely. These probabilities imply, again, that for a macroscopic system, a decrease in entropy is impossible. For example, for heat transfer to occur spontaneously from 1.00 kg of 0ºC ice to its 0ºC environment, there would be a decrease in entropy of 1.22×10^3 J/K. Given that a ΔS 10^{-21} J/K corresponds to about a 1 in 10^{30} chance, a decrease of this size (10^3 J/K) is an *utter* impossibility. Even for a milligram of melted ice to spontaneously refreeze is impossible.

Problem-Solving Strategies for Entropy

1. *Examine the situation to determine if entropy is involved.*
2. *Identify the system of interest and draw a labeled diagram of the system showing energy flow.*
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 can be inferred from the problem as stated (identify the knowns).* You must carefully identify the heat transfer, if any, and the temperature at which the process takes place. It is also important to identify the initial and final states.
5. *Solve the appropriate equation for the quantity to be determined (the unknown).* Note that the change in entropy can be determined between any states by calculating it for a reversible process.
6. *Substitute the known value along with their units into the appropriate equation, and obtain numerical solutions complete with units.*
7. *To see if it is reasonable: Does it make sense?* For example, total entropy should increase for any real process or be constant for a reversible process. Disordered states should be more probable and have greater entropy than ordered states.

Section Summary

- Disorder is far more likely than order, which can be seen statistically.
- The entropy of a system in a given state (a macrostate) can be written as $S = k \ln W$, where $k = 1.38 \times 10^{-23}$ J/K is Boltzmann's constant, and $\ln W$ is the natural logarithm of the number of microstates W corresponding to the given macrostate.

Conceptual Questions

1. Explain why a building made of bricks has smaller entropy than the same bricks in a disorganized pile. Do this by considering the number of ways that each could be formed (the number of microstates in each macrostate).

Problems & Exercises

1. Using Table 2, verify the contention that if you toss 100 coins each second, you can expect to get 100 heads or 100 tails once in 2×10^{22} years; calculate the time to two-digit accuracy.
2. What percent of the time will you get something in the range from 60 heads and 40 tails through 40 heads and 60 tails when tossing 100 coins? The total number of microstates in that range is 1.22×10^{30}. (Consult Table 2.)
3. (a) If tossing 100 coins, how many ways (microstates) are there to get the three most likely macrostates of 49 heads and 51 tails, 50 heads and 50 tails, and 51 heads and 49 tails? (b) What percent of the total possibilities is this? (Consult Table 2.)
4. (a) What is the change in entropy if you start with 100 coins in the 45 heads and 55 tails macrostate, toss them, and get 51 heads and 49 tails? (b) What if you get 75 heads and 25 tails? (c) How much more likely is 51 heads and 49 tails than 75 heads and 25 tails? (d) Does either outcome violate the second law of thermodynamics?
5. (a) What is the change in entropy if you start with 10 coins in the 5 heads and 5 tails macrostate, toss them, and get 2 heads and 8 tails? (b) How much more likely is 5 heads and 5 tails than 2 heads and 8 tails? (Take the ratio of the number of microstates to find out.) (c) If you were betting on 2 heads and 8 tails would you accept odds of 252 to 45? Explain why or why not.

Table 3. 10-Coin Toss

Macrostate		**Number of Microstates**
Heads	Tails	(*W*)
10	0	1
9	1	10
8	2	45
7	3	120
6	4	210
5	5	252
4	6	210
3	7	120
2	8	45
1	9	10
0	10	1
		Total: 1024

6. (a) If you toss 10 coins, what percent of the time will you get the three most likely macrostates (6 heads and 4 tails, 5 heads and 5 tails, 4 heads and 6 tails)? (b) You can realistically toss 10 coins and count the number of heads and tails about twice a minute. At that rate, how long will it take on average to get either 10 heads and 0 tails or 0 heads and 10 tails?

7. (a) Construct a table showing the macrostates and all of the individual microstates for tossing 6 coins. (Use Table 3 as a guide.) (b) How many macrostates are there? (c) What is the total number of microstates? (d) What percent chance is there of tossing 5 heads and 1 tail? (e) How much more likely are you to toss 3 heads and 3 tails than 5 heads and 1 tail? (Take the ratio of the number of microstates to find out.)
8. In an air conditioner, 12.65 MJ of heat transfer occurs from a cold environment in 1.00 h. (a) What mass of ice melting would involve the same heat transfer? (b) How many hours of operation would be equivalent to melting 900 kg of ice? (c) If ice costs 20 cents per kg, do you think the air conditioner could be operated more cheaply than by simply using ice? Describe in detail how you evaluate the relative costs.

Glossary

macrostate: an overall property of a system

microstate: each sequence within a larger macrostate

statistical analysis: using statistics to examine data, such as counting microstates and macrostates

Selected Solutions to Problems & Exercises

1. It should happen twice in every 1.27×10^{30} s or once in every 6.35×10^{29} s

$$\left(6.35 \times 10^{29}\text{s}\right)\left(\frac{1\text{ h}}{3600\text{ s}}\right)\quad\left(\frac{1\text{ d}}{24\text{ h}}\right)\left(\frac{1\text{ y}}{365.25\text{ d}}\right)$$
$$= \qquad\qquad 2.0 \times 10^{22}\text{y}$$

3. (a) 3.0×10^{29}; (b) 24%

5. (a) -2.38×10^{-23} J/K; (b) 5.6 times more likely; (c) If you were betting on two heads and 8 tails, the odds of breaking even are 252 to 45, so on average you would break even. So, no, you wouldn't bet on odds of 252 to 45.

7. (b) 7; (c) 64; (d) 9.38%; (e) 3.33 times more likely (20 to 6)

6. Oscillatory Motion and Waves

Introduction to Oscillatory Motion and Waves

Lumen Learning

Figure 1. There are at least four types of waves in this picture—only the water waves are evident. There are also sound waves, light waves, and waves on the guitar strings. (credit: John Norton)

What do an ocean buoy, a child in a swing, the cone inside a speaker, a guitar, atoms in a crystal, the motion of chest cavities, and the beating of hearts all have in common? They all *oscillate*—that is, they move back and forth between two points. Many systems oscillate, and they have certain characteristics in common. All oscillations involve force and energy. You push a child in a swing to get the motion started. The energy of atoms vibrating in a crystal can be increased with heat. You put energy into a guitar string when you pluck it.

Some oscillations create *waves*. A guitar creates sound waves. You can make water waves in a swimming pool by slapping the water with your hand. You can no doubt think of other types of waves. Some, such as water waves, are visible. Some, such as sound waves, are not. But *every wave is a disturbance that moves from its source and carries energy*. Other examples of waves include earthquakes and visible light. Even subatomic particles, such as electrons, can behave like waves.

By studying oscillatory motion and waves, we shall find that a small number of underlying principles describe all of them and that wave phenomena are more common than you have ever imagined. We begin by studying the type of force that underlies the simplest oscillations and waves. We will then expand our exploration of oscillatory motion and waves to include concepts such as simple harmonic motion,

uniform circular motion, and damped harmonic motion. Finally, we will explore what happens when two or more waves share the same space, in the phenomena known as superposition and interference.

Hooke's Law: Stress and Strain Revisited

Lumen Learning

Learning Objectives

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

- Explain Newton's third law of motion with respect to stress and deformation.
- Describe the restoration of force and displacement.
- Calculate the energy in Hook's Law of deformation, and the stored energy in a string.

Newton's first law implies that an object oscillating back and forth is experiencing forces. Without force, the object would move in a straight line at a constant speed rather than oscillate. Consider, for example, plucking a plastic ruler to the left as shown in Figure 1. The deformation of the ruler creates a force in the opposite direction, known as a *restoring force*. Once released, the restoring force causes the ruler to move back toward its stable equilibrium position, where the net force on it is zero. However, by the time the ruler gets there, it gains momentum and continues to move to the right, producing the opposite deformation. It is then forced to the left, back through equilibrium, and the process is repeated until dissipative forces dampen the motion. These forces remove mechanical energy from the system, gradually reducing the motion until the ruler comes to rest.

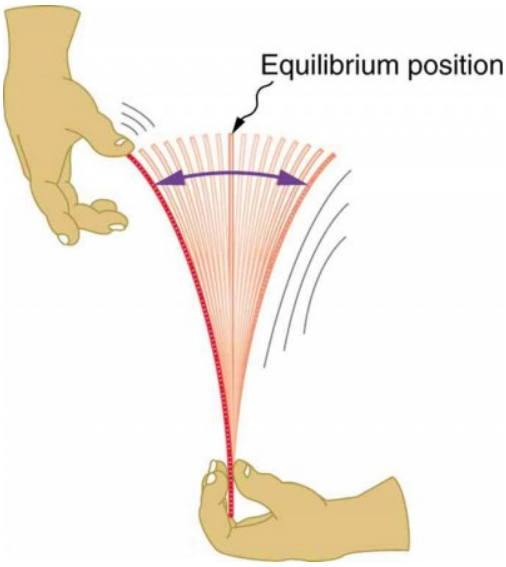


Figure 1. When displaced from its vertical equilibrium position, this plastic ruler oscillates back and forth because of the restoring force opposing displacement. When the ruler is on the left, there is a force to the right, and vice versa.

The simplest oscillations occur when the restoring force is directly proportional to displacement. When stress and strain were covered in Newton's Third Law of Motion, the name was given to this relationship between force and displacement was Hooke's law: $F = -kx$.

Here, F is the restoring force, x is the displacement from equilibrium or *deformation*, and k is a constant related to the difficulty in deforming the system. The minus sign indicates the restoring force is in the direction opposite to the displacement.

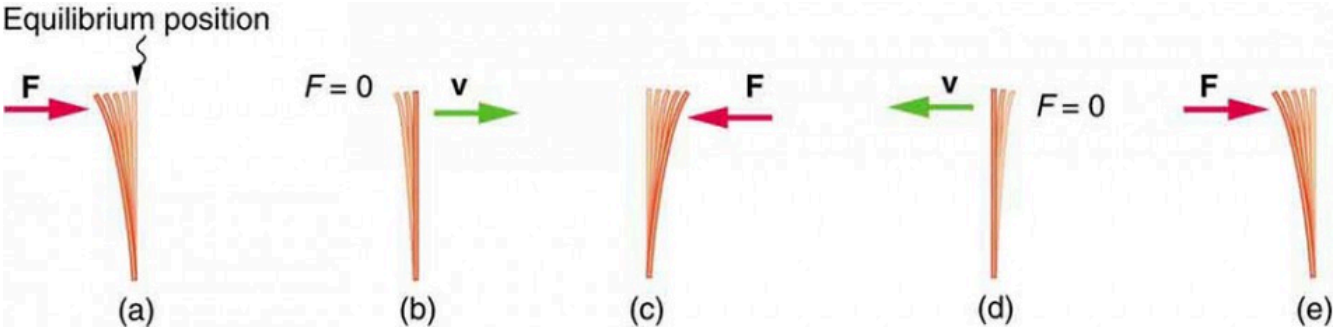
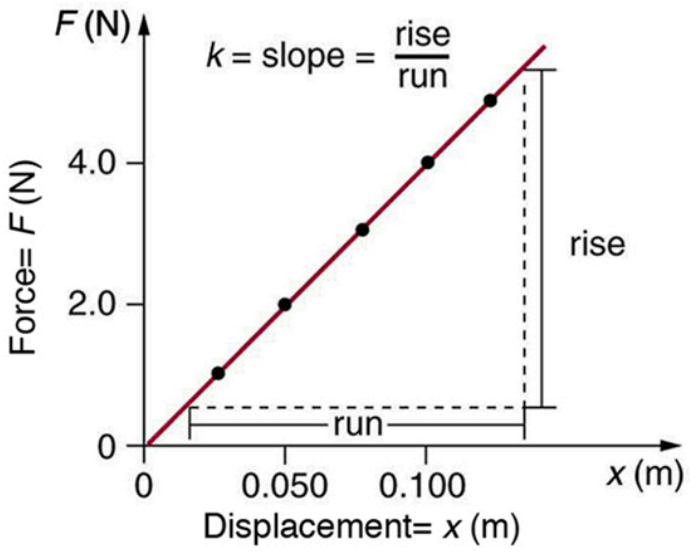


Figure 2. (a) The plastic ruler has been released, and the restoring force is returning the ruler to its equilibrium position. (b) The net force is zero at the equilibrium position, but the ruler has momentum and continues to move to the right. (c) The restoring force is in the opposite direction. It stops the ruler and moves it back toward equilibrium again. (d) Now the ruler has momentum to the left. (e) In the absence of damping (caused by frictional forces), the ruler reaches its original position. From there, the motion will repeat itself.

The *force constant k* is related to the rigidity (or stiffness) of a system—the larger the force constant, the greater the restoring force, and the stiffer the system. The units of *k* are newtons per meter (N/m). For example, *k* is directly related to Young's modulus when we stretch a string. Figure 3 shows a graph of the absolute value of the restoring force versus the displacement for a system that can be described by Hooke's law—a simple spring in this case. The slope of the graph equals the force constant *k* in newtons per meter. A common physics laboratory exercise is to measure restoring forces created by springs, determine if they follow Hooke's law, and calculate their force constants if they do.

a)



m (kg)	w (N)	x (m)
0.000	0.00	0.000
0.100	0.98	0.025
0.200	1.96	0.050
0.300	2.94	0.076
0.400	3.92	0.099
0.500	4.90	0.127

b)

m (kg)	w (N)	x (m)
0.000	0.00	0.000
0.100	0.98	0.025
0.200	1.96	0.050
0.300	2.94	0.076
0.400	3.92	0.099
0.500	4.90	0.127

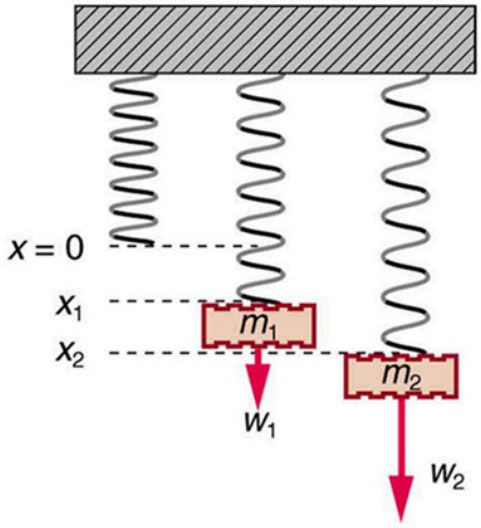


Figure 3. (a) A graph of absolute value of the restoring force versus displacement is displayed. The fact that the graph is a straight line means that the system obeys Hooke's law. The slope of the graph is the force constant k. (b) The data in the graph were generated by measuring the displacement of a spring from equilibrium while supporting various weights. The restoring force equals the weight supported, if the mass is stationary.

Example 1. How Stiff Are Car Springs?

What is the force constant for the suspension system of a car that settles 1.20 cm when an 80.0-kg person gets in?

Strategy

Consider the car to be in its equilibrium position x=0 before the person gets in. The car then settles down 1.20 cm, which means it is displaced to a position $x = -1.20 \times 10^{-2}$ m. At that point, the springs supply a restoring force F equal to the person's weight $w = mg$ = (80.0 kg)(9.80 m/s^2) = 784 N. We take this force to be F in Hooke's law. Knowing F and x, we can then solve the force constant k.

Solution

Solve Hooke's law, $F = -kx$, for k:

$$k = -\frac{F}{x}$$

Substitute known values and solve k:

$$\begin{aligned} k &= -\frac{784 \text{ N}}{-1.20 \times 10^{-2} \text{ m}} \\ &= 6.53 \times 10^4 \text{ N/m} \end{aligned}$$

Figure 4. The mass of a car increases due to the introduction of a passenger. This affects the displacement of the car on its suspension system. (credit: exfordy on Flickr)

Discussion

Note that F and x have opposite signs because they are in opposite directions—the restoring force is up, and the displacement is down. Also, note that the car would oscillate up and down when the person got in if it were not for damping (due to frictional forces) provided by shock absorbers. Bouncing cars are a sure sign of bad shock absorbers.

Energy in Hooke's Law of Deformation

In order to produce a deformation, work must be done. That is, a force must be exerted through a distance, whether you pluck a guitar string or compress a car spring. If the only result is deformation, and no work goes into thermal, sound, or kinetic energy, then all the work is initially stored in the deformed object as some form of potential energy. The potential energy stored in a spring is

$$\text{PE}_{\text{el}} = \tfrac{1}{2}kx^2$$

. Here, we generalize the idea to elastic potential energy for a deformation of any system that can be described by Hooke's law. Hence,

$$\text{PE}_{\text{el}} = \tfrac{1}{2}kx^2$$

, where PE_{el} is the *elastic potential energy* stored in any deformed system that obeys Hooke's law and has a displacement *x* from equilibrium and a force constant *k*.

It is possible to find the work done in deforming a system in order to find the energy stored. This work is performed by an applied force F_{app}. The applied force is exactly opposite to the restoring force (action-reaction), and so $F_{app} = kx$. Figure 5 shows a graph of the applied force versus deformation *x* for a system that can be described by Hooke's law. Work done on the system is force multiplied by distance, which equals the area under the curve or

$$\tfrac{1}{2}kx^2$$

(Method A in Figure 5). Another way to determine the work is to note that the force increases linearly from 0 to *kx*, so that the average force is

$$\tfrac{1}{2}kx$$

$$W = F_{\text{app}}d = \left(\tfrac{1}{2}kx\right)(x) = \tfrac{1}{2}kx^2$$

, the distance moved is *x*, and thus (Method B in Figure 5).

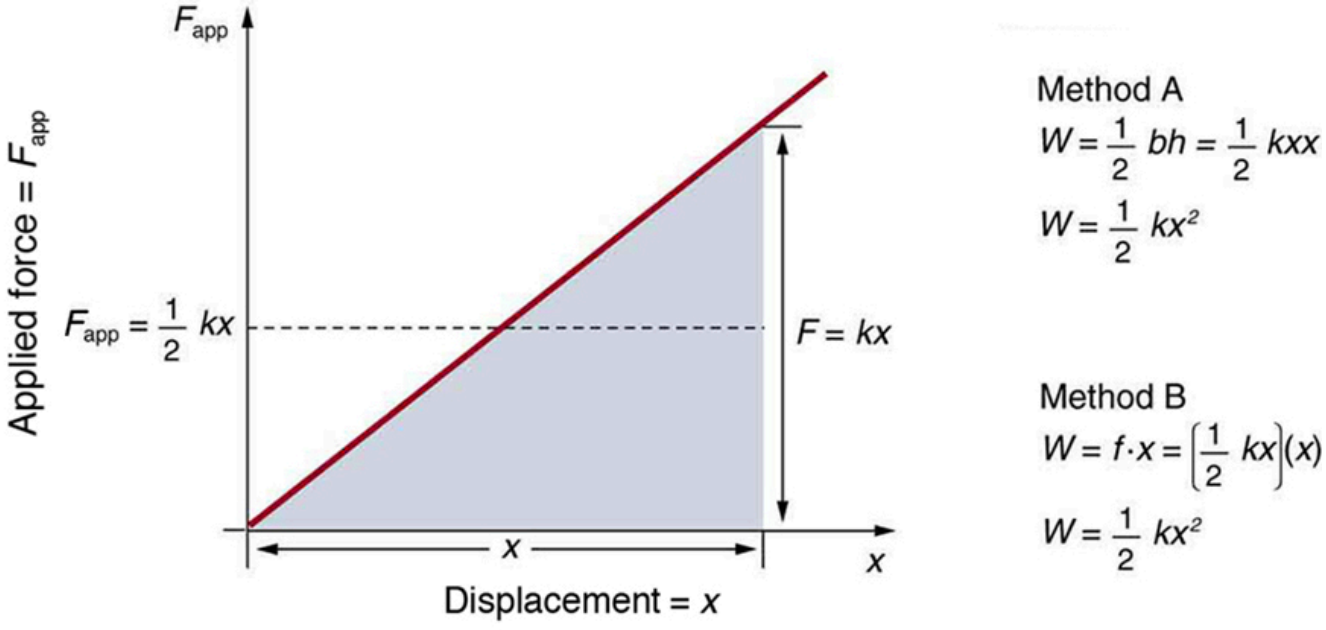


Figure 5. A graph of applied force versus distance for the deformation of a system that can be described by Hooke's law is displayed. The work done on the system equals the area under the graph or the area of the triangle, which is

$$W = \tfrac{1}{2}kx^2$$

half its base multiplied by its height, or .

Example 2. Calculating Stored Energy: A Tranquilizer Gun Spring

We can use a toy gun's spring mechanism to ask and answer two simple questions:

1. How much energy is stored in the spring of a tranquilizer gun that has a force constant of 50.0 N/m and is compressed 0.150 m?
2. If you neglect friction and the mass of the spring, at what speed will a 2.00-g projectile be ejected from the gun?

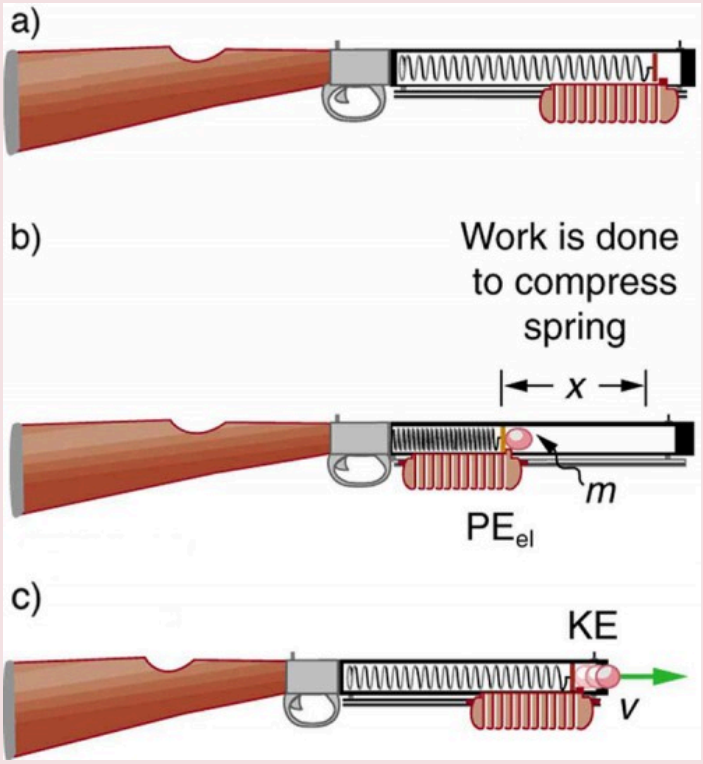


Figure 6. (a) In this image of the gun, the spring is uncompressed before being cocked. (b) The spring has been compressed a distance x, and the projectile is in place. (c) When released, the spring converts elastic potential energy PE_{el} into kinetic energy.

Strategy for Part 1

The energy stored in the spring can be found directly from elastic potential energy equation, because *k* and *x* are given.

Solution for Part 1

Entering the given values for *k* and *x* yields

$$\begin{aligned} \mathrm{PE_{el}} &= \tfrac{1}{2}kx^2 = \tfrac{1}{2}\,(50.0\ \mathrm{N/m})\,(0.150\ \mathrm{m})^2 = 0.563\ \mathrm{N}\cdot\ \mathrm{m} \\ &= 0.563\ \mathrm{J} \end{aligned}$$

Strategy for Part 2

Because there is no friction, the potential energy is converted entirely into kinetic energy. The expression for kinetic energy can be solved for the projectile's speed.

Solution for Part 2

Identify known quantities:

$$\text{KE}_\text{f} = \text{PE}_\text{el} \text{ or}$$
$$\tfrac{1}{2}mv^2 = \tfrac{1}{2}kx^2 = \text{PE}_\text{el} = 0.563 \text{ J}$$

Solve for *v*:

$$v = \left[\frac{2\text{PE}_\text{el}}{m}\right]^{1/2} = \left[\frac{2\,(0.563 \text{ J})}{0.002 \text{ kg}}\right]^{1/2} = 23.7\,(\text{J/kg})^{1/2}$$

Convert units: 23.7 m/s

Discussion

Parts 1 and 2: This projectile speed is impressive for a tranquilizer gun (more than 80 km/h). The numbers in this problem seem reasonable. The force needed to compress the spring is small enough for an adult to manage, and the energy imparted to the dart is small enough to limit the damage it might do. Yet, the speed of the dart is great enough for it to travel an acceptable distance.

Check your Understanding

Part 1

Envision holding the end of a ruler with one hand and deforming it with the other. When you let go, you can see the oscillations of the ruler. In what way could you modify this simple experiment to increase the rigidity of the system?

You could hold the ruler at its midpoint so that the part of the ruler that oscillates is half as long as in the original experiment.

Part 2

If you apply a deforming force on an object and let it come to equilibrium, what happened to the work you did on the system?

It was stored in the object as potential energy.

Section Summary

- An oscillation is a back and forth motion of an object between two points of deformation.
- An oscillation may create a wave, which is a disturbance that propagates from where it was created.
- The simplest type of oscillations and waves are related to systems that can be described by

Hooke's law: $F = -kx$,
where F is the restoring force, x is the displacement from equilibrium or deformation, and k is the force constant of the system.

- Elastic potential energy PE_{el} stored in the deformation of a system that can be described by
$$\mathrm{PE}_{\mathrm{el}} = \tfrac{1}{2}kx^2$$
Hooke's law is given by .

Conceptual Questions

1. Describe a system in which elastic potential energy is stored.

Problems & Exercises

1. Fish are hung on a spring scale to determine their mass (most fishermen feel no obligation to truthfully report the mass). (a) What is the force constant of the spring in such a scale if it the spring stretches 8.00 cm for a 10.0 kg load? (b) What is the mass of a fish that stretches the spring 5.50 cm? (c) How far apart are the half-kilogram marks on the scale?
2. It is weigh-in time for the local under-85-kg rugby team. The bathroom scale used to assess eligibility can be described by Hooke's law and is depressed 0.75 cm by its maximum load of 120 kg. (a) What is the spring's effective spring constant? (b) A player stands on the scales and depresses it by 0.48 cm. Is he eligible to play on this under-85 kg team?
3. One type of BB gun uses a spring-driven plunger to blow the BB from its barrel. (a) Calculate the force constant of its plunger's spring if you must compress it 0.150 m to drive the 0.0500-kg plunger to a top speed of 20.0 m/s. (b) What force must be exerted to compress the spring?
4. (a) The springs of a pickup truck act like a single spring with a force constant of 1.30×10^5 N/m. By how much will the truck be depressed by its maximum load of 1000 kg? (b) If the pickup truck has four identical springs, what is the force constant of each?
5. When an 80.0-kg man stands on a pogo stick, the spring is compressed 0.120 m. (a) What is the force constant of the spring? (b) Will the spring be compressed more when he hops down the road?
6. A spring has a length of 0.200 m when a 0.300-kg mass hangs from it, and a length of 0.750 m when a 1.95-kg mass hangs from it. (a) What is the force constant of the spring? (b) What is the unloaded length of the spring?

Glossary

deformation: displacement from equilibrium

elastic potential energy: potential energy stored as a result of deformation of an elastic object, such as the stretching of a spring

force constant: a constant related to the rigidity of a system: the larger the force constant, the more rigid the system; the force constant is represented by k

restoring force: force acting in opposition to the force caused by a deformation

Selected Solutions to Problems & Exercises

1. (a) 1.23×10^3 N/m; (b) 6.88 kg; (c) 4.00 mm
3. (a) 889 N/m; (b) 133 N
5. (a) 6.53×10^3 N/m; (b) Yes

Period and Frequency in Oscillations

Lumen Learning

Learning Objectives

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

- Observe the vibrations of a guitar string.
- Determine the frequency of oscillations.

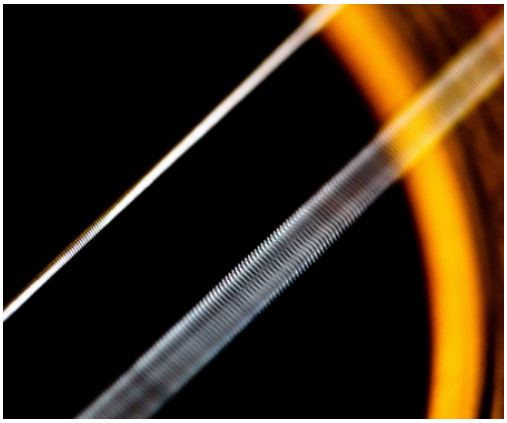


Figure 1. The strings on this guitar vibrate at regular time intervals. (credit: JAR)

When you pluck a guitar string, the resulting sound has a steady tone and lasts a long time. Each successive vibration of the string takes the same time as the previous one. We define *periodic motion* to be a motion that repeats itself at regular time intervals, such as exhibited by the guitar string or by an object on a spring moving up and down. The time to complete one oscillation remains constant and is called the *period T*. Its units are usually seconds, but may be any convenient unit of time. The word period refers to the time for some event whether repetitive or not; but we shall be primarily interested in periodic motion, which is by definition repetitive. A concept closely related to period is the frequency of an event. For example, if you get a paycheck twice a month, the frequency of payment is two per month and the period between checks is half a month. *Frequency f* is defined to be the number of events per unit time. For periodic motion, frequency is the number of oscillations per unit time. The relationship between frequency and period is

$$f = \frac{1}{T}$$

.

The SI unit for frequency is the *cycle per second*, which is defined to be a *hertz* (Hz):

$$1 \text{ Hz} = 1\frac{\text{cycle}}{\text{sec}} \text{ or } 1 \text{ Hz} = \frac{1}{\text{s}}$$

A cycle is one complete oscillation. Note that a vibration can be a single or multiple event, whereas oscillations are usually repetitive for a significant number of cycles.

Example 1. Determine the Frequency of Two Oscillations: Medical Ultrasound and the Period of Middle C

We can use the formulas presented in this module to determine both the frequency based on known oscillations and the oscillation based on a known frequency. Let's try one example of each.

1. A medical imaging device produces ultrasound by oscillating with a period of 0.400 μs. What is the frequency of this oscillation?
2. The frequency of middle C on a typical musical instrument is 264 Hz. What is the time for one complete oscillation?

Strategy

Both Parts 1 and 2 can be answered using the relationship between period and frequency. In Part 1, the period *T* is given and we are asked to find frequency *f*. In Part 2, the frequency *f* is given and we are asked to find the period *T*.

Solution for Part 1

Substitute 0.400 μs for *T* in

$$f = \frac{1}{T}$$

:

$$f = \frac{1}{T} = \frac{1}{0.400 \times 10^{-6}\ \text{s}}$$

Solve to find $f = 2.50 \times 10^6$ Hz.

Discussion for Part 1

The frequency of sound found in Part 1 is much higher than the highest frequency that humans can hear and, therefore, is called ultrasound. Appropriate oscillations at this frequency generate ultrasound used for noninvasive medical diagnoses, such as observations of a fetus in the womb.

Solution for Part 2

Identify the known values:The time for one complete oscillation is the period *T*:

$$f = \frac{1}{T}$$

.

Solve for *T*:

$$T = \frac{1}{f}$$

.

Substitute the given value for the frequency into the resulting expression:

$$T = \frac{1}{f} = \frac{1}{264\ \text{Hz}} = \frac{1}{264\ \text{cycles/s}} = 3.79 \times 10^{-3}\ \text{s} = 3.79\ \text{ms}$$

Discussion for Part 2

The period found in Part 2 is the time per cycle, but this value is often quoted as simply the time in convenient units (ms or milliseconds in this case).

Check your Understanding

Identify an event in your life (such as receiving a paycheck) that occurs regularly. Identify both the period and frequency of this event.

Solution

I visit my parents for dinner every other Sunday. The frequency of my visits is 26 per calendar year. The period is two weeks.

Section Summary

- Periodic motion is a repetitious oscillation.
- The time for one oscillation is the period T.
- The number of oscillations per unit time is the frequency f.

$$f = \frac{1}{T}$$

- These quantities are related by .

Problems & Exercises

1. What is the period of 60.0 Hz electrical power?
2. If your heart rate is 150 beats per minute during strenuous exercise, what is the time per beat in units of seconds?
3. Find the frequency of a tuning fork that takes 2.50×10^{-3} s to complete one oscillation.
4. A stroboscope is set to flash every 8.00×10^{-5} s. What is the frequency of the flashes?
5. A tire has a tread pattern with a crevice every 2.00 cm. Each crevice makes a single vibration as the tire moves. What is the frequency of these vibrations if the car moves at 30.0 m/s?
6. **Engineering Application.** Each piston of an engine makes a sharp sound every other revolution of the engine. (a) How fast is a race car going if its eight-cylinder engine emits a sound of frequency 750 Hz, given that the engine makes 2000 revolutions per kilometer? (b) At how many revolutions per minute is the engine rotating?

Glossary

period: time it takes to complete one oscillation

periodic motion: motion that repeats itself at regular time intervals

frequency: number of events per unit of time

Selected Solutions to Problems & Exercises

1. 16.7 ms
2. 0.400 s/beats
3. 400 Hz
4. 12,500 Hz
5. 1.50 kHz
6. (a) 93.8 m/s; (b) 11.3×10^3 rev/min

Simple Harmonic Motion: A Special Periodic Motion

Lumen Learning

Learning Objectives

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

- Describe a simple harmonic oscillator.
- Explain the link between simple harmonic motion and waves.

The oscillations of a system in which the net force can be described by Hooke's law are of special importance, because they are very common. They are also the simplest oscillatory systems. *Simple Harmonic Motion* (SHM) is the name given to oscillatory motion for a system where the net force can be described by Hooke's law, and such a system is called a *simple harmonic oscillator*. If the net force can be described by Hooke's law and there is no *damping* (by friction or other non-conservative forces), then a simple harmonic oscillator will oscillate with equal displacement on either side of the equilibrium position, as shown for an object on a spring in Figure 1. The maximum displacement from equilibrium is called the *amplitude X*. The units for amplitude and displacement are the same, but depend on the type of oscillation. For the object on the spring, the units of amplitude and displacement are meters; whereas for sound oscillations, they have units of pressure (and other types of oscillations have yet other units). Because amplitude is the maximum displacement, it is related to the energy in the oscillation.

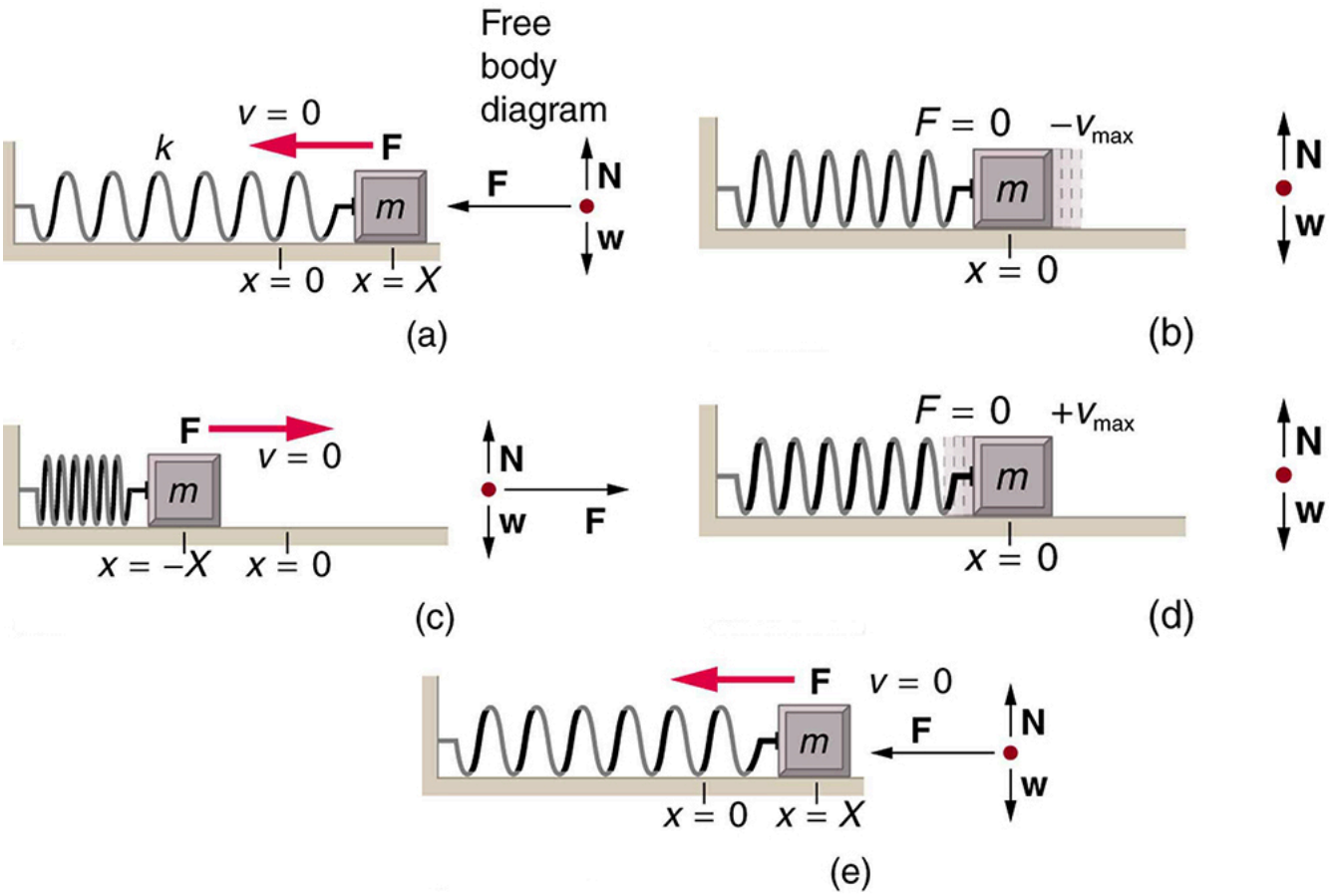


Figure 1. An object attached to a spring sliding on a frictionless surface is an uncomplicated simple harmonic oscillator. When displaced from equilibrium, the object performs simple harmonic motion that has an amplitude X and a period T. The object's maximum speed occurs as it passes through equilibrium. The stiffer the spring is, the smaller the period T. The greater the mass of the object is, the greater the period T.

Take-Home Experiment: SHM and the Marble

Find a bowl or basin that is shaped like a hemisphere on the inside. Place a marble inside the bowl and tilt the bowl periodically so the marble rolls from the bottom of the bowl to equally high points on the sides of the bowl. Get a feel for the force required to maintain this periodic motion. What is the restoring force and what role does the force you apply play in the simple harmonic motion (SHM) of the marble?

What is so significant about simple harmonic motion? One special thing is that the period T and frequency f of a simple harmonic oscillator are independent of amplitude. The string of a guitar, for example, will oscillate with the same frequency whether plucked gently or hard. Because the period is constant, a simple harmonic oscillator can be used as a clock.

Two important factors do affect the period of a simple harmonic oscillator. The period is related to how stiff the system is. A very stiff object has a large force constant k, which causes the system to have a smaller period. For example, you can adjust a diving board's stiffness—the stiffer it is, the faster it vibrates, and the shorter its period. Period also depends on the mass of the oscillating system. The more

massive the system is, the longer the period. For example, a heavy person on a diving board bounces up and down more slowly than a light one.

In fact, the mass *m* and the force constant *k* are the *only* factors that affect the period and frequency of simple harmonic motion.

Period of Simple Harmonic Oscillator

The *period of a simple harmonic oscillator* is given by

$$T = 2\pi\sqrt{\frac{m}{k}}$$

and, because

$$f = \frac{1}{T}$$

, the *frequency of a simple harmonic oscillator* is

$$f = \frac{1}{2\pi}\sqrt{\frac{k}{m}}$$

.

Note that neither *T* nor *f* has any dependence on amplitude.

Take-Home Experiment: Mass and Ruler Oscillations

Find two identical wooden or plastic rulers. Tape one end of each ruler firmly to the edge of a table so that the length of each ruler that protrudes from the table is the same. On the free end of one ruler tape a heavy object such as a few large coins. Pluck the ends of the rulers at the same time and observe which one undergoes more cycles in a time period, and measure the period of oscillation of each of the rulers.

Example 1. Calculate the Frequency and Period of Oscillations: Bad Shock Absorbers in a Car

If the shock absorbers in a car go bad, then the car will oscillate at the least provocation, such as when going over bumps in the road and after stopping (See Figure 2). Calculate the frequency and period of these oscillations for such a car if the car's mass (including its load) is 900 kg and the force constant (*k*) of the suspension system is 6.53×10^4 N/m.

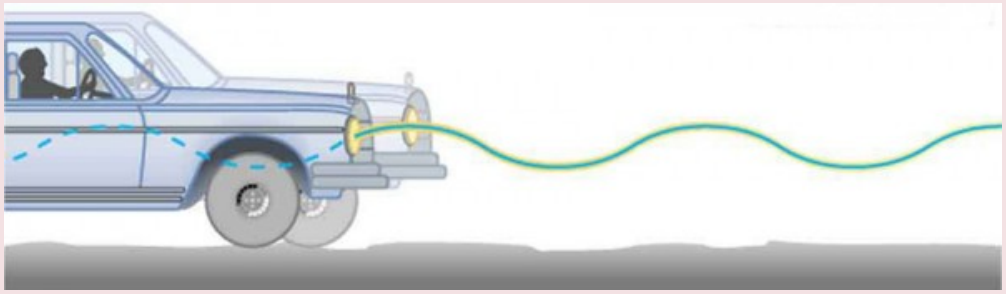


Figure 2. The bouncing car makes a wavelike motion. If the restoring force in the suspension system can be described only by Hooke's law, then the wave is a sine function. (The wave is the trace produced by the headlight as the car moves to the right.)

Strategy

The frequency of the car's oscillations will be that of a simple harmonic oscillator as given in the equation

$$f = \frac{1}{2\pi}\sqrt{\frac{k}{m}}$$

. The mass and the force constant are both given.

Solution

Enter the known values of *k* and *m*:

$$f = \frac{1}{2\pi}\sqrt{\frac{k}{m}} = \frac{1}{2\pi}\sqrt{\frac{6.53 \times 10^4 \text{ N/m}}{900 \text{ kg}}}$$

Calculate the frequency:

$$\frac{1}{2\pi}\sqrt{72.6/\text{s}^{-2}} = 1.3656/\text{s}^{-1} \approx 1.36/\text{s}^{-1} = 1.36 \text{ Hz}$$

You could use

$$T = 2\pi\sqrt{\frac{m}{k}}$$

to calculate the period, but it is simpler to use the relationship

$$T = \frac{1}{f}$$

and substitute the value just found for *f*:

$$T = \frac{1}{f} = \frac{1}{1.356 \text{ Hz}} = 0.738 \text{ s}$$

Discussion

The values of *T* and *f* both seem about right for a bouncing car. You can observe these oscillations if you push down hard on the end of a car and let go.

The Link between Simple Harmonic Motion and Waves

If a time-exposure photograph of the bouncing car were taken as it drove by, the headlight would make a wavelike streak, as shown in Figure 2. Similarly, Figure 3 shows an object bouncing on a spring as it leaves a wavelike "trace of its position on a moving strip of paper. Both waves are sine functions. All simple harmonic motion is intimately related to sine and cosine waves.

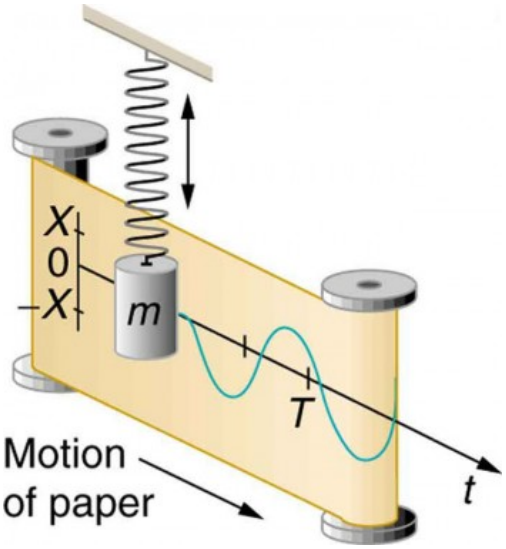


Figure 3. The vertical position of an object bouncing on a spring is recorded on a strip of moving paper, leaving a sine wave.

The displacement as a function of time *t* in any simple harmonic motion—that is, one in which the net restoring force can be described by Hooke's law, is given by

$$x(t) = X\cos\frac{2\pi t}{T}$$

,

where *X* is amplitude. At $t = 0$, the initial position is $x_0 = X$, and the displacement oscillates back and forth with a period *T*. (When $t = T$, we get $x = X$ again because cos 2π = 1.). Furthermore, from this expression for *x*, the velocity *v* as a function of time is given by

$$v(t) = -v_{\max}\sin\left(\frac{2\pi t}{T}\right)$$

, where

$$v_{\max} = \frac{2\pi X}{T} = X\sqrt{\frac{k}{m}}$$

.

The object has zero velocity at maximum displacement—for example, $v=0$ when $t=0$, and at that time $x=X$. The minus sign in the first equation for $v(t)$ gives the correct direction for the velocity. Just after the start of the motion, for instance, the velocity is negative because the system is moving back toward the equilibrium point. Finally, we can get an expression for acceleration using Newton's second law. [Then we have $x(t)$, $v(t)$, *t*, and $a(t)$, the quantities needed for kinematics and a description of simple harmonic motion.] According to Newton's second law, the acceleration is

$$a = \frac{F}{m} = \frac{kx}{m}$$

. So, $a(t)$ is also a cosine function:

$$a(t) = -\frac{kX}{m}\cos\frac{2\pi t}{T}$$

.

Hence, $a(t)$ is directly proportional to and in the opposite direction to $a(t)$.

Figure 4 shows the simple harmonic motion of an object on a spring and presents graphs of $x(t)$, $v(t)$, and $a(t)$ versus time.

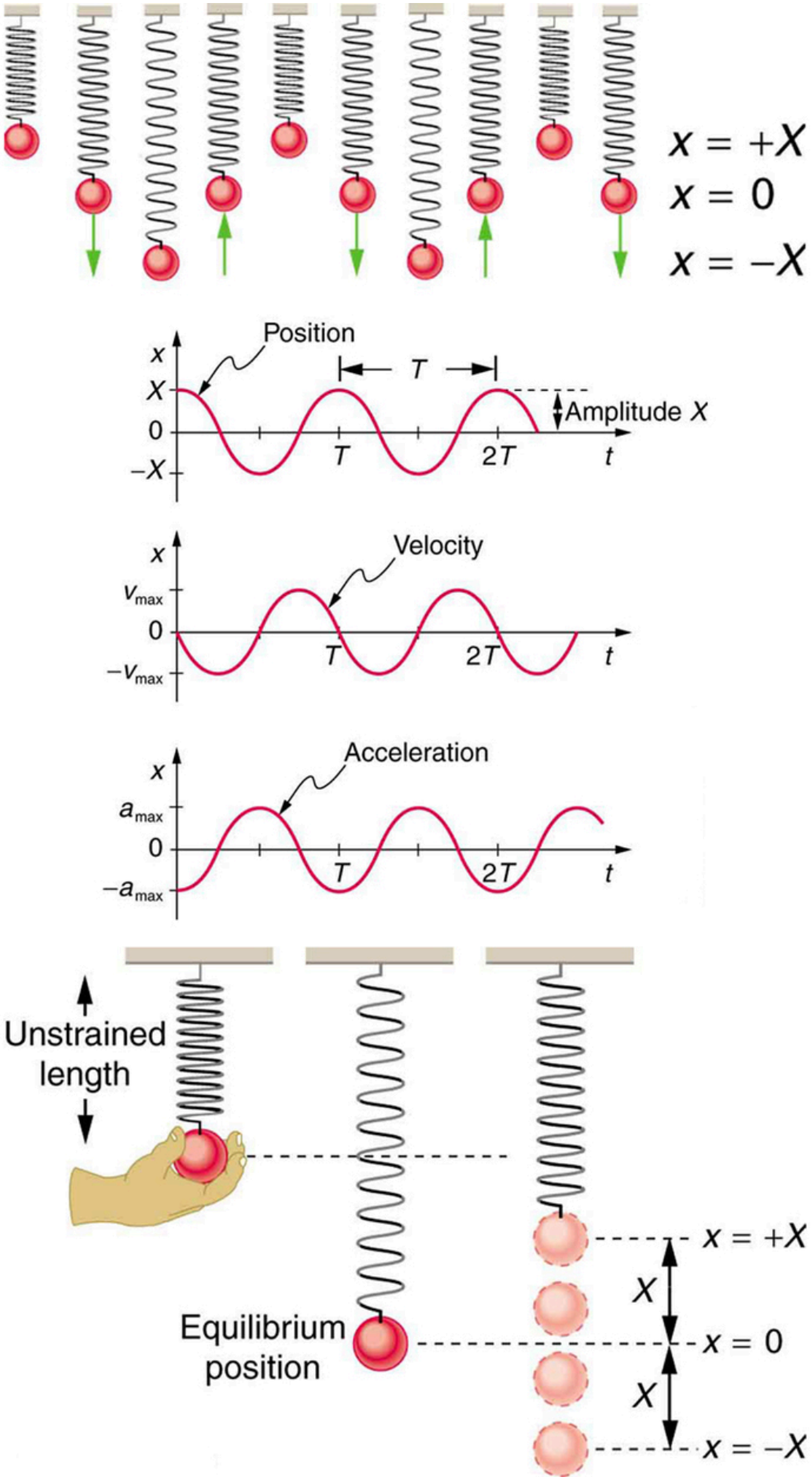

x = +X
x = 0
x = −X
Position
T
Amplitude X
x
X
0
−X
T
2T
t
Velocity
v_{max}
$-v_{max}$
Acceleration
a_{max}
$-a_{max}$
Unstrained length
Equilibrium position
X
X

Figure 4. Graphs of and versus t for the motion of an object on a spring. The net force on the object can be described by Hooke's law, and so the object undergoes simple harmonic motion. Note that the initial position has the vertical displacement at its maximum value X; v is initially zero and then negative as the object moves down; and the initial acceleration is negative, back toward the equilibrium position and becomes zero at that point.

The most important point here is that these equations are mathematically straightforward and are valid for all simple harmonic motion. They are very useful in visualizing waves associated with simple harmonic motion, including visualizing how waves add with one another.

Check Your Understanding

Part 1

Suppose you pluck a banjo string. You hear a single note that starts out loud and slowly quiets over time. Describe what happens to the sound waves in terms of period, frequency and amplitude as the sound decreases in volume.

Solution

Frequency and period remain essentially unchanged. Only amplitude decreases as volume decreases.

Part 2

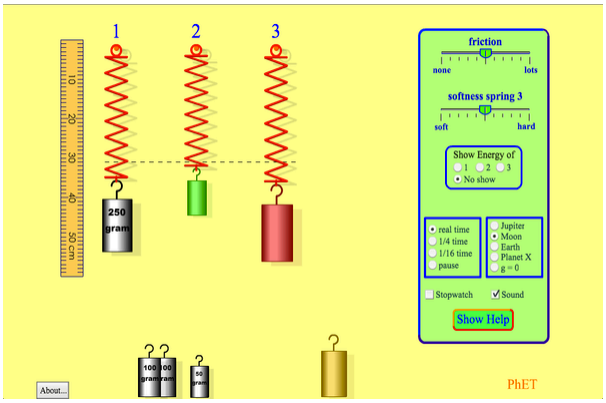
A babysitter is pushing a child on a swing. At the point where the swing reaches *x*, where would the corresponding point on a wave of this motion be located?

Solution

x is the maximum deformation, which corresponds to the amplitude of the wave. The point on the wave would either be at the very top or the very bottom of the curve.

PhET Explorations: Masses and Springs

A realistic mass and spring laboratory. Hang masses from springs and adjust the spring stiffness and damping. You can even slow time. Transport the lab to different planets. A chart shows the kinetic, potential, and thermal energy for each spring.



Click to run the simulation.

Selected Solutions

- Simple harmonic motion is oscillatory motion for a system that can be described only by Hooke's law. Such a system is also called a simple harmonic oscillator.
- Maximum displacement is the amplitude *X*. The period *T* and frequency *f* of a simple harmonic oscillator are given by $T = 2\pi\sqrt{\frac{m}{k}}$ and $f = \frac{1}{2\pi}\sqrt{\frac{k}{m}}$, where *m* is the mass of the system.
- Displacement in simple harmonic motion as a function of time is given by $x\left(t\right) = X\cos\frac{2\pi t}{T}$.
- The velocity is given by $v\left(t\right) = -v_{\max}\sin\frac{2\pi t}{T}$, where $v_{\max} = \sqrt{\frac{k}{m}}X$.
- The acceleration is found to be $a(t) = -\frac{kX}{m}\cos\frac{2\pi t}{T}$.

Conceptual Questions

1. What conditions must be met to produce simple harmonic motion?
2. (a) If frequency is not constant for some oscillation, can the oscillation be simple harmonic motion? (b) Can you think of any examples of harmonic motion where the frequency may depend on the amplitude?
3. Give an example of a simple harmonic oscillator, specifically noting how its frequency is independent of amplitude.

4. Explain why you expect an object made of a stiff material to vibrate at a higher frequency than a similar object made of a spongy material.
5. As you pass a freight truck with a trailer on a highway, you notice that its trailer is bouncing up and down slowly. Is it more likely that the trailer is heavily loaded or nearly empty? Explain your answer.
6. Some people modify cars to be much closer to the ground than when manufactured. Should they install stiffer springs? Explain your answer.

Problems & Exercises

1. A type of cuckoo clock keeps time by having a mass bouncing on a spring, usually something cute like a cherub in a chair. What force constant is needed to produce a period of 0.500 s for a 0.0150-kg mass?
2. If the spring constant of a simple harmonic oscillator is doubled, by what factor will the mass of the system need to change in order for the frequency of the motion to remain the same?
3. A 0.500-kg mass suspended from a spring oscillates with a period of 1.50 s. How much mass must be added to the object to change the period to 2.00 s?
4. By how much leeway (both percentage and mass) would you have in the selection of the mass of the object in the previous problem if you did not wish the new period to be greater than 2.01 s or less than 1.99 s?
5. Suppose you attach the object with mass m to a vertical spring originally at rest, and let it bounce up and down. You release the object from rest at the spring's original rest length. (a) Show that the spring exerts an upward force of 2.00 mg on the object at its lowest point. (b) If the spring has a force constant of 10.0 N/m and a 0.25-kg-mass object is set in motion as described, find the amplitude of the oscillations. (c) Find the maximum velocity.
6. A diver on a diving board is undergoing simple harmonic motion. Her mass is 55.0 kg and the period of her motion is 0.800 s. The next diver is a male whose period of simple harmonic oscillation is 1.05 s. What is his mass if the mass of the board is negligible?
7. Suppose a diving board with no one on it bounces up and down in a simple harmonic motion with a frequency of 4.00 Hz. The board has an effective mass of 10.0 kg. What is the frequency of the simple harmonic motion of a 75.0-kg diver on the board?
8. The device pictured in Figure 6 entertains infants while keeping them from wandering. The child bounces in a harness suspended from a door frame by a spring constant.

Figure 6. This child's toy relies on springs to keep infants entertained. (credit: By Humboldthead, Flickr)

(a) If the spring stretches 0.250 m while supporting an 8.0-kg child, what is its spring constant? (b) What is the time for one complete bounce of this child? (c) What is the child's maximum velocity if the amplitude of her bounce is 0.200 m?

9. A 90.0-kg skydiver hanging from a parachute bounces up and down with a period of 1.50 s. What is the new period of oscillation when a second skydiver, whose mass is 60.0 kg, hangs from the legs of the first, as seen in Figure 7.

Figure 7. The oscillations of one skydiver are about to be affected by a second skydiver. (credit: U.S. Army, www.army.mil)

Glossary

amplitude: the maximum displacement from the equilibrium position of an object oscillating around the equilibrium position

simple harmonic motion: the oscillatory motion in a system where the net force can be described by Hooke's law

simple harmonic oscillator: a device that implements Hooke's law, such as a mass that is attached to a spring, with the other end of the spring being connected to a rigid support such as a wall

Selected Solutions to Problems & Exercises

1. 2.37 N/m

3. 0.389 kg

6. 94.7 kg

9. 1.94 s

Video: Harmonic Motion

Lumen Learning

Watch the following Physics Concept Trailer to see the principles of harmonic motion applied to bungee jumping.

A YouTube element has been excluded from this version of the text. You can view it online here: https://pressbooks.nscc.ca/heatlightsound/?p=127

The Simple Pendulum

Lumen Learning

Learning Objectives

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

- Measure acceleration due to gravity.

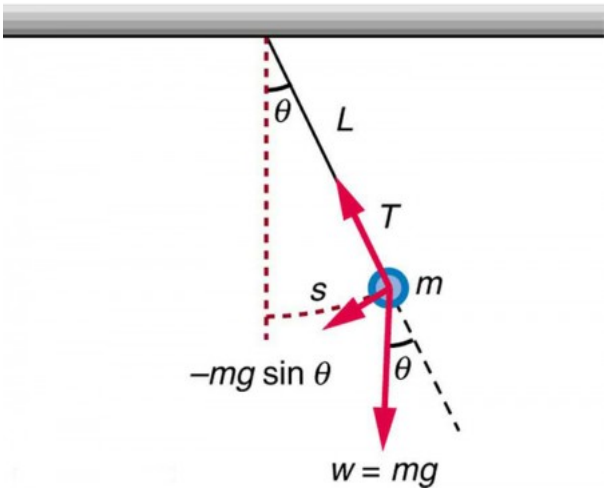


Figure 1.

In Figure 1 we see that a simple pendulum has a small-diameter bob and a string that has a very small mass but is strong enough not to stretch appreciably. The linear displacement from equilibrium is s, the length of the arc. Also shown are the forces on the bob, which result in a net force of $-mg\sin\theta$ toward the equilibrium position—that is, a restoring force.

Pendulums are in common usage. Some have crucial uses, such as in clocks; some are for fun, such as a child's swing; and some are just there, such as the sinker on a fishing line. For small displacements, a pendulum is a simple harmonic oscillator. A *simple pendulum* is defined to have an object that has a small mass, also known as the pendulum bob, which is suspended from a light wire or string, such as shown in Figure 1. Exploring the simple pendulum a bit further, we can discover the conditions under which it performs simple harmonic motion, and we can derive an interesting expression for its period.

We begin by defining the displacement to be the arc length s. We see from Figure 1 that the net force on the bob is tangent to the arc and equals $-mg\sin\theta$. (The weight mg has components $mg\cos\theta$ along the string and $mg\sin\theta$ tangent to the arc.) Tension in the string exactly cancels the component $mg\cos\theta$ parallel to the string. This leaves a *net* restoring force back toward the equilibrium position at $\theta = 0$.

Now, if we can show that the restoring force is directly proportional to the displacement, then we have a simple harmonic oscillator. In trying to determine if we have a simple harmonic oscillator, we should note that for small angles (less than about 15º), $\sin\theta \approx \theta$ ($\sin\theta$ and θ differ by about 1% or less at smaller angles). Thus, for angles less than about 15º, the restoring force F is

$$F \approx -mg\theta.$$

The displacement *s* is directly proportional to θ. When θ is expressed in radians, the arc length in a circle is related to its radius (*L* in this instance) by $s = L\theta$, so that

$$\theta = \frac{s}{L}$$

.

For small angles, then, the expression for the restoring force is:

$$F \approx -\frac{mg}{L}s$$

.

This expression is of the form: $F = -kx$, where the force constant is given by

$$k = \frac{mg}{L}$$

and the displacement is given by $x = s$. For angles less than about 15º, the restoring force is directly proportional to the displacement, and the simple pendulum is a simple harmonic oscillator.

Using this equation, we can find the period of a pendulum for amplitudes less than about 15º. For the simple pendulum:

$$T = 2\pi\sqrt{\frac{m}{k}} = 2\pi\sqrt{\frac{m}{\frac{mg}{L}}}$$

Thus,

$$T = 2\pi\sqrt{\frac{L}{g}}$$

for the period of a simple pendulum. This result is interesting because of its simplicity. The only things that affect the period of a simple pendulum are its length and the acceleration due to gravity. The period is completely independent of other factors, such as mass. As with simple harmonic oscillators, the period *T* for a pendulum is nearly independent of amplitude, especially if θ is less than about 15º. Even simple pendulum clocks can be finely adjusted and accurate.

Note the dependence of *T* on *g*. If the length of a pendulum is precisely known, it can actually be used to measure the acceleration due to gravity. Consider Example 1.

Example 1. Measuring Acceleration due to Gravity: The Period of a Pendulum

What is the acceleration due to gravity in a region where a simple pendulum having a length 75.000 cm has a period of 1.7357 s?

Strategy

We are asked to find g given the period T and the length L of a pendulum. We can solve

$$T = 2\pi\sqrt{\frac{L}{g}}$$

for g, assuming only that the angle of deflection is less than 15º.

Solution

Square

$$T = 2\pi\sqrt{\frac{L}{g}}$$

and solve for g:

$$g = 4\pi^2\frac{L}{T^2}$$

.

Substitute known values into the new equation:

$$g = 4\pi^2\frac{0.750000\text{ m}}{(1.7357\text{ s})^2}$$

.

Calculate to find g:

$$g = 9.8281\text{ m/s}^2.$$

Discussion

This method for determining g can be very accurate. This is why length and period are given to five digits in this example. For the precision of the approximation $\sin\theta \approx \theta$ to be better than the precision of the pendulum length and period, the maximum displacement angle should be kept below about 0.5º.

Making Career Connections

Knowing g can be important in geological exploration; for example, a map of g over large geographical regions aids the study of plate tectonics and helps in the search for oil fields and large mineral deposits.

Take Home Experiment: Determining *g*

Use a simple pendulum to determine the acceleration due to gravity g in your own locale. Cut a piece of a string or dental floss so that it is about 1 m long. Attach a small object of high density to the end of the string (for example, a metal nut or a car key). Starting at an angle of less than 10º, allow the pendulum to swing and

measure the pendulum's period for 10 oscillations using a stopwatch. Calculate g. How accurate is this measurement? How might it be improved?

Check Your Understanding

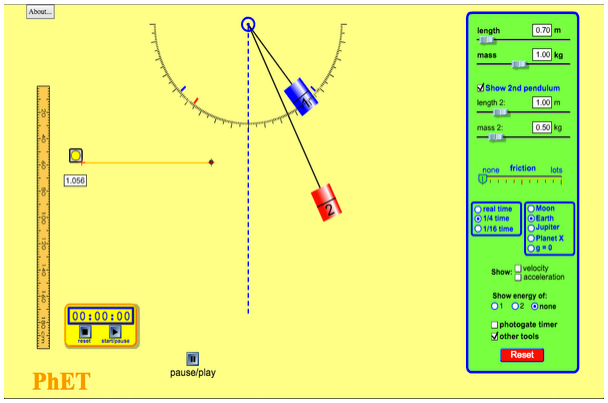
An engineer builds two simple pendula. Both are suspended from small wires secured to the ceiling of a room. Each pendulum hovers 2 cm above the floor. Pendulum 1 has a bob with a mass of 10 kg. Pendulum 2 has a bob with a mass of 100 kg. Describe how the motion of the pendula will differ if the bobs are both displaced by 12º.

Solution

The movement of the pendula will not differ at all because the mass of the bob has no effect on the motion of a simple pendulum. The pendula are only affected by the period (which is related to the pendulum's length) and by the acceleration due to gravity.

PhET Explorations: Pendulum Lab

Play with one or two pendulums and discover how the period of a simple pendulum depends on the length of the string, the mass of the pendulum bob, and the amplitude of the swing. It's easy to measure the period using the photogate timer. You can vary friction and the strength of gravity. Use the pendulum to find the value of g on planet X. Notice the anharmonic behavior at large amplitude.



Click to run the simulation.

Section Summary

- A mass m suspended by a wire of length L is a simple pendulum and undergoes simple harmonic motion for amplitudes less than about 15º.

$$T = 2\pi\sqrt{\frac{L}{g}}$$

- The period of a simple pendulum is , where L is the length of the string and g is the acceleration due to gravity.

Conceptual Questions

1. Pendulum clocks are made to run at the correct rate by adjusting the pendulum's length. Suppose you move from one city to another where the acceleration due to gravity is slightly greater, taking your pendulum clock with you, will you have to lengthen or shorten the pendulum to keep the correct time, other factors remaining constant? Explain your answer.

Problems & Exercises

As usual, the acceleration due to gravity in these problems is taken to be g = 9.80 m/s^2, unless otherwise specified.

1. What is the length of a pendulum that has a period of 0.500 s?
2. Some people think a pendulum with a period of 1.00 s can be driven with "mental energy" or psycho kinetically, because its period is the same as an average heartbeat. True or not, what is the length of such a pendulum?
3. What is the period of a 1.00-m-long pendulum?
4. How long does it take a child on a swing to complete one swing if her center of gravity is 4.00 m below the pivot?
5. The pendulum on a cuckoo clock is 5.00 cm long. What is its frequency?
6. Two parakeets sit on a swing with their combined center of mass 10.0 cm below the pivot. At what frequency do they swing?
7. (a) A pendulum that has a period of 3.00000 s and that is located where the acceleration due to gravity is 9.79 m/s^2 is moved to a location where it the acceleration due to gravity is 9.82 m/s^2. What is its new period? (b) Explain why so many digits are needed in the value for the period, based on the relation between the period and the acceleration due to gravity.
8. A pendulum with a period of 2.00000 s in one location (g = 9.80 m/s^2) is moved to a new location where the period is now 1.99796 s. What is the acceleration due to gravity at its new location?
9. (a) What is the effect on the period of a pendulum if you double its length? (b) What is the effect on the period of a pendulum if you decrease its length by 5.00%?
10. Find the ratio of the new/old periods of a pendulum if the pendulum were transported from Earth to the Moon, where the acceleration due to gravity is 1.63 m/s^2.
11. At what rate will a pendulum clock run on the Moon, where the acceleration due to gravity is 1.63 m/s^2, if it keeps time accurately on Earth? That is, find the time (in hours) it takes the clock's hour hand to make one revolution on the Moon.
12. Suppose the length of a clock's pendulum is changed by 1.000%, exactly at noon one day. What time will it read 24.00 hours later, assuming it the pendulum has kept perfect time before the

change? Note that there are two answers, and perform the calculation to four-digit precision.

13. If a pendulum-driven clock gains 5.00 s/day, what fractional change in pendulum length must be made for it to keep perfect time?

Glossary

simple pendulum: an object with a small mass suspended from a light wire or string

Selected Solutions to Problems & Exercises

1. 6.21 cm

3. 2.01 s

5. 2.23 Hz

7. (a) 2.99541 s; (b) Since the period is related to the square root of the acceleration of gravity, when the acceleration changes by 1% the period changes by $(0.01)^2$ =0.01% so it is necessary to have at least 4 digits after the decimal to see the changes.

9. (a) Period increases by a factor of 1.41

$$\left(\sqrt{2}\right)$$

; (b) Period decreases to 97.5% of old period

11. Slow by a factor of 2.45

13. length must increase by 0.0116%

Energy and the Simple Harmonic Oscillator

Lumen Learning

Learning Objectives

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

- Determine the maximum speed of an oscillating system.

To study the energy of a simple harmonic oscillator, we first consider all the forms of energy it can have We know from Hooke's Law: Stress and Strain Revisited that the energy stored in the deformation of a simple harmonic oscillator is a form of potential energy given by:

$$\text{PE}_{\text{el}} = \frac{1}{2}kx^2$$

.

Because a simple harmonic oscillator has no dissipative forces, the other important form of energy is kinetic energy KE. Conservation of energy for these two forms is:

$$\text{KE} + \text{PE}_{\text{el}} = \text{constant}$$

or

$$\frac{1}{2}mv^2 + \frac{1}{2}kx^2 = \text{constant}$$

.

This statement of conservation of energy is valid for *all* simple harmonic oscillators, including ones where the gravitational force plays a role

Namely, for a simple pendulum we replace the velocity with $v = L\omega$, the spring constant with

$$k = \frac{mg}{L}$$

, and the displacement term with $x = L\theta$. Thus

$$\frac{1}{2}mL^2\omega^2 + \frac{1}{2}mgL\theta^2 = \text{constant}$$

In the case of undamped simple harmonic motion, the energy oscillates back and forth between kinetic and potential, going completely from one to the other as the system oscillates. So for the simple example

of an object on a frictionless surface attached to a spring, as shown again in Figure 1, the motion starts with all of the energy stored in the spring. As the object starts to move, the elastic potential energy is converted to kinetic energy, becoming entirely kinetic energy at the equilibrium position. It is then converted back into elastic potential energy by the spring, the velocity becomes zero when the kinetic energy is completely converted, and so on. This concept provides extra insight here and in later applications of simple harmonic motion, such as alternating current circuits.

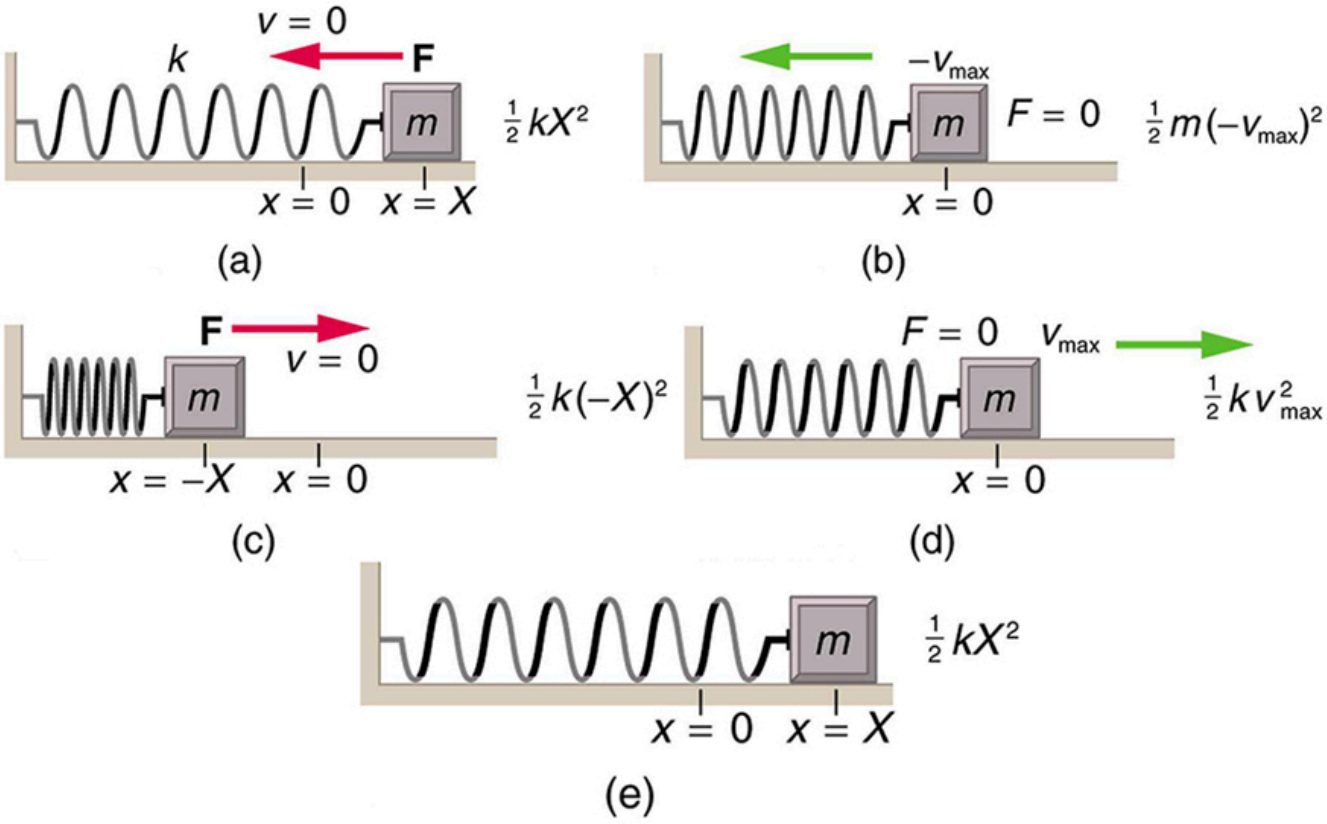


Figure 1. The transformation of energy in simple harmonic motion is illustrated for an object attached to a spring on a frictionless surface.

The conservation of energy principle can be used to derive an expression for velocity *v*. If we start our simple harmonic motion with zero velocity and maximum displacement ($x = X$), then the total energy is

$$\frac{1}{2}kX^2$$

.

This total energy is constant and is shifted back and forth between kinetic energy and potential energy, at most times being shared by each. The conservation of energy for this system in equation form is thus:

$$\frac{1}{2}mv^2 + \frac{1}{2}kx^2 = \frac{1}{2}kX^2$$

.

Solving this equation for *v* yields:

$$v = \pm\sqrt{\frac{k}{m}\left(X^2 - x^2\right)}$$

.

Manipulating this expression algebraically gives:

$$v = \pm\sqrt{\frac{k}{m}}X\sqrt{1 - \frac{x^2}{X^2}}$$

and so

$$v = \pm v_{\max}\sqrt{1 - \frac{x^2}{X^2}}$$

,

where

$$v_{\max} = \sqrt{\frac{k}{m}}X$$

.

From this expression, we see that the velocity is a maximum ($v_{\max}$) at $x = 0$, as stated earlier in

$$v(t) = -v_{\max}\sin\frac{2\pi t}{T}$$

. Notice that the maximum velocity depends on three factors. Maximum velocity is directly proportional to amplitude. As you might guess, the greater the maximum displacement the greater the maximum velocity. Maximum velocity is also greater for stiffer systems, because they exert greater force for the same displacement. This observation is seen in the expression for $v_{\max}$; it is proportional to the square root of the force constant k. Finally, the maximum velocity is smaller for objects that have larger masses, because the maximum velocity is inversely proportional to the square root of m. For a given force, objects that have large masses accelerate more slowly.

A similar calculation for the simple pendulum produces a similar result, namely:

$$\omega_{\max} = \sqrt{\frac{g}{L}}\theta_{\max}$$

Example 1. Determine the Maximum Speed of an Oscillating System: A Bumpy Road

Suppose that a car is 900 kg and has a suspension system that has a force constant $k = 6.53 \times 10^4$ N/m. The car hits a bump and bounces with an amplitude of 0.100 m. What is its maximum vertical velocity if you assume no damping occurs?

Strategy

We can use the expression for v_{max} given in

$$v_{\max} = \sqrt{\frac{k}{m}}X$$

to determine the maximum vertical velocity. The variables *m* and *k* are given in the problem statement, and the maximum displacement *X* is 0.100 m.

Solution

Identify knowns. Substitute known values into

$$v_{\max} = \sqrt{\frac{k}{m}}X$$

:

$$v_{\max} = \sqrt{\frac{6.53\times10^4\ \text{N/m}}{900\ \text{kg}}}\,(0.100\ \text{m})$$

Calculate to find v_{max}= 0.852 m/s.

Discussion

This answer seems reasonable for a bouncing car. There are other ways to use conservation of energy to find v_{max}. We could use it directly, as was done in the example featured in Hooke's Law: Stress and Strain Revisited.

The small vertical displacement *y* of an oscillating simple pendulum, starting from its equilibrium position, is given as *y*(*t*) *a* sin *ωt*, where *a* is the amplitude, *ω* is the angular velocity and *t* is the time taken. Substituting

$$\omega = \frac{2\pi}{T}$$

, we have

$$yt = a\sin\left(\frac{2\pi t}{T}\right)$$

.

Thus, the displacement of pendulum is a function of time as shown above.

Also the velocity of the pendulum is given by

$$v(t) = \frac{2a\pi}{T}\cos\left(\frac{2\pi t}{T}\right)$$

,

so the motion of the pendulum is a function of time.

Check Your Understanding

Part 1

Why does it hurt more if your hand is snapped with a ruler than with a loose spring, even if the displacement of each system is equal?

Solution

The ruler is a stiffer system, which carries greater force for the same amount of displacement. The ruler snaps your hand with greater force, which hurts more.

Part 2

You are observing a simple harmonic oscillator. Identify one way you could decrease the maximum velocity of the system.

Solution

You could increase the mass of the object that is oscillating.

Section Summary

- Energy in the simple harmonic oscillator is shared between elastic potential energy and kinetic energy, with the total being constant: $\frac{1}{2}\mathrm{mv}^2 + \frac{1}{2}\mathrm{kx}^2 = \text{constant}$.
- Maximum velocity depends on three factors: it is directly proportional to amplitude, it is greater for stiffer systems, and it is smaller for objects that have larger masses: $v_{\max} = \sqrt{\frac{k}{m}}X$.

Conceptual Questions

Explain in terms of energy how dissipative forces such as friction reduce the amplitude of a harmonic oscillator. Also explain how a driving mechanism can compensate. (A pendulum clock is such a system.)

Problems & Exercises

1. The length of nylon rope from which a mountain climber is suspended has a force constant of 1.40×10^4 N/m. (a) What is the frequency at which he bounces, given his mass plus and the mass of his equipment are 90.0 kg? (b) How much would this rope stretch to break the climber's fall if he free-falls 2.00 m before the rope runs out of slack? Hint: Use conservation of energy. (c)

Repeat both parts of this problem in the situation where twice this length of nylon rope is used.

2. **Engineering Application.** Near the top of the Citigroup Center building in New York City, there is an object with mass of 4.00×10^5 kg on springs that have adjustable force constants. Its function is to dampen wind-driven oscillations of the building by oscillating at the same frequency as the building is being driven—the driving force is transferred to the object, which oscillates instead of the entire building. (a) What effective force constant should the springs have to make the object oscillate with a period of 2.00 s? (b) What energy is stored in the springs for a 2.00-m displacement from equilibrium?

Solutions to Problems & Exercises

1. (a) 1.99 Hz; (b) 50.2 cm; (c) 1.41 Hz, 0.710 m
2. (a) 3.95×10^6 N/m; (b) 7.90×10^6 J

Uniform Circular Motion and Simple Harmonic Motion

Lumen Learning

Learning Objectives

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

- Compare simple harmonic motion with uniform circular motion.

Figure 1. The horses on this merry-go-round exhibit uniform circular motion. (credit: Wonderlane, Flickr)

There is an easy way to produce simple harmonic motion by using uniform circular motion. Figure 2 shows one way of using this method. A ball is attached to a uniformly rotating vertical turntable, and its shadow is projected on the floor as shown. The shadow undergoes simple harmonic motion. Hooke's law usually describes uniform circular motions (ω constant) rather than systems that have large visible displacements. So observing the projection of uniform circular motion, as in Figure 2, is often easier than observing a precise large-scale simple harmonic oscillator. If studied in sufficient depth, simple harmonic motion produced in this manner can give considerable insight into many aspects of oscillations and waves and is very useful mathematically. In our brief treatment, we shall indicate some of the major features of this relationship and how they might be useful.

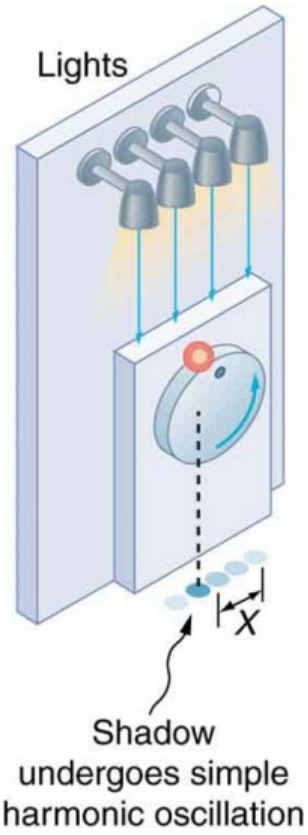


Figure 2. The shadow of a ball rotating at constant angular velocity ω on a turntable goes back and forth in precise simple harmonic motion.

Figure 3 shows the basic relationship between uniform circular motion and simple harmonic motion. The point P travels around the circle at constant angular velocity ω. The point P is analogous to an object on the merry-go-round. The projection of the position of P onto a fixed axis undergoes simple harmonic motion and is analogous to the shadow of the object. At the time shown in the figure, the projection has position x and moves to the left with velocity v. The velocity of the point P around the circle equals

$$\bar{v}_{\max}$$

.The projection of

$$\bar{v}_{\max}$$

on the x-axis is the velocity v of the simple harmonic motion along the x-axis.

In Figure 3 we see that a point P moving on a circular path with a constant angular velocity ω is undergoing uniform circular motion. Its projection on the x-axis undergoes simple harmonic motion. Also shown is the velocity of this point around the circle,

$$\bar{v}_{\max}$$

, and its projection, which is v. Note that these velocities form a similar triangle to the displacement triangle.

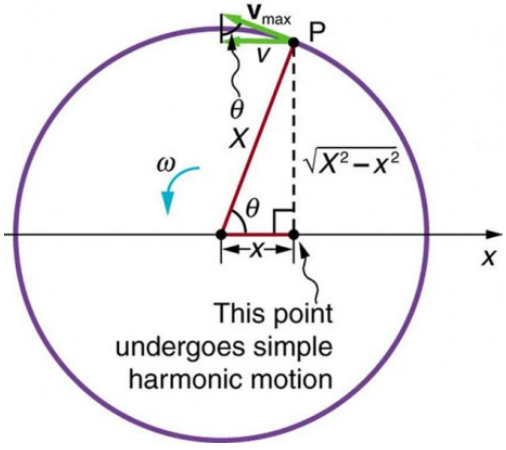


Figure 3. A point moving on a circular path

To see that the projection undergoes simple harmonic motion, note that its position x is given by

$$x = X \cos \theta,$$

where $\theta = \omega t$, ω is the constant angular velocity, and X is the radius of the circular path. Thus,

$$x = X \cos \omega t.$$

The angular velocity ω is in radians per unit time; in this case 2π radians is the time for one revolution T. That is,

$$\omega = \frac{2\pi}{T}$$

. Substituting this expression for ω, we see that the position x is given by:

$$x(t) = \cos\left(\frac{2\pi t}{T}\right)$$

.

This expression is the same one we had for the position of a simple harmonic oscillator in Simple Harmonic Motion: A Special Periodic Motion. If we make a graph of position versus time as in Figure 4, we see again the wavelike character (typical of simple harmonic motion) of the projection of uniform circular motion onto the x-axis.

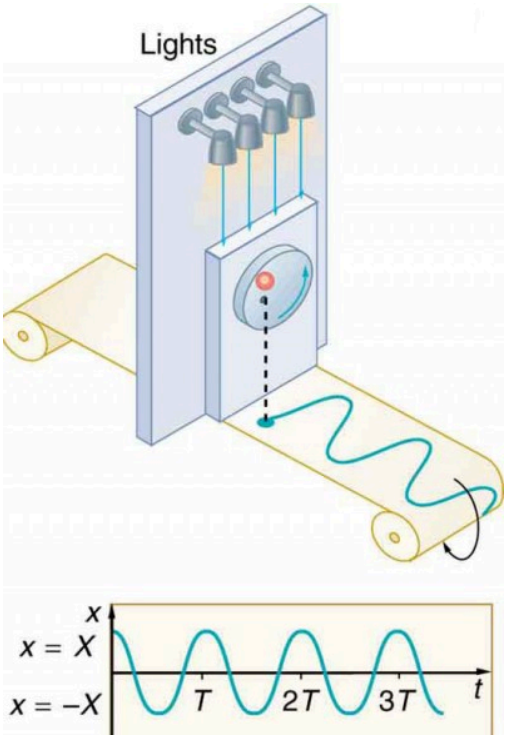


Figure 4. The position of the projection of uniform circular motion performs simple harmonic motion, as this wavelike graph of x versus t indicates.

Now let us use Figure 3 to do some further analysis of uniform circular motion as it relates to simple harmonic motion. The triangle formed by the velocities in the figure and the triangle formed by the displacements (X, x, and $\sqrt{X^2 - x^2}$) are similar right triangles. Taking ratios of similar sides, we see that

$$\frac{v}{v_{\max}} = \sqrt{\frac{X^2 - x^2}{X}} = \sqrt{1 - \frac{x^2}{X^2}}$$

We can solve this equation for the speed v or

$$v = v_{\max}\sqrt{1 - \frac{x^2}{X^2}}$$

.

This expression for the speed of a simple harmonic oscillator is exactly the same as the equation obtained from conservation of energy considerations in Energy and the Simple Harmonic Oscillator.You can begin to see that it is possible to get all of the characteristics of simple harmonic motion from an analysis of the projection of uniform circular motion.

Finally, let us consider the period T of the motion of the projection. This period is the time it takes

the point P to complete one revolution. That time is the circumference of the circle $2\pi X$ divided by the velocity around the circle, v_{max}. Thus, the period T is

$$T = \frac{2\pi X}{v_{\max}}$$

.

We know from conservation of energy considerations that

$$v_{\max} = \sqrt{\frac{k}{m}} X$$

.

Solving this equation for

$$\frac{X}{v_{\max}}$$

gives

$$\frac{X}{v_{\max}} = \sqrt{\frac{m}{k}}$$

.

Substituting this expression into the equation for T yields

$$T = 2\pi \sqrt{\frac{m}{k}}$$

.

Thus, the period of the motion is the same as for a simple harmonic oscillator. We have determined the period for any simple harmonic oscillator using the relationship between uniform circular motion and simple harmonic motion.

Some modules occasionally refer to the connection between uniform circular motion and simple harmonic motion. Moreover, if you carry your study of physics and its applications to greater depths, you will find this relationship useful. It can, for example, help to analyze how waves add when they are superimposed.

Check Your Understanding

Identify an object that undergoes uniform circular motion. Describe how you could trace the simple harmonic motion of this object as a wave.

Solution

A record player undergoes uniform circular motion. You could attach dowel rod to one point on the outside

edge of the turntable and attach a pen to the other end of the dowel. As the record player turns, the pen will move. You can drag a long piece of paper under the pen, capturing its motion as a wave.

Section Summary

A projection of uniform circular motion undergoes simple harmonic oscillation.

Problems & Exercises

1. (a)What is the maximum velocity of an 85.0-kg person bouncing on a bathroom scale having a force constant of 1.50×10^6 N/m, if the amplitude of the bounce is 0.200 cm? (b)What is the maximum energy stored in the spring?
2. A novelty clock has a 0.0100-kg mass object bouncing on a spring that has a force constant of 1.25 N/m. What is the maximum velocity of the object if the object bounces 3.00 cm above and below its equilibrium position? (b) How many joules of kinetic energy does the object have at its maximum velocity?
3. At what positions is the speed of a simple harmonic oscillator half its maximum? That is, what values of $\frac{x}{X}$ give $v = \pm\frac{v_{\max}}{2}$, where X is the amplitude of the motion?
4. A ladybug sits 12.0 cm from the center of a Beatles music album spinning at 33.33 rpm. What is the maximum velocity of its shadow on the wall behind the turntable, if illuminated parallel to the record by the parallel rays of the setting Sun?

Selected Solutions to Problems & Exercises

1. (a) 0.266 m/s; (b) 3.00 J

3. $$\pm\frac{\sqrt{3}}{2}$$

Damped Harmonic Motion

Lumen Learning

Learning Objectives

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

- Compare and discuss underdamped and overdamped oscillating systems.
- Explain critically damped system.

A guitar string stops oscillating a few seconds after being plucked. To keep a child happy on a swing, you must keep pushing. Although we can often make friction and other non-conservative forces negligibly small, completely undamped motion is rare. In fact, we may even want to damp oscillations, such as with car shock absorbers.

Figure 1. In order to counteract dampening forces, this dad needs to keep pushing the swing. (credit: Erik A. Johnson, Flickr)

For a system that has a small amount of damping, the period and frequency are nearly the same as for simple harmonic motion, but the amplitude gradually decreases as shown in Figure 2. This occurs because the non-conservative damping force removes energy from the system, usually in the form of thermal energy. In general, energy removal by non-conservative forces is described as $W_{nc} = \Delta(KE + PE)$, where W_{nc} is work done by a non-conservative force (here the damping force). For a damped harmonic oscillator, W_{nc} is negative because it removes mechanical energy (KE + PE) from the system.

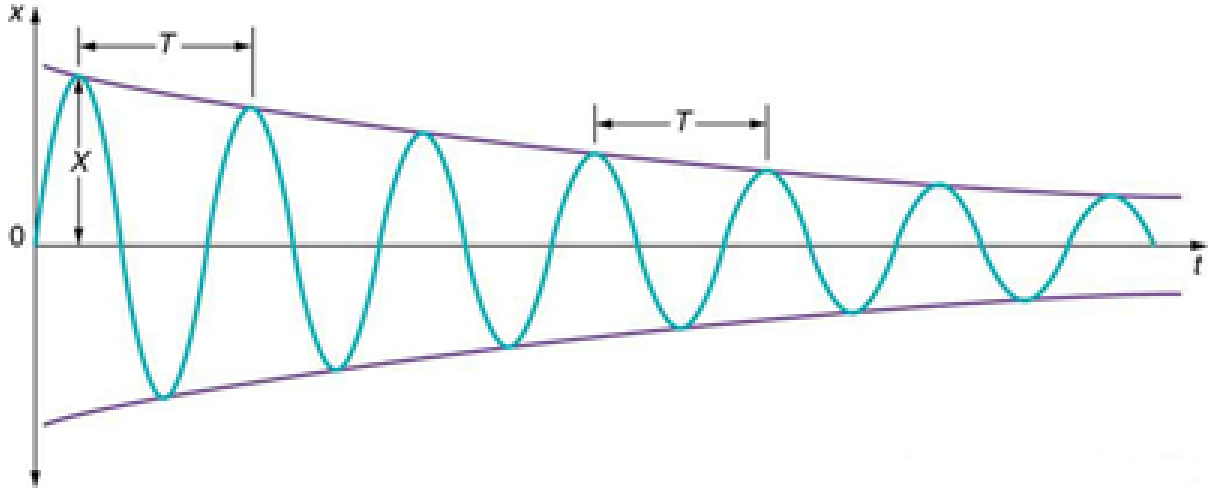


Figure 2. In this graph of displacement versus time for a harmonic oscillator with a small amount of damping, the amplitude slowly decreases, but the period and frequency are nearly the same as if the system were completely undamped.

If you gradually *increase* the amount of damping in a system, the period and frequency begin to be affected, because damping opposes and hence slows the back and forth motion. (The net force is smaller in both directions.) If there is very large damping, the system does not even oscillate—it slowly moves toward equilibrium. Figure 3 shows the displacement of a harmonic oscillator for different amounts of damping. When we want to damp out oscillations, such as in the suspension of a car, we may want the system to return to equilibrium as quickly as possible *Critical damping* is defined as the condition in which the damping of an oscillator results in it returning as quickly as possible to its equilibrium position The critically damped system may overshoot the equilibrium position, but if it does, it will do so only once. Critical damping is represented by Curve A in Figure 3. With less-than critical damping, the system will return to equilibrium faster but will overshoot and cross over one or more times. Such a system is *underdamped*; its displacement is represented by the curve in Figure 2. Curve B in Figure 3 represents an *overdamped* system. As with critical damping, it too may overshoot the equilibrium position, but will reach equilibrium over a longer period of time.

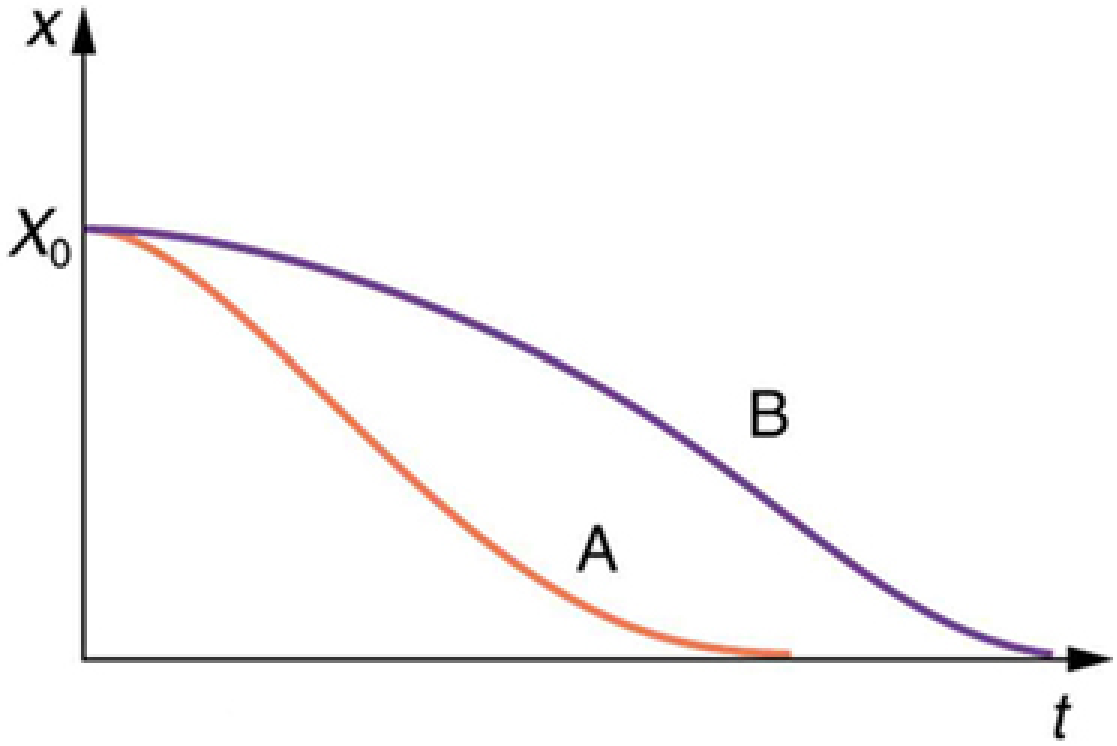


Figure 3. Displacement versus time for a critically damped harmonic oscillator (A) and an overdamped harmonic oscillator (B). The critically damped oscillator returns to equilibrium at X = 0 in the smallest time possible without overshooting.

Critical damping is often desired, because such a system returns to equilibrium rapidly and remains at equilibrium as well. In addition, a constant force applied to a critically damped system moves the system to a new equilibrium position in the shortest time possible without overshooting or oscillating about the new position. For example, when you stand on bathroom scales that have a needle gauge, the needle moves to its equilibrium position without oscillating. It would be quite inconvenient if the needle oscillated about the new equilibrium position for a long time before settling. Damping forces can vary greatly in character. Friction, for example, is sometimes independent of velocity (as assumed in most places in this text). But many damping forces depend on velocity—sometimes in complex ways, sometimes simply being proportional to velocity.

Example 1. Damping an Oscillatory Motion: Friction on an Object Connected to a Spring

Damping oscillatory motion is important in many systems, and the ability to control the damping is even more so. This is generally attained using non-conservative forces such as the friction between surfaces, and viscosity for objects moving through fluids. The following example considers friction. Suppose a 0.200-kg object is connected to a spring as shown in Figure 4, but there is simple friction between the object and the surface, and the coefficient of friction μ_k is equal to 0.0800.

1. What is the frictional force between the surfaces?
2. What total distance does the object travel if it is released 0.100 m from equilibrium, starting at v = 0? The force constant of the spring is k = 50.0 N/m.

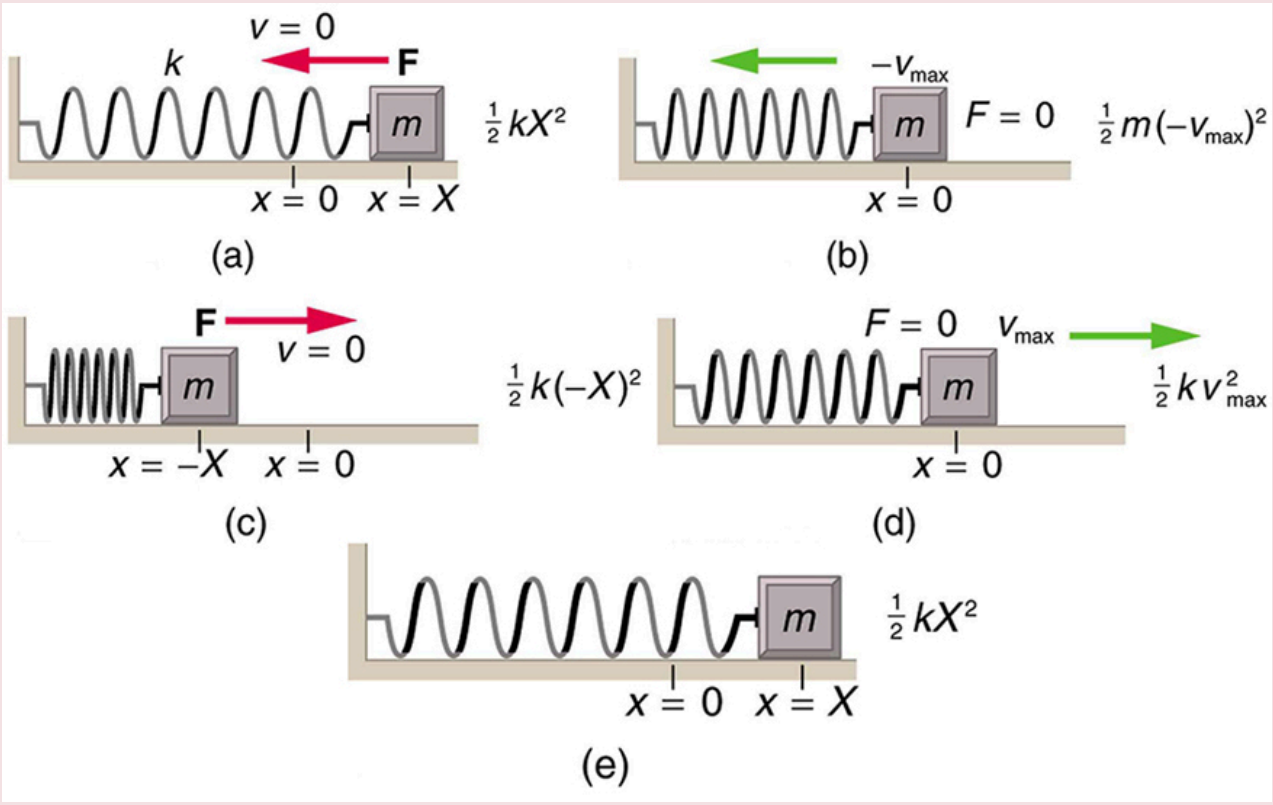


Figure 4. The transformation of energy in simple harmonic motion is illustrated for an object attached to a spring on a frictionless surface.

Strategy

This problem requires you to integrate your knowledge of various concepts regarding waves, oscillations, and damping. To solve an integrated concept problem, you must first identify the physical principles involved. Part 1 is about the frictional force. This is a topic involving the application of Newton's Laws. Part 2 requires an understanding of work and conservation of energy, as well as some understanding of horizontal oscillatory systems.

Now that we have identified the principles we must apply in order to solve the problems, we need to identify the knowns and unknowns for each part of the question, as well as the quantity that is constant in Part 1 and Part 2 of the question.

Solution to Part 1

Choose the proper equation: Friction is $f = \mu_k mg$. Identify the known values.

Enter the known values into the equation: $f = (0.0800)\ (0\ .200\ \text{kg})\ (9\ .80\ \text{m/ s}^{2})$.

Calculate and convert units: $f = 0.157\ \text{N}$.

Discussion for Part 1

The force here is small because the system and the coefficients are small.

Solution to Part 2

Identify the knowns:

- The system involves elastic potential energy as the spring compresses and expands, friction that is related to the work done, and the kinetic energy as the body speeds up and slows down.
- Energy is not conserved as the mass oscillates because friction is a non-conservative force.
- The motion is horizontal, so gravitational potential energy does not need to be considered.

$$\mathrm{PE}_{\mathrm{el,i}} = \tfrac{1}{2}kX^2$$

- Because the motion starts from rest, the energy in the system is initially . This energy is removed by work done by friction $W_{\mathrm{nc}} = -fd$, where d is the total distance traveled and $f = \mu_{\mathrm{k}}mg$ is the force of friction. When the system stops moving, the friction force will balance the

$$\mathrm{PE}_{\mathrm{el,f}} = \tfrac{1}{2}kx^2$$

force exerted by the spring, so where x is the final position and is given by

$$\begin{aligned} F_{\mathrm{el}} &= f \\ kx &= \mu_{\mathrm{k}}mg \\ x &= \tfrac{\mu_{\mathrm{k}}mg}{k} \end{aligned}$$

.

1. By equating the work done to the energy removed, solve for the distance d.

2. The work done by the non-conservative forces equals the initial, stored elastic potential energy. Identify the correct equation to use:

$$W_{\mathrm{nc}} = \Delta\left(\mathrm{KE} + \mathrm{PE}\right) = \mathrm{PE}_{\mathrm{el,f}} - \mathrm{PE}_{\mathrm{el,i}} = \tfrac{1}{2}k\left(\left(\tfrac{\mu_{\mathrm{k}}mg}{k}\right)^2 - X^2\right)$$

3. Recall that $W_{\mathrm{nc}} = -fd$.

4. Enter the friction as $f = \mu_{\mathrm{k}}mg$ into $W_{\mathrm{nc}} = -fd$, thus $W_{\mathrm{nc}} = -\mu_{\mathrm{k}}mgd$.

5. Combine these two equations to find

$$\tfrac{1}{2}k\left(\left(\tfrac{\mu_{\mathrm{k}}mg}{k}\right)^2 - X^2\right) = -\mu_{\mathrm{k}}mgd$$

6. Solve the equation for d:

$$d = \tfrac{k}{2\mu_{\mathrm{k}}mg}\left(X^2 - \left(\tfrac{\mu_{\mathrm{k}}mg}{k}\right)^2\right)$$

7. Enter the known values into the resulting equation:

$$d = \frac{50.0\ \mathrm{N/m}}{2(0.0800)(0.200\ \mathrm{kg})\left(9.80\ \mathrm{m/s^2}\right)}\left((0.100\ \mathrm{m})^2 - \left(\frac{(0.0800)(0.200\ \mathrm{kg})\left(9.80\ \mathrm{m/s^2}\right)}{50.0\ \mathrm{N/m}}\right)^2\right)$$

8. Calculate d and convert units $d = 1.59$ m.

Discussion for Part 2

This is the total distance traveled back and forth across $x = 0$, which is the undamped equilibrium position. The number of oscillations about the equilibrium position will be more than

$$\frac{d}{X} = \frac{1.59 \text{ m}}{0.100 \text{ m}} = 15.9$$

because the amplitude of the oscillations is decreasing with time. At the end of the motion, this system will not return to $x = 0$ for this type of damping force, because static friction will exceed the restoring force. This system is underdamped. In contrast, an overdamped system with a simple constant damping force would not cross the equilibrium position $x = 0$ a single time. For example, if this system had a damping force 20 times greater, it would only move 0.0484 m toward the equilibrium position from its original 0.100-m position.

This worked example illustrates how to apply problem-solving strategies to situations that integrate the different concepts you have learned. The first step is to identify the physical principles involved in the problem. The second step is to solve for the unknowns using familiar problem-solving strategies. These are found throughout the text, and many worked examples show how to use them for single topics. In this integrated concepts example, you can see how to apply them across several topics. You will find these techniques useful in applications of physics outside a physics course, such as in your profession, in other science disciplines, and in everyday life.

Check Your Understanding

Part 1

Why are completely undamped harmonic oscillators so rare?

Solution

Friction often comes into play whenever an object is moving. Friction causes damping in a harmonic oscillator.

Part 2

Describe the difference between overdamping, underdamping, and critical damping.

Solution

An overdamped system moves slowly toward equilibrium. An underdamped system moves quickly to equilibrium, but will oscillate about the equilibrium point as it does so. A critically damped system moves as quickly as possible toward equilibrium without oscillating about the equilibrium.

Section Summary

- Damped harmonic oscillators have non-conservative forces that dissipate their energy.
- Critical damping returns the system to equilibrium as fast as possible without overshooting.

- An underdamped system will oscillate through the equilibrium position.
- An overdamped system moves more slowly toward equilibrium than one that is critically damped.

Conceptual Questions

1. Give an example of a damped harmonic oscillator. (They are more common than undamped or simple harmonic oscillators.)
2. How would a car bounce after a bump under each of these conditions? (a) overdamping; (b) underdamping; (c) critical damping.
3. Most harmonic oscillators are damped and, if undriven, eventually come to a stop. How is this observation related to the second law of thermodynamics?

Problems & Exercises

1. The amplitude of a lightly damped oscillator decreases by 3.0% during each cycle. What percentage of the mechanical energy of the oscillator is lost in each cycle?

Glossary

critical damping: the condition in which the damping of an oscillator causes it to return as quickly as possible to its equilibrium position without oscillating back and forth about this position

over damping: the condition in which damping of an oscillator causes it to return to equilibrium without oscillating; oscillator moves more slowly toward equilibrium than in the critically damped system

under damping: the condition in which damping of an oscillator causes it to return to equilibrium with the amplitude gradually decreasing to zero; system returns to equilibrium faster but overshoots and crosses the equilibrium position one or more times

Forced Oscillations and Resonance

Lumen Learning

Learning Objectives

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

- Observe resonance of a paddle ball on a string.
- Observe amplitude of a damped harmonic oscillator.

Sit in front of a piano sometime and sing a loud brief note at it with the dampers off its strings. It will sing the same note back at you—the strings, having the same frequencies as your voice, are resonating in response to the forces from the sound waves that you sent to them. Your voice and a piano's strings is a good example of the fact that objects—in this case, piano strings—can be forced to oscillate but oscillate best at their natural frequency. In this section, we shall briefly explore applying a *periodic driving force* acting on a simple harmonic oscillator. The driving force puts energy into the system at a certain frequency, not necessarily the same as the natural frequency of the system. The *natural frequency* is the frequency at which a system would oscillate if there were no driving and no damping force.

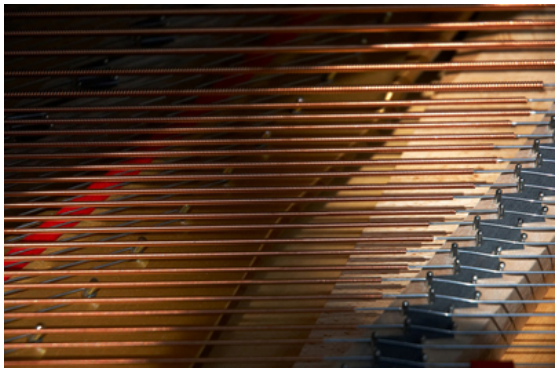


Figure 1. You can cause the strings in a piano to vibrate simply by producing sound waves from your voice. (credit: Matt Billings, Flickr)

Most of us have played with toys involving an object supported on an elastic band, something like the paddle ball suspended from a finger in Figure 2. Imagine the finger in the figure is your finger. At first you hold your finger steady, and the ball bounces up and down with a small amount of damping. If you move your finger up and down slowly, the ball will follow along without bouncing much on its own. As you increase the frequency at which you move your finger up and down, the ball will respond by oscillating with increasing amplitude. When you drive the ball at its natural frequency, the ball's oscillations increase in amplitude with each oscillation for as long as you drive it. The phenomenon of driving a system with a frequency equal to its natural frequency is called *resonance*. A system being driven at its natural frequency is said to *resonate*. As the driving frequency gets progressively higher than the resonant or natural frequency, the amplitude of the oscillations becomes smaller, until the oscillations nearly disappear and your finger simply moves up and down with little effect on the ball.

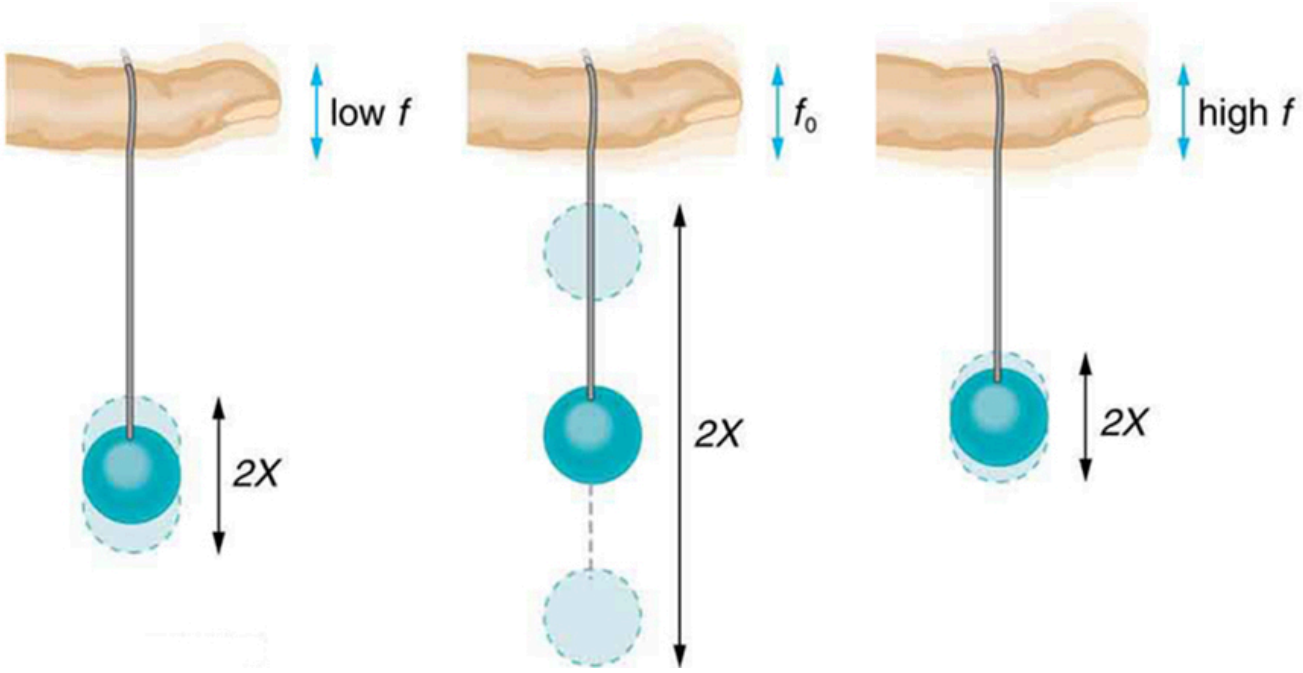


Figure 2. The paddle ball on its rubber band moves in response to the finger supporting it. If the finger moves with the natural frequency f0 of the ball on the rubber band, then a resonance is achieved, and the amplitude of the ball's oscillations increases dramatically. At higher and lower driving frequencies, energy is transferred to the ball less efficiently, and it responds with lower-amplitude oscillations.

Figure 3 shows a graph of the amplitude of a damped harmonic oscillator as a function of the frequency of the periodic force driving it. There are three curves on the graph, each representing a different amount of damping. All three curves peak at the point where the frequency of the driving force equals the natural frequency of the harmonic oscillator. The highest peak, or greatest response, is for the least amount of damping, because less energy is removed by the damping force.

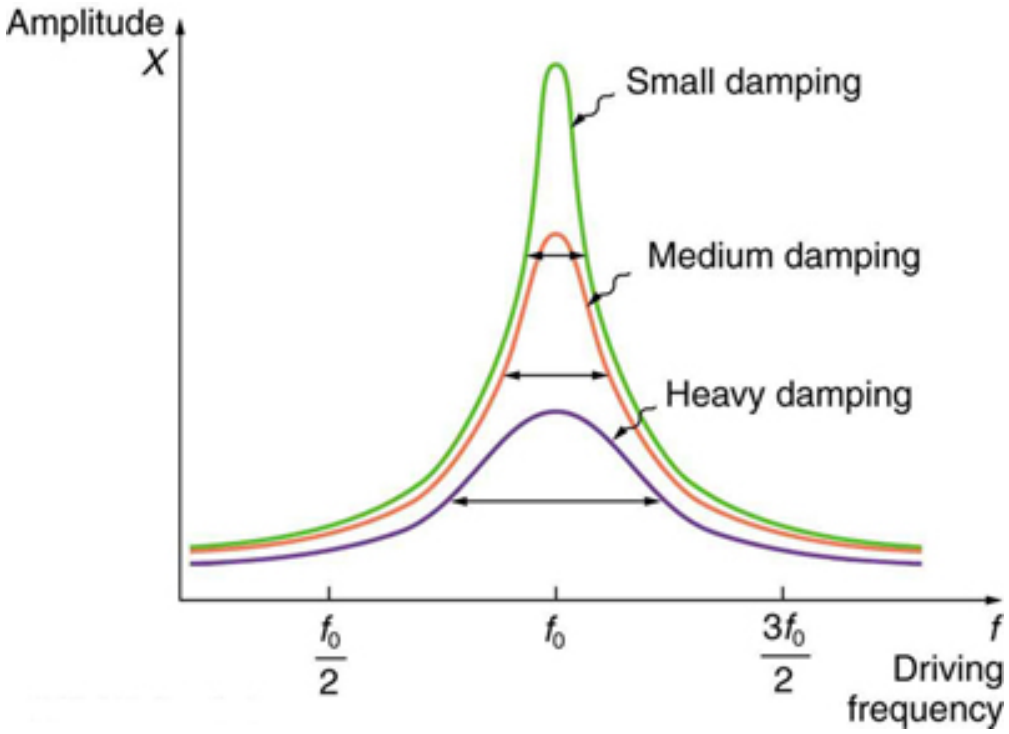


Figure 3. Amplitude of a harmonic oscillator as a function of the frequency of the driving force. The curves represent the same oscillator with the same natural frequency but with different amounts of damping. Resonance occurs when the driving frequency equals the natural frequency, and the greatest response is for the least amount of damping. The narrowest response is also for the least damping.

It is interesting that the widths of the resonance curves shown in Figure 3 depend on damping: the less the damping, the narrower the resonance. The message is that if you want a driven oscillator to resonate at a very specific frequency, you need as little damping as possible. Little damping is the case for piano strings and many other musical instruments. Conversely, if you want small-amplitude oscillations, such as in a car's suspension system, then you want heavy damping. Heavy damping reduces the amplitude, but the tradeoff is that the system responds at more frequencies.

These features of driven harmonic oscillators apply to a huge variety of systems. When you tune a radio, for example, you are adjusting its resonant frequency so that it only oscillates to the desired station's broadcast (driving) frequency. The more selective the radio is in discriminating between stations, the smaller its damping. Magnetic resonance imaging (MRI) is a widely used medical diagnostic tool in which atomic nuclei (mostly hydrogen nuclei) are made to resonate by incoming radio waves (on the order of 100 MHz). A child on a swing is driven by a parent at the swing's natural frequency to achieve maximum amplitude. In all of these cases, the efficiency of energy transfer from the driving force into the oscillator is best at resonance.

Speed bumps and gravel roads prove that even a car's suspension system is not immune to resonance. In spite of finely engineered shock absorbers, which ordinarily convert mechanical energy to thermal energy almost as fast as it comes in, speed bumps still cause a large-amplitude oscillation. On gravel roads that are corrugated, you may have noticed that if you travel at the "wrong" speed, the bumps are very noticeable whereas at other speeds you may hardly feel the bumps at all. Figure 4 shows a photograph of a famous example (the Tacoma Narrows Bridge) of the destructive effects of a driven harmonic oscillation. The Millennium Bridge in London was closed for a short period of time for the same reason while inspections were carried out.

Figure 4. In 1940, the Tacoma Narrows Bridge in Washington state collapsed. Heavy cross winds drove the bridge into oscillations at its resonant frequency. Damping decreased when support cables broke loose and started to slip over the towers, allowing increasingly greater amplitudes until the structure failed (credit: PRI's Studio 360, via Flickr)

In our bodies, the chest cavity is a clear example of a system at resonance. The diaphragm and chest wall drive the oscillations of the chest cavity which result in the lungs inflating and deflating. The system is critically damped and the muscular diaphragm oscillates at the resonant value for the system, making it highly efficient.

Check Your Understanding

A famous magic trick involves a performer singing a note toward a crystal glass until the glass shatters. Explain why the trick works in terms of resonance and natural frequency.

Solution

The performer must be singing a note that corresponds to the natural frequency of the glass. As the sound wave is directed at the glass, the glass responds by resonating at the same frequency as the sound wave. With enough energy introduced into the system, the glass begins to vibrate and eventually shatters.

Section Summary

- A system's natural frequency is the frequency at which the system will oscillate if not affected by driving or damping forces.
- A periodic force driving a harmonic oscillator at its natural frequency produces resonance. The system is said to resonate.
- The less damping a system has, the higher the amplitude of the forced oscillations near resonance. The more damping a system has, the broader response it has to varying driving frequencies.

Conceptual Questions

1. Why are soldiers in general ordered to “route step” (walk out of step) across a bridge?

Problems & Exercises

1. How much energy must the shock absorbers of a 1200-kg car dissipate in order to damp a bounce that initially has a velocity of 0.800 m/s at the equilibrium position? Assume the car returns to its original vertical position.
2. If a car has a suspension system with a force constant of 5.00×10^4 N/m, how much energy must the car’s shocks remove to dampen an oscillation starting with a maximum displacement of 0.0750 m?
3. (a) How much will a spring that has a force constant of 40.0 N/m be stretched by an object with a mass of 0.500 kg when hung motionless from the spring? (b) Calculate the decrease in gravitational potential energy of the 0.500-kg object when it descends this distance. (c) Part of this gravitational energy goes into the spring. Calculate the energy stored in the spring by this stretch, and compare it with the gravitational potential energy. Explain where the rest of the energy might go.
4. Suppose you have a 0.750-kg object on a horizontal surface connected to a spring that has a force constant of 150 N/m. There is simple friction between the object and surface with a static coefficient of friction $\mu_s = 0.100$. (a) How far can the spring be stretched without moving the mass? (b) If the object is set into oscillation with an amplitude twice the distance found in part (a), and the kinetic coefficient of friction is $\mu_k = 0.0850$, what total distance does it travel before stopping? Assume it starts at the maximum amplitude.
5. **Engineering Application.** A suspension bridge oscillates with an effective force constant of 1.00×10^8 N/m. (a) How much energy is needed to make it oscillate with an amplitude of 0.100 m? (b) If soldiers march across the bridge with a cadence equal to the bridge’s natural frequency and impart 1.00×10^4 J of energy each second, how long does it take for the bridge’s oscillations to go from 0.100 m to 0.500 m amplitude.

Glossary

natural frequency: the frequency at which a system would oscillate if there were no driving and no damping forces

resonance: the phenomenon of driving a system with a frequency equal to the system’s natural frequency

resonate: a system being driven at its natural frequency

Selected Solutions to Problems & Exercises

1. 384 J

3. (a). 0.123 m; (b). −0.600 J; (c). 0.300 J. The rest of the energy may go into heat caused by friction and other damping forces.

5. (a) 5.00×10^5 J; (b) 1.20×10^3 s

Waves

Lumen Learning

Learning Objectives

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

- State the characteristics of a wave.
- Calculate the velocity of wave propagation.

Figure 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. Even radio waves are most easily understood using an analogy with water waves. Visualizing water waves is useful because there is more to it than just a mental image. Water waves exhibit characteristics common to all waves, such as amplitude, period, frequency and energy. All wave characteristics can be described by a small set of underlying principles.

A wave is a disturbance that propagates, or moves from the place it was created. The simplest waves repeat themselves for several cycles and are associated with simple harmonic motion. Let us start by considering the simplified water wave in Figure 2. The wave is an up and down disturbance of the water surface. It causes a sea gull to move up and down in simple harmonic motion as the wave crests and troughs (peaks and valleys) pass under the bird. The time for one complete up and down motion is the wave's period *T*. The wave's frequency is

$$f = \frac{1}{T}$$

, as usual. The wave itself moves to the right in Figure 2. This movement of the wave is actually the disturbance moving to the right, not the water itself (or the bird would move to the right). We define *wave velocity* v_w to be the speed at which the disturbance moves. Wave velocity is sometimes also called

the *propagation velocity or propagation speed*, because the disturbance propagates from one location to another.

Misconception Alert

Many people think that water waves push water from one direction to another. In fact, the particles of water tend to stay in one location, save for moving up and down due to the energy in the wave. The energy moves forward through the water, but the water stays in one place. If you feel yourself pushed in an ocean, what you feel is the energy of the wave, not a rush of water.

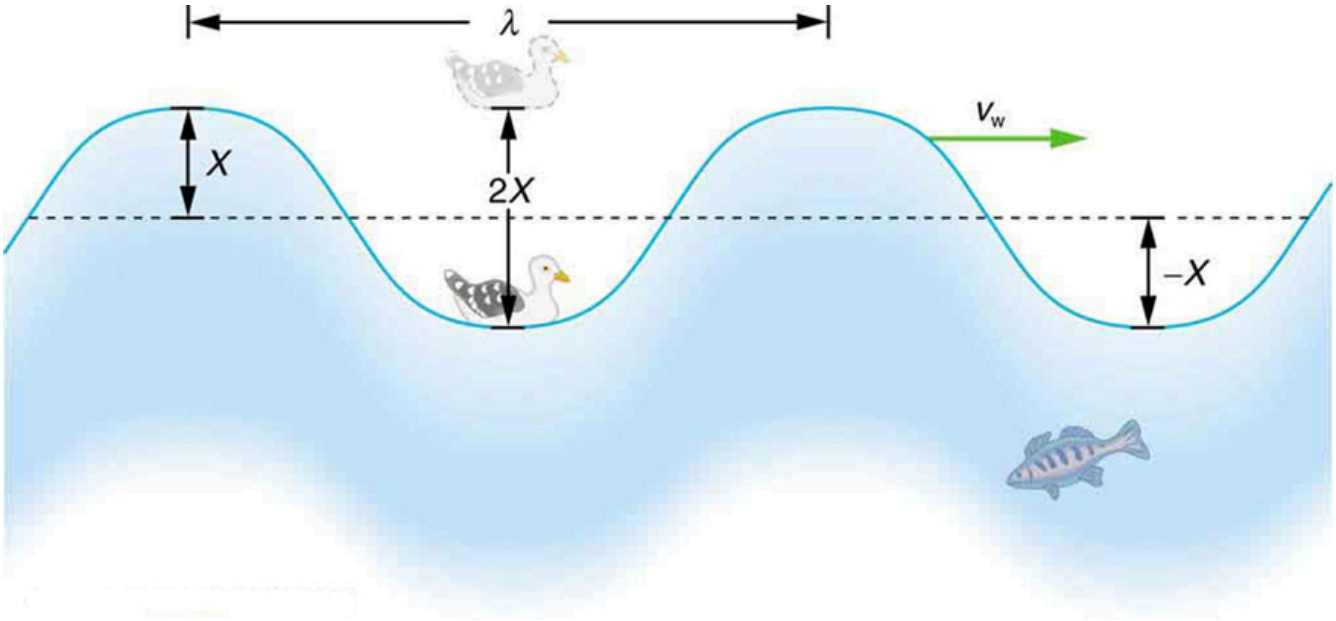


Figure 2. An idealized ocean wave passes under a sea gull that bobs up and down in simple harmonic motion. The wave has a wavelength λ, which is the distance between adjacent identical parts of the wave. The up and down disturbance of the surface propagates parallel to the surface at a speed V_w.

The water wave in the figure also has a length associated with it, called its *wavelength* λ, the distance between adjacent identical parts of a wave. (λ is the distance parallel to the direction of propagation.) The speed of propagation v_w is the distance the wave travels in a given time, which is one wavelength in the time of one period. In equation form, that is

$$v_w = \frac{\lambda}{T}$$

or $v_w = f\lambda$.

This fundamental relationship holds for all types of waves. For water waves, v_w is the speed of a surface wave; for sound, v_w is the speed of sound; and for visible light, v_w is the speed of light, for example.

Take-Home Experiment: Waves in a Bowl

Fill a large bowl or basin with water and wait for the water to settle so there are no ripples. Gently drop a cork into the middle of the bowl. Estimate the wavelength and period of oscillation of the water wave that propagates away from the cork. Remove the cork from the bowl and wait for the water to settle again. Gently drop the cork at a height that is different from the first drop. Does the wavelength depend upon how high above the water the cork is dropped?

Example 1. Calculate the Velocity of Wave Propagation: Gull in the Ocean

Calculate the wave velocity of the ocean wave in Figure 2 if the distance between wave crests is 10.0 m and the time for a sea gull to bob up and down is 5.00 s.

Strategy

We are asked to find v_w. The given information tells us that λ = 10.0m and T = 5.00 s. Therefore, we can use

$$v_w = \frac{\lambda}{T}$$

to find the wave velocity.

Solution

Enter the known values into

$$v_w = \frac{\lambda}{T}$$

:

$$v_w = \frac{10.0 \text{ m}}{5.00 \text{ s}}$$

Solve for v_w to find v_w= 2.00 m/s.

Discussion

This slow speed seems reasonable for an ocean wave. Note that the wave moves to the right in the figure at this speed, not the varying speed at which the sea gull moves up and down.

Transverse and Longitudinal Waves

A simple wave consists of a periodic disturbance that propagates from one place to another. The wave in Figure 3 propagates in the horizontal direction while the surface is disturbed in the vertical direction. Such a wave is called a *transverse wave* or shear wave; in such a wave, the disturbance is perpendicular to the direction of propagation. In contrast, in a *longitudinal wave* or compressional wave, the disturbance is parallel to the direction of propagation. Figure 4 shows an example of a longitudinal wave. The size of the disturbance is its amplitude X and is completely independent of the speed of propagation v_w.

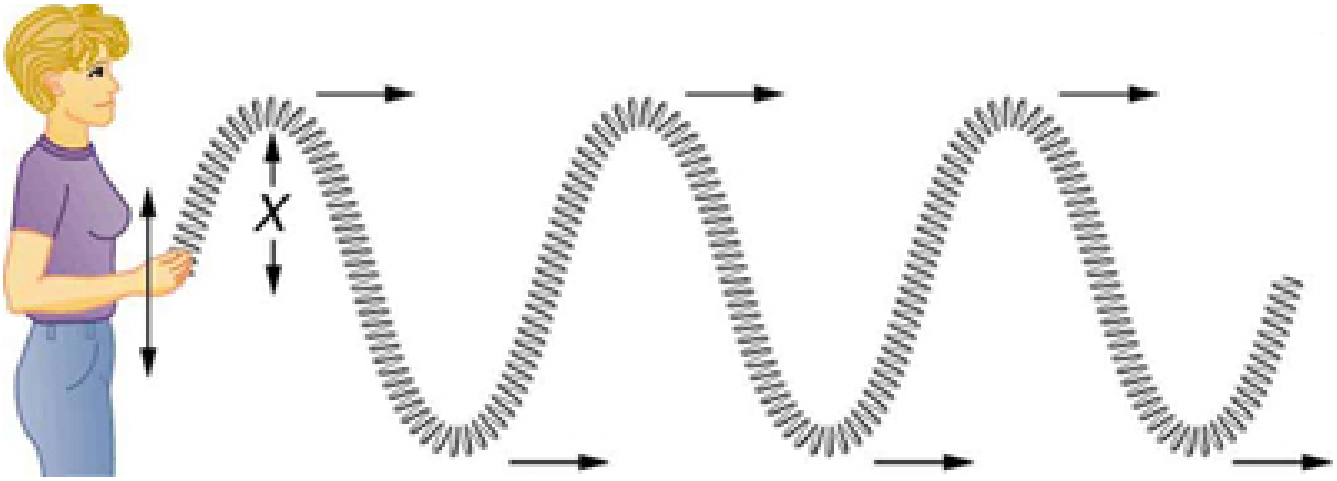


Figure 3. In this example of a transverse wave, the wave propagates horizontally, and the disturbance in the cord is in the vertical direction.

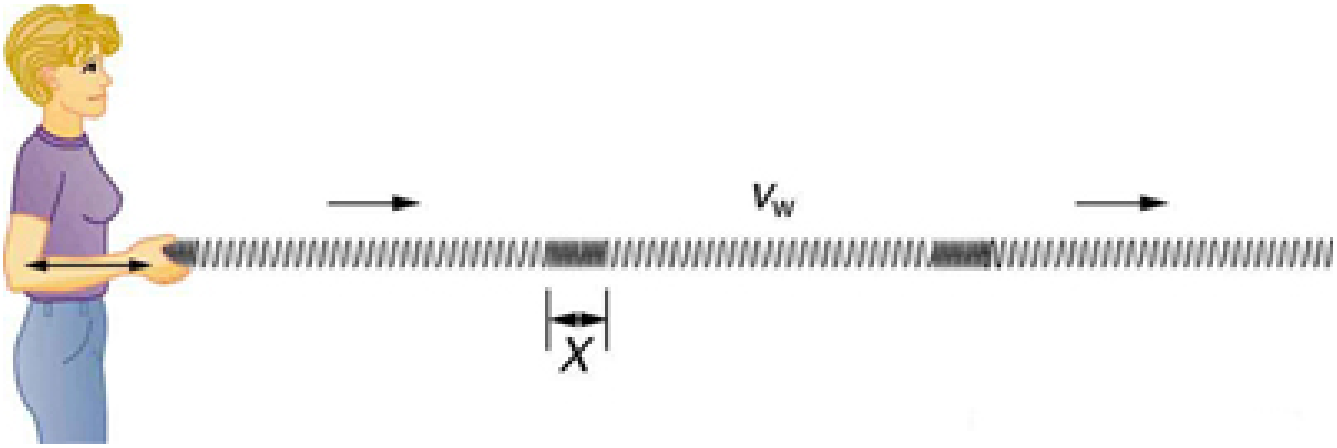


Figure 4. In this example of a longitudinal wave, the wave propagates horizontally, and the disturbance in the cord is also in the horizontal direction.

Waves may be transverse, longitudinal, or *a combination of the two*. (Water waves are actually a combination of transverse and longitudinal. The simplified water wave illustrated in Figure 2 shows no longitudinal motion of the bird.) The waves on the strings of musical instruments are transverse—so are electromagnetic waves, such as visible light.

Sound waves in air and water are longitudinal. Their disturbances are periodic variations in pressure that are transmitted in fluids. Fluids do not have appreciable shear strength, and thus the sound waves in them must be longitudinal or compressional. Sound in solids can be both longitudinal and transverse.

Earthquake waves under Earth's surface also have both longitudinal and transverse components (called compressional or P-waves and shear or S-waves, respectively). These components have important individual characteristics—they propagate at different speeds, for example. Earthquakes also have surface waves that are similar to surface waves on water.

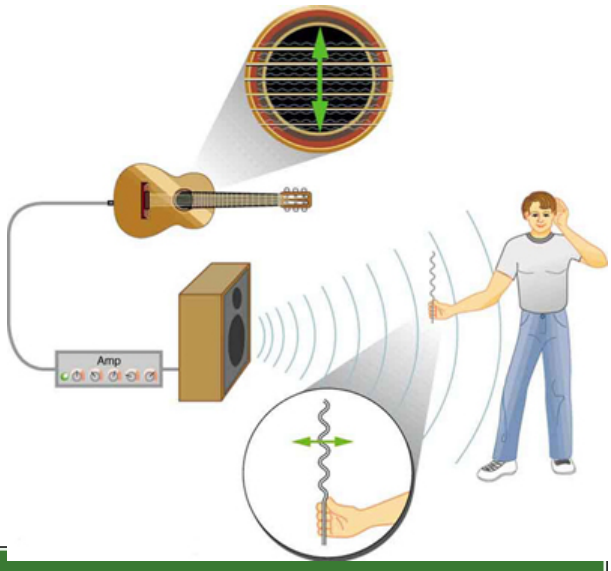


Figure 5. The wave on a guitar string is transverse. The sound wave rattles a sheet of paper in a direction that shows the sound wave is longitudinal.

Check Your Understanding

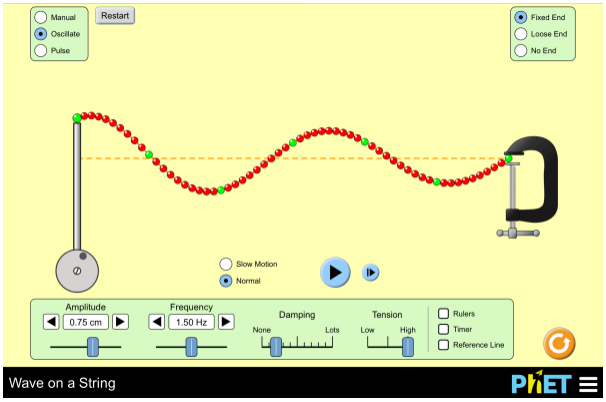
Why is it important to differentiate between longitudinal and transverse waves?

Solution

In the different types of waves, energy can propagate in a different direction relative to the motion of the wave. This is important to understand how different types of waves affect the materials around them.

PhET Explorations: Wave on a String

Watch a string vibrate in slow motion. Wiggle the end of the string and make waves, or adjust the frequency and amplitude of an oscillator. Adjust the damping and tension. The end can be fixed, loose, or open.



Click to run the simulation.

Section Summary

- A wave is a disturbance that moves from the point of creation with a wave velocity v_w.
- A wave has a wavelength λ, which is the distance between adjacent identical parts of the wave.
- Wave velocity and wavelength are related to the wave's frequency and period by $v_w = \frac{\lambda}{T}$ or $v_w = f\lambda$.
- A transverse wave has a disturbance perpendicular to its direction of propagation, whereas a longitudinal wave has a disturbance parallel to its direction of propagation.

Conceptual Questions

1. Give one example of a transverse wave and another of a longitudinal wave, being careful to note the relative directions of the disturbance and wave propagation in each.
2. What is the difference between propagation speed and the frequency of a wave? Does one or both affect wavelength? If so, how?

Problems & Exercises

1. Storms in the South Pacific can create waves that travel all the way to the California coast, which are 12,000 km away. How long does it take them if they travel at 15.0 m/s?
2. Waves on a swimming pool propagate at 0.750 m/s. You splash the water at one end of the pool

and observe the wave go to the opposite end, reflect, and return in 30.0 s. How far away is the other end of the pool?

3. Wind gusts create ripples on the ocean that have a wavelength of 5.00 cm and propagate at 2.00 m/s. What is their frequency?
4. How many times a minute does a boat bob up and down on ocean waves that have a wavelength of 40.0 m and a propagation speed of 5.00 m/s?
5. Scouts at a camp shake the rope bridge they have just crossed and observe the wave crests to be 8.00 m apart. If they shake it the bridge twice per second, what is the propagation speed of the waves?
6. What is the wavelength of the waves you create in a swimming pool if you splash your hand at a rate of 2.00 Hz and the waves propagate at 0.800 m/s?
7. What is the wavelength of an earthquake that shakes you with a frequency of 10.0 Hz and gets to another city 84.0 km away in 12.0 s?
8. Radio waves transmitted through space at 3.00×10^8 m/s by the Voyager spacecraft have a wavelength of 0.120 m. What is their frequency?
9. Your ear is capable of differentiating sounds that arrive at the ear just 1.00 ms apart. What is the minimum distance between two speakers that produce sounds that arrive at noticeably different times on a day when the speed of sound is 340 m/s?
10. (a) Seismographs measure the arrival times of earthquakes with a precision of 0.100 s. To get the distance to the epicenter of the quake, they compare the arrival times of S- and P-waves, which travel at different speeds. [link]) If S- and P-waves travel at 4.00 and 7.20 km/s, respectively, in the region considered, how precisely can the distance to the source of the earthquake be determined? (b) Seismic waves from underground detonations of nuclear bombs can be used to locate the test site and detect violations of test bans. Discuss whether your answer to (a) implies a serious limit to such detection. (Note also that the uncertainty is greater if there is an uncertainty in the propagation speeds of the S- and P-waves.)

Figure 7. A seismograph as described in above problem.(credit: Oleg Alexandrov)

Glossary

longitudinal wave: a wave in which the disturbance is parallel to the direction of propagation

transverse wave: a wave in which the disturbance is perpendicular to the direction of propagation

wave velocity: the speed at which the disturbance moves. Also called the propagation velocity or propagation speed

wavelength: the distance between adjacent identical parts of a wave

Selected Solutions to Problems & Exercises

1. $t = 9.26$ d
3. $f = 40.0$ Hz
5. $v_w = 16.0$ m/s
7. $\lambda = 700$ m
9. $d = 34.0$ cm

Superposition and Interference

Lumen Learning

Learning Objectives

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

- Explain standing waves.
- Describe the mathematical representation of overtones and beat frequency.

Most waves do not look very simple. They look more like the waves in Figure 1 than like the simple water wave considered in Waves. (Simple waves may be created by a simple harmonic oscillation, and thus have a sinusoidal shape). Complex waves are more interesting, even beautiful, but they look formidable. Most waves appear complex because they result from several simple waves adding together. Luckily, the rules for adding waves are quite simple.

Figure 1. These waves result from the superposition of several waves from different sources, producing a complex pattern. (credit: waterborough, Wikimedia Commons)

When two or more waves arrive at the same point, they superimpose themselves on one another. More specifically, the disturbances of waves are superimposed when they come together—a phenomenon called *superposition*. Each disturbance corresponds to a force, and forces add. If the disturbances are along the same line, then the resulting wave is a simple addition of the disturbances of the individual waves—that is, their amplitudes add. Figure 2 and Figure 3 illustrate superposition in two special cases, both of which produce simple results.

Figure 2 shows two identical waves that arrive at the same point exactly in phase. The crests of the two waves are precisely aligned, as are the troughs. This superposition produces pure *constructive interference*. Because the disturbances add, pure constructive interference produces a wave that has twice the amplitude of the individual waves, but has the same wavelength.

Figure 3 shows two identical waves that arrive exactly out of phase—that is, precisely aligned crest to trough—producing pure *destructive interference*. Because the disturbances are in the opposite direction for this superposition, the resulting amplitude is zero for pure destructive interference—the waves completely cancel.

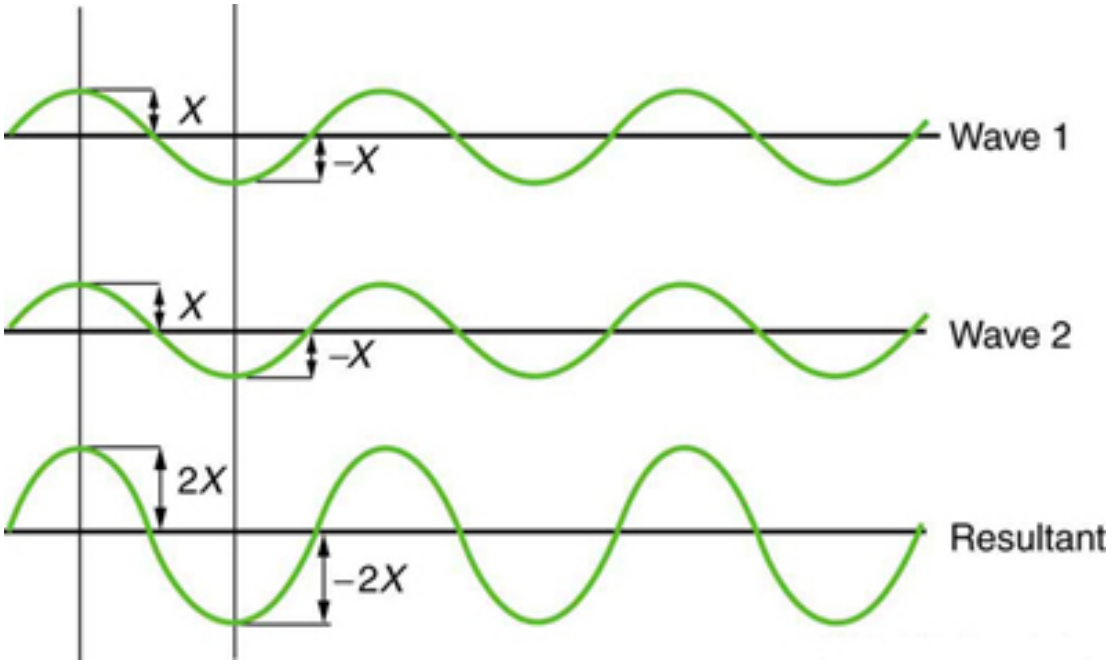


Figure 2. Pure constructive interference of two identical waves produces one with twice the amplitude, but the same wavelength.

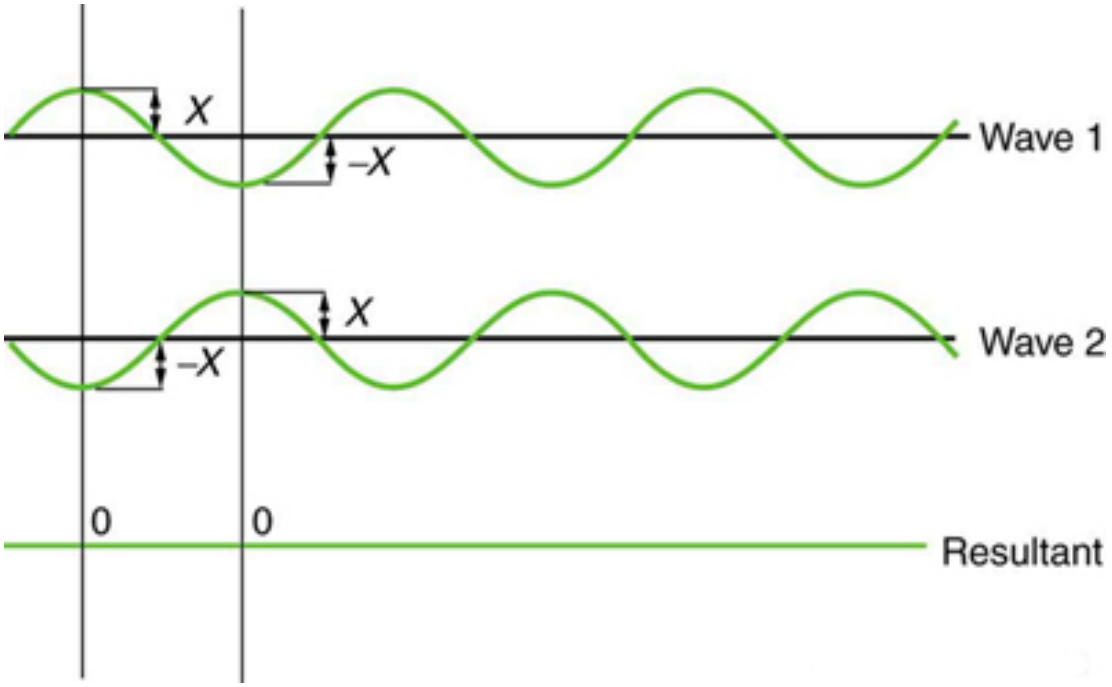


Figure 3. Pure destructive interference of two identical waves produces zero amplitude, or complete cancellation.

While pure constructive and pure destructive interference do occur, they require precisely aligned identical waves. The superposition of most waves produces a combination of constructive and destructive interference and can vary from place to place and time to time. Sound from a stereo, for example, can be loud in one spot and quiet in another. Varying loudness means the sound waves add partially constructively and partially destructively at different locations. A stereo has at least two speakers creating sound waves, and waves can reflect from walls. All these waves superimpose. An example of sounds that vary over time from constructive to destructive is found in the combined whine of airplane jets heard by a stationary passenger. The combined sound can fluctuate up and down in volume as the sound from the two engines varies in time from constructive to destructive. These examples are of waves that are similar.An example of the superposition of two dissimilar waves is shown in Figure 4. Here again, the disturbances add and subtract, producing a more complicated looking wave.

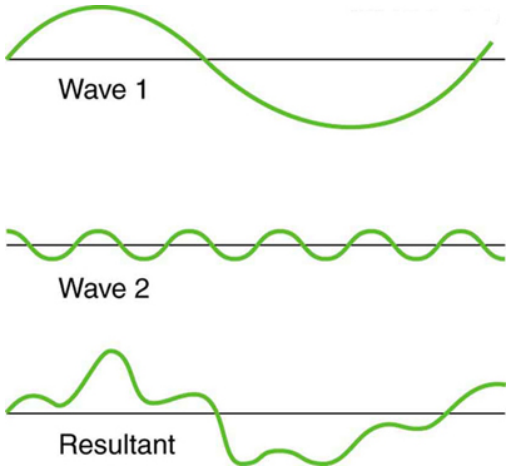


Figure 4. Superposition of non-identical waves exhibits both constructive and destructive interference.

Standing Waves

Sometimes waves do not seem to move; rather, they just vibrate in place. Unmoving waves can be seen on the surface of a glass of milk in a refrigerator, for example. Vibrations from the refrigerator motor create waves on the milk that oscillate up and down but do not seem to move across the surface. These waves are formed by the superposition of two or more moving waves, such as illustrated in Figure 5 for two identical waves moving in opposite directions. The waves move through each other with their disturbances adding as they go by. If the two waves have the same amplitude and wavelength, then they alternate between constructive and destructive interference. The resultant looks like a wave standing in place and, thus, is called a *standing wave*. Waves on the glass of milk are one example of standing waves. There are other standing waves, such as on guitar strings and in organ pipes. With the glass of milk, the two waves that produce standing waves may come from reflections from the side of the glass.

A closer look at earthquakes provides evidence for conditions appropriate for resonance, standing waves, and constructive and destructive interference. A building may be vibrated for several seconds with a driving frequency matching that of the natural frequency of vibration of the building—producing a resonance resulting in one building collapsing while neighboring buildings do not. Often buildings of a certain height are devastated while other taller buildings remain intact. The building height matches the condition for setting up a standing wave for that particular height. As the earthquake waves travel along the surface of Earth and reflect off denser rocks, constructive interference occurs at certain points. Often areas closer to the epicenter are not damaged while areas farther away are damaged.

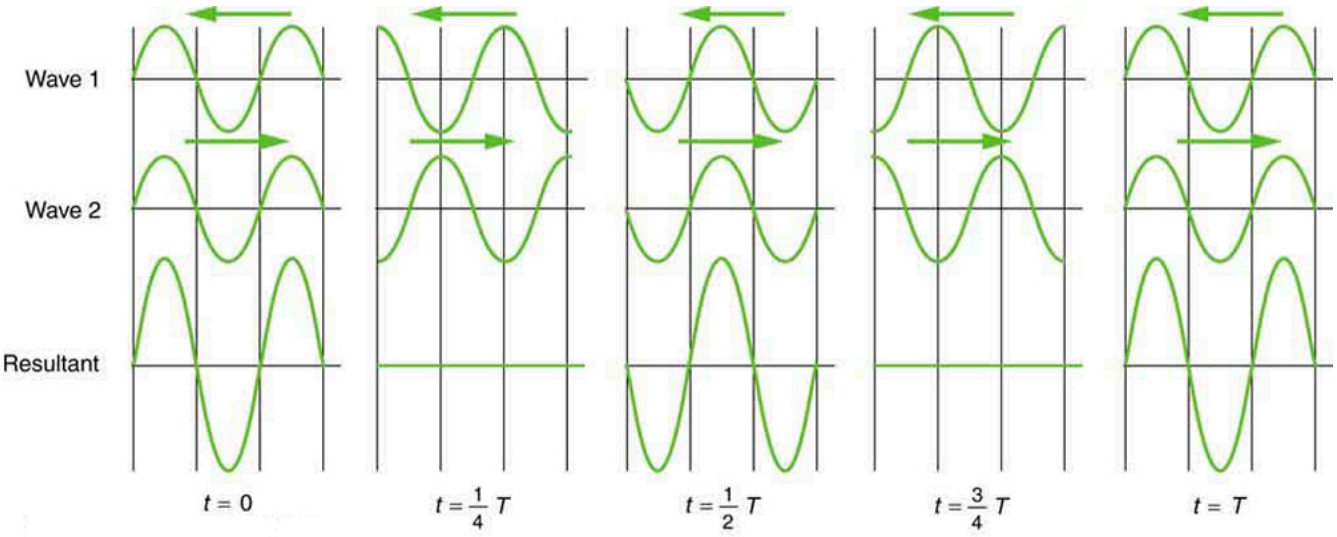


Figure 5. Standing wave created by the superposition of two identical waves moving in opposite directions. The oscillations are at fixed locations in space and result from alternately constructive and destructive interference.

Standing waves are also found on the strings of musical instruments and are due to reflections of waves from the ends of the string. Figures 6 and 7 show three standing waves that can be created on a string that is fixed at both ends. *Nodes* are the points where the string does not move; more generally, nodes are where the wave disturbance is zero in a standing wave. The fixed ends of strings must be nodes, too, because the string cannot move there. The word *antinode* is used to denote the location of maximum amplitude in standing waves. Standing waves on strings have a frequency that is related to the propagation speed v_w of the disturbance on the string. The wavelength λ is determined by the distance between the points where the string is fixed in place.

The lowest frequency, called the *fundamental frequency*, is thus for the longest wavelength, which is seen to be $\lambda_1 = 2L$. Therefore, the fundamental frequency is

$$f_1 = \frac{v_w}{\lambda_1} = \frac{v_w}{2L}$$

. In this case, the *overtones* or harmonics are multiples of the fundamental frequency. As seen in Figure 7, the first harmonic can easily be calculated since $\lambda_2 = L$. Thus,

$$f_2 = \frac{v_w}{\lambda_2} = \frac{v_w}{2L} = 2f_1$$

. Similarly, $f_3 = 3f_1$, and so on. All of these frequencies can be changed by adjusting the tension in the string. The greater the tension, the greater v_w is and the higher the frequencies. This observation is familiar to anyone who has ever observed a string instrument being tuned. We will see in later chapters that standing waves are crucial to many resonance phenomena, such as in sounding boxes on string instruments.

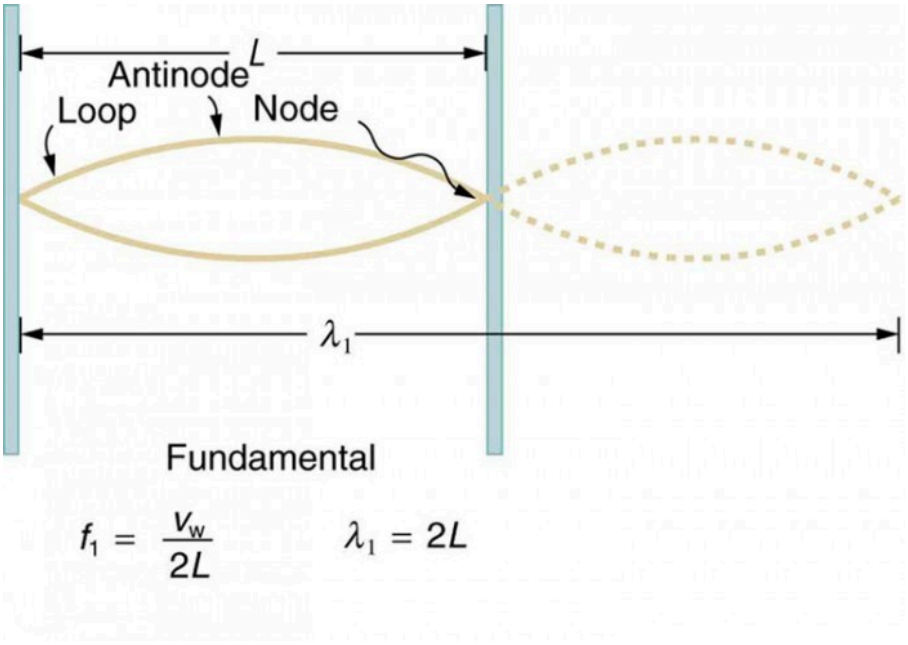


Figure 6. The figure shows a string oscillating at its fundamental frequency.

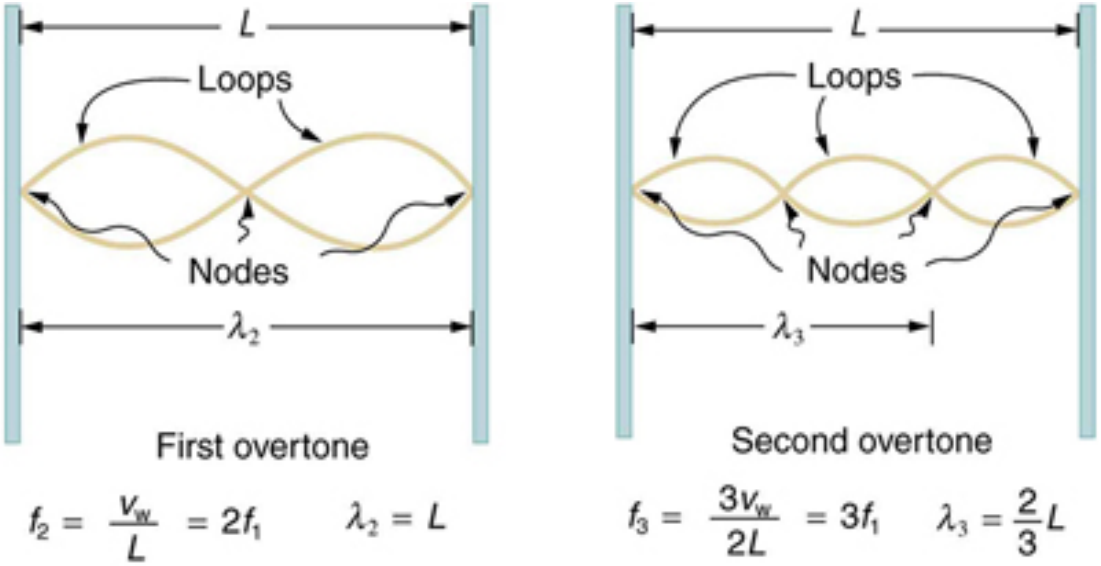


Figure 7. First and second harmonic frequencies are shown.

Beats

Striking two adjacent keys on a piano produces a warbling combination usually considered to be unpleasant. The superposition of two waves of similar but not identical frequencies is the culprit. Another example is often noticeable in jet aircraft, particularly the two-engine variety, while taxiing. The combined sound of the engines goes up and down in loudness. This varying loudness happens because the sound waves have similar but not identical frequencies. The discordant warbling of the piano and the fluctuating loudness of the jet engine noise are both due to alternately constructive and destructive interference as the two waves go in and out of phase. Figure 8 illustrates this graphically.

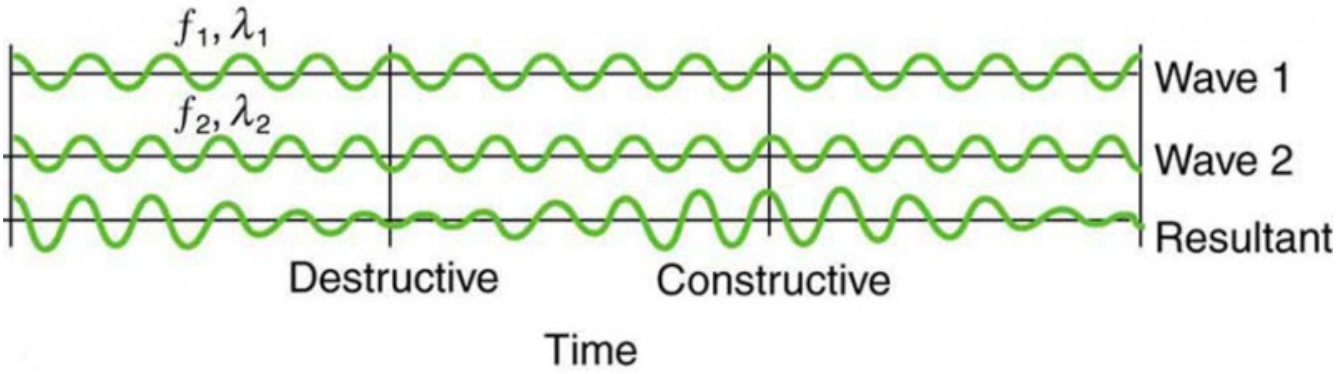


Figure 8. Beats are produced by the superposition of two waves of slightly different frequencies but identical amplitudes. The waves alternate in time between constructive interference and destructive interference, giving the resulting wave a time-varying amplitude.

The wave resulting from the superposition of two similar-frequency waves has a frequency that is the average of the two. This wave fluctuates in amplitude, or *beats*, with a frequency called the *beat frequency*. We can determine the beat frequency by adding two waves together mathematically. Note that a wave can be represented at one point in space as

$$x = X\cos\left(\frac{2\pi t}{T}\right) = X\cos\left(2\pi ft\right)$$

, where

$$f = \frac{1}{T}$$

is the frequency of the wave. Adding two waves that have different frequencies but identical amplitudes produces a resultant $x = x_1 + x_2$. More specifically, $x = X\cos(2\pi f_1 t) + X\cos(2\pi f_2 t)$.

Using a trigonometric identity, it can be shown that $x = 2X\cos(\pi f_B t)\cos(2\pi f_{ave} t)$, where $f_B = |f_1 - f_2|$ is the beat frequency, and f_{ave} is the average of f_1 and f_2. These results mean that the resultant wave has twice the amplitude and the average frequency of the two superimposed waves, but it also fluctuates in overall amplitude at the beat frequency f_B. The first cosine term in the expression effectively causes the amplitude to go up and down. The second cosine term is the wave with frequency f_{ave}. This result is valid for all types of waves. However, if it is a sound wave, providing the two frequencies are similar, then what we hear is an average frequency that gets louder and softer (or warbles) at the beat frequency.

Making Career Connections

Piano tuners use beats routinely in their work. When comparing a note with a tuning fork, they listen for beats and adjust the string until the beats go away (to zero frequency). For example, if the tuning fork has a 256 Hz frequency and two beats per second are heard, then the other frequency is either 254 or 258 Hz. Most keys hit multiple strings, and these strings are actually adjusted until they have nearly the same frequency and give a slow beat for richness. Twelve-string guitars and mandolins are also tuned using beats.

While beats may sometimes be annoying in audible sounds, we will find that beats have many

applications. Observing beats is a very useful way to compare similar frequencies. There are applications of beats as apparently disparate as in ultrasonic imaging and radar speed traps.

Check Your Understanding

Part 1

Imagine you are holding one end of a jump rope, and your friend holds the other. If your friend holds her end still, you can move your end up and down, creating a transverse wave. If your friend then begins to move her end up and down, generating a wave in the opposite direction, what resultant wave forms would you expect to see in the jump rope?

Solution

The rope would alternate between having waves with amplitudes two times the original amplitude and reaching equilibrium with no amplitude at all. The wavelengths will result in both constructive and destructive interference

Part 2

Define nodes and antinodes.

Solution

Nodes are areas of wave interference where there is no motion. Antinodes are areas of wave interference where the motion is at its maximum point.

Part 3

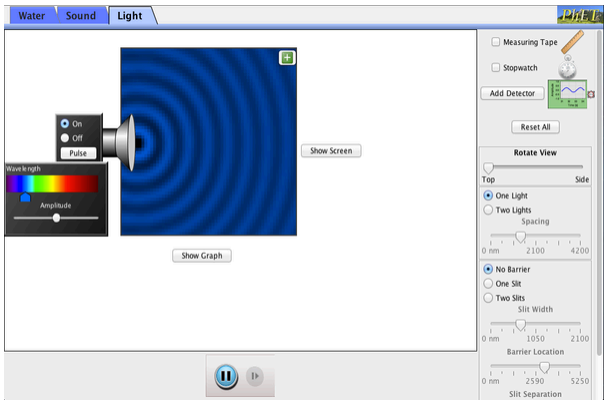
You hook up a stereo system. When you test the system, you notice that in one corner of the room, the sounds seem dull. In another area, the sounds seem excessively loud. Describe how the sound moving about the room could result in these effects.

Solution

With multiple speakers putting out sounds into the room, and these sounds bouncing off walls, there is bound to be some wave interference. In the dull areas, the interference is probably mostly destructive. In the louder areas, the interference is probably mostly constructive.

PhET Explorations: Wave Interference

Make waves with a dripping faucet, audio speaker, or laser! Add a second source or a pair of slits to create an interference pattern.



Click to download the simulation. Run using Java.

Section Summary

- Superposition is the combination of two waves at the same location.
- Constructive interference occurs when two identical waves are superimposed in phase.
- Destructive interference occurs when two identical waves are superimposed exactly out of phase.
- A standing wave is one in which two waves superimpose to produce a wave that varies in amplitude but does not propagate.
- Nodes are points of no motion in standing waves.
- An antinode is the location of maximum amplitude of a standing wave.
- Waves on a string are resonant standing waves with a fundamental frequency and can occur at higher multiples of the fundamental, called overtones or harmonics.
- Beats occur when waves of similar frequencies f_1 and f_2 are superimposed. The resulting amplitude oscillates with a beat frequency given by $f_B = |f_1 - f_2|$.

Conceptual Questions

1. Speakers in stereo systems have two color-coded terminals to indicate how to hook up the wires. If the wires are reversed, the speaker moves in a direction opposite that of a properly connected speaker. Explain why it is important to have both speakers connected the same way.

Problems & Exercises

1. A car has two horns, one emitting a frequency of 199 Hz and the other emitting a frequency of 203 Hz. What beat frequency do they produce?
2. The middle-C hammer of a piano hits two strings, producing beats of 1.50 Hz. One of the strings is tuned to 260.00 Hz. What frequencies could the other string have?
3. Two tuning forks having frequencies of 460 and 464 Hz are struck simultaneously. What average frequency will you hear, and what will the beat frequency be?
4. Twin jet engines on an airplane are producing an average sound frequency of 4100 Hz with a beat frequency of 0.500 Hz. What are their individual frequencies?
5. A wave traveling on a Slinky® that is stretched to 4 m takes 2.4 s to travel the length of the Slinky and back again. (a) What is the speed of the wave? (b) Using the same Slinky stretched to the same length, a standing wave is created which consists of three antinodes and four nodes. At what frequency must the Slinky be oscillating?
6. Three adjacent keys on a piano (F, F-sharp, and G) are struck simultaneously, producing frequencies of 349, 370, and 392 Hz. What beat frequencies are produced by this discordant combination?

Glossary

antinode: the location of maximum amplitude in standing waves

beat frequency: the frequency of the amplitude fluctuations of a wave

constructive interference: when two waves arrive at the same point exactly in phase; that is, the crests of the two waves are precisely aligned, as are the troughs

destructive interference: when two identical waves arrive at the same point exactly out of phase; that is, precisely aligned crest to trough

fundamental frequency: the lowest frequency of a periodic waveform

nodes: the points where the string does not move; more generally, nodes are where the wave disturbance is zero in a standing wave

overtones: multiples of the fundamental frequency of a sound

superposition: the phenomenon that occurs when two or more waves arrive at the same point

Selected Solutions to Problems & Exercises

1. $f = 4$ Hz

3. 462 Hz, 4 Hz

5. (a) 3.33 m/s; (b) 1.25 Hz

Energy in Waves: Intensity

Lumen Learning

Learning Objectives

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

- Calculate the intensity and the power of rays and waves.

Figure 1. The destructive effect of an earthquake is palpable evidence of the energy carried in these waves. The Richter scale rating of earthquakes is related to both their amplitude and the energy they carry. (credit: Petty Officer 2nd Class Candice Villarreal, U.S. Navy)

All waves carry energy. The energy of some waves can be directly observed. Earthquakes can shake whole cities to the ground, performing the work of thousands of wrecking balls.

Loud sounds pulverize nerve cells in the inner ear, causing permanent hearing loss. Ultrasound is used for deep-heat treatment of muscle strains. A laser beam can burn away a malignancy. Water waves chew up beaches.

The amount of energy in a wave is related to its amplitude. Large-amplitude earthquakes produce large ground displacements. Loud sounds have higher pressure amplitudes and come from larger-amplitude source vibrations than soft sounds. Large ocean breakers churn up the shore more than small ones. More quantitatively, a wave is a displacement that is resisted by a restoring force. The larger the displacement x, the larger the force $F = kx$ needed to create it. Because work W is related to force multiplied by

distance (Fx) and energy is put into the wave by the work done to create it, the energy in a wave is related to amplitude. In fact, a wave's energy is directly proportional to its amplitude squared because $W \propto Fx = kx^2$.

The energy effects of a wave depend on time as well as amplitude. For example, the longer deep-heat ultrasound is applied, the more energy it transfers. Waves can also be concentrated or spread out. Sunlight, for example, can be focused to burn wood. Earthquakes spread out, so they do less damage the farther they get from the source. In both cases, changing the area the waves cover has important effects. All these pertinent factors are included in the definition of *intensity I* as power per unit area:

$$I = \frac{P}{A}$$

, where P is the power carried by the wave through area A. The definition of intensity is valid for any energy in transit, including that carried by waves. The SI unit for intensity is watts per square meter (W/m^2). For example, infrared and visible energy from the Sun impinge on Earth at an intensity of 1300 W/m^2 just above the atmosphere. There are other intensity-related units in use, too. The most common is the decibel. For example, a 90 decibel sound level corresponds to an intensity of 10^{-3} W/m^2. (This quantity is not much power per unit area considering that 90 decibels is a relatively high sound level. Decibels will be discussed in some detail in a later chapter.

Example 1. Calculating intensity and power: How much energy is in a ray of sunlight?

The average intensity of sunlight on Earth's surface is about 500 W/m^2.

1. Calculate the amount of energy that falls on a solar collector having an area of 0.500 m^2 in 4.00 h.
2. What intensity would such sunlight have if concentrated by a magnifying glass onto an area 200 times smaller than its own?

Strategy for Part 1

Because power is energy per unit time or

$$P = \frac{E}{t}$$

, the definition of intensity can be written as

$$I = \frac{P}{A} = \frac{\frac{E}{t}}{A}$$

,

and this equation can be solved for E with the given information.

Solution to Part 1

Begin with the equation that states the definition of intensity:

$$I = \frac{P}{A}$$

Replace P with its equivalent

$$\frac{E}{t}$$

:

$$I = \frac{\frac{E}{t}}{A}$$

.

Solve for E: $E = IAt$.

Substitute known values into the equation: $E = (700\ \text{W/m}^2)(0.500\ \text{m}^2)[(4.00\ \text{h})(3600\ \text{s/h})]$.

Calculate to find E and convert units: 5.04×10^6 J.

Discussion for Part 1

The energy falling on the solar collector in 4 h in part is enough to be useful—for example, for heating a significant amount of water.

Strategy for Part 2

Taking a ratio of new intensity to old intensity and using primes for the new quantities, we will find that it depends on the ratio of the areas. All other quantities will cancel.

Solution to Part 2

Take the ratio of intensities, which yields:

$$\frac{I\prime}{I} = \frac{\frac{P\prime}{A\prime}}{\frac{P}{A}} = \frac{A}{A\prime}$$

(The powers cancel because $P' = P$.)

Identify the knowns:

- $A = 200A'$

  $\frac{I\prime}{I} = 200$
-

Substitute known quantities: $I' = 200I = 200(700\ \text{W/m}^2)$.

Calculate to find I': $I' = 1.40 \times 10^5\ \text{W/m}^2$.

Discussion for Part 2

Decreasing the area increases the intensity considerably. The intensity of the concentrated sunlight could even start a fire.

Example 2. Determine the combined intensity of two waves: Perfect constructive interference

If two identical waves, each having an intensity of 1.00 W/m^2, interfere perfectly constructively, what is the intensity of the resulting wave?

Strategy

We know from Superposition and Interference that when two identical waves, which have equal amplitudes X, interfere perfectly constructively, the resulting wave has an amplitude of $2X$. Because a wave's intensity is proportional to amplitude squared, the intensity of the resulting wave is four times as great as in the individual waves.

Solution

Recall that intensity is proportional to amplitude squared.

Calculate the new amplitude: $I' \propto (X')^2 = (2X)^2 = 4X^2$.

Recall that the intensity of the old amplitude was $I \propto X^2$.

Take the ratio of new intensity to the old intensity. This gives

$$\frac{I\prime}{I} = 4$$

.

Calculate to find $\mathbf{I'}$: $I' = 4I = 4.00$ W/m^2.

Discussion

The intensity goes up by a factor of 4 when the amplitude doubles. This answer is a little disquieting. The two individual waves each have intensities of 1.00 W/m^2, yet their sum has an intensity of 4.00 W/m^2, which may appear to violate conservation of energy. This violation, of course, cannot happen. What does happen is intriguing. The area over which the intensity is 4.00 W/m^2 is much less than the area covered by the two waves before they interfered. There are other areas where the intensity is zero. The addition of waves is not as simple as our first look in Superposition and Interference suggested. We actually get a pattern of both constructive interference and destructive interference whenever two waves are added. For example, if we have two stereo speakers putting out 1.00 W/m^2 each, there will be places in the room where the intensity is 4.00 W/m^2, other places where the intensity is zero, and others in between. Figure 2 shows what this interference might look like. We will pursue interference patterns elsewhere in this text.

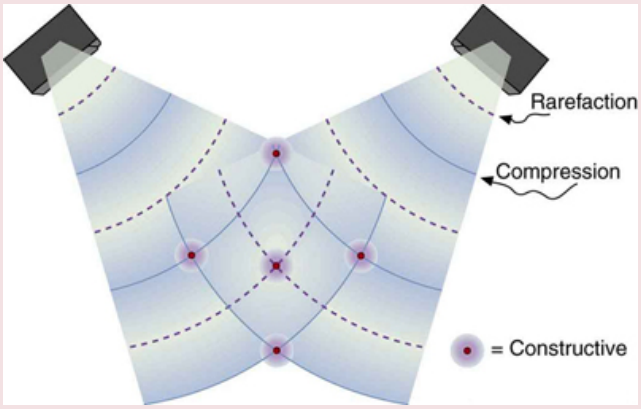


Figure 2. These stereo speakers produce both constructive interference and destructive interference in the room, a property common to the superposition of all types of waves. The shading is proportional to intensity.

Check Your Understanding

Which measurement of a wave is most important when determining the wave's intensity?

Solution

Amplitude, because a wave's energy is directly proportional to its amplitude squared.

Section Summary

- Intensity is defined to be the power per unit area: $I = \frac{P}{A}$ and has units of W/m^2.

Conceptual Questions

1. Two identical waves undergo pure constructive interference. Is the resultant intensity twice that of the individual waves? Explain your answer.
2. Circular water waves decrease in amplitude as they move away from where a rock is dropped. Explain why.

Problems & Exercises

1. **Medical Application.** Ultrasound of intensity 1.50×10^2 W/m^2 is produced by the rectangular head of a medical imaging device measuring 3.00 by 5.00 cm. What is its power output?
2. The low-frequency speaker of a stereo set has a surface area of 0.05 m^2 and produces 1W of acoustical power. What is the intensity at the speaker? If the speaker projects sound uniformly in all directions, at what distance from the speaker is the intensity 0.1 W/m^2?
3. To increase intensity of a wave by a factor of 50, by what factor should the amplitude be increased?
4. **Engineering Application.** A device called an insolation meter is used to measure the intensity of sunlight has an area of 100 cm2 and registers 6.50 W. What is the intensity in W/m^2?
5. **Astronomy Application.** Energy from the Sun arrives at the top of the Earth's atmosphere with an intensity of 1.30 kW/m^2. How long does it take for 1.8×10^9 J to arrive on an area of 1.00 m^2?
6. Suppose you have a device that extracts energy from ocean breakers in direct proportion to their intensity. If the device produces 10.0 kW of power on a day when the breakers are 1.20 m high, how much will it produce when they are 0.600 m high?
7. **Engineering Application.** (a) A photovoltaic array of (solar cells) is 10.0% efficient in gathering solar energy and converting it to electricity. If the average intensity of sunlight on one day is 700 W/m^2, what area should your array have to gather energy at the rate of 100 W? (b) What is the maximum cost of the array if it must pay for itself in two years of operation averaging 10.0 hours

per day? Assume that it earns money at the rate of 9.00 ¢ per kilowatt-hour.

8. A microphone receiving a pure sound tone feeds an oscilloscope, producing a wave on its screen. If the sound intensity is originally 2.00×10^{-5} W/m^2, but is turned up until the amplitude increases by 30.0%, what is the new intensity?
9. **Medical Application.** (a) What is the intensity in W/m^2 of a laser beam used to burn away cancerous tissue that, when 90.0% absorbed, puts 500 J of energy into a circular spot 2.00 mm in diameter in 4.00 s? (b) Discuss how this intensity compares to the average intensity of sunlight (about 700 W/m^2) and the implications that would have if the laser beam entered your eye. Note how your answer depends on the time duration of the exposure.

Glossary

intensity: power per unit area

Selected Solutions to Problems & Exercises

1. 0.225 W
3. 7.07
5. 16.0 d
6. 2.50 kW
8. 3.38×10^{-5} W/m^2

7. Physics of Hearing

Introduction to the Physics of Hearing

Lumen Learning

Figure 1. This tree fell some time ago. When it fell, atoms in the air were disturbed. Physicists would call this disturbance sound whether someone was around to hear it or not. (credit: B.A. Bowen Photography)

If a tree falls in the forest and no one is there to hear it, does it make a sound? The answer to this old philosophical question depends on how you define sound. If sound only exists when someone is around to perceive it, then there was no sound. However, if we define sound in terms of physics; that is, a disturbance of the atoms in matter transmitted from its origin outward (in other words, a wave), then there *was* a sound, even if nobody was around to hear it.

Such a wave is the physical phenomenon we call *sound*. Its perception is hearing. Both the physical phenomenon and its perception are interesting and will be considered in this text. We shall explore both sound and hearing; they are related, but are not the same thing. We will also explore the many practical uses of sound waves, such as in medical imaging.

Video: Waves and Sound

Lumen Learning

Watch the following Physics Concept Trailer to learn more about sound waves.

A YouTube element has been excluded from this version of the text. You can view it online here: https://pressbooks.nscc.ca/heatlightsound/?p=149

Sound

Lumen Learning

Learning Objectives

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

- Define sound and hearing.
- Describe sound as a longitudinal wave.

Sound can be used as a familiar illustration of waves. Because hearing is one of our most important senses, it is interesting to see how the physical properties of sound correspond to our perceptions of it. *Hearing* is the perception of sound, just as vision is the perception of visible light. But sound has important applications beyond hearing. Ultrasound, for example, is not heard but can be employed to form medical images and is also used in treatment.

Figure 1. This glass has been shattered by a high-intensity sound wave of the same frequency as the resonant frequency of the glass. While the sound is not visible, the effects of the sound prove its existence. (credit: ||read||, Flickr)

The physical phenomenon of *sound* is defined to be a disturbance of matter that is transmitted from its source outward. Sound is a wave. On the atomic scale, it is a disturbance of atoms that is far more ordered than their thermal motions. In many instances, sound is a periodic wave, and the atoms undergo simple harmonic motion. In this text, we shall explore such periodic sound waves.

A vibrating string produces a sound wave as illustrated in Figure 2. As the string oscillates back and forth, it transfers energy to the air, mostly as thermal energy created by turbulence. But a small part of the string's energy goes into compressing and expanding the surrounding air, creating slightly higher and lower local pressures. These compressions (high pressure regions) and rarefactions (low pressure regions) move out as longitudinal pressure waves having the same frequency as the string—they are the disturbance that is a sound wave. (Sound waves in air and most fluids are longitudinal, because fluids have almost no shear strength. In solids, sound waves can be both transverse and longitudinal.) Figure 2c shows a graph of gauge pressure versus distance from the vibrating string.

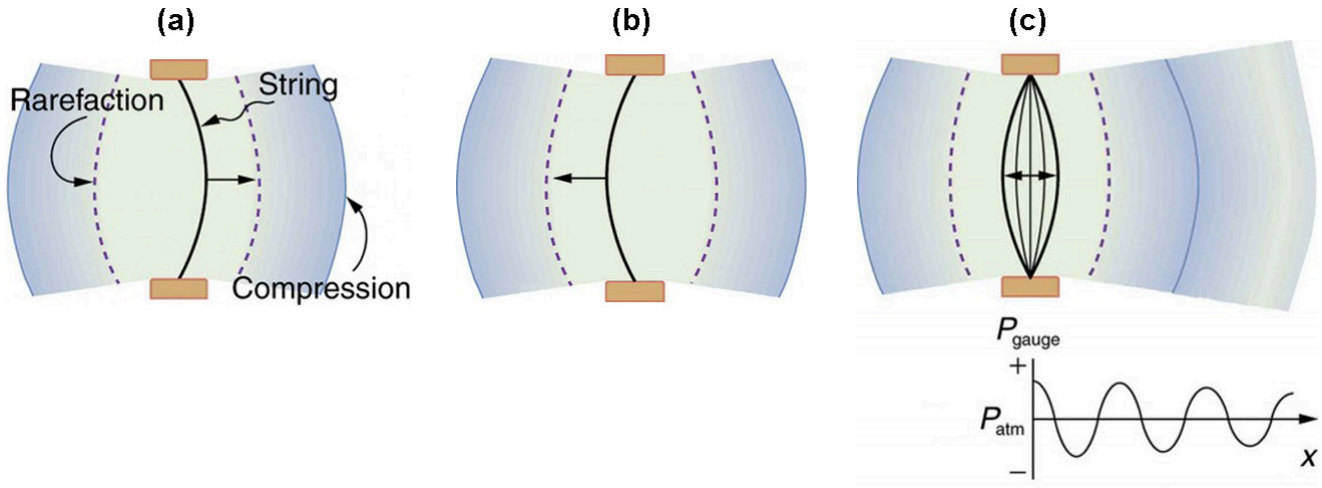


Figure 2. (a) A vibrating string moving to the right compresses the air in front of it and expands the air behind it. (b) As the string moves to the left, it creates another compression and rarefaction as the ones on the right move away from the string. (c) After many vibrations, there are a series of compressions and rarefactions moving out from the string as a sound wave. The graph shows gauge pressure versus distance from the source. Pressures vary only slightly from atmospheric for ordinary sounds.

The amplitude of a sound wave decreases with distance from its source, because the energy of the wave is spread over a larger and larger area. But it is also absorbed by objects, such as the eardrum in Figure 3, and converted to thermal energy by the viscosity of air. In addition, during each compression a little heat transfers to the air and during each rarefaction even less heat transfers from the air, so that the heat transfer reduces the organized disturbance into random thermal motions. (These processes can be viewed as a manifestation of the second law of thermodynamics presented in Introduction to the Second Law of Thermodynamics: Heat Engines and Their Efficiency.)

Whether the heat transfer from compression to rarefaction is significant depends on how far apart they are—that is, it depends on wavelength. Wavelength, frequency, amplitude, and speed of propagation are important for sound, as they are for all waves.

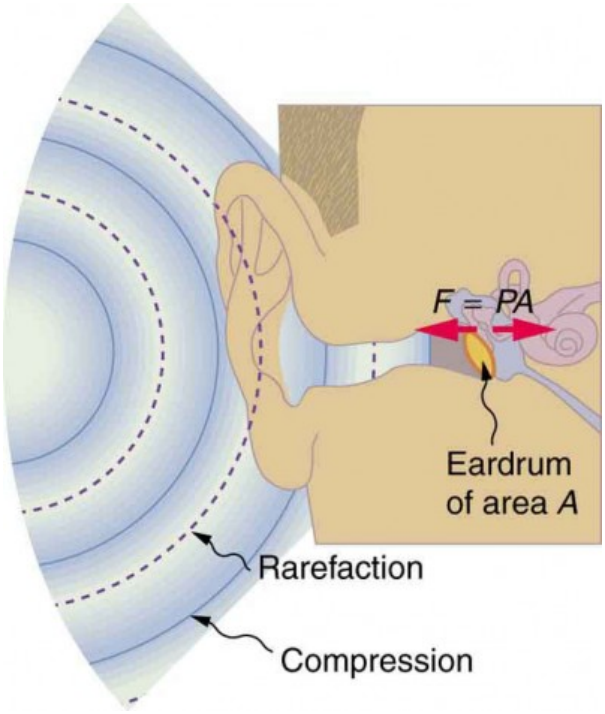
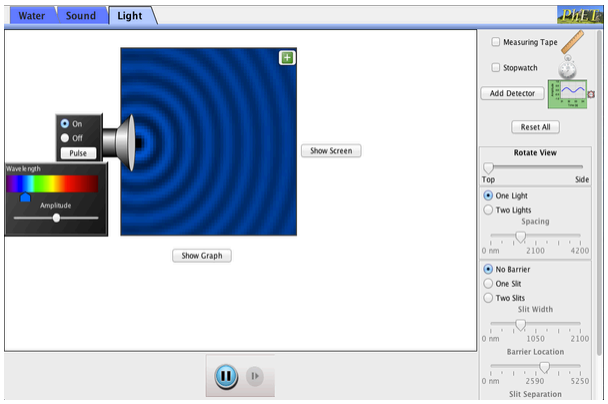


Figure 3. Sound wave compressions and rarefactions travel up the ear canal and force the eardrum to vibrate. There is a net force on the eardrum, since the sound wave pressures differ from the atmospheric pressure found behind the eardrum. A complicated mechanism converts the vibrations to nerve impulses, which are perceived by the person.

PhET Explorations: Wave Interference

Make waves with a dripping faucet, audio speaker, or laser! Add a second source or a pair of slits to create an interference pattern.



Click to download the simulation. Run using Java.

Section Summary

- Sound is a disturbance of matter that is transmitted from its source outward.
- Sound is one type of wave.
- Hearing is the perception of sound.

Glossary

sound: a disturbance of matter that is transmitted from its source outward

hearing: the perception of sound

Speed of Sound, Frequency, and Wavelength

Lumen Learning

Learning Objectives

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

- Define pitch.
- Describe the relationship between the speed of sound, its frequency, and its wavelength.
- Describe the effects on the speed of sound as it travels through various media.
- Describe the effects of temperature on the speed of sound.

Sound, like all waves, travels at a certain speed and has the properties of frequency and wavelength. You can observe direct evidence of the speed of sound while watching a fireworks display. The flash of an explosion is seen well before its sound is heard, implying both that sound travels at a finite speed and that it is much slower than light. You can also directly sense the frequency of a sound. Perception of frequency is called *pitch*. The wavelength of sound is not directly sensed, but indirect evidence is found in the correlation of the size of musical instruments with their pitch. Small instruments, such as a piccolo, typically make high-pitch sounds, while large instruments, such as a tuba, typically make low-pitch sounds. High pitch means small wavelength, and the size of a musical instrument is directly related to the wavelengths of sound it produces. So a small instrument creates short-wavelength sounds. Similar arguments hold that a large instrument creates long-wavelength sounds.

Figure 1. When a firework explodes, the light energy is perceived before the sound energy. Sound travels more slowly than light does. (credit: Dominic Alves, Flickr)

The relationship of the speed of sound, its frequency, and wavelength is the same as for all waves: $v_w = f\lambda$, where v_w is the speed of sound, f is its frequency, and λ is its wavelength. The wavelength of a sound is the distance between adjacent identical parts of a wave—for example, between adjacent compressions as illustrated in Figure 2. The frequency is the same as that of the source and is the number of waves that pass a point per unit time.

Table 1 makes it apparent that the speed of sound varies greatly in different media. The speed of sound in a medium is determined by a combination of the medium's rigidity (or compressibility in gases) and its density. The more rigid (or less compressible) the medium, the faster the speed of sound. This observation is analogous to the fact that the frequency of a simple harmonic motion is directly proportional to the stiffness of the oscillating object. The greater the density of a medium, the slower the speed of sound. This observation is analogous to the fact that the frequency of a simple harmonic motion is inversely proportional to the mass of the oscillating object. The speed of sound in air is low, because air is compressible. Because liquids and solids are relatively rigid and very difficult to compress, the speed of sound in such media is generally greater than in gases.

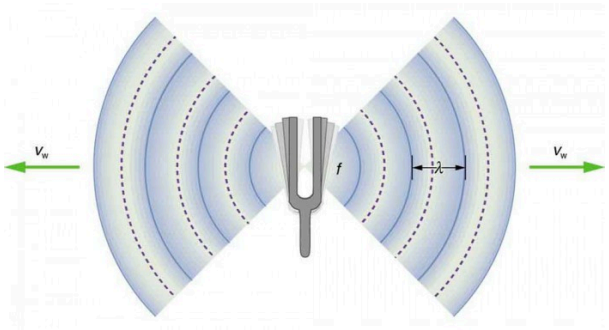


Figure 2. A sound wave emanates from a source vibrating at a frequency f, propagates at V_w, and has a wavelength λ.

Table 1. Speed of Sound in Various Media

Medium	**v_w(m/s)**
Gases at 0°C	
Air	331
Carbon dioxide	259
Oxygen	316
Helium	965
Hydrogen	1290
Liquids at 20°C	
Ethanol	1160
Mercury	1450
Water, fresh	1480
Sea water	1540
Human tissue	1540
Solids (longitudinal or bulk)	
Vulcanized rubber	54
Polyethylene	920
Marble	3810
Glass, Pyrex	5640
Lead	1960
Aluminum	5120
Steel	5960

Earthquakes, essentially sound waves in Earth's crust, are an interesting example of how the speed of sound depends on the rigidity of the medium. Earthquakes have both longitudinal and transverse components, and these travel at different speeds. The bulk modulus of granite is greater than its shear modulus. For that reason, the speed of longitudinal or pressure waves (P-waves) in earthquakes in granite is significantly higher than the speed of transverse or shear waves (S-waves). Both components of earthquakes travel slower in less rigid material, such as sediments. P-waves have speeds of 4 to 7 km/s, and S-waves correspondingly range in speed from 2 to 5 km/s, both being faster in more rigid material. The P-wave gets progressively farther ahead of the S-wave as they travel through Earth's crust. The time between the P- and S-waves is routinely used to determine the distance to their source, the epicenter of the earthquake.

The speed of sound is affected by temperature in a given medium. For air at sea level, the speed of sound is given by

$$v_{\text{w}} = (331\ \text{m/s})\sqrt{\frac{T}{273\ \text{K}}}$$

,

where the temperature (denoted as T) is in units of kelvin. The speed of sound in gases is related to the average speed of particles in the gas, v_{rms}, and that

$$v_{\text{rms}} = \sqrt{\frac{3kT}{m}}$$

,

where k is the Boltzmann constant (1.38×10^{-23} J/K) and m is the mass of each (identical) particle in the gas. So, it is reasonable that the speed of sound in air and other gases should depend on the square root of temperature. While not negligible, this is not a strong dependence. At 0ºC , the speed of sound is 331 m/s, whereas at 20.0ºC it is 343 m/s, less than a 4% increase. Figure 3 shows a use of the speed of sound by a bat to sense distances. Echoes are also used in medical imaging.

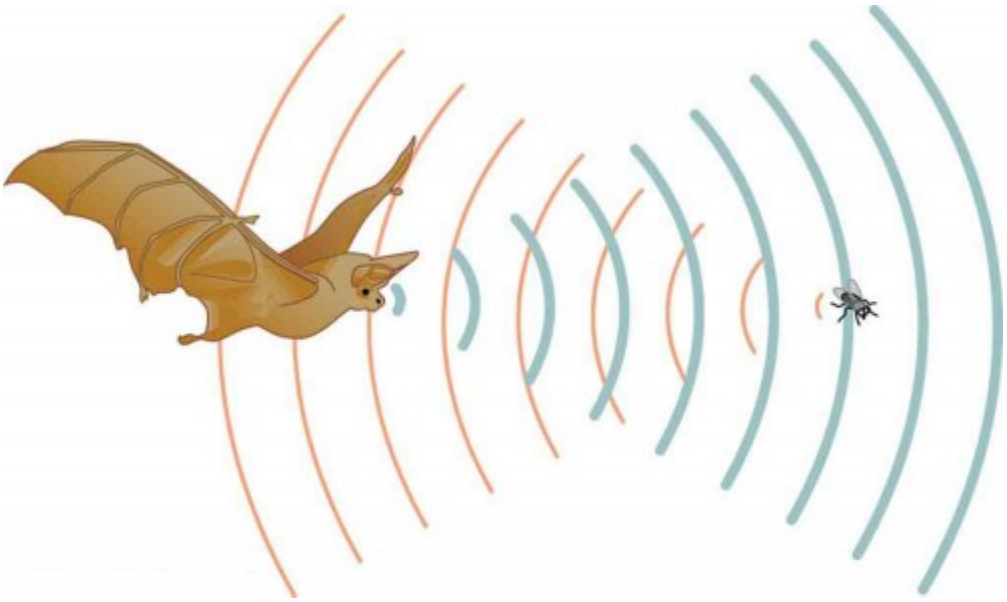


Figure 3. A bat uses sound echoes to find its way about and to catch prey. The time for the echo to return is directly proportional to the distance.

One of the more important properties of sound is that its speed is nearly independent of frequency. This independence is certainly true in open air for sounds in the audible range of 20 to 20,000 Hz. If this independence were not true, you would certainly notice it for music played by a marching band in a football stadium, for example. Suppose that high-frequency sounds traveled faster—then the farther you were from the band, the more the sound from the low-pitch instruments would lag that from the high-pitch ones. But the music from all instruments arrives in cadence independent of distance, and so all frequencies must travel at nearly the same speed. Recall that

$$v_{\text{w}} = f\lambda.$$

In a given medium under fixed conditions, v_w is constant, so that there is a relationship between f and λ; the higher the frequency, the smaller the wavelength. See Figure 4 and consider the following example.

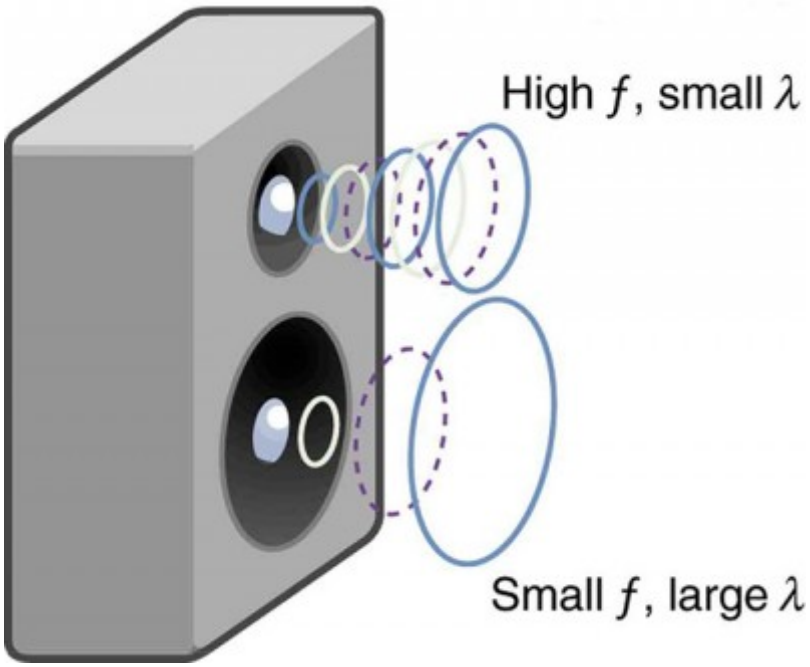


Figure 4. Because they travel at the same speed in a given medium, low-frequency sounds must have a greater wavelength than high-frequency sounds. Here, the lower-frequency sounds are emitted by the large speaker, called a woofer, while the higher-frequency sounds are emitted by the small speaker, called a tweeter.

Example 1. Calculating Wavelengths: What Are the Wavelengths of Audible Sounds?

Calculate the wavelengths of sounds at the extremes of the audible range, 20 and 20,000 Hz, in 30.0ºC air. (Assume that the frequency values are accurate to two significant figures.)

Strategy

To find wavelength from frequency, we can use $v_w = f\lambda$.

Solution

1. Identify knowns. The value for v_w, is given by

$$v_w = (331 \text{ m/s}) \sqrt{\frac{T}{273 \text{ K}}}$$

.

2. Convert the temperature into kelvin and then enter the temperature into the equation

$$v_w = (331 \text{ m/s}) \sqrt{\frac{303 \text{ K}}{273 \text{ K}}} = 348.7 \text{ m/s}$$

.

3. Solve the relationship between speed and wavelength for λ:

$$\lambda = \frac{v_{\text{w}}}{f}$$

.

4. Enter the speed and the minimum frequency to give the maximum wavelength:

$$\lambda_{\text{max}} = \frac{348.7 \text{ m/s}}{20 \text{ Hz}} = 17 \text{ m}$$

.

5. Enter the speed and the maximum frequency to give the minimum wavelength:

$$\lambda_{\text{min}} = \frac{348.7 \text{ m/s}}{20{,}000 \text{ Hz}} = 0.017 \text{ m} = 1.7 \text{ cm}$$

.

Discussion

Because the product of f multiplied by λ equals a constant, the smaller f is, the larger λ must be, and vice versa.

The speed of sound can change when sound travels from one medium to another. However, the frequency usually remains the same because it is like a driven oscillation and has the frequency of the original source. If v_{w} changes and f remains the same, then the wavelength λ must change. That is, because $v_{\text{w}} = f\lambda$, the higher the speed of a sound, the greater its wavelength for a given frequency.

Making Connections: Take-Home Investigation—Voice as a Sound Wave

Suspend a sheet of paper so that the top edge of the paper is fixed and the bottom edge is free to move. You could tape the top edge of the paper to the edge of a table. Gently blow near the edge of the bottom of the sheet and note how the sheet moves. Speak softly and then louder such that the sounds hit the edge of the bottom of the paper, and note how the sheet moves. Explain the effects.

Check Your Understanding

Part 1

Imagine you observe two fireworks explode. You hear the explosion of one as soon as you see it. However, you see the other firework for several milliseconds before you hear the explosion. Explain why this is so.

Solution

Sound and light both travel at definite speeds. The speed of sound is slower than the speed of light. The first firework is probably very close by, so the speed difference is not noticeable. The second firework is farther away, so the light arrives at your eyes noticeably sooner than the sound wave arrives at your ears.

Part 2

You observe two musical instruments that you cannot identify. One plays high-pitch sounds and the other plays low-pitch sounds. How could you determine which is which without hearing either of them play?

Solution

Compare their sizes. High-pitch instruments are generally smaller than low-pitch instruments because they generate a smaller wavelength.

Section Summary

- The relationship of the speed of sound v_w, its frequency f, and its wavelength λ is given by $v_w f\lambda$, which is the same relationship given for all waves.
- In air, the speed of sound is related to air temperature T by $v_w = (331\text{m/s})\sqrt{\frac{T}{273\text{K}}}$. v_w is the same for all frequencies and wavelengths.

Conceptual Questions

1. How do sound vibrations of atoms differ from thermal motion?
2. When sound passes from one medium to another where its propagation speed is different, does its frequency or wavelength change? Explain your answer briefly.

Problems & Exercises

1. When poked by a spear, an operatic soprano lets out a 1200-Hz shriek. What is its wavelength if the speed of sound is 345 m/s?
2. What frequency sound has a 0.10-m wavelength when the speed of sound is 340 m/s?
3. Calculate the speed of sound on a day when a 1500 Hz frequency has a wavelength of 0.221 m.
4. (a) What is the speed of sound in a medium where a 100-kHz frequency produces a 5.96-cm wavelength? (b) Which substance in Table 1 is this likely to be?
5. Show that the speed of sound in 20.0ºC air is 343 m/s, as claimed in the text.
6. Air temperature in the Sahara Desert can reach 56.0ºC (about 134ºF). What is the speed of sound in air at that temperature?
7. Dolphins make sounds in air and water. What is the ratio of the wavelength of a sound in air to its wavelength in seawater? Assume air temperature is 20.0ºC.
8. A sonar echo returns to a submarine 1.20 s after being emitted. What is the distance to the object

creating the echo? (Assume that the submarine is in the ocean, not in fresh water.)

9. (a) If a submarine's sonar can measure echo times with a precision of 0.0100 s, what is the smallest difference in distances it can detect? (Assume that the submarine is in the ocean, not in fresh water.) (b) Discuss the limits this time resolution imposes on the ability of the sonar system to detect the size and shape of the object creating the echo.
10. A physicist at a fireworks display times the lag between seeing an explosion and hearing its sound, and finds it to be 0.400 s. (a) How far away is the explosion if air temperature is 24.0°C and if you neglect the time taken for light to reach the physicist? (b) Calculate the distance to the explosion taking the speed of light into account. Note that this distance is negligibly greater.
11. Suppose a bat uses sound echoes to locate its insect prey, 3.00 m away. (See Figure 3.) (a) Calculate the echo times for temperatures of 5.00ºC and 35.0ºC. (b) What percent uncertainty does this cause for the bat in locating the insect? (c) Discuss the significance of this uncertainty and whether it could cause difficulties for the bat. (In practice, the bat continues to use sound as it closes in, eliminating most of any difficulties imposed by this and other effects, such as motion of the prey.)

Glossary

pitch: the perception of the frequency of a sound

Selected Solutions to Problems & Exercises

1. 0.288 m

3. 332 m/s

5.

$$\begin{aligned} v_{\text{w}} &= (331\ \text{m/s})\sqrt{\frac{T}{273\ \text{K}}} = (331\ \text{m/s})\sqrt{\frac{293\ \text{K}}{273\ \text{K}}} \\ &= 343\ \text{m/s} \end{aligned}$$

7. 0.223

9. (a) 7.70 m; (b) This means that sonar is good for spotting and locating large objects, but it isn't able to resolve smaller objects, or detect the detailed shapes of objects. Objects like ships or large pieces of airplanes can be found by sonar, while smaller pieces must be found by other means.

11. (a) 18.0 ms, 17.1 ms; (b) 5.00%; (c) This uncertainty could definitely cause difficulties for the bat, if it didn't continue to use sound as it closed in on its prey. A 5% uncertainty could be the difference between catching the prey around the neck or around the chest, which means that it could miss grabbing its prey.

Sound Intensity and Sound Level

Lumen Learning

Learning Objectives

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

- Define intensity, sound intensity, and sound pressure level.
- Calculate sound intensity levels in decibels (dB).

In a quiet forest, you can sometimes hear a single leaf fall to the ground. After settling into bed, you may hear your blood pulsing through your ears. But when a passing motorist has his stereo turned up, you cannot even hear what the person next to you in your car is saying. We are all very familiar with the loudness of sounds and aware that they are related to how energetically the source is vibrating. In cartoons depicting a screaming person (or an animal making a loud noise), the cartoonist often shows an open mouth with a vibrating uvula, the hanging tissue at the back of the mouth, to suggest a loud sound coming from the throat Figure 2. High noise exposure is hazardous to hearing, and it is common for musicians to have hearing losses that are sufficiently severe that they interfere with the musicians' abilities to perform. The relevant physical quantity is sound intensity, a concept that is valid for all sounds whether or not they are in the audible range.

Figure 1. Noise on crowded roadways like this one in Delhi makes it hard to hear others unless they shout. (credit: Lingaraj G J, Flickr)

Intensity is defined to be the power per unit area carried by a wave. Power is the rate at which energy is transferred by the wave. In equation form, *intensity I* is

$$I = \frac{P}{A}$$

, where P is the power through an area A. The SI unit for I is W/m^2. The intensity of a sound wave is related to its amplitude squared by the following relationship:

$$I = \frac{(\Delta p)^2}{2\rho v_{\text{w}}}$$

.

Here Δp is the pressure variation or pressure amplitude (half the difference between the maximum and

minimum pressure in the sound wave) in units of pascals (Pa) or N/m^2. (We are using a lower case *p* for pressure to distinguish it from power, denoted by *P* above.) The energy (as kinetic energy $\frac{mv^2}{2}$) of an oscillating element of air due to a traveling sound wave is proportional to its amplitude squared. In this equation, ρ is the density of the material in which the sound wave travels, in units of kg/m^3, and v_w is the speed of sound in the medium, in units of m/s. The pressure variation is proportional to the amplitude of the oscillation, and so *I* varies as $(\Delta p)^2$ (Figure 2). This relationship is consistent with the fact that the sound wave is produced by some vibration; the greater its pressure amplitude, the more the air is compressed in the sound it creates.

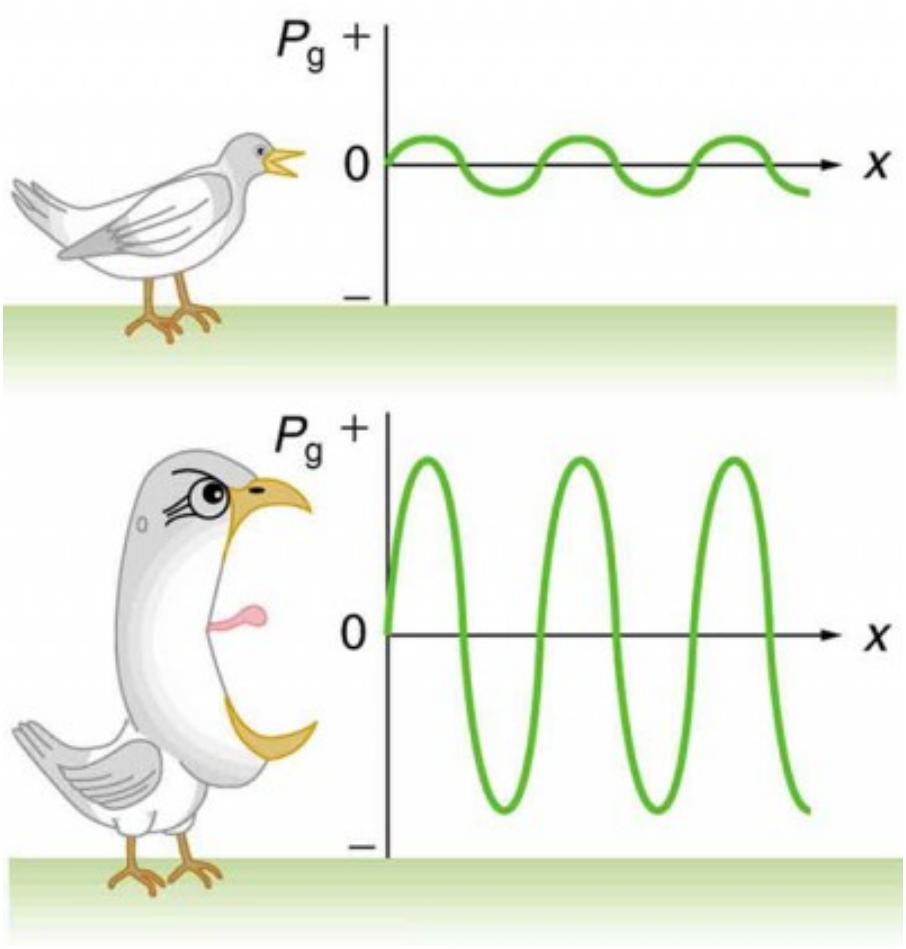


Figure 2. Graphs of the gauge pressures in two sound waves of different intensities. The more intense sound is produced by a source that has larger-amplitude oscillations and has greater pressure maxima and minima. Because pressures are higher in the greater-intensity sound, it can exert larger forces on the objects it encounters.

Sound intensity levels are quoted in decibels (dB) much more often than sound intensities in watts per meter squared. Decibels are the unit of choice in the scientific literature as well as in the popular media. The reasons for this choice of units are related to how we perceive sounds. How our ears perceive sound can be more accurately described by the logarithm of the intensity rather than directly to the intensity. The *sound intensity level* β in decibels of a sound having an intensity *I* in watts per meter squared is defined to be

$$\beta\,(\mathrm{dB}) = 10\log_{10}\left(\frac{I}{I_0}\right)$$

, where $I_0 = 10^{-12}$ W/m^2 is a reference intensity. In particular, I_0 is the lowest or threshold intensity of sound a person with normal hearing can perceive at a frequency of 1000 Hz. Sound intensity level is not the same as intensity. Because β is defined in terms of a ratio, it is a unitless quantity telling you the *level* of the sound relative to a fixed standard (10^{-12} W/m^2, in this case). The units of decibels (dB) are used to indicate this ratio is multiplied by 10 in its definition. The bel, upon which the decibel is based, is named for Alexander Graham Bell, the inventor of the telephone.

Table 1. Sound Intensity Levels and Intensities

Sound intensity level β (dB)	**Intensity I(W/m^2)**	**Example/effect**
0	1×10^{-12}	Threshold of hearing at 1000 Hz
10	1×10^{-11}	Rustle of leaves
20	1×10^{-10}	Whisper at 1 m distance
30	1×10^{-9}	Quiet home
40	1×10^{-8}	Average home
50	1×10^{-7}	Average office, soft music
60	1×10^{-6}	Normal conversation
70	1×10^{-5}	Noisy office, busy traffic
80	1×10^{-4}	Loud radio, classroom lecture
90	1×10^{-3}	Inside a heavy truck; damage from prolonged exposure[1]
100	1×10^{-2}	Noisy factory, siren at 30 m; damage from 8 h per day exposure
110	1×10^{-1}	Damage from 30 min per day exposure
120	1	Loud rock concert, pneumatic chipper at 2 m; threshold of pain
140	1×10^{2}	Jet airplane at 30 m; severe pain, damage in seconds
160	1×10^{4}	Bursting of eardrums

The decibel level of a sound having the threshold intensity of 10^{-12} W/m^2 is $\beta = 0$ dB, because $\log_{10}1 = 0$. That is, the threshold of hearing is 0 decibels. Table 1 gives levels in decibels and intensities in watts per meter squared for some familiar sounds.

One of the more striking things about the intensities in Table 1 is that the intensity in watts per meter

1. Several government agencies and health-related professional associations recommend that 85 dB not be exceeded for 8-hour daily exposures in the absence of hearing protection.

squared is quite small for most sounds. The ear is sensitive to as little as a trillionth of a watt per meter squared—even more impressive when you realize that the area of the eardrum is only about 1 cm^2, so that only 10^{-16} W falls on it at the threshold of hearing! Air molecules in a sound wave of this intensity vibrate over a distance of less than one molecular diameter, and the gauge pressures involved are less than 10^{-9} atm.

Another impressive feature of the sounds in Table 1 is their numerical range. Sound intensity varies by a factor of 10^{12} from threshold to a sound that causes damage in seconds. You are unaware of this tremendous range in sound intensity because how your ears respond can be described approximately as the logarithm of intensity. Thus, sound intensity levels in decibels fit your experience better than intensities in watts per meter squared. The decibel scale is also easier to relate to because most people are more accustomed to dealing with numbers such as 0, 53, or 120 than numbers such as 1.00×10^{-11}.

One more observation readily verified by examining Table 1 or using

$$I = \frac{(\Delta p)^2}{2\rho v_{\text{w}}}$$

is that each factor of 10 in intensity corresponds to 10 dB. For example, a 90 dB sound compared with a 60 dB sound is 30 dB greater, or three factors of 10 (that is, 10^3 times) as intense. Another example is that if one sound is 10^7 as intense as another, it is 70 dB higher. See Table 2.

Table 2. Ratios of Intensities and Corresponding Differences in Sound Intensity Levels

$\frac{I_2}{I_1}$	$\beta_2 - \beta_1$
2.0	3.0 dB
5.0	7.0 dB
10.0	10.0 dB

Example 1. Calculating Sound Intensity Levels: Sound Waves

Calculate the sound intensity level in decibels for a sound wave traveling in air at 0ºC and having a pressure amplitude of 0.656 Pa.

Strategy

We are given Δp, so we can calculate I using the equation

$$I = \frac{(\Delta p)^2}{(2\rho v_{\text{w}})^2}$$

. Using I, we can calculate β straight from its definition in

$$\beta\,(\text{dB}) = 10 \log_{10}\left(\frac{I}{I_0}\right)$$

.

Solution

1. Identify knowns: Sound travels at 331 m/s in air at 0ºC. Air has a density of 1.29 kg/m^3 at atmospheric pressure and 0ºC.

2. Enter these values and the pressure amplitude into

$$I = \frac{(\Delta p)^2}{2\rho v_{\mathrm{w}}}$$

:

$$I = \frac{(\Delta p)^2}{2\rho v_{\mathrm{w}}} = \frac{(0.656\ \mathrm{Pa})^2}{2\left(1.29\ \mathrm{kg/m^3}\right)(331\ \mathrm{m/s})} = 5.04 \times 10^{-4}\ \mathrm{W/m^2}$$

3. Enter the value for I and the known value for I_0 into

$$\beta\,(\mathrm{dB}) = 10 \log_{10}\left(\frac{I}{I_0}\right)$$

. Calculate to find the sound intensity level in decibels:

$10 \log_{10}(5.04 \times 10^8) = 10(8.70)\mathrm{dB} = 87\ \mathrm{dB}$.

Discussion

This 87 dB sound has an intensity five times as great as an 80 dB sound. So a factor of five in intensity corresponds to a difference of 7 dB in sound intensity level. This value is true for any intensities differing by a factor of five.

Example 2. Change Intensity Levels of a Sound: What Happens to the Decibel Level?

Show that if one sound is twice as intense as another, it has a sound level about 3 dB higher.

Strategy

You are given that the ratio of two intensities is 2 to 1, and are then asked to find the difference in their sound levels in decibels. You can solve this problem using of the properties of logarithms.

Solution

1. Identify knowns.

The ratio of the two intensities is 2 to 1, or:

$$\frac{I_2}{I_1} = 2.00$$

.

We wish to show that the difference in sound levels is about 3 dB. That is, we want to show

$\beta_2 - \beta_1 = 3$ dB.

Note that

$$\log_{10} b - \log_{10} a = \log_{10} \left(\frac{b}{a}\right)$$

.

2. Use the definition of β to get:

$$\beta_2 - \beta_1 = 10 \log_{10} \left(\frac{I_2}{I_1}\right) = 10 \log_{10} 2.00 = 10\,(0.301)\ \text{dB}$$

Thus,

$$\beta_2 - \beta_1 = 3.01\ \text{dB}.$$

Discussion

This means that the two sound intensity levels differ by 3.01 dB, or about 3 dB, as advertised. Note that because only the ratio

$$\frac{I_2}{I_1}$$

is given (and not the actual intensities), this result is true for any intensities that differ by a factor of two. For example, a 56.0 dB sound is twice as intense as a 53.0 dB sound, a 97.0 dB sound is half as intense as a 100 dB sound, and so on.

It should be noted at this point that there is another decibel scale in use, called the *sound pressure level*, based on the ratio of the pressure amplitude to a reference pressure. This scale is used particularly in applications where sound travels in water. It is beyond the scope of most introductory texts to treat this scale because it is not commonly used for sounds in air, but it is important to note that very different decibel levels may be encountered when sound pressure levels are quoted. For example, ocean noise pollution produced by ships may be as great as 200 dB expressed in the sound pressure level, where the more familiar sound intensity level we use here would be something under 140 dB for the same sound.

Take-Home Investigation: Feeling Sound

Find a CD player and a CD that has rock music. Place the player on a light table, insert the CD into the player, and start playing the CD. Place your hand gently on the table next to the speakers. Increase the volume and note the level when the table just begins to vibrate as the rock music plays. Increase the reading on the volume control until it doubles. What has happened to the vibrations?

Check Your Understanding

Part 1

Describe how amplitude is related to the loudness of a sound.

Solution

Amplitude is directly proportional to the experience of loudness. As amplitude increases, loudness increases.

Part 2

Identify common sounds at the levels of 10 dB, 50 dB, and 100 dB.

Solution

10 dB: Running fingers through your hair.

50 dB: Inside a quiet home with no television or radio.

100 dB: Take-off of a jet plane.

Section Summary

- Intensity is the same for a sound wave as was defined for all waves; it is $I = \frac{P}{A}$, where P is the power crossing area A. The SI unit for I is watts per meter squared. The intensity of a sound wave is also related to the pressure amplitude Δp, $I = \frac{(\Delta p)^2}{2\rho v_w}$, where ρ is the density of the medium in which the sound wave travels and v_w is the speed of sound in the medium.
- Sound intensity level in units of decibels (dB) is $\beta\,(\text{dB}) = 10 \log_{10}\left(\frac{I}{I_0}\right)$, where $I_0 = 10^{-12}$ W/m^2 is the threshold intensity of hearing.

Conceptual Questions

1. Six members of a synchronized swim team wear earplugs to protect themselves against water pressure at depths, but they can still hear the music and perform the combinations in the water perfectly. One day, they were asked to leave the pool so the dive team could practice a few dives, and they tried to practice on a mat, but seemed to have a lot more difficulty. Why might this be?
2. A community is concerned about a plan to bring train service to their downtown from the town's outskirts. The current sound intensity level, even though the rail yard is blocks away, is 70 dB downtown. The mayor assures the public that there will be a difference of only 30 dB in sound in the downtown area. Should the townspeople be concerned? Why?

Problems & Exercises

1. What is the intensity in watts per meter squared of 85.0-dB sound?
2. The warning tag on a lawn mower states that it produces noise at a level of 91.0 dB. What is this in watts per meter squared?
3. A sound wave traveling in 20ºC air has a pressure amplitude of 0.5 Pa. What is the intensity of the wave?
4. What intensity level does the sound in the preceding problem correspond to?
5. What sound intensity level in dB is produced by earphones that create an intensity of 4.00×10^{-2} W/m^2?
6. Show that an intensity of 10^{-12} W/m^2 is the same as 10^{-16} W/m^2.
7. (a) What is the decibel level of a sound that is twice as intense as a 90.0-dB sound? (b) What is the decibel level of a sound that is one-fifth as intense as a 90.0-dB sound?
8. (a) What is the intensity of a sound that has a level 7.00 dB lower than a 4.00×10^{-9} W/m^2 sound? (b) What is the intensity of a sound that is 3.00 dB higher than a 4.00×10^{-9} W/m^2 sound?
9. (a) How much more intense is a sound that has a level 17.0 dB higher than another? (b) If one sound has a level 23.0 dB less than another, what is the ratio of their intensities?
10. People with good hearing can perceive sounds as low in level as −8.00 dB at a frequency of 3000 Hz. What is the intensity of this sound in watts per meter squared?
11. If a large housefly 3.0 m away from you makes a noise of 40.0 dB, what is the noise level of 1000 flies at that distance, assuming interference has a negligible effect?
12. Ten cars in a circle at a boom box competition produce a 120-dB sound intensity level at the center of the circle. What is the average sound intensity level produced there by each stereo, assuming interference effects can be neglected?
13. The amplitude of a sound wave is measured in terms of its maximum gauge pressure. By what factor does the amplitude of a sound wave increase if the sound intensity level goes up by 40.0 dB?
14. If a sound intensity level of 0 dB at 1000 Hz corresponds to a maximum gauge pressure (sound amplitude) of 10^{-9} atm, what is the maximum gauge pressure in a 60-dB sound? What is the maximum gauge pressure in a 120-dB sound?
15. An 8-hour exposure to a sound intensity level of 90.0 dB may cause hearing damage. What energy in joules falls on a 0.800-cm-diameter eardrum so exposed?
16. (a) Ear trumpets were never very common, but they did aid people with hearing losses by gathering sound over a large area and concentrating it on the smaller area of the eardrum. What decibel increase does an ear trumpet produce if its sound gathering area is 900 cm^2 and the area of the eardrum is 0.500 cm^2, but the trumpet only has an efficiency of 5.00% in transmitting the sound to the eardrum? (b) Comment on the usefulness of the decibel increase found in part (a).
17. Sound is more effectively transmitted into a stethoscope by direct contact than through the air, and it is further intensified by being concentrated on the smaller area of the eardrum. It is reasonable to assume that sound is transmitted into a stethoscope 100 times as effectively compared with transmission though the air. What, then, is the gain in decibels produced by a

stethoscope that has a sound gathering area of 15.0 cm^2, and concentrates the sound onto two eardrums with a total area of 0.900 cm^2 with an efficiency of 40.0%?

18. Loudspeakers can produce intense sounds with surprisingly small energy input in spite of their low efficiencies. Calculate the power input needed to produce a 90.0-dB sound intensity level for a 12.0-cm-diameter speaker that has an efficiency of 1.00%. (This value is the sound intensity level right at the speaker.)

Glossary

intensity: the power per unit area carried by a wave

sound intensity level: a unitless quantity telling you the level of the sound relative to a fixed standard

sound pressure level: the ratio of the pressure amplitude to a reference pressure

Selected Solutions to Problems & Exercises

1. 3.16×10^{-4} W/m^2
3. 3.04×10^{-4} W/m^2
5. 106 dB
7. (a) 93 dB; (b) 83 dB
9. (a) 50.1; (b) 5.01×10^{-3} or $\frac{1}{200}$
11. 70.0 dB
13. 100
15. 1.45×10^{-3} J
17. 28.2 dB

Doppler Effect and Sonic Booms

Lumen Learning

Learning Objectives

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

- Define Doppler effect, Doppler shift, and sonic boom.
- Calculate the frequency of a sound heard by someone observing Doppler shift.
- Describe the sounds produced by objects moving faster than the speed of sound.

The characteristic sound of a motorcycle buzzing by is an example of the *Doppler effect*. The high-pitch scream shifts dramatically to a lower-pitch roar as the motorcycle passes by a stationary observer. The closer the motorcycle brushes by, the more abrupt the shift. The faster the motorcycle moves, the greater the shift. We also hear this characteristic shift in frequency for passing race cars, airplanes, and trains. It is so familiar that it is used to imply motion and children often mimic it in play.

The Doppler effect is an alteration in the observed frequency of a sound due to motion of either the source or the observer. Although less familiar, this effect is easily noticed for a stationary source and moving observer. For example, if you ride a train past a stationary warning bell, you will hear the bell's frequency shift from high to low as you pass by. The actual change in frequency due to relative motion of source and observer is called a *Doppler shift*. The Doppler effect and Doppler shift are named for the Austrian physicist and mathematician Christian Johann Doppler (1803–1853), who did experiments with both moving sources and moving observers. Doppler, for example, had musicians play on a moving open train car and also play standing next to the train tracks as a train passed by. Their music was observed both on and off the train, and changes in frequency were measured.

What causes the Doppler shift? Figure 1, Figure 2, and Figure 3 compare sound waves emitted by stationary and moving sources in a stationary air mass. Each disturbance spreads out spherically from the point where the sound was emitted. If the source is stationary, then all of the spheres representing the air compressions in the sound wave centered on the same point, and the stationary observers on either side see the same wavelength and frequency as emitted by the source, as in Figure 1. If the source is moving, as in Figure 2, then the situation is different. Each compression of the air moves out in a sphere from the point where it was emitted, but the point of emission moves. This moving emission point causes the air compressions to be closer together on one side and farther apart on the other. Thus, the wavelength is shorter in the direction the source is moving (on the right in Figure 2), and longer in the opposite direction (on the left in Figure 2). Finally, if the observers move, as in Figure 3, the frequency at which they receive the compressions changes. The observer moving toward the source receives them at a higher frequency, and the person moving away from the source receives them at a lower frequency.

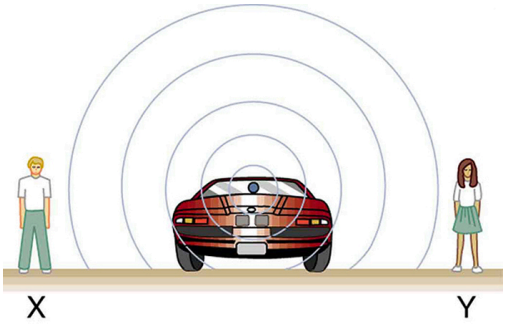


Figure 1. Sounds emitted by a source spread out in spherical waves. Because the source, observers, and air are stationary, the wavelength and frequency are the same in all directions and to all observers.

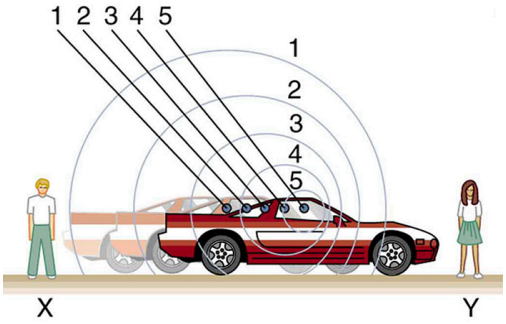


Figure 2. Sounds emitted by a source moving to the right spread out from the points at which they were emitted. The wavelength is reduced and, consequently, the frequency is increased in the direction of motion, so that the observer on the right hears a higher-pitch sound. The opposite is true for the observer on the left, where the wavelength is increased and the frequency is reduced.

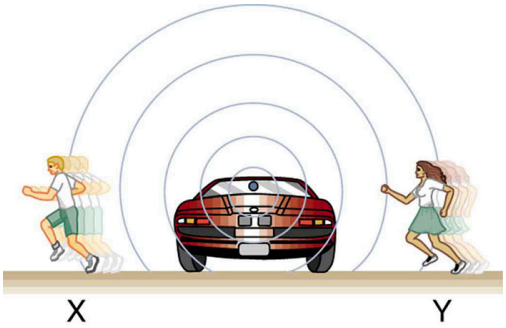


Figure 3. The same effect is produced when the observers move relative to the source. Motion toward the source increases frequency as the observer on the right passes through more wave crests than she would if stationary. Motion away from the source decreases frequency as the observer on the left passes through fewer wave crests than he would if stationary.

We know that wavelength and frequency are related by $v_w = f\lambda$, where v_w is the fixed speed of sound. The sound moves in a medium and has the same speed v_w in that medium whether the source is moving or not. Thus f multiplied by λ is a constant. Because the observer on the right in Figure 2 receives a shorter wavelength, the frequency she receives must be higher. Similarly, the observer on the left receives a longer wavelength, and hence he hears a lower frequency. The same thing happens in Figure 3. A higher frequency is received by the observer moving toward the source, and a lower frequency is received by an observer moving away from the source. In general, then, relative motion of source and observer toward one another increases the received frequency. Relative motion apart decreases frequency. The greater the relative speed is, the greater the effect.

The Doppler Effect

The Doppler effect occurs not only for sound but for any wave when there is relative motion between the observer and the source. There are Doppler shifts in the frequency of sound, light, and water waves, for example. Doppler shifts can be used to determine velocity, such as when ultrasound is reflected from blood in a medical diagnostic. The recession of galaxies is determined by the shift in the frequencies of light received from them and has implied much about the origins of the universe. Modern physics has been profoundly affected by observations of Doppler shifts.

For a stationary observer and a moving source, the frequency f_{obs} received by the observer can be shown to be

$$f_{obs} = f_s \left(\frac{v_w}{v_w \pm v_s} \right)$$

,

where f_s is the frequency of the source, v_s is the speed of the source along a line joining the source and observer, and v_w is the speed of sound. The minus sign is used for motion toward the observer and the plus sign for motion away from the observer, producing the appropriate shifts up and down in frequency. Note that the greater the speed of the source, the greater the effect. Similarly, for a stationary source and moving observer, the frequency received by the observer f_{obs} is given by

$$f_{obs} = f_s \left(\frac{v_w \pm v_{obs}}{v_w} \right)$$

,

where v_{obs} is the speed of the observer along a line joining the source and observer. Here the plus sign is for motion toward the source, and the minus is for motion away from the source.

Example 1. Calculate Doppler Shift: A Train Horn

Suppose a train that has a 150-Hz horn is moving at 35.0 m/s in still air on a day when the speed of sound is 340 m/s.

1. What frequencies are observed by a stationary person at the side of the tracks as the train approaches and after it passes?
2. What frequency is observed by the train's engineer traveling on the train?

Strategy

To find the observed frequency in Part 1,

$$f_{obs} = f_s \left(\frac{v_w}{v_w \pm v_s} \right)$$

, must be used because the source is moving. The minus sign is used for the approaching train, and the plus sign for the receding train. In Part 2, there are two Doppler shifts—one for a moving source and the other for a moving observer.

Solution for Part 1

Enter known values into

$$f_{obs} = f_s \left(\frac{v_w}{v_w - v_s} \right)$$

:

$$f_{obs} = f_s \left(\frac{v_w}{v_w - v_s} \right) = (150 \text{ Hz}) \left(\frac{340 \text{ m/s}}{340 \text{ m/s} - 35.0 \text{ m/s}} \right)$$

Calculate the frequency observed by a stationary person as the train approaches: f_{obs} = (150 Hz) (1.11) = 167 Hz

Use the same equation with the plus sign to find the frequency heard by a stationary person as the train recedes.

$$f_{\text{obs}} = f_{\text{s}}\left(\frac{v_{\text{w}}}{v_{\text{w}} + v_{\text{s}}}\right) = (150\ \text{Hz})\left(\frac{340\ \text{m/s}}{340\ \text{m/s} + 35.0\ \text{m/s}}\right)$$

Calculate the second frequency: f_{obs} = (150 Hz) (0.907) = 136 Hz

Discussion on Part 1

The numbers calculated are valid when the train is far enough away that the motion is nearly along the line joining train and observer. In both cases, the shift is significant and easily noticed. Note that the shift is 17.0 Hz for motion toward and 14.0 Hz for motion away. The shifts are not symmetric.

Solution for Part 2

Identify knowns:

- It seems reasonable that the engineer would receive the same frequency as emitted by the horn, because the relative velocity between them is zero.
- Relative to the medium (air), the speeds are $v_{\text{s}} = v_{\text{obs}}$ = 35.0 m/s.
- The first Doppler shift is for the moving observer; the second is for the moving source.

Use the following equation:

$$f_{\text{obs}} = \left[f_{\text{s}}\left(\frac{v_{\text{w}} \pm v_{\text{obs}}}{v_{\text{w}}}\right)\right]\left(\frac{v_{\text{w}}}{v_{\text{w}} \pm v_{\text{s}}}\right)$$

The quantity in the square brackets is the Doppler-shifted frequency due to a moving observer. The factor on the right is the effect of the moving source.

Because the train engineer is moving in the direction toward the horn, we must use the plus sign for v_{obs}; however, because the horn is also moving in the direction away from the engineer, we also use the plus sign for v_{s}. But the train is carrying both the engineer and the horn at the same velocity, so $v_{\text{s}} = v_{\text{obs}}$. As a result, everything but f_{s} cancels, yielding $f_{\text{obs}} = f_{\text{s}}$.

Discussion for Part 2

We may expect that there is no change in frequency when source and observer move together because it fits your experience. For example, there is no Doppler shift in the frequency of conversations between driver and passenger on a motorcycle. People talking when a wind moves the air between them also observe no Doppler shift in their conversation. The crucial point is that source and observer are not moving relative to each other.

Sonic Booms to Bow Wakes

What happens to the sound produced by a moving source, such as a jet airplane, that approaches or even exceeds the speed of sound? The answer to this question applies not only to sound but to all other waves as well.

Suppose a jet airplane is coming nearly straight at you, emitting a sound of frequency f_{s}. The greater

the plane's speed v_s, the greater the Doppler shift and the greater the value observed for f_{obs}. Now, as v_s approaches the speed of sound, f_{obs} approaches infinity, because the denominator in

$$f_{obs} = f_s \left(\frac{v_w}{v_w \pm v_s} \right)$$

approaches zero. At the speed of sound, this result means that in front of the source, each successive wave is superimposed on the previous one because the source moves forward at the speed of sound. The observer gets them all at the same instant, and so the frequency is infinite. (Before airplanes exceeded the speed of sound, some people argued it would be impossible because such constructive superposition would produce pressures great enough to destroy the airplane.) If the source exceeds the speed of sound, no sound is received by the observer until the source has passed, so that the sounds from the approaching source are mixed with those from it when receding. This mixing appears messy, but something interesting happens—a sonic boom is created. (See Figure 4.)

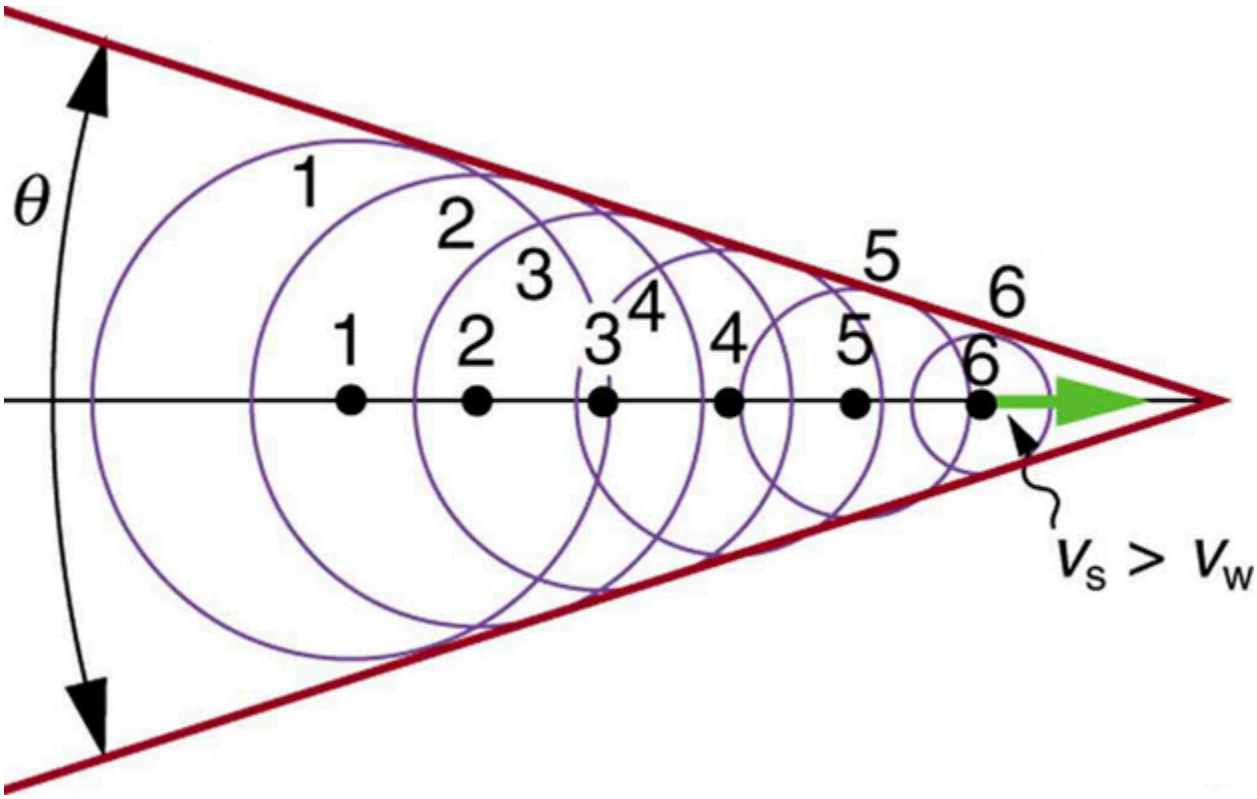


Figure 4. Sound waves from a source that moves faster than the speed of sound spread spherically from the point where they are emitted, but the source moves ahead of each. Constructive interference along the lines shown (actually a cone in three dimensions) creates a shock wave called a sonic boom. The faster the speed of the source, the smaller the angle θ.

There is constructive interference along the lines shown (a cone in three dimensions) from similar sound waves arriving there simultaneously. This superposition forms a disturbance called a *sonic boom*, a constructive interference of sound created by an object moving faster than sound. Inside the cone, the interference is mostly destructive, and so the sound intensity there is much less than on the shock wave. An aircraft creates two sonic booms, one from its nose and one from its tail. (See Figure 5.) During television coverage of space shuttle landings, two distinct booms could often be heard. These were separated by exactly the time it would take the shuttle to pass by a point. Observers on the ground often do not see the aircraft creating the sonic boom, because it has passed by before the shock wave reaches them, as seen in Figure 5. If the aircraft flies close by at low altitude, pressures in the sonic boom can be destructive and break windows as well as rattle nerves. Because of how destructive sonic booms can be, supersonic flights are banned over populated areas of the United States.

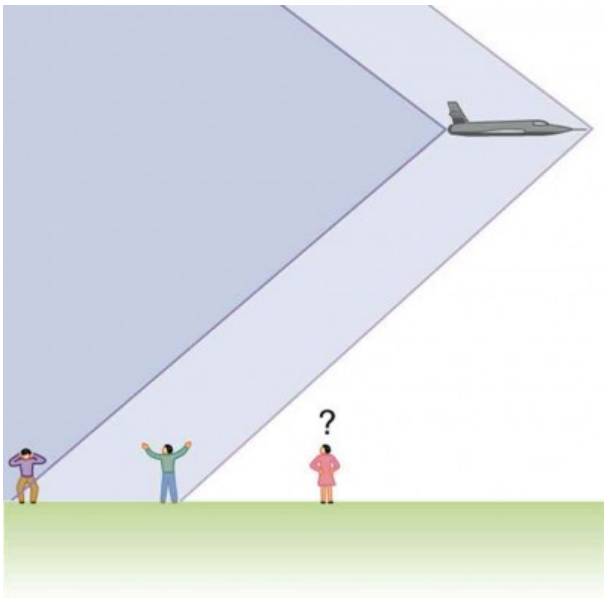


Figure 5. Two sonic booms, created by the nose and tail of an aircraft, are observed on the ground after the plane has passed by.

Sonic booms are one example of a broader phenomenon called bow wakes. A *bow wake*, such as the one in Figure 6, is created when the wave source moves faster than the wave propagation speed. Water waves spread out in circles from the point where created, and the bow wake is the familiar V-shaped wake trailing the source. A more exotic bow wake is created when a subatomic particle travels through a medium faster than the speed of light travels in that medium. (In a vacuum, the maximum speed of light will be $c = 3.00 \times 10^8$ m/s; in the medium of water, the speed of light is closer to $0.75c$. If the particle creates light in its passage, that light spreads on a cone with an angle indicative of the speed of the particle, as illustrated in Figure 7. Such a bow wake is called Cerenkov radiation and is commonly observed in particle physics.

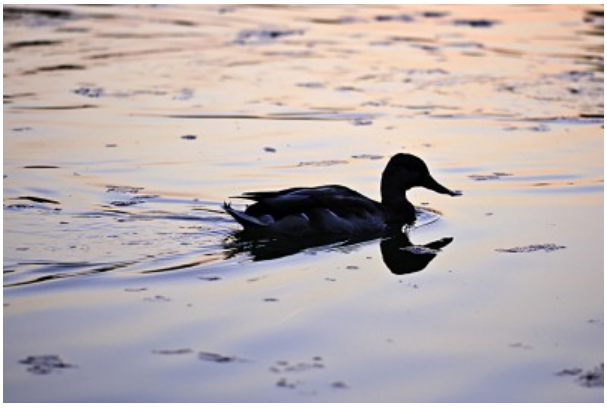


Figure 6. Bow wake created by a duck. Constructive interference produces the rather structured wake, while there is relatively little wave action inside the wake, where interference is mostly destructive. (credit: Horia Varlan, Flickr)

Doppler shifts and sonic booms are interesting sound phenomena that occur in all types of waves. They can be of considerable use. For example, the Doppler shift in ultrasound can be used to measure blood velocity, while police use the Doppler shift in radar (a microwave) to measure car velocities. In meteorology, the Doppler shift is used to track the motion of storm clouds; such "Doppler Radar" can give velocity and direction and rain or snow potential of imposing weather fronts. In astronomy, we can examine the light emitted from distant galaxies and determine their speed relative to ours. As galaxies move away from us, their light is shifted to a lower frequency, and so to a longer wavelength—the so-called red shift. Such information from galaxies far, far away has allowed us to estimate the age of the universe (from the Big Bang) as about 14 billion years.

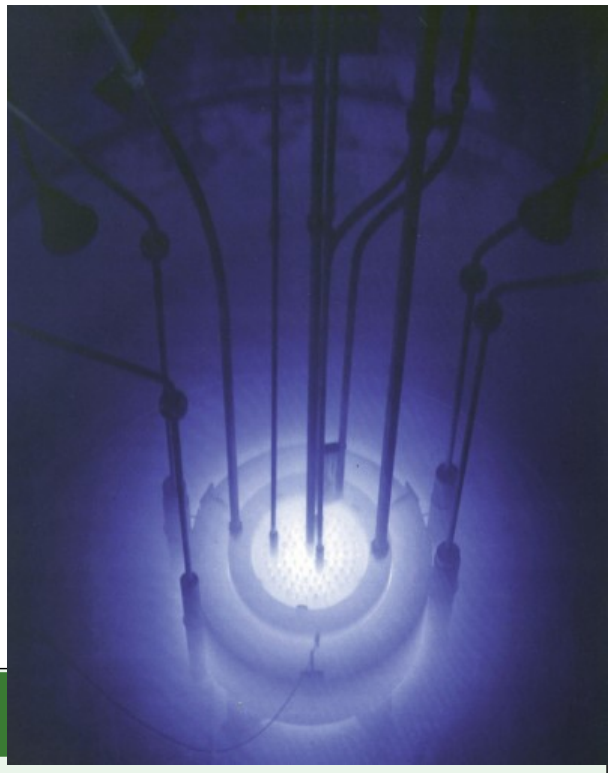


Figure 7. The blue glow in this research reactor pool is Cerenkov radiation caused by subatomic particles traveling faster than the speed of light in water. (credit: U.S. Nuclear Regulatory Commission)

Check Your Understanding

Part 1

Why did scientist Christian Doppler observe musicians both on a moving train and also from a stationary point not on the train?

Solution

Doppler needed to compare the perception of sound when the observer is stationary and the sound source moves, as well as when the sound source and the observer are both in motion.

Part 2

Describe a situation in your life when you might rely on the Doppler shift to help you either while driving a car or walking near traffic.

Solution

If I am driving and I hear Doppler shift in an ambulance siren, I would be able to tell when it was getting closer and also if it has passed by. This would help me to know whether I needed to pull over and let the ambulance through.

Section Summary

- The Doppler effect is an alteration in the observed frequency of a sound due to motion of either the source or the observer.
- The actual change in frequency is called the Doppler shift.

- A sonic boom is constructive interference of sound created by an object moving faster than sound.
- A sonic boom is a type of bow wake created when any wave source moves faster than the wave propagation speed.
- For a stationary observer and a moving source, the observed frequency f_{obs} is: $f_{\text{obs}} = f_s \left(\frac{v_w}{v_w \pm v_s} \right)$, where f_s is the frequency of the source, v_s is the speed of the source, and v_w is the speed of sound. The minus sign is used for motion toward the observer and the plus sign for motion away.
- For a stationary source and moving observer, the observed frequency is: $f_{\text{obs}} = f_s \left(\frac{v_w \pm v_{\text{obs}}}{v_w} \right)$, where v_{obs} is the speed of the observer.

Conceptual Questions

1. Is the Doppler shift real or just a sensory illusion?
2. Due to efficiency considerations related to its bow wake, the supersonic transport aircraft must maintain a cruising speed that is a constant ratio to the speed of sound (a constant Mach number). If the aircraft flies from warm air into colder air, should it increase or decrease its speed? Explain your answer.
3. When you hear a sonic boom, you often cannot see the plane that made it. Why is that?

Problems & Exercises

1. (a) What frequency is received by a person watching an oncoming ambulance moving at 110 km/h and emitting a steady 800-Hz sound from its siren? The speed of sound on this day is 345 m/s. (b) What frequency does she receive after the ambulance has passed?
2. (a) At an air show a jet flies directly toward the stands at a speed of 1200 km/h, emitting a frequency of 3500 Hz, on a day when the speed of sound is 342 m/s. What frequency is received by the observers? (b) What frequency do they receive as the plane flies directly away from them?
3. What frequency is received by a mouse just before being dispatched by a hawk flying at it at 25.0 m/s and emitting a screech of frequency 3500 Hz? Take the speed of sound to be 331 m/s.
4. A spectator at a parade receives an 888-Hz tone from an oncoming trumpeter who is playing an 880-Hz note. At what speed is the musician approaching if the speed of sound is 338 m/s?
5. A commuter train blows its 200-Hz horn as it approaches a crossing. The speed of sound is 335 m/s. (a) An observer waiting at the crossing receives a frequency of 208 Hz. What is the speed of the train? (b) What frequency does the observer receive as the train moves away?
6. Can you perceive the shift in frequency produced when you pull a tuning fork toward you at 10.0 m/s on a day when the speed of sound is 344 m/s? To answer this question, calculate the factor by

which the frequency shifts and see if it is greater than 0.300%.

7. Two eagles fly directly toward one another, the first at 15.0 m/s and the second at 20.0 m/s. Both screech, the first one emitting a frequency of 3200 Hz and the second one emitting a frequency of 3800 Hz. What frequencies do they receive if the speed of sound is 330 m/s?
8. What is the minimum speed at which a source must travel toward you for you to be able to hear that its frequency is Doppler shifted? That is, what speed produces a shift of 0.300% on a day when the speed of sound is 331 m/s?

Glossary

Doppler effect: an alteration in the observed frequency of a sound due to motion of either the source or the observer

Doppler shift: the actual change in frequency due to relative motion of source and observer

sonic boom: a constructive interference of sound created by an object moving faster than sound

bow wake: V-shaped disturbance created when the wave source moves faster than the wave propagation speed

Selected Solutions to Problems & Exercises

1. (a) 878 Hz; (b) 735 Hz
3. 3.79×10^3 Hz
5. (a) 12.9 m/s; (b) 193 Hz
7. First eagle hears 4.23×10^3 Hz; Second eagle hears 3.56×10^3 Hz

Sound Interference and Resonance: Standing Waves in Air Columns

Lumen Learning

Learning Objectives

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

- Define antinode, node, fundamental, overtones, and harmonics.
- Identify instances of sound interference in everyday situations.
- Describe how sound interference occurring inside open and closed tubes changes the characteristics of the sound, and how this applies to sounds produced by musical instruments.
- Calculate the length of a tube using sound wave measurements.

Interference is the hallmark of waves, all of which exhibit constructive and destructive interference exactly analogous to that seen for water waves. In fact, one way to prove something “is a wave” is to observe interference effects. So, sound being a wave, we expect it to exhibit interference; we have already mentioned a few such effects, such as the beats from two similar notes played simultaneously.

Figure 2 shows a clever use of sound interference to cancel noise. Larger-scale applications of active noise reduction by destructive interference are contemplated for entire passenger compartments in commercial aircraft. To obtain destructive interference, a fast electronic analysis is performed, and a second sound is introduced with its maxima and minima exactly reversed from the incoming noise. Sound waves in fluids are pressure waves and consistent with Pascal’s principle; pressures from two different sources add and subtract like simple numbers; that is, positive and negative gauge pressures add to a much smaller pressure, producing a lower-intensity sound. Although completely destructive interference is possible only under the simplest conditions, it is possible to reduce noise levels by 30 dB or more using this technique.

Figure 1. Some types of headphones use the phenomena of constructive and destructive interference to cancel out outside noises. (credit: JVC America, Flickr)

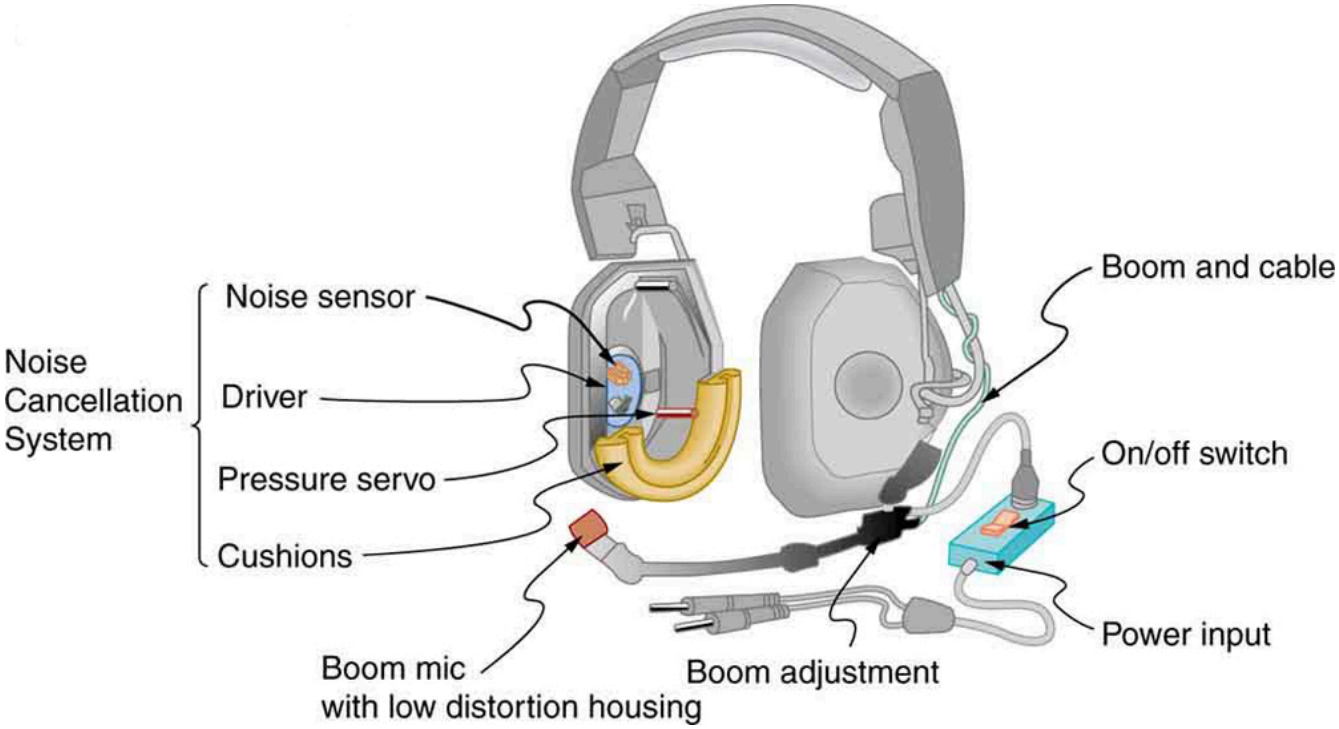


Figure 2. Headphones designed to cancel noise with destructive interference create a sound wave exactly opposite to the incoming sound. These headphones can be more effective than the simple passive attenuation used in most ear protection. Such headphones were used on the record-setting, around the world nonstop flight of the Voyager aircraft to protect the pilots' hearing from engine noise.

Where else can we observe sound interference? All sound resonances, such as in musical instruments, are due to constructive and destructive interference. Only the resonant frequencies interfere constructively to form standing waves, while others interfere destructively and are absent. From the toot made by blowing over a bottle, to the characteristic flavor of a violin's sounding box, to the recognizability of a great singer's voice, resonance and standing waves play a vital role.

Interference

Interference is such a fundamental aspect of waves that observing interference is proof that something is a wave. The wave nature of light was established by experiments showing interference. Similarly, when electrons scattered from crystals exhibited interference, their wave nature was confirmed to be exactly as predicted by symmetry with certain wave characteristics of light.

Suppose we hold a tuning fork near the end of a tube that is closed at the other end, as shown in Figure 3, Figure 4, Figure 5, and Figure 6. If the tuning fork has just the right frequency, the air column in the tube resonates loudly, but at most frequencies it vibrates very little. This observation just means that the air column has only certain natural frequencies. The figures show how a resonance at the lowest of these natural frequencies is formed. A disturbance travels down the tube at the speed of sound and bounces off the closed end. If the tube is just the right length, the reflected sound arrives back at the tuning fork

exactly half a cycle later, and it interferes constructively with the continuing sound produced by the tuning fork. The incoming and reflected sounds form a standing wave in the tube as shown.

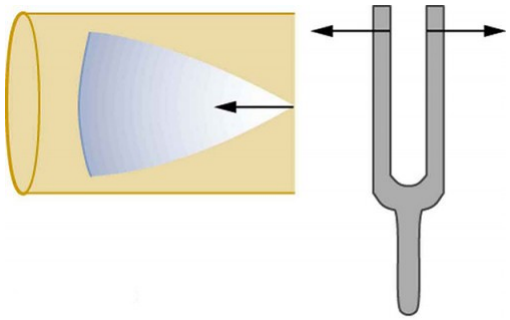


Figure 3. Resonance of air in a tube closed at one end, caused by a tuning fork. A disturbance moves down the tube.

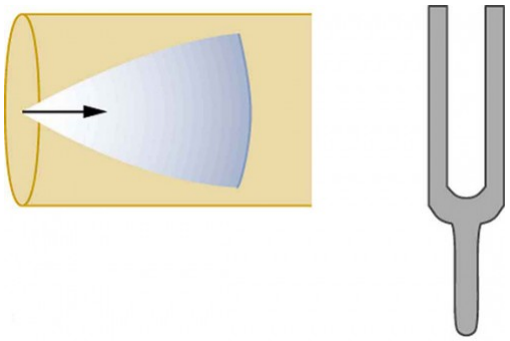


Figure 4. Resonance of air in a tube closed at one end, caused by a tuning fork. The disturbance reflects from the closed end of the tube.

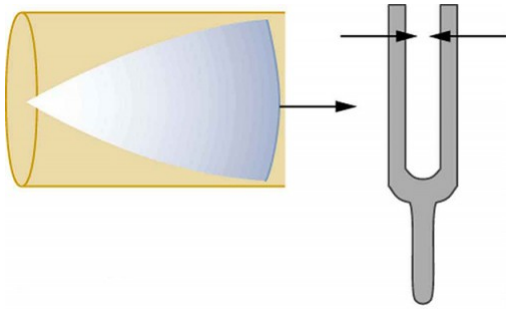


Figure 5. Resonance of air in a tube closed at one end, caused by a tuning fork. If the length of the tube L is just right, the disturbance gets back to the tuning fork half a cycle later and interferes constructively with the continuing sound from the tuning fork. This interference forms a standing wave, and the air column resonates.

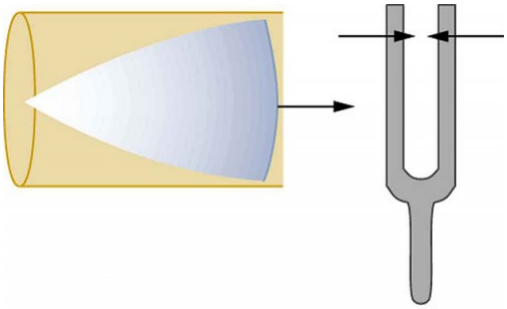


Figure 6. Resonance of air in a tube closed at one end, caused by a tuning fork. A graph of air displacement along the length of the tube shows none at the closed end, where the motion is constrained, and a maximum at the open end. This standing wave has one-fourth of its wavelength in the tube, so that $\lambda = 4L$.

The standing wave formed in the tube has its maximum air displacement (an *antinode*) at the open end, where motion is unconstrained, and no displacement (a *node*) at the closed end, where air movement is halted. The distance from a node to an antinode is one-fourth of a wavelength, and this equals the length of the tube; thus, $\lambda = 4L$. This same resonance can be produced by a vibration introduced at or near the closed end of the tube, as shown in Figure 7. It is best to consider this a natural vibration of the air column independently of how it is induced.

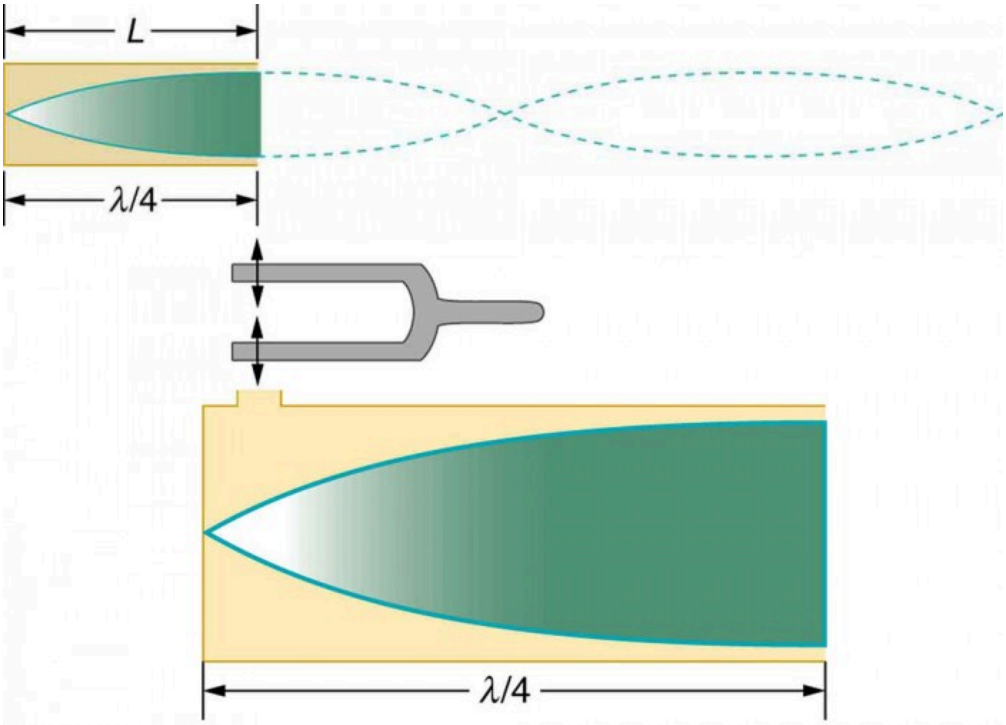


Figure 7. The same standing wave is created in the tube by a vibration introduced near its closed end.

Given that maximum air displacements are possible at the open end and none at the closed end, there are other, shorter wavelengths that can resonate in the tube, such as the one shown in Figure 8. Here the standing wave has three-fourths of its wavelength in the tube, or

$$L = \tfrac{3}{4}\lambda'$$

, so that

$$\lambda' = \tfrac{4L}{3}$$

. Continuing this process reveals a whole series of shorter-wavelength and higher-frequency sounds that resonate in the tube. We use specific terms for the resonances in any system. The lowest resonant frequency is called the *fundamental*, while all higher resonant frequencies are called *overtones*. All resonant frequencies are integral multiples of the fundamental, and they are collectively called *harmonics*. The fundamental is the first harmonic, the first overtone is the second harmonic, and so on. Figure 9 shows the fundamental and the first three overtones (the first four harmonics) in a tube closed at one end.

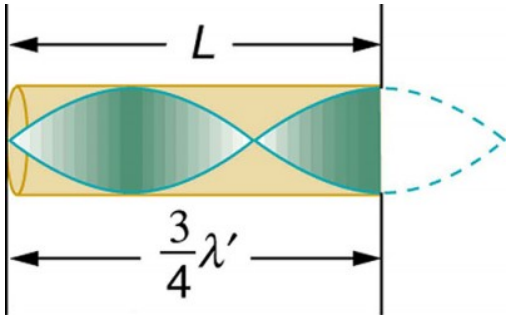


Figure 8. Another resonance for a tube closed at one end. This has maximum air displacements at the open end, and none at the closed end. The wavelength is shorter, with three-fourths λ′ equaling the length of

$$\lambda' = \tfrac{4L}{3}$$

the tube, so that . This higher-frequency vibration is the first overtone.

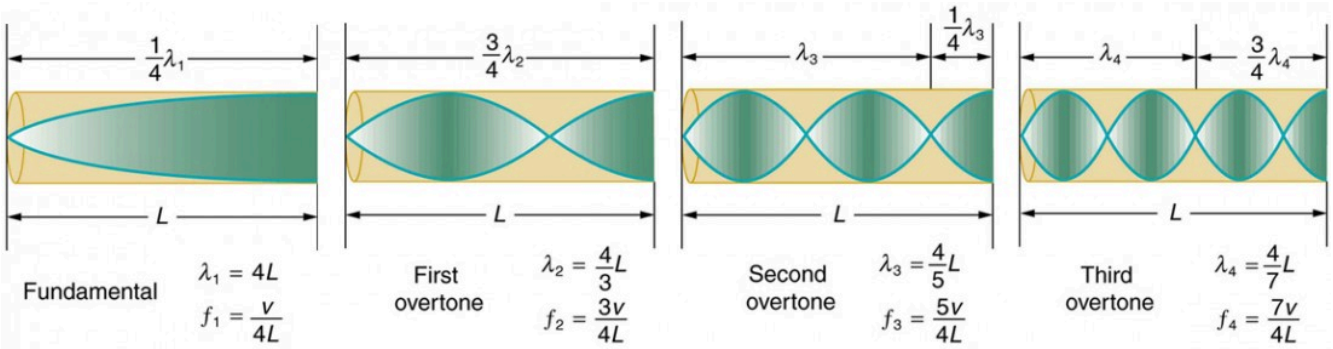


Figure 9. The fundamental and three lowest overtones for a tube closed at one end. All have maximum air displacements at the open end and none at the closed end.

The fundamental and overtones can be present simultaneously in a variety of combinations. For example, middle C on a trumpet has a sound distinctively different from middle C on a clarinet, both instruments being modified versions of a tube closed at one end. The fundamental frequency is the same (and usually the most intense), but the overtones and their mix of intensities are different and subject to shading by the musician. This mix is what gives various musical instruments (and human voices) their distinctive characteristics, whether they have air columns, strings, sounding boxes, or drumheads. In fact, much of our speech is determined by shaping the cavity formed by the throat and mouth and positioning the tongue to adjust the fundamental and combination of overtones. Simple resonant cavities can be made to resonate with the sound of the vowels, for example. (See Figure 10.) In boys, at puberty, the larynx grows and the shape of the resonant cavity changes giving rise to the difference in predominant frequencies in speech between men and women.

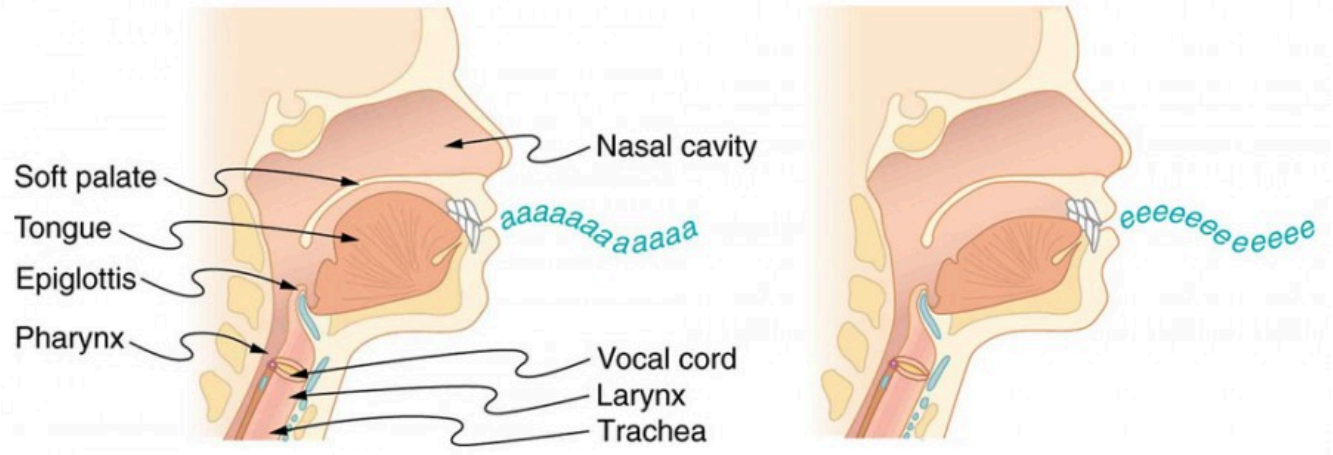


Figure 10. The throat and mouth form an air column closed at one end that resonates in response to vibrations in the voice box. The spectrum of overtones and their intensities vary with mouth shaping and tongue position to form different sounds. The voice box can be replaced with a mechanical vibrator, and understandable speech is still possible. Variations in basic shapes make different voices recognizable.

Now let us look for a pattern in the resonant frequencies for a simple tube that is closed at one end. The fundamental has $\lambda = 4L$, and frequency is related to wavelength and the speed of sound as given by $v_w = f\lambda$.

Solving for f in this equation gives

$$f = \frac{v_{\text{w}}}{\lambda} = \frac{v_{\text{w}}}{4L}$$

,

where v_{w} is the speed of sound in air. Similarly, the first overtone has

$$\lambda' = \frac{4L}{3}$$

(see Figure 9), so that

$$f' = 3\frac{v_{\text{w}}}{4L} = 3f$$

.

Because $f' = 3f$, we call the first overtone the third harmonic. Continuing this process, we see a pattern that can be generalized in a single expression. The resonant frequencies of a tube closed at one end are

$$f_n = n\frac{v_{\text{w}}}{4L}, n = 1, 3, 5$$

,

where f_1 is the fundamental, f_3 is the first overtone, and so on. It is interesting that the resonant frequencies depend on the speed of sound and, hence, on temperature. This dependence poses a noticeable problem for organs in old unheated cathedrals, and it is also the reason why musicians commonly bring their wind instruments to room temperature before playing them.

Example 1. Find the Length of a Tube with a 128 Hz Fundamental

1. What length should a tube closed at one end have on a day when the air temperature, is 22.0ºC, if its fundamental frequency is to be 128 Hz (C below middle C)?
2. What is the frequency of its fourth overtone?

Strategy

The length L can be found from the relationship in

$$f_n = n\frac{v_{\text{w}}}{4L}$$

, but we will first need to find the speed of sound v_{w}.

Solution for Part 1

Identify knowns:

- the fundamental frequency is 128 Hz
- the air temperature is 22.0ºC

Use

$$f_n = n\frac{v_\text{w}}{4L}$$

to find the fundamental frequency (n = 1):

$$f_1 = \frac{v_\text{w}}{4L}$$

Solve this equation for length:

$$L = \frac{v_\text{w}}{4f_1}$$

.

Find the speed of sound using

$$v_\text{w} = (331\ \text{m/s})\sqrt{\frac{T}{273\ \text{K}}}$$

.

$$v_\text{w} = (331\ \text{m/s})\sqrt{\frac{295\ \text{K}}{273\ \text{K}}} = 344\ \text{m/s}$$

Enter the values of the speed of sound and frequency into the expression for L.

$$L = \frac{v_\text{w}}{4f_1} = \frac{344\ \text{m/s}}{4(128\ \text{Hz})} = 0.672\ \text{m}$$

Discussion on Part 1

Many wind instruments are modified tubes that have finger holes, valves, and other devices for changing the length of the resonating air column and hence, the frequency of the note played. Horns producing very low frequencies, such as tubas, require tubes so long that they are coiled into loops.

Solution for Part 2

Identify knowns:

- the first overtone has n = 3
- the second overtone has n = 5
- the third overtone has n = 7
- the fourth overtone has n = 9

Enter the value for the fourth overtone into

$$f_n = n\frac{v_\text{w}}{4L}$$

:

$$f_9 = 9\frac{v_\text{w}}{4L} = 9f_1 = 1.15\ \text{kHz}$$

Discussion on Part 2

Whether this overtone occurs in a simple tube or a musical instrument depends on how it is stimulated to vibrate and the details of its shape. The trombone, for example, does not produce its fundamental frequency and only makes overtones.

Another type of tube is one that is *open* at both ends. Examples are some organ pipes, flutes, and oboes. The resonances of tubes open at both ends can be analyzed in a very similar fashion to those for tubes closed at one end. The air columns in tubes open at both ends have maximum air displacements at both ends, as illustrated in Figure 11. Standing waves form as shown.

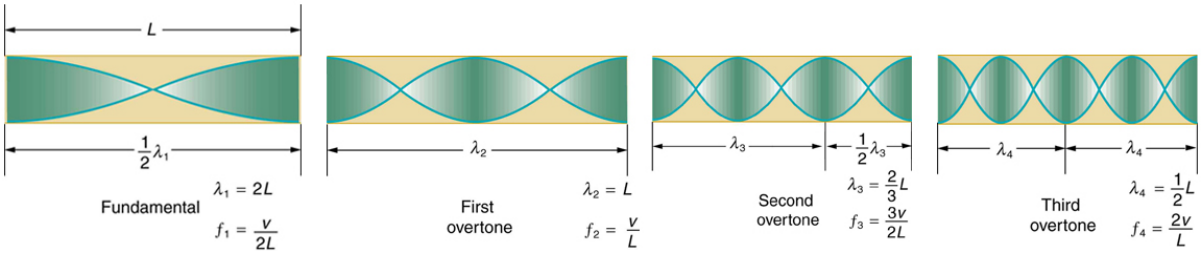


Figure 11. The resonant frequencies of a tube open at both ends are shown, including the fundamental and the first three overtones. In all cases the maximum air displacements occur at both ends of the tube, giving it different natural frequencies than a tube closed at one end.

Based on the fact that a tube open at both ends has maximum air displacements at both ends, and using Figure 11 as a guide, we can see that the resonant frequencies of a tube open at both ends are:

$$f_n = n\frac{v_w}{2L}, n = 1, 2, 3, \ldots,$$

where f_1 is the fundamental, f_2 is the first overtone, f_3 is the second overtone, and so on. Note that a tube open at both ends has a fundamental frequency twice what it would have if closed at one end. It also has a different spectrum of overtones than a tube closed at one end. So if you had two tubes with the same fundamental frequency but one was open at both ends and the other was closed at one end, they would sound different when played because they have different overtones. Middle C, for example, would sound richer played on an open tube, because it has even multiples of the fundamental as well as odd. A closed tube has only odd multiples.

Real-World Applications: Resonance in Everyday Systems

Resonance occurs in many different systems, including strings, air columns, and atoms. Resonance is the driven or forced oscillation of a system at its natural frequency. At resonance, energy is transferred rapidly to the oscillating system, and the amplitude of its oscillations grows until the system can no longer be described by Hooke's law. An example of this is the distorted sound intentionally produced in certain types of rock music.

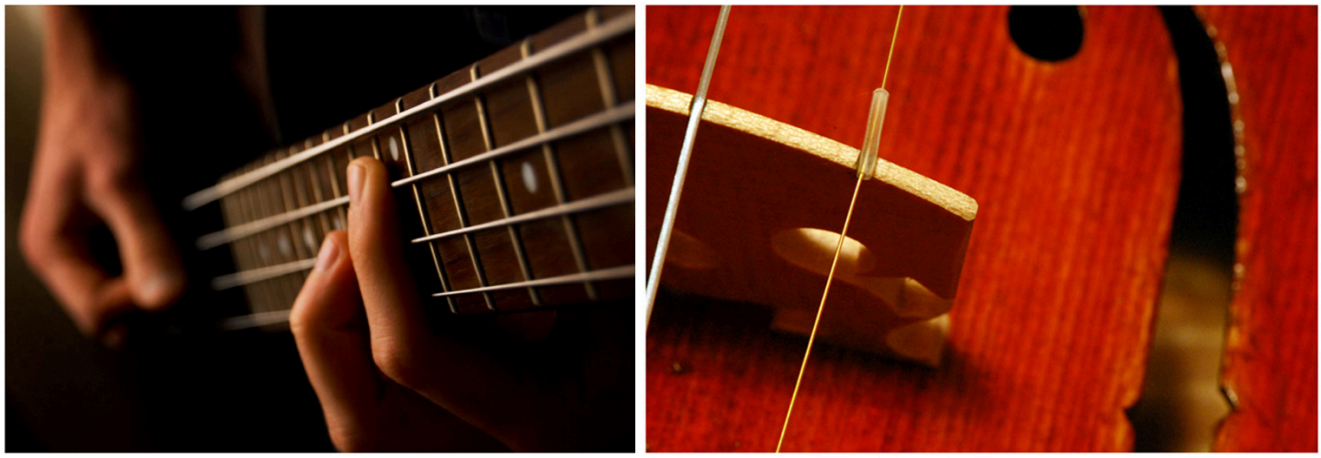


Figure 12. String instruments such as violins and guitars use resonance in their sounding boxes to amplify and enrich the sound created by their vibrating strings. The bridge and supports couple the string vibrations to the sounding boxes and air within. (credits: guitar, Feliciano Guimares, Fotopedia; violin, Steve Snodgrass, Flickr)

Wind instruments use resonance in air columns to amplify tones made by lips or vibrating reeds. Other instruments also use air resonance in clever ways to amplify sound. Figure 12 shows a violin and a guitar, both of which have sounding boxes but with different shapes, resulting in different overtone structures. The vibrating string creates a sound that resonates in the sounding box, greatly amplifying the sound and creating overtones that give the instrument its characteristic flavor. The more complex the shape of the sounding box, the greater its ability to resonate over a wide range of frequencies. The marimba, like the one shown in Figure 13 uses pots or gourds below the wooden slats to amplify their tones. The resonance of the pot can be adjusted by adding water.

Figure 13. Resonance has been used in musical instruments since prehistoric times. This marimba uses gourds as resonance chambers to amplify its sound. (credit: APC Events, Flickr)

We have emphasized sound applications in our discussions of resonance and standing waves, but these ideas apply to any system that has wave characteristics. Vibrating strings, for example, are actually resonating and have fundamentals and overtones similar to those for air columns. More subtle are the resonances in atoms due to the wave character of their electrons. Their orbitals can be viewed as standing waves, which have a fundamental (ground state) and overtones (excited states). It is fascinating that wave characteristics apply to such a wide range of physical systems.

Check Your Understanding

Part 1

Describe how noise-canceling headphones differ from standard headphones used to block outside sounds.

Solution

Regular headphones only block sound waves with a physical barrier. Noise-canceling headphones use destructive interference to reduce the loudness of outside sounds.

Part 2

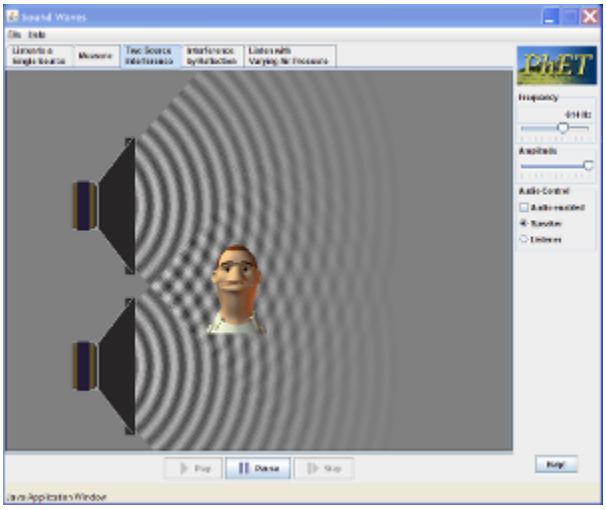
How is it possible to use a standing wave's node and antinode to determine the length of a closed-end tube?

Solution

When the tube resonates at its natural frequency, the wave's node is located at the closed end of the tube, and the antinode is located at the open end. The length of the tube is equal to one-fourth of the wavelength of this wave. Thus, if we know the wavelength of the wave, we can determine the length of the tube.

PhET Explorations: Sound

This simulation lets you see sound waves. Adjust the frequency or volume and you can see and hear how the wave changes. Move the listener around and hear what she hears.



Click to download the simulation. Run using Java.

Section Summary

- Sound interference and resonance have the same properties as defined for all waves.
- In air columns, the lowest-frequency resonance is called the fundamental, whereas all higher resonant frequencies are called overtones. Collectively, they are called harmonics.

- The resonant frequencies of a tube closed at one end are: $f_n = n\frac{v_w}{4L},\ n = 1, 3, 5\ldots$, f_1 is the fundamental and L is the length of the tube.
- The resonant frequencies of a tube open at both ends are: $f_n = n\frac{v_w}{2L},\ n = 1, 2, 3\ldots$

Conceptual Questions

1. How does an unamplified guitar produce sounds so much more intense than those of a plucked string held taut by a simple stick?
2. You are given two wind instruments of identical length. One is open at both ends, whereas the other is closed at one end. Which is able to produce the lowest frequency?
3. What is the difference between an overtone and a harmonic? Are all harmonics overtones? Are all overtones harmonics?

Problems & Exercises

1. A “showy” custom-built car has two brass horns that are supposed to produce the same frequency but actually emit 263.8 and 264.5 Hz. What beat frequency is produced?

2. What beat frequencies will be present: (a) If the musical notes A and C are played together (frequencies of 220 and 264 Hz)? (b) If D and F are played together (frequencies of 297 and 352 Hz)? (c) If all four are played together?

3. What beat frequencies result if a piano hammer hits three strings that emit frequencies of 127.8, 128.1, and 128.3 Hz?

4. A piano tuner hears a beat every 2.00 s when listening to a 264.0-Hz tuning fork and a single piano string. What are the two possible frequencies of the string?

5. (a) What is the fundamental frequency of a 0.672-m-long tube, open at both ends, on a day when the speed of sound is 344 m/s? (b) What is the frequency of its second harmonic?

6. If a wind instrument, such as a tuba, has a fundamental frequency of 32.0 Hz, what are its first three overtones? It is closed at one end. (The overtones of a real tuba are more complex than this example, because it is a tapered tube.)

7. What are the first three overtones of a bassoon that has a fundamental frequency of 90.0 Hz? It is open at both ends. (The overtones of a real bassoon are more complex than this example, because its double reed makes it act more like a tube closed at one end.)

8. How long must a flute be in order to have a fundamental frequency of 262 Hz (this frequency corresponds to middle C on the evenly tempered chromatic scale) on a day when air temperature is 20.0ºC? It is open at both ends.

9. What length should an oboe have to produce a fundamental frequency of 110 Hz on a day when the speed of sound is 343 m/s? It is open at both ends.

10. What is the length of a tube that has a fundamental frequency of 176 Hz and a first overtone of 352 Hz if the speed of sound is 343 m/s?

11. (a) Find the length of an organ pipe closed at one end that produces a fundamental frequency of 256 Hz when air temperature is 18.0ºC. (b) What is its fundamental frequency at 25.0ºC?

12. By what fraction will the frequencies produced by a wind instrument change when air temperature goes from 10.0ºC to 30.0ºC? That is, find the ratio of the frequencies at those temperatures.

13. The ear canal resonates like a tube closed at one end. (See Figure 5 in Hearing.) If ear canals range in length from 1.80 to 2.60 cm in an average population, what is the range of fundamental resonant frequencies? Take air temperature to be 37.0ºC, which is the same as body temperature. How does this result correlate with the intensity versus frequency graph of the human ear (Figure 14)?

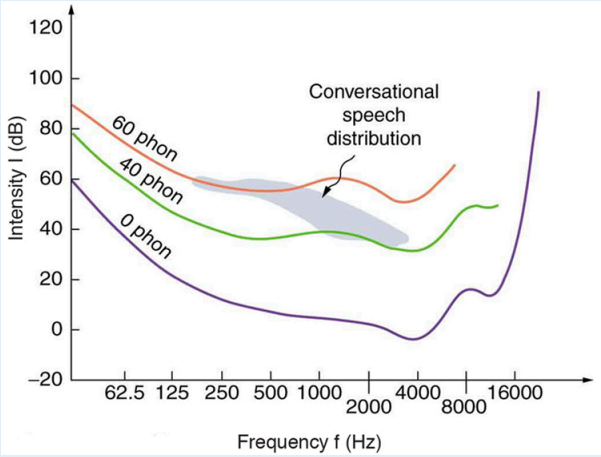


Figure 14. The shaded region represents frequencies and intensity levels found in normal conversational speech. The 0-phon line represents the normal hearing threshold, while those at 40 and 60 represent thresholds for people with 40- and 60-phon hearing losses, respectively.

14. Calculate the first overtone in an ear canal, which resonates like a 2.40-cm-long tube closed at one end, by taking air temperature to be 37.0ºC. Is the ear particularly sensitive to such a frequency? (The resonances of the ear canal are complicated by its nonuniform shape, which we shall ignore.)

15. A crude approximation of voice production is to consider the breathing passages and mouth to be a resonating tube closed at one end. (See Figure 10.) (a) What is the fundamental frequency if the tube is 0.240-m long, by taking air temperature to be 37.0ºC? (b) What would this frequency become if the person replaced the air with helium? Assume the same temperature dependence for helium as for air.

16. (a) Students in a physics lab are asked to find the length of an air column in a tube closed at one end that has a fundamental frequency of 256 Hz. They hold the tube vertically and fill it with water to the top, then lower the water while a 256-Hz tuning fork is rung and listen for the first resonance. What is the air temperature if the resonance occurs for a length of 0.336 m? (b) At what length will they observe the second resonance (first overtone)?

17. What frequencies will a 1.80-m-long tube produce in the audible range at 20.0ºC if: (a) The tube is closed at one end? (b) It is open at both ends?

Glossary

antinode: point of maximum displacement

node: point of zero displacement

fundamental: the lowest-frequency resonance

overtones: all resonant frequencies higher than the fundamental

harmonics: the term used to refer collectively to the fundamental and its overtones

Selected Solutions to Problems & Exercises

1. 0.7 Hz

3. 0.3 Hz, 0.2 Hz, 0.5 Hz

5. (a) 256 Hz; (b) 512 Hz

7. 180 Hz, 270 Hz, 360 Hz

9. 1.56 m

11. (a) 0.334 m; (b) 259 Hz

13. 3.39 to 4.90 kHz

15. (a) 367 Hz; (b) 1.07 kHz

17. (a) $f_n = n(47.6\text{ Hz})$, $n = 1, 3, 5, \ldots, 419$; (b) $f_n = n(95.3\text{ Hz})$, $n = 1, 2, 3, \ldots, 210$

Hearing

Lumen Learning

Learning Objectives

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

- Define hearing, pitch, loudness, timbre, note, tone, phon, ultrasound, and infrasound.
- Compare loudness to frequency and intensity of a sound.
- Identify structures of the inner ear and explain how they relate to sound perception.

The human ear has a tremendous range and sensitivity. It can give us a wealth of simple information—such as pitch, loudness, and direction. And from its input we can detect musical quality and nuances of voiced emotion. How is our hearing related to the physical qualities of sound, and how does the hearing mechanism work?

Figure 1. Hearing allows this vocalist, his band, and his fans to enjoy music. (credit: West Point Public Affairs, Flickr)

Hearing is the perception of sound. (Perception is commonly defined to be awareness through the senses, a typically circular definition of higher-level processes in living organisms.) Normal human hearing encompasses frequencies from 20 to 20,000 Hz, an impressive range. Sounds below 20 Hz are called *infrasound*, whereas those above 20,000 Hz are *ultrasound*. Neither is perceived by the ear, although infrasound can sometimes be felt as vibrations. When we do hear low-frequency vibrations, such as the sounds of a diving board, we hear the individual vibrations only because there are higher-frequency sounds in each. Other animals have hearing ranges different from that of humans. Dogs can hear sounds as high as 30,000 Hz, whereas bats and dolphins can hear up to 100,000-Hz sounds. You may have noticed that dogs respond to the sound of a dog whistle which produces sound out of the range of human hearing. Elephants are known to respond to frequencies below 20 Hz.

The perception of frequency is called *pitch*. Most of us have excellent relative pitch, which means that we can tell whether one sound has a different frequency from another. Typically, we can discriminate between two sounds if their frequencies differ by 0.3% or more. For example, 500.0 and 501.5 Hz are noticeably different. Pitch perception is directly related to frequency and is not greatly affected by other physical quantities such as intensity. Musical *notes* are particular sounds that can be produced by most instruments and in Western music have particular names. Combinations of notes constitute music. Some people can identify musical notes, such as A-sharp, C, or E-flat, just by listening to them. This uncommon ability is called perfect pitch.

The ear is remarkably sensitive to low-intensity sounds. The lowest audible intensity or threshold is about 10^{-12} W/m^2 or 0 dB. Sounds as much as 10^{12} more intense can be briefly tolerated. Very few measuring devices are capable of observations over a range of a trillion. The perception of intensity is called *loudness*. At a given frequency, it is possible to discern differences of about 1 dB, and a change of 3 dB is easily noticed. But loudness is not related to intensity alone. Frequency has a major effect on how loud a sound seems. The ear has its maximum sensitivity to frequencies in the range of 2000 to 5000 Hz, so that sounds in this range are perceived as being louder than, say, those at 500 or 10,000 Hz, even when they all have the same intensity. Sounds near the high- and low-frequency extremes of the hearing range seem even less loud, because the ear is even less sensitive at those frequencies. Table 1 gives the dependence of certain human hearing perceptions on physical quantities.

Table 1. Sound Perceptions

Perception	**Physical quantity**
Pitch	Frequency
Loudness	Intensity and Frequency
Timbre	Number and relative intensity of multiple frequencies. Subtle craftsmanship leads to non-linear effects and more detail.
Note	Basic unit of music with specific names, combined to generate tunes
Tone	Number and relative intensity of multiple frequencies.

When a violin plays middle C, there is no mistaking it for a piano playing the same note. The reason is that each instrument produces a distinctive set of frequencies and intensities. We call our perception of these combinations of frequencies and intensities *tone* quality, or more commonly the *timbre* of the sound. It is more difficult to correlate timbre perception to physical quantities than it is for loudness or pitch perception. Timbre is more subjective. Terms such as dull, brilliant, warm, cold, pure, and rich are employed to describe the timbre of a sound. So the consideration of timbre takes us into the realm of perceptual psychology, where higher-level processes in the brain are dominant. This is true for other perceptions of sound, such as music and noise. We shall not delve further into them; rather, we will concentrate on the question of loudness perception.

A unit called a *phon* is used to express loudness numerically. Phons differ from decibels because the phon is a unit of loudness perception, whereas the decibel is a unit of physical intensity. Figure 2 shows the relationship of loudness to intensity (or intensity level) and frequency for persons with normal hearing. The curved lines are equal-loudness curves. Each curve is labeled with its loudness in phons. Any sound along a given curve will be perceived as equally loud by the average person. The curves were determined by having large numbers of people compare the loudness of sounds at different frequencies and sound intensity levels. At a frequency of 1000 Hz, phons are taken to be numerically equal to decibels. The following example helps illustrate how to use the graph:

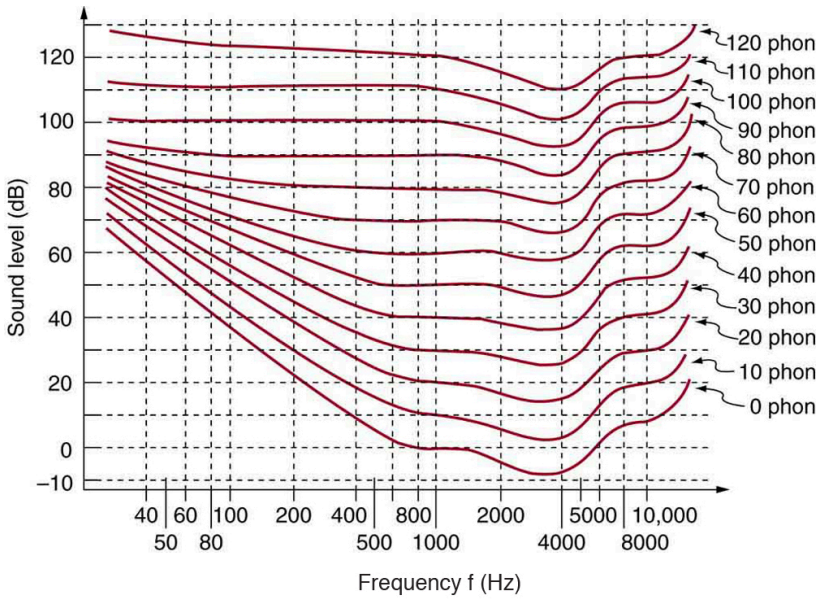


Figure 2. The relationship of loudness in phons to intensity level (in decibels) and intensity (in watts per meter squared) for persons with normal hearing. The curved lines are equal-loudness curves—all sounds on a given curve are perceived as equally loud. Phons and decibels are defined to be the same at 1000 Hz.

Example 1. Measuring Loudness: Loudness Versus Intensity Level and Frequency

1. What is the loudness in phons of a 100-Hz sound that has an intensity level of 80 dB?
2. What is the intensity level in decibels of a 4000-Hz sound having a loudness of 70 phons?
3. At what intensity level will an 8000-Hz sound have the same loudness as a 200-Hz sound at 60 dB?

Strategy for Part 1

The graph in Figure 2 should be referenced in order to solve this example. To find the loudness of a given sound, you must know its frequency and intensity level and locate that point on the square grid, then interpolate between loudness curves to get the loudness in phons.

Solution for Part 1

Identify knowns:

- The square grid of the graph relating phons and decibels is a plot of intensity level versus frequency—both physical quantities.
- 100 Hz at 80 dB lies halfway between the curves marked 70 and 80 phons.

Find the loudness: 75 phons.

Strategy for Part 2

The graph in Figure 2 should be referenced in order to solve this example. To find the intensity level of a

sound, you must have its frequency and loudness. Once that point is located, the intensity level can be determined from the vertical axis.

Solution for Part 2

Identify knowns; Values are given to be 4000 Hz at 70 phons.

Follow the 70-phon curve until it reaches 4000 Hz. At that point, it is below the 70 dB line at about 67 dB.

Find the intensity level: 67 dB

Strategy for Part 3

The graph in Figure 2 should be referenced in order to solve this example.

Solution for Part 3

Locate the point for a 200 Hz and 60 dB sound. Find the loudness: This point lies just slightly above the 50-phon curve, and so its loudness is 51 phons. Look for the 51-phon level is at 8000 Hz: 63 dB.

Discussion

These answers, like all information extracted from Figure 2, have uncertainties of several phons or several decibels, partly due to difficulties in interpolation, but mostly related to uncertainties in the equal-loudness curves.

Further examination of the graph in Figure 2 reveals some interesting facts about human hearing. First, sounds below the 0-phon curve are not perceived by most people. So, for example, a 60 Hz sound at 40 dB is inaudible. The 0-phon curve represents the threshold of normal hearing. We can hear some sounds at intensity levels below 0 dB. For example, a 3-dB, 5000-Hz sound is audible, because it lies above the 0-phon curve. The loudness curves all have dips in them between about 2000 and 5000 Hz. These dips mean the ear is most sensitive to frequencies in that range. For example, a 15-dB sound at 4000 Hz has a loudness of 20 phons, the same as a 20-dB sound at 1000 Hz. The curves rise at both extremes of the frequency range, indicating that a greater-intensity level sound is needed at those frequencies to be perceived to be as loud as at middle frequencies. For example, a sound at 10,000 Hz must have an intensity level of 30 dB to seem as loud as a 20 dB sound at 1000 Hz. Sounds above 120 phons are painful as well as damaging.

We do not often utilize our full range of hearing. This is particularly true for frequencies above 8000 Hz, which are rare in the environment and are unnecessary for understanding conversation or appreciating music. In fact, people who have lost the ability to hear such high frequencies are usually unaware of their loss until tested. The shaded region in Figure 3 is the frequency and intensity region where most conversational sounds fall. The curved lines indicate what effect hearing losses of 40 and 60 phons will have. A 40-phon hearing loss at all frequencies still allows a person to understand conversation, although it will seem very quiet. A person with a 60-phon loss at all frequencies will hear only the lowest frequencies and will not be able to understand speech unless it is much louder than normal. Even so, speech may seem indistinct, because higher frequencies are not as well perceived. The conversational speech region also has a gender component, in that female voices are usually characterized by higher frequencies. So the person with a 60-phon hearing impediment might have difficulty understanding the normal conversation of a woman.

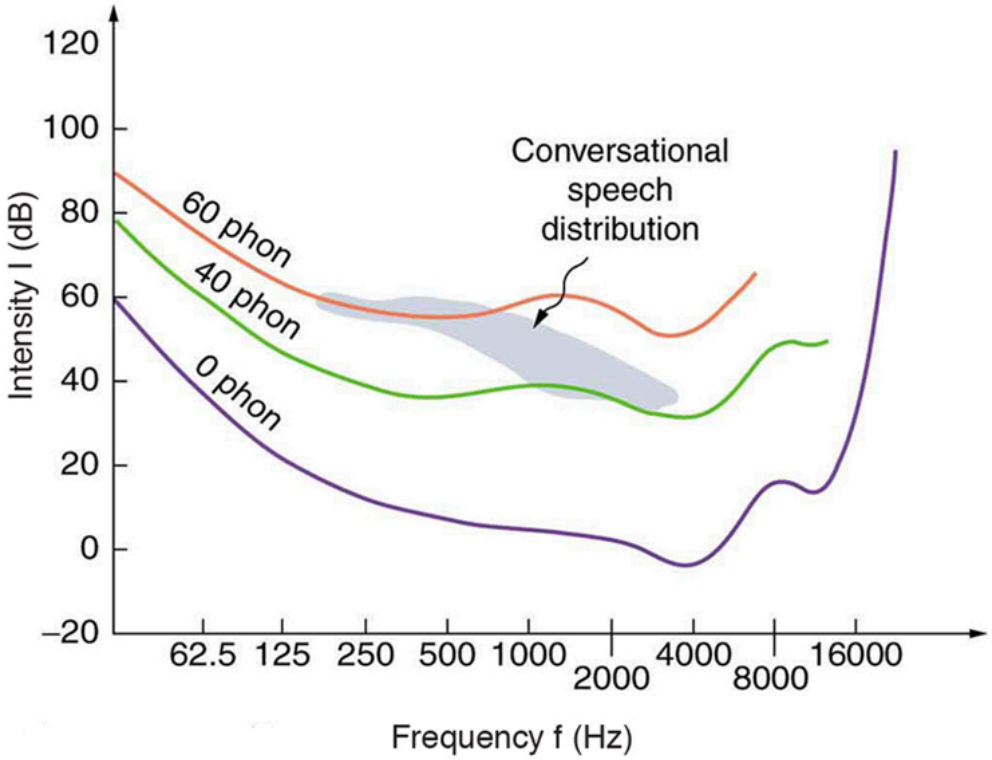
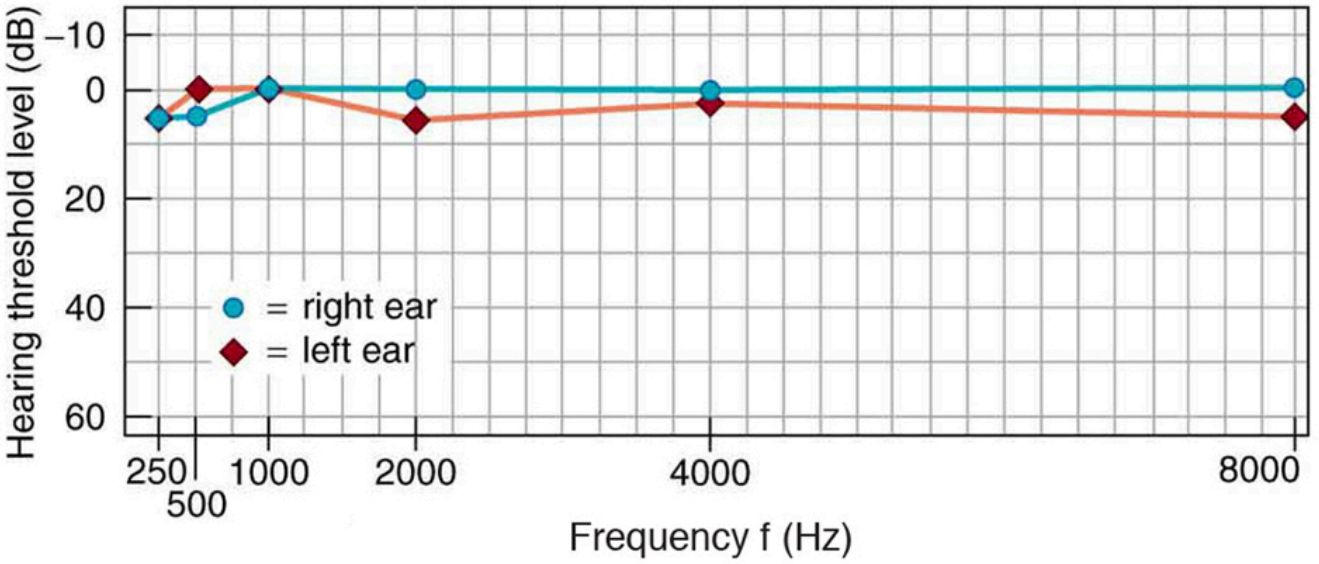
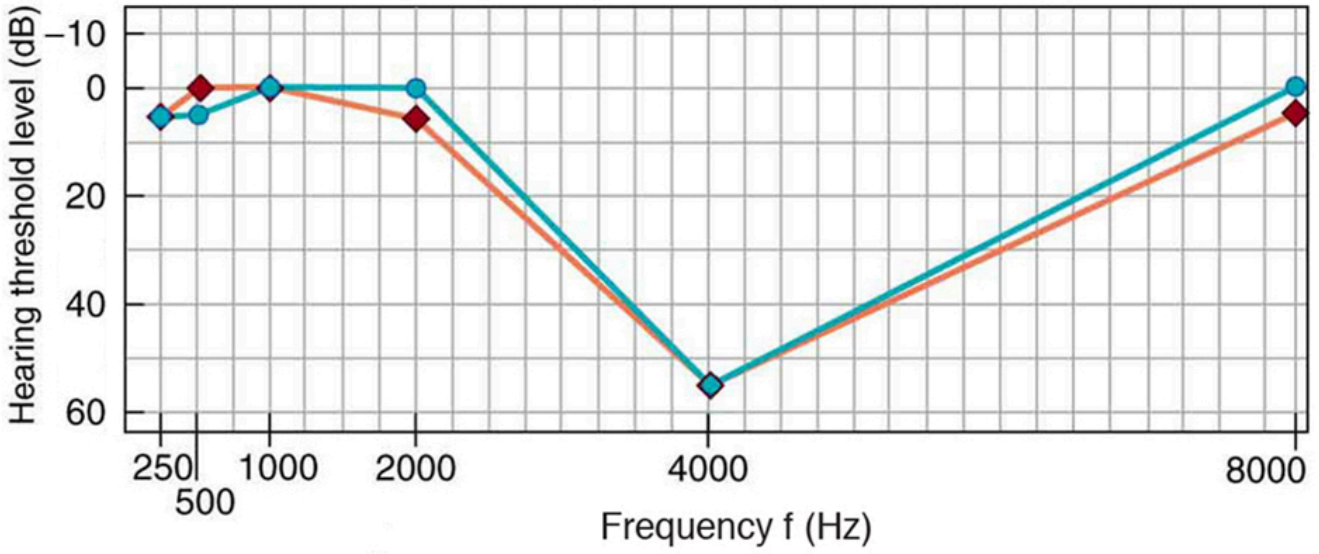


Figure 3. The shaded region represents frequencies and intensity levels found in normal conversational speech. The 0-phon line represents the normal hearing threshold, while those at 40 and 60 represent thresholds for people with 40- and 60-phon hearing losses, respectively.

Hearing tests are performed over a range of frequencies, usually from 250 to 8000 Hz, and can be displayed graphically in an audiogram like that in Figure 4. The hearing threshold is measured in dB *relative to the normal threshold*, so that normal hearing registers as 0 dB at all frequencies. Hearing loss caused by noise typically shows a dip near the 4000 Hz frequency, irrespective of the frequency that caused the loss and often affects both ears. The most common form of hearing loss comes with age and is called *presbycusis*—literally elder ear. Such loss is increasingly severe at higher frequencies, and interferes with music appreciation and speech recognition.


Hearing threshold level (dB)
−10
0
20
40
60
= right ear
= left ear
250
500
1000
2000
4000
8000
Frequency f (Hz)


Hearing threshold level (dB)
−10
0
20
40
60
250
500
1000
2000
4000
8000
Frequency f (Hz)

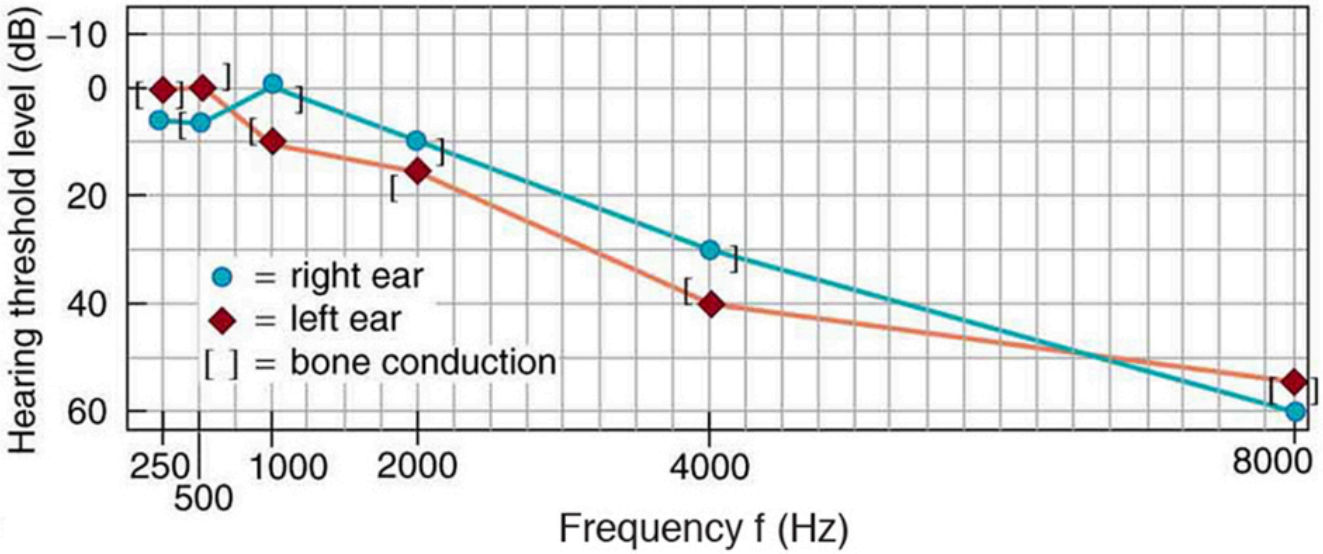

Hearing threshold level (dB)
−10
0
20
40
60
= right ear
= left ear
[ ] = bone conduction
250
500
1000
2000
4000
8000
Frequency f (Hz)

Figure 4. Audiograms showing the threshold in intensity level versus frequency for three different individuals. Intensity level is measured relative to the normal threshold. The top left graph is that of a person with normal hearing. The graph to its right has a dip at 4000 Hz and is that of a child who suffered hearing loss due to a cap gun. The third graph is typical of presbycusis, the progressive loss of higher frequency hearing with age. Tests performed by bone conduction (brackets) can distinguish nerve damage from middle ear damage.

The Hearing Mechanism

The hearing mechanism involves some interesting physics. The sound wave that impinges upon our ear is a pressure wave. The ear is a transducer that converts sound waves into electrical nerve impulses in a manner much more sophisticated than, but analogous to, a microphone. Figure 5 shows the gross anatomy of the ear with its division into three parts: the outer ear or ear canal; the middle ear, which runs from the eardrum to the cochlea; and the inner ear, which is the cochlea itself. The body part normally referred to as the ear is technically called the pinna.

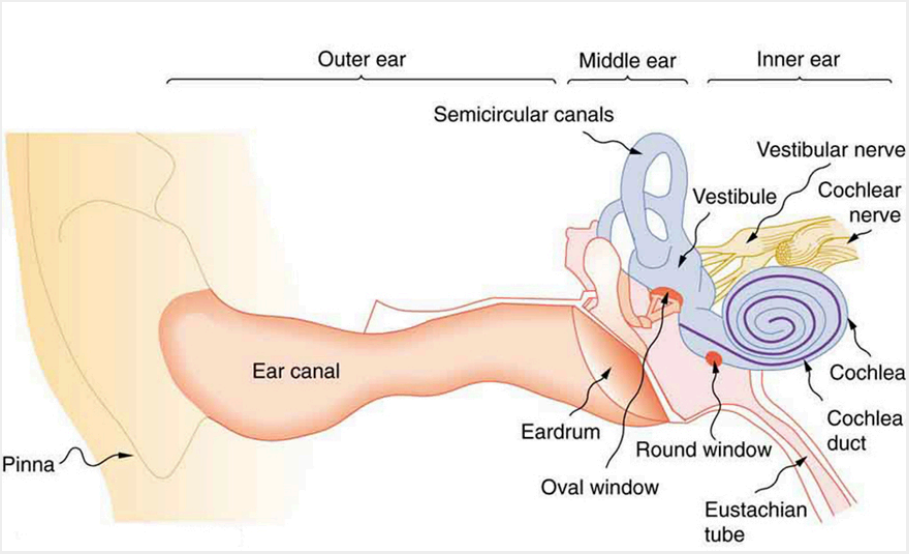


Figure 5. The illustration shows the gross anatomy of the human ear.

The outer ear, or ear canal, carries sound to the recessed protected eardrum. The air column in the ear canal resonates and is partially responsible for the sensitivity of the ear to sounds in the 2000 to 5000 Hz range. The middle ear converts sound into mechanical vibrations and applies these vibrations to the cochlea. The lever system of the middle ear takes the force exerted on the eardrum by sound pressure variations, amplifies it and transmits it to the inner ear via the oval window, creating pressure waves in the cochlea approximately 40 times greater than those impinging on the eardrum. (See Figure 6.) Two muscles in the middle ear (not shown) protect the inner ear from very intense sounds. They react to intense sound in a few milliseconds and reduce the force transmitted to the cochlea. This protective reaction can also be triggered by your own voice, so that humming while shooting a gun, for example, can reduce noise damage.

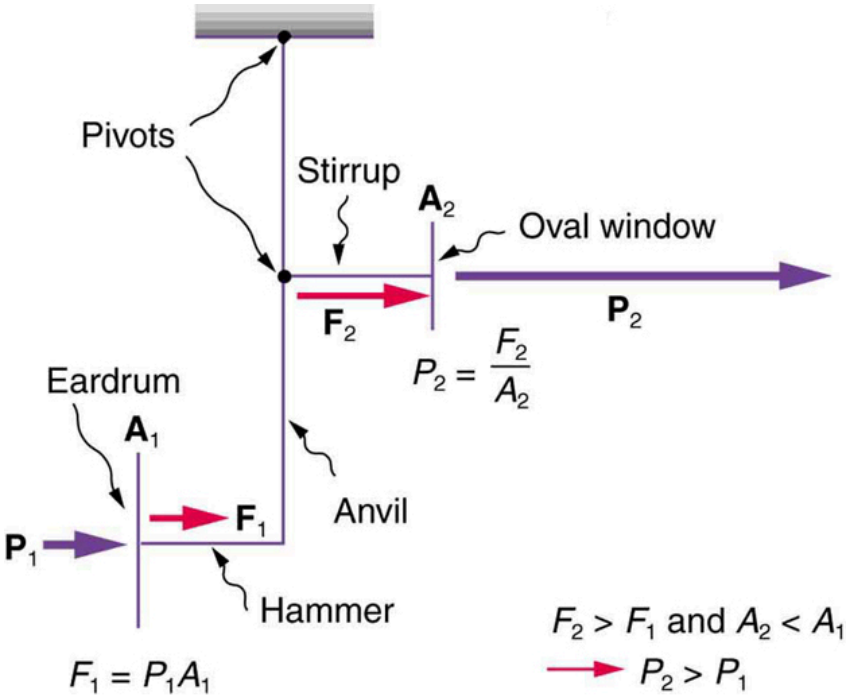


Figure 6. This schematic shows the middle ear's system for converting sound pressure into force, increasing that force through a lever system, and applying the increased force to a small area of the cochlea, thereby creating a pressure about 40 times that in the original sound wave. A protective muscle reaction to intense sounds greatly reduces the mechanical advantage of the lever system.

Figure 7 shows the middle and inner ear in greater detail. Pressure waves moving through the cochlea cause the tectorial membrane to vibrate, rubbing cilia (called hair cells), which stimulate nerves that send electrical signals to the brain. The membrane resonates at different positions for different frequencies, with high frequencies stimulating nerves at the near end and low frequencies at the far end. The complete operation of the cochlea is still not understood, but several mechanisms for sending information to the brain are known to be involved. For sounds below about 1000 Hz, the nerves send signals at the same frequency as the sound. For frequencies greater than about 1000 Hz, the nerves signal frequency by position. There is a structure to the cilia, and there are connections between nerve cells that perform signal processing before information is sent to the brain. Intensity information is partly indicated by the number of nerve signals and by volleys of signals. The brain processes the cochlear nerve signals to provide additional information such as source direction (based on time and intensity comparisons of sounds from both ears). Higher-level processing produces many nuances, such as music appreciation.

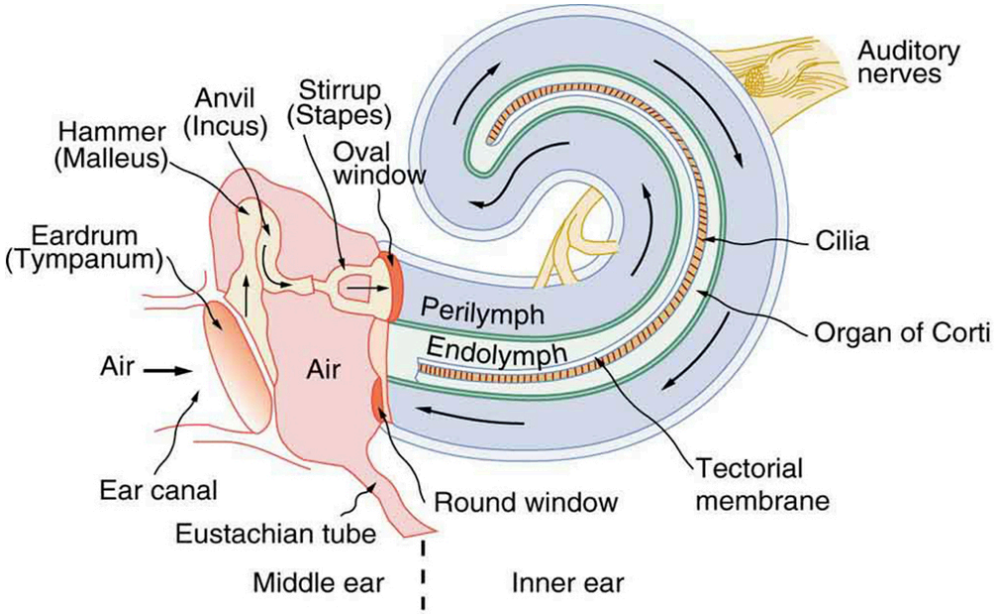


Figure 7. The inner ear, or cochlea, is a coiled tube about 3 mm in diameter and 3 cm in length if uncoiled. When the oval window is forced inward, as shown, a pressure wave travels through the perilymph in the direction of the arrows, stimulating nerves at the base of cilia in the organ of Corti.

Hearing losses can occur because of problems in the middle or inner ear. Conductive losses in the middle ear can be partially overcome by sending sound vibrations to the cochlea through the skull. Hearing aids for this purpose usually press against the bone behind the ear, rather than simply amplifying the sound sent into the ear canal as many hearing aids do. Damage to the nerves in the cochlea is not repairable, but amplification can partially compensate. There is a risk that amplification will produce further damage. Another common failure in the cochlea is damage or loss of the cilia but with nerves remaining functional. Cochlear implants that stimulate the nerves directly are now available and widely accepted. Over 100,000 implants are in use, in about equal numbers of adults and children.

The cochlear implant was pioneered in Melbourne, Australia, by Graeme Clark in the 1970s for his deaf father. The implant consists of three external components and two internal components. The external components are a microphone for picking up sound and converting it into an electrical signal, a speech processor to select certain frequencies and a transmitter to transfer the signal to the internal components through electromagnetic induction. The internal components consist of a receiver/transmitter secured in the bone beneath the skin, which converts the signals into electric impulses and sends them through an internal cable to the cochlea and an array of about 24 electrodes wound through the cochlea. These electrodes in turn send the impulses directly into the brain. The electrodes basically emulate the cilia.

Check Your Understanding

Are ultrasound and infrasound imperceptible to all hearing organisms? Explain your answer.

Solution

No, the range of perceptible sound is based in the range of human hearing. Many other organisms perceive either infrasound or ultrasound.

Section Summary

- The range of audible frequencies is 20 to 20,000 Hz.
- Those sounds above 20,000 Hz are ultrasound, whereas those below 20 Hz are infrasound.
- The perception of frequency is pitch.
- The perception of intensity is loudness.
- Loudness has units of phons.

Conceptual Questions

1. Why can a hearing test show that your threshold of hearing is 0 dB at 250 Hz, when Figure 3 implies that no one can hear such a frequency at less than 20 dB?

Problems & Exercises

1. The factor of 10^{-12} in the range of intensities to which the ear can respond, from threshold to that causing damage after brief exposure, is truly remarkable. If you could measure distances over the same range with a single instrument and the smallest distance you could measure was 1 mm, what would the largest be?
2. The frequencies to which the ear responds vary by a factor of 10^3. Suppose the speedometer on your car measured speeds differing by the same factor of 10^3, and the greatest speed it reads is 90.0 mi/h. What would be the slowest nonzero speed it could read?
3. What are the closest frequencies to 500 Hz that an average person can clearly distinguish as being different in frequency from 500 Hz? The sounds are not present simultaneously.
4. Can the average person tell that a 2002-Hz sound has a different frequency than a 1999-Hz sound without playing them simultaneously?
5. If your radio is producing an average sound intensity level of 85 dB, what is the next lowest sound intensity level that is clearly less intense?
6. Can you tell that your roommate turned up the sound on the TV if its average sound intensity level goes from 70 to 73 dB?
7. Based on the graph in Figure 2, what is the threshold of hearing in decibels for frequencies of 60, 400, 1000, 4000, and 15,000 Hz? Note that many AC electrical appliances produce 60 Hz, music

is commonly 400 Hz, a reference frequency is 1000 Hz, your maximum sensitivity is near 4000 Hz, and many older TVs produce a 15,750 Hz whine.

8. What sound intensity levels must sounds of frequencies 60, 3000, and 8000 Hz have in order to have the same loudness as a 40-dB sound of frequency 1000 Hz (that is, to have a loudness of 40 phons)?
9. What is the approximate sound intensity level in decibels of a 600-Hz tone if it has a loudness of 20 phons? If it has a loudness of 70 phons?
10. (a) What are the loudnesses in phons of sounds having frequencies of 200, 1000, 5000, and 10,000 Hz, if they are all at the same 60.0-dB sound intensity level? (b) If they are all at 110 dB? (c) If they are all at 20.0 dB?
11. Suppose a person has a 50-dB hearing loss at all frequencies. By how many factors of 10 will low-intensity sounds need to be amplified to seem normal to this person? Note that smaller amplification is appropriate for more intense sounds to avoid further hearing damage.
12. If a woman needs an amplification of 5.0×10^{12} times the threshold intensity to enable her to hear at all frequencies, what is her overall hearing loss in dB? Note that smaller amplification is appropriate for more intense sounds to avoid further damage to her hearing from levels above 90 dB.
13. (a) What is the intensity in watts per meter squared of a just barely audible 200-Hz sound? (b) What is the intensity in watts per meter squared of a barely audible 4000-Hz sound?
14. (a) Find the intensity in watts per meter squared of a 60.0-Hz sound having a loudness of 60 phons. (b) Find the intensity in watts per meter squared of a 10,000-Hz sound having a loudness of 60 phons.
15. A person has a hearing threshold 10 dB above normal at 100 Hz and 50 dB above normal at 4000 Hz. How much more intense must a 100-Hz tone be than a 4000-Hz tone if they are both barely audible to this person?
16. A child has a hearing loss of 60 dB near 5000 Hz, due to noise exposure, and normal hearing elsewhere. How much more intense is a 5000-Hz tone than a 400-Hz tone if they are both barely audible to the child?
17. What is the ratio of intensities of two sounds of identical frequency if the first is just barely discernible as louder to a person than the second?

Glossary

loudness: the perception of sound intensity

timbre: number and relative intensity of multiple sound frequencies

note: basic unit of music with specific names, combined to generate tunes

tone: number and relative intensity of multiple sound frequencies

phon: the numerical unit of loudness

ultrasound: sounds above 20,000 Hz

infrasound: sounds below 20 Hz

Selected Solutions to Problems & Exercises

1. 1×10^6 km

3. 498.5 or 501.5 Hz

5. 82 dB

7. approximately 48, 9, 0, –7, and 20 dB, respectively

9. (a) 23 dB; (b) 70 dB

11. Five factors of 10

13. (a) 2×10^{-10} W/m^2; (b) 2×10^{-13} W/m^2

15. 2.5

17. 1.26

Ultrasound

Lumen Learning

Learning Objectives

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

- Define acoustic impedance and intensity reflection coefficient.
- Describe medical and other uses of ultrasound technology.
- Calculate acoustic impedance using density values and the speed of ultrasound.
- Calculate the velocity of a moving object using Doppler-shifted ultrasound.

Any sound with a frequency above 20,000 Hz (or 20 kHz)—that is, above the highest audible frequency—is defined to be ultrasound. In practice, it is possible to create ultrasound frequencies up to more than a gigahertz. (Higher frequencies are difficult to create; furthermore, they propagate poorly because they are very strongly absorbed.) Ultrasound has a tremendous number of applications, which range from burglar alarms to use in cleaning delicate objects to the guidance systems of bats. We begin our discussion of ultrasound with some of its applications in medicine, in which it is used extensively both for diagnosis and for therapy.

Figure 1. Ultrasound is used in medicine to painlessly and noninvasively monitor patient health and diagnose a wide range of disorders. (credit: abbybatchelder, Flickr)

Characteristics of Ultrasound

The characteristics of ultrasound, such as frequency and intensity, are wave properties common to all types of waves. Ultrasound also has a wavelength that limits the fineness of detail it can detect. This characteristic is true of all waves. We can never observe details significantly smaller than the wavelength of our probe; for example, we will never see individual atoms with visible light, because the atoms are so small compared with the wavelength of light.

Ultrasound in Medical Therapy

Ultrasound, like any wave, carries energy that can be absorbed by the medium carrying it, producing effects that vary with intensity. When focused to intensities of 10^3 to 10^5 W/m^2, ultrasound can be used to shatter gallstones or pulverize cancerous tissue in surgical procedures. (See Figure 2.) Intensities this great can damage individual cells, variously causing their protoplasm to stream inside them, altering their permeability, or rupturing their walls through *cavitation*. Cavitation is the creation of vapor cavities in a fluid—the longitudinal vibrations in ultrasound alternatively compress and expand the medium, and at sufficient amplitudes the expansion separates molecules. Most cavitation damage is done when the cavities collapse, producing even greater shock pressures.

Most of the energy carried by high-intensity ultrasound in tissue is converted to thermal energy. In fact, intensities of 10^3 to 10^4 W/m^2 are commonly used for deep-heat treatments called ultrasound diathermy. Frequencies of 0.8 to 1 MHz are typical. In both athletics and physical therapy, ultrasound diathermy is most often applied to injured or overworked muscles to relieve pain and improve flexibility. Skill is needed by the therapist to avoid "bone burns" and other tissue damage caused by overheating and cavitation, sometimes made worse by reflection and focusing of the ultrasound by joint and bone tissue.

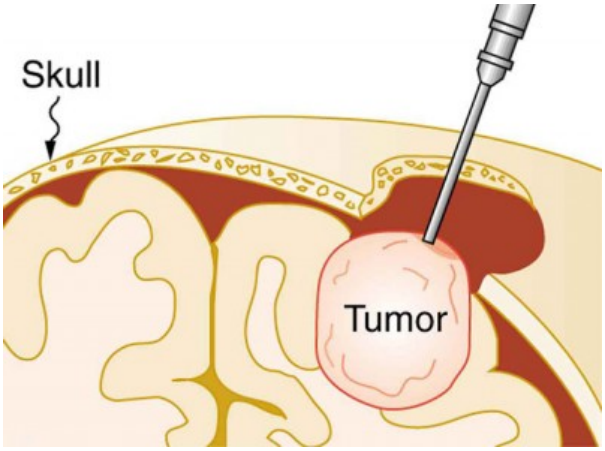


Figure 2. The tip of this small probe oscillates at 23 kHz with such a large amplitude that it pulverizes tissue on contact. The debris is then aspirated. The speed of the tip may exceed the speed of sound in tissue, thus creating shock waves and cavitation, rather than a smooth simple harmonic oscillator–type wave.

In some instances, you may encounter a different decibel scale, called the sound *pressure* level, when ultrasound travels in water or in human and other biological tissues. We shall not use the scale here, but it is notable that numbers for sound pressure levels range 60 to 70 dB higher than you would quote for β, the sound intensity level used in this text. Should you encounter a sound pressure level of 220 decibels, then, it is not an astronomically high intensity, but equivalent to about 155 dB—high enough to destroy tissue, but not as unreasonably high as it might seem at first.

Ultrasound in Medical Diagnostics

When used for imaging, ultrasonic waves are emitted from a transducer, a crystal exhibiting the piezoelectric effect (the expansion and contraction of a substance when a voltage is applied across it, causing a vibration of the crystal). These high-frequency vibrations are transmitted into any tissue in contact with the transducer. Similarly, if a pressure is applied to the crystal (in the form of a wave reflected off tissue layers), a voltage is produced which can be recorded. The crystal therefore acts as both a transmitter and a receiver of sound. Ultrasound is also partially absorbed by tissue on its path, both on its journey away from the transducer and on its return journey. From the time between when the original signal is sent and when the reflections from various boundaries between media are received, (as well as a measure of the intensity loss of the signal), the nature and position of each boundary between tissues and organs may be deduced.

Reflections at boundaries between two different media occur because of differences in a characteristic known as the *acoustic impedance Z* of each substance. Impedance is defined as $Z = \rho v$, where ρ is the density of the medium (in kg/m^3) and v is the speed of sound through the medium (in m/s). The units for Z are therefore kg/(m$^2 \cdot$ s).

Table 1 shows the density and speed of sound through various media (including various soft tissues) and the associated acoustic impedances. Note that the acoustic impedances for soft tissue do not vary much but that there is a big difference between the acoustic impedance of soft tissue and air and also between soft tissue and bone.

Table 1. The Ultrasound Properties of Various Media, Including Soft Tissue Found in the Body

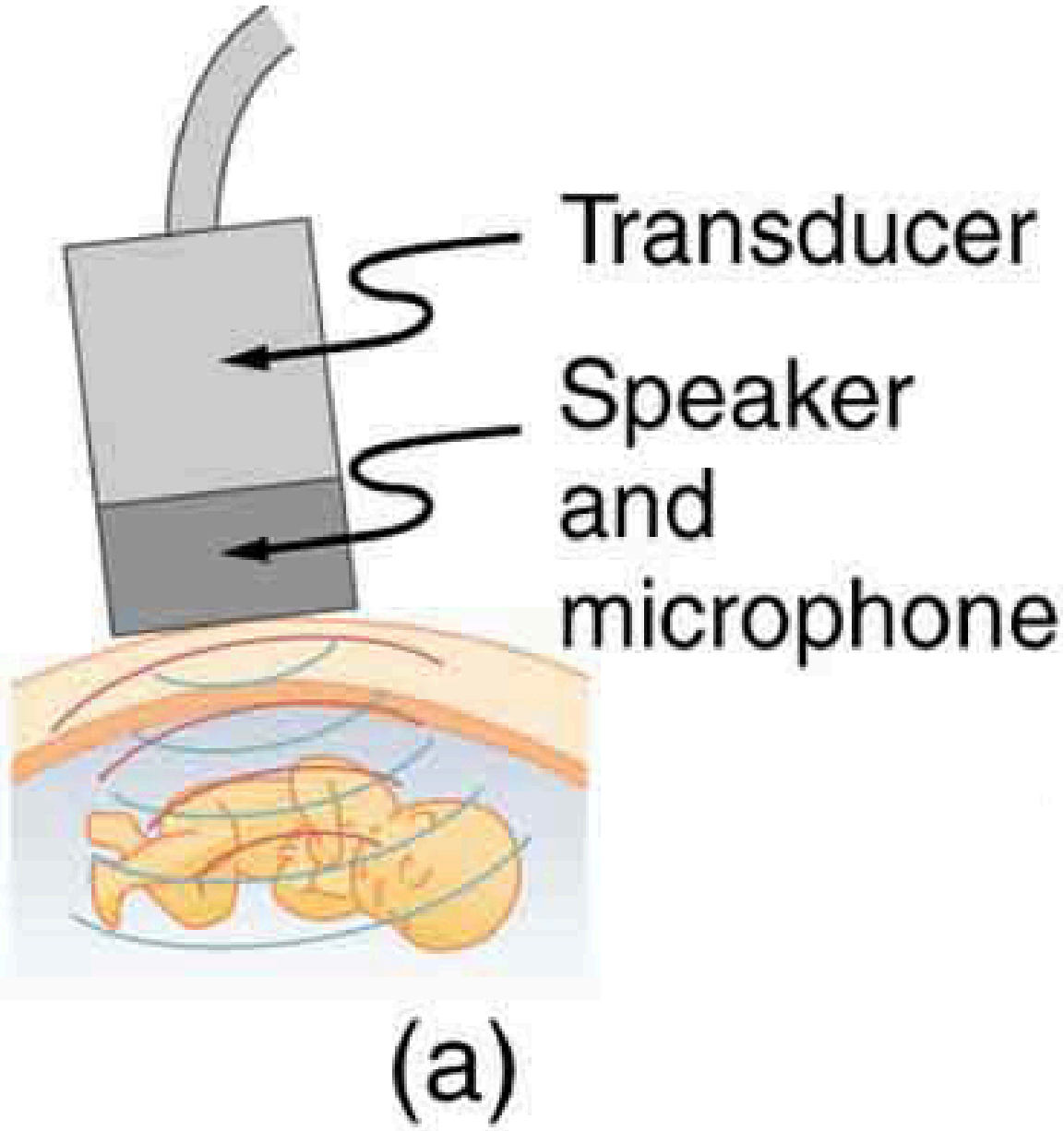
Medium	**Density (kg/m^3)**	**Speed of Ultrasound (m/s)**	**Acoustic Impedance (kg/(m$^2 \cdot$ s))**
Air	1.3	330	429
Water	1000	1500	1.5×10^6
Blood	1060	1570	1.66×10^6
Fat	925	1450	1.34×10^6
Muscle (average)	1075	1590	1.70×10^6
Bone (varies)	1400–1900	4080	5.7×10^6 to 7.8×10^6
Barium titanate (transducer material)	5600	5500	30.8×10^6

At the boundary between media of different acoustic impedances, some of the wave energy is reflected and some is transmitted. The greater the *difference* in acoustic impedance between the two media, the greater the reflection and the smaller the transmission.

The *intensity reflection coefficient a* is defined as the ratio of the intensity of the reflected wave relative to the incident (transmitted) wave. This statement can be written mathematically as

$$a = \frac{(Z_2 - Z_1)^2}{(Z_1 + Z_2)^2}$$

, where Z_1 and Z_2 are the acoustic impedances of the two media making up the boundary. A reflection coefficient of zero (corresponding to total transmission and no reflection) occurs when the acoustic impedances of the two media are the same. An impedance “match” (no reflection) provides an efficient coupling of sound energy from one medium to another. The image formed in an ultrasound is made by tracking reflections (as shown in Figure 3) and mapping the intensity of the reflected sound waves in a two-dimensional plane.



(a)

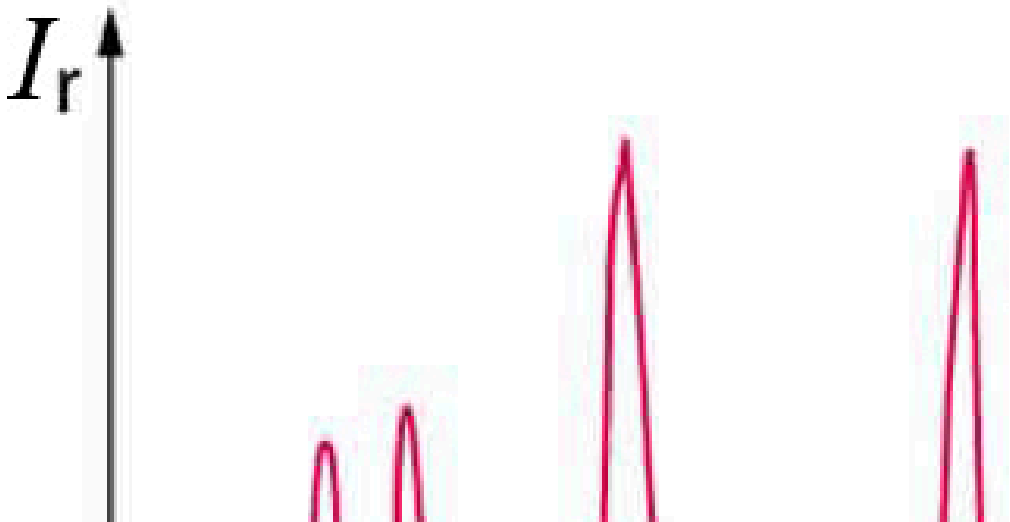


Figure 3. (a) An ultrasound speaker doubles as a microphone. Brief bleeps are broadcast, and echoes are recorded from various depths. (b) Graph of echo intensity versus time. The time for echoes to return is directly proportional to the distance of the reflector, yielding this information noninvasively.

Example 1. Calculate Acoustic Impedance and Intensity Reflection Coefficient: Ultrasound and Fat Tissue

1. Using the values for density and the speed of ultrasound given in Table 1, show that the acoustic impedance of fat tissue is indeed 1.34×10^6 kg/(m^2 · s).
2. Calculate the intensity reflection coefficient of ultrasound when going from fat to muscle tissue.

Strategy for Part 1

The acoustic impedance can be calculated using $Z = \rho v$ and the values for ρ and v found in Table 1.

Solution for Part 1

Substitute known values from Table 1 into $Z = \rho v$: $Z = \rho v$ = (925 kg/m^3)(1450 m/s)

Calculate to find the acoustic impedance of fat tissue: 1.34×10^6 kg/(m^2 · s)

This value is the same as the value given for the acoustic impedance of fat tissue.

Strategy for Part 2

The intensity reflection coefficient for any boundary between two media is given by

$$a = \frac{(Z_2 - Z_1)^2}{(Z_1 + Z_2)^2}$$

, and the acoustic impedance of muscle is given in Table 1.

Solution for Part 2

Substitute known values into

$$a = \frac{(Z_2 - Z_1)^2}{(Z_1 + Z_2)^2}$$

to find the intensity reflection coefficient:

$$a = \frac{(Z_2 - Z_1)^2}{(Z_1 + Z_2)^2} = \frac{\left(1.34 \times 10^6 \text{ kg/(m}^2 \cdot \text{s)} - 1.70 \times 10^6 \text{ kg/(m}^2 \cdot \text{s)}\right)^2}{\left(1.70 \times 10^6 \text{ kg/(m}^2 \cdot \text{s)} + 1.34 \times 10^6 \text{ kg/(m}^2 \cdot \text{s)}\right)^2} =$$

$$0.014$$

Discussion

This result means that only 1.4% of the incident intensity is reflected, with the remaining being transmitted.

The applications of ultrasound in medical diagnostics have produced untold benefits with no known risks. Diagnostic intensities are too low (about 10^{-2} W/m^2) to cause thermal damage. More significantly,

ultrasound has been in use for several decades and detailed follow-up studies do not show evidence of ill effects, quite unlike the case for x-rays.

The most common ultrasound applications produce an image like that shown in Figure 4. The speaker-microphone broadcasts a directional beam, sweeping the beam across the area of interest. This is accomplished by having multiple ultrasound sources in the probe's head, which are phased to interfere constructively in a given, adjustable direction. Echoes are measured as a function of position as well as depth. A computer constructs an image that reveals the shape and density of internal structures.

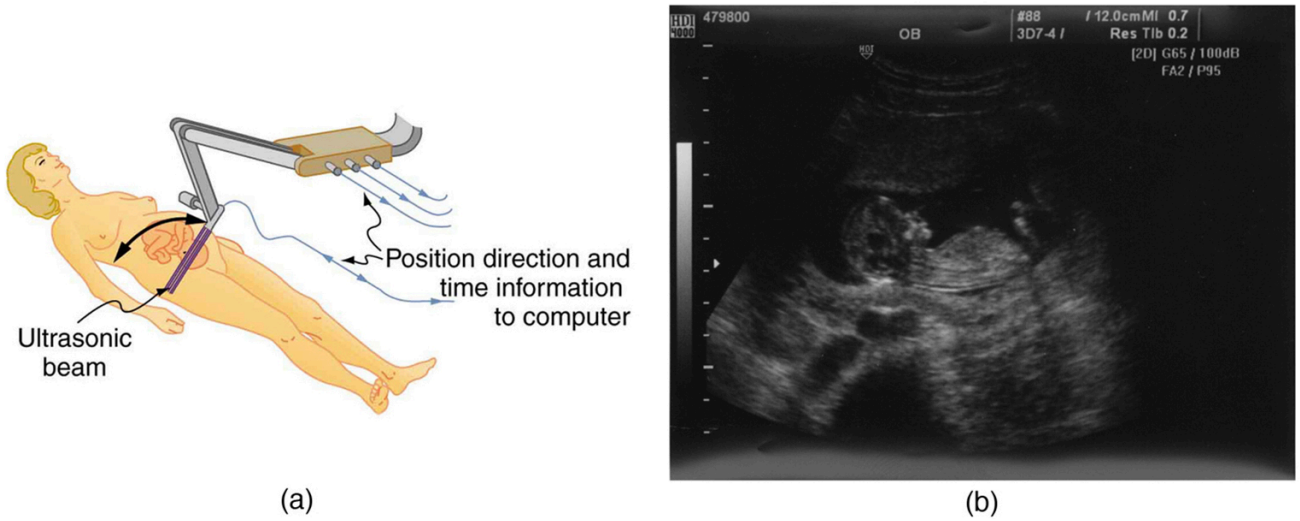


Figure 4. (a) An ultrasonic image is produced by sweeping the ultrasonic beam across the area of interest, in this case the woman's abdomen. Data are recorded and analyzed in a computer, providing a two-dimensional image. (b) Ultrasound image of 12-week-old fetus. (credit: Margaret W. Carruthers, Flickr)

How much detail can ultrasound reveal? The image in Figure 4 is typical of low-cost systems, but that in Figure 5 shows the remarkable detail possible with more advanced systems, including 3D imaging. Ultrasound today is commonly used in prenatal care. Such imaging can be used to see if the fetus is developing at a normal rate, and help in the determination of serious problems early in the pregnancy. Ultrasound is also in wide use to image the chambers of the heart and the flow of blood within the beating heart, using the Doppler effect (echocardiology).

Whenever a wave is used as a probe, it is very difficult to detect details smaller than its wavelength λ. Indeed, current technology cannot do quite this well. Abdominal scans may use a 7-MHz frequency, and the speed of sound in tissue is about 1540 m/s—so the wavelength limit to detail would be

$$\lambda = \frac{v_w}{f} = \frac{1540 \text{ m/s}}{7 \times 10^6 \text{ Hz}} = 0.22 \text{ mm}$$

. In practice, 1-mm detail is attainable, which is sufficient for many purposes. Higher-frequency ultrasound would allow greater detail, but it does not penetrate as well as lower frequencies do. The accepted rule of thumb is that you can effectively scan to a depth of about 500λ into tissue. For 7 MHz, this penetration limit is 500 × 0.22 mm, which is 0.11 m. Higher frequencies may be employed in smaller organs, such as the eye, but are not practical for looking deep into the body.

Figure 5. A 3D ultrasound image of a fetus. As well as for the detection of any abnormalities, such scans have also been shown to be useful for strengthening the emotional bonding between parents and their unborn child. (credit: Jennie Cu, Wikimedia Commons)

In addition to shape information, ultrasonic scans can produce density information superior to that found in X-rays, because the intensity of a reflected sound is related to changes in density. Sound is most strongly reflected at places where density changes are greatest.

Another major use of ultrasound in medical diagnostics is to detect motion and determine velocity through the Doppler shift of an echo, known as *Doppler-shifted ultrasound*. This technique is used to monitor fetal heartbeat, measure blood velocity, and detect occlusions in blood vessels, for example. (See Figure 6.) The magnitude of the Doppler shift in an echo is directly proportional to the velocity of whatever reflects the sound. Because an echo is involved, there is actually a double shift. The first occurs because the reflector (say a fetal heart) is a moving observer and receives a Doppler-shifted frequency. The reflector then acts as a moving source, producing a second Doppler shift.

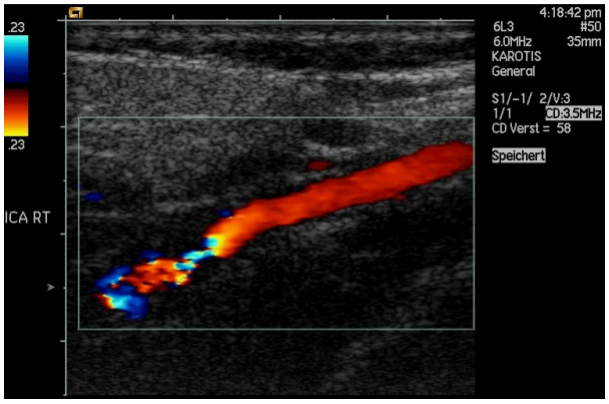


Figure 6. This Doppler-shifted ultrasonic image of a partially occluded artery uses color to indicate velocity. The highest velocities are in red, while the lowest are blue. The blood must move faster through the constriction to carry the same flow. (credit: Arning C, Grzyska U, Wikimedia Commons)

A clever technique is used to measure the Doppler shift in an echo. The frequency of the echoed sound is superimposed on the broadcast frequency, producing beats. The beat frequency is $F_B = |f_1 - f_2|$, and so it is directly proportional to the Doppler shift ($f_1 - f_2$) and hence, the reflector's velocity. The advantage in this technique is that the Doppler shift is small (because the reflector's velocity is small), so that great accuracy would be needed to measure the shift directly. But measuring the beat frequency is easy, and it is not affected if the broadcast frequency varies somewhat. Furthermore, the beat frequency is in the audible range and can be amplified for audio feedback to the medical observer.

Uses for Doppler-Shifted Radar

Doppler-shifted radar echoes are used to measure wind velocities in storms as well as aircraft and automobile speeds. The principle is the same as for Doppler-shifted ultrasound. There is evidence that bats and dolphins may also sense the velocity of an object (such as prey) reflecting their ultrasound signals by observing its Doppler shift.

Example 2. Calculate Velocity of Blood: Doppler-Shifted Ultrasound

Ultrasound that has a frequency of 2.50 MHz is sent toward blood in an artery that is moving toward the source at 20.0 cm/s, as illustrated in Figure 7. Use the speed of sound in human tissue as 1540 m/s. (Assume that the frequency of 2.50 MHz is accurate to seven significant figures.)

1. What frequency does the blood receive?
2. What frequency returns to the source?
3. What beat frequency is produced if the source and returning frequencies are mixed?

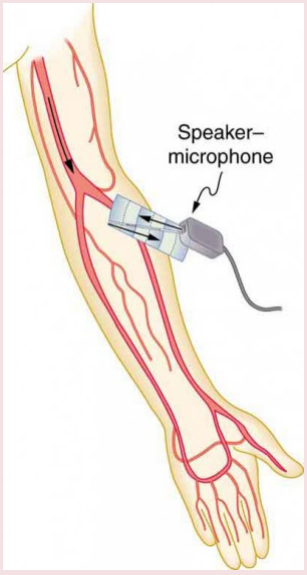


Figure 7. Ultrasound is partly reflected by blood cells and plasma back toward the speaker-microphone. Because the cells are moving, two Doppler shifts are produced—one for blood as a moving observer, and the other for the reflected sound coming from a moving source. The magnitude of the shift is directly proportional to blood velocity.

Strategy

The first two questions can be answered using

$$f_{\text{obs}} = f_{\text{s}}\left(\frac{v_{\text{w}}}{v_{\text{w}} \pm v_{\text{s}}}\right)$$

and

$$f_{\text{obs}} = f_{\text{s}}\left(\frac{v_{\text{w}} \pm v_{\text{obs}}}{v_{\text{w}}}\right)$$

for the Doppler shift. The last question asks for beat frequency, which is the difference between the original and returning frequencies.

Solution for Part 1

Identify knowns:

- The blood is a moving observer, and so the frequency it receives is given by

$$f_{\text{obs}} = f_{\text{s}}\left(\frac{v_{\text{w}} \pm v_{\text{obs}}}{v_{\text{w}}}\right)$$

.

- v_{b} is the blood velocity (v_{obs} here) and the plus sign is chosen because the motion is toward the

source.

Enter the given values into the equation.

$$f_{\text{obs}} = (2,500,000\ \text{Hz}) \left(\frac{1540\ \text{m/s} + 0.2\ \text{m/s}}{1540\ \text{m/s}} \right)$$

Calculate to find the frequency: 20,500,325 Hz.

Solution for Part 2

Identify knowns:

- The blood acts as a moving source.
- The microphone acts as a stationary observer.
  - The frequency leaving the blood is 2,500,325 Hz, but it is shifted upward as given by $f_{\text{obs}} = f_{\text{s}} \left(\frac{v_{\text{w}}}{v_{\text{w}} - v_{\text{b}}} \right)$. f_{obs} is the frequency received by the speaker-microphone.
- The source velocity is v_{b}.
- The minus sign is used because the motion is toward the observer.

The minus sign is used because the motion is toward the observer.

Enter the given values into the equation:

$$f_{\text{obs}} = (2,500,325\ \text{Hz}) \left(\frac{1540\ \text{m/s}}{1540\ \text{m/s} - 0.200\ \text{m/s}} \right)$$

Calculate to find the frequency returning to the source: 2,500,649 Hz.

Solution for Part 3

Identify knowns. The beat frequency is simply the absolute value of the difference between f_{s} and f_{obs}, as stated in:

$$f_{\text{B}} = |f_{\text{obs}} - f_{\text{s}}|.$$

Substitute known values:

$$|2{,}500{,}649\ \text{Hz} - 2{,}500{,}000\ \text{Hz}|$$

Calculate to find the beat frequency: 649 Hz.

Discussion

The Doppler shifts are quite small compared with the original frequency of 2.50 MHz. It is far easier to measure the beat frequency than it is to measure the echo frequency with an accuracy great enough to see shifts of a few hundred hertz out of a couple of megahertz. Furthermore, variations in the source frequency do not greatly affect the beat frequency, because both f_{s} and f_{obs}would increase or decrease. Those changes subtract out in $f_{\text{B}} = |f_{\text{obs}} - f_{\text{s}}|$.

Industrial and Other Applications of Ultrasound

Industrial, retail, and research applications of ultrasound are common. A few are discussed here. Ultrasonic cleaners have many uses. Jewelry, machined parts, and other objects that have odd shapes and crevices are immersed in a cleaning fluid that is agitated with ultrasound typically about 40 kHz in frequency. The intensity is great enough to cause cavitation, which is responsible for most of the cleansing action. Because cavitation-produced shock pressures are large and well transmitted in a fluid, they reach into small crevices where even a low-surface-tension cleaning fluid might not penetrate.

Sonar is a familiar application of ultrasound. Sonar typically employs ultrasonic frequencies in the range from 30.0 to 100 kHz. Bats, dolphins, submarines, and even some birds use ultrasonic sonar. Echoes are analyzed to give distance and size information both for guidance and finding prey. In most sonar applications, the sound reflects quite well because the objects of interest have significantly different density than the medium in which they travel. When the Doppler shift is observed, velocity information can also be obtained. Submarine sonar can be used to obtain such information, and there is evidence that some bats also sense velocity from their echoes.

Similarly, there are a range of relatively inexpensive devices that measure distance by timing ultrasonic echoes. Many cameras, for example, use such information to focus automatically. Some doors open when their ultrasonic ranging devices detect a nearby object, and certain home security lights turn on when their ultrasonic rangers observe motion. Ultrasonic “measuring tapes” also exist to measure such things as room dimensions. Sinks in public restrooms are sometimes automated with ultrasound devices to turn faucets on and off when people wash their hands. These devices reduce the spread of germs and can conserve water.

Ultrasound is used for nondestructive testing in industry and by the military. Because ultrasound reflects well from any large change in density, it can reveal cracks and voids in solids, such as aircraft wings, that are too small to be seen with x-rays. For similar reasons, ultrasound is also good for measuring the thickness of coatings, particularly where there are several layers involved.

Basic research in solid state physics employs ultrasound. Its attenuation is related to a number of physical characteristics, making it a useful probe. Among these characteristics are structural changes such as those found in liquid crystals, the transition of a material to a superconducting phase, as well as density and other properties.

These examples of the uses of ultrasound are meant to whet the appetites of the curious, as well as to illustrate the underlying physics of ultrasound. There are many more applications, as you can easily discover for yourself.

Check Your Understanding

Why is it possible to use ultrasound both to observe a fetus in the womb and also to destroy cancerous tumors in the body?

Solution

Ultrasound can be used medically at different intensities. Lower intensities do not cause damage and are used for medical imaging. Higher intensities can pulverize and destroy targeted substances in the body, such as tumors.

Section Summary

- The acoustic impedance is defined as $Z = \rho v$, ρ is the density of a medium through which the sound travels and v is the speed of sound through that medium.
- The intensity reflection coefficient a, a measure of the ratio of the intensity of the wave reflected off a boundary between two media relative to the intensity of the incident wave, is given by

$$a = \frac{(Z_2 - Z_1)^2}{(Z_1 + Z_2)^2}$$

.

- The intensity reflection coefficient is a unitless quantity.

Conceptual Questions

1. If audible sound follows a rule of thumb similar to that for ultrasound, in terms of its absorption, would you expect the high or low frequencies from your neighbor's stereo to penetrate into your house? How does this expectation compare with your experience?
2. Elephants and whales are known to use infrasound to communicate over very large distances. What are the advantages of infrasound for long distance communication?
3. It is more difficult to obtain a high-resolution ultrasound image in the abdominal region of someone who is overweight than for someone who has a slight build. Explain why this statement is accurate.
4. Suppose you read that 210-dB ultrasound is being used to pulverize cancerous tumors. You calculate the intensity in watts per centimeter squared and find it is unreasonably high (10^5 W/cm^2). What is a possible explanation?

Problems & Exercises

Unless otherwise indicated, for problems in this section, assume that the speed of sound through human tissues is 1540 m/s.

1. What is the sound intensity level in decibels of ultrasound of intensity 10^5 W/m^2, used to pulverize tissue during surgery?
2. Is 155-dB ultrasound in the range of intensities used for deep heating? Calculate the intensity of this ultrasound and compare this intensity with values quoted in the text.
3. Find the sound intensity level in decibels of 2.00×10^{-2} W/m^2 ultrasound used in medical diagnostics.
4. The time delay between transmission and the arrival of the reflected wave of a signal using ultrasound traveling through a piece of fat tissue was 0.13 ms. At what depth did this reflection

occur?

5. In the clinical use of ultrasound, transducers are always coupled to the skin by a thin layer of gel or oil, replacing the air that would otherwise exist between the transducer and the skin. (a) Using the values of acoustic impedance given in Table 1 calculate the intensity reflection coefficient between transducer material and air. (b) Calculate the intensity reflection coefficient between transducer material and gel (assuming for this problem that its acoustic impedance is identical to that of water). (c) Based on the results of your calculations, explain why the gel is used.
6. (a) Calculate the minimum frequency of ultrasound that will allow you to see details as small as 0.250 mm in human tissue. (b) What is the effective depth to which this sound is effective as a diagnostic probe?
7. (a) Find the size of the smallest detail observable in human tissue with 20.0-MHz ultrasound. (b) Is its effective penetration depth great enough to examine the entire eye (about 3.00 cm is needed)? (c) What is the wavelength of such ultrasound in 0ºC air?
8. (a) Echo times are measured by diagnostic ultrasound scanners to determine distances to reflecting surfaces in a patient. What is the difference in echo times for tissues that are 3.50 and 3.60 cm beneath the surface? (This difference is the minimum resolving time for the scanner to see details as small as 0.100 cm, or 1.00 mm. Discrimination of smaller time differences is needed to see smaller details.) (b) Discuss whether the period *T* of this ultrasound must be smaller than the minimum time resolution. If so, what is the minimum frequency of the ultrasound and is that out of the normal range for diagnostic ultrasound?
9. (a) How far apart are two layers of tissue that produce echoes having round-trip times (used to measure distances) that differ by 0.750 μs? (b) What minimum frequency must the ultrasound have to see detail this small?
10. (a) A bat uses ultrasound to find its way among trees. If this bat can detect echoes 1.00 ms apart, what minimum distance between objects can it detect? (b) Could this distance explain the difficulty that bats have finding an open door when they accidentally get into a house?
11. A dolphin is able to tell in the dark that the ultrasound echoes received from two sharks come from two different objects only if the sharks are separated by 3.50 m, one being that much farther away than the other. (a) If the ultrasound has a frequency of 100 kHz, show this ability is not limited by its wavelength. (b) If this ability is due to the dolphin's ability to detect the arrival times of echoes, what is the minimum time difference the dolphin can perceive?
12. A diagnostic ultrasound echo is reflected from moving blood and returns with a frequency 500 Hz higher than its original 2.00 MHz. What is the velocity of the blood? (Assume that the frequency of 2.00 MHz is accurate to seven significant figures and 500 Hz is accurate to three significant figures.)
13. Ultrasound reflected from an oncoming bloodstream that is moving at 30.0 cm/s is mixed with the original frequency of 2.50 MHz to produce beats. What is the beat frequency? (Assume that the frequency of 2.50 MHz is accurate to seven significant figures.)

Glossary

acoustic impedance: property of medium that makes the propagation of sound waves more difficult

intensity reflection coefficient: a measure of the ratio of the intensity of the wave reflected off a boundary between two media relative to the intensity of the incident wave

Doppler-shifted ultrasound: a medical technique to detect motion and determine velocity through the Doppler shift of an echo

Selected Solutions to Problems & Exercises

1. 170 dB

3. 103 dB

5. (a) 1.00; (b) 0.823; (c) Gel is used to facilitate the transmission of the ultrasound between the transducer and the patient's body.

7. (a) 77.0 μm; (b) Effective penetration depth = 3.85 cm, which is enough to examine the eye; (c) 16.6 μm

9. (a) 5.78×10^{-4} m; (b) 2.67×10^{6} Hz

11. (a)

$$v_{\text{w}} = 1540 \text{ m/s} = f\lambda \Rightarrow \lambda = \frac{1540 \text{ m/s}}{100 \times 10^{3} \text{Hz}} = 0.0154 \text{ m} < 3.50 \text{ m}$$

. Because the wavelength is much shorter than the distance in question, the wavelength is not the limiting factor; (b) 4.55 ms

13. 974 Hz (Note: extra digits were retained in order to show the difference.)

8. Electromagnetic Waves

Introduction to Electromagnetic Waves

Lumen Learning

Figure 1. Human eyes detect these orange "sea goldie" fish swimming over a coral reef in the blue waters of the Gulf of Eilat (Red Sea) using visible light. (credit: Daviddarom, Wikimedia Commons)

The beauty of a coral reef, the warm radiance of sunshine, the sting of sunburn, the X-ray revealing a broken bone, even microwave popcorn—all are brought to us by *electromagnetic waves*. The list of the various types of electromagnetic waves, ranging from radio transmission waves to nuclear gamma-ray (γ-ray) emissions, is interesting in itself.

Even more intriguing is that all of these widely varied phenomena are different manifestations of the same thing—electromagnetic waves. (See Figure 2.) What are electromagnetic waves? How are they created, and how do they travel? How can we understand and organize their widely varying properties? What is their relationship to electric and magnetic effects? These and other questions will be explored.

Misconception Alert: Sound Waves vs. Radio Waves

Many people confuse sound waves with *radio waves*, one type of electromagnetic (EM) wave. However, sound and radio waves are completely different phenomena. Sound creates pressure variations (waves) in matter, such as air or water, or your eardrum. Conversely, radio waves are *electromagnetic waves*, like visible light, infrared, ultraviolet, X-rays, and gamma rays. EM waves don't need a medium in which to propagate; they can travel through a vacuum, such as outer space.

A radio works because sound waves played by the D.J. at the radio station are converted into electromagnetic waves, then encoded and transmitted in the radio-frequency range. The radio in your car receives the radio waves, decodes the information, and uses a speaker to change it back into a sound wave, bringing sweet music to your ears.

Discovering a New Phenomenon

It is worth noting at the outset that the general phenomenon of electromagnetic waves was predicted by theory before it was realized that light is a form of electromagnetic wave. The prediction was made by James Clerk Maxwell in the mid-19th century when he formulated a single theory combining all the electric and magnetic effects known by scientists at that time. "Electromagnetic waves" was the name he gave to the phenomena his theory predicted.

Such a theoretical prediction followed by experimental verification is an indication of the power of science in general, and physics in particular. The underlying connections and unity of physics allow certain great minds to solve puzzles without having all the pieces. The prediction of electromagnetic waves is one of the most spectacular examples of this power. Certain others, such as the prediction of antimatter, will be discussed in later modules.

Figure 2. The electromagnetic waves sent and received by this 50-foot radar dish antenna at Kennedy Space Center in Florida are not visible, but help track expendable launch vehicles with high-definition imagery. The first use of this C-band radar dish was for the launch of the Atlas V rocket sending the New Horizons probe toward Pluto. (credit: NASA)

Maxwell's Equations: Electromagnetic Waves Predicted and Observed

Lumen Learning

Learning Objective

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

- Restate Maxwell's equations.

The Scotsman James Clerk Maxwell (1831–1879) is regarded as the greatest theoretical physicist of the 19th century. (See Figure 1.) Although he died young, Maxwell not only formulated a complete electromagnetic theory, represented by *Maxwell's equations*, he also developed the kinetic theory of gases and made significant contributions to the understanding of color vision and the nature of Saturn's rings.

Figure 1. James Clerk Maxwell, a 19th-century physicist, developed a theory that explained the relationship between electricity and magnetism and correctly predicted that visible light is caused by electromagnetic waves. (credit: G. J. Stodart)

Maxwell brought together all the work that had been done by brilliant physicists such as Oersted, Coulomb, Gauss, and Faraday, and added his own insights to develop the overarching theory of electromagnetism. Maxwell's equations are paraphrased here in words because their mathematical statement is beyond the level of this text. However, the equations illustrate how apparently simple mathematical statements can elegantly unite and express a multitude of concepts—why mathematics is the language of science.

Maxwell's Equations

1. *Electric field lines* originate on positive charges and terminate on negative charges. The electric field is defined as the force per unit charge on a test charge, and the strength of the force is related to the electric constant ε_0, also known as the permittivity of free space. From Maxwell's first equation we obtain a special form of Coulomb's law known as Gauss's law for electricity.
2. *Magnetic field lines* are continuous, having no beginning or end. No magnetic monopoles are known to exist. The strength of the magnetic force is related to the magnetic constant

μ_0, also known as the permeability of free space. This second of Maxwell's equations is known as Gauss's law for magnetism.

3. A changing magnetic field induces an electromotive force (emf) and, hence, an electric field. The direction of the emf opposes the change. This third of Maxwell's equations is Faraday's law of induction, and includes Lenz's law.
4. Magnetic fields are generated by moving charges or by changing electric fields. This fourth of Maxwell's equations encompasses Ampere's law and adds another source of magnetism—changing electric fields.

Maxwell's equations encompass the major laws of electricity and magnetism. What is not so apparent is the symmetry that Maxwell introduced in his mathematical framework. Especially important is his addition of the hypothesis that changing electric fields create magnetic fields. This is exactly analogous (and symmetric) to Faraday's law of induction and had been suspected for some time, but fits beautifully into Maxwell's equations.

Symmetry is apparent in nature in a wide range of situations. In contemporary research, symmetry plays a major part in the search for sub-atomic particles using massive multinational particle accelerators such as the new Large Hadron Collider at CERN.

Making Connections: Unification of Forces

Maxwell's complete and symmetric theory showed that electric and magnetic forces are not separate, but different manifestations of the same thing—the electromagnetic force. This classical unification of forces is one motivation for current attempts to unify the four basic forces in nature—the gravitational, electrical, strong, and weak nuclear forces.

Since changing electric fields create relatively weak magnetic fields, they could not be easily detected at the time of Maxwell's hypothesis. Maxwell realized, however, that oscillating charges, like those in AC circuits, produce changing electric fields. He predicted that these changing fields would propagate from the source like waves generated on a lake by a jumping fish.

The waves predicted by Maxwell would consist of oscillating electric and magnetic fields—defined to be an electromagnetic wave (EM wave). Electromagnetic waves would be capable of exerting forces on charges great distances from their source, and they might thus be detectable. Maxwell calculated that electromagnetic waves would propagate at a speed given by the equation

$$c = \frac{1}{\sqrt{\mu_0 \epsilon_0}}$$

When the values for μ_0 and ε_0 are entered into the equation for c , we find that

$$c = \frac{1}{\sqrt{\left(8.85 \times 10^{-12} \frac{\mathrm{C}^2}{\mathrm{N \cdot m}^2}\right)\left(4\pi \times 10^{-7} \frac{\mathrm{T \cdot m}}{\mathrm{A}}\right)}} = 300 \times 10^8 \text{ m/s}$$

which is the speed of light. In fact, Maxwell concluded that light is an electromagnetic wave having such wavelengths that it can be detected by the eye.

Other wavelengths should exist—it remained to be seen if they did. If so, Maxwell's theory and remarkable predictions would be verified, the greatest triumph of physics since Newton. Experimental verification came within a few years, but not before Maxwell's death.

Hertz's Observations

The German physicist Heinrich Hertz (1857–1894) was the first to generate and detect certain types of electromagnetic waves in the laboratory. Starting in 1887, he performed a series of experiments that not only confirmed the existence of electromagnetic waves, but also verified that they travel at the speed of light.

Hertz used an AC *RLC* (resistor-inductor-capacitor) circuit that resonates at a known frequency

$$f_0 = \frac{1}{2\pi\sqrt{LC}}$$

and connected it to a loop of wire as shown in Figure 2. High voltages induced across the gap in the loop produced sparks that were visible evidence of the current in the circuit and that helped generate electromagnetic waves.

Across the laboratory, Hertz had another loop attached to another *RLC* circuit, which could be tuned (as the dial on a radio) to the same resonant frequency as the first and could, thus, be made to receive electromagnetic waves. This loop also had a gap across which sparks were generated, giving solid evidence that electromagnetic waves had been received.

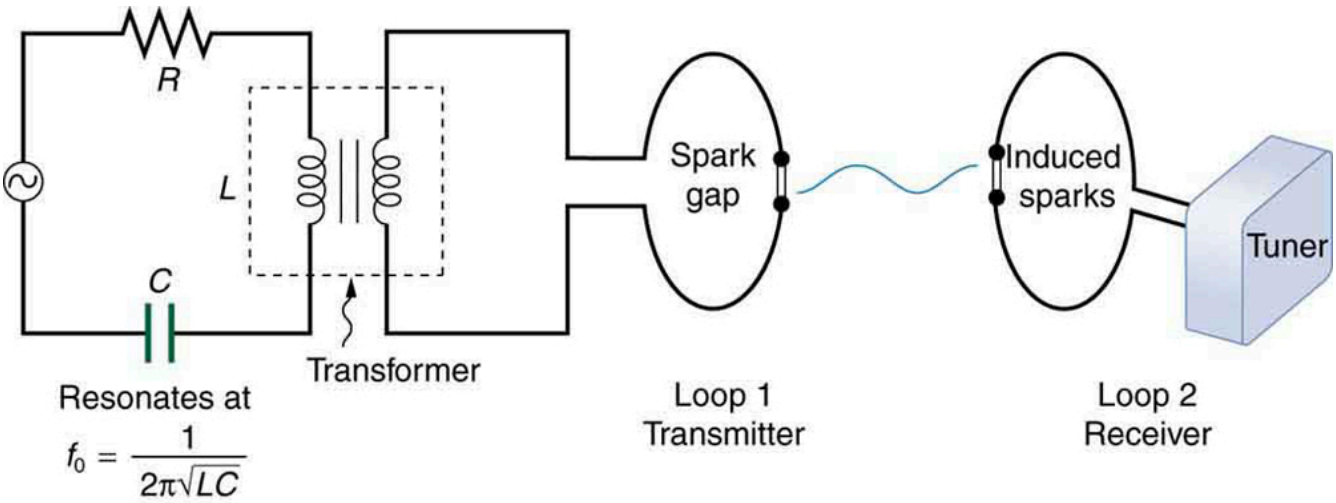


Figure 2. The apparatus used by Hertz in 1887 to generate and detect electromagnetic waves. An RLC circuit connected to the first loop caused sparks across a gap in the wire loop and generated electromagnetic waves. Sparks across a gap in the second loop located across the laboratory gave evidence that the waves had been received.

Hertz also studied the reflection, refraction, and interference patterns of the electromagnetic waves he generated, verifying their wave character. He was able to determine wavelength from the interference patterns, and knowing their frequency, he could calculate the propagation speed using the equation $v = f\lambda$ (velocity—or speed—equals frequency times wavelength). Hertz was thus able to prove that electromagnetic waves travel at the speed of light. The SI unit for frequency, the hertz (1 Hz = 1 cycle/sec), is named in his honor.

Section Summary

- Electromagnetic waves consist of oscillating electric and magnetic fields and propagate at the speed of light c. They were predicted by Maxwell, who also showed that
  $$c = \frac{1}{\sqrt{\mu_0 \epsilon_0}},$$
  where μ_0 is the permeability of free space and ε_0 is the permittivity of free space.
- Maxwell's prediction of electromagnetic waves resulted from his formulation of a complete and symmetric theory of electricity and magnetism, known as Maxwell's equations.
- These four equations are paraphrased in this text, rather than presented numerically, and encompass the major laws of electricity and magnetism. First is Gauss's law for electricity, second is Gauss's law for magnetism, third is Faraday's law of induction, including Lenz's law, and fourth is Ampere's law in a symmetric formulation that adds another source of magnetism—changing electric fields.

Problems & Exercises

1. Verify that the correct value for the speed of light c is obtained when numerical values for the

permeability and permittivity of free space (μ_0 and ε_0) are entered into the equation $c = \frac{1}{\sqrt{\mu_0 \epsilon_0}}$.

2. Show that, when SI units for μ_0 and ε_0 are entered, the units given by the right-hand side of the equation in the problem above are m/s.

Glossary

electromagnetic waves: radiation in the form of waves of electric and magnetic energy

Maxwell's equations: a set of four equations that comprise a complete, overarching theory of electromagnetism

RLC circuit: an electric circuit that includes a resistor, capacitor and inductor

hertz: an SI unit denoting the frequency of an electromagnetic wave, in cycles per second

speed of light: in a vacuum, such as space, the speed of light is a constant 3×10^8 m/s

electromotive force (emf): energy produced per unit charge, drawn from a source that produces an electrical current

electric field lines: a pattern of imaginary lines that extend between an electric source and charged objects in the surrounding area, with arrows pointed away from positively charged objects and toward negatively charged objects. The more lines in the pattern, the stronger the electric field in that region

magnetic field lines: a pattern of continuous, imaginary lines that emerge from and enter into opposite magnetic poles. The density of the lines indicates the magnitude of the magnetic field

Production of Electromagnetic Waves

Lumen Learning

Learning Objectives

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

- Describe the electric and magnetic waves as they move out from a source, such as an AC generator.
- Explain the mathematical relationship between the magnetic field strength and the electrical field strength.
- Calculate the maximum strength of the magnetic field in an electromagnetic wave, given the maximum electric field strength.

We can get a good understanding of *electromagnetic waves* (EM) by considering how they are produced. Whenever a current varies, associated electric and magnetic fields vary, moving out from the source like waves. Perhaps the easiest situation to visualize is a varying current in a long straight wire, produced by an AC generator at its center, as illustrated in Figure 1.

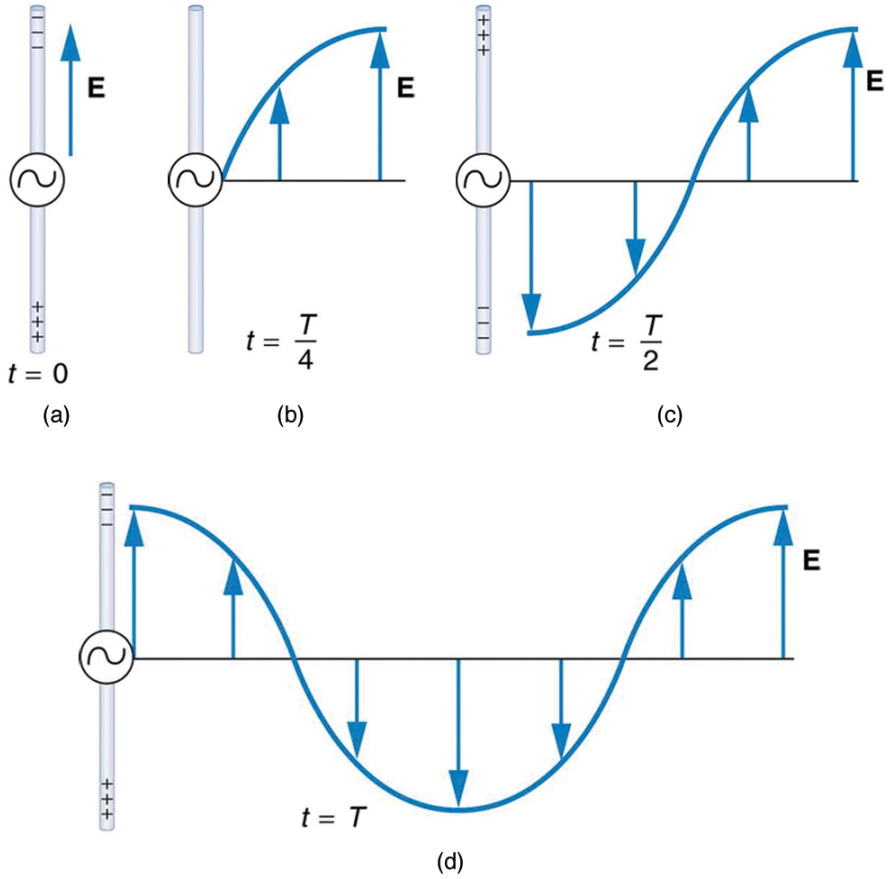


Figure 1. This long straight gray wire with an AC generator at its center becomes a broadcast antenna for electromagnetic waves. Shown here are the charge distributions at four different times. The electric field (E) propagates away from the antenna at the speed of light, forming part of an electromagnetic wave.

The *electric field* (**E**) shown surrounding the wire is produced by the charge distribution on the wire. Both the **E** and the charge distribution vary as the current changes. The changing field propagates outward at the speed of light.

There is an associated *magnetic field* (**B**) which propagates outward as well (see Figure 2). The electric and magnetic fields are closely related and propagate as an electromagnetic wave. This is what happens in broadcast antennae such as those in radio and TV stations.

Closer examination of the one complete cycle shown in Figure 1 reveals the periodic nature of the generator-driven charges oscillating up and down in the antenna and the electric field produced. At time t=0, there is the maximum separation of charge, with negative charges at the top and positive charges at the bottom, producing the maximum magnitude of the electric field (or E-field) in the upward direction. One-fourth of a cycle later, there is no charge separation and the field next to the antenna is zero, while the maximum E-field has moved away at speed c.

As the process continues, the charge separation reverses and the field reaches its maximum downward value, returns to zero, and rises to its maximum upward value at the end of one complete cycle. The outgoing wave has an *amplitude* proportional to the maximum separation of charge. Its *wavelength* (λ) is

proportional to the period of the oscillation and, hence, is smaller for short periods or high frequencies. (As usual, wavelength and *frequency*(*f*) are inversely proportional.)

Electric and Magnetic Waves: Moving Together

Following Ampere's law, current in the antenna produces a magnetic field, as shown in Figure 2. The relationship between **E** and **B** is shown at one instant in Figure 2a. As the current varies, the magnetic field varies in magnitude and direction.

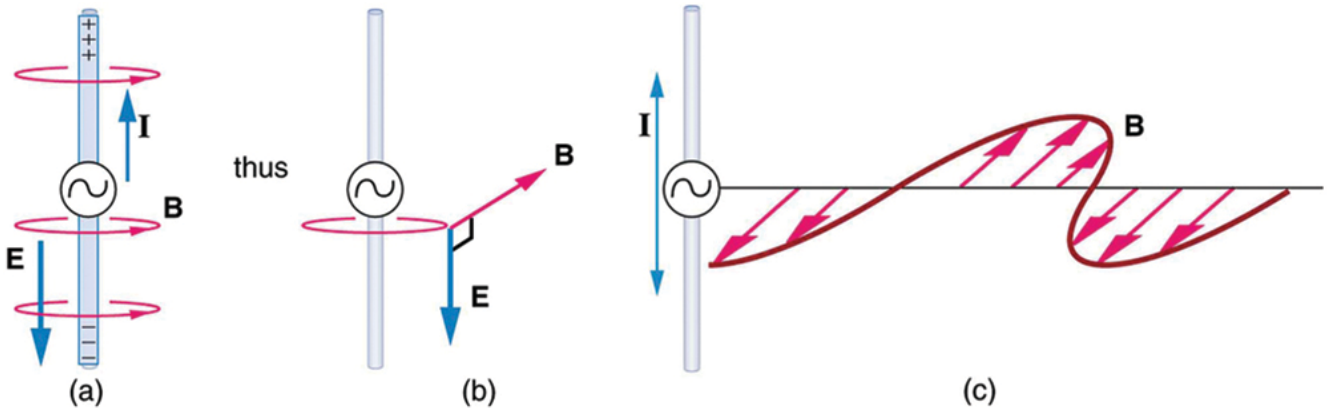


*Figure 2. (a) The current in the antenna produces the circular magnetic field lines. The current (I) produces the separation of charge along the wire, which in turn creates the electric field as shown. (b) The electric and magnetic fields (**E** and **B**) near the wire are perpendicular; they are shown here for one point in space. (c) The magnetic field varies with current and propagates away from the antenna at the speed of light.*

The magnetic field lines also propagate away from the antenna at the speed of light, forming the other part of the electromagnetic wave, as seen in Figure 2b. The magnetic part of the wave has the same period and wavelength as the electric part, since they are both produced by the same movement and separation of charges in the antenna.

The electric and magnetic waves are shown together at one instant in time in Figure 3. The electric and magnetic fields produced by a long straight wire antenna are exactly in phase. Note that they are perpendicular to one another and to the direction of propagation, making this a *transverse wave*

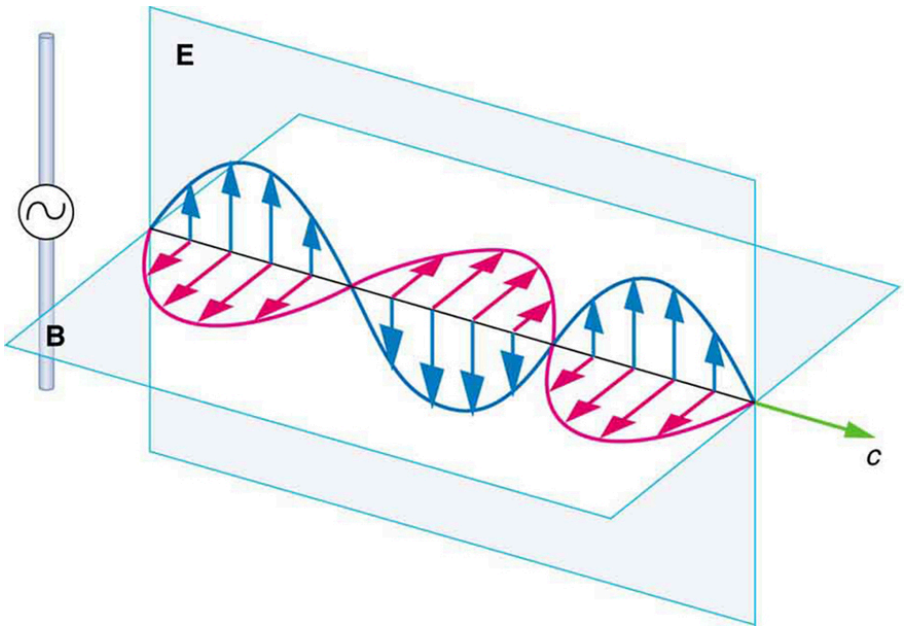


*Figure 3. A part of the electromagnetic wave sent out from the antenna at one instant in time. The electric and magnetic fields (**E** and **B**) are in phase, and they are perpendicular to one another and the direction of propagation. For clarity, the waves are shown only along one direction, but they propagate out in other directions too.*

Electromagnetic waves generally propagate out from a source in all directions, sometimes forming a complex radiation pattern. A linear antenna like this one will not radiate parallel to its length, for example. The wave is shown in one direction from the antenna in Figure 3 to illustrate its basic characteristics.

Instead of the AC generator, the antenna can also be driven by an AC circuit. In fact, charges radiate whenever they are accelerated. But while a current in a circuit needs a complete path, an antenna has a varying charge distribution forming a *standing wave*, driven by the AC. The dimensions of the antenna are critical for determining the frequency of the radiated electromagnetic waves. This is a *resonant* phenomenon and when we tune radios or TV, we vary electrical properties to achieve appropriate resonant conditions in the antenna.

Receiving Electromagnetic Waves

Electromagnetic waves carry energy away from their source, similar to a sound wave carrying energy away from a standing wave on a guitar string. An antenna for receiving EM signals works in reverse. And like antennas that produce EM waves, receiver antennas are specially designed to resonate at particular frequencies.

An incoming electromagnetic wave accelerates electrons in the antenna, setting up a standing wave. If the radio or TV is switched on, electrical components pick up and amplify the signal formed by the accelerating electrons. The signal is then converted to audio and/or video format. Sometimes big receiver dishes are used to focus the signal onto an antenna.

In fact, charges radiate whenever they are accelerated. When designing circuits, we often assume that

energy does not quickly escape AC circuits, and mostly this is true. A broadcast antenna is specially designed to enhance the rate of electromagnetic radiation, and shielding is necessary to keep the radiation close to zero. Some familiar phenomena are based on the production of electromagnetic waves by varying currents. Your microwave oven, for example, sends electromagnetic waves, called microwaves, from a concealed antenna that has an oscillating current imposed on it.

Relating *E*-Field and *B*-Field Strengths

There is a relationship between the *E*– and *B*-field strengths in an electromagnetic wave. This can be understood by again considering the antenna just described. The stronger the *E*-field created by a separation of charge, the greater the current and, hence, the greater the *B*-field created.

Since current is directly proportional to voltage (Ohm's law) and voltage is directly proportional to *E*-field strength, the two should be directly proportional. It can be shown that the magnitudes of the fields do have a constant ratio, equal to the speed of light. That is,

$$\frac{E}{B} = c$$

is the ratio of *E*-field strength to *B*-field strength in any electromagnetic wave. This is true at all times and at all locations in space. A simple and elegant result.

Example 1. Calculating *B*-Field Strength in an Electromagnetic Wave

What is the maximum strength of the *B*-field in an electromagnetic wave that has a maximum *E*-field strength of 1000 V/m?

Strategy

To find the *B*-field strength, we rearrange the above equation to solve for *B*, yielding

$$B = \frac{E}{c}$$

.

Solution

We are given *E*, and *c* is the speed of light. Entering these into the expression for *B* yields

$$B = \frac{1000 \text{ V/m}}{3.00 \times 10^8 \text{ m/s}} = 3.33 \times 10^{-6} \text{ T}$$

,

Where T stands for Tesla, a measure of magnetic field strength.

Discussion

The *B*-field strength is less than a tenth of the Earth's admittedly weak magnetic field. This means that a

relatively strong electric field of 1000 V/m is accompanied by a relatively weak magnetic field. Note that as this wave spreads out, say with distance from an antenna, its field strengths become progressively weaker.

The result of this example is consistent with the statement made in the module Maxwell's Equations: Electromagnetic Waves Predicted and Observed that changing electric fields create relatively weak magnetic fields. They can be detected in electromagnetic waves, however, by taking advantage of the phenomenon of resonance, as Hertz did. A system with the same natural frequency as the electromagnetic wave can be made to oscillate. All radio and TV receivers use this principle to pick up and then amplify weak electromagnetic waves, while rejecting all others not at their resonant frequency.

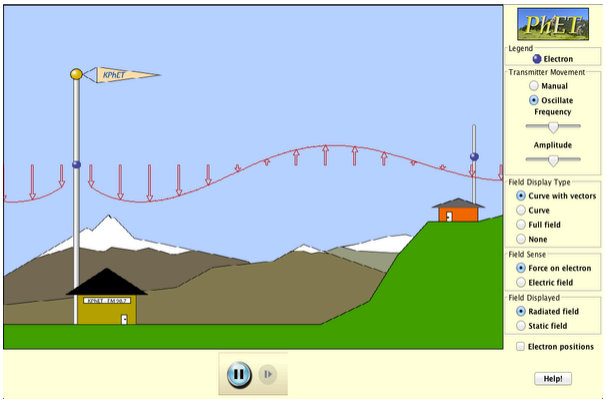
Take-Home Experiment: Antennas

For your TV or radio at home, identify the antenna, and sketch its shape. If you don't have cable, you might have an outdoor or indoor TV antenna. Estimate its size. If the TV signal is between 60 and 216 MHz for basic channels, then what is the wavelength of those EM waves?

Try tuning the radio and note the small range of frequencies at which a reasonable signal for that station is received. (This is easier with digital readout.) If you have a car with a radio and extendable antenna, note the quality of reception as the length of the antenna is changed.

PhET Explorations: Radio Waves and Electromagnetic Fields

Broadcast radio waves from KPhET. Wiggle the transmitter electron manually or have it oscillate automatically. Display the field as a curve or vectors. The strip chart shows the electron positions at the transmitter and at the receiver.



Click to download the simulation. Run using Java.

Section Summary

- Electromagnetic waves are created by oscillating charges (which radiate whenever accelerated) and have the same frequency as the oscillation.
- Since the electric and magnetic fields in most electromagnetic waves are perpendicular to the direction in which the wave moves, it is ordinarily a transverse wave.
- The strengths of the electric and magnetic parts of the wave are related by $\frac{E}{B} = \mathrm{c}$, which implies that the magnetic field *B* is very weak relative to the electric field *E*.

Conceptual Questions

1. The direction of the electric field shown in each part of Figure 1 is that produced by the charge distribution in the wire. Justify the direction shown in each part, using the Coulomb force law and the definition of $\mathbf{E} = \frac{\mathbf{F}}{q}$, where *q* is a positive test charge.
2. Is the direction of the magnetic field shown in Figure 2a consistent with the right-hand rule for current (RHR-2) in the direction shown in the figure?
3. Why is the direction of the current shown in each part of Figure 2 opposite to the electric field produced by the wire's charge separation?
4. In which situation shown in Figure 4 will the electromagnetic wave be more successful in inducing a current in the wire? Explain.

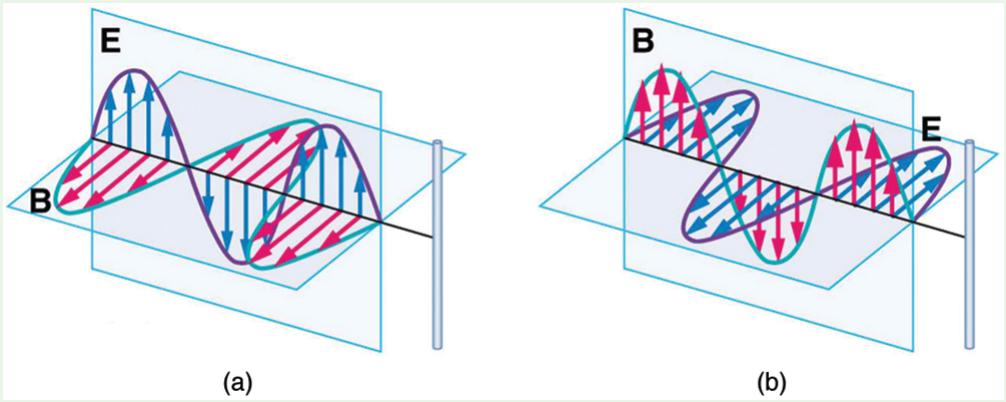


Figure 4. Electromagnetic waves approaching long straight wires.

5. In which situation shown in Figure 5 will the electromagnetic wave be more successful in inducing a current in the loop? Explain.

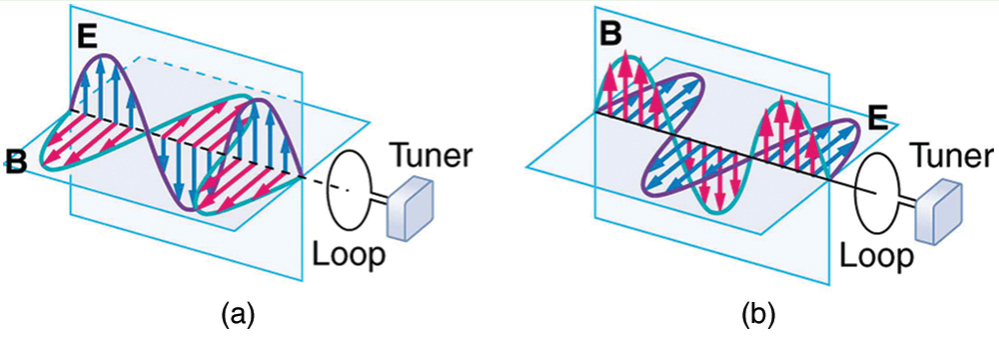


Figure 5. Electromagnetic waves approaching a wire loop.

6. Should the straight wire antenna of a radio be vertical or horizontal to best receive radio waves broadcast by a vertical transmitter antenna? How should a loop antenna be aligned to best receive the signals? (Note that the direction of the loop that produces the best reception can be used to determine the location of the source. It is used for that purpose in tracking tagged animals in nature studies, for example.)
7. Under what conditions might wires in a DC circuit emit electromagnetic waves?
8. Give an example of interference of electromagnetic waves.
9. Figure 6 shows the interference pattern of two radio antennas broadcasting the same signal. Explain how this is analogous to the interference pattern for sound produced by two speakers. Could this be used to make a directional antenna system that broadcasts preferentially in certain directions? Explain.

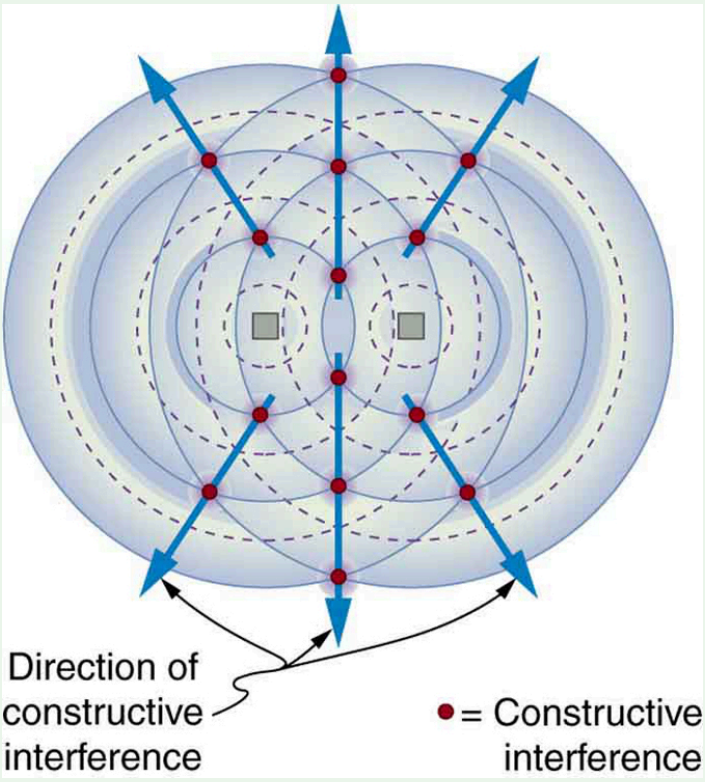


Figure 6. An overhead view of two radio broadcast antennas sending the same signal, and the interference pattern they produce.

10. Can an antenna be any length? Explain your answer.

Problems & Exercises

1. What is the maximum electric field strength in an electromagnetic wave that has a maximum magnetic field strength of 5.00×10^{-4} T (about 10 times the Earth's)?
2. The maximum magnetic field strength of an electromagnetic field is 5×10^{-6} T. Calculate the maximum electric field strength if the wave is traveling in a medium in which the speed of the wave is
   0.75 *c*.
3. Verify the units obtained for magnetic field strength *B* in Example 1 (using the equation $B = \frac{E}{c}$) are in fact teslas (T).

Glossary

electric field: a vector quantity (**E**); the lines of electric force per unit charge, moving radially outward from a positive charge and in toward a negative charge

electric field strength: the magnitude of the electric field, denoted E-field

magnetic field: a vector quantity (**B**); can be used to determine the magnetic force on a moving charged particle

magnetic field strength: the magnitude of the magnetic field, denoted B-field

transverse wave: a wave, such as an electromagnetic wave, which oscillates perpendicular to the axis along the line of travel

standing wave: a wave that oscillates in place, with nodes where no motion happens

wavelength: the distance from one peak to the next in a wave

amplitude: the height, or magnitude, of an electromagnetic wave

frequency: the number of complete wave cycles (up-down-up) passing a given point within one second (cycles/second)

resonant: a system that displays enhanced oscillation when subjected to a periodic disturbance of the same frequency as its natural frequency

oscillate: to fluctuate back and forth in a steady beat

Selected Solutions to Problems & Exercises

1. 150 kV/m

The Electromagnetic Spectrum

Lumen Learning

Learning Objectives

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

- List three “rules of thumb” that apply to the different frequencies along the electromagnetic spectrum.
- Explain why the higher the frequency, the shorter the wavelength of an electromagnetic wave.
- Draw a simplified electromagnetic spectrum, indicating the relative positions, frequencies, and spacing of the different types of radiation bands.
- List and explain the different methods by which electromagnetic waves are produced across the spectrum.

In this module we examine how electromagnetic waves are classified into categories such as radio, infrared, ultraviolet, and so on, so that we can understand some of their similarities as well as some of their differences. We will also find that there are many connections with previously discussed topics, such as wavelength and resonance. A brief overview of the production and utilization of electromagnetic waves is found in Table 1.

Table 1. Electromagnetic Waves

Type of EM wave	Production	Applications	Life sciences aspect	Issues
Radio & TV	Accelerating charges	Communications remote controls	MRI	Requires controls for band use
Microwaves	Accelerating charges & thermal agitation	Communications, ovens, radar	Deep heating	Cell phone use
Infrared	Thermal agitations & electronic transitions	Thermal imaging, heating	Absorbed by atmosphere	Greenhouse effect
Visible light	Thermal agitations & electronic transitions	All pervasive	Photosynthesis, Human vision	
Ultraviolet	Thermal agitations & electronic transitions	Sterilization, Cancer control	Vitamin D production	Ozone depletion, Cancer causing
X-rays	Inner electronic transitions and fast collisions	Medical Security	Medical diagnosis, Cancer therapy	Cancer causing
Gamma rays	Nuclear decay	Nuclear medicine, Security	Medical diagnosis, Cancer therapy	Cancer causing, Radiation damage

Connections: Waves

There are many types of waves, such as water waves and even earthquakes. Among the many shared attributes of waves are propagation speed, frequency, and wavelength. These are always related by the expression $v_W=f\lambda$. This module concentrates on EM waves, but other modules contain examples of all of these characteristics for sound waves and submicroscopic particles.

As noted before, an electromagnetic wave has a frequency and a wavelength associated with it and travels at the speed of light, or c. The relationship among these wave characteristics can be described by $v_W = f\lambda$, where v_W is the propagation speed of the wave, f is the frequency, and λ is the wavelength. Here $v_W = c$, so that for all electromagnetic waves, $c = f\lambda$.

Thus, for all electromagnetic waves, the greater the frequency, the smaller the wavelength.

Figure 1 shows how the various types of electromagnetic waves are categorized according to their wavelengths and frequencies—that is, it shows the electromagnetic spectrum. Many of the characteristics of the various types of electromagnetic waves are related to their frequencies and wavelengths, as we shall see.

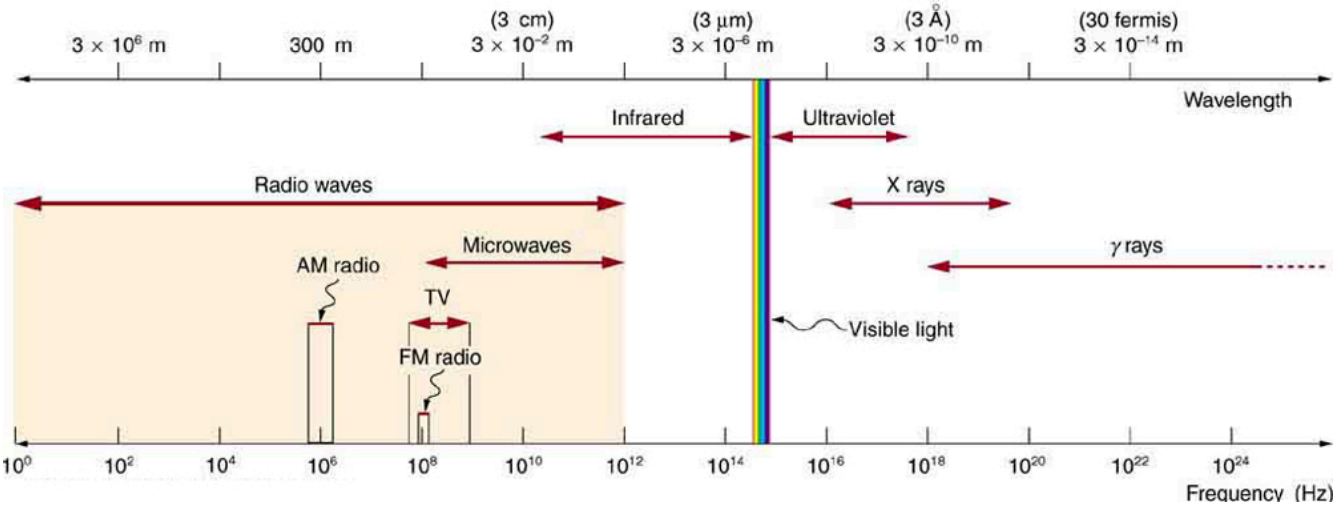


Figure 1. The electromagnetic spectrum, showing the major categories of electromagnetic waves. The range of frequencies and wavelengths is remarkable. The dividing line between some categories is distinct, whereas other categories overlap.

Electromagnetic Spectrum: Rules of Thumb

Three rules that apply to electromagnetic waves in general are as follows:

- High-frequency electromagnetic waves are more energetic and are more able to penetrate than low-frequency waves.
- High-frequency electromagnetic waves can carry more information per unit time than low-frequency waves.
- The shorter the wavelength of any electromagnetic wave probing a material, the smaller the detail it is possible to resolve.

Note that there are exceptions to these rules of thumb.

Transmission, Reflection, and Absorption

What happens when an electromagnetic wave impinges on a material? If the material is transparent to the particular frequency, then the wave can largely be transmitted. If the material is opaque to the frequency, then the wave can be totally reflected. The wave can also be absorbed by the material, indicating that there is some interaction between the wave and the material, such as the thermal agitation of molecules.

Of course it is possible to have partial transmission, reflection, and absorption. We normally associate these properties with visible light, but they do apply to all electromagnetic waves. What is not obvious is that something that is transparent to light may be opaque at other frequencies. For example, ordinary glass is transparent to visible light but largely opaque to ultraviolet radiation. Human skin is opaque to visible light—we cannot see through people—but transparent to X-rays.

Radio and TV Waves

The broad category of *radio waves* is defined to contain any electromagnetic wave produced by currents in wires and circuits. Its name derives from their most common use as a carrier of audio information (i.e., radio). The name is applied to electromagnetic waves of similar frequencies regardless of source. Radio waves from outer space, for example, do not come from alien radio stations. They are created by many astronomical phenomena, and their study has revealed much about nature on the largest scales.

There are many uses for radio waves, and so the category is divided into many subcategories, including microwaves and those electromagnetic waves used for AM and FM radio, cellular telephones, and TV.

The lowest commonly encountered radio frequencies are produced by high-voltage AC power transmission lines at frequencies of 50 or 60 Hz. (See Figure 2.) These extremely long wavelength electromagnetic waves (about 6000 km!) are one means of energy loss in long-distance power transmission.

Figure 2. This high-voltage traction power line running to Eutingen Railway Substation in Germany radiates electromagnetic waves with very long wavelengths. (credit: Zonk43, Wikimedia Commons)

There is an ongoing controversy regarding potential health hazards associated with exposure to these electromagnetic fields (*E*-fields). Some people suspect that living near such transmission lines may cause a variety of illnesses, including cancer. But demographic data are either inconclusive or simply do not support the hazard theory. Recent reports that have looked at many European and American epidemiological studies have found no increase in risk for cancer due to exposure to *E*-fields.

Extremely low frequency (ELF) radio waves of about 1 kHz are used to communicate with submerged submarines. The ability of radio waves to penetrate salt water is related to their wavelength (much like ultrasound penetrating tissue)—the longer the wavelength, the farther they penetrate. Since salt water is a good conductor, radio waves are strongly absorbed by it, and very long wavelengths are needed to reach a submarine under the surface. (See Figure 3.)

AM radio waves are used to carry commercial radio signals in the frequency range from 540 to 1600 kHz. The abbreviation AM stands for *amplitude modulation*, which is the method for placing information on these waves. (See Figure 4.) A *carrier wave* having the basic frequency of the radio station, say 1530 kHz, is varied or modulated in amplitude by an audio signal. The resulting wave has a constant frequency, but a varying amplitude.

A radio receiver tuned to have the same resonant frequency as the carrier wave can pick up the signal, while rejecting the many other frequencies impinging on its antenna. The receiver's circuitry is designed to respond to variations in amplitude of the carrier wave to replicate the original audio signal. That audio signal is amplified to drive a speaker or perhaps to be recorded.

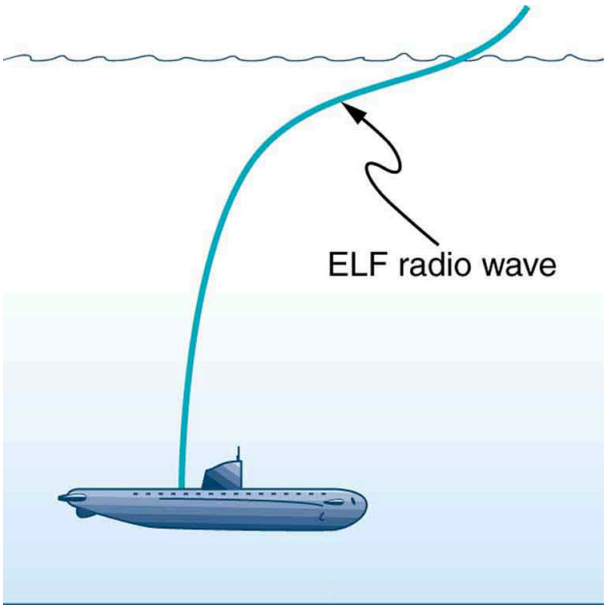


Figure 3. Very long wavelength radio waves are needed to reach this submarine, requiring extremely low frequency signals (ELF). Shorter wavelengths do not penetrate to any significant depth.

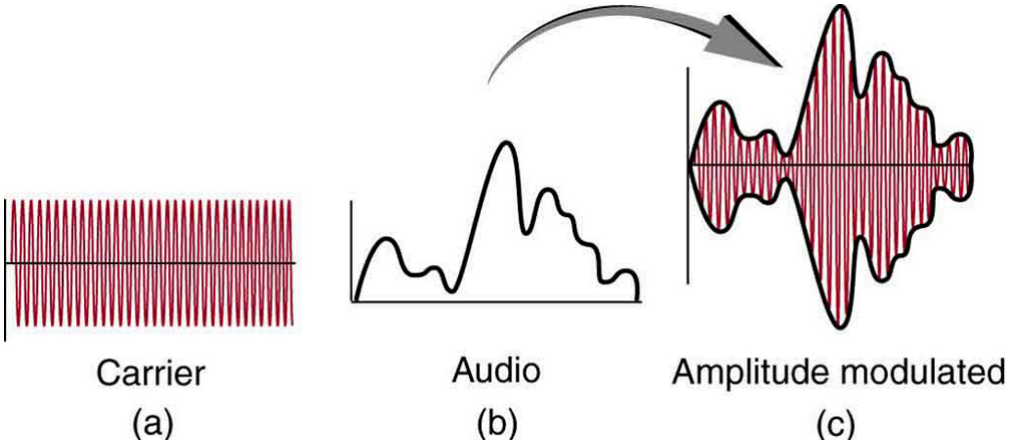


Figure 4. Amplitude modulation for AM radio. (a) A carrier wave at the station's basic frequency. (b) An audio signal at much lower audible frequencies. (c) The amplitude of the carrier is modulated by the audio signal without changing its basic frequency.

FM Radio Waves

FM radio waves are also used for commercial radio transmission, but in the frequency range of 88 to 108 MHz. FM stands for *frequency modulation*, another method of carrying information. (See Figure 5.) Here a carrier wave having the basic frequency of the radio station, perhaps 105.1 MHz, is modulated in frequency by the audio signal, producing a wave of constant amplitude but varying frequency.

Since audible frequencies range up to 20 kHz (or 0.020 MHz) at most, the frequency of the FM radio wave can vary from the carrier by as much as 0.020 MHz. Thus the carrier frequencies of two different radio stations cannot be closer than 0.020 MHz. An FM receiver is tuned to resonate at the carrier frequency and has circuitry that responds to variations in frequency, reproducing the audio information.

FM radio is inherently less subject to noise from stray radio sources than AM radio. The reason is that amplitudes of waves add. So an AM receiver would interpret noise added onto the amplitude of its carrier wave as part of the information. An FM receiver can be made to reject amplitudes other than that of the basic carrier wave and only look for variations in frequency. It is thus easier to reject noise from FM, since noise produces a variation in amplitude.

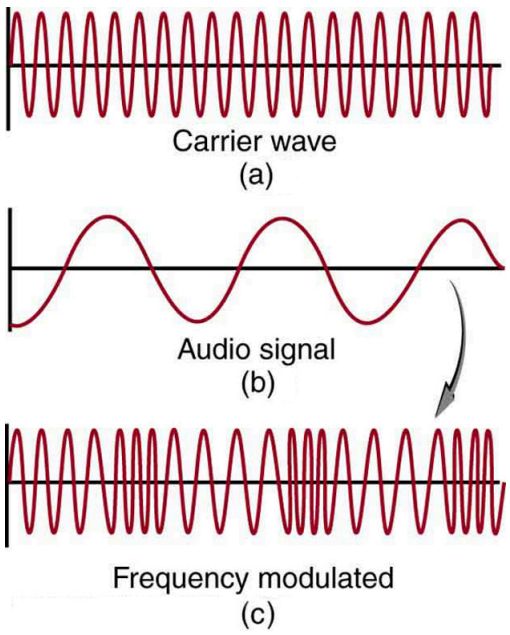


Figure 5. Frequency modulation for FM radio. (a) A carrier wave at the station's basic frequency. (b) An audio signal at much lower audible frequencies. (c) The frequency of the carrier is modulated by the audio signal without changing its amplitude.

Television is also broadcast on electromagnetic waves. Since the waves must carry a great deal of visual as well as audio information, each channel requires a larger range of frequencies than simple radio transmission. TV channels utilize frequencies in the range of 54 to 88 MHz and 174 to 222 MHz. (The entire FM radio band lies between channels 88 MHz and 174 MHz.) These TV channels are called VHF (for *very high frequency*). Other channels called UHF (for *ultra high frequency*) utilize an even higher frequency range of 470 to 1000 MHz.

The TV video signal is AM, while the TV audio is FM. Note that these frequencies are those of free transmission with the user utilizing an old-fashioned roof antenna. Satellite dishes and cable transmission of TV occurs at significantly higher frequencies and is rapidly evolving with the use of the high-definition or HD format.

Example 1. Calculating Wavelengths of Radio Waves

Calculate the wavelengths of a 1530-kHz AM radio signal, a 105.1-MHz FM radio signal, and a 1.90-GHz cell phone signal.

Strategy

The relationship between wavelength and frequency is $c = f\lambda$, where $c = 3.00 \times 10^8$ m/s is the speed of light (the speed of light is only very slightly smaller in air than it is in a vacuum). We can rearrange this equation to find the wavelength for all three frequencies.

Solution

Rearranging gives

$$\lambda = \frac{c}{f}$$

.

For the f = 1530 kHz AM radio signal:

$$\begin{aligned} \lambda &= \frac{3.00 \times 10^8 \text{ m/s}}{1530 \times 10^3 \text{ cycles/s}} \\ &= 196 \text{ m} \end{aligned}$$

For the f = 105.1 MHz FM radio signal:

$$\begin{aligned} \lambda &= \frac{3.00 \times 10^8 \text{ m/s}}{105.1 \times 10^6 \text{ cycles/s}} \\ &= 2.85 \text{ m} \end{aligned}$$

And for the f = 1.90 GHz cell phone:

$$\begin{aligned} \lambda &= \frac{3.00 \times 10^8 \text{ m/s}}{1.90 \times 10^9 \text{ cycles/s}} \\ &= 0.158 \text{ m} \end{aligned}$$

Discussion

These wavelengths are consistent with the spectrum in Figure 1. The wavelengths are also related to other properties of these electromagnetic waves, as we shall see.

The wavelengths found in the preceding example are representative of AM, FM, and cell phones, and account for some of the differences in how they are broadcast and how well they travel. The most efficient length for a linear antenna, such as discussed in Production of Electromagnetic Waves, is

$$\frac{\lambda}{2}$$

, half the wavelength of the electromagnetic wave. Thus a very large antenna is needed to efficiently broadcast typical AM radio with its carrier wavelengths on the order of hundreds of meters.

One benefit to these long AM wavelengths is that they can go over and around rather large obstacles (like buildings and hills), just as ocean waves can go around large rocks. FM and TV are best received when there is a line of sight between the broadcast antenna and receiver, and they are often sent from very tall structures. FM, TV, and mobile phone antennas themselves are much smaller than those used for AM, but they are elevated to achieve an unobstructed line of sight. (See Figure 6.)

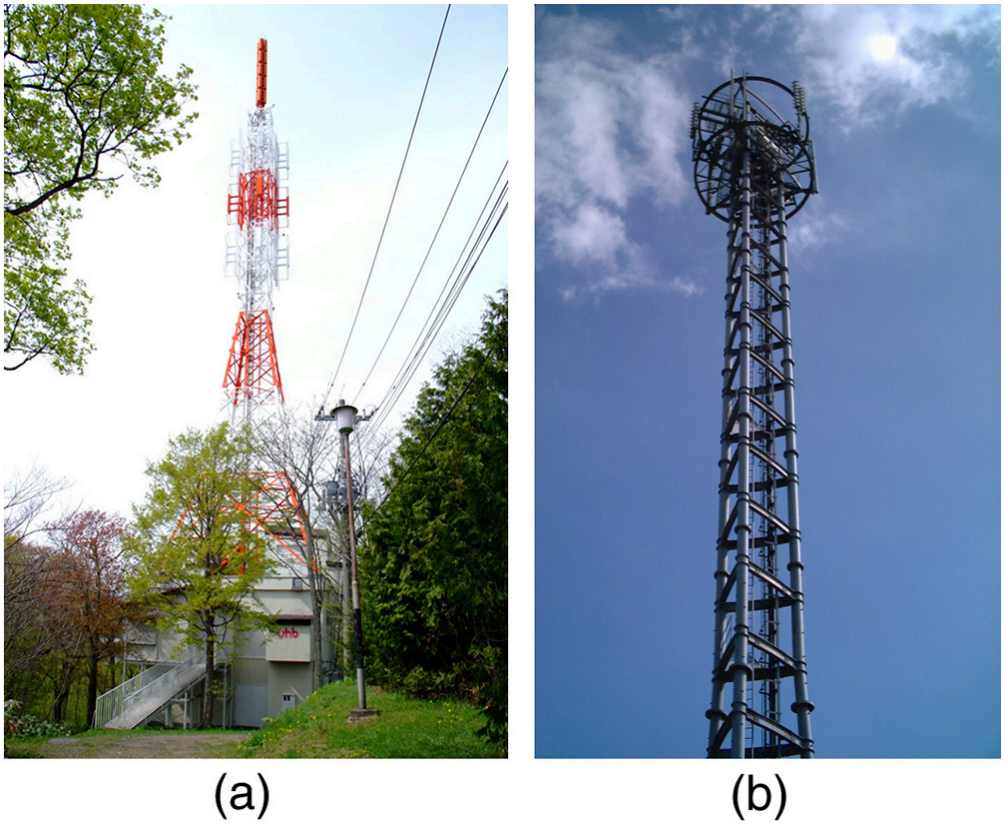


Figure 6. (a) A large tower is used to broadcast TV signals. The actual antennas are small structures on top of the tower—they are placed at great heights to have a clear line of sight over a large broadcast area. (credit: Ozizo, Wikimedia Commons) (b) The NTT Dokomo mobile phone tower at Tokorozawa City, Japan. (credit: tokoroten, Wikimedia Commons)

Radio Wave Interference

Astronomers and astrophysicists collect signals from outer space using electromagnetic waves. A common problem for astrophysicists is the "pollution" from electromagnetic radiation pervading our surroundings from communication systems in general. Even everyday gadgets like our car keys having the facility to lock car doors remotely and being able to turn TVs on and off using remotes involve radio-wave frequencies. In order to prevent interference between all these electromagnetic signals, strict regulations are drawn up for different organizations to utilize different radio frequency bands.

One reason why we are sometimes asked to switch off our mobile phones (operating in the range of 1.9 GHz) on airplanes and in hospitals is that important communications or medical equipment often uses similar radio frequencies and their operation can be affected by frequencies used in the communication devices.

For example, radio waves used in magnetic resonance imaging (MRI) have frequencies on the order of 100 MHz, although this varies significantly depending on the strength of the magnetic field used and the nuclear type being scanned. MRI is an important medical imaging and research tool, producing highly detailed two- and three-dimensional images. Radio waves are broadcast, absorbed, and reemitted in a resonance process that is sensitive to the density of nuclei (usually protons or hydrogen nuclei).

The wavelength of 100-MHz radio waves is 3 m, yet using the sensitivity of the resonant frequency to the magnetic field strength, details smaller than a millimeter can be imaged. This is a good example of an exception to a rule of thumb (in this case, the rubric that details much smaller than the probe's wavelength cannot be detected). The intensity of the radio waves used in MRI presents little or no hazard to human health.

Microwaves

Microwaves are the highest-frequency electromagnetic waves that can be produced by currents in macroscopic circuits and devices. Microwave frequencies range from about 10^9 Hz to the highest practical *LC* resonance at nearly 10^{12} Hz. Since they have high frequencies, their wavelengths are short compared with those of other radio waves—hence the name "microwave."

Microwaves can also be produced by atoms and molecules. They are, for example, a component of electromagnetic radiation generated by *thermal agitation*. The thermal motion of atoms and molecules in any object at a temperature above absolute zero causes them to emit and absorb radiation.

Since it is possible to carry more information per unit time on high frequencies, microwaves are quite suitable for communications. Most satellite-transmitted information is carried on microwaves, as are land-based long-distance transmissions. A clear line of sight between transmitter and receiver is needed because of the short wavelengths involved.

Radar is a common application of microwaves that was first developed in World War II. By detecting and timing microwave echoes, radar systems can determine the distance to objects as diverse as clouds and aircraft. A Doppler shift in the radar echo can be used to determine the speed of a car or the intensity of a rainstorm. Sophisticated radar systems are used to map the Earth and other planets, with a resolution limited by wavelength. (See Figure 7.) The shorter the wavelength of any probe, the smaller the detail it is possible to observe.

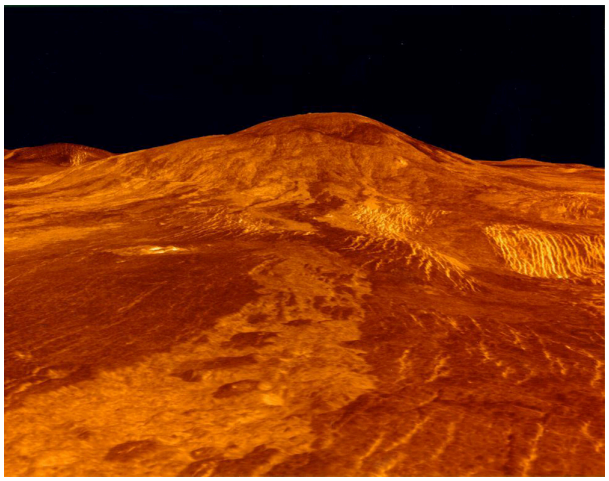


Figure 7. An image of Sif Mons with lava flows on Venus, based on Magellan synthetic aperture radar data combined with radar altimetry to produce a three-dimensional map of the surface. The Venusian atmosphere is opaque to visible light, but not to the microwaves that were used to create this image. (credit: NSSDC, NASA/JPL)

Heating with Microwaves

How does the ubiquitous microwave oven produce microwaves electronically, and why does food absorb them preferentially? Microwaves at a frequency of 2.45 GHz are produced by accelerating electrons. The microwaves are then used to induce an alternating electric field in the oven.

Water and some other constituents of food have a slightly negative charge at one end and a slightly positive charge at one end (called polar molecules). The range of microwave frequencies is specially

selected so that the polar molecules, in trying to keep orienting themselves with the electric field, absorb these energies and increase their temperatures—called dielectric heating.

The energy thereby absorbed results in thermal agitation heating food and not the plate, which does not contain water. Hot spots in the food are related to constructive and destructive interference patterns. Rotating antennas and food turntables help spread out the hot spots.

Another use of microwaves for heating is within the human body. Microwaves will penetrate more than shorter wavelengths into tissue and so can accomplish "deep heating" (called microwave diathermy). This is used for treating muscular pains, spasms, tendonitis, and rheumatoid arthritis.

Take-Home Experiment—Microwave Ovens

1. Look at the door of a microwave oven. Describe the structure of the door. Why is there a metal grid on the door? How does the size of the holes in the grid compare with the wavelengths of microwaves used in microwave ovens? What is this wavelength?
2. Place a glass of water (about 250 ml) in the microwave and heat it for 30 seconds. Measure the temperature gain (the ΔT). Assuming that the power output of the oven is 1000 W, calculate the efficiency of the heat-transfer process.
3. Remove the rotating turntable or moving plate and place a cup of water in several places along a line parallel with the opening. Heat for 30 seconds and measure the ΔT for each position. Do you see cases of destructive interference?

Microwaves generated by atoms and molecules far away in time and space can be received and detected by electronic circuits. Deep space acts like a blackbody with a 2.7 K temperature, radiating most of its energy in the microwave frequency range. In 1964, Penzias and Wilson detected this radiation and eventually recognized that it was the radiation of the Big Bang's cooled remnants.

Infrared Radiation

The microwave and infrared regions of the electromagnetic spectrum overlap (see Figure 1). *Infrared radiation* is generally produced by thermal motion and the vibration and rotation of atoms and molecules. Electronic transitions in atoms and molecules can also produce infrared radiation.

The range of infrared frequencies extends up to the lower limit of visible light, just below red. In fact, infrared means "below red." Frequencies at its upper limit are too high to be produced by accelerating electrons in circuits, but small systems, such as atoms and molecules, can vibrate fast enough to produce these waves.

Water molecules rotate and vibrate particularly well at infrared frequencies, emitting and absorbing them so efficiently that the emissivity for skin is $e = 0.97$ in the infrared. Night-vision scopes can detect the infrared emitted by various warm objects, including humans, and convert it to visible light.

We can examine radiant heat transfer from a house by using a camera capable of detecting infrared radiation. Reconnaissance satellites can detect buildings, vehicles, and even individual humans by

their infrared emissions, whose power radiation is proportional to the fourth power of the absolute temperature. More mundanely, we use infrared lamps, some of which are called quartz heaters, to preferentially warm us because we absorb infrared better than our surroundings.

The Sun radiates like a nearly perfect blackbody (that is, it has $e = 1$), with a 6000 K surface temperature. About half of the solar energy arriving at the Earth is in the infrared region, with most of the rest in the visible part of the spectrum, and a relatively small amount in the ultraviolet. On average, 50 percent of the incident solar energy is absorbed by the Earth.

The relatively constant temperature of the Earth is a result of the energy balance between the incoming solar radiation and the energy radiated from the Earth. Most of the infrared radiation emitted from the Earth is absorbed by CO_2 and H_2O in the atmosphere and then radiated back to Earth or into outer space. This radiation back to Earth is known as the greenhouse effect, and it maintains the surface temperature of the Earth about 40ºC higher than it would be if there is no absorption. Some scientists think that the increased concentration of CO_2 and other greenhouse gases in the atmosphere, resulting from increases in fossil fuel burning, has increased global average temperatures.

Visible Light

Visible light is the narrow segment of the electromagnetic spectrum to which the normal human eye responds. Visible light is produced by vibrations and rotations of atoms and molecules, as well as by electronic transitions within atoms and molecules. The receivers or detectors of light largely utilize electronic transitions. We say the atoms and molecules are excited when they absorb and relax when they emit through electronic transitions.

Figure 8 shows this part of the spectrum, together with the colors associated with particular pure wavelengths. We usually refer to visible light as having wavelengths of between 400 nm and 750 nm. (The retina of the eye actually responds to the lowest ultraviolet frequencies, but these do not normally reach the retina because they are absorbed by the cornea and lens of the eye.)

Red light has the lowest frequencies and longest wavelengths, while violet has the highest frequencies and shortest wavelengths. Blackbody radiation from the Sun peaks in the visible part of the spectrum but is more intense in the red than in the violet, making the Sun yellowish in appearance.

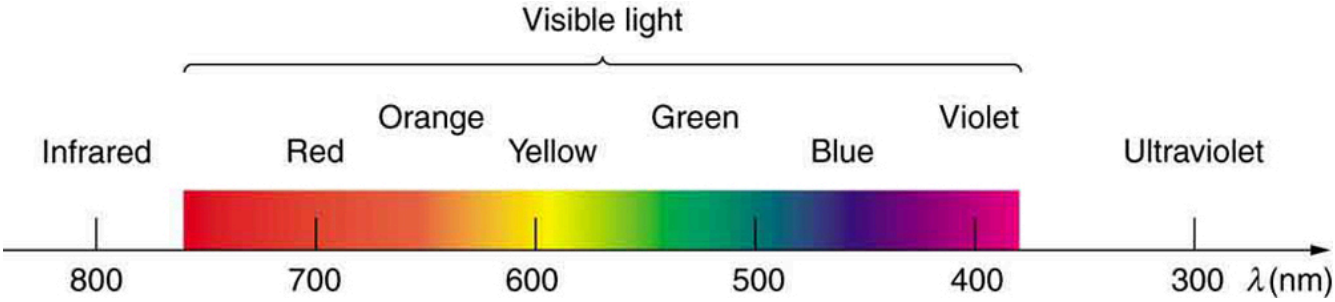


Figure 8. A small part of the electromagnetic spectrum that includes its visible components. The divisions between infrared, visible, and ultraviolet are not perfectly distinct, nor are those between the seven rainbow colors.

Living things—plants and animals—have evolved to utilize and respond to parts of the electromagnetic spectrum they are embedded in. Visible light is the most predominant and we enjoy the beauty of nature

through visible light. Plants are more selective. Photosynthesis makes use of parts of the visible spectrum to make sugars.

Example 2. Integrated Concept Problem: Correcting Vision with Lasers

During laser vision correction, a brief burst of 193-nm ultraviolet light is projected onto the cornea of a patient. It makes a spot 0.80 mm in diameter and evaporates a layer of cornea 0.30 μm thick. Calculate the energy absorbed, assuming the corneal tissue has the same properties as water; it is initially at 34ºC. Assume the evaporated tissue leaves at a temperature of 100ºC.

Strategy

The energy from the laser light goes toward raising the temperature of the tissue and also toward evaporating it. Thus we have two amounts of heat to add together. Also, we need to find the mass of corneal tissue involved.

Solution

To Figure out the heat required to raise the temperature of the tissue to 100ºC, we can apply concepts of thermal energy. We know that

$$Q = mc\Delta T,$$

where Q is the heat required to raise the temperature, ΔT is the desired change in temperature, m is the mass of tissue to be heated, and c is the specific heat of water equal to 4186 J/kg/K.

Without knowing the mass m at this point, we have

$$Q = m(4186 \text{ J/kg/K})(100^\circ\text{C}–34^\circ\text{C}) = m(276{,}276 \text{ J/kg}) = m(276 \text{ kJ/kg}).$$

The latent heat of vaporization of water is 2256 kJ/kg, so that the energy needed to evaporate mass m is

$$Q_\text{v} = mL_\text{v} = m(2256 \text{ kJ/kg}).$$

To find the mass m, we use the equation

$$\rho = \frac{m}{V}$$

, where ρ is the density of the tissue and V is its volume. For this case,

$$\begin{aligned} m &= \rho V \\ &= \left(1000 \text{ kg/m}^3\right)\left(\text{area} \times \text{thickness}\left(m^3\right)\right) \\ &= \left(1000 \text{ kg/m}^3\right)\left(\pi\left(0.80 \times 10^{-3} \text{ m}\right)^2/4\right)\left(0.30 \times 10^{-6} \text{ m}\right) \\ &= 0.151 \times 10^{-9} \text{ kg} \end{aligned}$$

Therefore, the total energy absorbed by the tissue in the eye is the sum of Q and Q_v:

$$Q_\text{tot} = m(c\Delta T + L_\text{v}) = (0.151 \times 10^{-9} \text{ kg})(276 \cdot \text{kJ/kg} + 2256 \text{ kJ/kg}) = 382 \times 10^{-9} \text{ kJ}.$$

Discussion

The lasers used for this eye surgery are excimer lasers, whose light is well absorbed by biological tissue. They evaporate rather than burn the tissue, and can be used for precision work. Most lasers used for this type of eye surgery have an average power rating of about one watt. For our example, if we assume that each laser

burst from this pulsed laser lasts for 10 ns, and there are 400 bursts per second, then the average power is $Q_{tot} \times 400 = 150$ mW.

Optics is the study of the behavior of visible light and other forms of electromagnetic waves. Optics falls into two distinct categories. When electromagnetic radiation, such as visible light, interacts with objects that are large compared with its wavelength, its motion can be represented by straight lines like rays. Ray optics is the study of such situations and includes lenses and mirrors.

When electromagnetic radiation interacts with objects about the same size as the wavelength or smaller, its wave nature becomes apparent. For example, observable detail is limited by the wavelength, and so visible light can never detect individual atoms, because they are so much smaller than its wavelength. Physical or wave optics is the study of such situations and includes all wave characteristics.

Take-Home Experiment: Colors That Match

When you light a match you see largely orange light; when you light a gas stove you see blue light. Why are the colors different? What other colors are present in these?

Ultraviolet Radiation

Ultraviolet means "above violet." The electromagnetic frequencies of *ultraviolet radiation (UV)* extend upward from violet, the highest-frequency visible light. Ultraviolet is also produced by atomic and molecular motions and electronic transitions. The wavelengths of ultraviolet extend from 400 nm down to about 10 nm at its highest frequencies, which overlap with the lowest X-ray frequencies. It was recognized as early as 1801 by Johann Ritter that the solar spectrum had an invisible component beyond the violet range.

Solar UV radiation is broadly subdivided into three regions: UV-A (320–400 nm), UV-B (290–320 nm), and UV-C (220–290 nm), ranked from long to shorter wavelengths (from smaller to larger energies). Most UV-B and all UV-C is absorbed by ozone (O_3) molecules in the upper atmosphere. Consequently, 99% of the solar UV radiation reaching the Earth's surface is UV-A.

Human Exposure to UV Radiation

It is largely exposure to UV-B that causes skin cancer. It is estimated that as many as 20% of adults will develop skin cancer over the course of their lifetime. Again, treatment is often successful if caught early. Despite very little UV-B reaching the Earth's surface, there are substantial increases in skin-cancer rates in countries such as Australia, indicating how important it is that UV-B and UV-C continue to be absorbed by the upper atmosphere.

All UV radiation can damage collagen fibers, resulting in an acceleration of the aging process of skin and the formation of wrinkles. Because there is so little UV-B and UV-C reaching the Earth's surface,

sunburn is caused by large exposures, and skin cancer from repeated exposure. Some studies indicate a link between overexposure to the Sun when young and melanoma later in life.

The tanning response is a defense mechanism in which the body produces pigments to absorb future exposures in inert skin layers above living cells. Basically UV-B radiation excites DNA molecules, distorting the DNA helix, leading to mutations and the possible formation of cancerous cells.

Repeated exposure to UV-B may also lead to the formation of cataracts in the eyes—a cause of blindness among people living in the equatorial belt where medical treatment is limited. Cataracts, clouding in the eye's lens and a loss of vision, are age related; 60% of those between the ages of 65 and 74 will develop cataracts. However, treatment is easy and successful, as one replaces the lens of the eye with a plastic lens. Prevention is important. Eye protection from UV is more effective with plastic sunglasses than those made of glass.

A major acute effect of extreme UV exposure is the suppression of the immune system, both locally and throughout the body.

Low-intensity ultraviolet is used to sterilize haircutting implements, implying that the energy associated with ultraviolet is deposited in a manner different from lower-frequency electromagnetic waves. (Actually this is true for all electromagnetic waves with frequencies greater than visible light.)

Flash photography is generally not allowed of precious artworks and colored prints because the UV radiation from the flash can cause photo-degradation in the artworks. Often artworks will have an extra-thick layer of glass in front of them, which is especially designed to absorb UV radiation.

UV Light and the Ozone Layer

If all of the Sun's ultraviolet radiation reached the Earth's surface, there would be extremely grave effects on the biosphere from the severe cell damage it causes. However, the layer of ozone (O_3) in our upper atmosphere (10 to 50 km above the Earth) protects life by absorbing most of the dangerous UV radiation.

Unfortunately, today we are observing a depletion in ozone concentrations in the upper atmosphere. This depletion has led to the formation of an "ozone hole" in the upper atmosphere. The hole is more centered over the southern hemisphere, and changes with the seasons, being largest in the spring. This depletion is attributed to the breakdown of ozone molecules by refrigerant gases called chlorofluorocarbons (CFCs).

The UV radiation helps dissociate the CFC's, releasing highly reactive chlorine (Cl) atoms, which catalyze the destruction of the ozone layer. For example, the reaction of $CFCl_3$ with a photon of light ($h\nu$) can be written as

$$CFCl_3 + h\nu \rightarrow CFCl_2 + Cl.$$

The Cl atom then catalyzes the breakdown of ozone as follows:

$$Cl + O_3 \rightarrow ClO + O_2 \text{ and } ClO + O_3 \rightarrow Cl + 2O_2.$$

A single chlorine atom could destroy ozone molecules for up to two years before being transported down to the surface. The CFCs are relatively stable and will contribute to ozone depletion for years to come. CFCs are found in refrigerants, air conditioning systems, foams, and aerosols.

International concern over this problem led to the establishment of the "Montreal Protocol" agreement (1987) to phase out CFC production in most countries. However, developing-country participation is needed if worldwide production and elimination of CFCs is to be achieved. Probably the largest contributor to CFC emissions today is India. But the protocol seems to be working, as there are signs of an ozone recovery. (See Figure 9.)

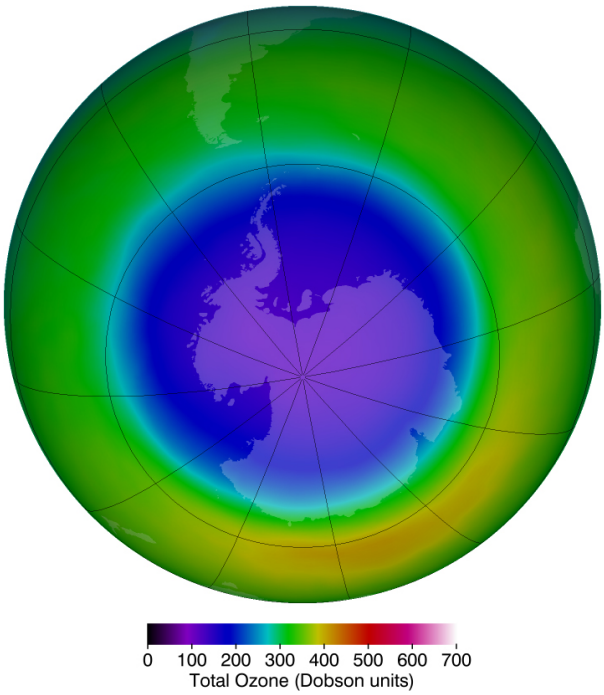


Figure 9. This map of ozone concentration over Antarctica in October 2011 shows severe depletion suspected to be caused by CFCs. Less dramatic but more general depletion has been observed over northern latitudes, suggesting the effect is global. With less ozone, more ultraviolet radiation from the Sun reaches the surface, causing more damage. (credit: NASA Ozone Watch)

Benefits of UV Light

Besides the adverse effects of ultraviolet radiation, there are also benefits of exposure in nature and uses in technology. Vitamin D production in the skin (epidermis) results from exposure to UVB radiation, generally from sunlight. A number of studies indicate lack of vitamin D can result in the development of a range of cancers (prostate, breast, colon), so a certain amount of UV exposure is helpful. Lack of vitamin D is also linked to osteoporosis. Exposures (with no sunscreen) of 10 minutes a day to arms, face, and legs might be sufficient to provide the accepted dietary level. However, in the winter time north of about 37º latitude, most UVB gets blocked by the atmosphere.

UV radiation is used in the treatment of infantile jaundice and in some skin conditions. It is also used in sterilizing workspaces and tools, and killing germs in a wide range of applications. It is also used as an analytical tool to identify substances.

When exposed to ultraviolet, some substances, such as minerals, glow in characteristic visible wavelengths, a process called fluorescence. So-called black lights emit ultraviolet to cause posters and clothing to fluoresce in the visible. Ultraviolet is also used in special microscopes to detect details smaller than those observable with longer-wavelength visible-light microscopes.

Things Great and Small: A Submicroscopic View of X-Ray Production

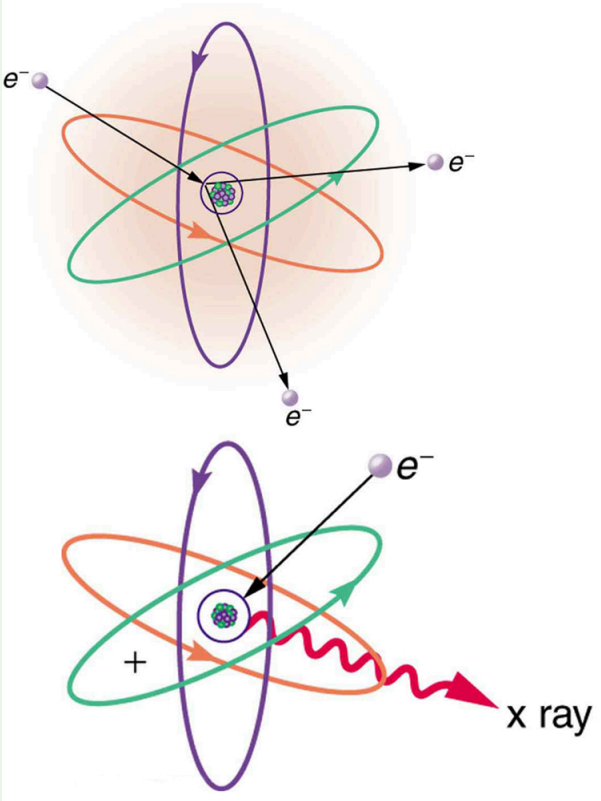


Figure 10. Artist's conception of an electron ionizing an atom followed by the recapture of an electron and emission of an X-ray. An energetic electron strikes an atom and knocks an electron out of one of the orbits closest to the nucleus. Later, the atom captures another electron, and the energy released by its fall into a low orbit generates a high-energy EM wave called an X-ray.

X-rays can be created in a high-voltage discharge. They are emitted in the material struck by electrons in the discharge current. There are two mechanisms by which the electrons create X-rays.

The first method is illustrated in Figure 10. An electron is accelerated in an evacuated tube by a high positive voltage. The electron strikes a metal plate (e.g., copper) and produces X-rays. Since this is a high-voltage discharge, the electron gains sufficient energy to ionize the atom.

In the case shown, an inner-shell electron (one in an orbit relatively close to and tightly bound to the nucleus) is ejected. A short time later, another electron is captured and falls into the orbit in a single great plunge. The energy released by this fall is given to an EM wave known as an X-ray. Since the orbits of the atom are unique to the type of atom, the energy of the X-ray is characteristic of the atom, hence the name characteristic X-ray.

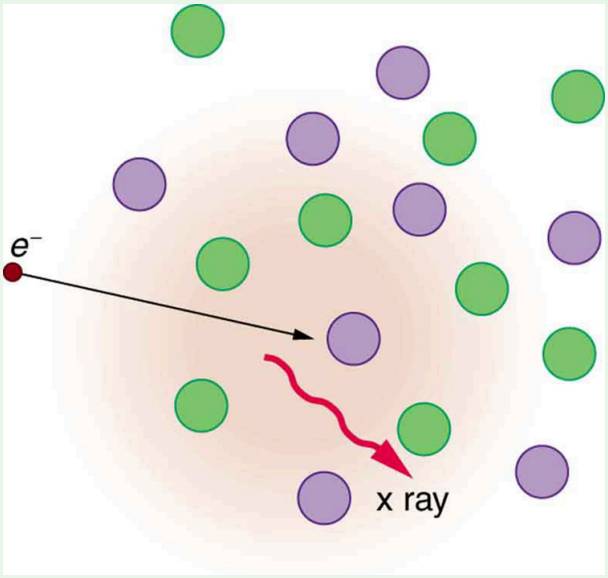


Figure 11. Artist's conception of an electron being slowed by collisions in a material and emitting X-ray radiation. This energetic electron makes numerous collisions with electrons and atoms in a material it penetrates. An accelerated charge radiates EM waves, a second method by which X-rays are created.

The second method by which an energetic electron creates an X-ray when it strikes a material is illustrated in Figure 11. The electron interacts with charges in the material as it penetrates. These collisions transfer kinetic energy from the electron to the electrons and atoms in the material.

A loss of kinetic energy implies an acceleration, in this case decreasing the electron's velocity. Whenever a charge is accelerated, it radiates EM waves. Given the high energy of the electron, these EM waves can have high energy. We call them X-rays. Since the process is random, a broad spectrum of X-ray energy is emitted that is more characteristic of the electron energy than the type of material the electron encounters. Such EM radiation is called "bremsstrahlung" (German for "braking radiation").

X-Rays

In the 1850s, scientists (such as Faraday) began experimenting with high-voltage electrical discharges in tubes filled with rarefied gases. It was later found that these discharges created an invisible, penetrating form of very high frequency electromagnetic radiation. This radiation was called an *X-ray*, because its identity and nature were unknown.

As described in "Things Great and Small" feature, there are two methods by which X-rays are created—both are submicroscopic processes and can be caused by high-voltage discharges. While the low-frequency end of the X-ray range overlaps with the ultraviolet, X-rays extend to much higher frequencies (and energies).

X-rays have adverse effects on living cells similar to those of ultraviolet radiation, and they have the additional liability of being more penetrating, affecting more than the surface layers of cells. Cancer and

genetic defects can be induced by exposure to X-rays. Because of their effect on rapidly dividing cells, X-rays can also be used to treat and even cure cancer.

The widest use of X-rays is for imaging objects that are opaque to visible light, such as the human body or aircraft parts. In humans, the risk of cell damage is weighed carefully against the benefit of the diagnostic information obtained. However, questions have risen in recent years as to accidental overexposure of some people during CT scans—a mistake at least in part due to poor monitoring of radiation dose.

The ability of X-rays to penetrate matter depends on density, and so an X-ray image can reveal very detailed density information. Figure 12 shows an example of the simplest type of X-ray image, an X-ray shadow on film. The amount of information in a simple X-ray image is impressive, but more sophisticated techniques, such as CT scans, can reveal three-dimensional information with details smaller than a millimeter.

The use of X-ray technology in medicine is called radiology—an established and relatively cheap tool in comparison to more sophisticated technologies. Consequently, X-rays are widely available and used extensively in medical diagnostics. During World War I, mobile X-ray units, advocated by Madame Marie Curie, were used to diagnose soldiers.

Because they can have wavelengths less than 0.01 nm, X-rays can be scattered (a process called X-ray diffraction) to detect the shape of molecules and the structure of crystals. X-ray diffraction was crucial to Crick, Watson, and Wilkins in the determination of the shape of the double-helix DNA molecule.

X-rays are also used as a precise tool for trace-metal analysis in X-ray induced fluorescence, in which the energy of the X-ray emissions are related to the specific types of elements and amounts of materials present.

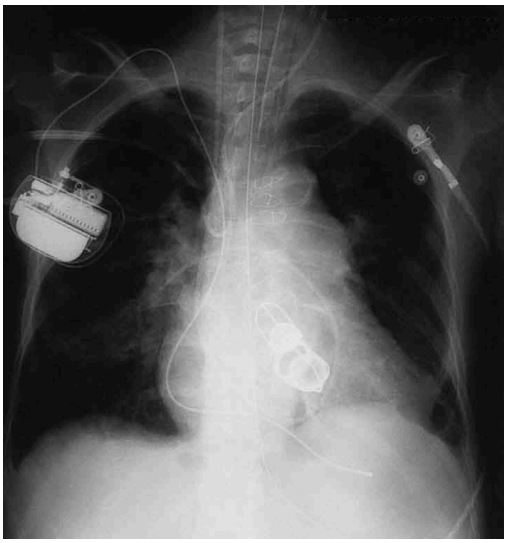


Figure 12. This shadow X-ray image shows many interesting features, such as artificial heart valves, a pacemaker, and the wires used to close the sternum. (credit: P. P. Urone)

Gamma Rays

Soon after nuclear radioactivity was first detected in 1896, it was found that at least three distinct types of radiation were being emitted. The most penetrating nuclear radiation was called a *gamma ray* (**γ** ray) (again a name given because its identity and character were unknown), and it was later found to be an extremely high frequency electromagnetic wave.

In fact, *γ* rays are any electromagnetic radiation emitted by a nucleus. This can be from natural nuclear decay or induced nuclear processes in nuclear reactors and weapons. The lower end of the *γ*-ray frequency range overlaps the upper end of the X-ray range, but *γ* rays can have the highest frequency of any electromagnetic radiation.

Gamma rays have characteristics identical to X-rays of the same frequency—they differ only in source. At higher frequencies, *γ* rays are more penetrating and more damaging to living tissue. They have many of the same uses as X-rays, including cancer therapy. Gamma radiation from radioactive materials is used in nuclear medicine.

Figure 13 shows a medical image based on *γ* rays. Food spoilage can be greatly inhibited by exposing it to large doses of *γ* radiation, thereby obliterating responsible microorganisms. Damage to food cells through irradiation occurs as well, and the long-term hazards of consuming radiation-preserved food are unknown and controversial for some groups. Both X-ray and *γ*-ray technologies are also used in scanning luggage at airports.

Figure 13. This is an image of the γ-rays emitted by nuclei in a compound that is concentrated in the bones and eliminated through the kidneys. Bone cancer is evidenced by nonuniform concentration in similar structures. For example, some ribs are darker than others. (credit: P. P. Urone)

Detecting Electromagnetic Waves from Space

A final note on star gazing. The entire electromagnetic spectrum is used by researchers for investigating stars, space, and time. As noted earlier, Penzias and Wilson detected microwaves to identify the background radiation originating from the Big Bang. Radio telescopes such as the Arecibo Radio Telescope in Puerto Rico and Parkes Observatory in Australia were designed to detect radio waves.

Infrared telescopes need to have their detectors cooled by liquid nitrogen to be able to gather useful signals. Since infrared radiation is predominantly from thermal agitation, if the detectors were not cooled, the vibrations of the molecules in the antenna would be stronger than the signal being collected.

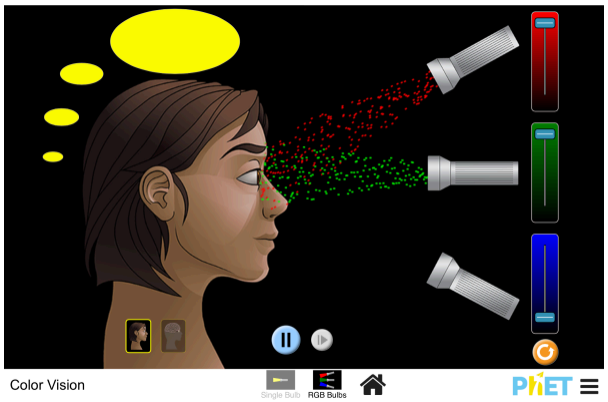
The most famous of these infrared sensitive telescopes is the James Clerk Maxwell Telescope in Hawaii. The earliest telescopes, developed in the seventeenth century, were optical telescopes,

collecting visible light. Telescopes in the ultraviolet, X-ray, and γ-ray regions are placed outside the atmosphere on satellites orbiting the Earth.

The Hubble Space Telescope (launched in 1990) gathers ultraviolet radiation as well as visible light. In the X-ray region, there is the Chandra X-ray Observatory (launched in 1999), and in the γ-ray region, there is the new Fermi Gamma-ray Space Telescope (launched in 2008—taking the place of the Compton Gamma Ray Observatory, 1991–2000.).

PhET Explorations: Color Vision

Make a whole rainbow by mixing red, green, and blue light. Change the wavelength of a monochromatic beam or filter white light. View the light as a solid beam, or see the individual photons.



Click to run the simulation.

Section Summary

- The relationship among the speed of propagation, wavelength, and frequency for any wave is given by $v_W = f\lambda$, so that for electromagnetic waves, $c = f\lambda$, where f is the frequency, λ is the wavelength, and c is the speed of light.
- The electromagnetic spectrum is separated into many categories and subcategories, based on the frequency and wavelength, source, and uses of the electromagnetic waves.
- Any electromagnetic wave produced by currents in wires is classified as a radio wave, the lowest frequency electromagnetic waves. Radio waves are divided into many types, depending on their applications, ranging up to microwaves at their highest frequencies.
- Infrared radiation lies below visible light in frequency and is produced by thermal motion and the vibration and rotation of atoms and molecules. Infrared's lower frequencies overlap with the highest-frequency microwaves.
- Visible light is largely produced by electronic transitions in atoms and molecules, and is

defined as being detectable by the human eye. Its colors vary with frequency, from red at the lowest to violet at the highest.

- Ultraviolet radiation starts with frequencies just above violet in the visible range and is produced primarily by electronic transitions in atoms and molecules.
- X-rays are created in high-voltage discharges and by electron bombardment of metal targets. Their lowest frequencies overlap the ultraviolet range but extend to much higher values, overlapping at the high end with gamma rays.
- Gamma rays are nuclear in origin and are defined to include the highest-frequency electromagnetic radiation of any type.

Conceptual Questions

1. If you live in a region that has a particular TV station, you can sometimes pick up some of its audio portion on your FM radio receiver. Explain how this is possible. Does it imply that TV audio is broadcast as FM?
2. Explain why people who have the lens of their eye removed because of cataracts are able to see low-frequency ultraviolet.
3. How do fluorescent soap residues make clothing look “brighter and whiter” in outdoor light? Would this be effective in candlelight?
4. Give an example of resonance in the reception of electromagnetic waves.
5. Illustrate that the size of details of an object that can be detected with electromagnetic waves is related to their wavelength, by comparing details observable with two different types (for example, radar and visible light or infrared and X-rays).
6. Why don’t buildings block radio waves as completely as they do visible light?
7. Make a list of some everyday objects and decide whether they are transparent or opaque to each of the types of electromagnetic waves.
8. Your friend says that more patterns and colors can be seen on the wings of birds if viewed in ultraviolet light. Would you agree with your friend? Explain your answer.
9. The rate at which information can be transmitted on an electromagnetic wave is proportional to the frequency of the wave. Is this consistent with the fact that laser telephone transmission at visible frequencies carries far more conversations per optical fiber than conventional electronic transmission in a wire? What is the implication for ELF radio communication with submarines?
10. Give an example of energy carried by an electromagnetic wave.
11. In an MRI scan, a higher magnetic field requires higher frequency radio waves to resonate with the nuclear type whose density and location is being imaged. What effect does going to a larger magnetic field have on the most efficient antenna to broadcast those radio waves? Does it favor a smaller or larger antenna?
12. Laser vision correction often uses an excimer laser that produces 193-nm electromagnetic radiation. This wavelength is extremely strongly absorbed by the cornea and ablates it in a manner that reshapes the cornea to correct vision defects. Explain how the strong absorption helps concentrate the energy in a thin layer and thus give greater accuracy in shaping the cornea. Also explain how this strong absorption limits damage to the lens and retina of the eye.

Problems & Exercises

1. (a) Two microwave frequencies are authorized for use in microwave ovens: 900 and 2560 MHz. Calculate the wavelength of each. (b) Which frequency would produce smaller hot spots in foods due to interference effects?
2. (a) Calculate the range of wavelengths for AM radio given its frequency range is 540 to 1600 kHz. (b) Do the same for the FM frequency range of 88.0 to 108 MHz.
3. A radio station utilizes frequencies between commercial AM and FM. What is the frequency of a 11.12-m-wavelength channel?
4. Find the frequency range of visible light, given that it encompasses wavelengths from 380 to 760 nm.
5. Combing your hair leads to excess electrons on the comb. How fast would you have to move the comb up and down to produce red light?
6. Electromagnetic radiation having a 15.0−μm wavelength is classified as infrared radiation. What is its frequency?
7. Approximately what is the smallest detail observable with a microscope that uses ultraviolet light of frequency 1.20×10^{15} Hz?
8. A radar used to detect the presence of aircraft receives a pulse that has reflected off an object 6×10^{-5} s after it was transmitted. What is the distance from the radar station to the reflecting object?
9. Some radar systems detect the size and shape of objects such as aircraft and geological terrain. Approximately what is the smallest observable detail utilizing 500-MHz radar?
10. Determine the amount of time it takes for X-rays of frequency 3×10^{18} Hz to travel (a) 1 mm and (b) 1 cm.
11. If you wish to detect details of the size of atoms (about 1×10^{-10} m) with electromagnetic radiation, it must have a wavelength of about this size. (a) What is its frequency? (b) What type of electromagnetic radiation might this be?
12. If the Sun suddenly turned off, we would not know it until its light stopped coming. How long would that be, given that the Sun is 1.50×10^{11} m away?
13. Distances in space are often quoted in units of light years, the distance light travels in one year. (a) How many meters is a light year? (b) How many meters is it to Andromeda, the nearest large galaxy, given that it is 2.00×10^{6} light years away? (c) The most distant galaxy yet discovered is 12.0×10^{9} light years away. How far is this in meters?
14. A certain 50.0-Hz AC power line radiates an electromagnetic wave having a maximum electric field strength of 13.0 kV/m. (a) What is the wavelength of this very low frequency electromagnetic wave? (b) What is its maximum magnetic field strength?
15. During normal beating, the heart creates a maximum 4.00-mV potential across 0.300 m of a person's chest, creating a 1.00-Hz electromagnetic wave. (a) What is the maximum electric field strength created? (b) What is the corresponding maximum magnetic field strength in the electromagnetic wave? (c) What is the wavelength of the electromagnetic wave?
16. (a) The ideal size (most efficient) for a broadcast antenna with one end on the ground is one-fourth the wavelength $\left(\frac{\lambda}{4}\right)$ of the electromagnetic radiation being sent out. If a new radio station has such an antenna that is 50.0 m high, what frequency does it broadcast most efficiently? Is this

in the AM or FM band? (b) Discuss the analogy of the fundamental resonant mode of an air column closed at one end to the resonance of currents on an antenna that is one-fourth their wavelength.

17. (a) What is the wavelength of 100-MHz radio waves used in an MRI unit? (b) If the frequencies are swept over a±1.00 range centered on 100 MHz, what is the range of wavelengths broadcast?

18. (a) What is the frequency of the 193-nm ultraviolet radiation used in laser eye surgery? (b) Assuming the accuracy with which this EM radiation can ablate the cornea is directly proportional to wavelength, how much more accurate can this UV be than the shortest visible wavelength of light?

19. TV-reception antennas for VHF are constructed with cross wires supported at their centers, as shown in [link]. The ideal length for the cross wires is one-half the wavelength to be received, with the more expensive antennas having one for each channel. Suppose you measure the lengths of the wires for particular channels and find them to be 1.94 and 0.753 m long, respectively. What are the frequencies for these channels?

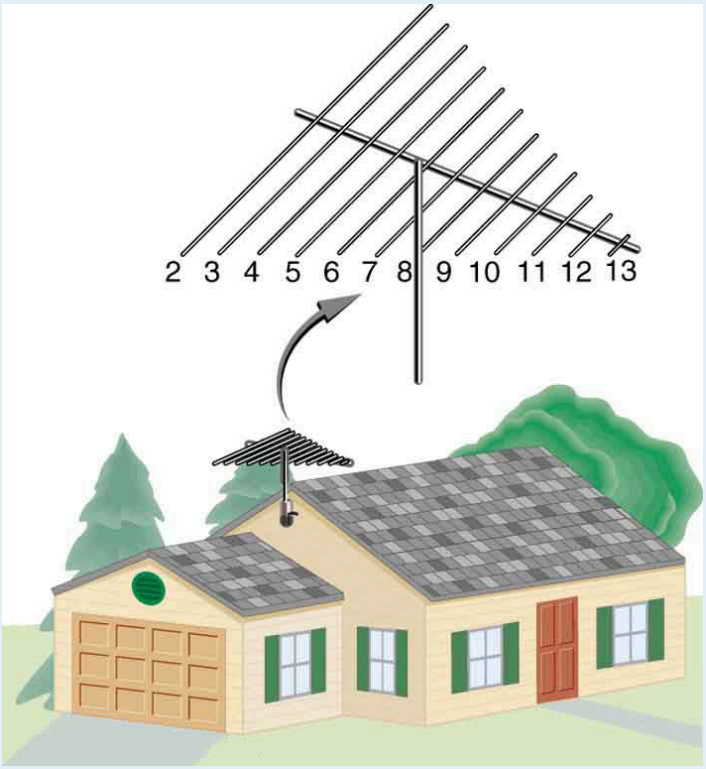


Figure 14. A television reception antenna has cross wires of various lengths to most efficiently receive different wavelengths.

20. Conversations with astronauts on lunar walks had an echo that was used to estimate the distance to the Moon. The sound spoken by the person on Earth was transformed into a radio signal sent to the Moon, and transformed back into sound on a speaker inside the astronaut's space suit. This sound was picked up by the microphone in the space suit (intended for the astronaut's voice) and sent back to Earth as a radio echo of sorts. If the round-trip time was 2.60 s, what was the approximate distance to the Moon, neglecting any delays in the electronics?

21. Lunar astronauts placed a reflector on the Moon's surface, off which a laser beam is periodically reflected. The distance to the Moon is calculated from the round-trip time. (a) To what accuracy

in meters can the distance to the Moon be determined, if this time can be measured to 0.100 ns? (b) What percent accuracy is this, given the average distance to the Moon is 3.84×10^8 m?

22. Radar is used to determine distances to various objects by measuring the round-trip time for an echo from the object. (a) How far away is the planet Venus if the echo time is 1000 s? (b) What is the echo time for a car 75.0 m from a Highway Police radar unit? (c) How accurately (in nanoseconds) must you be able to measure the echo time to an airplane 12.0 km away to determine its distance within 10.0 m?

23. **Integrated Concepts.** (a) Calculate the ratio of the highest to lowest frequencies of electromagnetic waves the eye can see, given the wavelength range of visible light is from 380 to 760 nm. (b) Compare this with the ratio of highest to lowest frequencies the ear can hear.

24. **Integrated Concepts.** (a) Calculate the rate in watts at which heat transfer through radiation occurs (almost entirely in the infrared) from 1.0 m^2 of the Earth's surface at night. Assume the emissivity is 0.90, the temperature of the Earth is 15ºC, and that of outer space is 2.7 K. (b) Compare the intensity of this radiation with that coming to the Earth from the Sun during the day, which averages about 800 W/m^2, only half of which is absorbed. (c) What is the maximum magnetic field strength in the outgoing radiation, assuming it is a continuous wave?

Glossary

electromagnetic spectrum: the full range of wavelengths or frequencies of electromagnetic radiation

radio waves: electromagnetic waves with wavelengths in the range from 1 mm to 100 km; they are produced by currents in wires and circuits and by astronomical phenomena

microwaves: electromagnetic waves with wavelengths in the range from 1 mm to 1 m; they can be produced by currents in macroscopic circuits and devices

thermal agitation: the thermal motion of atoms and molecules in any object at a temperature above absolute zero, which causes them to emit and absorb radiation

radar: a common application of microwaves. Radar can determine the distance to objects as diverse as clouds and aircraft, as well as determine the speed of a car or the intensity of a rainstorm

infrared radiation (IR): a region of the electromagnetic spectrum with a frequency range that extends from just below the red region of the visible light spectrum up to the microwave region, or from 0.74μm to 300μm

ultraviolet radiation (UV): electromagnetic radiation in the range extending upward in frequency from violet light and overlapping with the lowest X-ray frequencies, with wavelengths from 400 nm down to about 10 nm

visible light: the narrow segment of the electromagnetic spectrum to which the normal human eye responds

amplitude modulation (AM): a method for placing information on electromagnetic waves by

modulating the amplitude of a carrier wave with an audio signal, resulting in a wave with constant frequency but varying amplitude

extremely low frequency (ELF): electromagnetic radiation with wavelengths usually in the range of 0 to 300 Hz, but also about 1kHz

carrier wave: an electromagnetic wave that carries a signal by modulation of its amplitude or frequency

frequency modulation (FM): a method of placing information on electromagnetic waves by modulating the frequency of a carrier wave with an audio signal, producing a wave of constant amplitude but varying frequency

TV: video and audio signals broadcast on electromagnetic waves

very high frequency (VHF): TV channels utilizing frequencies in the two ranges of 54 to 88 MHz and 174 to 222 MHz

ultra-high frequency (UHF): TV channels in an even higher frequency range than VHF, of 470 to 1000 MHz

X-ray: invisible, penetrating form of very high frequency electromagnetic radiation, overlapping both the ultraviolet range and the *γ*-ray range

gamma ray: (*γ* ray); extremely high frequency electromagnetic radiation emitted by the nucleus of an atom, either from natural nuclear decay or induced nuclear processes in nuclear reactors and weapons. The lower end of the *γ*-ray frequency range overlaps the upper end of the X-ray range, but *γ* rays can have the highest frequency of any electromagnetic radiation

Selected Solutions to Problems & Exercises

1. (a) 33.3 cm (900 MHz) 11.7 cm (2560 MHz); (b) The microwave oven with the smaller wavelength would produce smaller hot spots in foods, corresponding to the one with the frequency 2560 MHz.

3. 26.96 MHz

5. 5.0×10^{14} Hz

7.

$$\lambda = \frac{c}{f} = \frac{3.00 \times 10^8 \text{ m/s}}{1.20 \times 10^{15} \text{ Hz}} = 2.50 \times 10^{-7} \text{ m}$$

9. 0.600 m

11. (a)

$$f = \frac{c}{\lambda} = \frac{3.00 \times 10^8 \text{ m/s}}{1 \times 10^{-10} \text{ m}} = 3 \times 10^{18} \text{ Hz}$$

; (b) X-rays

14. (a) 6.00×10^6 m; (b) 4.33×10^{-5} T

16. (a) 1.50 × 10 6 Hz, AM band; (b) The resonance of currents on an antenna that is 1/4 their wavelength is analogous to the fundamental resonant mode of an air column closed at one end, since the tube also has a length equal to 1/4 the wavelength of the fundamental oscillation.

18. (a) 1.55×10^{15} Hz; (b) The shortest wavelength of visible light is 380 nm, so that

$$\begin{aligned} & \frac{\lambda_{\text{visible}}}{\lambda_{\text{UV}}} \\ = & \frac{380 \text{ nm}}{193 \text{ nm}} \\ = & 1.97 \end{aligned}$$

In other words, the UV radiation is 97% more accurate than the shortest wavelength of visible light, or almost twice as accurate!

20. 3.90×10^{8} m

22. (a) 1.50×10^{11} m; (b) 0.500 μs; (c) 66.7 ns

24. (a) -3.5×10^{2} W/m^2; (b) 88%; (c) 1.7 μT

Energy in Electromagnetic Waves

Lumen Learning

Learning Objectives

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

- Explain how the energy and amplitude of an electromagnetic wave are related.
- Given its power output and the heating area, calculate the intensity of a microwave oven's electromagnetic field, as well as its peak electric and magnetic field strengths

Anyone who has used a microwave oven knows there is energy in *electromagnetic waves*. Sometimes this energy is obvious, such as in the warmth of the summer sun. Other times it is subtle, such as the unfelt energy of gamma rays, which can destroy living cells.

Electromagnetic waves can bring energy into a system by virtue of their *electric and magnetic fields*. These fields can exert forces and move charges in the system and, thus, do work on them. If the frequency of the electromagnetic wave is the same as the natural frequencies of the system (such as microwaves at the resonant frequency of water molecules), the transfer of energy is much more efficient.

Making Connections: Waves and Particles

The behavior of electromagnetic radiation clearly exhibits wave characteristics. But we shall find in later modules that at high frequencies, electromagnetic radiation also exhibits particle characteristics. These particle characteristics will be used to explain more of the properties of the electromagnetic spectrum and to introduce the formal study of modern physics.

Another startling discovery of modern physics is that particles, such as electrons and protons, exhibit wave characteristics. This simultaneous sharing of wave and particle properties for all submicroscopic entities is one of the great symmetries in nature.

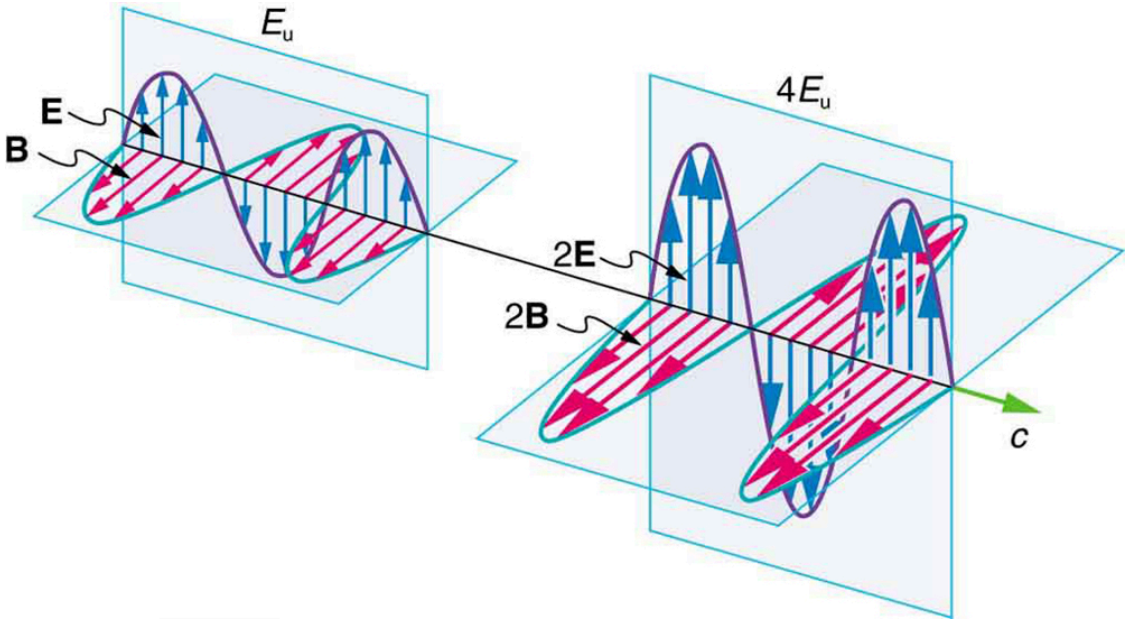


Figure 1. Energy carried by a wave is proportional to its amplitude squared. With electromagnetic waves, larger E-fields and B-fields exert larger forces and can do more work.

But there is energy in an electromagnetic wave, whether it is absorbed or not. Once created, the fields carry energy away from a source. If absorbed, the field strengths are diminished and anything left travels on. Clearly, the larger the strength of the electric and magnetic fields, the more work they can do and the greater the energy the electromagnetic wave carries.

A wave's energy is proportional to its *amplitude* squared (E^2 or B^2). This is true for waves on guitar strings, for water waves, and for sound waves, where amplitude is proportional to pressure. In electromagnetic waves, the amplitude is the *maximum field strength* of the electric and magnetic fields. (See Figure 1.)

Thus the energy carried and the *intensity I* of an electromagnetic wave is proportional to E^2 and B^2. In fact, for a continuous sinusoidal electromagnetic wave, the average intensity I_{ave} is given by

$$I_{\text{ave}} = \frac{c\epsilon_0 E_0^2}{2}$$

where c is the speed of light, ε_0 is the permittivity of free space, and E_0 is the maximum electric field strength; intensity, as always, is power per unit area (here in W/m^2).

The average intensity of an electromagnetic wave I_{ave} can also be expressed in terms of the magnetic field strength by using the relationship

$$B = \frac{E}{c}$$

, and the fact that

$$\epsilon_0 = \frac{1}{\mu_0 c^2}$$

, where μ_0 is the permeability of free space. Algebraic manipulation produces the relationship

$$I_{\text{ave}} = \frac{cB_0^2}{2\mu_0}$$

,

where B_0 is the maximum magnetic field strength.

One more expression for I_{ave} in terms of both electric and magnetic field strengths is useful. Substituting the fact that $c \cdot B_0 = E_0$, the previous expression becomes

$$I_{\text{ave}} = \frac{E_0 B_0^2}{2\mu_0}$$

.

Whichever of the three preceding equations is most convenient can be used, since they are really just different versions of the same principle: Energy in a wave is related to amplitude squared. Furthermore, since these equations are based on the assumption that the electromagnetic waves are sinusoidal, peak intensity is twice the average; that is, $I_0 = 2I_{\text{ave}}$.

Example 1. Calculate Microwave Intensities and Fields

On its highest power setting, a certain microwave oven projects 1.00 kW of microwaves onto a 30.0 by 40.0 cm area.

1. What is the intensity in W/m^2?
2. Calculate the peak electric field strength E_0 in these waves.
3. What is the peak magnetic field strength B_0?

Strategy

In Part 1, we can find intensity from its definition as power per unit area. Once the intensity is known, we can use the equations below to find the field strengths asked for in Parts 2 and 3.

Solution for Part 1

Entering the given power into the definition of intensity, and noting the area is 0.300 by 0.400 m, yields

$$I = \frac{P}{A} = \frac{1.00 \text{ kW}}{0.300 \text{ m} \times 0.400 \text{ m}}$$

Here $I = I_{\text{ave}}$, so that

$$I_{\text{ave}} = \frac{1000 \text{ W}}{0.120 \text{ m}^2} = 8.33 \times 10^3 \text{ W/m}^2$$

Note that the peak intensity is twice the average: $I_0 = 2I_{\text{ave}} = 1.67 \times 10^4$ W/m^2.

Solution for Part 2

To find E_0, we can rearrange the first equation given above for I_{ave} to give

$$E_0 = \left(\frac{2I_{\text{ave}}}{c\epsilon_0}\right)^{1/2}$$

Entering known values gives

$$\begin{aligned} E_0 &= \sqrt{\frac{2\left(8.33\times10^3\ \text{W/m}^2\right)}{(3.00\times10^8\ \text{m/s})\left(8.85\times10^{-12}\ \text{C}^2/\text{N}\cdot\text{m}^2\right)}} \\ &= 2.51\times10^3\ \text{V/m} \end{aligned}$$

Solution for Part 3

Perhaps the easiest way to find magnetic field strength, now that the electric field strength is known, is to use the relationship given by

$$B_0 = \frac{E_0}{c}$$

.

Entering known values gives

$$\begin{aligned} B_0 &= \frac{2.51\times10^3\ \text{V/m}}{3.0\times10^8\ \text{m/s}} \\ &= 8.35\times10^{-6}\ \text{T} \end{aligned}$$

Discussion

As before, a relatively strong electric field is accompanied by a relatively weak magnetic field in an electromagnetic wave, since

$$B = \frac{E}{c}$$

, and c is a large number.

Section Summary

- The energy carried by any wave is proportional to its amplitude squared. For electromagnetic waves, this means intensity can be expressed as $I_{\text{ave}} = \frac{c\epsilon_0 E_0^2}{2}$, where I_{ave} is the average intensity in W/m^2, and E_0 is the maximum electric field strength of a continuous sinusoidal wave.
- This can also be expressed in terms of the maximum magnetic field strength B_0 as $I_{\text{ave}} = \frac{cB_0^2}{2\mu_0}$ and in terms of both electric and magnetic fields as $I_{\text{ave}} = \frac{E_0 B_0}{2\mu_0}$.

- The three expressions for I_{ave} are all equivalent.

Problems & Exercises

1. What is the intensity of an electromagnetic wave with a peak electric field strength of 125 V/m?
2. Find the intensity of an electromagnetic wave having a peak magnetic field strength of 4.00×10^{-9} T.
3. Assume the helium-neon lasers commonly used in student physics laboratories have power outputs of 0.250 mW. (a) If such a laser beam is projected onto a circular spot 1.00 mm in diameter, what is its intensity? (b) Find the peak magnetic field strength. (c) Find the peak electric field strength.
4. An AM radio transmitter broadcasts 50.0 kW of power uniformly in all directions. (a) Assuming all of the radio waves that strike the ground are completely absorbed, and that there is no absorption by the atmosphere or other objects, what is the intensity 30.0 km away? (Hint: Half the power will be spread over the area of a hemisphere.) (b) What is the maximum electric field strength at this distance?
5. Suppose the maximum safe intensity of microwaves for human exposure is taken to be 1.00 W/m2. (a) If a radar unit leaks 10.0 W of microwaves (other than those sent by its antenna) uniformly in all directions, how far away must you be to be exposed to an intensity considered to be safe? Assume that the power spreads uniformly over the area of a sphere with no complications from absorption or reflection. (b) What is the maximum electric field strength at the safe intensity? (Note that early radar units leaked more than modern ones do. This caused identifiable health problems, such as cataracts, for people who worked near them.)
6. A 2.50-m-diameter university communications satellite dish receives TV signals that have a maximum electric field strength (for one channel) of 7.50 μV/m. (See Figure 2.) (a) What is the intensity of this wave? (b) What is the power received by the antenna? (c) If the orbiting satellite broadcasts uniformly over an area of 1.50×10^{13} m^2 (a large fraction of North America), how much power does it radiate?

Figure 2. Satellite dishes receive TV signals sent from orbit. Although the signals are quite weak, the receiver can detect them by being tuned to resonate at their frequency.

7. Lasers can be constructed that produce an extremely high intensity electromagnetic wave for a brief time—called pulsed lasers. They are used to ignite nuclear fusion, for example. Such a laser may produce an electromagnetic wave with a maximum electric field strength of 1.00×10^{11} V/m for a time of 1.00 ns. (a) What is the maximum magnetic field strength in the wave? (b) What is the intensity of the beam? (c) What energy does it deliver on a 1.00-mm^2 area?

8. Show that for a continuous sinusoidal electromagnetic wave, the peak intensity is twice the average intensity ($I_0 = 2I_{ave}$), using either the fact that $E_0 = \sqrt{2}E_{rms}$ or $B_0 = \sqrt{2}B_{rms}$, where rms means average (actually root mean square, a type of average).

9. Suppose a source of electromagnetic waves radiates uniformly in all directions in empty space where there are no absorption or interference effects. (a) Show that the intensity is inversely proportional to r^2, the distance from the source squared. (b) Show that the magnitudes of the electric and magnetic fields are inversely proportional to r.

10. **Integrated Concepts.** An *LC* circuit with a 5.00-pF capacitor oscillates in such a manner as to radiate at a wavelength of 3.30 m. (a) What is the resonant frequency? (b) What inductance is in

series with the capacitor?

11. **Integrated Concepts.** What capacitance is needed in series with an 800-μH inductor to form a circuit that radiates a wavelength of 196 m?

12. **Integrated Concepts.** Police radar determines the speed of motor vehicles using the same Doppler-shift technique employed for ultrasound in medical diagnostics. Beats are produced by mixing the double Doppler-shifted echo with the original frequency. If 1.50×10^9-Hz microwaves are used and a beat frequency of 150 Hz is produced, what is the speed of the vehicle? (Assume the same Doppler-shift formulas are valid with the speed of sound replaced by the speed of light.)

13. **Integrated Concepts.**Assume the mostly infrared radiation from a heat lamp acts like a continuous wave with wavelength 1.50μm. (a) If the lamp's 200-W output is focused on a person's shoulder, over a circular area 25.0 cm in diameter, what is the intensity in W/m^2? (b) What is the peak electric field strength? (c) Find the peak magnetic field strength. (d) How long will it take to increase the temperature of the 4.00-kg shoulder by 2.00º C, assuming no other heat transfer and given that its specific heat is 3.47×10^3 J/kg·ºC?

14. **Integrated Concepts.** On its highest power setting, a microwave oven increases the temperature of 0.400 kg of spaghetti by 45.0ºC in 120 s. (a) What was the rate of power absorption by the spaghetti, given that its specific heat is 3.76×10^3 J/kg · ºC? (b) Find the average intensity of the microwaves, given that they are absorbed over a circular area 20.0 cm in diameter. (c) What is the peak electric field strength of the microwave? (d) What is its peak magnetic field strength?

15. **Integrated Concepts.** Electromagnetic radiation from a 5.00-mW laser is concentrated on a 1.00-mm^2 area. (a) What is the intensity in W/m^2? (b) Suppose a 2.00-nC static charge is in the beam. What is the maximum electric force it experiences? (c) If the static charge moves at 400 m/s, what maximum magnetic force can it feel?

16. **Integrated Concepts.** A 200-turn flat coil of wire 30.0 cm in diameter acts as an antenna for FM radio at a frequency of 100 MHz. The magnetic field of the incoming electromagnetic wave is perpendicular to the coil and has a maximum strength of 1.00×10^{-12} T. (a) What power is incident on the coil? (b) What average emf is induced in the coil over one-fourth of a cycle? (c) If the radio receiver has an inductance of 2.50 μH, what capacitance must it have to resonate at 100 MHz?

17. **Integrated Concepts.** If electric and magnetic field strengths vary sinusoidally in time, being zero at $t = 0$, then $E = E_0 \sin 2\pi ft$ and $B = B_0 \sin 2\pi ft$. Let $f = 1.00$ GHz here. (a) When are the field strengths first zero? (b) When do they reach their most negative value? (c) How much time is needed for them to complete one cycle?

18. **Unreasonable Results.** A researcher measures the wavelength of a 1.20-GHz electromagnetic wave to be 0.500 m. (a) Calculate the speed at which this wave propagates. (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

19. **Unreasonable Results.** The peak magnetic field strength in a residential microwave oven is 9.20×10^{-5} T. (a) What is the intensity of the microwave? (b) What is unreasonable about this result? (c) What is wrong about the premise?

20. **Unreasonable Results.** An LC circuit containing a 2.00-H inductor oscillates at such a frequency that it radiates at a 1.00-m wavelength. (a) What is the capacitance of the circuit? (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

21. **Unreasonable Results.** An LC circuit containing a 1.00-pF capacitor oscillates at such a frequency that it radiates at a 300-nm wavelength. (a) What is the inductance of the circuit? (b)

What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

22. **Create Your Own Problem.** Consider electromagnetic fields produced by high voltage power lines. Construct a problem in which you calculate the intensity of this electromagnetic radiation in W/m2 based on the measured magnetic field strength of the radiation in a home near the power lines. Assume these magnetic field strengths are known to average less than a μT. The intensity is small enough that it is difficult to imagine mechanisms for biological damage due to it. Discuss how much energy may be radiating from a section of power line several hundred meters long and compare this to the power likely to be carried by the lines. An idea of how much power this is can be obtained by calculating the approximate current responsible for μT fields at distances of tens of meters.

23. **Create Your Own Problem.** Consider the most recent generation of residential satellite dishes that are a little less than half a meter in diameter. Construct a problem in which you calculate the power received by the dish and the maximum electric field strength of the microwave signals for a single channel received by the dish. Among the things to be considered are the power broadcast by the satellite and the area over which the power is spread, as well as the area of the receiving dish.

Glossary

maximum field strength: the maximum amplitude an electromagnetic wave can reach, representing the maximum amount of electric force and/or magnetic flux that the wave can exert

intensity: the power of an electric or magnetic field per unit area, for example, Watts per square meter

Selected Solutions to Problems & Exercises

1.

$$\begin{aligned} I &= \frac{c\epsilon_0 E_0^2}{2} \\ &= \frac{\left(3.00\times10^8\text{m/s}\right)\left(8.85\times10^{-12}\text{C}^2/\text{N}\cdot\text{m}^2\right)(125\text{ V/m})^2}{2} \\ &= 20.7\text{ W/m}^2 \end{aligned}$$

3. (a)

$$I = \frac{P}{A} = \frac{P}{\pi r^2} = \frac{0.250\times10^{-3}\text{W}}{\pi\left(0.500\times10^{-3}\text{m}\right)^2} = 318\text{ W/m}^2$$

(b)

$$\begin{aligned} I_{\text{ave}} &= \frac{cB_0^2}{2\mu_0} \Rightarrow B_0 = \left(\frac{2\mu_0 I}{c}\right)^{1/2} \\ &= \left(\frac{2\left(4\pi\times10^{-7}\text{T}\cdot\text{m/A}\right)\left(318.3\text{ W/m}^2\right)}{3.00\times10^8\text{m/s}}\right)^{1/2} \\ &= 1.63\times10^{-6}\text{T} \end{aligned}$$

(c)

$$\begin{aligned} E_0 &= \mathrm{cB}_0 = \left(3.00 \times 10^8 \mathrm{m/s}\right)\left(1.633 \times 10^{-6} \mathrm{T}\right) \\ &= 4.90 \times 10^2 \mathrm{V/m} \end{aligned}$$

5. (a) 89.2 cm; (b) 27.4 V/m

7. (a) 333 T; (b) 1.33×10^{19} W/m^2; (c) 13.3 kJ

9. (a)

$$I = \frac{P}{A} = \frac{P}{4\pi r^2} \propto \frac{1}{r^2}$$

; (b)

$$I \propto {E_0}^2, {B_0}^2 \Rightarrow {E_0}^2, {B_0}^2 \propto \frac{1}{r^2} \Rightarrow E_0, B_0 \propto \frac{1}{r}$$

11. 13.5 pF

13. (a) 4.07 kW/m^2; (b) 1.75 kV/m; (c) 5.84 μT; (d) 2 min 19 s

15. (a) 5.00×10^3 W/m^2; (b) 3.88×10^{-6} N; (c) 5.18×10^{-12} N

17. (a) $t = 0$; (b) 7.50×10^{-10} s; (c) 1.00×10^{-9} s

19. (a) 1.01×10^6 W/m^2; (b) Much too great for an oven; (c) The assumed magnetic field is unreasonably large.

21. (a) 2.53×10^{-20} H; (b) L is much too small; (c) The wavelength is unreasonably small.

9. Geometric Optics

Introduction to Geometric Optics

Lumen Learning

Figure 1. Image seen as a result of reflection of light on a plane smooth surface. (credit: NASA Goddard Photo and Video, via Flickr)

Light from this page or screen is formed into an image by the lens of your eye, much as the lens of the camera that made this photograph. Mirrors, like lenses, can also form images that in turn are captured by your eye.

Our lives are filled with light. Through vision, the most valued of our senses, light can evoke spiritual emotions, such as when we view a magnificent sunset or glimpse a rainbow breaking through the clouds. Light can also simply amuse us in a theater, or warn us to stop at an intersection. It has innumerable uses beyond vision. Light can carry telephone signals through glass fibers or cook a meal in a solar oven. Life itself could not exist without light’s energy. From photosynthesis in plants to the sun warming a cold-blooded animal, its supply of energy is vital.

Figure 2. Double Rainbow over the bay of Pocitos in Montevideo, Uruguay. (credit: Madrax, Wikimedia Commons)

We already know that visible light is the type of electromagnetic waves to which our eyes respond. That knowledge still leaves many questions regarding the nature of light and vision. What is color, and how do our eyes detect it? Why do diamonds sparkle? How does light travel? How do lenses and mirrors form images? These are but a few of the questions that are answered by the study of optics. Optics is the branch of physics that deals with the behavior of visible light and other electromagnetic waves. In particular, optics is concerned with the generation and propagation of light and its interaction with matter. What we have already learned about the generation of light in our study of heat transfer by radiation will be expanded upon in later topics, especially those on atomic physics. Now, we will concentrate on the propagation of light and its interaction with matter.

It is convenient to divide optics into two major parts based on the size of objects that light encounters. When light interacts with an object that is several times as large as the light's wavelength, its observable behavior is like that of a ray; it does not prominently display its wave characteristics. We call this part of optics "geometric optics." This chapter will concentrate on such situations. When light interacts with smaller objects, it has very prominent wave characteristics, such as constructive and destructive interference. Wave Optics will concentrate on such situations.

The Ray Aspect of Light

Lumen Learning

Learning Objectives

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

- List the ways by which light travels from a source to another location.

There are three ways in which light can travel from a source to another location. (See Figure 1.) 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 person. Light can also arrive after being reflected, such as by a mirror. In all of these cases, light is modeled as traveling in straight lines called rays. 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 (or even science fiction depictions of ray guns).

Ray

The word “ray” comes from mathematics and here means a straight line that originates at some point.

Experiments, as well as our own experiences, show that when light interacts with objects several times as large as 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 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 light encounters anything we can observe with unaided eyes, such as a mirror, it acts like a ray, with only subtle wave characteristics. We will concentrate on the ray characteristics in this chapter.

Since light moves in straight lines, changing directions when it interacts with materials, it is described by geometry and simple trigonometry. This part of optics, where the ray aspect of light dominates, is therefore called *geometric optics*. There are two laws that 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.

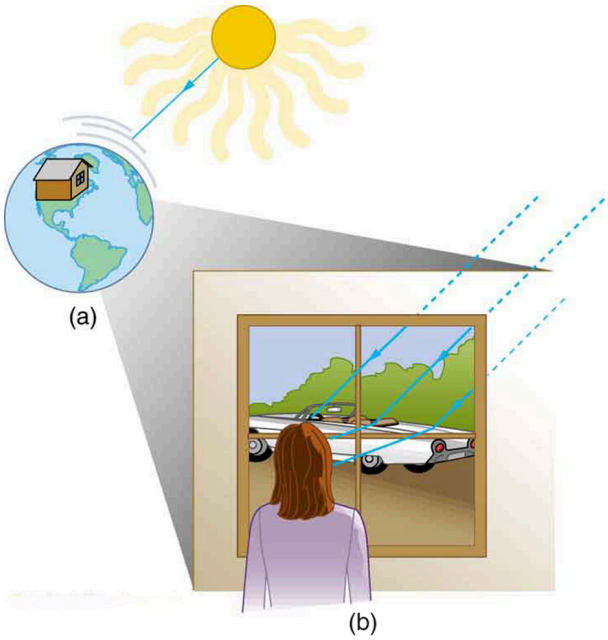


Figure 1. Three methods for light to travel from a source to another location. (a) Light reaches the upper atmosphere of Earth traveling through empty space directly from the source. (b) Light can reach a person in one of two ways. It can travel through media like air and glass. It can also reflect from 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.

Geometric Optics

The part of optics dealing with the ray aspect of light is called geometric optics.

Section Summary

A straight line that originates at some point is called a ray.
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; (3) after being reflected from a mirror.

Problems & Exercises

Suppose a man stands in front of a mirror as shown in Figure 2. 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?

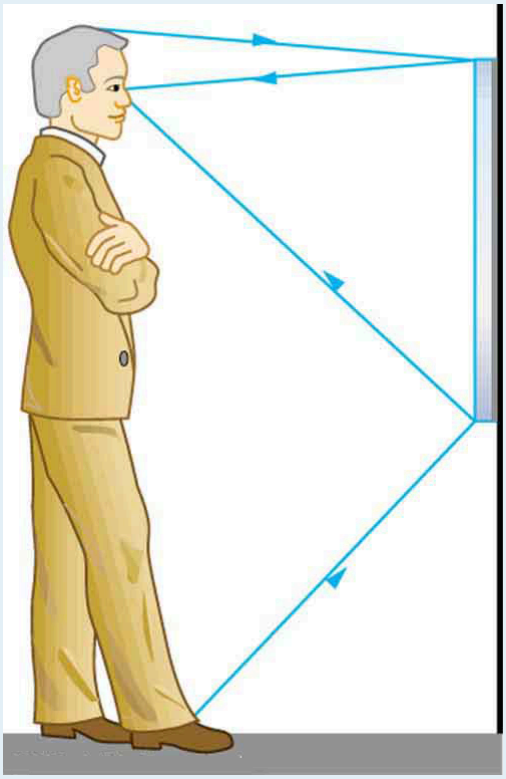


Figure 2. A man standing in front of a mirror on a wall at a distance of several feet. The mirror's top is at eye level, but its bottom is only waist high. Arrows illustrate how the man can see his reflection from head to toe in the mirror.

Glossary

ray: straight line that originates at some point

geometric optics: part of optics dealing with the ray aspect of light

Solution to Problems & Exercises

Top 1.715 m from floor, bottom 0.825 m from floor. Height of mirror is 0.890 m, or precisely one-half the height of the person.

The Law of Reflection

Lumen Learning

Learning Objectives

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

- Explain reflection of light from polished and rough surfaces.

Whenever we look into a mirror, or squint at sunlight glinting from a lake, we are seeing a reflection. When you look at this page, too, you are seeing light reflected from it. Large telescopes use reflection to form an image of stars and other astronomical objects.

The law of reflection is illustrated in Figure 1, which also shows how the angles are measured relative to the perpendicular to the surface at the point where the light ray strikes. We expect to see reflections from smooth surfaces, but Figure 2 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 illustrated in Figure 3. Many objects, such as people, clothing, leaves, and walls, 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 4. When the moon reflects from a lake, as shown in Figure 5, a combination of these effects takes place.

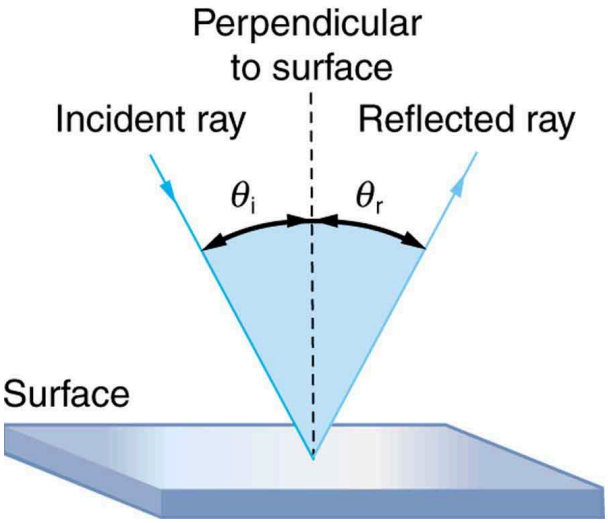


Figure 1. The law of reflection states that the angle of reflection equals the angle of incidence—θr = θi. The angles are measured relative to the perpendicular to the surface at the point where the ray strikes the surface.

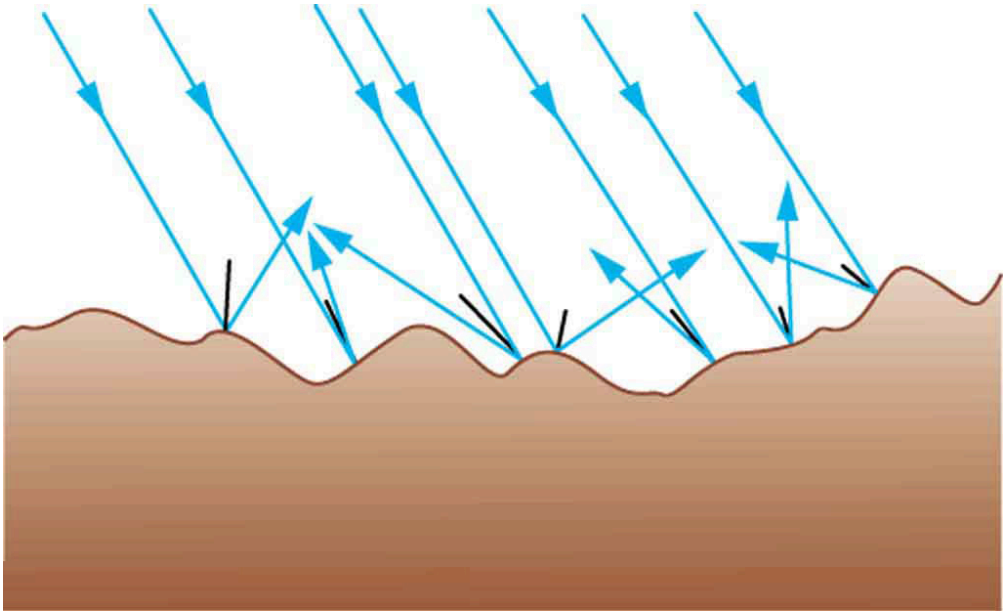


Figure 2. Light is diffused when it reflects from a rough surface. Here many parallel rays are incident, but they are reflected at many different angles since the surface is rough.

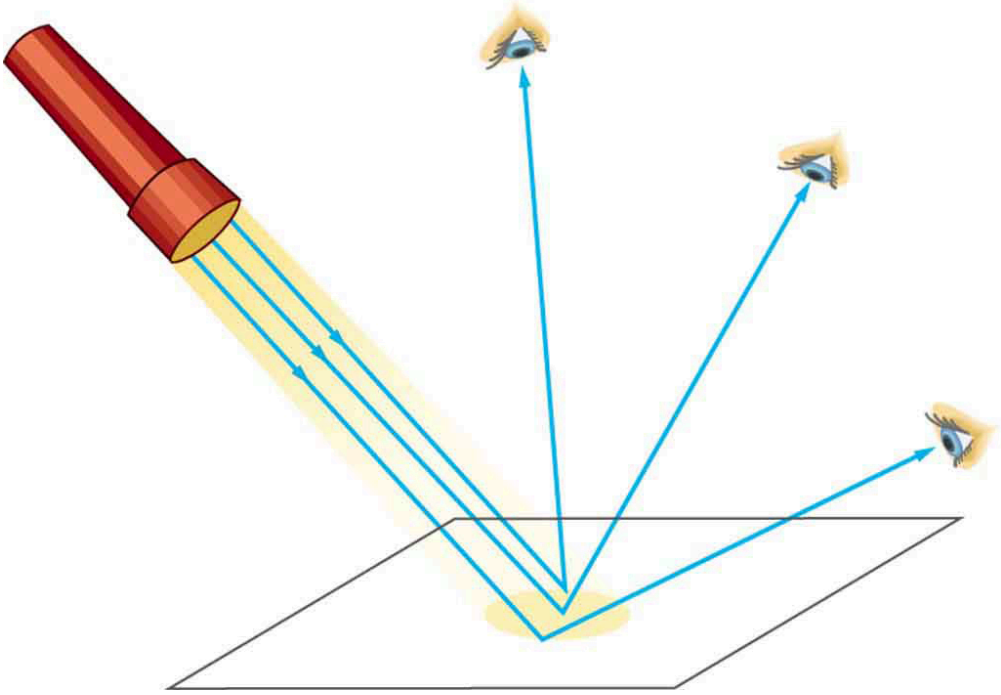


Figure 3. When a sheet of paper is illuminated with many parallel incident rays, it can be seen at many different angles, because its surface is rough and diffuses the light.

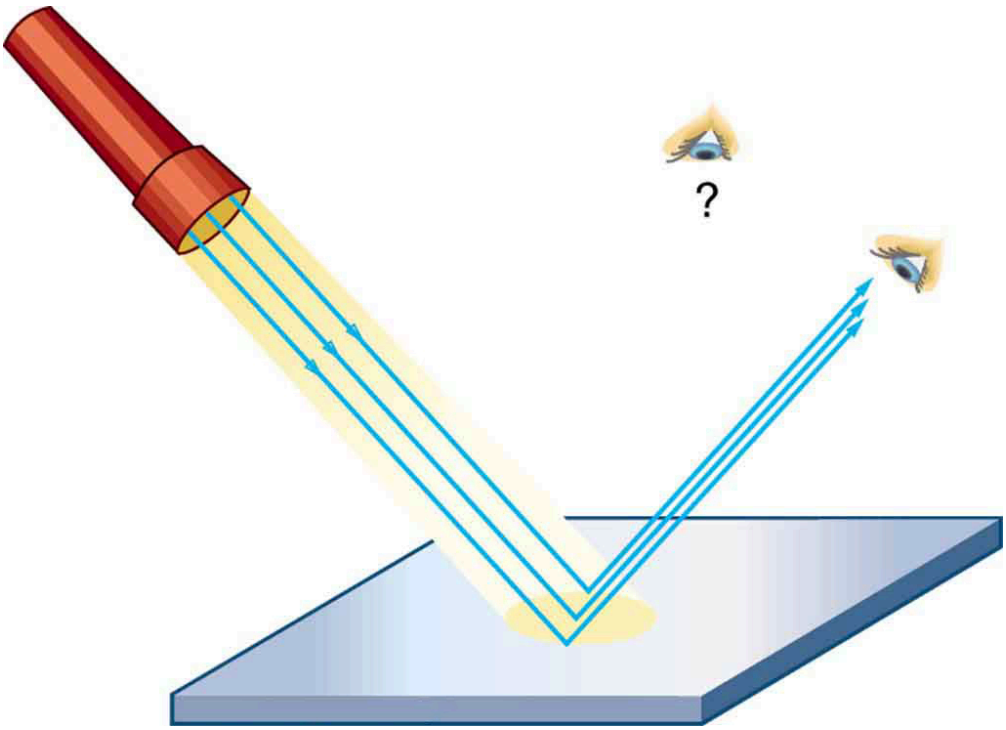


Figure 4. A mirror illuminated by many parallel rays reflects them in only one direction, since its surface is very smooth. Only the observer at a particular angle will see the reflected light.

Figure 5. Moonlight is spread out when it is reflected by the lake, since the surface is shiny but uneven. (credit: Diego Torres Silvestre, Flickr)

The law of reflection is very simple: The angle of reflection equals the angle of incidence.

The Law of Reflection

The angle of reflection equals the angle of incidence.

When we see ourselves in a mirror, it appears that our image is actually behind the mirror. This is illustrated in Figure 6. We see the light coming from a direction determined by the law of reflection. The angles are such that our image is exactly the same distance behind the mirror as we stand away from 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 our imagination. Mirror images can be photographed and videotaped by instruments and look just as they do with our eyes (optical instruments themselves). The precise manner in which images are formed by mirrors and lenses will be treated in later sections of this chapter.

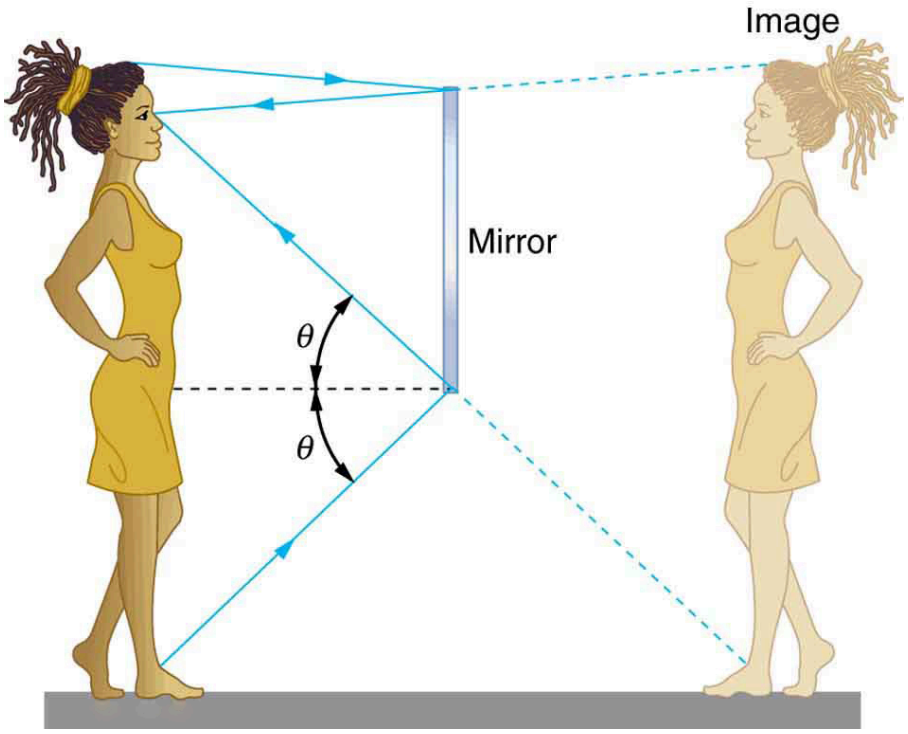


Figure 6. Our image in a mirror is behind the mirror. The two rays shown are those that strike the mirror at just the correct angles to be reflected into the eyes of the person. The image appears to be in the direction the rays are coming from when they enter the eyes.

Take-Home Experiment: Law of Reflection

Take a piece of paper and shine a flashlight at an angle at the paper, as shown in Figure 3. Now shine the

flashlight at a mirror at an angle. Do your observations confirm the predictions in Figure 3 and Figure 4? Shine the flashlight on various surfaces and determine whether the reflected light is diffuse or not. You can choose a shiny metallic lid of a pot or your skin. Using the mirror and flashlight, can you confirm the law of reflection? You will need to draw lines on a piece of paper showing the incident and reflected rays. (This part works even better if you use a laser pencil.)

Section Summary

- 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.
- Mirror images can be photographed and videotaped by instruments.

Conceptual Question

1. 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?

Problems & Exercises

1. 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 in the following figure.

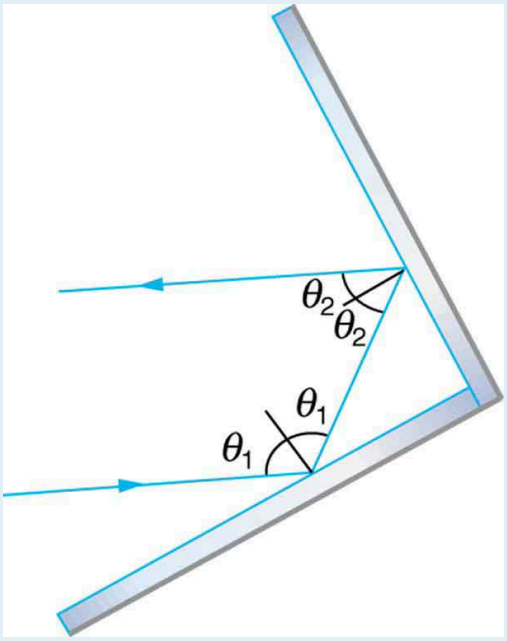


Figure 7. A corner reflector sends the reflected ray back in a direction parallel to the incident ray, independent of incoming direction.

2. 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.
3. 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. Show that after striking a plane mirror, the angle between their directions remains θ.

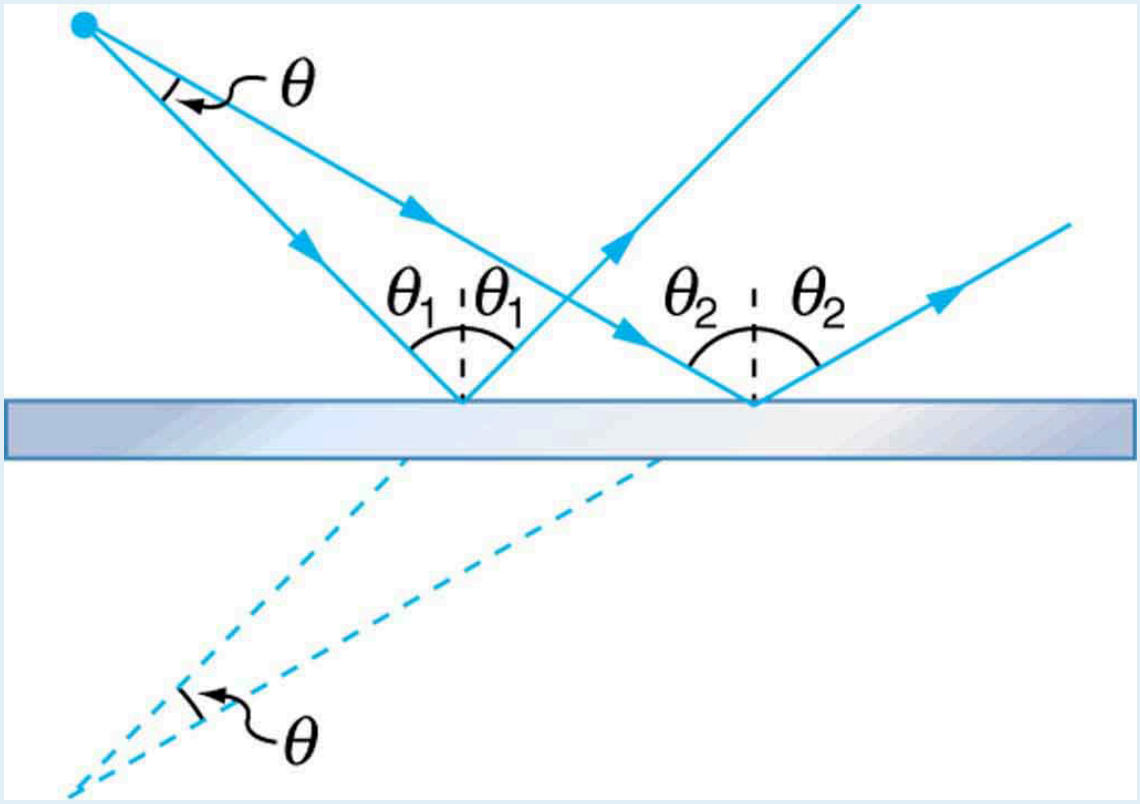


Figure 8. A flat mirror neither converges nor diverges light rays. Two rays continue to diverge at the same angle after reflection.

Glossary

mirror: smooth surface that reflects light at specific angles, forming an image of the person or object in front of it

law of reflection: angle of reflection equals the angle of incidence

The Law of Refraction

Lumen Learning

Learning Objective

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

- Determine the index of refraction, given the speed of light in a medium.

It is easy to 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. (See Figure 1.) This is because light coming from the fish to us changes direction when it leaves the tank, and in this case, it can travel two different paths to get to our eyes. The changing of a light ray's direction (loosely called bending) when it passes through variations in matter is called *refraction*. Refraction is responsible for a tremendous range of optical phenomena, from the action of lenses to voice transmission through optical fibers.

Refraction

The changing of a light ray's direction (loosely called bending) when it passes through variations in matter is called refraction.

Speed of Light

The speed of light *c* not only affects refraction, it is one of the central concepts of Einstein's theory of relativity. As the accuracy of the measurements of the speed of light were improved, *c* was found not to depend on the velocity of the source or the observer. 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 the chapter Special Relativity. It makes connections between space and time and alters our expectations that all observers measure the same time for the same event, for example. The speed of light is so important that its value in a vacuum is one of the most fundamental constants in nature as well as being one of the four fundamental SI units.

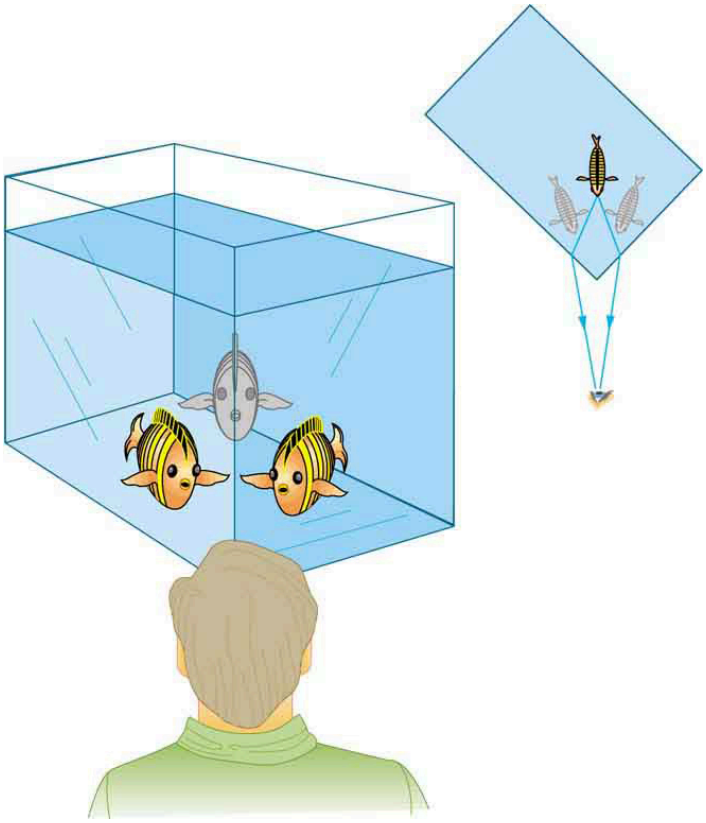


Figure 1. Looking at the fish tank as shown, we can see the same fish in two different locations, because light changes directions when it passes from water to air. In this case, the light can reach the observer by two different paths, and so the fish seems to be in two different places. This bending of light is called refraction and is responsible for many optical phenomena.

Why does light change direction when passing from one material (medium) to another? It is because light changes speed when going from one material to another. So before we study the law of refraction, it is useful to discuss the speed of light and how it varies in different media.

The Speed of Light

Early attempts to measure the speed of light, such as those made by Galileo, determined that light moved extremely fast, perhaps instantaneously. The first real evidence that light traveled at a finite speed came from the Danish astronomer Ole Roemer in the late 17th century. Roemer had noted that the average orbital period of one of Jupiter's moons, as measured from Earth, varied depending on whether Earth was moving toward or away from Jupiter. He correctly concluded that the apparent change in period was due to the change in distance between Earth and Jupiter and the time it took light to travel this distance. From his 1676 data, a value of the speed of light was calculated to be 2.26×10^8 m/s (only 25% different from today's accepted value). In more recent times, physicists have measured the speed of light in numerous ways and with increasing accuracy. One particularly direct method, used in 1887 by the American physicist Albert Michelson (1852–1931), is illustrated in Figure 2. Light reflected from a rotating set of mirrors was reflected from a stationary mirror 35 km away and returned to the rotating

mirrors. The time for the light to travel can be determined by how fast the mirrors must rotate for the light to be returned to the observer's eye.

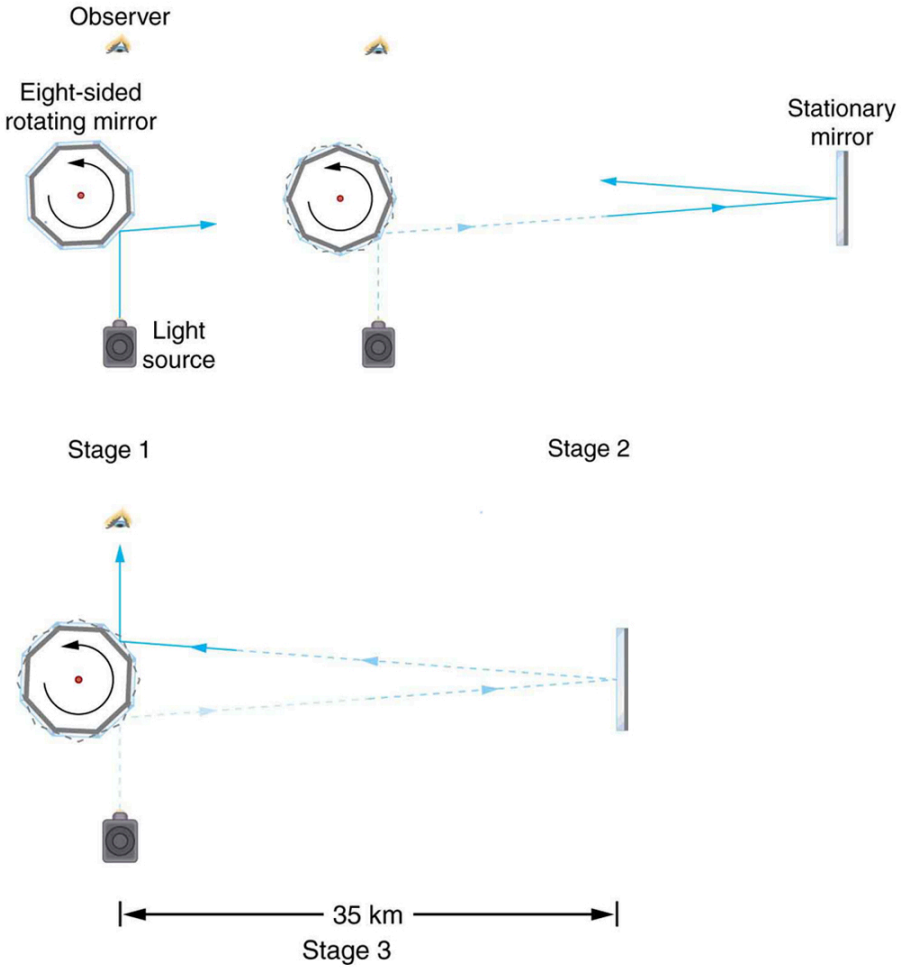


Figure 2. A schematic of early apparatus used by Michelson and others to determine the speed of light. As the mirrors rotate, the reflected ray is only briefly directed at the stationary mirror. The returning ray will be reflected into the observer's eye only if the next mirror has rotated into the correct position just as the ray returns. By measuring the correct rotation rate, the time for the round trip can be measured and the speed of light calculated. Michelson's calculated value of the speed of light was only 0.04% different from the value used today.

The speed of light is now 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 fixed value $c = 2.9972458 \times 10^8$ m/s $\approx 3.00 \times 10^8$ m/s, where the approximate value of 3.00×10^8 m/s is used whenever three-digit accuracy is sufficient. 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 with different atoms, crystal lattices, and other substructures varies. We define the *index of refraction n* of a material to be $n = \frac{c}{v}$, 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.

Value of the Speed of Light

$c = 2.9972458 \times 10^8$ m/s ≈ 3.00×10^8 m/s

Index of Refraction

$$n = \frac{c}{v}$$

That is, $n \geq 1$. Table 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 produced by a prism.) 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.

Table 1. Index of Refraction in Various Media

Medium	***n***
Gases at 0ºC, 1 atm	
Air	1.000293
Carbon dioxide	1.00045
Hydrogen	1.000139
Oxygen	1.000271
Liquids at 20ºC	
Benzene	1.501
Carbon disulfide	1.628
Carbon tetrachloride	1.461
Ethanol	1.361
Glycerine	1.473
Water, fresh	1.333
Solids at 20ºC	
Diamond	2.419
Fluorite	1.434
Glass, crown	1.52
Glass, flint	1.66
Ice at 20ºC	1.309
Polystyrene	1.49
Plexiglas	1.51
Quartz, crystalline	1.544
Quartz, fused	1.458
Sodium chloride	1.544
Zircon	1.923

Example 1. Speed of Light in Matter

Calculate the speed of light in zircon, a material used in jewelry to imitate diamond.

Strategy

The speed of light in a material, *v*, can be calculated from the index of refraction *n* of the material using the equation

$$n = \frac{c}{v}$$

.

Solution

The equation for index of refraction states that

$$n = \frac{c}{v}$$

. Rearranging this to determine *v* gives

$$v = \frac{c}{n}$$

.

The index of refraction for zircon is given as 1.923 in Table 1, and *c* is given in the equation for speed of light. Entering these values in the last expression gives

$$\begin{aligned} v &= \frac{3.00 \times 10^8 \text{ m/s}}{1.923} \\ &= 1.56 \times 10^8 \text{ m/s} \end{aligned}$$

Discussion

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 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.

Law of Refraction

Figure 3 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 will be reflected from the surface, but for now we will concentrate on the light that is transmitted.) The change in direction of the light ray depends on how the speed of light changes. The change in the speed of light is related to the indices of refraction of the media involved. In the situations shown in Figure 3, medium 2 has a greater index of refraction than medium 1. This means that the speed of light is less in medium 2 than in medium 1. Note that as shown in Figure 3a, the direction of the ray moves closer to the perpendicular when it slows down. Conversely, as shown in Figure 3b, the direction of the ray moves away from the perpendicular when it speeds up. The path is exactly reversible. In both cases, you can imagine what happens by thinking about pushing a lawn mower from a footpath onto grass, and vice versa. Going from the footpath to grass, the front wheels are slowed and pulled to the side as shown. This is the same change in direction as for light when it goes from a fast medium to a slow one. When going from the grass to the footpath, the front wheels can move

faster and the mower changes direction as shown. This, too, is the same change in direction as for light going from slow to fast.

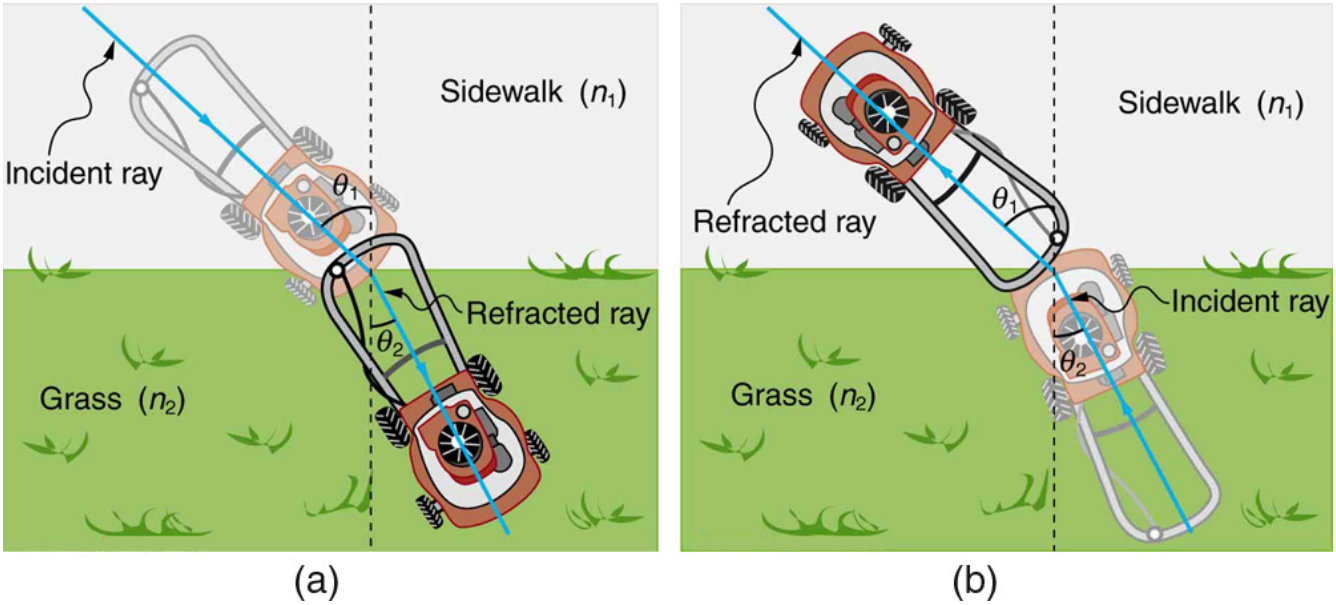


Figure 3. The change in direction of a light ray depends on how the speed of light changes when it crosses from one medium to another. The speed of light is greater in medium 1 than in medium 2 in the situations shown here. (a) A ray of light moves closer to the perpendicular when it slows down. This is analogous to what happens when a lawn mower goes from a footpath to grass. (b) A ray of light moves away from the perpendicular when it speeds up. This is analogous to what happens when a lawn mower goes from grass to footpath. The paths are exactly reversible.

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," which is stated in equation form as $n_1 \sin\theta_1 = n_2 \sin\theta_2$.

Here n_1 and n_2 are the indices of refraction for medium 1 and 2, and θ_1 and θ_2 are the angles between the rays and the perpendicular in medium 1 and 2, as shown in Figure 3. The incoming ray is called the incident ray and the outgoing ray the refracted ray, and the associated angles the incident angle and the refracted angle. The law of refraction is also called Snell's law after the Dutch mathematician Willebrord Snell (1591–1626), who discovered it in 1621. Snell's experiments showed that the law of refraction was obeyed and that a characteristic index of refraction n could be assigned to a given medium. Snell was not aware that the speed of light varied in different media, but through experiments he was able to determine indices of refraction from the way light rays changed direction.

The Law of Refraction

$$n_1 \sin\theta_1 = n_2 \sin\theta_2$$

Take-Home Experiment: A Broken Pencil

A classic observation of refraction occurs when a pencil is placed in a glass half filled with water. Do this and observe the shape of the pencil when you look at the pencil sideways, that is, through air, glass, water. Explain your observations. Draw ray diagrams for the situation.

Example 2. Determine the Index of Refraction from Refraction Data

Find the index of refraction for medium 2 in Figure 3a, assuming medium 1 is air and given the incident angle is 30.0º and the angle of refraction is 22.0º.

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, θ_1 = 30.0º and θ_2 = 22.0º. With this information, the only unknown in Snell's law is n_2, so that it can be used to find this unknown.

Solution

Snell's law is $n_1 \sin\theta_1 = n_2 \sin\theta_2$.

Rearranging to isolate n_2 gives

$$n_2 = n_1 \frac{\sin\theta_1}{\sin\theta_2}$$

Entering known values,

$$\begin{aligned} n_2 &= 1.00 \frac{\sin 30.0^\circ}{\sin 22.0^\circ} = \frac{0.500}{0.375} \\ &= 1.33 \end{aligned}$$

Discussion

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.

Example 3. A Larger Change in Direction

Suppose that in a situation like that in Example 2, light goes from air to diamond and that the incident angle is 30.0º. 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 θ_1 = 30.0º. We can look up the

index of refraction for diamond in Table 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,

$$\sin \theta_2 = \frac{1.00}{2.419} \sin 30.0^\circ = (0.413)\,(0.500) = 0.207$$

.

The angle is thus $\theta_2 = \sin^{-1} 0.207 = 11.9^\circ$.

Discussion

For the same 30º angle of incidence, the angle of refraction in diamond is significantly smaller than in water (11.9º rather than 22º—see the preceding example). 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.

Section Summary

- The changing of a light ray's direction when it passes through variations in matter is called refraction.
  The speed of light in vacuum $c = 2.9972458 \times 10^8$ m/s ≈ 3.00×10^8 m/s.
  Index of refraction $n = \frac{c}{v}$, where v is the speed of light in the material, c is the speed of light in vacuum, and n is the index of refraction.
  Snell's law, the law of refraction, is stated in equation form as $n_1 \sin \theta_1 = n_2 \sin\theta_2$.

Conceptual Questions

1. 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.
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. Will light change direction toward or away from the perpendicular when it goes from air to

water? Water to glass? Glass to air?

5. Explain why an object in water always appears to be at a depth shallower than it actually is? Why do people sometimes sustain neck and spinal injuries when diving into unfamiliar ponds or waters?
6. 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.
7. Why is the front surface of a thermometer curved as shown?

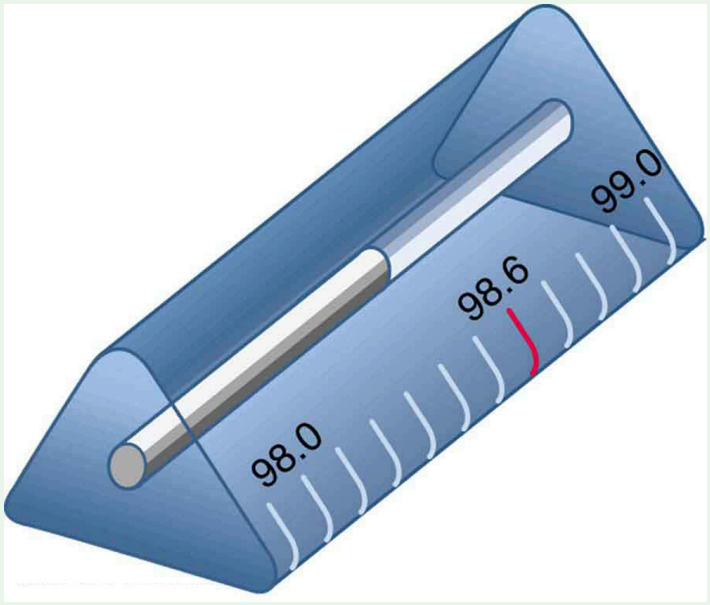


Figure 4. The curved surface of the thermometer serves a purpose.

8. Suppose light were incident from air onto a material that had a negative index of refraction, say –1.3; where does the refracted light ray go?

Problems & Exercises

1. What is the speed of light in water? In glycerine?
2. What is the speed of light in air? In crown glass?
3. 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 1.
4. In what substance in Table 1 is the speed of light 2.290×10^8 m/s?
5. 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?
6. A scuba diver training in a pool looks at his instructor as shown in Figure 5. 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 25.0º.

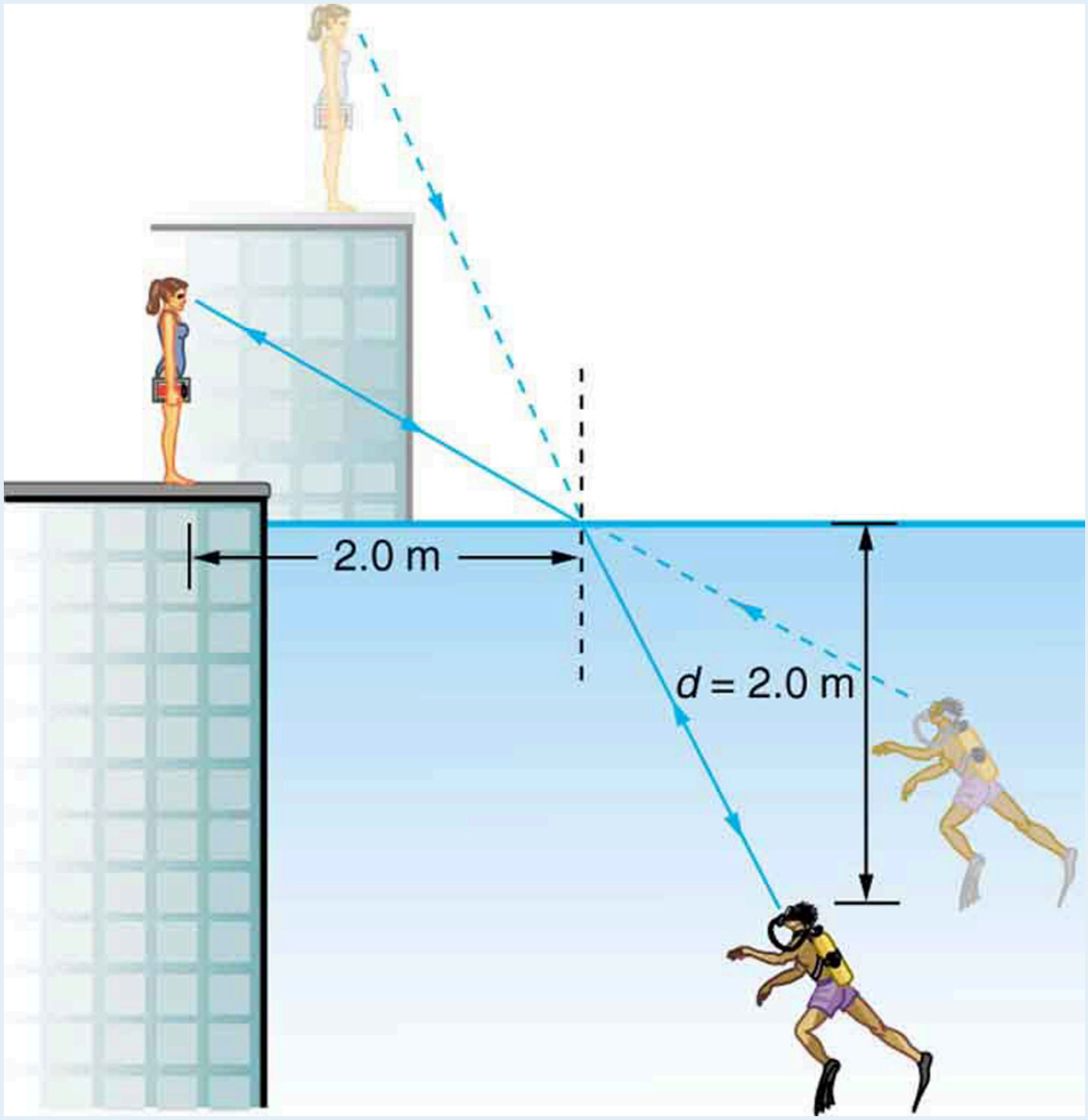


Figure 5. A scuba diver in a pool and his trainer look at each other.

7. 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?
8. (a) Using information in Figure 5, 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.
9. 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 45.0º, and you observe the angle of refraction to be 40.3º. What is the index of refraction of the substance and its likely identity?

10. 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 $n = 1.000293$.

11. Suppose Figure 6 represents 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 40.0º and the glass is 1.00 cm thick.

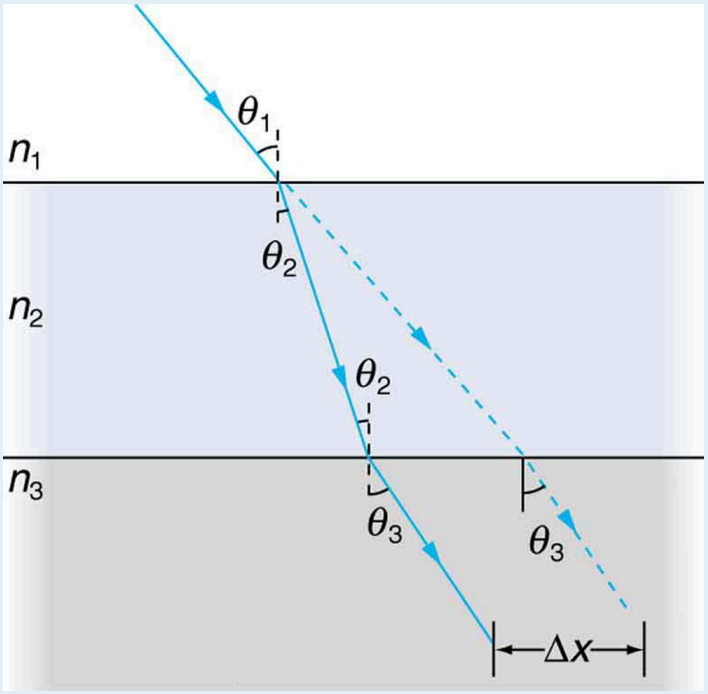


Figure 6. A ray of light passes from one medium to a third by traveling through a second. The final direction is the same as if the second medium were not present, but the ray is displaced by Δx (shown exaggerated).

12. Figure 6 shows a ray of light passing from one medium into a second and then a third. Show that θ_3 is the same as it would be if the second medium were not present (provided total internal reflection does not occur).

13. **Unreasonable Results.** Suppose light travels from water to another substance, with an angle of incidence of 10.0º and an angle of refraction of 14.9º. (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?

14. **Construct Your Own Problem.** Consider sunlight entering the Earth's atmosphere at sunrise and sunset—that is, at a 90º 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.

15. **Unreasonable Results.** Light traveling from water to a gemstone strikes the surface at an angle of 80.0º and has an angle of refraction of 15.2º. (a) What is the speed of light in the gemstone? (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

Glossary

refraction: changing of a light ray's direction when it passes through variations in matter

index of refraction: for a material, the ratio of the speed of light in vacuum to that in the material

Selected Solutions to Problems & Exercises

1. 2.25×10^8 m/s in water; 2.04×10^8 m/s in glycerine

3. 1.490, polystyrene

5. 1.28 s

7. 1.03 ns

9. $n = 1.46$, fused quartz

13. (a) 0.898; (b) Can't have $n < 1.00$ since this would imply a speed greater than c; (c) Refracted angle is too big relative to the angle of incidence.

15. (a)

$$\frac{c}{5.00}$$

; (b) Speed of light too slow, since index is much greater than that of diamond; (c) Angle of refraction is unreasonable relative to the angle of incidence.

Total Internal Reflection

Lumen Learning

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 fiber optics.
- 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, such as is shown in Figure 1a. 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, $\theta_1 > \theta_2$.) Now imagine what happens as the incident angle is increased. This causes θ_2 to increase also. The largest the angle of refraction θ_2 can be is 90^o, as shown in Figure 1b.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^o. That is, θ_c is the incident angle for which $\theta_2 = 90^o$. If the incident angle θ_1 is greater than the critical angle, as shown in Figure 1c, then all of the light is reflected back into medium 1, a condition called *total internal reflection*.

Critical Angle

The incident angle θ_1 that produces an angle of refraction of 90^o is called the critical angle, θ_c.

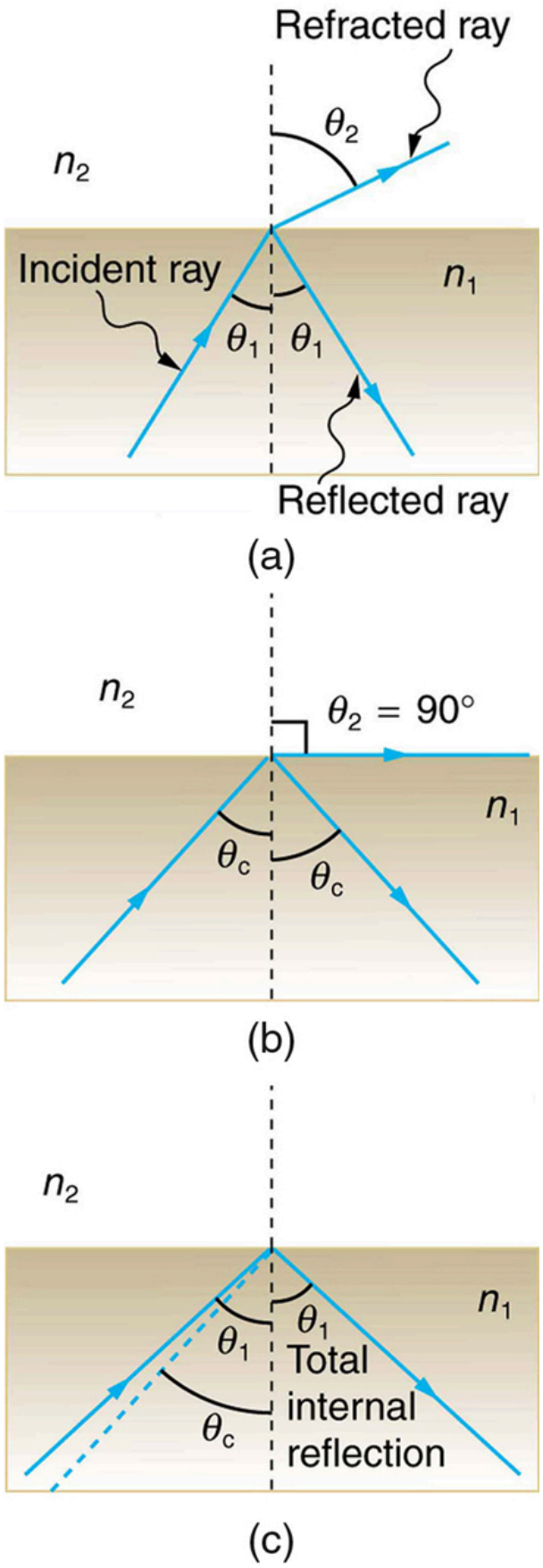


Figure 1. (a) A ray of light crosses a boundary where

the speed of light increases and the index of refraction decreases. That is, $n_2 < n_1$. The ray bends away from the perpendicular. (b) The critical angle θc is the one for which the angle of refraction is. (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 *90º* ($\theta_2 = 90^\circ$). Noting that $\sin 90^\circ = 1$, 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

$$\theta_c = \sin^{-1}\left(\frac{n_2}{n_1}\right)$$

for $n_1 > n_2$.

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 the above equation is written for a light ray that travels in medium 1 and reflects from medium 2, as shown in the figure.

Example 1. How Big is the Critical Angle Here?

What is the critical angle for light traveling in a polystyrene (a type of plastic) pipe surrounded by air?

Strategy

The index of refraction for polystyrene is found to be 1.49 in Figure 2, and 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 the equation

$$\theta_c = \sin^{-1}\left(\frac{n_2}{n_1}\right)$$

can be used to find the critical angle θ_c. Here, then, $n_2 = 1.00$ and $n_1 = 1.49$.

Solution

The critical angle is given by

$$\theta_c = \sin^{-1}\left(\frac{n_2}{n_1}\right)$$

Substituting the identified values gives

$$\begin{aligned}\theta_c &= \sin^{-1}\left(\tfrac{1.00}{1.49}\right)\\ &= \sin^{-1}(0.671)\\ &= 42.2^\circ\end{aligned}$$

Discussion

This means that any ray of light inside the plastic that strikes the surface at an angle greater than 42.2º will be totally reflected. This will make 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 48.6º, while that from diamond to air is 24.4º, and that from flint glass to crown glass is 66.3º. There is no total reflection for rays going in the other direction—for example, from air to water—since the condition that the second medium must have a smaller index of refraction is not satisfied. A number of interesting applications of total internal reflection follow.

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 (See Figure 2.) The index of refraction outside the fiber must be smaller than inside, a condition that is easily satisfied by coating the outside of the fiber with a material having an appropriate refractive index. 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.

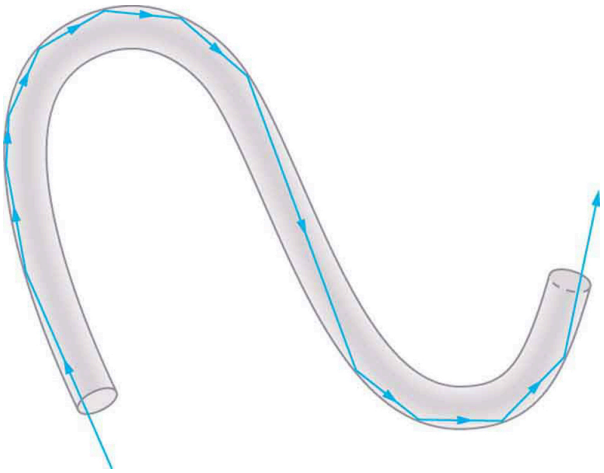


Figure 2. Light entering a thin fiber may strike the inside surface at large or grazing angles and is completely reflected if these angles exceed the critical angle. Such rays continue down the fiber, even following it around corners, since the angles of reflection and incidence remain large.

Bundles of fibers can be used to transmit an image without a lens, as illustrated in Figure 3. The output of a device called an *endoscope* is shown in Figure 3b. Endoscopes are used to explore the body through various 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 to be observed. Surgery can be performed, such as arthroscopic surgery on the knee 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.

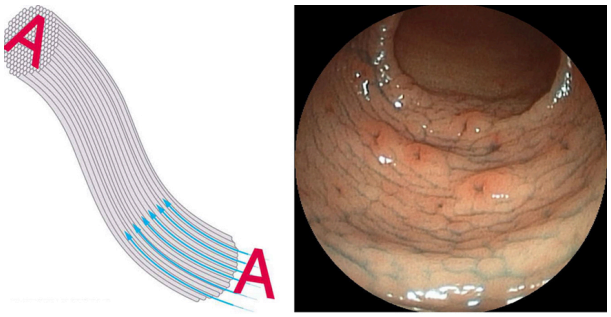


Figure 3. (a) An image is transmitted by a bundle of fibers that have fixed neighbors. (b) An endoscope is used to probe the body, both transmitting light to the interior and returning an image such as the one shown. (credit: Med_Chaos, Wikimedia Commons)

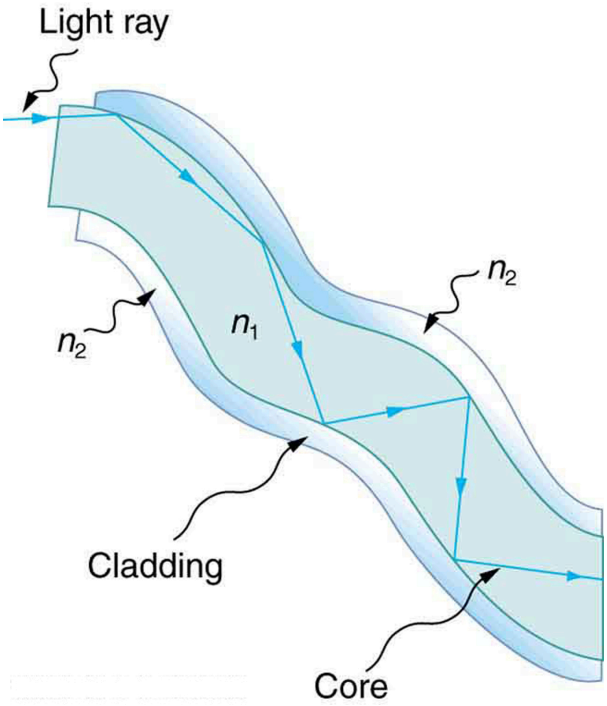
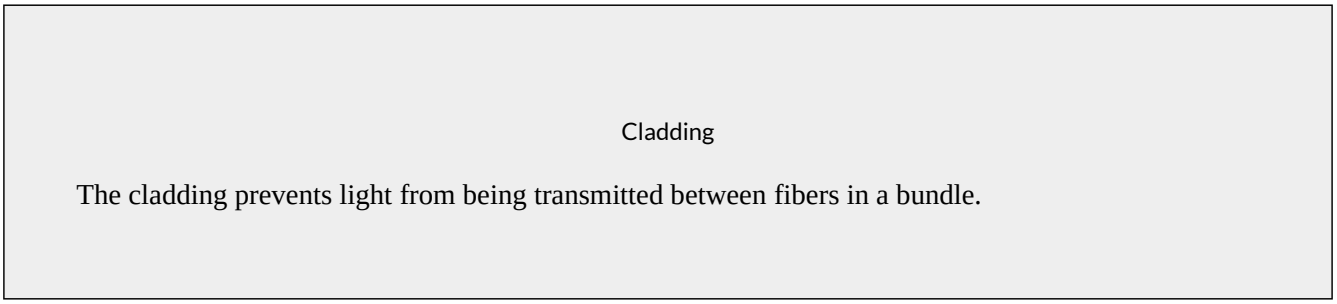


Figure 4. Fibers in bundles are clad by a material that has a lower index of refraction than the core to ensure total internal reflection, even when fibers are in contact with one another. This shows a single fiber with its cladding.

Fiber optics has revolutionized surgical techniques and observations within the body. There are a host of medical diagnostic and therapeutic uses. The flexibility of the fiber optic bundle allows it to navigate around difficult and small 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.

Fibers in bundles are surrounded by a cladding material that has a lower index of refraction than the core. (See Figure 4.) 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. The cladding prevents light from escaping out of the fiber; instead most of 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 flexible and durable.



The cladding prevents light from being transmitted between fibers in a bundle.

Special tiny lenses that can be attached to the ends of bundles of fibers are being designed and fabricated. Light emerging from a fiber bundle can be focused and a tiny spot can be imaged. 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 tens of microns below the surface without cutting the surface—non-intrusive diagnostics. This is particularly useful for determining the extent of cancers in the stomach and bowel.

Most telephone conversations and Internet communications are now carried by laser signals along optical fibers. 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

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. This is true whenever the reflecting surfaces are perpendicular, and it is independent of the angle of incidence. Such an object, shown in Figure 5, is called a *corner reflector*, since the light bounces from its inside corner. Many inexpensive reflector buttons on bicycles, cars, and warning signs have corner reflectors designed to return light in the direction from which it originated. It was more expensive for astronauts to place one on the moon. Laser signals can be bounced from that corner reflector to measure the gradually increasing distance to the moon with great precision.

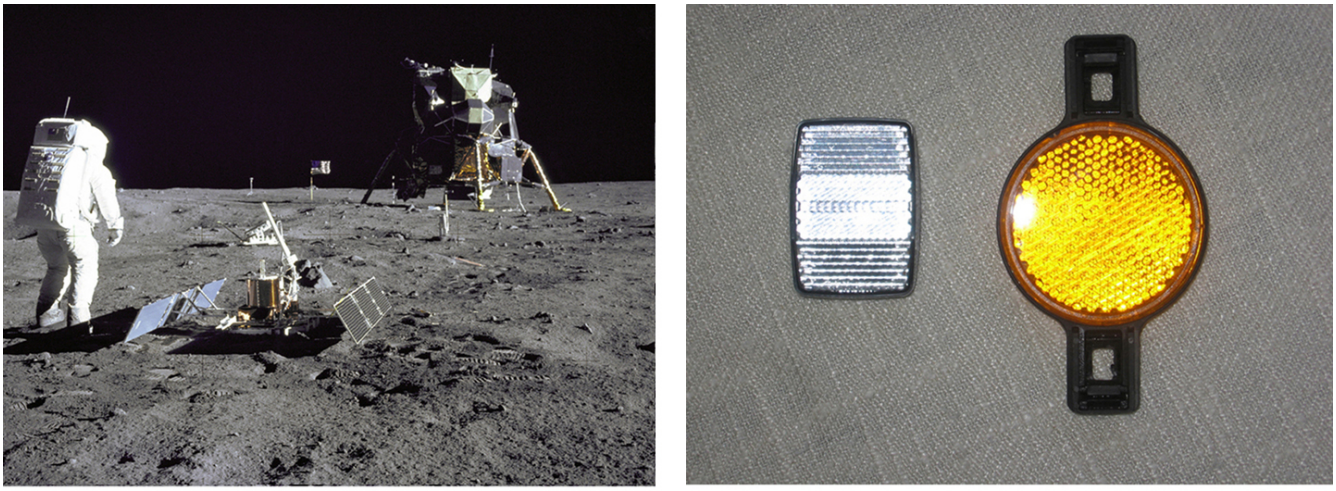


Figure 5. (a) Astronauts placed a corner reflector on the moon to measure its gradually increasing orbital distance. (credit: NASA) (b) The bright spots on these bicycle safety reflectors are reflections of the flash of the camera that took this picture on a dark night. (credit: Julo, Wikimedia Commons)

Corner reflectors 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 45º. One use of these perfect mirrors is in binoculars, as shown in Figure 6. Another use is in periscopes found in submarines.

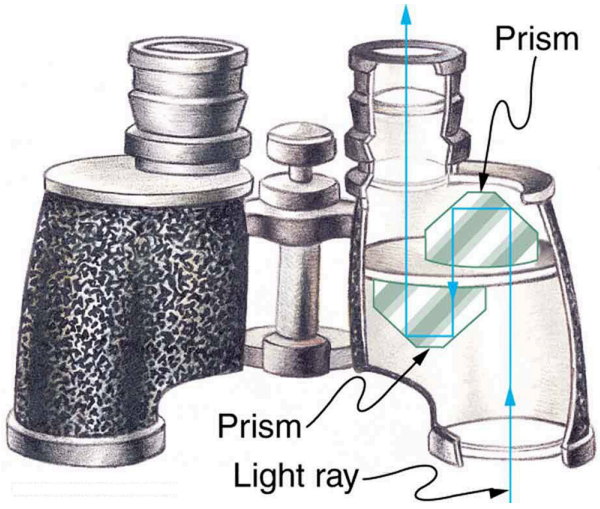


Figure 6. These binoculars employ corner reflectors with total internal reflection to get light to the observer's eyes.

The Sparkle of Diamonds

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 24.4º, and so when light enters a diamond, it has trouble getting back out. (See Figure 7.) Although light freely enters the diamond, it can exit only if it makes an angle less than 24.4º. Facets on diamonds are specifically intended to make this unlikely, so that the light can exit only in certain places. Good diamonds are very clear, so that the light makes many internal reflections and is concentrated at the few places it can exit—hence the sparkle. (Zircon is a natural gemstone that has an exceptionally large index of refraction, but 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 still less than that of diamond.) The colors you see emerging from a sparkling diamond are not due to the diamond's color, which is usually nearly colorless. Those colors result from dispersion, the topic of Dispersion: The Rainbow and Prisms. 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, while around 50% of the world's clear diamonds come from central and southern Africa.

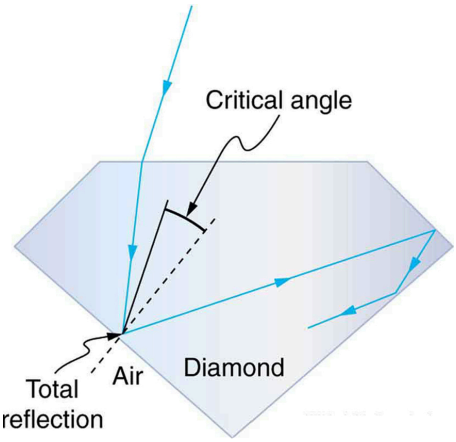
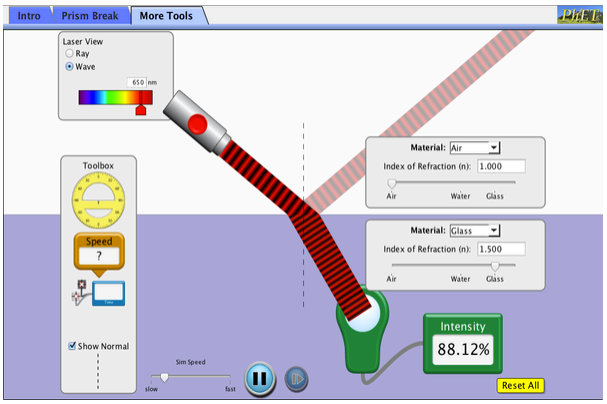


Figure 7. Light cannot easily escape a diamond, because its critical angle with air is so small. Most reflections are total, and the facets are placed so that light can exit only in particular ways—thus concentrating the light and making the diamond sparkle.

PhET Explorations: Bending Light

Explore bending of light between two media with different indices of refraction. See how changing from air to water to glass changes the bending angle. Play with prisms of different shapes and make rainbows.



Click to download the simulation. Run using Java.

Section Summary

- The incident angle that produces an angle of refraction of 90º is called critical angle.
- Total internal reflection is a phenomenon that occurs at the boundary between two mediums, 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.
- Endoscopes are used to explore the body through various orifices or minor incisions, based on the transmission of light through optical fibers.
- 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.

Conceptual Questions

1. A ring with a colorless gemstone is dropped into water. The gemstone becomes invisible when submerged. Can it be a diamond? Explain.
2. 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.
3. Is it possible that total internal reflection plays a role in rainbows? Explain in terms of indices of refraction and angles, perhaps referring to Figure 8. Some of us have seen the formation of a double rainbow. Is it physically possible to observe a triple rainbow?

Figure 8. Double rainbows are not a very common observance. (credit: InvictusOU812, Flickr)

4. 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.

Problems & Exercises

1. Verify that the critical angle for light going from water to air is 48.6º, as discussed at the end of Example 1, regarding the critical angle for light traveling in a polystyrene (a type of plastic) pipe surrounded by air.
2. (a) At the end of Example 1, it was stated that the critical angle for light going from diamond to air is 24.4º. Verify this. (b) What is the critical angle for light going from zircon to air?
3. An optical fiber uses flint glass clad with crown glass. What is the critical angle?
4. At what minimum angle will you get total internal reflection of light traveling in water and reflected from ice?
5. Suppose you are using total internal reflection to make an efficient corner reflector. If there is air outside and the incident angle is 45.0º, what must be the minimum index of refraction of the material from which the reflector is made?
6. 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 68.4º when submerged in water? What is the substance, based on Figure 9? (b) What would the critical angle be for this substance in air?

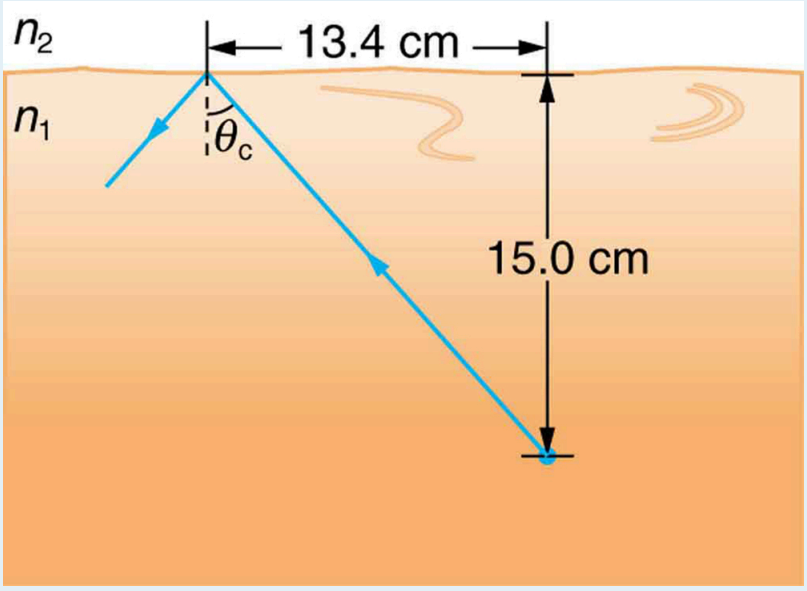


Figure 9. A light ray inside a liquid strikes the surface at the critical angle and undergoes total internal reflection.

7. A ray of light, emitted beneath the surface of an unknown liquid with air above it, undergoes total internal reflection as shown in Figure 9. What is the index of refraction for the liquid and its likely identification?
8. A light ray entering an optical fiber surrounded by air is first refracted and then reflected as

shown in Figure 10. Show that if the fiber is made from crown glass, any incident ray will be totally internally reflected.

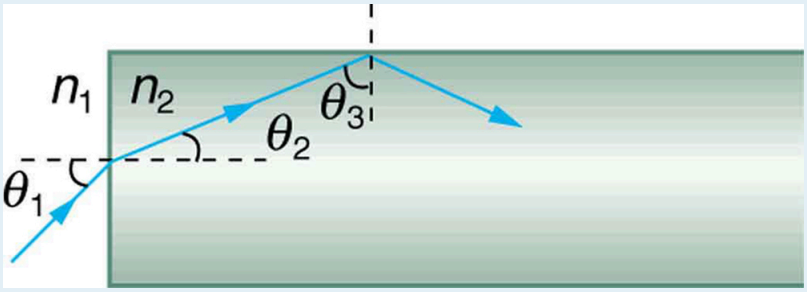


Figure 10. A light ray enters the end of a fiber, the surface of which is perpendicular to its sides. Examine the conditions under which it may be totally internally reflected.

Glossary

critical angle: incident angle that produces an angle of refraction of 90º

fiber optics: transmission of light down fibers of plastic or glass, applying the principle of total internal reflection

corner reflector: an object consisting of two mutually perpendicular reflecting surfaces, so that the light that enters is reflected back exactly parallel to the direction from which it came

zircon: natural gemstone with a large index of refraction

Selected Solutions to Problems & Exercises

3. 66.3º

5. > 1.414

7. 1.50, benzene

Dispersion: The Rainbow and Prisms

Lumen Learning

Learning Objective

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

- Explain the phenomenon of dispersion and discuss its advantages and disadvantages.

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. (See Figure 1.)

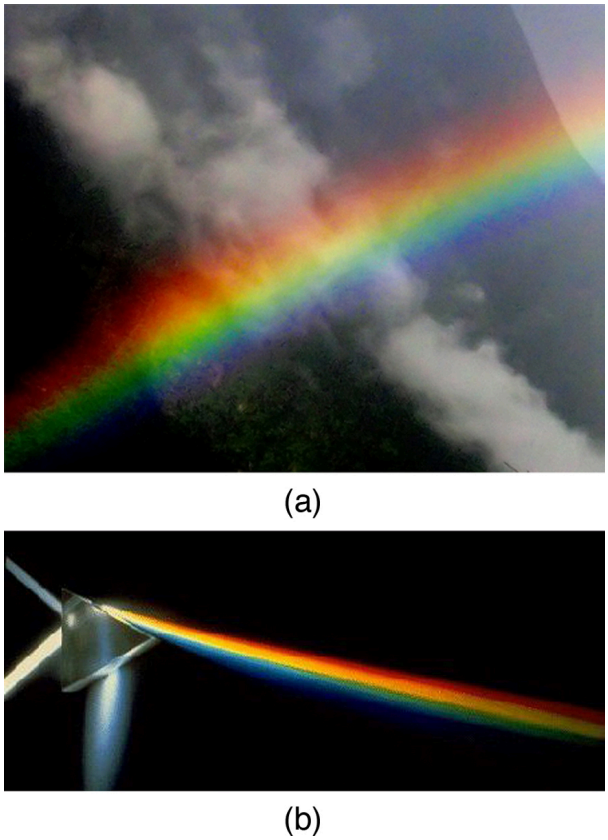


Figure 1. The colors of the rainbow (a) and those produced by a prism (b) are identical. (credit: Alfredo55, Wikimedia Commons; NASA)

We see about six colors in a rainbow—red, orange, yellow, green, blue, and violet; sometimes indigo is listed, too. Those colors are associated with different wavelengths of light, as shown in Figure 2. 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 plotted versus wavelength in Figure 2. What this implies is that white light is spread out according to wavelength in a rainbow. *Dispersion* is defined as the spreading of white light into its full spectrum of wavelengths. More technically, dispersion occurs whenever there is a process that changes the direction of light in a manner that depends on wavelength. Dispersion, as a general phenomenon, can occur for any type of wave and always involves wavelength-dependent processes.

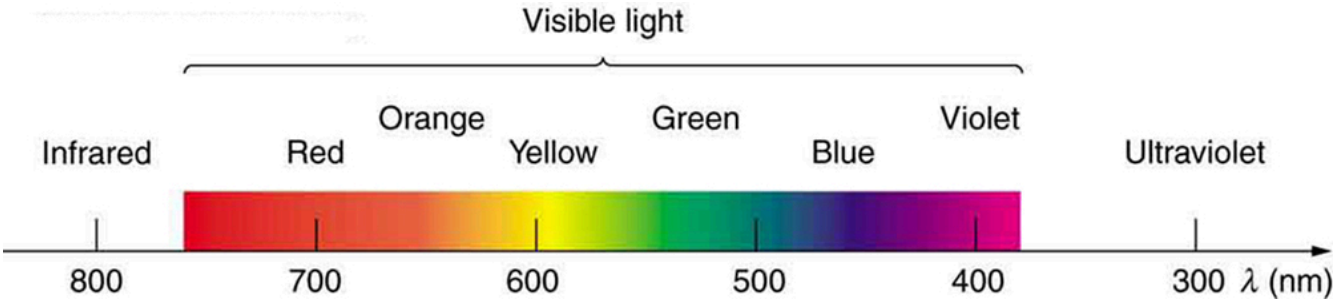


Figure 2. Even though rainbows are associated with seven colors, the rainbow is a continuous distribution of colors according to wavelengths.

Dispersion

Dispersion is defined to be the spreading of white light into its full spectrum of wavelengths.

Refraction is responsible for dispersion in rainbows and many other situations. The angle of refraction depends on the index of refraction, as we saw in The Law of Refraction. We know that the index of refraction *n* depends on the medium. But for a given medium, *n* also depends on wavelength. (See Table 1. 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 3b, and the light is dispersed into the same sequence of wavelengths as seen in Figure 1 and Figure 2.

Table 1. Index of Refraction *n* in Selected Media at Various Wavelengths

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

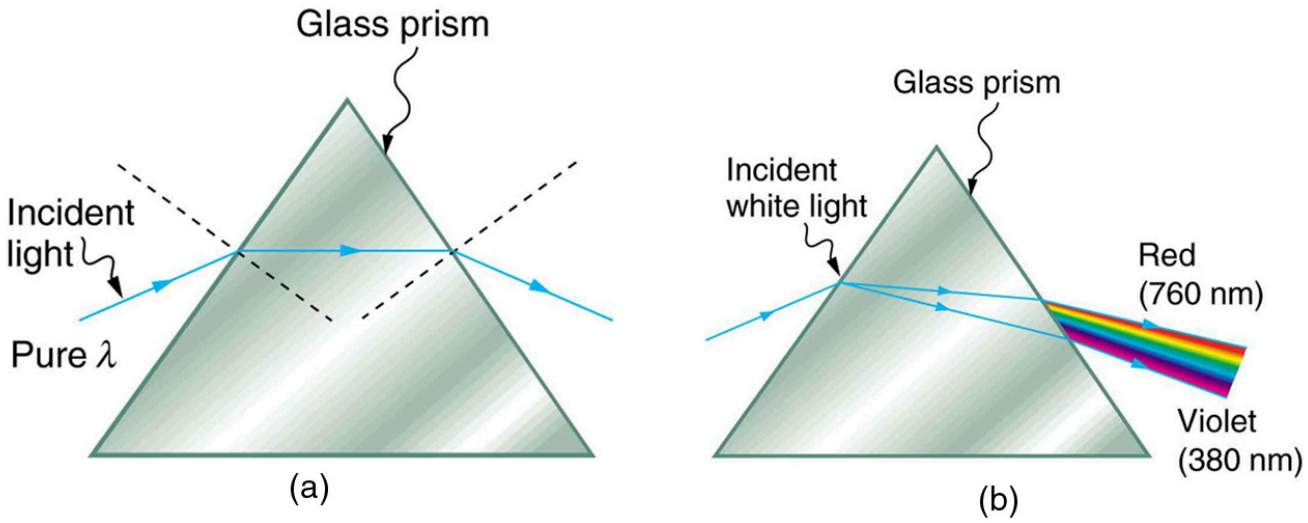


Figure 3. (a) A pure wavelength of light falls onto a prism and is refracted at both surfaces. (b) White light is dispersed by the prism (shown exaggerated). Since the index of refraction varies with wavelength, the angles of refraction vary with wavelength. A sequence of red to violet is produced, because the index of refraction increases steadily with decreasing wavelength.

Making Connections: Dispersion

Any type of wave can exhibit dispersion. Sound waves, all types of electromagnetic waves, and water waves can be dispersed according to wavelength. Dispersion occurs whenever the speed of propagation depends on wavelength, thus separating and spreading out various wavelengths. 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 is 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 empty space.

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, as shown in Figure 4. 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, as shown in Figure 5a. (There is no dispersion caused by reflection at the back surface, since the law of reflection does not depend on wavelength.) The actual rainbow of colors seen by an observer depends on the myriad of 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 Figure 5b. (If there are two reflections of light within the water drop, another "secondary" rainbow is produced. This rare event produces an arc that lies above the primary rainbow arc—see Figure 5c.)

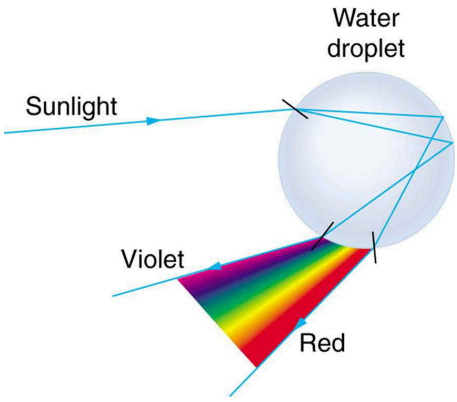


Figure 4. Part of the light falling on this water drop enters and is reflected from the back of the drop. This light is refracted and dispersed both as it enters and as it leaves the drop.

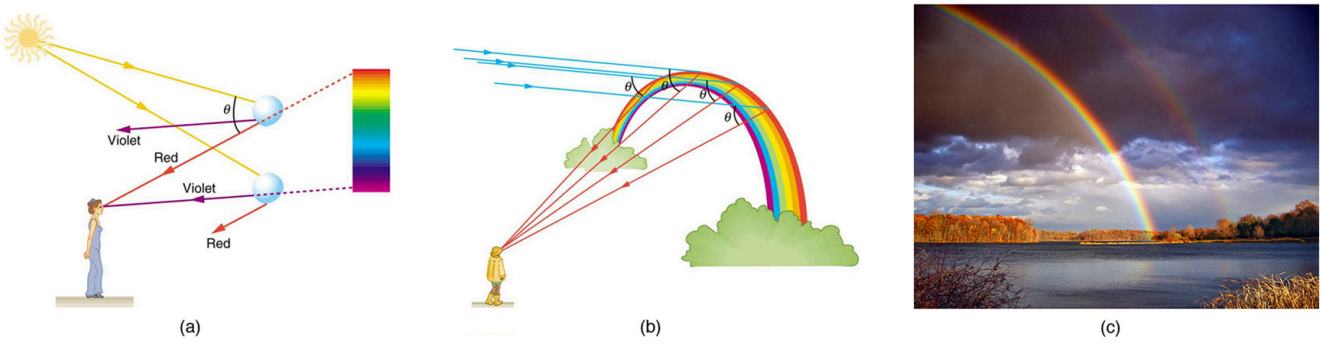


Figure 5. (a) Different colors emerge in different directions, and so you must look at different locations to see the various colors of a rainbow. (b) The arc of a rainbow results from the fact that a line between the observer and any point on the arc must make the correct angle with the parallel rays of sunlight to receive the refracted rays. (c) Double rainbow. (credit: Nicholas, Wikimedia Commons)

Rainbows

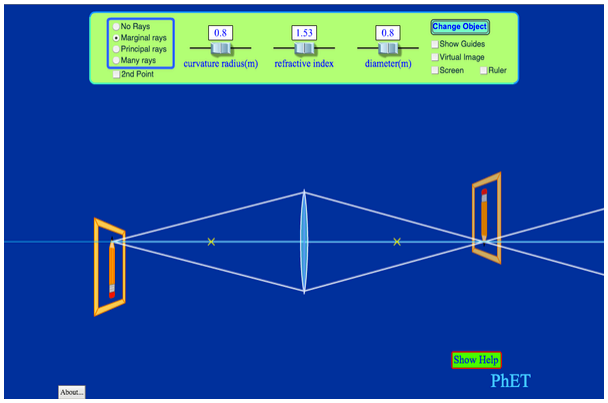
Rainbows are produced by a combination of refraction and reflection.

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. As with many phenomena, dispersion can be useful or a nuisance, depending on the situation and our human goals.

PhET Explorations: Geometric Optics

How does a lens form an image? See how light rays are refracted by a lens. Watch how the image changes when you adjust the focal length of the lens, move the object, move the lens, or move the screen.



Click to run the simulation.

Section Summary

- 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.

Problems & Exercises

1. (a) What is the ratio of the speed of red light to violet light in diamond, based on [link]? (b) What is this ratio in polystyrene? (c) Which is more dispersive?
2. A beam of white light goes from air into water at an incident angle of 75.0º. At what angles are the red (660 nm) and violet (410 nm) parts of the light refracted?
3. By how much do the critical angles for red (660 nm) and violet (410 nm) light differ in a diamond surrounded by air?
4. (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 30.0º 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?

5. 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 60.0º incident angle. What is the angle between the two colors in water?
6. A ray of 610 nm light goes from air into fused quartz at an incident angle of 55.0º. At what incident angle must 470 nm light enter flint glass to have the same angle of refraction?
7. 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 30.0º incident angle. (a) At what angles do the two colors emerge? (b) By what distance are the red and blue separated when they emerge?
8. A narrow beam of white light enters a prism made of crown glass at a 45.0º incident angle, as shown in Figure 6. 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?

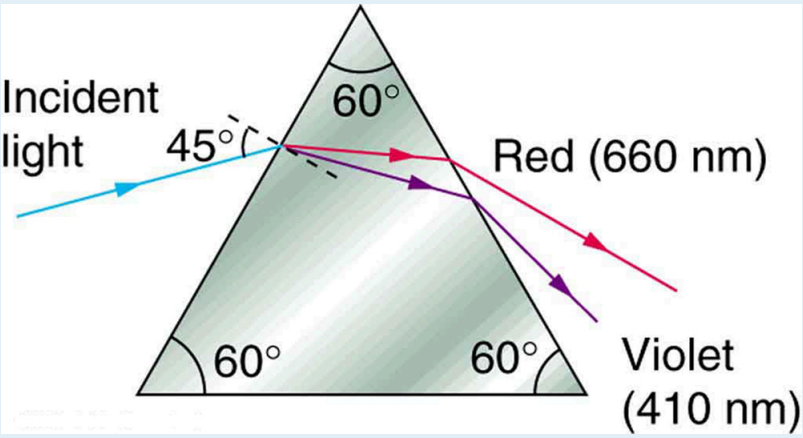


Figure 6. This prism will disperse the white light into a rainbow of colors. The incident angle is 45.0º, and the angles at which the red and violet light emerge are θ_R and θ_V.

Glossary

dispersion: spreading of white light into its full spectrum of wavelengths

rainbow: dispersion of sunlight into a continuous distribution of colors according to wavelength, produced by the refraction and reflection of sunlight by water droplets in the sky

Selected Solutions to Problems & Exercises

2. 46.5º, red; 46.0º, violet

4. (a) 0.043º; (b) 1.33 m

6. 71.3º

8. 53.5º, red; 55.2º, violet

Image Formation by Lenses

Lumen Learning

Learning Objectives

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

- List the rules for ray tracking for thin lenses.
- Illustrate the formation of images using the technique of ray tracking.
- Determine power of a lens given the focal length.

Lenses are found in a huge array of optical instruments, ranging from a simple magnifying glass to the eye to a camera's zoom lens. In this section, we will use the law of refraction 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 the convex lens in Figure 1. The convex lens shown has been shaped so that all light rays that enter it parallel to its axis cross one another at a single point on the opposite side of the lens. (The axis is defined to be a line normal to the lens at its center, as shown in Figure 1.) Such a lens is called a *converging (or convex) lens* for the converging effect it has on light rays. An expanded view of the path of one ray through the lens is shown, to illustrate how the ray changes direction both as it enters and as it leaves the lens. Since the index of refraction of the lens is greater than that of air, the ray moves towards the perpendicular as it enters and away from the perpendicular as it leaves. (This is in accordance with the law of refraction.) Due to the lens's shape, light is thus bent toward the axis at both surfaces. The point at which the rays cross is defined to be the *focal point* F of the lens. The distance from the center of the lens to its focal point is defined to be the *focal length f* of the lens. Figure 2 shows how a converging lens, such as that in a magnifying glass, can converge the nearly parallel light rays from the sun to a small spot.

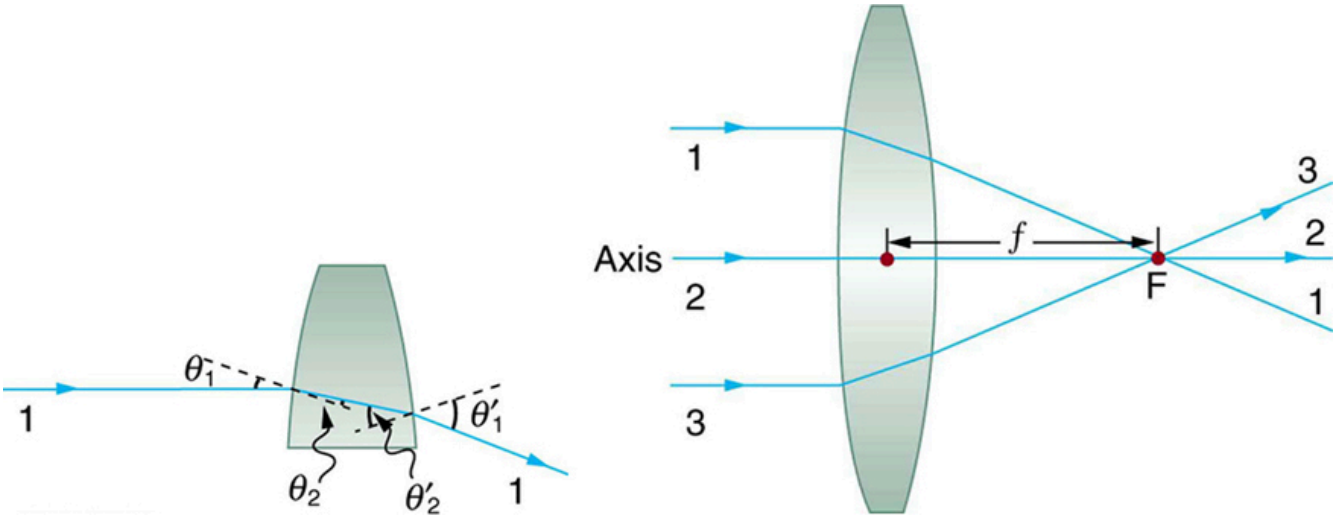


Figure 1. Rays of light entering a converging lens parallel to its axis converge at its focal point F. (Ray 2 lies on the axis of the lens.) The distance from the center of the lens to the focal point is the lens's focal length f. An expanded view of the path taken by ray 1 shows the perpendiculars and the angles of incidence and refraction at both surfaces.

Converging or Convex Lens

The lens in which light rays that enter it parallel to its axis cross one another at a single point on the opposite side with a converging effect is called converging lens.

Focal Point F

The point at which the light rays cross is called the focal point F of the lens.

Focal Length *f*

The distance from the center of the lens to its focal point is called focal length *f*.

Figure 2. Sunlight focused by a converging magnifying glass can burn paper. Light rays from the sun are nearly parallel and cross at the focal point of the lens. The more powerful the lens, the closer to the lens the rays will cross.

The greater effect a lens has on light rays, the more powerful it is said to be. For example, a powerful converging lens will focus parallel light rays closer to itself and will have a smaller focal length than a weak lens. The light will also focus into a smaller and more intense spot for a more powerful lens. The *power P* of a lens is defined to be the inverse of its focal length. In equation form, this is

$$P = \frac{1}{f}$$

.

Power *P*

The **power** *P* of a lens is defined to be the inverse of its focal length. In equation form, this is

$$P = \frac{1}{f}$$

, where *f* is the focal length of the lens, which must be given in meters (and not cm or mm). The power of a lens *P* has the unit diopters (D), provided that the focal length is given in meters. That is,

$$1\text{D} = \frac{1}{\text{m}}, \text{ or } 1\text{m}^{-1}$$

. (Note that this power (optical power, actually) is not the same as power in watts defined in the chapter Work, Energy, and Energy Resources. It is a concept related to the effect of optical devices on light.) Optometrists prescribe common spectacles and contact lenses in units of diopters.

Example 1. What is the Power of a Common Magnifying Glass?

Suppose you take a magnifying glass out on a sunny day and you find that it concentrates sunlight to a small spot 8.00 cm away from the lens. What are the focal length and power of the lens?

Strategy

The situation here is the same as those shown in Figure 1 and Figure 2. The Sun is so far away that the Sun's rays are nearly parallel when they reach Earth. The magnifying glass is a convex (or converging) lens, focusing the nearly parallel rays of sunlight. Thus the focal length of the lens is the distance from the lens to the spot, and its power is the inverse of this distance (in m).

Solution

The focal length of the lens is the distance from the center of the lens to the spot, given to be 8.00 cm. Thus,

$$f = 8.00 \text{ cm}.$$

To find the power of the lens, we must first convert the focal length to meters; then, we substitute this value into the equation for power. This gives

$$P = \frac{1}{f} = \frac{1}{0.0800 \text{ m}} = 12.5 \text{ D}$$

.

Discussion

This is a relatively powerful lens. The power of a lens in diopters should not be confused with the familiar concept of power in watts. It is an unfortunate fact that the word "power" is used for two completely different concepts. If you examine a prescription for eyeglasses, you will note lens powers given in diopters. If you examine the label on a motor, you will note energy consumption rate given as a power in watts.

Figure 3 shows a concave lens and the effect it has on rays of light that enter it parallel to its axis (the path taken by ray 2 in the Figure is the axis of the lens). The concave lens is a *diverging lens*, because it causes the light rays to bend away (diverge) from its axis. In this case, the lens has been shaped so that all light rays entering it parallel to its axis appear to originate from the same point, F, defined to be the focal point of a diverging lens. The distance from the center of the lens to the focal point is again called the focal length *f* of the lens. Note that the focal length and power of a diverging lens are defined to be negative.

For example, if the distance to *F* in Figure 3 is 5.00 cm, then the focal length is $f = -5.00$ cm and the power of the lens is $P = -20$ D. An expanded view of the path of one ray through the lens is shown in the Figure to illustrate how the shape of the lens, together with the law of refraction, causes the ray to follow its particular path and be diverged.

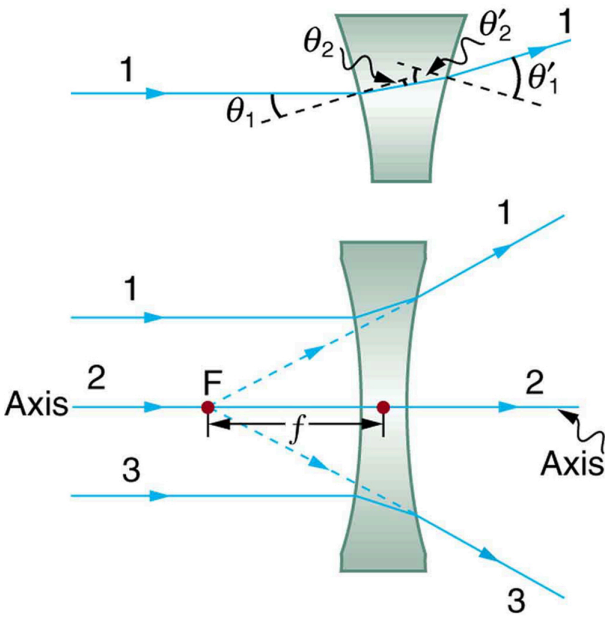


Figure 3. Rays of light entering a diverging lens parallel to its axis are diverged, and all appear to originate at its focal point F. The dashed lines are not rays—they indicate the directions from which the rays appear to come. The focal length f of a diverging lens is negative. An expanded view of the path taken by ray 1 shows the perpendiculars and the angles of incidence and refraction at both surfaces.

Diverging Lens

A lens that causes the light rays to bend away from its axis is called a diverging lens.

As noted in the initial discussion of the law of refraction in The Law of Refraction, 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 1 and Figure 3. For example, if a point light source is placed at the focal point of a convex lens, as shown in Figure 4, parallel light rays emerge from the other side.

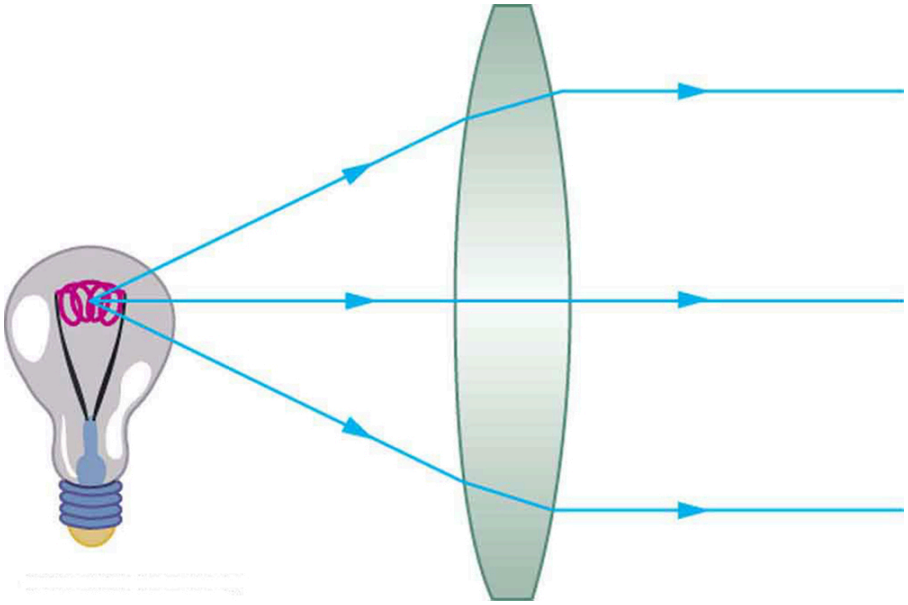


Figure 4. A small light source, like a light bulb filament, placed at the focal point of a convex lens, results in parallel rays of light emerging from the other side. The paths are exactly the reverse of those shown in Figure 1. This technique is used in lighthouses and sometimes in traffic lights to 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 that light rays take. For rays passing through matter, the law of refraction is used to trace the paths. Here we use ray tracing to help us understand the action of lenses in situations ranging from forming images on film to magnifying small print to correcting nearsightedness. While ray tracing for complicated lenses, such as those found in sophisticated cameras, may require computer techniques, there is a set of simple rules for tracing rays through thin lenses.

A *thin lens* is defined to be one whose thickness allows rays to refract, as illustrated in Figure 1, but does not allow properties such as dispersion and aberrations. An ideal thin lens has two refracting surfaces but the lens is thin enough to assume that light rays bend only once. A thin symmetrical lens has two focal points, one on either side and both at the same distance from the lens. (See Figure 6.)

Another important characteristic of a thin lens is that light rays through its center are deflected by a negligible amount, as seen in Figure 5.

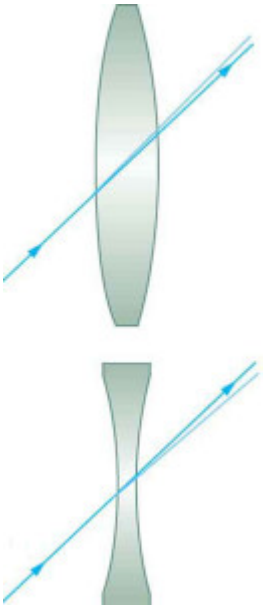


Figure 6. The light ray through the center of a thin lens is deflected by a negligible amount and is assumed to emerge parallel to its original path (shown as a shaded line).

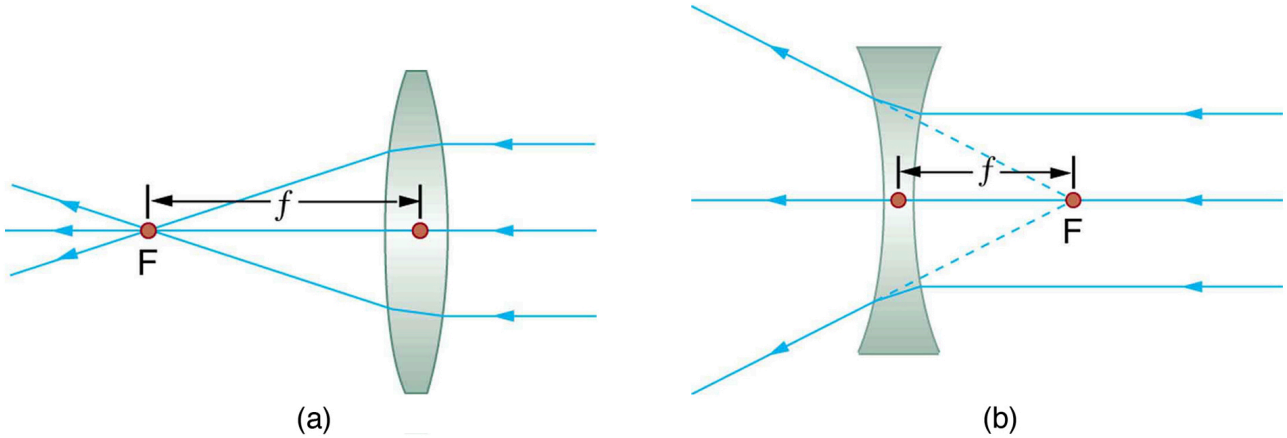


Figure 6. Thin lenses have the same focal length on either side. (a) Parallel light rays entering a converging lens from the right cross at its focal point on the left. (b) Parallel light rays entering a diverging lens from the right seem to come from the focal point on the right.

Thin Lens

A thin lens is defined to be one whose thickness allows rays to refract but does not allow properties such as dispersion and aberrations.

Take-Home Experiment: A Visit to the Optician

Look through your eyeglasses (or those of a friend) backward and forward and comment on whether they act like thin lenses.

Using paper, pencil, and a straight edge, ray tracing can accurately describe the operation of a lens. The rules for ray tracing for thin lenses are based on the illustrations already discussed:

1. A ray entering a converging lens parallel to its axis passes through the focal point F of the lens on the other side. (See rays 1 and 3 in Figure 1.)
2. A ray entering a diverging lens parallel to its axis seems to come from the focal point F. (See rays 1 and 3 in Figure 2.)
3. A ray passing through the center of either a converging or a diverging lens does not change direction. (See Figure 5, and see ray 2 in Figure 1 and Figure 2.)
4. A ray entering a converging lens through its focal point exits parallel to its axis. (The reverse of rays 1 and 3 in Figure 1.)
5. A ray that enters a diverging lens by heading toward the focal point on the opposite side exits parallel to the axis. (The reverse of rays 1 and 3 in Figure 2.)

Rules for Ray Tracing

1. A ray entering a converging lens parallel to its axis passes through the focal point F of the lens on the other side.
2. A ray entering a diverging lens parallel to its axis seems to come from the focal point F.
3. A ray passing through the center of either a converging or a diverging lens does not change direction.
4. A ray entering a converging lens through its focal point exits parallel to its axis.
5. A ray that enters a diverging lens by heading toward the focal point on the opposite side

exits parallel to the axis.

Image Formation by Thin Lenses

In some circumstances, a lens forms an obvious image, such as when a movie projector casts an image onto a screen. In other cases, the image is less obvious. Where, for example, is the image formed by eyeglasses? We will use ray tracing for thin lenses to illustrate how they form images, and we will develop equations to describe the image formation quantitatively.

Consider an object some distance away from a converging lens, as shown in Figure 7. To find the location and size of the image formed, we trace the paths of selected light rays originating from one point on the object, in this case the top of the person's head. The Figure shows three rays from the top of the object that can be traced using the ray tracing rules given above. (Rays leave this point going in many directions, but we concentrate on only a few with paths that are easy to trace.) The first ray is one that enters the lens parallel to its axis and passes through the focal point on the other side (rule 1). The second ray passes through the center of the lens without changing direction (rule 3). The third ray passes through the nearer focal point on its way into the lens and leaves the lens parallel to its axis (rule 4). The three rays cross at the same point on the other side of the lens. The image of the top of the person's head is located at this point. All rays that come from the same point on the top of the person's head are refracted in such a way as to cross at the point shown. Rays from another point on the object, such as her belt buckle, will also cross at another common point, forming a complete image, as shown. Although three rays are traced in Figure 7, only two are necessary to locate the image. It is best to trace rays for which there are simple ray tracing rules. Before applying ray tracing to other situations, let us consider the example shown in Figure 7 in more detail.

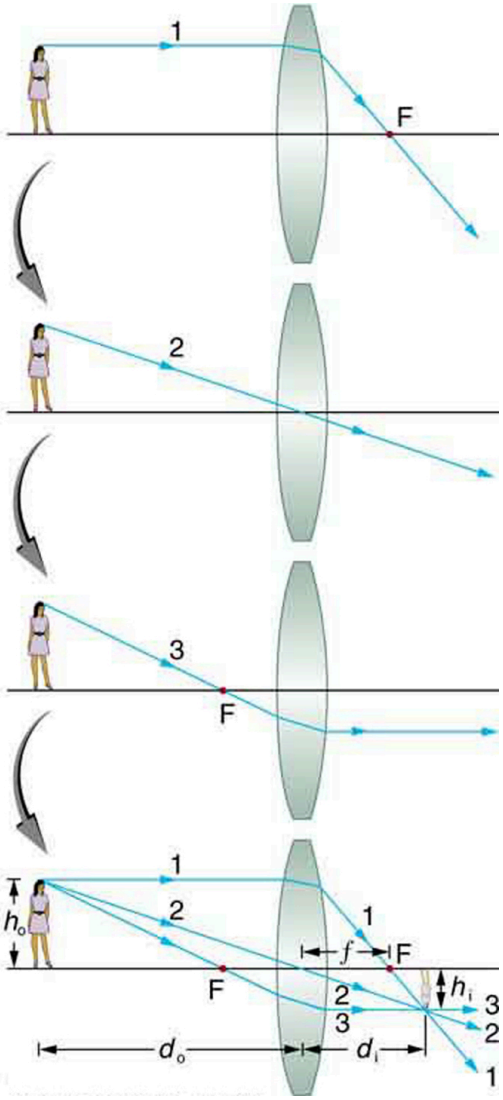


Figure 7. Ray tracing is used to locate the image formed by a lens. Rays originating from the same point on the object are traced—the three chosen rays each follow one of the rules for ray tracing, so that their paths are easy to determine. The image is located at the point where the rays cross. In this case, a real image—one that can be projected on a screen—is formed.

The image formed in Figure 7 is a *real image*, meaning that it can be projected. That is, light rays from one point on the object actually cross at the location of the image and can be projected onto a screen, a piece of film, or the retina of an eye, for example. Figure 8 shows how such an image would be projected onto film by a camera lens. This Figure also shows how a real image is projected onto the retina by the lens of an eye. Note that the image is there whether it is projected onto a screen or not.

Real Image

The image in which light rays from one point on the object actually cross at the location of the image and can be projected onto a screen, a piece of film, or the retina of an eye is called a real image.

Several important distances appear in Figure 7. We define d_o to be the object distance, the distance of an object from the center of a lens. *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 height of the image are given the symbols h_o and h_i, respectively. Images that appear upright relative to the object have heights that are positive and those that are inverted have negative heights. Using the rules of ray tracing and making a scale drawing with paper and pencil, like that in Figure 7, 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. To obtain numerical information, we use a pair of equations that can be derived from a geometric analysis of ray tracing for thin lenses. The *thin lens equations* are

$$\frac{1}{d_o} + \frac{1}{d_i} = \frac{1}{f}$$

and

$$\frac{h_i}{h_o} = \frac{d_i}{d_o} = m$$

.

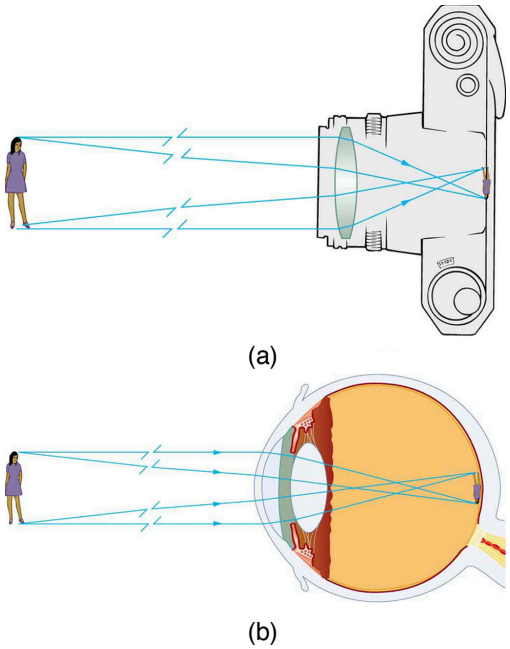


Figure 8. Real images can be projected. (a) A real image of the person is projected onto film. (b) The converging nature of the multiple surfaces that make up the eye result in the projection of a real image on the retina.

We define the ratio of image height to object height $\left(\frac{h_i}{h_o}\right)$ to be the *magnification m*. (The minus sign in the equation above will be discussed shortly.) The thin lens equations are broadly applicable to all situations involving thin lenses (and "thin" mirrors, as we will see later). We will explore many features of image formation in the following worked examples.

Image Distance

The distance of the image from the center of the lens is called image distance.

Thin Lens Equations and Magnification

$$\frac{1}{d_o} + \frac{1}{d_i} = \frac{1}{f}$$

$$\frac{h_i}{h_o} = \frac{d_i}{d_o} = m$$

Example 2. Finding the Image of a Light Bulb Filament by Ray Tracing and by the Thin Lens Equations

A clear glass light bulb is placed 0.750 m from a convex lens having a 0.500 m focal length, as shown in Figure 9. Use ray tracing to get an approximate location for the image. Then use the thin lens equations to calculate both the location of the image and its magnification. Verify that ray tracing and the thin lens equations produce consistent results.

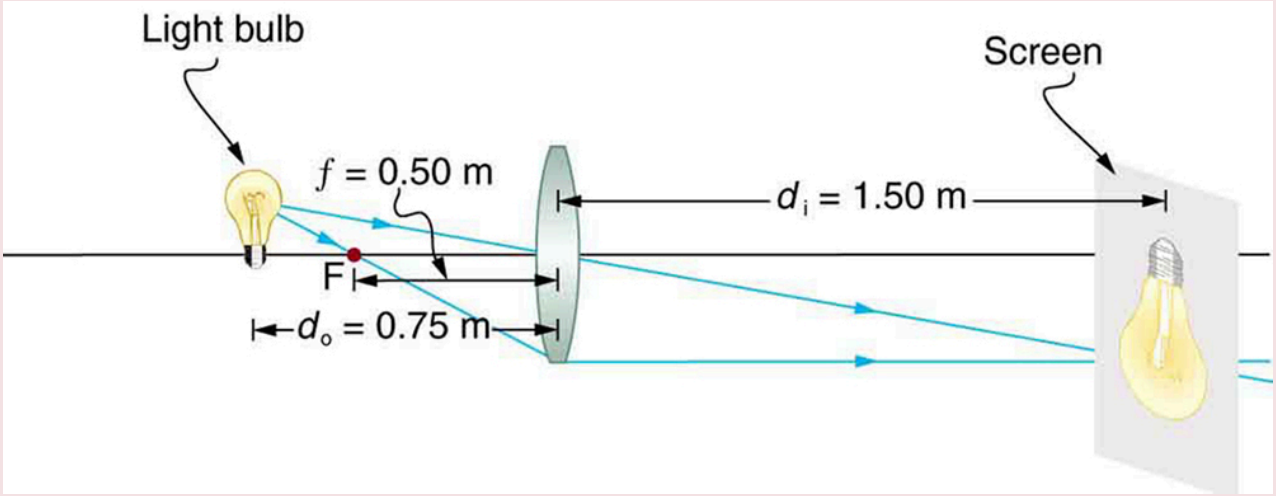


Figure 9. A light bulb placed 0.750 m from a lens having a 0.500 m focal length produces a real image on a poster board as discussed in the example above. Ray tracing predicts the image location and size.

Strategy and Concept

Since the object is placed farther away from a converging lens than the focal length of the lens, this situation is analogous to those illustrated in Figure 7 and Figure 8. Ray tracing to scale should produce similar results for d_i. Numerical solutions for d_i and m can be obtained using the thin lens equations, noting that d_o = 0.750 m and f = 0.500 m.

Solutions (Ray Tracing)

The ray tracing to scale in Figure 9 shows two rays from a point on the bulb's filament crossing about 1.50 m on the far side of the lens. Thus the image distance d_i is about 1.50 m. Similarly, the image height based on

ray tracing is greater than the object height by about a factor of 2, and the image is inverted. Thus m is about –2. The minus sign indicates that the image is inverted.

The thin lens equations can be used to find d_i from the given information:

$$\frac{1}{d_\mathrm{o}} + \frac{1}{d_\mathrm{i}} = \frac{1}{f}$$

.

Rearranging to isolate d_i gives

$$\frac{1}{d_\mathrm{i}} = \frac{1}{f} - \frac{1}{d_\mathrm{o}}$$

.

Entering known quantities gives a value for

$$\frac{1}{d_\mathrm{i}}$$

:

$$\frac{1}{d_\mathrm{i}} = \frac{1}{0.500\ \mathrm{m}} - \frac{1}{0.750\ \mathrm{m}} = \frac{0.667}{\mathrm{m}}$$

.

This must be inverted to find d_i:

$$d_\mathrm{i} = \frac{\mathrm{m}}{0.667} = 1.50\ \mathrm{m}$$

.

Note that another way to find d_i is to rearrange the equation:

$$\frac{1}{d_\mathrm{i}} = \frac{1}{f} - \frac{1}{d_\mathrm{o}}$$

.

This yields the equation for the image distance as:

$$d_\mathrm{i} = \frac{f d_\mathrm{o}}{d_\mathrm{o} - f}$$

Note that there is no inverting here.

The thin lens equations can be used to find the magnification m, since both d_i and d_o are known. Entering their values gives

$$m = -\frac{d_\mathrm{i}}{d_\mathrm{o}} = -\frac{1.50\ \mathrm{m}}{0.750\ \mathrm{m}} = -2.00$$

Discussion

Note that the minus sign causes the magnification to be negative when the image is inverted. Ray tracing and the use of the thin lens equations produce consistent results. The thin lens equations give the most precise results, being limited only by the accuracy of the given information. Ray tracing is limited by the accuracy with which you can draw, but it is highly useful both conceptually and visually.

Real images, such as the one considered in the previous example, are formed by converging lenses whenever an object is farther from the lens than its focal length. This is true for movie projectors, cameras, and the eye. We shall refer to these as *case 1* images. A case 1 image is formed when $d_o > f$ and f is positive, as in Figure 10a. (A summary of the three cases or types of image formation appears at the end of this section.)

A different type of image is formed when an object, such as a person's face, is held close to a convex lens. The image is upright and larger than the object, as seen in Figure 10b, and so the lens is called a magnifier. If you slowly pull the magnifier away from the face, you will see that the magnification steadily increases until the image begins to blur. Pulling the magnifier even farther away produces an inverted image as seen in Figure 10a. The distance at which the image blurs, and beyond which it inverts, is the focal length of the lens. To use a convex lens as a magnifier, the object must be closer to the converging lens than its focal length. This is called a *case 2* image. A case 2 image is formed when $d_o < f$ and f is positive.

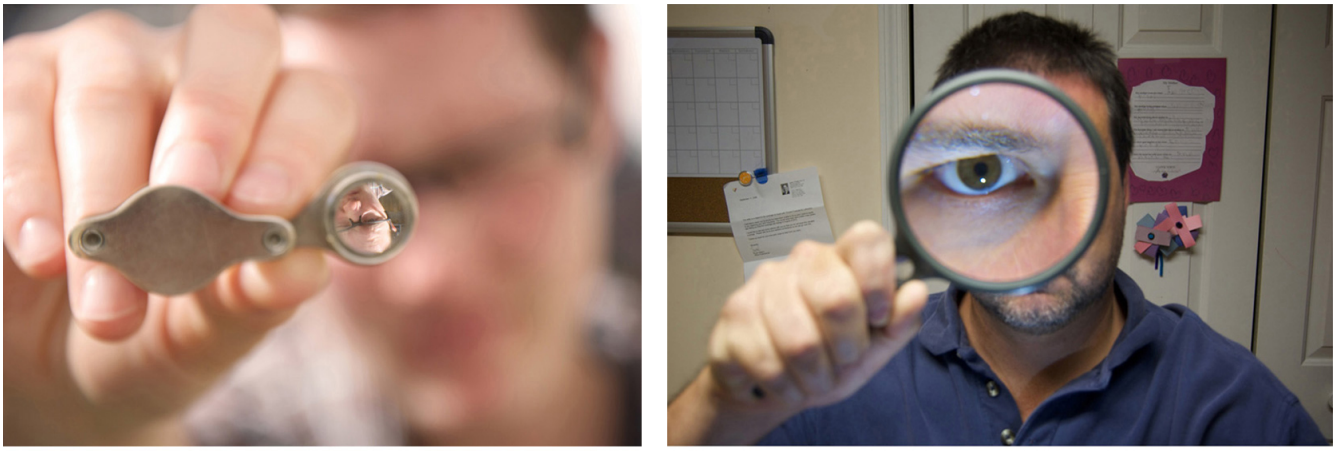


Figure 10. (a) When a converging lens is held farther away from the face than the lens's focal length, an inverted image is formed. This is a case 1 image. Note that the image is in focus but the face is not, because the image is much closer to the camera taking this photograph than the face. (credit: DaMongMan, Flickr) (b) A magnified image of a face is produced by placing it closer to the converging lens than its focal length. This is a case 2 image. (credit: Casey Fleser, Flickr)

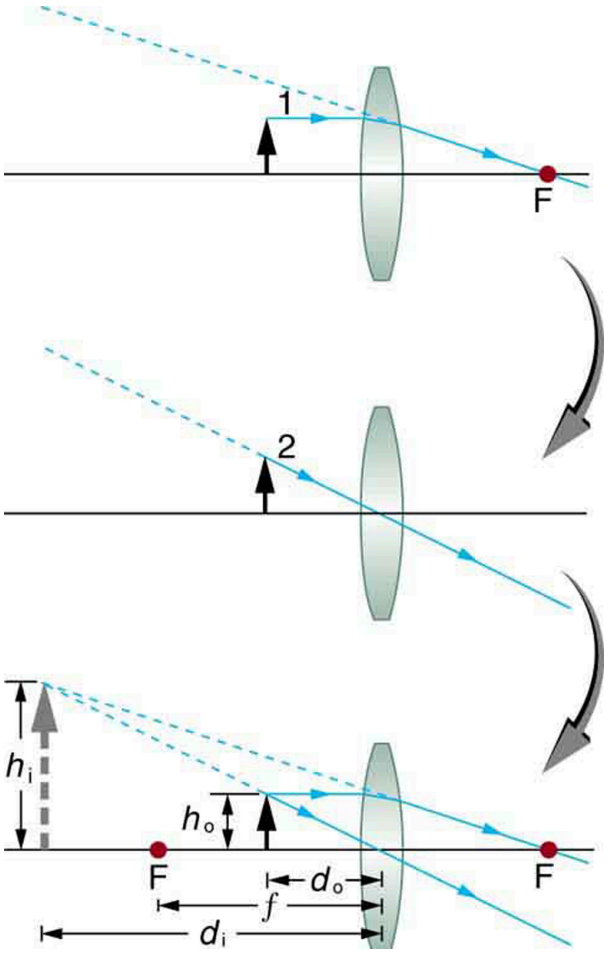


Figure 11. Ray tracing predicts the image location and size for an object held closer to a converging lens than its focal length. Ray 1 enters parallel to the axis and exits through the focal point on the opposite side, while ray 2 passes through the center of the lens without changing path. The two rays continue to diverge on the other side of the lens, but both appear to come from a common point, locating the upright, magnified, virtual image. This is a case 2 image.

Figure 11 uses ray tracing to show how an image is formed when an object is held closer to a converging lens than its focal length. Rays coming from a common point on the object continue to diverge after passing through the lens, but all appear to originate from a point at the location of the image. The image is on the same side of the lens as the object and is farther away from the lens than the object. This image, like all case 2 images, cannot be projected and, hence, is called a *virtual image*.

Light rays only appear to originate at a virtual image; they do not actually pass through that location in space. A screen placed at the location of a virtual image will receive only diffuse light from the object, not focused rays from the lens. Additionally, a screen placed on the opposite side of the lens will receive rays that are still diverging, and so no image will be projected on it. We can see the magnified image with our eyes, because the lens of the eye converges the rays into a real image projected on our retina. Finally, we note that a virtual image is upright and larger than the object, meaning that the magnification is positive and greater than 1.

Virtual Image

An image that is on the same side of the lens as the object and cannot be projected on a screen is called a virtual image.

Example 3. Image Produced by a Magnifying Glass

Suppose the book page in Figure 11a is held 7.50 cm from a convex lens of focal length 10.0 cm, such as a typical magnifying glass might have. What magnification is produced?

Strategy and Concept

We are given that d_o = 7.50 cm and f = 10.0 cm, so we have a situation where the object is placed closer to the lens than its focal length. We therefore expect to get a case 2 virtual image with a positive magnification that is greater than 1. Ray tracing produces an image like that shown in Figure 11, but we will use the thin lens equations to get numerical solutions in this example.

Solution

To find the magnification m, we try to use magnification equation,

$$m = -\frac{d_i}{d_o}$$

. We do not have a value for d_i, so that we must first find the location of the image using lens equation. (The procedure is the same as followed in the preceding example, where d_o and f were known.) Rearranging the magnification equation to isolate d_i gives

$$\frac{1}{d_i} = \frac{1}{f} - \frac{1}{d_o}$$

.

Entering known values, we obtain a value for

$$\frac{1}{d_i}$$

:

$$\frac{1}{d_i} = \frac{1}{10.0\ \text{cm}} - \frac{1}{7.50\ \text{cm}} = \frac{-0.0333}{\text{cm}}$$

.

This must be inverted to find d_i:

$$d_i = -\frac{\text{cm}}{0.0333} = -30.0\ \text{cm}$$

.

Now the thin lens equation can be used to find the magnification m, since both d_i and d_o are known. Entering their values gives

$$m = -\frac{d_\text{i}}{d_\text{o}} = -\frac{-30.0\text{ cm}}{10.0\text{ cm}} = 3.00$$

Discussion

A number of results in this example are true of all case 2 images, as well as being consistent with Figure 11. Magnification is indeed positive (as predicted), meaning the image is upright. The magnification is also greater than 1, meaning that the image is larger than the object—in this case, by a factor of 3. Note that the image distance is negative. This means the image is on the same side of the lens as the object. Thus the image cannot be projected and is virtual. (Negative values of d_i occur for virtual images.) The image is farther from the lens than the object, since the image distance is greater in magnitude than the object distance. The location of the image is not obvious when you look through a magnifier. In fact, since the image is bigger than the object, you may think the image is closer than the object. But the image is farther away, a fact that is useful in correcting farsightedness, as we shall see in a later section.

A third type of image is formed by a diverging or concave lens. Try looking through eyeglasses meant to correct nearsightedness. (See Figure 12.) You will see an image that is upright but smaller than the object. This means that the magnification is positive but less than 1. The ray diagram in Figure 13 shows that the image is on the same side of the lens as the object and, hence, cannot be projected—it is a virtual image. Note that the image is closer to the lens than the object. This is a *case 3* image, formed for any object by a negative focal length or diverging lens.

Figure 12. A car viewed through a concave or diverging lens looks upright. This is a case 3 image. (credit: Daniel Oines, Flickr)

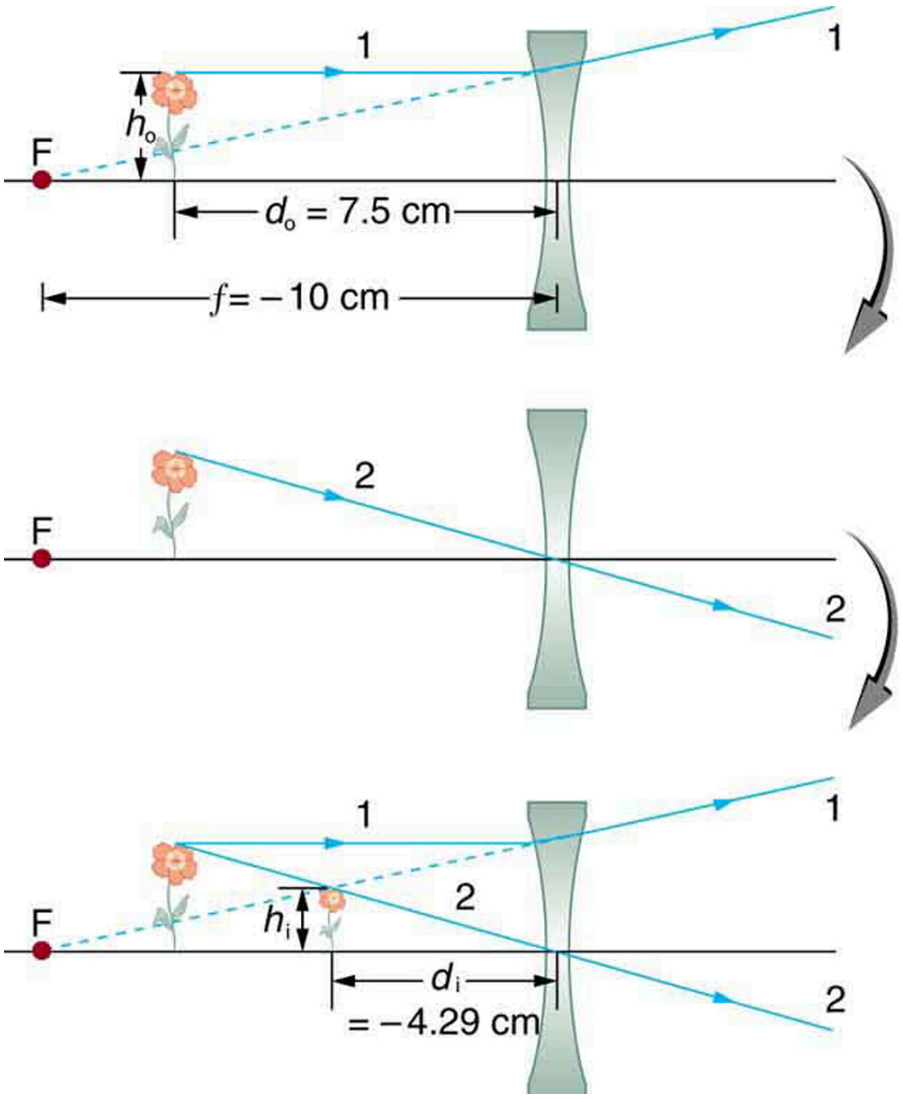


Figure 13. Ray tracing predicts the image location and size for a concave or diverging lens. Ray 1 enters parallel to the axis and is bent so that it appears to originate from the focal point. Ray 2 passes through the center of the lens without changing path. The two rays appear to come from a common point, locating the upright image. This is a case 3 image, which is closer to the lens than the object and smaller in height.

Example 4. Image Produced by a Concave Lens

Suppose an object such as a book page is held 7.50 cm from a concave lens of focal length –10.0 cm. Such a lens could be used in eyeglasses to correct pronounced nearsightedness. What magnification is produced?

Strategy and Concept

This example is identical to the preceding one, except that the focal length is negative for a concave or diverging lens. The method of solution is thus the same, but the results are different in important ways.

Solution

To find the magnification *m*, we must first find the image distance d_i using thin lens equation

$$\frac{1}{d_\text{i}} = \frac{1}{f} - \frac{1}{d_\text{o}}$$

, or its alternative rearrangement

$$d_\text{i} = \frac{f d_\text{o}}{d_\text{o} - f}$$

.

We are given that f = –10.0 cm and d_o = 7.50 cm. Entering these yields a value for

$$\frac{1}{d_\text{i}}$$

:

$$\frac{1}{d_\text{i}} = \frac{1}{-10.0\text{ cm}} - \frac{1}{7.50\text{ cm}} = \frac{-0.2333}{\text{cm}}$$

This must be inverted to find d_i:

$$d_\text{i} = -\frac{\text{cm}}{0.2333} = -4.29\text{ cm}$$

Or

$$d_\text{i} = \frac{(7.5)(-10)}{(7.5 - (-10))} = -\frac{75}{17.5} = -4.29\text{ cm}$$

Now the magnification equation can be used to find the magnification *m*, since both d_i and d_o are known. Entering their values gives

$$m = -\frac{d_\text{i}}{d_\text{o}} = -\frac{-4.29\text{ cm}}{7.50\text{ cm}} = 0.571$$

Discussion

A number of results in this example are true of all case 3 images, as well as being consistent with Figure 13. Magnification is positive (as predicted), meaning the image is upright. The magnification is also less than 1, meaning the image is smaller than the object—in this case, a little over half its size. The image distance is negative, meaning the image is on the same side of the lens as the object. (The image is virtual.) The image is closer to the lens than the object, since the image distance is smaller in magnitude than the object distance. The location of the image is not obvious when you look through a concave lens. In fact, since the image is smaller than the object, you may think it is farther away. But the image is closer than the object, a fact that is useful in correcting nearsightedness, as we shall see in a later section.

Table 1 summarizes the three types of images formed by single thin lenses. These are referred to as case 1, 2, and 3 images. Convex (converging) lenses can form either real or virtual images (cases 1 and 2, respectively), whereas concave (diverging) lenses can form only virtual images (always case 3). Real images are always inverted, but they can be either larger or smaller than the object. For example, a slide

projector forms an image larger than the slide, whereas a camera makes an image smaller than the object being photographed. Virtual images are always upright and cannot be projected. Virtual images are larger than the object only in case 2, where a convex lens is used. The virtual image produced by a concave lens is always smaller than the object—a case 3 image. We can see and photograph virtual images only by using an additional lens to form a real image.

Table 1. Three Types of Images Formed By Thin Lenses

Type	Formed when	Image type	d_i	m
Case 1	f positive, $d_o > f$	real	positive	negative
Case 2	f positive, $d_o < f$	virtual	negative	positive $m > 1$
Case 3	f negative	virtual	negative	positive $m < 1$

In Image Formation by Mirrors, we shall see that mirrors can form exactly the same types of images as lenses.

Take-Home Experiment: Concentrating Sunlight

Find several lenses and determine whether they are converging or diverging. In general those that are thicker near the edges are diverging and those that are thicker near the center are converging. On a bright sunny day take the converging lenses outside and try focusing the sunlight onto a piece of paper. Determine the focal lengths of the lenses. Be careful because the paper may start to burn, depending on the type of lens you have selected.

Problem-Solving Strategies for Lenses

Step 1. Examine the situation to determine that image formation by a lens is involved.

Step 2. Determine whether ray tracing, the thin lens equations, or both are to be employed. A sketch is very useful even if ray tracing is not specifically required by the problem. Write symbols and values on the sketch.

Step 3. Identify exactly what needs to be determined in the problem (identify the unknowns).

Step 4. Make alist of what is given or can be inferred from the problem as stated (identify the knowns). It is helpful to determine whether the situation involves a case 1, 2, or 3 image. While these are just names for types of images, they have certain characteristics (given in Table 1) that can be of great use in solving problems.

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 the use of the thin lens equations. These are solved in the usual manner by substituting knowns and solving for unknowns. Several worked examples serve as guides.

Step 7. Check to see if the answer is reasonable: Does it make sense? If you have identified the type of image (case 1, 2, or 3), you should assess whether your answer is consistent with the type of image, magnification, and so on.

Misconception Alert

We do not realize that light rays are coming from every part of the object, passing through every part of the lens, and all can be used to form the final image.

We generally feel the entire lens, or mirror, is needed to form an image. Actually, half a lens will form the same, though a fainter, image.

Section Summary

- Light rays entering a converging lens parallel to its axis cross one another at a single point on the opposite side.
- For a converging lens, the focal point is the point at which converging light rays cross; for a diverging lens, the focal point is the point from which diverging light rays appear to originate.
- The distance from the center of the lens to its focal point is called the focal length *f*.
  Power *P* of a lens is defined to be the inverse of its focal length, $P = \frac{1}{f}$.
  A lens that causes the light rays to bend away from its axis is called a diverging lens.
  Ray tracing is the technique of graphically determining the paths that light rays take.
  The image in which light rays from one point on the object actually cross at the location of the image and can be projected onto a screen, a piece of film, or the retina of an eye is called a real image.
- Thin lens equations are

$$\frac{1}{d_o} + \frac{1}{d_i} = \frac{1}{f}$$

and

$$\frac{h_i}{h_o} = -\frac{d_i}{d_o} = m$$

(magnification).

- The distance of the image from the center of the lens is called image distance.
  An image that is on the same side of the lens as the object and cannot be projected on a screen is called a virtual image.

Conceptual Questions

1. It can be argued 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?

2. You can often see a reflection when looking at a sheet of glass, particularly if it is darker on the other side. Explain why you can often see a double image in such circumstances.
3. 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?
4. A thin lens has two focal points, one on either side, at equal distances from its center, and should behave the same for light entering from either side. Look through your eyeglasses (or those of a friend) backward and forward and comment on whether they are thin lenses.
5. Will the focal length of a lens change when it is submerged in water? Explain.

Problems & Exercises

1. What is the power in diopters of a camera lens that has a 50.0 mm focal length?
2. Your camera's zoom lens has an adjustable focal length ranging from 80.0 to 200 mm. What is its range of powers?
3. What is the focal length of 1.75 D reading glasses found on the rack in a pharmacy?
4. You note that your prescription for new eyeglasses is –4.50 D. What will their focal length be?
5. 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 for lenses.
6. 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 for Lenses* (above).
7. 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?
8. How far from a piece of paper must you hold your father's 2.25 D reading glasses to try to burn a hole in the paper with sunlight?
9. 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.
10. 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?
11. 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?
12. (a) 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? (b) Calculate the power of the magnifier in diopters. (c) Discuss how this power compares to those for store-bought reading glasses (typically 1.0 to 4.0 D). Is the magnifier's power greater, and should it be?

13. What magnification will be produced by a lens of power –4.00 D (such as might be used to correct myopia) if an object is held 25.0 cm away?
14. In Example 3, the magnification of a book held 7.50 cm from a 10.0 cm focal length lens was found to be 3.00. (a) Find the magnification for the book when it is held 8.50 cm from the magnifier. (b) Do the same for when it is held 9.50 cm from the magnifier. (c) Comment on the trend in m as the object distance increases as in these two calculations.
15. 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?
16. A camera with a 100 mm focal length lens is used to photograph the sun and moon. 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?
17. Combine thin lens equations to show that the magnification for a thin lens is determined by its focal length and the object distance and is given by $m = \frac{f}{(f - d_o)}$.

Glossary

converging lens: a convex lens in which light rays that enter it parallel to its axis converge at a single point on the opposite side

diverging lens: a concave lens in which light rays that enter it parallel to its axis bend away (diverge) from its 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

focal length: distance from the center of a lens or curved mirror to its focal point

magnification: ratio of image height to object height

power: inverse of focal length

real image: image that can be projected

virtual image: image that cannot be projected

Selected Solutions to Problems & Exercises

2. 5.00 to 12.5 D

4. –0.222 m

6. (a) 3.43 m; (b) 0.800 by 1.20 m

7. (a) −1.35 m (on the object side of the lens); (b) +10.0; (c) 5.00 cm

8. 44.4 cm

10. (a) 6.60 cm; (b) –0.333

12. (a) +7.50 cm; (b) 13.3 D; (c) Much greater

14. (a) +6.67; (b) +20.0; (c) The magnification increases without limit (to infinity) as the object distance increases to the limit of the focal distance.

16. −0.933 mm

Image Formation by Mirrors

Lumen Learning

Learning Objectives

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

- Illustrate image formation in a flat mirror.
- Explain with ray diagrams the formation of an image using spherical mirrors.
- Determine focal length and magnification given radius of curvature, distance of object and image.

We only have to look as far as the nearest bathroom to find an example of an image formed by a mirror. Images in flat mirrors are the same size as the object and are located behind the mirror. Like lenses, mirrors can form a variety of images. For example, dental mirrors may produce a magnified image, just as makeup mirrors do. Security mirrors in shops, on the other hand, form images that are smaller than the object. We will use the law of reflection to understand how mirrors form images, and we will find that mirror images are analogous to those formed by lenses.

Figure 1 helps illustrate how a flat mirror forms an image. Two rays are shown emerging from the same point, striking the mirror, and being reflected into the observer's eye. The rays can diverge slightly, and both still get into the eye. If the rays are extrapolated backward, they seem to originate from a common point behind the mirror, locating the image. (The paths of the reflected rays into the eye are the same as if they had come directly from that point behind the mirror.) Using the law of reflection—the angle of reflection equals the angle of incidence—we can see that the image and object are the same distance from the mirror. This is a virtual image, since it cannot be projected—the rays only appear to originate from a common point behind the mirror. Obviously, if you walk behind the mirror, you cannot see the image, since the rays do not go there. But in front of the mirror, the rays behave exactly as if they had come from behind the mirror, so that is where the image is situated.

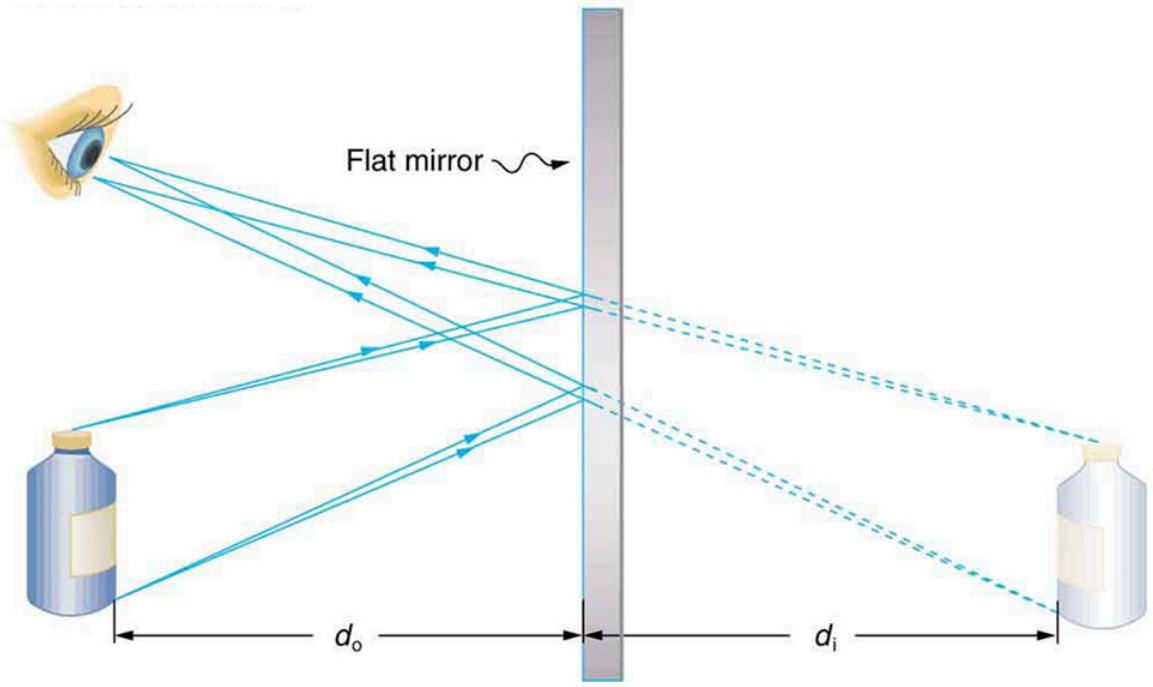


Figure 1. Two sets of rays from common points on an object are reflected by a flat mirror into the eye of an observer. The reflected rays seem to originate from behind the mirror, locating the virtual image.

Now let us consider the focal length of a mirror—for example, the concave spherical mirrors in Figure 2. Rays of light that strike the surface follow the law of reflection. For a mirror that is large compared with its radius of curvature, as in Figure 2a, we see that the reflected rays do not cross at the same point, and the mirror does not have a well-defined focal point. If the mirror had the shape of a parabola, the rays would all cross at a single point, and the mirror would have a well-defined focal point. But parabolic mirrors are much more expensive to make than spherical mirrors. The solution is to use a mirror that is small compared with its radius of curvature, as shown in Figure 2b. (This is the mirror equivalent of the thin lens approximation.) To a very good approximation, this mirror has a well-defined focal point at F that is the focal distance *f* from the center of the mirror. The focal length *f* of a concave mirror is positive, since it is a converging mirror.

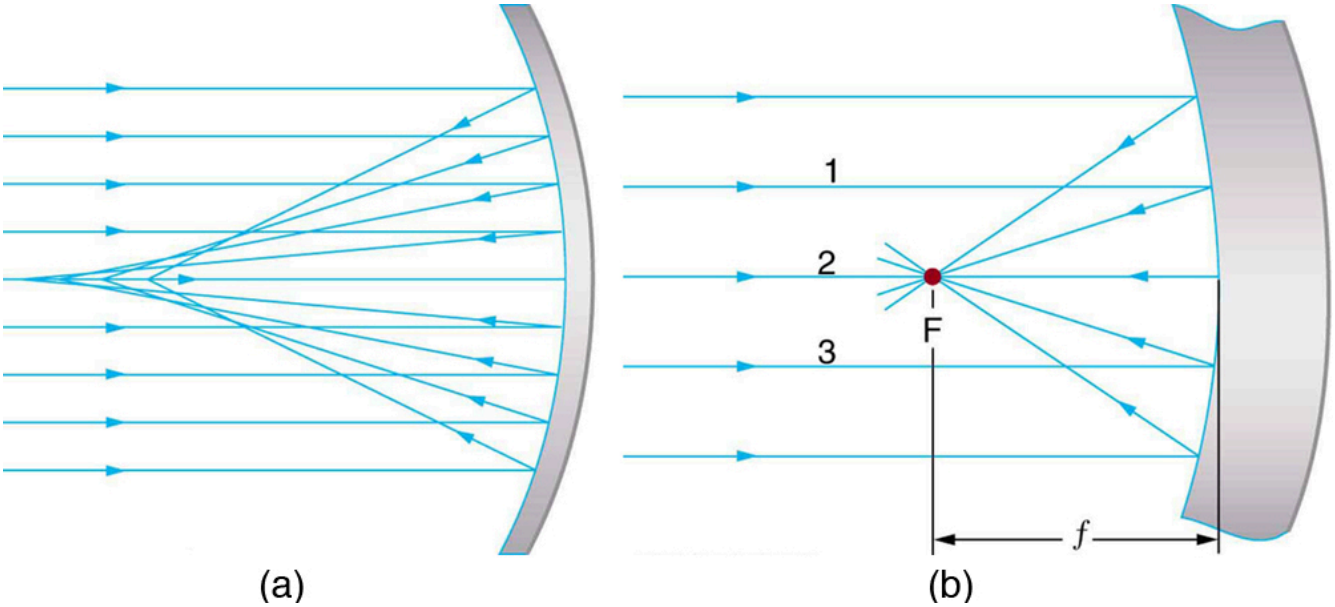


Figure 2. (a) Parallel rays reflected from a large spherical mirror do not all cross at a common point. (b) If a spherical mirror is small compared with its radius of curvature, parallel rays are focused to a common point. The distance of the focal point from the center of the mirror is its focal length f. Since this mirror is converging, it has a positive focal length.

Just as for lenses, the shorter the focal length, the more powerful the mirror; thus,

$$P = \frac{1}{f}$$

for a mirror, too. A more strongly curved mirror has a shorter focal length and a greater power. Using the law of reflection and some simple trigonometry, it can be shown that the focal length is half the radius of curvature, or

$$f = \frac{R}{2}$$

, where *R* is the radius of curvature of a spherical mirror. The smaller the radius of curvature, the smaller the focal length and, thus, the more powerful the mirror

The convex mirror shown in Figure 3 also has a focal point. Parallel rays of light reflected from the mirror seem to originate from the point F at the focal distance *f* behind the mirror. The focal length and power of a convex mirror are negative, since it is a diverging mirror.

Ray tracing is as useful for mirrors as for lenses. The rules for ray tracing for mirrors are based on the illustrations just discussed:

1. A ray approaching a concave converging mirror parallel to its axis is reflected through the focal point F of the mirror on the same side. (See rays 1 and 3 in Figure 2b.)
2. A ray approaching a convex diverging mirror parallel to its axis is reflected so that it seems to come from the focal point F behind the mirror. (See rays 1 and 3 in Figure 3.)
3. Any ray striking the center of a mirror is followed by applying the law of reflection; it makes the same angle with the axis when leaving as when approaching. (See ray 2 in Figure 4.)
4. A ray approaching a concave converging mirror through its focal point is reflected parallel to its axis. (The reverse of rays 1 and 3 in Figure 2.)
5. A ray approaching a convex diverging mirror by heading toward its focal point on the opposite side is reflected parallel to the axis. (The reverse of rays 1 and 3 in Figure 3.)

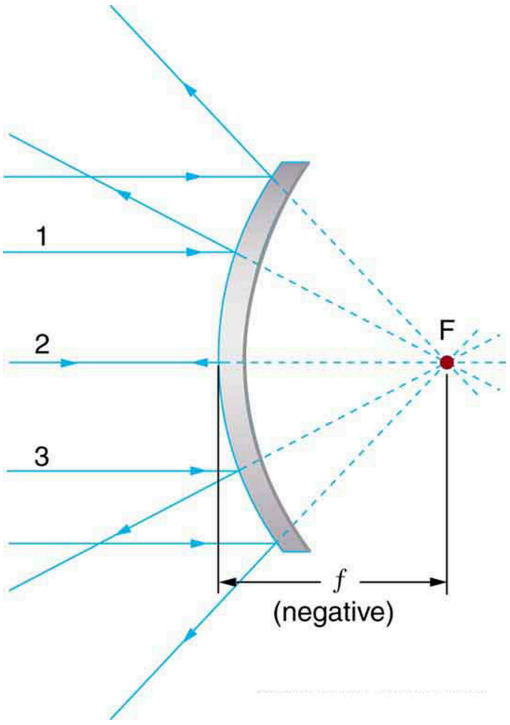


Figure 3. Parallel rays of light reflected from a convex spherical mirror (small in size compared with its radius of curvature) seem to originate from a well-defined focal point at the focal distance f behind the mirror. Convex mirrors diverge light rays and, thus, have a negative focal length.

We will use ray tracing to illustrate how images are formed by mirrors, and we can use ray tracing quantitatively to obtain numerical information. But since we assume each mirror is small compared with its radius of curvature, we can use the thin lens equations for mirrors just as we did for lenses.

Consider the situation shown in Figure 4, concave spherical mirror reflection, in which an object is placed farther from a concave (converging) mirror than its focal length. That is, f is positive and $d_o > f$, so that we may expect an image similar to the case 1 real image formed by a converging lens. Ray tracing in Figure 4 shows that the rays from a common point on the object all cross at a point on the same side of the mirror as the object. Thus a real image can be projected onto a screen placed at this location. The image distance is positive, and the image is inverted, so its magnification is negative. This is a *case 1 image for mirrors*. It differs from the case 1 image for lenses only in that the image is on the same side of the mirror as the object. It is otherwise identical.

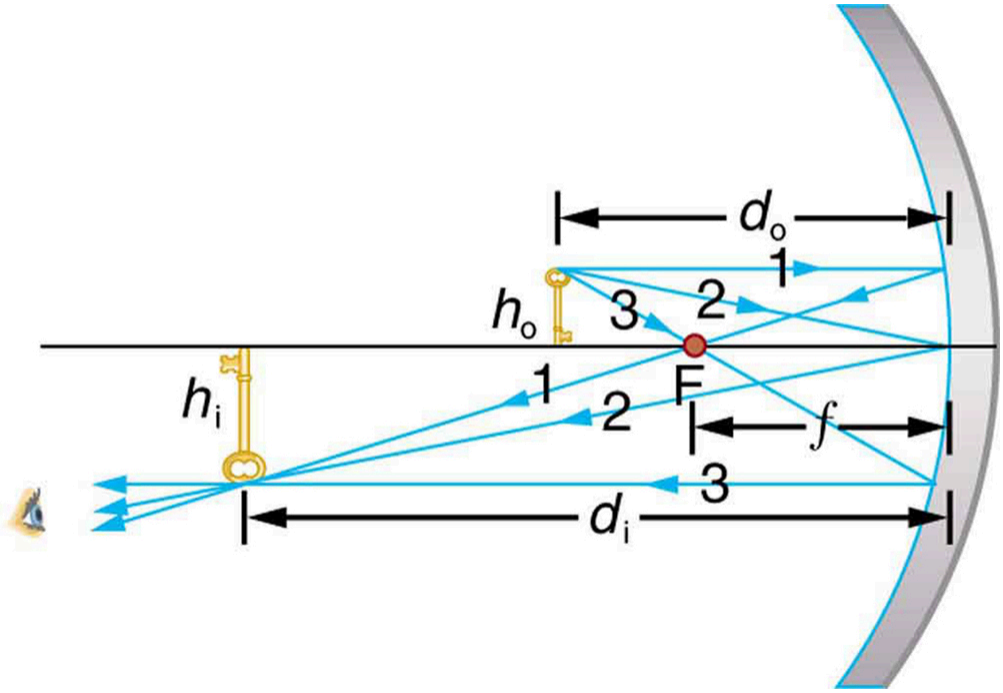


Figure 4. A case 1 image for a mirror. An object is farther from the converging mirror than its focal length. Rays from a common point on the object are traced using the rules in the text. Ray 1 approaches parallel to the axis, ray 2 strikes the center of the mirror, and ray 3 goes through the focal point on the way toward the mirror. All three rays cross at the same point after being reflected, locating the inverted real image. Although three rays are shown, only two of the three are needed to locate the image and determine its height.

Example 1. A Concave Reflector

Electric room heaters use a concave mirror to reflect infrared (IR) radiation from hot coils. Note that IR 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?

Strategy and Concept

We are given that the concave mirror projects a real image of the coils at an image distance d_i=3.00 m. The coils are the object, and we are asked to find their location—that is, to find the object distance d_o. We are also given the radius of curvature of the mirror, so that its focal length is

$$f = \frac{R}{2} = 25.0 \text{ cm}$$

(positive since the mirror is concave or converging). Assuming the mirror is small compared with its radius of curvature, we can use the thin lens equations, to solve this problem.

Solution

Since d_i and f are known, thin lens equation can be used to find d_o:

$$\frac{1}{d_o} + \frac{1}{d_i} = \frac{1}{f}$$

.

Rearranging to isolate d_o gives

$$\frac{1}{d_o} = \frac{1}{f} - \frac{1}{d_i}$$

.

Entering known quantities gives a value for

$$\frac{1}{d_o}$$

:

$$\frac{1}{d_o} = \frac{1}{0.250\ \text{m}} - \frac{1}{3.00\ \text{m}} = \frac{3.667}{\text{m}}$$

.

This must be inverted to find d_o:

$$d_o = \frac{1\ \text{m}}{3.667} = 27.3\ \text{cm}$$

.

Discussion

Note that the object (the filament) is farther from the mirror than the mirror's focal length. This is a case 1 image ($d_o > f$ and f positive), consistent with the fact that a real image is formed. You will get the most concentrated thermal energy directly in front of the mirror and 3.00 m away from it. Generally, this is not desirable, since it could cause burns. Usually, you want the rays to emerge parallel, and this is accomplished by having the filament at the focal point of the mirror.

Note that the filament here is not much farther from the mirror than its focal length and that the image produced is considerably farther away. This is exactly analogous to a slide projector. Placing a slide only slightly farther away from the projector lens than its focal length produces an image significantly farther away. As the object gets closer to the focal distance, the image gets farther away. In fact, as the object distance approaches the focal length, the image distance approaches infinity and the rays are sent out parallel to one another.

Example 2. Solar Electric Generating System

One of the solar technologies used today for generating electricity is a device (called a parabolic trough or concentrating collector) that concentrates the sunlight onto a blackened pipe that contains a fluid. This heated fluid is pumped to a heat exchanger, where its heat energy is transferred to another system that is used to generate steam—and so generate electricity through a conventional steam cycle. Figure 5 shows such a working system in southern California. Concave mirrors are used to concentrate the sunlight onto the pipe. The mirror has the approximate shape of a section of a cylinder. For the problem, assume that the mirror is exactly one-quarter of a full cylinder.

1. If we wish to place the fluid-carrying pipe 40.0 cm from the concave mirror at the mirror's focal

point, what will be the radius of curvature of the mirror?

2. Per meter of pipe, what will be the amount of sunlight concentrated onto the pipe, assuming the insolation (incident solar radiation) is 0.900 k W/m^2?
3. If the fluid-carrying pipe has a 2.00-cm diameter, what will be the temperature increase of the fluid per meter of pipe over a period of one minute? Assume all the solar radiation incident on the reflector is absorbed by the pipe, and that the fluid is mineral oil.

Strategy

To solve an *Integrated Concept Problem* we must first identify the physical principles involved. Part 1 is related to the current topic. Part 2 involves a little math, primarily geometry. Part 3 requires an understanding of heat and density.

Solution to Part 1

To a good approximation for a concave or semi-spherical surface, the point where the parallel rays from the sun converge will be at the focal point, so $R = 2f = 80.0$ cm.

Solution to Part 2

The insolation is 900 W /m^2. We must find the cross-sectional area A of the concave mirror, since the power delivered is 900 W /m^2 × A. The mirror in this case is a quarter-section of a cylinder, so the area for a length

$$\mathrm{A} = \tfrac{1}{4}\left(2\pi \mathrm{R}\right)\mathrm{L}$$

L of the mirror is . The area for a length of 1.00 m is then

$$\mathrm{A} = \frac{\pi}{2} R\left(1.00\ \mathrm{m}\right) = \frac{(3.14)}{2}\left(0.800\ \mathrm{m}\right)\left(1.00\ \mathrm{m}\right) = 1.26^2$$

The insolation on the 1.00-m length of pipe is then

$$\left(9.00 \times 10^2 \frac{\mathrm{W}}{\mathrm{m}^2}\right)\left(1.26\ \mathrm{m}^2\right) = 1130\ \mathrm{W}$$

Solution to Part 3

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

$$\begin{aligned} m &= \rho \mathrm{V} = \rho\pi\left(\tfrac{d}{2}\right)^2\left(1.00\ \mathrm{m}\right) \\ &= \left(8.00 \times 10^2\ \mathrm{kg/m}^3\right)(3.14)\left(0.0100\ \mathrm{m}\right)^2\left(1.00\ \mathrm{m}\right) \\ &= 0.251\ \mathrm{kg} \end{aligned}$$

Therefore, the increase in temperature in one minute is

$$\begin{aligned} \Delta T &= \tfrac{Q}{mc} \\ &= \frac{(1130\ \mathrm{W})(60.0\ \mathrm{s})}{(0.251\ \mathrm{kg})(1670\ \mathrm{J\cdot\ kg/^\circ\ C})} \\ &= 162^\circ\mathrm{C} \end{aligned}$$

Discussion for Part 3

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 400ºC. We are considering only one meter of pipe here, and ignoring heat losses along the pipe.

Figure 5. Parabolic trough collectors are used to generate electricity in southern California. (credit: kjkolb, Wikimedia Commons)

What happens if an object is closer to a concave mirror than its focal length? This is analogous to a case 2 image for lenses ($d_o < f$ and f positive), which is a magnifier. In fact, this is how makeup mirrors act as magnifiers. Figure 6a uses ray tracing to locate the image of an object placed close to a concave mirror. Rays from a common point on the object are reflected in such a manner that they appear to be coming from behind the mirror, meaning that the image is virtual and cannot be projected. As with a magnifying glass, the image is upright and larger than the object. This is a *case 2 image for mirrors* and is exactly analogous to that for lenses.

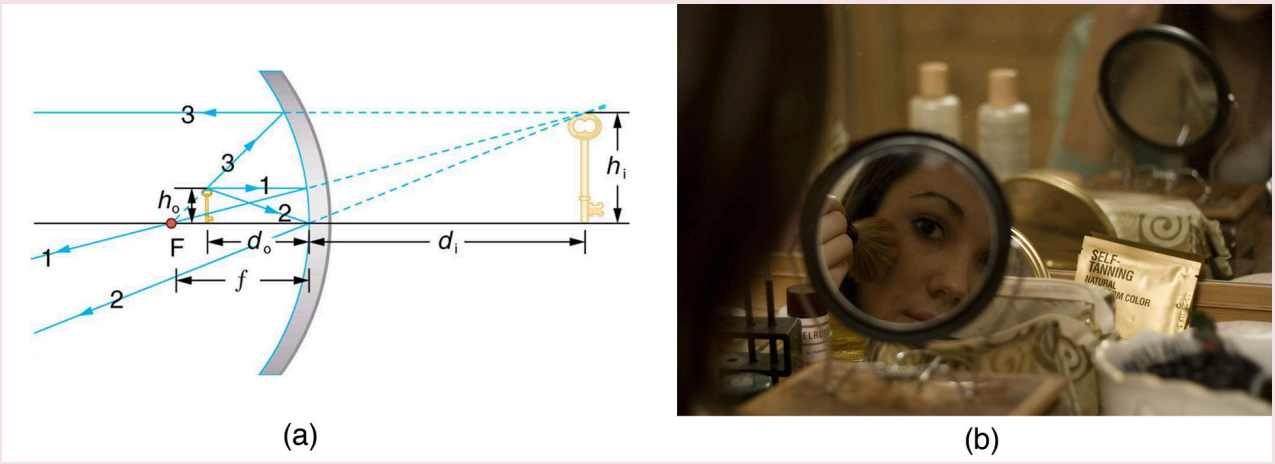


Figure 6. (a) Case 2 images for mirrors are formed when a converging mirror has an object closer to it than its focal length. Ray 1 approaches parallel to the axis, ray 2 strikes the center of the mirror, and ray 3 approaches the mirror as if it came from the focal point. (b) A magnifying mirror showing the reflection. (credit: Mike Melrose, Flickr)

All three rays appear to originate from the same point after being reflected, locating the upright virtual image behind the mirror and showing it to be larger than the object. (b) Makeup mirrors are perhaps the most common use of a concave mirror to produce a larger, upright image.

A convex mirror is a diverging mirror (f is negative) and forms only one type of image. It is a *case 3* image—one that is upright and smaller than the object, just as for diverging lenses. Figure 7a uses ray tracing to illustrate the location and size of the case 3 image for mirrors. Since the image is behind the mirror, it cannot be projected and is thus a virtual image. It is also seen to be smaller than the object.

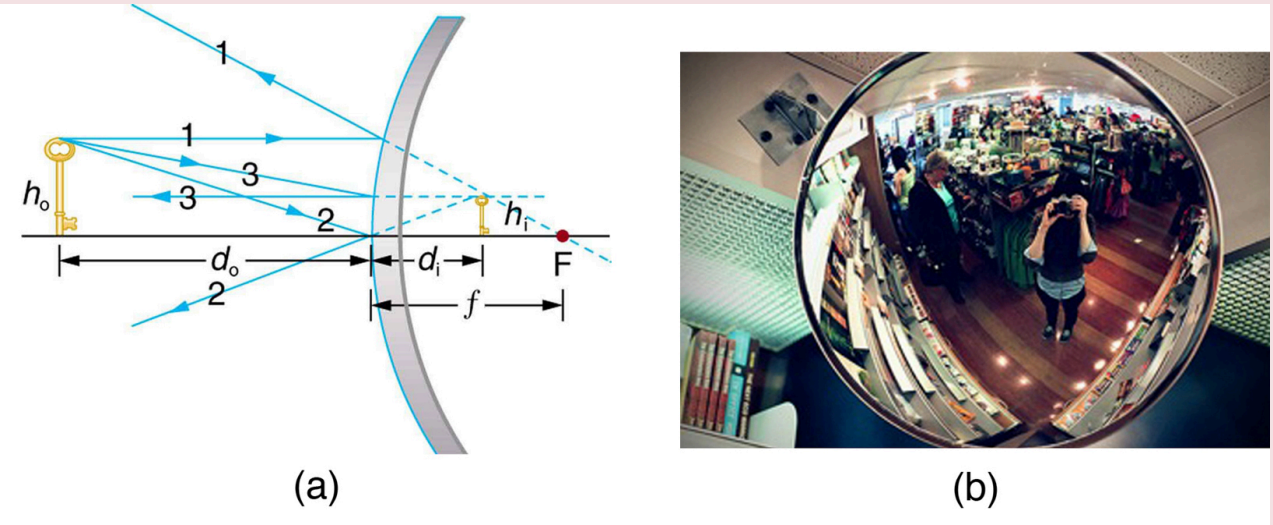


Figure 7. Case 3 images for mirrors are formed by any convex mirror. Ray 1 approaches parallel to the axis, ray 2 strikes the center of the mirror, and ray 3 approaches toward the focal point. All three rays appear to originate from the same point after being reflected, locating the upright virtual image behind the mirror and showing it to be smaller than the object. (b) Security mirrors are convex, producing a smaller, upright image. Because the image is smaller, a larger area is imaged compared to what would be observed for a flat mirror (and hence security is improved). (credit: Laura D'Alessandro, Flickr)

Example 3. Image in a Convex Mirror

A keratometer is a device used to measure the curvature of the cornea, 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 12.0 cm from the cornea and the image's magnification is 0.0320, what is the cornea's radius of curvature?

Strategy

If we can find the focal length of the convex mirror formed by the cornea, we can find its radius of curvature (the radius of curvature is twice the focal length of a spherical mirror). We are given that the object distance is d_o = 12.0 cm and that m = 0.0320. We first solve for the image distance d_i, and then for f.

Solution

$$m = -\frac{d_i}{d_o}$$

. Solving this expression for d_i gives $d_i = -md_o$.

Entering known values yields $d_i = -(0.0320)(12.0\text{ cm}) = -0.384$ cm.

$$\frac{1}{f} = \frac{1}{d_o} + \frac{1}{d_i}$$

Substituting known values,

$$\frac{1}{f} = \frac{1}{12.0\ \text{cm}} + \frac{1}{-0.384\ \text{cm}} = \frac{-2.52}{\text{cm}}$$

.

This must be inverted to find *f*:

$$f = \frac{\text{cm}}{-2.52} = -0.400\ \text{cm}$$

The radius of curvature is twice the focal length, so that $R = 2|f| = 0.800$ cm.

Discussion

Although the focal length *f* of a convex mirror is defined to be negative, we take the absolute value to give us a positive value for *R*. 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, many corneas are not spherical, complicating 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, where it cannot be projected. In this section's Problems and Exercises, you will show that for a fixed object distance, the smaller the radius of curvature, the smaller the magnification.

The three types of images formed by mirrors (cases 1, 2, and 3) are exactly analogous to those formed by lenses, as summarized in the table at the end of Image Formation by Lenses. It is easiest to concentrate on only three types of images—then remember that concave mirrors act like convex lenses, whereas convex mirrors act like concave lenses.

Take-Home Experiment: Concave Mirrors Close to Home

Find a flashlight and identify the curved mirror used in it. Find another flashlight and shine the first flashlight onto the second one, which is turned off. Estimate the focal length of the mirror. You might try shining a flashlight on the curved mirror behind the headlight of a car, keeping the headlight switched off, and determine its focal length.

Problem-Solving Strategy for Mirrors

Step 1. Examine the situation to determine that image formation by a mirror is involved.

Step 2. Refer to the Problem-Solving Strategies for Lenses. The same strategies are valid for mirrors as for lenses with one qualification—use the ray tracing rules for mirrors listed earlier in this section.

Section Summary

- The characteristics of an image formed by a flat mirror are: (a) The image and object are the same distance from the mirror, (b) The image is a virtual image, and (c) The image is situated behind the mirror.

$$f = \frac{R}{2}$$

- Image length is half the radius of curvature:
- A convex mirror is a diverging mirror and forms only one type of image, namely a virtual image.

Conceptual Questions

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? Can you photograph one? Can one be projected onto a screen with additional lenses or mirrors? Explain your responses.
3. Is it necessary to project a real image onto a screen for it to exist?
4. At what distance is an image always located—at d_o, d_i, or f?
5. Under what circumstances will an image be located at the focal point of a lens or mirror?
6. What is meant by a negative magnification? What is meant by a magnification that is less than 1 in magnitude?
7. Can a case 1 image be larger than the object even though its magnification is always negative? Explain.
8. Figure 8 shows a light bulb between two 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?

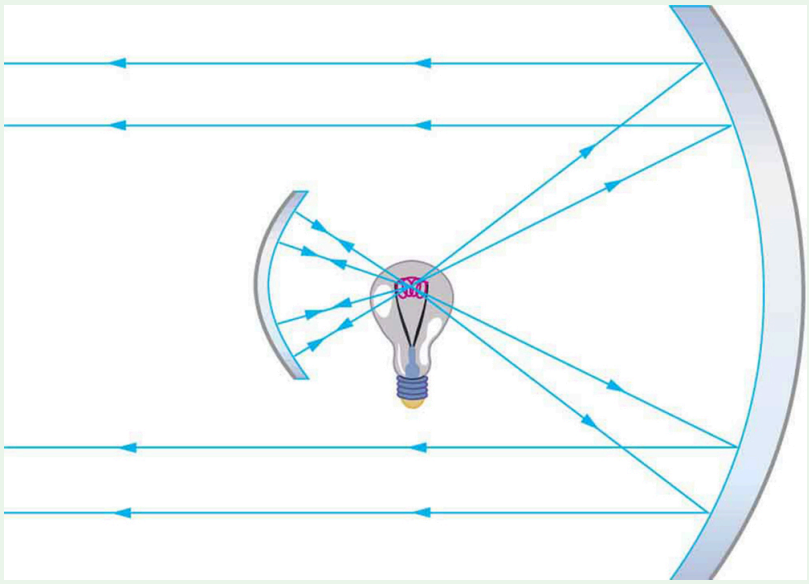


Figure 8. The two mirrors trap most of the bulb's light and form a directional beam as in a headlight.

9. The two mirrors trap most of the bulb's light and form a directional beam as in a headlight.
10. Two concave mirrors of different sizes are placed facing one another. A filament bulb is placed at the focus of the larger mirror. The rays after reflection from the larger mirror travel parallel to one another. The rays falling on the smaller mirror retrace their paths.

11. 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?
12. 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.
13. It can be argued that a flat mirror has an infinite focal length. If so, where does it form an image? That is, how are d_i and d_o related?
14. Why are diverging mirrors often used for rear-view mirrors in vehicles? What is the main disadvantage of using such a mirror compared with a flat one?

Problems & Exercises

1. What is the focal length of a makeup mirror that has a power of 1.50 D?
2. 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?
3. (a) Calculate the focal length of the mirror formed by the shiny back of a spoon that has a 3.00 cm radius of curvature. (b) What is its power in diopters?
4. Find the magnification of the heater element in Example 1. Note that its large magnitude helps spread out the reflected energy.
5. 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?
6. 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?
7. 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.)
8. 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 $d_i = -d_o$, since this is a negative image distance (it is a virtual image). (a) What is the focal length of a flat mirror? (b) What is its power?
9. Show that for a flat mirror $h_i = h_o$, knowing that the image is a distance behind the mirror equal in magnitude to the distance of the object from the mirror.
10. Use the law of reflection to prove that the focal length of a mirror is half its radius of curvature.
$$f = \frac{R}{2}$$
That is, prove that . Note this is true for a spherical mirror only if its diameter is small compared with its radius of curvature.
11. Referring to the electric room heater considered in the first example in this section, 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 cm^2, and that half of the

radiated power is reflected and focused by the mirror.

12. Consider a 250-W heat lamp fixed to the ceiling in a bathroom. If the filament in one light burns out then the remaining three still work. Construct a problem in which you determine the resistance of each filament in order to obtain a certain intensity projected on the bathroom floor. The ceiling is 3.0 m high. The problem will need to involve concave mirrors behind the filaments. Your instructor may wish to guide you on the level of complexity to consider in the electrical components.

Glossary

converging mirror: a concave mirror in which light rays that strike it parallel to its axis converge at one or more points along the axis

diverging mirror: a convex mirror in which light rays that strike it parallel to its axis bend away (diverge) from its axis

law of reflection: angle of reflection equals the angle of incidence

Selected Solutions to Problems & Exercises

1. +0.667 m

3. (a) -1.5×10^{-2} m; (b) −66.7 D

5. +0.360 m (concave)

7. (a) +0.111; (b) −0.334 cm (behind “mirror”); (c) 0.752cm

9.

$$m = \frac{h_\text{i}}{h_\text{o}} = -\frac{d_\text{i}}{d_\text{o}} = -\frac{-d_\text{o}}{d_\text{o}} = \frac{d_\text{o}}{d_\text{o}} = 1 \Rightarrow h_\text{i} = h_\text{o}$$

11. 6.82 kW/m^2

10. Vision and Optical Instruments

Introduction to Vision and Optical Instruments

Lumen Learning

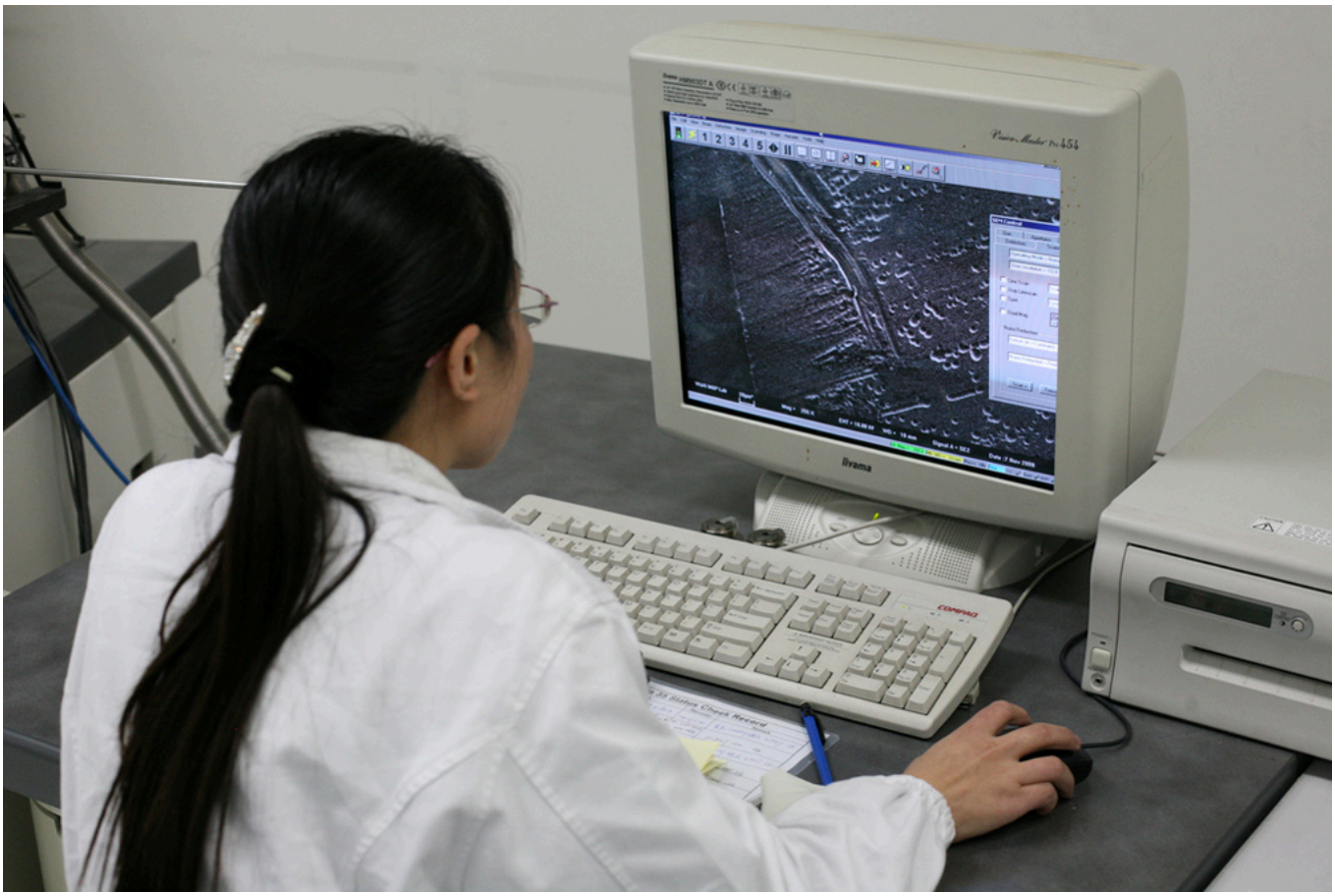


Figure 1. A scientist examines minute details on the surface of a disk drive at a magnification of 100,000 times. The image was produced using an electron microscope. (credit: Robert Scoble)

Explore how the image on the computer screen is formed. How is the image formation on the computer screen different from the image formation in your eye as you look down the microscope? How can videos of living cell processes be taken for viewing later on, and by many different people?

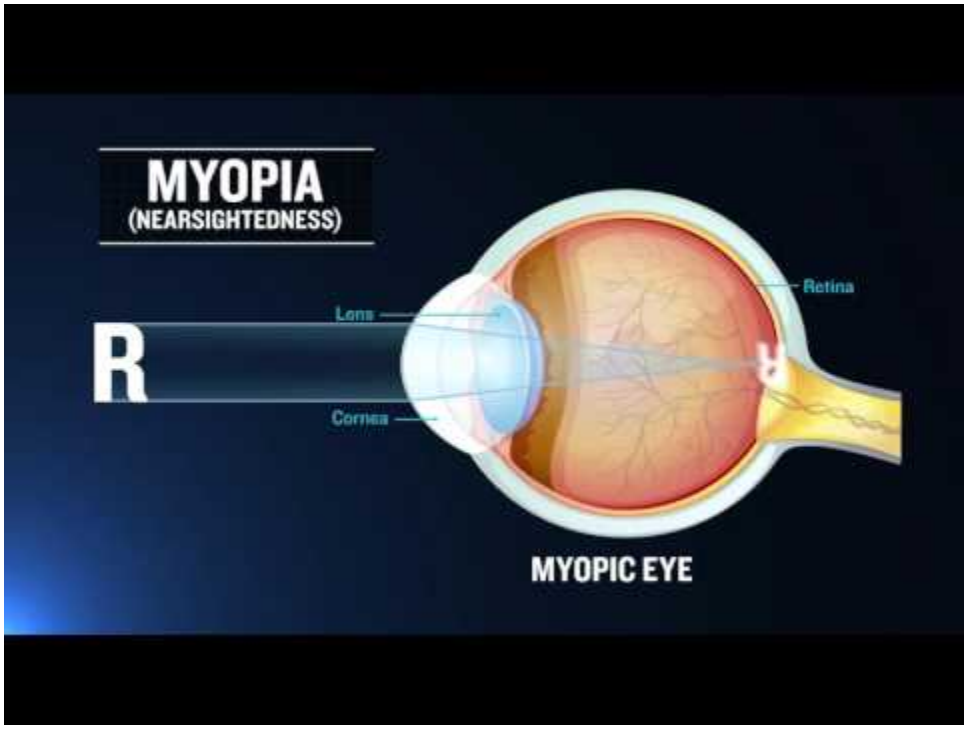
Seeing faces and objects we love and cherish is a delight—one's favorite teddy bear, a picture on the wall, or the sun rising over the mountains. Intricate images help us understand nature and are invaluable for developing techniques and technologies in order to improve the quality of life. The image of a red blood cell that almost fills the cross-sectional area of a tiny capillary makes us wonder how blood makes it through and not get stuck. We are able to see bacteria and viruses and understand their structure. It is the knowledge of physics that provides fundamental understanding and models required to develop new techniques and instruments. Therefore, physics is called an *enabling science*—a science that enables development and advancement in other areas. It is through optics and imaging that physics enables advancement in major areas of biosciences. This chapter illustrates the enabling nature of physics through an understanding of how a human eye is able to see and how we are able to use

optical instruments to see beyond what is possible with the naked eye. It is convenient to categorize these instruments on the basis of geometric optics and wave optics.

Video: Refraction

Lumen Learning

Watch the following Physics Concept Trailer to see how LASIK eye surgery changes the shape of the cornea so that light can refract correctly on the cornea.



A YouTube element has been excluded from this version of the text. You can view it online here: https://pressbooks.nscc.ca/heatlightsound/?p=207

Physics of the Eye

Lumen Learning

Learning Objectives

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

- Explain the image formation by the eye.
- Explain why peripheral images lack detail and color.
- Define refractive indices.
- Analyze the accommodation of the eye for distant and near vision.

The eye is perhaps the most interesting of all optical instruments. The eye is remarkable in how it forms images and in the richness of detail and color it can detect. However, our eyes commonly need some correction, to reach what is called “normal” vision, but should be called ideal rather than normal. Image formation by our eyes and common vision correction are easy to analyze with the optics discussed in Geometric Optics.

Figure 1 shows the basic anatomy of the eye. The cornea and lens form a system that, to a good approximation, acts as a single thin lens. For clear vision, a real image must be projected onto the light-sensitive retina, which lies at a fixed distance from the lens. The lens of the eye adjusts its power to produce an image on the retina for objects at different distances. The center of the image falls on the fovea, which has the greatest density of light receptors and the greatest acuity (sharpness) in the visual field. The variable opening (or pupil) of the eye along with chemical adaptation allows the eye to detect light intensities from the lowest observable to 10^{10} times greater (without damage). This is an incredible range of detection. Our eyes perform a vast number of functions, such as sense direction, movement, sophisticated colors, and distance. Processing of visual nerve impulses begins with interconnections in the retina and continues in the brain. The optic nerve conveys signals received by the eye to the brain.

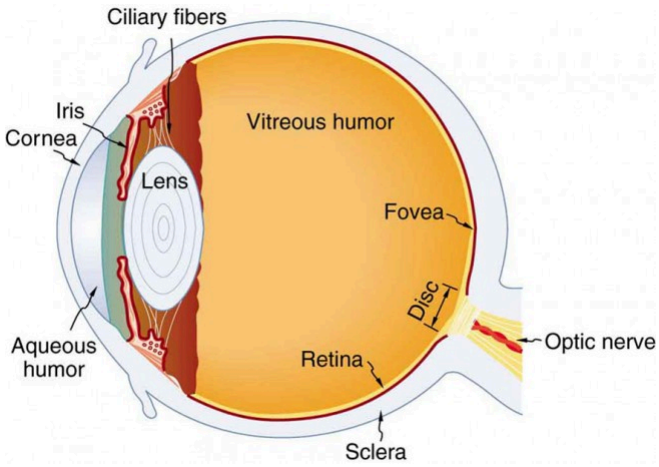


Figure 1. The cornea and lens of an eye act together to form a real image on the light-sensing retina, which has its densest concentration of receptors in the fovea and a blind spot over the optic nerve. The power of the lens of an eye is adjustable to provide an image on the retina for varying object distances. Layers of tissues with varying indices of refraction in the lens are shown here. However, they have been omitted from other pictures for clarity.

Refractive indices are crucial to image formation using lenses. Table 1 shows refractive indices relevant to the eye. The biggest change in the refractive index, and bending of rays, occurs at the cornea rather than the lens. The ray diagram in Figure 2 shows image formation by the cornea and lens of the eye. The rays bend according to the refractive indices provided in Table 1. The cornea provides about two-thirds of the power of the eye, owing to the fact that speed of light changes considerably while traveling from air into cornea. The lens provides the remaining power needed to produce an image on the retina. The cornea and lens can be treated as a single thin lens, even though the light rays pass through several layers of material (such as cornea, aqueous humor, several layers in the lens, and vitreous humor), changing direction at each interface. The image formed is much like the one produced by a single convex lens. This is a case 1 image. Images formed in the eye are inverted but the brain inverts them once more to make them seem upright.

Table 1. Refractive Indices Relevant to the Eye

Material	Index of Refraction
Water	1.33
Air	1.0
Cornea	1.38
Aqueous humor	1.34
Lens	1.41 average (varies throughout the lens, greatest in center)
Vitreous humor	1.34

As noted, the image must fall precisely on the retina to produce clear vision—that is, the image distance d_i must equal the lens-to-retina distance. Because the lens-to-retina distance does not change, the image distance d_i must be the same for objects at all distances. The eye manages this by varying the power (and focal length) of the lens to accommodate for objects at various distances. The process of adjusting the eye's focal length is called *accommodation*. A person with normal (ideal) vision can see objects clearly at distances ranging from 25 cm to essentially infinity. However, although the near point (the shortest distance at which a sharp focus can be obtained) increases with age (becoming meters for some older people), we will consider it to be 25 cm in our treatment here.

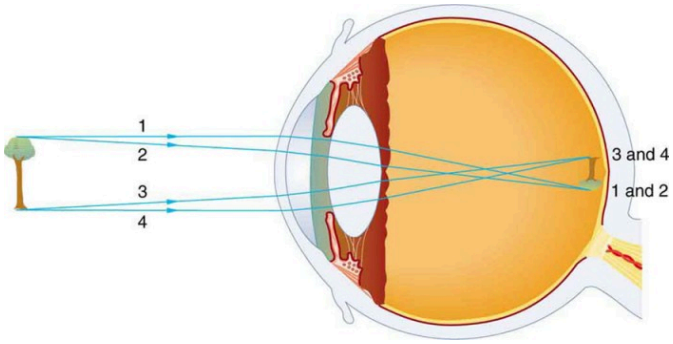


Figure 2. An image is formed on the retina with light rays converging most at the cornea and upon entering and exiting the lens. Rays from the top and bottom of the object are traced and produce an inverted real image on the retina. The distance to the object is drawn smaller than scale.

Figure 3 shows the accommodation of the eye for distant and near vision. Since light rays from a nearby object can diverge and still enter the eye, the lens must be more converging (more powerful) for close vision than for distant vision. To be more converging, the lens is made thicker by the action of the ciliary muscle surrounding it. The eye is most relaxed when viewing distant objects, one reason that

microscopes and telescopes are designed to produce distant images. Vision of very distant objects is called *totally relaxed*, while close vision is termed *accommodated*, with the closest vision being *fully accommodated*.

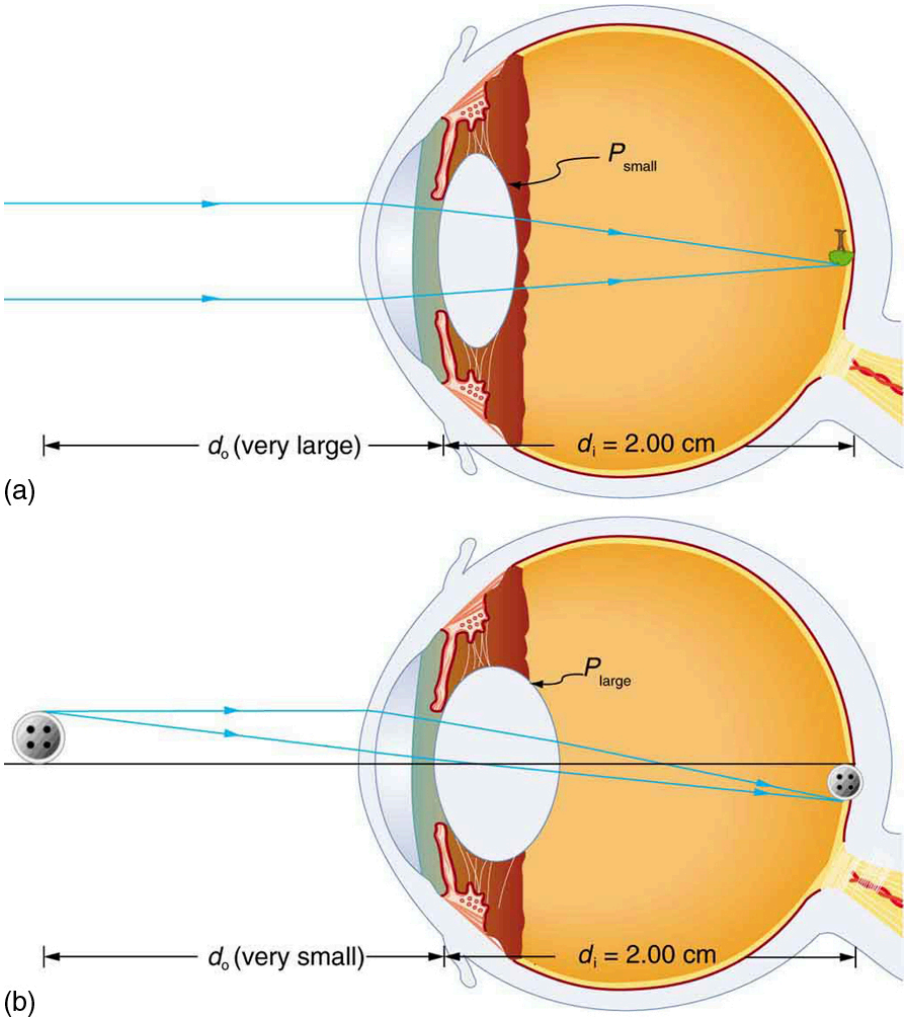


Figure 3. Relaxed and accommodated vision for distant and close objects. (a) Light rays from the same point on a distant object must be nearly parallel while entering the eye and more easily converge to produce an image on the retina. (b) Light rays from a nearby object can diverge more and still enter the eye. A more powerful lens is needed to converge them on the retina than if they were parallel.

We will use the thin lens equations to examine image formation by the eye quantitatively. First, note the power of a lens is given as

$$p = \frac{1}{f}$$

, so we rewrite the thin lens equations as

$$P = \frac{1}{d_o} + \frac{1}{d_i}$$

$\frac{h_\text{i}}{h_\text{o}} = -\frac{d_\text{i}}{d_\text{o}} = m$
and .

We understand that d_i must equal the lens-to-retina distance to obtain clear vision, and that normal vision is possible for objects at distances d_o = 25 cm to infinity.

Take-Home Experiment: The Pupil

Look at the central transparent area of someone's eye, the pupil, in normal room light. Estimate the diameter of the pupil. Now turn off the lights and darken the room. After a few minutes turn on the lights and promptly estimate the diameter of the pupil. What happens to the pupil as the eye adjusts to the room light? Explain your observations.

The eye can detect an impressive amount of detail, considering how small the image is on the retina. To get some idea of how small the image can be, consider the following example.

Example 1. Size of Image on Retina

What is the size of the image on the retina of a 1.20×10^{-2} cm diameter human hair, held at arm's length (60.0 cm) away? Take the lens-to-retina distance to be 2.00 cm.

Strategy

We want to find the height of the image h_i, given the height of the object is $h_\text{o} = 1.20 \times 10^{-2}$ cm. We also know that the object is 60.0 cm away, so that d_o=60.0 cm. For clear vision, the image distance must equal the

$$\frac{h_\text{i}}{h_\text{o}} = -\frac{d_\text{i}}{d_\text{o}} = m$$

lens-to-retina distance, and so d_i = 2.00 cm . The equation can be used to find h_i with the known information.

Solution

The only unknown variable in the equation

$$\frac{h_\text{i}}{h_\text{o}} = -\frac{d_\text{i}}{d_\text{o}} = m$$

is h_i:

$$\frac{h_\text{i}}{h_\text{o}} = -\frac{d_\text{i}}{d_\text{o}}$$

Rearranging to isolate h_i yields

$$h_\text{i} = -h_\text{o} \cdot \frac{d_\text{i}}{d_\text{o}}$$

.

Substituting the known values gives

$$\begin{aligned} h_{\text{i}} &= -\left(1.20 \times 10^{-2}\ \text{cm}\right) \frac{2.00\ \text{cm}}{60.0\ \text{cm}} \\ &= -4.00 \times 10^{-4}\ \text{cm} \end{aligned}$$

Discussion

This truly small image is not the smallest discernible—that is, the limit to visual acuity is even smaller than this. Limitations on visual acuity have to do with the wave properties of light and will be discussed in the next chapter. Some limitation is also due to the inherent anatomy of the eye and processing that occurs in our brain.

Example 2. Power Range of the Eye

Calculate the power of the eye when viewing objects at the greatest and smallest distances possible with normal vision, assuming a lens-to-retina distance of 2.00 cm (a typical value).

Strategy

For clear vision, the image must be on the retina, and so d_{i} = 2.00 cm here. For distant vision, $d_{\text{o}} \approx \infty$, and for close vision, d_{o} = 25.0 cm, as discussed earlier. The equation

$$P = \frac{1}{d_{\text{o}}} + \frac{1}{d_{\text{i}}}$$

as written just above, can be used directly to solve for P in both cases, since we know d_{i} and d_{o}. Power has units of diopters, where

$$1\ \text{D} = \frac{1}{\text{m}}$$

, and so we should express all distances in meters.

Solution

For distant vision,

$$P = \frac{1}{d_{\text{o}}} + \frac{1}{d_{\text{i}}} = \frac{1}{\infty} + \frac{1}{0.0200\ \text{m}}$$

Since

$$\frac{1}{\infty} = 0$$

, this gives

$$P = 0 + \frac{50.0}{\text{m}} = 50.0\ \text{D}$$

(distant vision).

Now, for close vision,

$$\begin{aligned} P &= \frac{1}{d_{\text{o}}} + \frac{1}{d_{\text{i}}} = \frac{1}{0.250\ \text{m}} + \frac{1}{0.0200\ \text{m}} \\ &= \frac{4.00}{\text{m}} + \frac{50.0}{\text{m}} = 4.00\ \text{D} + 50.0\ \text{D} \\ &= 54.0\ \text{D}\ \text{(close vision)} \end{aligned}$$

Discussion

For an eye with this typical 2.00 cm lens-to-retina distance, the power of the eye ranges from 50.0 D (for distant totally relaxed vision) to 54.0 D (for close fully accommodated vision), which is an 8% increase. This increase in power for close vision is consistent with the preceding discussion and the ray tracing in Figure 3. An 8% ability to accommodate is considered normal but is typical for people who are about 40 years old. Younger people have greater accommodation ability, whereas older people gradually lose the ability to accommodate. When an optometrist identifies accommodation as a problem in elder people, it is most likely due to stiffening of the lens. The lens of the eye changes with age in ways that tend to preserve the ability to see distant objects clearly but do not allow the eye to accommodate for close vision, a condition called *presbyopia* (literally, elder eye). To correct this vision defect, we place a converging, positive power lens in front of the eye, such as found in reading glasses. Commonly available reading glasses are rated by their power in diopters, typically ranging from 1.0 to 3.5 D.

Section Summary

- Image formation by the eye is adequately described by the thin lens equations: $P = \dfrac{1}{d_{\text{o}}} + \dfrac{1}{d_{\text{i}}}$ and $\dfrac{h_{\text{i}}}{h_{\text{o}}} = -\dfrac{d_{\text{i}}}{d_{\text{o}}} = m$.
- The eye produces a real image on the retina by adjusting its focal length and power in a process called accommodation.
- For close vision, the eye is fully accommodated and has its greatest power, whereas for distant vision, it is totally relaxed and has its smallest power.
- The loss of the ability to accommodate with age is called presbyopia, which is corrected by the use of a converging lens to add power for close vision.

Conceptual Questions

1. If the lens of a person's eye is removed because of cataracts (as has been done since ancient times), why would you expect a spectacle lens of about 16 D to be prescribed?
2. A cataract is cloudiness in the lens of the eye. Is light dispersed or diffused by it?
3. When laser light is shone into a relaxed normal-vision eye to repair a tear by spot-welding the retina to the back of the eye, the rays entering the eye must be parallel. Why?
4. How does the power of a dry contact lens compare with its power when resting on the tear layer of the eye? Explain.
5. Why is your vision so blurry when you open your eyes while swimming under water? How does a face mask enable clear vision?

Problems & Exercises

Unless otherwise stated, the lens-to-retina distance is 2.00 cm.

1. What is the power of the eye when viewing an object 50.0 cm away?
2. Calculate the power of the eye when viewing an object 3.00 m away.
3. (a) The print in many books averages 3.50 mm in height. How high is the image of the print on the retina when the book is held 30.0 cm from the eye? (b) Compare the size of the print to the sizes of rods and cones in the fovea and discuss the possible details observable in the letters. (The eye-brain system can perform better because of interconnections and higher order image processing.)
4. Suppose a certain person's visual acuity is such that he can see objects clearly that form an image 4.00 μm high on his retina. What is the maximum distance at which he can read the 75.0 cm high letters on the side of an airplane?
5. People who do very detailed work close up, such as jewellers, often can see objects clearly at much closer distance than the normal 25 cm. (a) What is the power of the eyes of a woman who can see an object clearly at a distance of only 8.00 cm? (b) What is the size of an image of a 1.00 mm object, such as lettering inside a ring, held at this distance? (c) What would the size of the image be if the object were held at the normal 25.0 cm distance?

Glossary

accommodation: the ability of the eye to adjust its focal length is known as accommodation

presbyopia: a condition in which the lens of the eye becomes progressively unable to focus on objects close to the viewer

Selected Solutions to Problems & Exercises

1. 52.0 D

3. (a) −0.233 mm; (b) The size of the rods and the cones is smaller than the image height, so we can distinguish letters on a page.

5. (a) +62.5 D; (b) –0.250 mm; (c) –0.0800 mm

Telescopes

Lumen Learning

Learning Objectives

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

- Outline the invention of a telescope.
- Describe the working of a telescope.

Telescopes are meant for viewing distant objects, producing an image that is larger than the image that can be seen with the unaided eye. Telescopes gather far more light than the eye, allowing dim objects to be observed with greater magnification and better resolution. Although Galileo is often credited with inventing the telescope, he actually did not. What he did was more important. He constructed several early telescopes, was the first to study the heavens with them, and made monumental discoveries using them. Among these are 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 vast numbers of individual stars.

Figure 1a shows a telescope made of two lenses, the convex objective and the concave eyepiece, the same construction used by Galileo. Such an arrangement produces an upright image and is used in spyglasses and opera glasses.

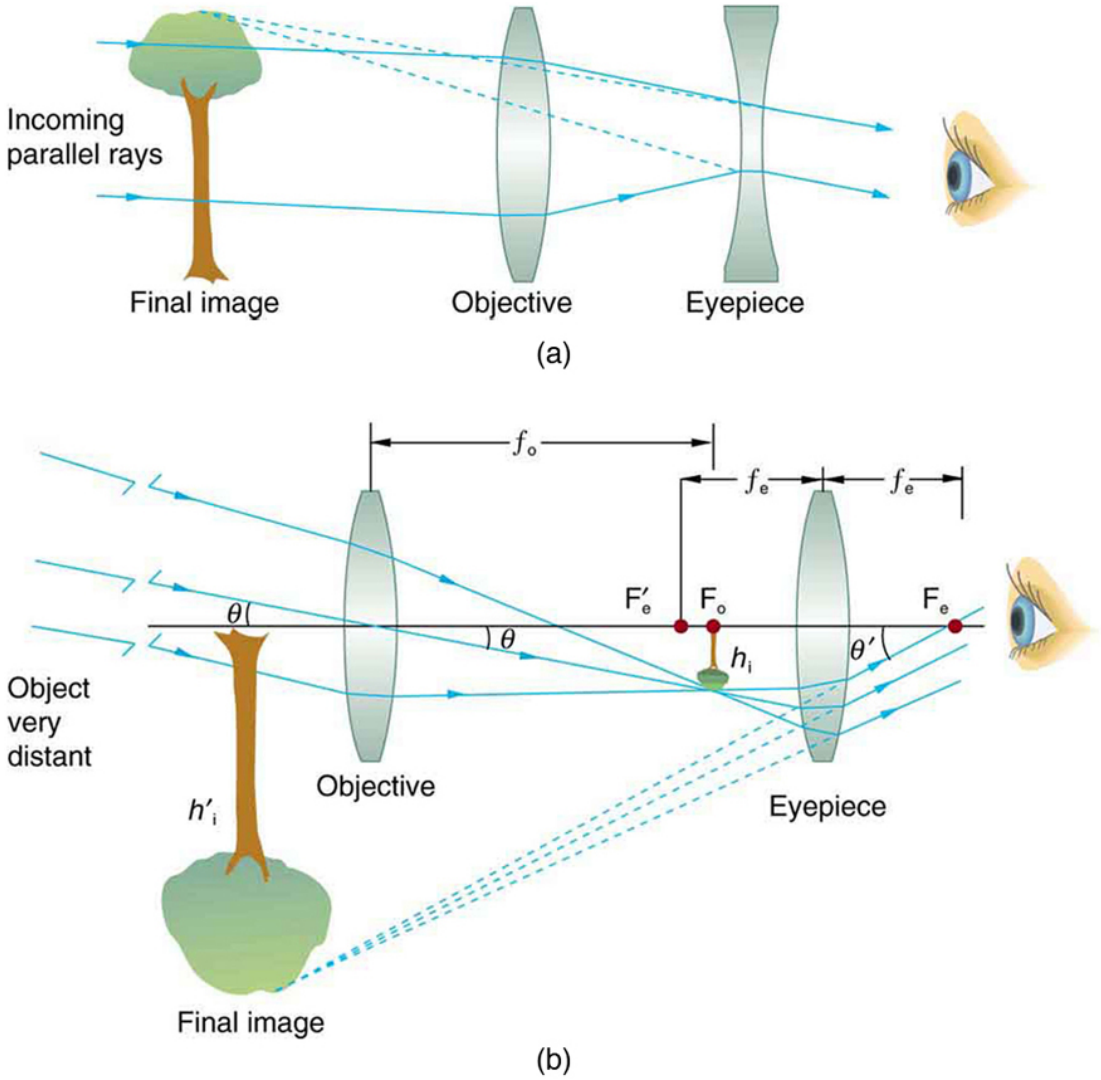


Figure 1. (a) Galileo made telescopes with a convex objective and a concave eyepiece. These produce an upright image and are used in spyglasses. (b) Most simple telescopes have two convex lenses. The objective forms a case 1 image that is the object for the eyepiece. The eyepiece forms a case 2 final image that is magnified.

The most common two-lens telescope, like the simple microscope, uses two convex lenses and is shown in Figure 1b. The object is so far away from the telescope that it is essentially at infinity compared with the focal lengths of the lenses ($d_o \approx \infty$). The first image is thus produced at $d_i = f_o$, as shown in the figure. To prove this, note that

$$\frac{1}{d_i} = \frac{1}{f_o} - \frac{1}{d_o} = \frac{1}{f_o} - \frac{1}{\infty}$$

Because

$$\tfrac{1}{\infty} = 0$$

, this simplifies to

$$\frac{1}{d_\mathrm{i}} = \frac{1}{f_\mathrm{o}}$$

, which implies that $d_\mathrm{i} = f_\mathrm{o}$, as claimed. It is true that for any distant object and any lens or mirror, the image is at the focal length.

The first image formed by a telescope objective as seen in Figure 1b will not be large compared with what you might see by looking at the object directly. For example, the spot formed by sunlight focused on a piece of paper by a magnifying glass is the image of the Sun, and it is small. The telescope eyepiece (like the microscope eyepiece) magnifies this first image. The distance between the eyepiece and the objective lens is made slightly less than the sum of their focal lengths so that the first image is closer to the eyepiece than its focal length. That is, d_o' is less than f_e, and so the eyepiece forms a case 2 image that is large and to the left for easy viewing. If the angle subtended by an object as viewed by the unaided eye is θ, and the angle subtended by the telescope image is θ', then the *angular magnification M* is defined to be their ratio. That is,

$$M = \frac{\theta'}{\theta}$$

. It can be shown that the angular magnification of a telescope is related to the focal lengths of the objective and eyepiece; and is given by

$$M = \frac{\theta'}{\theta} = -\frac{f_\mathrm{o}}{f_\mathrm{e}}$$

The minus sign indicates the image is inverted. To obtain the greatest angular magnification, it is best to have a long focal length objective and a short focal length eyepiece. 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.

The image in most telescopes is inverted, which is unimportant for observing the stars but 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 Figure 1a 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 2.

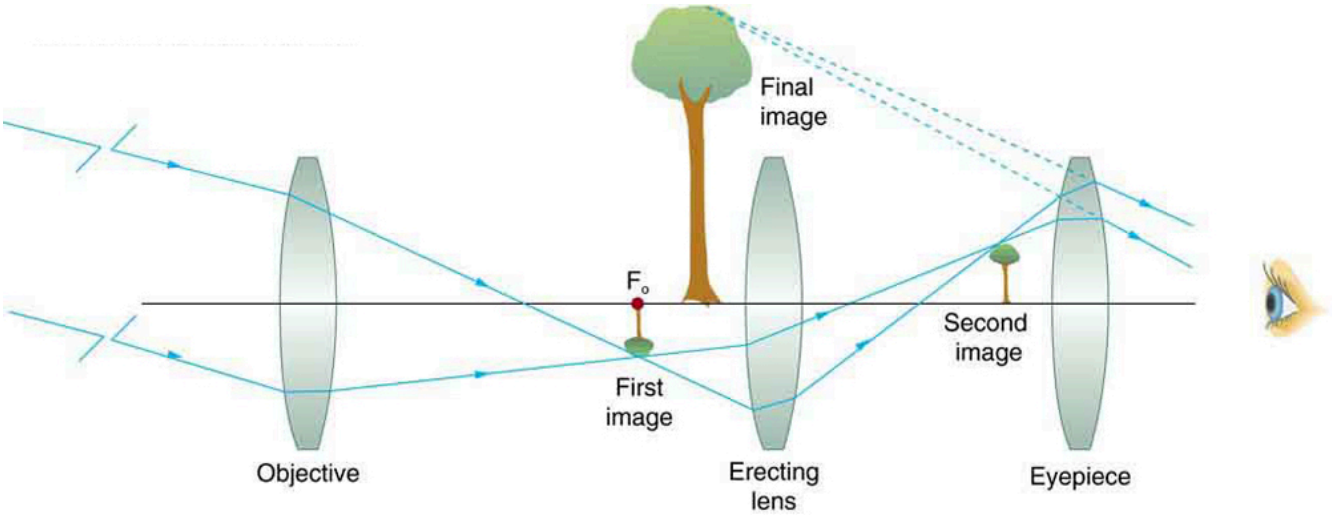


Figure 2. This arrangement of three lenses in a telescope produces an upright final image. The first two lenses are far enough apart that the second lens inverts the image of the first one more time. The third lens acts as a magnifier and keeps the image upright and in a location that is easy to view.

A telescope can also be made with a concave mirror as its first element or objective, since a concave mirror acts like a convex lens as seen in Figure 3. Flat mirrors are often employed in optical instruments to make them more compact or to send light to cameras and other sensing devices. There are many advantages to using mirrors rather than lenses for telescope objectives. Mirrors can be constructed much larger than lenses and can, thus, gather large amounts of light, as needed to view distant galaxies, for example. Large and relatively flat mirrors have very long focal lengths, so that great angular magnification is possible.

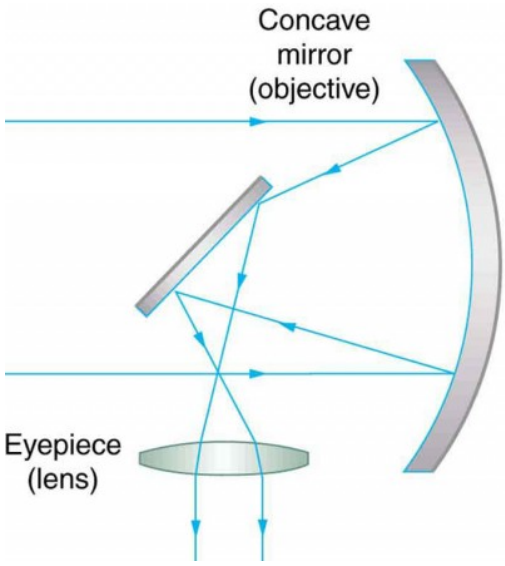


Figure 3. A two-element telescope composed of a mirror as the objective and a lens for the eyepiece is shown. This telescope forms an image in the same manner as the two-convex-lens telescope already discussed, but it does not suffer from chromatic aberrations. Such telescopes can gather more light, since larger mirrors than lenses can be constructed.

Telescopes, like microscopes, can utilize a range of frequencies from the electromagnetic spectrum. Figure 4a shows the Australia Telescope Compact Array, which uses six 22-m antennas for mapping the southern skies using radio waves. Figure 4b shows the focusing of x rays on the Chandra X-ray Observatory—a satellite orbiting earth since 1999 and looking at high temperature events as exploding stars, quasars, and black holes. X rays, with much more energy and shorter wavelengths than RF and light, are mainly absorbed and not reflected when incident perpendicular to the medium. But they can be reflected when incident at small glancing angles, much like a rock will skip on a lake if thrown at a small angle. The mirrors for the Chandra consist of a long barrelled pathway and 4 pairs of mirrors to focus the rays at a point 10 meters away from the entrance. The mirrors are extremely smooth and consist of a glass ceramic base with a thin coating of metal (iridium). Four pairs of precision manufactured mirrors are exquisitely shaped and aligned so that x rays ricochet off the mirrors like bullets off a wall, focusing on a spot.

A current exciting development is a collaborative effort involving 17 countries to construct a Square Kilometre Array (SKA) of telescopes capable of covering from 80 MHz to 2 GHz. The initial stage of the project is the construction of the Australian Square Kilometre Array Pathfinder in Western Australia (see Figure 5). The project will use cutting-edge technologies such as *adaptive optics* in which the lens or mirror is constructed from lots of carefully aligned tiny lenses and mirrors that can be manipulated using computers. A range of rapidly changing distortions can be minimized by deforming or tilting the tiny lenses and mirrors. The use of adaptive optics in vision correction is a current area of research.

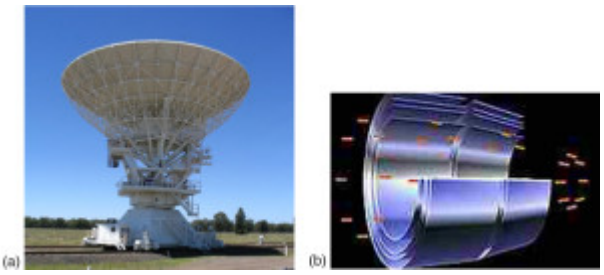


Figure 4. (a) The Australia Telescope Compact Array at Narrabri (500 km NW of Sydney). (credit: Ian Bailey) (b) The focusing of x rays on the Chandra Observatory, a satellite orbiting earth. X rays ricochet off 4 pairs of mirrors forming a barrelled pathway leading to the focus point. (credit: NASA)

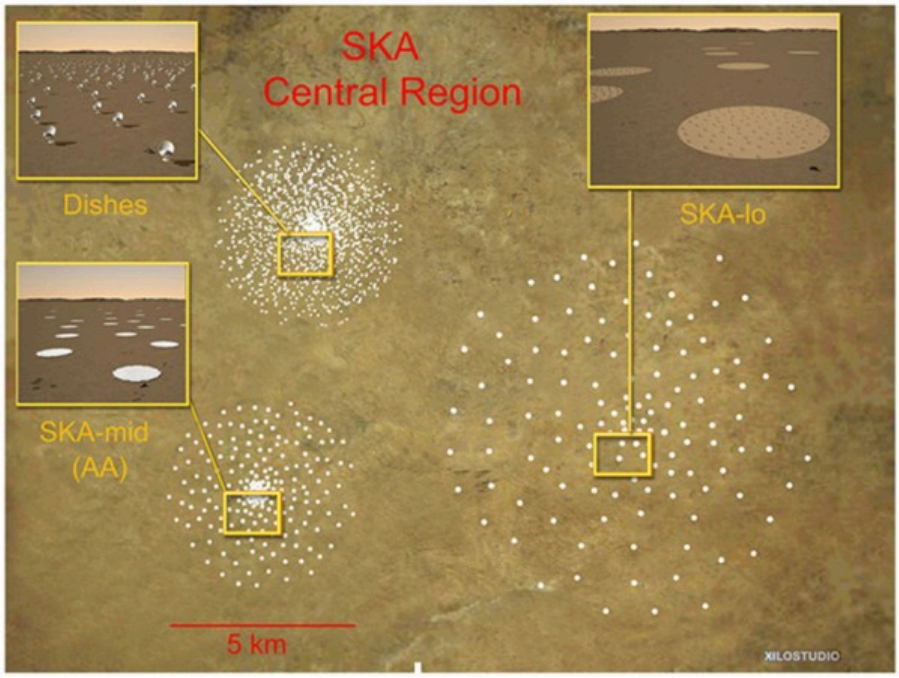


Figure 5. An artist's impression of the Australian Square Kilometre Array Pathfinder in Western Australia is displayed. (credit: SPDO, XILOSTUDIOS)

Section Summary

- Simple telescopes can be made with two lenses. They are used for viewing objects at large distances and utilize the entire range of the electromagnetic spectrum.
- The angular magnification M for a telescope is given by $$M = \frac{\theta'}{\theta} = -\frac{f_{\mathrm{o}}}{f_{\mathrm{e}}}$$, where θ is the angle subtended by an object viewed by the unaided eye, θ' is the angle subtended by a magnified image, and f_{o} and f_{e} are the focal lengths of the objective and the eyepiece.

Conceptual Questions

1. 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 & Exercises

Unless otherwise stated, the lens-to-retina distance is 2.00 cm.

1. What is the angular magnification of a telescope that has a 100 cm focal length objective and a 2.50 cm focal length eyepiece?
2. 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.
3. 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?
4. 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?
5. A 7.5× 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?
6. **Construct Your Own Problem.** Consider a telescope of the type used by Galileo, having a convex objective and a concave eyepiece as illustrated in Figure 1a. 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.

Glossary

adaptive optics: optical technology in which computers adjust the lenses and mirrors in a device to correct for image distortions

angular magnification: a ratio related to the focal lengths of the objective and eyepiece and given as $M = -\frac{f_o}{f_e}$

Selected Solutions to Problems & Exercises

1. −40.0
3. −1.67
5. +10.0 cm

Vision Correction

Lumen Learning

Learning Objectives

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

- Identify and discuss common vision defects.
- Explain nearsightedness and farsightedness corrections.
- Explain laser vision correction.

The need for some type of vision correction is very common. Common vision defects are easy to understand, and some are simple to correct. Figure 1 illustrates two common vision defects. *Nearsightedness*, or *myopia*, is the inability to see distant objects clearly while close objects are clear. The eye overconverges the nearly parallel rays from a distant object, and the rays cross in front of the retina. More divergent rays from a close object are converged on the retina for a clear image. The distance to the farthest object that can be seen clearly is called the *far point* of the eye (normally infinity). *Farsightedness*, or *hyperopia*, is the inability to see close objects clearly while distant objects may be clear. A farsighted eye does not converge sufficient rays from a close object to make the rays meet on the retina. Less diverging rays from a distant object can be converged for a clear image. The distance to the closest object that can be seen clearly is called the *near point* of the eye (normally 25 cm).

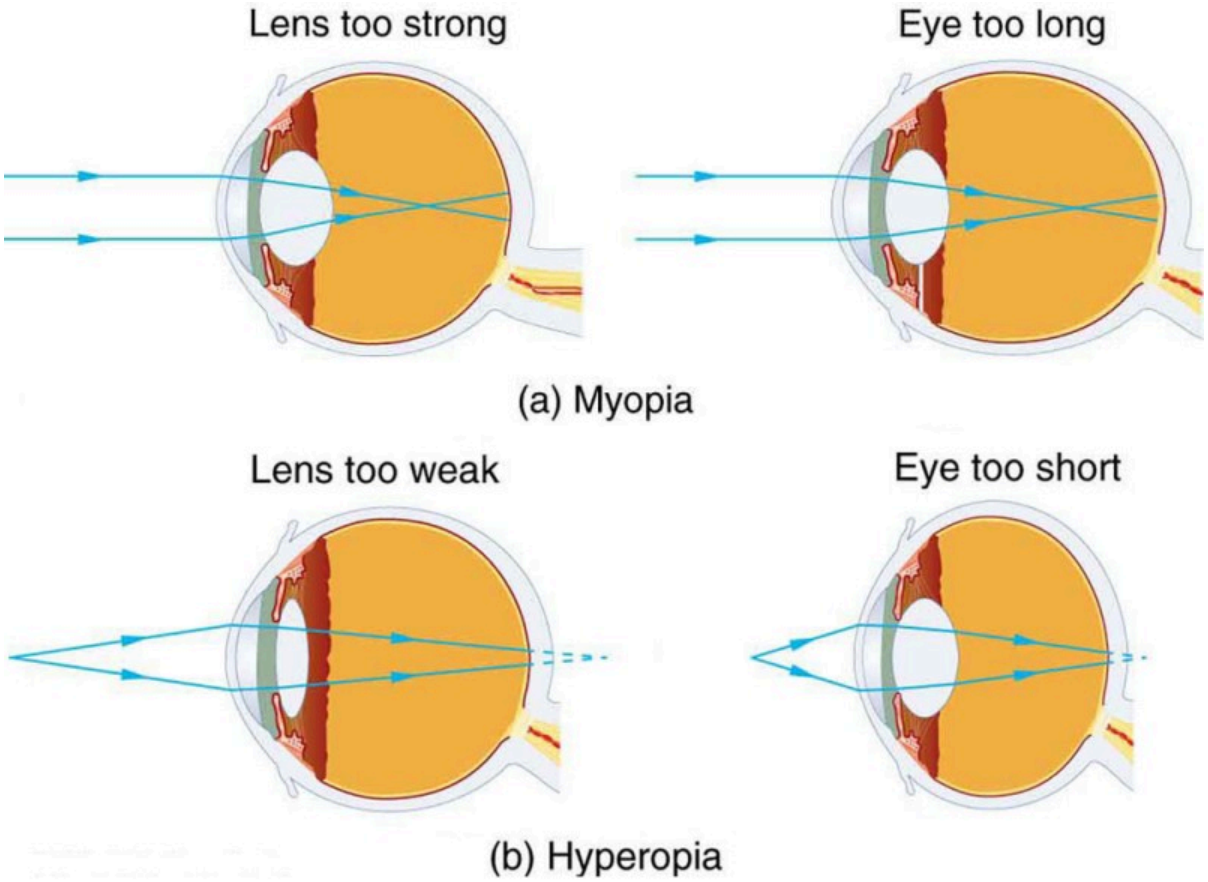


Figure 1. (a) The nearsighted (myopic) eye converges rays from a distant object in front of the retina; thus, they are diverging when they strike the retina, producing a blurry image. This can be caused by the lens of the eye being too powerful or the length of the eye being too great. (b) The farsighted (hyperopic) eye is unable to converge the rays from a close object by the time they strike the retina, producing blurry close vision. This can be caused by insufficient power in the lens or by the eye being too short.

Since the nearsighted eye over converges light rays, the correction for nearsightedness is to place a diverging spectacle lens in front of the eye. This reduces the power of an eye that is too powerful. Another way of thinking about this is that a diverging spectacle lens produces a case 3 image, which is closer to the eye than the object (see Figure 2). To determine the spectacle power needed for correction, you must know the person's far point—that is, you must know the greatest distance at which the person can see clearly. Then the image produced by a spectacle lens must be at this distance or closer for the nearsighted person to be able to see it clearly. It is worth noting that wearing glasses does not change the eye in any way. The eyeglass lens is simply used to create an image of the object at a distance where the nearsighted person can see it clearly. Whereas someone not wearing glasses can see clearly *objects* that fall between their near point and their far point, someone wearing glasses can see *images* that fall between their near point and their far point.

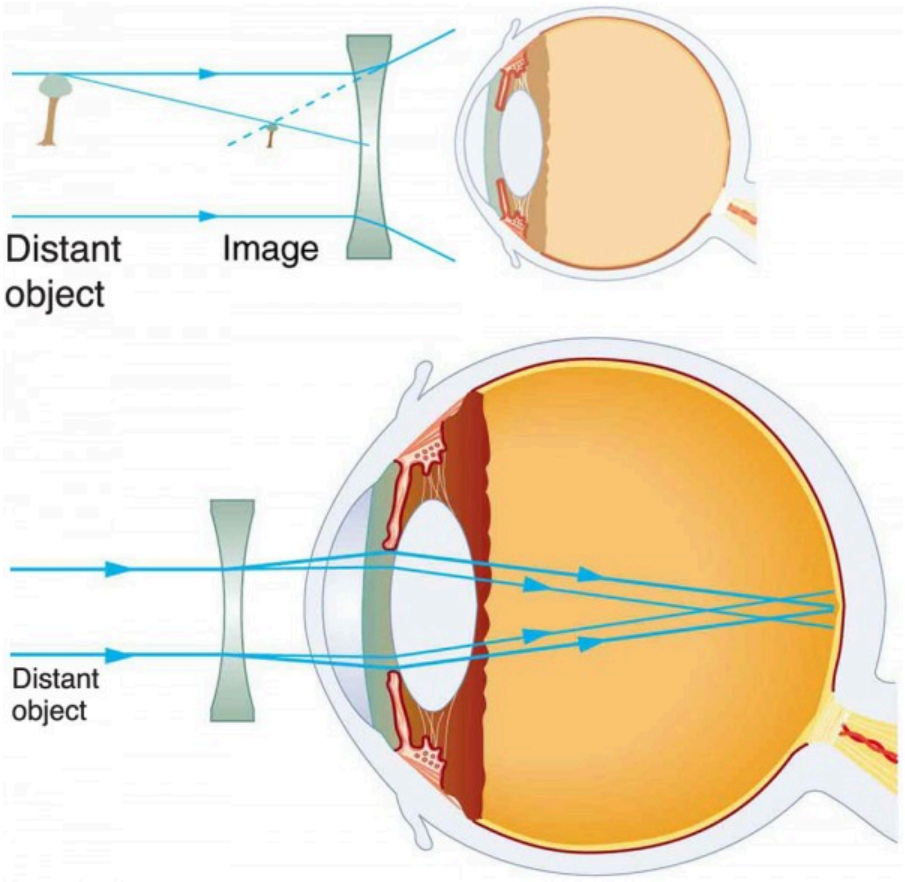


Figure 2. Correction of nearsightedness requires a diverging lens that compensates for the overconvergence by the eye. The diverging lens produces an image closer to the eye than the object, so that the nearsighted person can see it clearly.

Example 1. Correcting Nearsightedness

What power of spectacle lens is needed to correct the vision of a nearsighted person whose far point is 30.0 cm? Assume the spectacle (corrective) lens is held 1.50 cm away from the eye by eyeglass frames.

Strategy

You want this nearsighted person to be able to see very distant objects clearly. That means the spectacle lens must produce an image 30.0 cm from the eye for an object very far away. An image 30.0 cm from the eye will be 28.5 cm to the left of the spectacle lens (see Figure 2). Therefore, we must get $d_i = -28.5$ cm when $d_o \approx \infty$. The image distance is negative, because it is on the same side of the spectacle as the object.

Solution

Since d_i and d_o are known, the power of the spectacle lens can be found using

$$P = \frac{1}{d_o} + \frac{1}{d_i}$$

as written earlier:

$$P = \frac{1}{d_{\text{o}}} + \frac{1}{d_{\text{i}}} = \frac{1}{\infty} + \frac{1}{-0.285\ \text{m}}$$

Since

$$\frac{1}{\infty} = 0$$

, we obtain:

$$P = 0 - \frac{3.51}{\text{m}} = -3.51\ \text{D}$$

.

Discussion

The negative power indicates a diverging (or concave) lens, as expected. The spectacle produces a case 3 image closer to the eye, where the person can see it. If you examine eyeglasses for nearsighted people, you will find the lenses are thinnest in the center. Additionally, if you examine a prescription for eyeglasses for nearsighted people, you will find that the prescribed power is negative and given in units of diopters.

Since the farsighted eye under converges light rays, the correction for farsightedness is to place a converging spectacle lens in front of the eye. This increases the power of an eye that is too weak. Another way of thinking about this is that a converging spectacle lens produces a case 2 image, which is farther from the eye than the object (see Figure 3). To determine the spectacle power needed for correction, you must know the person's near point—that is, you must know the smallest distance at which the person can see clearly. Then the image produced by a spectacle lens must be at this distance or farther for the farsighted person to be able to see it clearly.

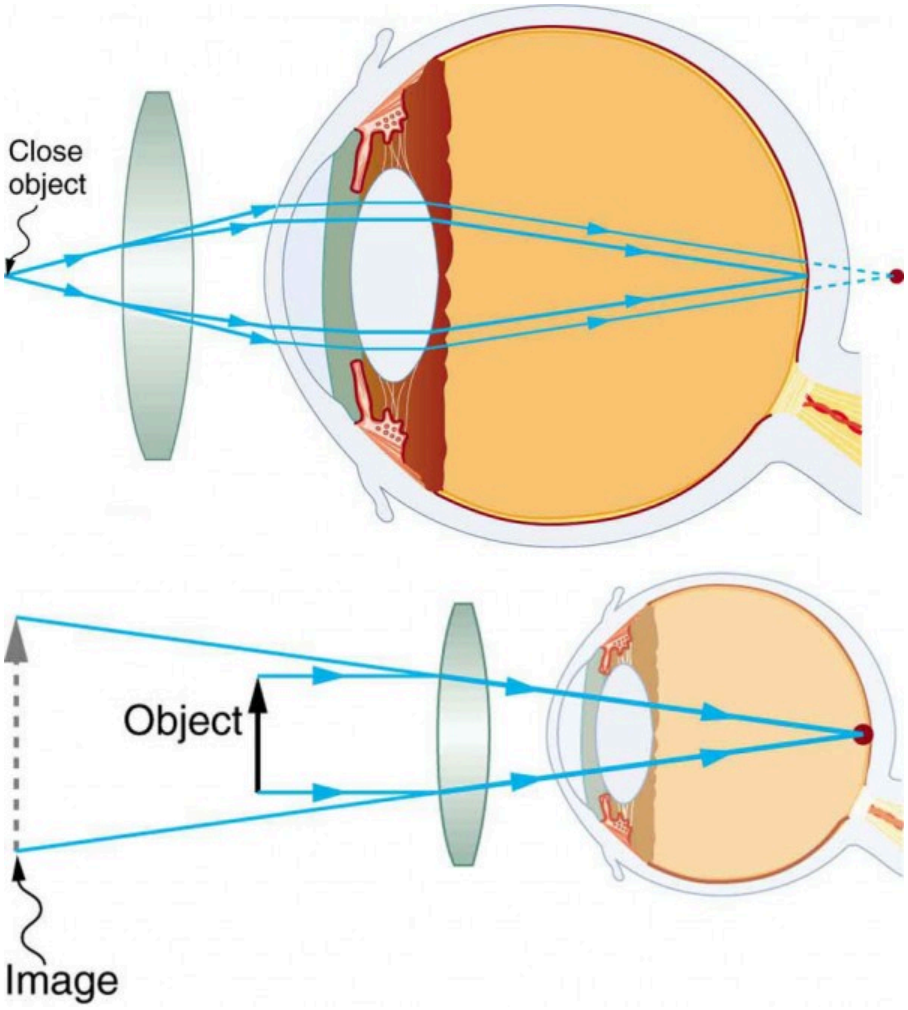


Figure 3. Correction of farsightedness uses a converging lens that compensates for the under convergence by the eye. The converging lens produces an image farther from the eye than the object, so that the farsighted person can see it clearly.

Example 2. Correcting Farsightedness

What power of spectacle lens is needed to allow a farsighted person, whose near point is 1.00 m, to see an object clearly that is 25.0 cm away? Assume the spectacle (corrective) lens is held 1.50 cm away from the eye by eyeglass frames.

Strategy

When an object is held 25.0 cm from the person's eyes, the spectacle lens must produce an image 1.00 m away (the near point). An image 1.00 m from the eye will be 98.5 cm to the left of the spectacle lens because the spectacle lens is 1.50 cm from the eye (see Figure 3). Therefore, d_i=−98.5 cm. The image distance is negative, because it is on the same side of the spectacle as the object. The object is 23.5 cm to the left of the spectacle, so that d_o=23.5 cm.

Solution

Since d_i and d_o are known, the power of the spectacle lens can be found using

$$P = \frac{1}{d_\text{o}} + \frac{1}{d_\text{i}}$$

:

$$\begin{aligned} P &= \frac{1}{d_\text{o}} + \frac{1}{d_\text{i}} = \frac{1}{0.235\ \text{m}} + \frac{1}{-0.985\ \text{m}} \\ &= 4.26\ \text{D} - 1.02\ \text{D} = 3.24\ \text{D} \end{aligned}$$

Discussion

The positive power indicates a converging (convex) lens, as expected. The convex spectacle produces a case 2 image farther from the eye, where the person can see it. If you examine eyeglasses of farsighted people, you will find the lenses to be thickest in the center. In addition, a prescription of eyeglasses for farsighted people has a prescribed power that is positive.

Another common vision defect is *astigmatism*, an unevenness or asymmetry in the focus of the eye. For example, rays passing through a vertical region of the eye may focus closer than rays passing through a horizontal region, resulting in the image appearing elongated. This is mostly due to irregularities in the shape of the cornea but can also be due to lens irregularities or unevenness in the retina. Because of these irregularities, different parts of the lens system produce images at different locations. The eye-brain system can compensate for some of these irregularities, but they generally manifest themselves as less distinct vision or sharper images along certain axes. Figure 4 shows a chart used to detect astigmatism. Astigmatism can be at least partially corrected with a spectacle having the opposite irregularity of the eye. If an eyeglass prescription has a cylindrical correction, it is there to correct astigmatism. The normal corrections for short- or farsightedness are spherical corrections, uniform along all axes.

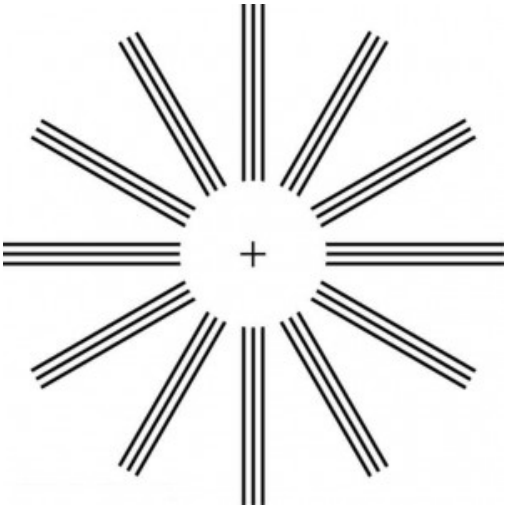


Figure 4. This chart can detect astigmatism, unevenness in the focus of the eye. Check each of your eyes separately by looking at the center cross (without spectacles if you wear them). If lines along some axes appear darker or clearer than others, you have an astigmatism.

Contact lenses have advantages over glasses beyond their cosmetic aspects. One problem with glasses is that as the eye moves, it is not at a fixed distance from the spectacle lens. Contacts rest on and move with the eye, eliminating this problem. Because contacts cover a significant portion of the cornea, they provide superior peripheral vision compared with eyeglasses. Contacts also correct some corneal astigmatism caused by surface irregularities. The tear layer between the smooth contact and the cornea fills in the irregularities. Since the index of refraction of the tear layer and the cornea are very similar, you now have a regular optical surface in place of an irregular one. If the curvature of a contact lens is not the same as the cornea (as may be necessary with some individuals to obtain a comfortable fit), the tear layer

between the contact and cornea acts as a lens. If the tear layer is thinner in the center than at the edges, it has a negative power, for example. Skilled optometrists will adjust the power of the contact to compensate.

Laser vision correction has progressed rapidly in the last few years. It is the latest and by far the most successful in a series of procedures that correct vision by reshaping the cornea. As noted at the beginning of this section, the cornea accounts for about two-thirds of the power of the eye. Thus, small adjustments of its curvature have the same effect as putting a lens in front of the eye. To a reasonable approximation, the power of multiple lenses placed close together equals the sum of their powers. For example, a concave spectacle lens (for nearsightedness) having $P=-3.00$ D has the same effect on vision as reducing the power of the eye itself by 3.00 D. So to correct the eye for nearsightedness, the cornea is flattened to reduce its power. Similarly, to correct for farsightedness, the curvature of the cornea is enhanced to increase the power of the eye—the same effect as the positive power spectacle lens used for farsightedness. Laser vision correction uses high intensity electromagnetic radiation to ablate (to remove material from the surface) and reshape the corneal surfaces.

Today, the most commonly used laser vision correction procedure is *Laser in situ Keratomileusis (LASIK)*. The top layer of the cornea is surgically peeled back and the underlying tissue ablated by multiple bursts of finely controlled ultraviolet radiation produced by an excimer laser. Lasers are used because they not only produce well-focused intense light, but they also emit very pure wavelength electromagnetic radiation that can be controlled more accurately than mixed wavelength light. The 193 nm wavelength UV commonly used is extremely and strongly absorbed by corneal tissue, allowing precise evaporation of very thin layers. A computer controlled program applies more bursts, usually at a rate of 10 per second, to the areas that require deeper removal. Typically a spot less than 1 mm in diameter and about 0.3 μm in thickness is removed by each burst. Nearsightedness, farsightedness, and astigmatism can be corrected with an accuracy that produces normal distant vision in more than 90% of the patients, in many cases right away. The corneal flap is replaced; healing takes place rapidly and is nearly painless. More than 1 million Americans per year undergo LASIK (see Figure 5).

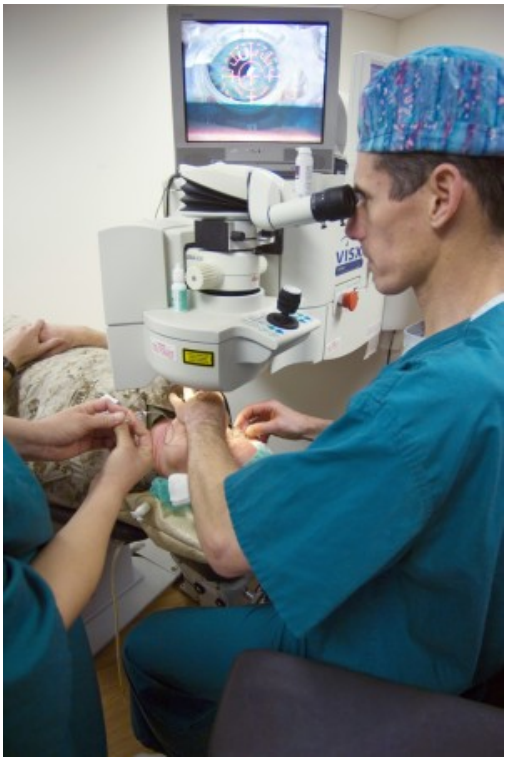


Figure 5. Laser vision correction is being performed using the LASIK procedure. Reshaping of the cornea by laser ablation is based on a careful assessment of the patient's vision and is computer controlled. The upper corneal layer is temporarily peeled back and minimally disturbed in LASIK, providing for more rapid and less painful healing of the less sensitive tissues below. (credit: U.S. Navy photo by Mass Communication Specialist 1st Class Brien Aho)

Section Summary

- Nearsightedness, or myopia, is the inability to see distant objects and is corrected with a diverging lens to reduce power.
- Farsightedness, or hyperopia, is the inability to see close objects and is corrected with a converging lens to increase power.
- In myopia and hyperopia, the corrective lenses

produce images at a distance that the person can see clearly—the far point and near point, respectively.

Conceptual Questions

1. It has become common to replace the cataract-clouded lens of the eye with an internal lens. This intraocular lens can be chosen so that the person has perfect distant vision. Will the person be able to read without glasses? If the person was nearsighted, is the power of the intraocular lens greater or less than the removed lens?
2. If the cornea is to be reshaped (this can be done surgically or with contact lenses) to correct myopia, should its curvature be made greater or smaller? Explain. Also explain how hyperopia can be corrected.
3. If there is a fixed percent uncertainty in LASIK reshaping of the cornea, why would you expect those people with the greatest correction to have a poorer chance of normal distant vision after the procedure?
4. A person with presbyopia has lost some or all of the ability to accommodate the power of the eye. If such a person's distant vision is corrected with LASIK, will she still need reading glasses? Explain.

Problems & Exercises

1. What is the far point of a person whose eyes have a relaxed power of 50.5 D?
2. What is the near point of a person whose eyes have an accommodated power of 53.5 D?
3. (a) A laser vision correction reshaping the cornea of a myopic patient reduces the power of his eye by 9.00 D, with a ±5.0% uncertainty in the final correction. What is the range of diopters for spectacle lenses that this person might need after LASIK procedure? (b) Was the person nearsighted or farsighted before the procedure? How do you know?
4. In a LASIK vision correction, the power of a patient's eye is increased by 3.00 D. Assuming this produces normal close vision, what was the patient's near point before the procedure?
5. What was the previous far point of a patient who had laser vision correction that reduced the power of her eye by 7.00 D, producing normal distant vision for her?
6. A severely myopic patient has a far point of 5.00 cm. By how many diopters should the power of his eye be reduced in laser vision correction to obtain normal distant vision for him?
7. A student's eyes, while reading the blackboard, have a power of 51.0 D. How far is the board from his eyes?
8. The power of a physician's eyes is 53.0 D while examining a patient. How far from her eyes is the feature being examined?
9. A young woman with normal distant vision has a 10.0% ability to accommodate (that is, increase) the power of her eyes. What is the closest object she can see clearly?
10. The far point of a myopic administrator is 50.0 cm. (a) What is the relaxed power of his eyes? (b) If he has the normal 8.00% ability to accommodate, what is the closest object he can see clearly?

11. A very myopic man has a far point of 20.0 cm. What power contact lens (when on the eye) will correct his distant vision?
12. Repeat the previous problem for eyeglasses held 1.50 cm from the eyes.
13. A myopic person sees that her contact lens prescription is −4.00 D. What is her far point?
14. Repeat the previous problem for glasses that are 1.75 cm from the eyes.
15. The contact lens prescription for a mildly farsighted person is 0.750 D, and the person has a near point of 29.0 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?
16. 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?
17. A mother sees that her child's contact lens prescription is 0.750 D. What is the child's near point?
18. Repeat the previous problem for glasses that are 2.20 cm from the eyes.
19. 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?
20. **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?

Glossary

nearsightedness: another term for myopia, a visual defect in which distant objects appear blurred because their images are focused in front of the retina rather than being focused on the retina

myopia: a visual defect in which distant objects appear blurred because their images are focused in front of the retina rather than being focused on the retina

far point: the object point imaged by the eye onto the retina in an unaccommodated eye

farsightedness: another term for hyperopia, the condition of an eye where incoming rays of light reach the retina before they converge into a focused image

hyperopia: the condition of an eye where incoming rays of light reach the retina before they converge into a focused image

near point: the point nearest the eye at which an object is accurately focused on the retina at full accommodation

astigmatism: the result of an inability of the cornea to properly focus an image onto the retina

laser vision correction: a medical procedure used to correct astigmatism and eyesight deficiencies such as myopia and hyperopia

Selected Solutions to Problems & Exercises

1. 2.00 m

3. (a) ±0.45 D; (b) The person was nearsighted because the patient was myopic and the power was reduced.

5. 0.143 m

7. 1.00 m

9. 20.0 cm

11. −5.00 D

13. 25.0 cm

15. −0.198 D

17. 30.8 cm

19. −0.444 D

Microscopes

Lumen Learning

Learning Objectives

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

- Investigate different types of microscopes.
- Learn how image is formed in a compound microscope.

Although the eye is marvelous in its ability to see objects large and small, it obviously has limitations to the smallest details it can detect. Human desire to see beyond what is possible with the naked eye led to the use of optical instruments. In this section we will examine microscopes, instruments for enlarging the detail that we cannot see with the unaided eye. The microscope is a multiple-element system having more than a single lens or mirror. (See Figure 1.) A microscope can be made from two convex lenses. The image formed by the first element becomes the object for the second element. The second element forms its own image, which is the object for the third element, and so on. Ray tracing helps to visualize the image formed. If the device is composed of thin lenses and mirrors that obey the thin lens equations, then it is not difficult to describe their behavior numerically.

Figure 1. Multiple lenses and mirrors are used in this microscope. (credit: U.S. Navy photo by Tom Watanabe)

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 as shown schematically in Figure 2. The first lens is called the *objective lens*, and has typical magnification values from 5× to 100×. In standard microscopes, the objectives are mounted such that when you switch between objectives, the sample remains in focus. Objectives arranged in this way are described as parfocal. The second, the *eyepiece*, also referred to as the ocular, has several lenses which slide inside a cylindrical barrel. The focusing ability is provided by the movement of both the objective lens and the eyepiece. The purpose of a microscope is to magnify small objects, and both lenses contribute to the final magnification. Additionally, the final enlarged image is produced in a location far enough from the observer to be easily viewed, since the eye cannot focus on objects or images that are too close.

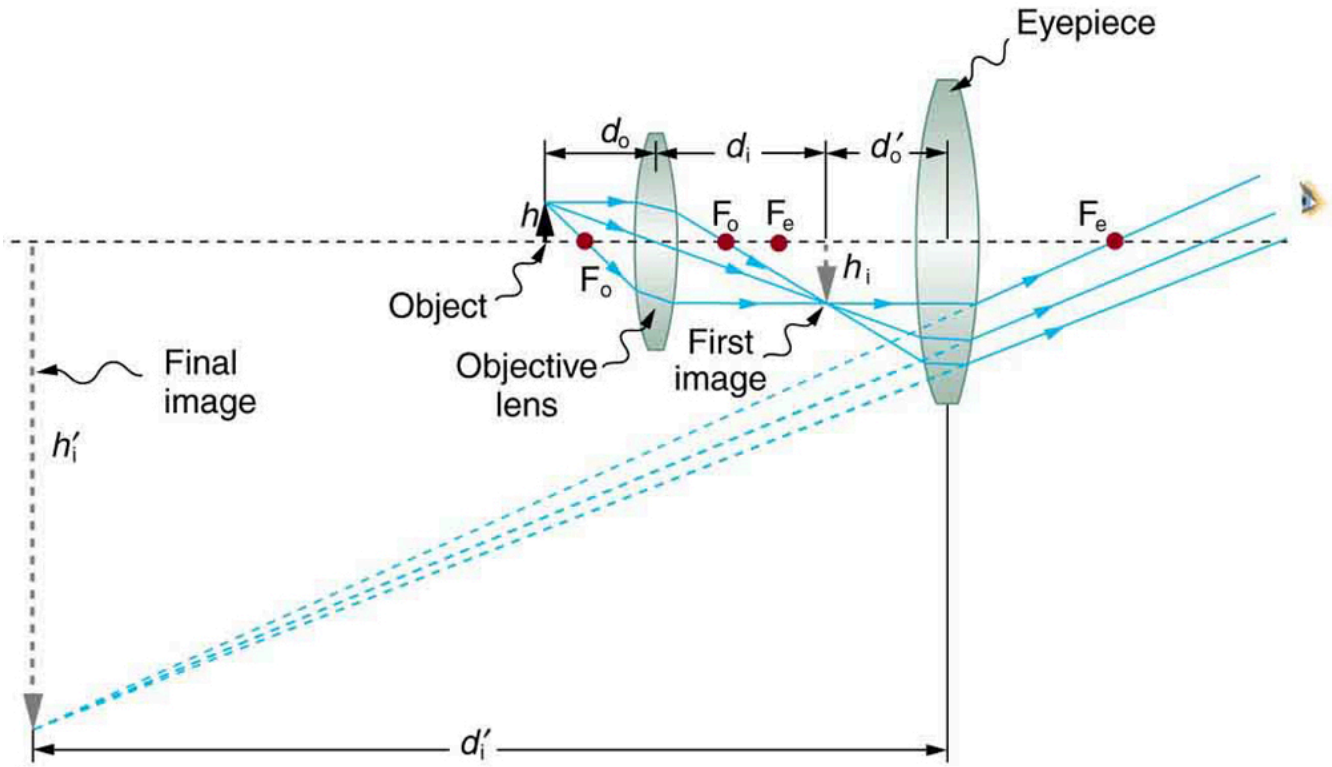


Figure 2. A compound microscope composed of two lenses, an objective and an eyepiece. The objective forms a case 1 image that is larger than the object. This first image is the object for the eyepiece. The eyepiece forms a case 2 final image that is further magnified.

To see how the microscope in Figure 2 forms an image, we consider its two lenses in succession. The object is slightly farther away from the objective lens than its focal length f_o, producing a case 1 image that is larger than the object. This first image is the object for the second lens, or eyepiece. The eyepiece is intentionally located so it can further magnify the image. The eyepiece is placed so that the first image is closer to it than its focal length f_e. Thus the eyepiece acts as a magnifying glass, and the final image is made even larger. The final image remains inverted, but it is farther from the observer, making it easy to view (the eye is most relaxed when viewing distant objects and normally cannot focus closer than 25 cm). Since each lens produces a magnification that multiplies the height of the image, it is apparent that the overall magnification m is the product of the individual magnifications: $m = m_o m_e$, where m_o is the magnification of the objective and m_e is the magnification of the eyepiece. This equation can be generalized for any combination of thin lenses and mirrors that obey the thin lens equations.

Overall Magnification

The overall magnification of a multiple-element system is the product of the individual magnifications of its elements.

Example 1. 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 and Concept

This situation is similar to that shown in Figure 2. To find the overall magnification, we must find the magnification of the objective, then the magnification of the eyepiece. This involves using the thin lens equation.

Solution

The magnification of the objective lens is given as

$$m_{\mathrm{o}} = -\frac{d_{\mathrm{i}}}{d_{\mathrm{o}}}$$

,

where d_{o} and d_{i} are the object and image distances, respectively, for the objective lens as labeled in Figure 2. The object distance is given to be d_{o}=6.20 mm, but the image distance d_{i} is not known. Isolating d_{i}, we have

$$\frac{1}{d_{\mathrm{i}}} = \frac{1}{f_{\mathrm{o}}} - \frac{1}{d_{\mathrm{o}}}$$

,

where f_{o} is the focal length of the objective lens. Substituting known values gives

$$\frac{1}{d_{\mathrm{i}}} = \frac{1}{6.00\ \mathrm{mm}} - \frac{1}{6.20\ \mathrm{mm}} = \frac{0.00538}{\mathrm{mm}}$$

.

We invert this to find d_{i}: d_{i} = 186 mm.

Substituting this into the expression for m_{o} gives

$$m_{\mathrm{o}} = -\frac{d_{\mathrm{i}}}{d_{\mathrm{o}}} = -\frac{186\ \mathrm{mm}}{6.20\ \mathrm{mm}} = -30.0$$

.

Now we must find the magnification of the eyepiece, which is given by

$$m_{\mathrm{e}} = -\frac{d_{\mathrm{i}}\prime}{d_{\mathrm{o}}\prime}$$

, where d_{i}' and d_{o}' are the image and object distances for the eyepiece (see Figure 2). The object distance is the distance of the first image from the eyepiece. Since the first image is 186 mm to the right of the objective and the eyepiece is 230 mm to the right of the objective, the object distance is d_{o}' = 230 mm − 186 mm = 44.0 mm. This places the first image closer to the eyepiece than its focal length, so that the eyepiece will form a case 2 image as shown in the figure. We still need to find the location of the final image d_{i}' in order to find the magnification. This is done as before to obtain a value for

$$\frac{1}{d_{\mathrm{i}}\prime}$$

:

$$\frac{1}{d_{\mathrm{i}}\prime} = \frac{1}{f_{\mathrm{e}}} - \frac{1}{d_{\mathrm{o}}\prime} = \frac{1}{50.0\ \mathrm{mm}} - \frac{1}{44.0\ \mathrm{mm}} = \frac{0.00273}{\mathrm{mm}}$$

Inverting gives

$$d_{\mathrm{i}}\prime = -\frac{\mathrm{mm}}{0.00273} = -367\ \mathrm{mm}$$

.

The eyepiece's magnification is thus

$$m_{\mathrm{e}} = -\frac{d_{\mathrm{i}}\prime}{d_{\mathrm{o}}\prime} = -\frac{-367\ \mathrm{mm}}{44.0\ \mathrm{mm}} = 8.33$$

.

So the overall magnification is $m = m_{\mathrm{o}}m_{\mathrm{e}} = (-30.0)(8.33) = -250$.

Discussion

Both the objective and the eyepiece contribute to the overall magnification, which is large and negative, consistent with Figure 2, 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 (case 2 and case 3 images for single elements are virtual and upright). The final image is 367 mm (0.367 m) to the left of the eyepiece. Had the eyepiece been placed farther from the objective, it could have formed a case 1 image to the right. Such an image could be projected on a screen, but it would be behind the head of the person in the figure and not appropriate for direct viewing. The procedure used to solve this example is applicable in any multiple-element system. Each element is treated in turn, with each forming an image that becomes the object for the next element. The process is not more difficult than for single lenses or mirrors, only lengthier.

Normal optical microscopes can magnify up to 1500× with a theoretical resolution of −0.2 μm. The lenses can be quite complicated and are composed of multiple elements to reduce aberrations. Microscope objective lenses are particularly important as they primarily gather light from the specimen. Three parameters describe microscope objectives: the *numerical aperture* (*NA*), the magnification (*m*), and the working distance. The *NA* is related to the light gathering ability of a lens and is obtained using the angle of acceptance θ formed by the maximum cone of rays focusing on the specimen (see Figure 3a) and is given by $NA = n \sin \alpha$, where n is the refractive index of the medium between the lens and the specimen and

$$\alpha = \frac{\theta}{2}$$

. As the angle of acceptance given by θ increases, *NA* becomes larger and more light is gathered from a smaller focal region giving higher resolution. A 0.75 *NA* objective gives more detail than a 0.10 *NA* objective.

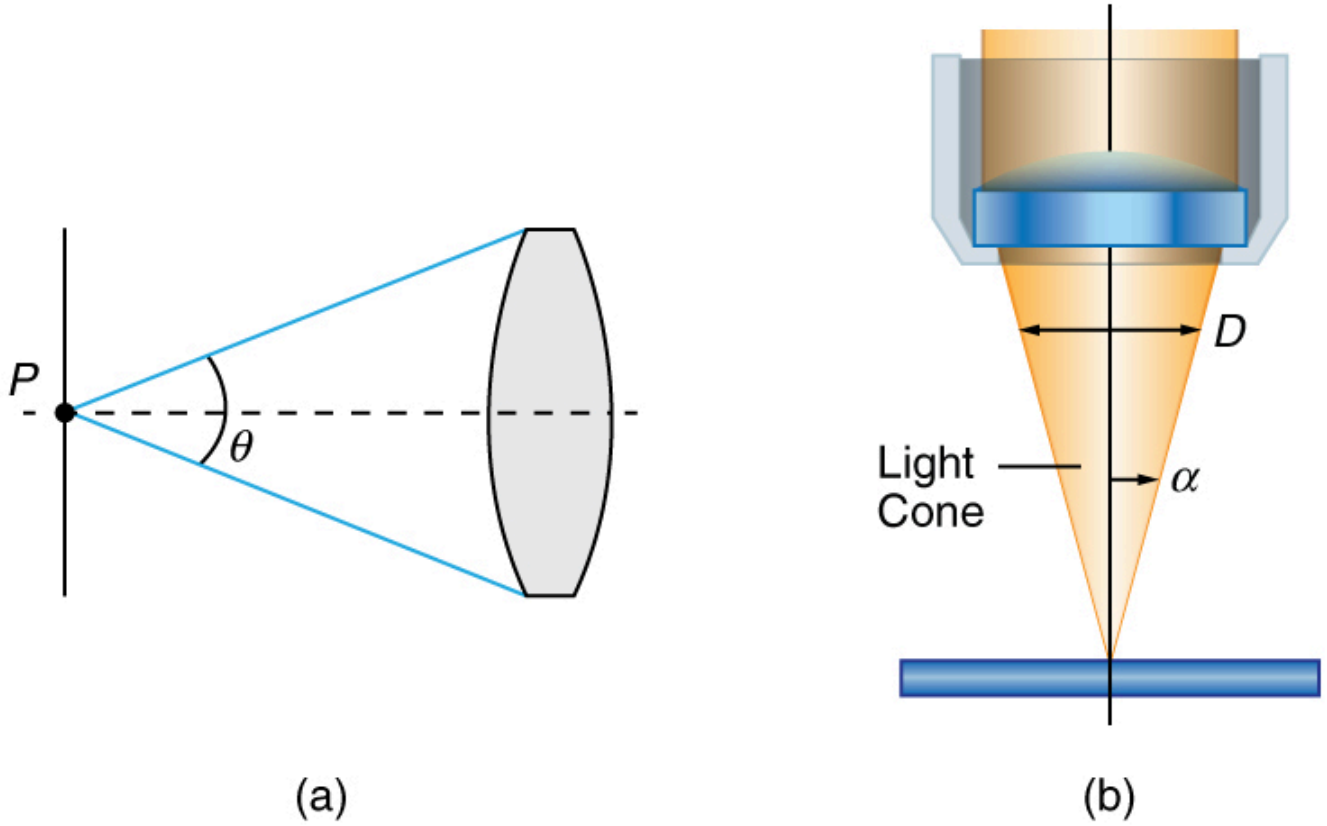


Figure 3. (a) The numerical aperture of a microscope objective lens refers to the light-gathering ability of the lens and is calculated using half the angle of acceptance . (b) Here, is half the acceptance angle for light rays from a specimen entering a camera lens, and is the diameter of the aperture that controls the light entering the lens.

While the numerical aperture can be used to compare resolutions of various objectives, it does not indicate how far the lens could be from the specimen. This is specified by the "working distance," which is the distance (in mm usually) from the front lens element of the objective to the specimen, or cover glass. The higher the *NA* the closer the lens will be to the specimen and the more chances there are of breaking the cover slip and damaging both the specimen and the lens. The focal length of an objective lens is different than the working distance. This is because objective lenses are made of a combination of lenses and the focal length is measured from inside the barrel. The working distance is a parameter that microscopists can use more readily as it is measured from the outermost lens. The working distance decreases as the *NA* and magnification both increase.

The term *f*/# in general is called the *f*-number and is used to denote the light per unit area reaching the image plane. In photography, an image of an object at infinity is formed at the focal point and the *f*-number is given by the ratio of the focal length *f* of the lens and the diameter *D* of the aperture controlling the light into the lens (see Figure 3b). If the acceptance angle is small the *NA* of the lens can also be used as given below.

$$f/\# = \frac{f}{D} \approx \frac{1}{2NA}$$

As the *f*-number decreases, the camera is able to gather light from a larger angle, giving wide-angle photography. As usual there is a trade-off. A greater *f*/# means less light reaches the image plane. A setting of *f*/16 usually allows one to take pictures in bright sunlight as the aperture diameter is small. In

optical fibers, light needs to be focused into the fiber. Figure 4 shows the angle used in calculating the *NA* of an optical fiber.

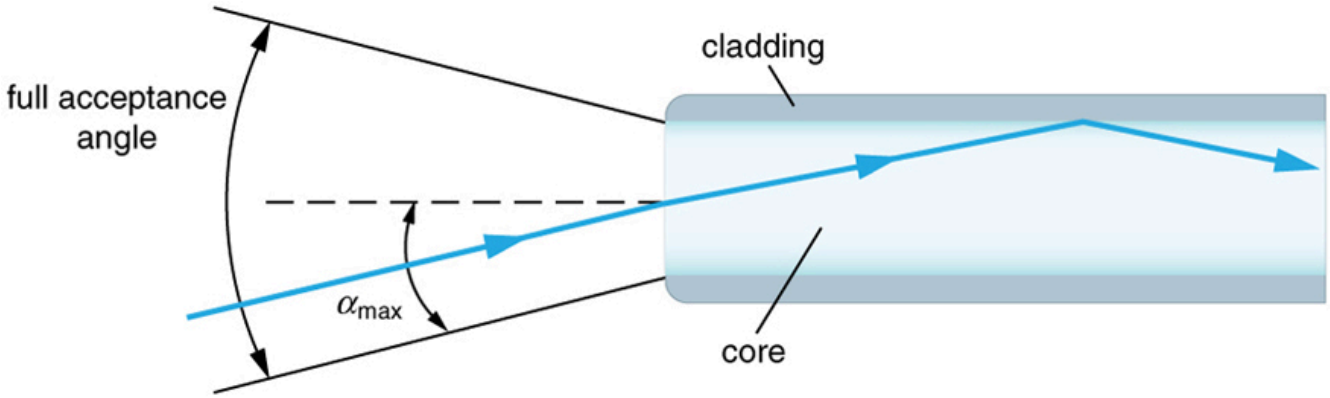


Figure 4. Light rays enter an optical fiber. The numerical aperture of the optical fiber can be determined by using the angle αmax.

Can the *NA* be larger than 1.00? The answer is 'yes' if we use immersion lenses in which a medium such as oil, glycerine or water is placed between the objective and the microscope cover slip. This minimizes the mismatch in refractive indices as light rays go through different media, generally providing a greater light-gathering ability and an increase in resolution. Figure 5 shows light rays when using air and immersion lenses.

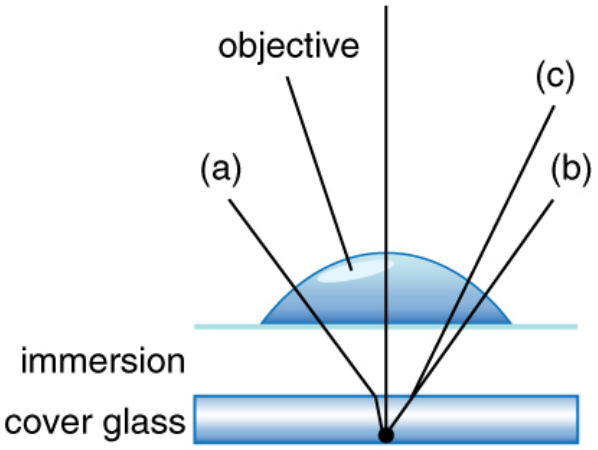


Figure 5. Light rays from a specimen entering the objective. Paths for immersion medium of air (a), water (b) (n = 1.33), and oil (c) (n = 1.51) are shown. The water and oil immersions allow more rays to enter the objective, increasing the resolution.

When using a microscope we do not see the entire extent of the sample. Depending on the eyepiece and objective lens we see a restricted region which we say is the field of view. The objective is then manipulated in two-dimensions above the sample to view other regions of the sample. Electronic scanning of either the objective or the sample is used in scanning microscopy. The image formed at each point during the scanning is combined using a computer to generate an image of a larger region of the sample at a selected magnification.

When using a microscope, we rely on gathering light to form an image. Hence most specimens need to be illuminated, particularly at higher magnifications, when observing details that are so small that they reflect only small amounts of light. To make such objects easily visible, the intensity of light falling on them needs to be increased. Special illuminating systems called condensers are used for this purpose. The type of condenser that is suitable for an application depends on how the specimen is examined, whether by transmission, scattering or reflecting. See Figure 6 for an example of each. White light sources are common and lasers are often used. Laser light illumination tends to be quite intense and it is important to ensure that the light does not result in the degradation of the specimen.

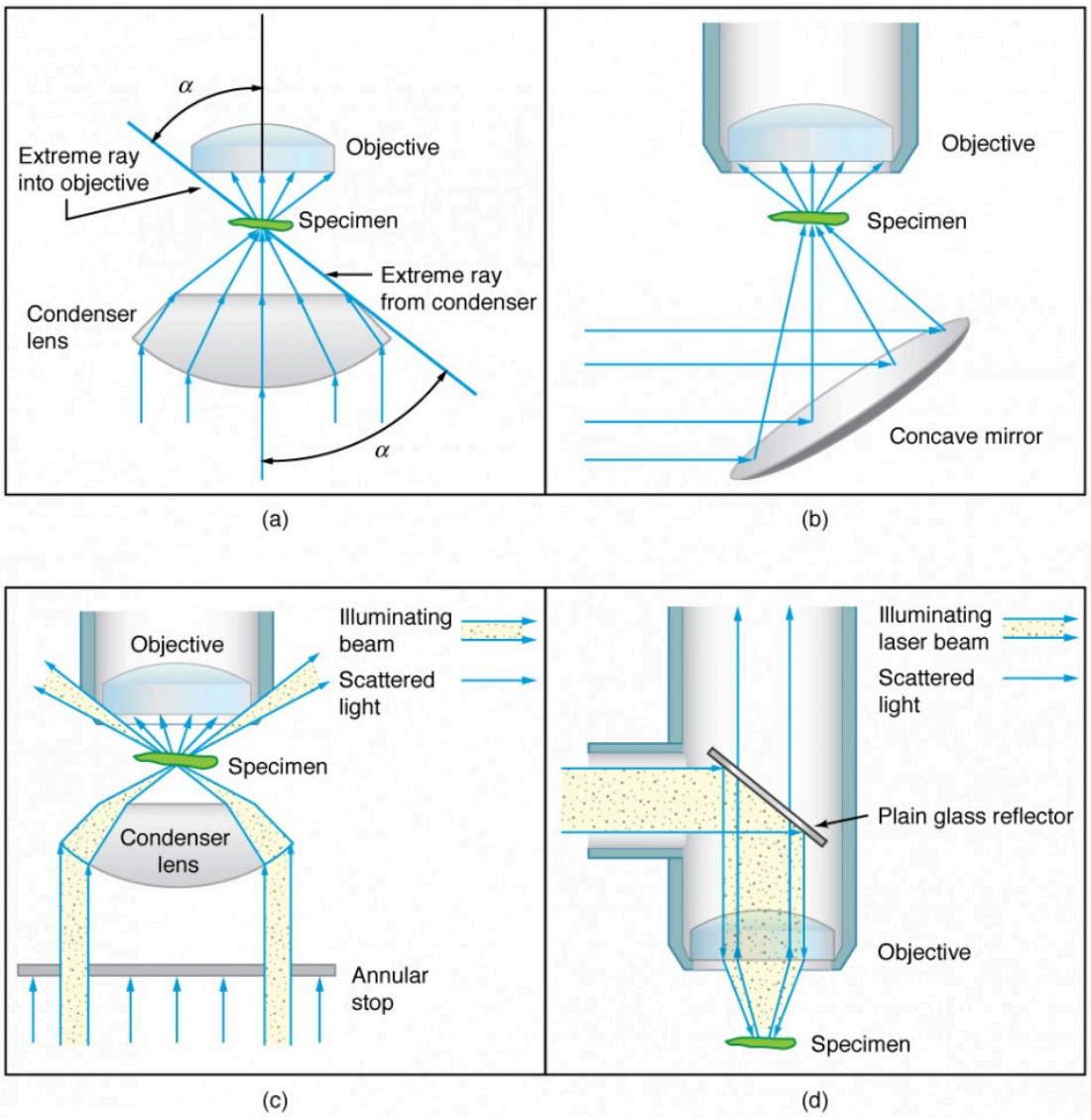


Figure 6. Illumination of a specimen in a microscope. (a) Transmitted light from a condenser lens. (b) Transmitted light from a mirror condenser. (c) Dark field illumination by scattering (the illuminating beam misses the objective lens). (d) High magnification illumination with reflected light – normally laser light.

We normally associate microscopes with visible light but x ray and electron microscopes provide greater resolution. The focusing and basic physics is the same as that just described, even though the lenses require different technology. The electron microscope requires vacuum chambers so that the electrons can proceed unheeded. Magnifications of 50 million times provide the ability to determine positions of individual atoms within materials. An electron microscope is shown in Figure 7.

Figure 7. An electron microscope has the capability to image individual atoms on a material. The microscope uses vacuum technology, sophisticated detectors and state of the art image processing software. (credit: Dave Pape)

We do not use our eyes to form images; rather images are recorded electronically and displayed on computers. In fact observing and saving images formed by optical microscopes on computers is now done routinely. Video recordings of what occurs in a microscope can be made for viewing by many people at later dates. Physics provides the science and tools needed to generate the sequence of time-lapse images of meiosis similar to the sequence sketched in Figure 8.

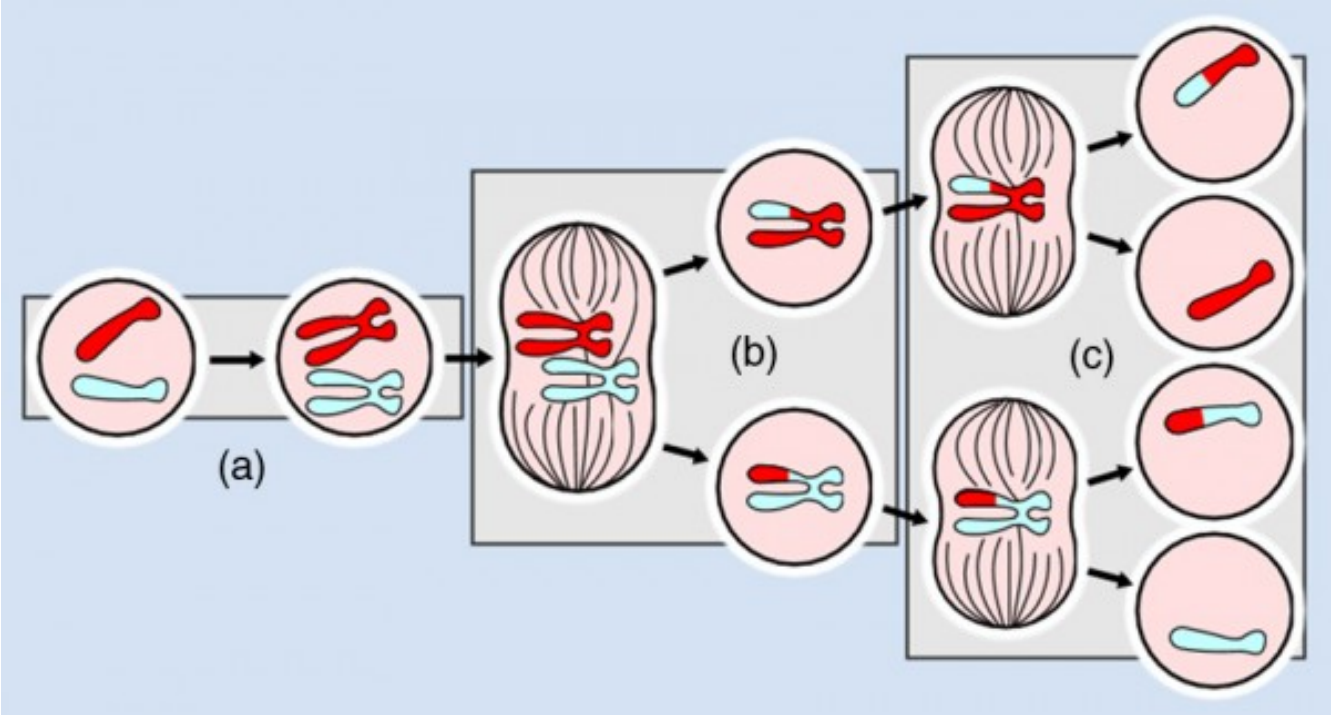


Figure 8. The image shows a sequence of events that takes place during meiosis. (credit: PatríciaR, Wikimedia Commons; National Center for Biotechnology Information)

Take-Home Experiment: Make a Lens

Look through a clear glass or plastic bottle and describe what you see. Now fill the bottle with water and describe what you see. Use the water bottle as a lens to produce the image of a bright object and estimate the focal length of the water bottle lens. How is the focal length a function of the depth of water in the bottle?

Section Summary

- The microscope is a multiple-element system having more than a single lens or mirror.
- Many optical devices contain more than a single lens or mirror. These are analysed 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 apply to each lens element.
- The overall magnification of a multiple-element system is the product of the magnifications of its individual elements. For a two-element system with an objective and an eyepiece, this is $m = m_o m_e$, where m_o is the magnification of the objective and m_e is the magnification of the eyepiece, such as for a microscope.
- Microscopes are instruments for allowing us to see detail we would not be able to see with the unaided eye and consist of a range of components.
- The eyepiece and objective contribute to the magnification. The numerical aperture *NA* of an objective is given by
  $NA = n \sin \alpha$ where n is the refractive index and α the angle of acceptance.
- Immersion techniques are often used to improve the light gathering ability of microscopes. The specimen is illuminated by transmitted, scattered or reflected light though a condenser.
- The *f*/# describes the light gathering ability of a lens. It is given by $f/\# = \frac{f}{D} \approx \frac{1}{2NA}$.

Conceptual Questions

1. Geometric optics describes the interaction of light with macroscopic objects. Why, then, is it correct to use geometric optics to analyse a microscope's image?
2. The image produced by the microscope in Figure 2 cannot be projected. Could extra lenses or mirrors project it? Explain.
3. Why not have the objective of a microscope form a case 2 image with a large magnification? (Hint: Consider the location of that image and the difficulty that would pose for using the eyepiece as a magnifier.)
4. What advantages do oil immersion objectives offer?
5. How does the *NA* of a microscope compare with the *NA* of an optical fiber?

Problems & Exercises

1. A microscope with an overall magnification of 800 has an objective that magnifies by 200. (a) What is the 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?
2. (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× eyepiece (one

that produces a magnification of 8.00) is used?

3. (a) Where does an object need to be placed relative to a microscope for its 0.500 cm focal length objective to produce a magnification of −400? (b) Where should the 5.00 cm focal length eyepiece be placed to produce a further fourfold (4.00) magnification?
4. You switch from a 1.40 *NA* 60× oil immersion objective to a 1.40 *NA* 60× oil immersion objective. What are the acceptance angles for each? Compare and comment on the values. Which would you use first to locate the target area on your specimen?
5. 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 magnification is produced by the eyepiece? (e) What is the overall magnification? (See Figure 2.)
6. You are using a standard microscope with a 0.10 *NA* 4× objective and switch to a 0.65 *NA* 40× objective. What are the acceptance angles for each? Compare and comment on the values. Which would you use first to locate the target area on of your specimen? (See Figure 3.)
7. **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?

Glossary

compound microscope: a microscope constructed from two convex lenses, the first serving as the ocular lens(close to the eye) and the second serving as the objective lens

objective lens: the lens nearest to the object being examined

eyepiece: the lens or combination of lenses in an optical instrument nearest to the eye of the observer

numerical aperture: a number or measure that expresses the ability of a lens to resolve fine detail in an object being observed. Derived by mathematical formula $NA = n \sin \alpha$, where n is the refractive index of the medium between the lens and the specimen and

$$\alpha = \frac{\theta}{2}$$

Selected Solutions to Problems & Exercises

1. (a) 4.00; (b) 1600

3. (a) 0.501 cm; (b) Eyepiece should be 204 cm behind the objective lens.

5. (a) +18.3 cm (on the eyepiece side of the objective lens); (b) −60.0; (c) −11.3 cm (on the objective side of the eyepiece); (d) +6.67; (e) −400

Color and Color Vision

Lumen Learning

Learning Objectives

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

- Explain the simple theory of color vision.
- Outline the coloring properties of light sources.
- Describe the retinex theory of color vision.

The gift of vision is made richer by the existence of color. Objects and lights abound with thousands of hues that stimulate our eyes, brains, and emotions. Two basic questions are addressed in this brief treatment—what does color mean in scientific terms, and how do we, as humans, perceive it?

Simple Theory of Color Vision

We have already noted that color is associated with the wavelength of visible electromagnetic radiation. When our eyes receive pure-wavelength light, we tend to see only a few colors. Six of these (most often listed) are red, orange, yellow, green, blue, and violet. These are the rainbow of colors produced when white light is dispersed according to different wavelengths. There are thousands of other *hues* that we can perceive. These include brown, teal, gold, pink, and white. One simple theory of color vision implies that all these hues are our eye's response to different combinations of wavelengths. This is true to an extent, but we find that color perception is even subtler than our eye's response for various wavelengths of light.

The two major types of light-sensing cells (photoreceptors) in the retina are *rods and cones*

Take-Home Experiment: Rods and Cones

1. Go into a darkened room from a brightly lit room, or from outside in the Sun. How long did it take to start seeing shapes more clearly? What about color? Return to the bright room. Did it take a few minutes before you could see things clearly?
2. Demonstrate the sensitivity of foveal vision. Look at the letter G in the word ROGERS. What about the clarity of the letters on either side of G?

Cones are most concentrated in the fovea, the central region of the retina. There are no rods here. The fovea is at the center of the macula, a 5 mm diameter region responsible for our central vision. The cones work best in bright light and are responsible for high resolution vision. There are about 6 million cones in the human retina. There are three types of cones, and each type is sensitive to different ranges of wavelengths, as illustrated in Figure 1.

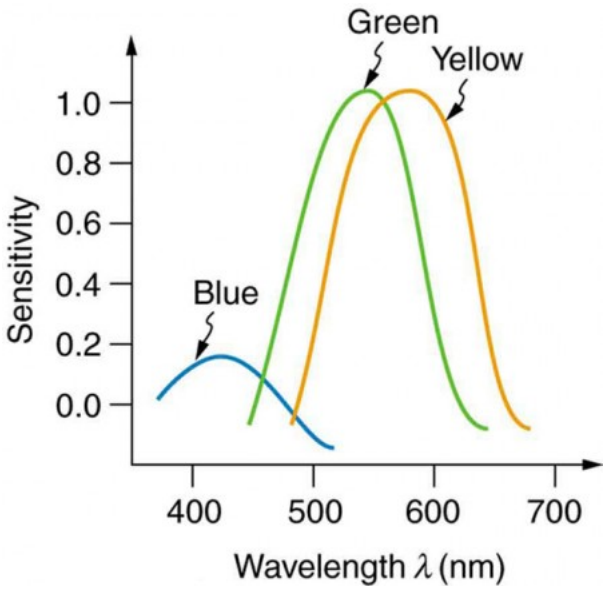


Figure 1. The image shows the relative sensitivity of the three types of cones, which are named according to wavelengths of greatest sensitivity. Rods are about 1000 times more sensitive, and their curve peaks at about 500 nm. Evidence for the three types of cones comes from direct measurements in animal and human eyes and testing of color blind people.

A *simplified theory of color vision* is that there are three *primary colors* corresponding to the three types of cones. The thousands of other hues that we can distinguish among are created by various combinations of stimulations of the three types of cones. Color television uses a three-color system in which the screen is covered with equal numbers of red, green, and blue phosphor dots. The broad range of hues a viewer sees is produced by various combinations of these three colors. For example, you will perceive yellow when red and green are illuminated with the correct ratio of intensities. White may be sensed when all three are illuminated. Then, it would seem that all hues can be produced by adding three primary colors in various proportions. But there is an indication that color vision is more sophisticated. There is no unique set of three primary colors. Another set that works is yellow, green, and blue. A further indication of the need for a more complex theory of color vision is that various different combinations can produce the same hue. Yellow can be sensed with yellow light, or with a combination of red and green, and also with white light from which violet has been removed. The three-primary-colors aspect of color vision is well established; more sophisticated theories expand on it rather than deny it.

Consider why various objects display color—that is, why are feathers blue and red in a crimson rosella? The *true color of an object* is defined by its absorptive or reflective characteristics. Figure 2 shows white light falling on three different objects, one pure blue, one pure red, and one black, as well as pure red light falling on a white object. Other hues are created by more complex absorption characteristics. Pink, for example on a galah cockatoo, can be due to weak absorption of all colors except red. An object can appear a different color under non-white illumination. For example, a pure blue object illuminated with pure red light will *appear* black, because it absorbs all the red light falling on it. But, the true color of the object is blue, which is independent of illumination.

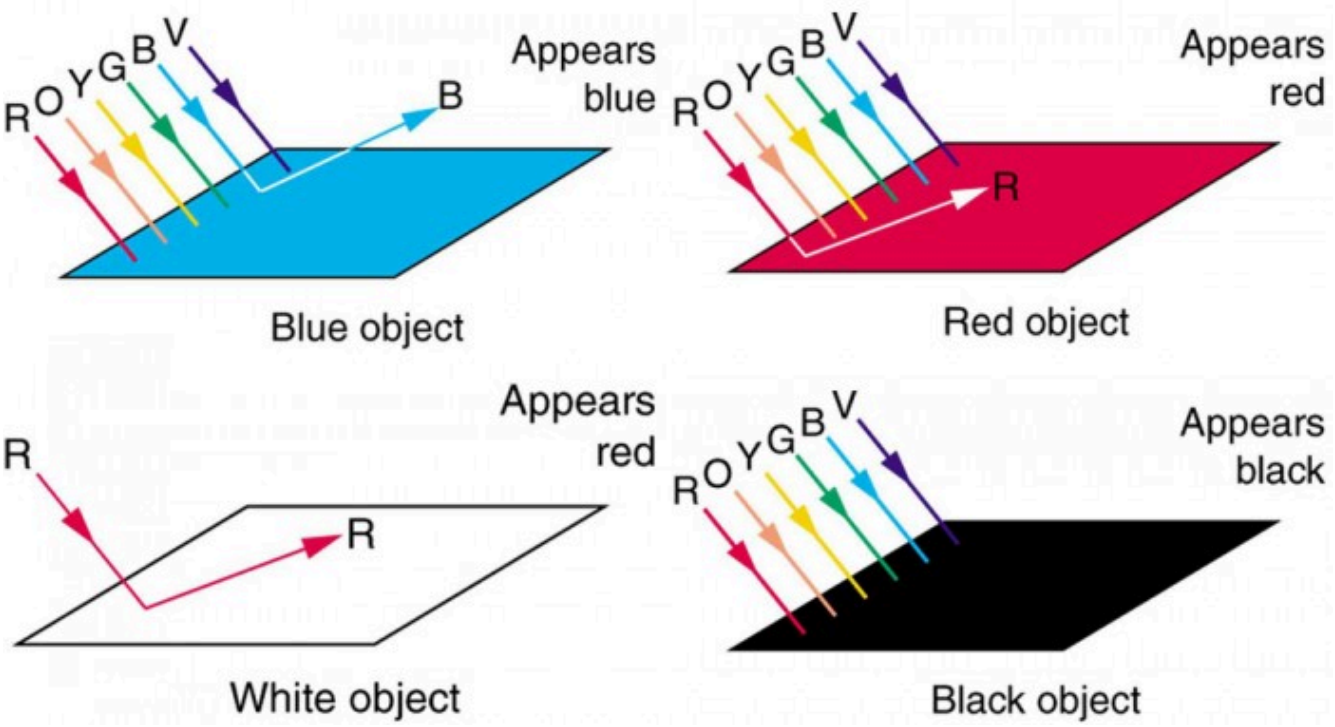


Figure 2. Absorption characteristics determine the true color of an object. Here, three objects are illuminated by white light, and one by pure red light. White is the equal mixture of all visible wavelengths; black is the absence of light.

Similarly, *light sources have colors* that are defined by the wavelengths they produce. A helium-neon laser emits pure red light. In fact, the phrase "pure red light" is defined by having a sharp constrained spectrum, a characteristic of laser light. The Sun produces a broad yellowish spectrum, fluorescent lights emit bluish-white light, and incandescent lights emit reddish-white hues as seen in Figure 3. As you would expect, you sense these colors when viewing the light source directly or when illuminating a white object with them. All of this fits neatly into the simplified theory that a combination of wavelengths produces various hues.

Take-Home Experiment: Exploring Color Addition

This activity is best done with plastic sheets of different colors as they allow more light to pass through to our eyes. However, thin sheets of paper and fabric can also be used. Overlay different colors of the material and hold them up to a white light. Using the theory described above, explain the colors you observe. You could also try mixing different crayon colors.

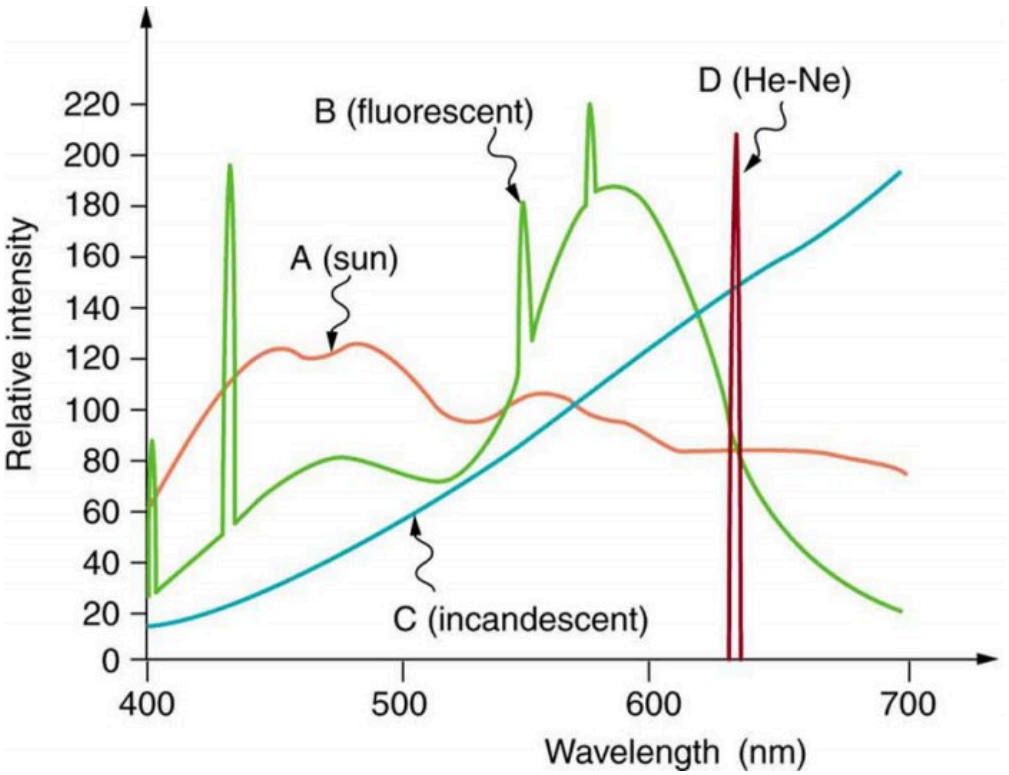


Figure 3. Emission spectra for various light sources are shown. Curve A is average sunlight at Earth's surface, curve B is light from a fluorescent lamp, and curve C is the output of an incandescent light. The spike for a helium-neon laser (curve D) is due to its pure wavelength emission. The spikes in the fluorescent output are due to atomic spectra—a topic that will be explored later.

Color Constancy and a Modified Theory of Color Vision

The eye-brain color-sensing system can, by comparing various objects in its view, perceive the true color of an object under varying lighting conditions—an ability that is called *color constancy*. We can sense that a white tablecloth, for example, is white whether it is illuminated by sunlight, fluorescent light, or candlelight. The wavelengths entering the eye are quite different in each case, as the graphs in Figure 3 imply, but our color vision can detect the true color by comparing the tablecloth with its surroundings.

Theories that take color constancy into account are based on a large body of anatomical evidence as well as perceptual studies. There are nerve connections among the light receptors on the retina, and there are far fewer nerve connections to the brain than there are rods and cones. This means that there is signal processing in the eye before information is sent to the brain. For example, the eye makes comparisons between adjacent light receptors and is very sensitive to edges as seen in Figure 4. Rather than responding simply to the light entering the eye, which is uniform in the various rectangles in this figure, the eye responds to the edges and senses false darkness variations.

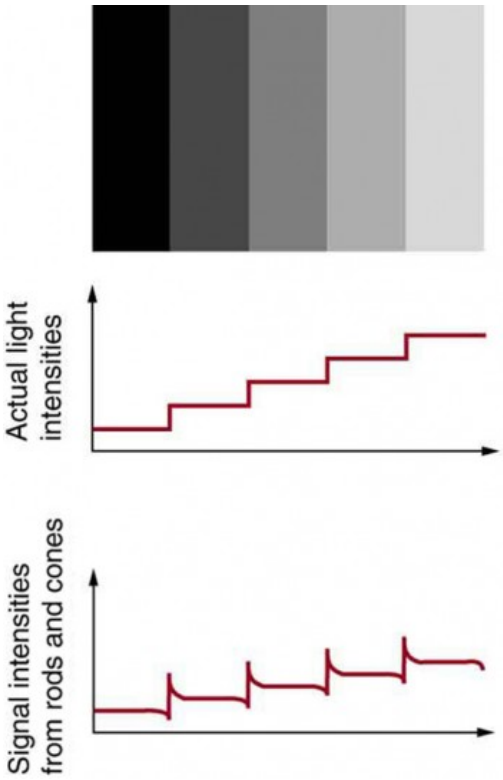
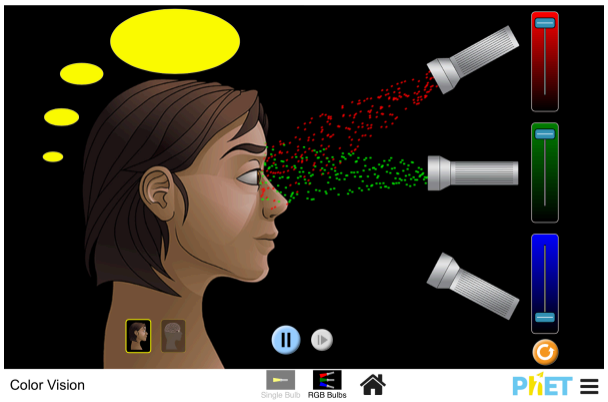


Figure 4. The importance of edges is shown. Although the grey strips are uniformly shaded, as indicated by the graph immediately below them, they do not appear uniform at all. Instead, they are perceived darker on the dark side and lighter on the light side of the edge, as shown in the bottom graph. This is due to nerve impulse processing in the eye.

One theory that takes various factors into account was advanced by Edwin Land (1909–1991), the creative founder of the Polaroid Corporation. Land proposed, based partly on his many elegant experiments, that the three types of cones are organized into systems called *retinexes*. Each retinex forms an image that is compared with the others, and the eye-brain system thus can compare a candle-illuminated white table cloth with its generally reddish surroundings and determine that it is actually white. This *retinex theory of color vision* is an example of modified theories of color vision that attempt to account for its subtleties. One striking experiment performed by Land demonstrates that some type of image comparison may produce color vision. Two pictures are taken of a scene on black-and-white film, one using a red filter, the other a blue filter. Resulting black-and-white slides are then projected and superimposed on a screen, producing a black-and-white image, as expected. Then a red filter is placed in front of the slide taken with a red filter, and the images are again superimposed on a screen. You would expect an image in various shades of pink, but instead, the image appears to humans in full color with all the hues of the original scene. This implies that color vision can be induced by comparison of the black-and-white and red images. Color vision is not completely understood or explained, and the retinex theory is not totally accepted. It is apparent that color vision is much subtler than what a first look might imply.

PhET Explorations: Color Vision

Make a whole rainbow by mixing red, green, and blue light. Change the wavelength of a monochromatic beam or filter white light. View the light as a solid beam, or see the individual photons.



Click to run the simulation.

Section Summary

- The eye has four types of light receptors—rods and three types of color-sensitive cones.
- The rods are good for night vision, peripheral vision, and motion changes, while the cones are responsible for central vision and color.
- We perceive many hues, from light having mixtures of wavelengths.
- A simplified theory of color vision states that there are three primary colors, which correspond to the three types of cones, and that various combinations of the primary colors produce all the hues.
- The true color of an object is related to its relative absorption of various wavelengths of light. The color of a light source is related to the wavelengths it produces.
- Color constancy is the ability of the eye-brain system to discern the true color of an object illuminated by various light sources.
- The retinex theory of color vision explains color constancy by postulating the existence of three retinexes or image systems, associated with the three types of cones that are compared to obtain sophisticated information.

Conceptual Questions

1. A pure red object on a black background seems to disappear when illuminated with pure green light. Explain why.
2. What is color constancy, and what are its limitations?
3. There are different types of color blindness related to the malfunction of different types of cones. Why would it be particularly useful to study those rare individuals who are color blind only in

one eye or who have a different type of color blindness in each eye?

4. Propose a way to study the function of the rods alone, given they can sense light about 1000 times dimmer than the cones.

Glossary

hues: identity of a color as it relates specifically to the spectrum

rods and cones: two types of photoreceptors in the human retina; rods are responsible for vision at low light levels, while cones are active at higher light levels

simplified theory of color vision: a theory that states that there are three primary colors, which correspond to the three types of cones

color constancy: a part of the visual perception system that allows people to perceive color in a variety of conditions and to see some consistency in the color

retinex: a theory proposed to explain color and brightness perception and constancies; is a combination of the words retina and cortex, which are the two areas responsible for the processing of visual information

retinex theory of color vision: the ability to perceive color in an ambient-colored environment

Aberrations

Lumen Learning

Learning Objective

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

- Describe optical aberration.

Real lenses behave somewhat differently from how they are modeled using the thin lens equations, producing *aberrations*. An aberration is a distortion in an image. There are a variety of aberrations due to a lens size, material, thickness, and position of the object. One common type of aberration is chromatic aberration, which is related to color. Since the index of refraction of lenses depends on color or wavelength, images are produced at different places and with different magnifications for different colors. (The law of reflection is independent of wavelength, and so mirrors do not have this problem. This is another advantage for mirrors in optical systems such as telescopes.)

Figure 1a shows chromatic aberration for a single convex lens and its partial correction with a two-lens system. Violet rays are bent more than red, since they have a higher index of refraction and are thus focused closer to the lens. The diverging lens partially corrects this, although it is usually not possible to do so completely. Lenses of different materials and having different dispersions may be used. For example an achromatic doublet consisting of a converging lens made of crown glass and a diverging lens made of flint glass in contact can dramatically reduce chromatic aberration (see Figure 1b).

Quite often in an imaging system the object is off-center. Consequently, different parts of a lens or mirror do not refract or reflect the image to the same point. This type of aberration is called a coma and is shown in Figure 2. The image in this case often appears pear-shaped. Another common aberration is spherical aberration where rays converging from the outer edges of a lens converge to a focus closer to the lens and rays closer to the axis focus further (see Figure 3). Aberrations due to astigmatism in the lenses of the eyes are discussed in Vision Correction, and a chart used to detect astigmatism is shown in Figure 4. Such aberrations and can also be an issue with manufactured lenses.

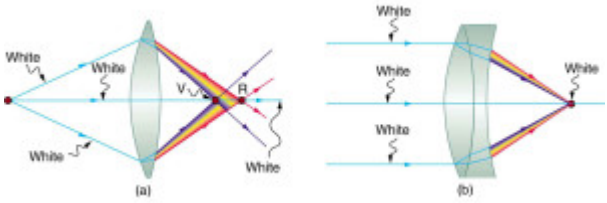


Figure 1. (a) Chromatic aberration is caused by the dependence of a lens's index of refraction on color (wavelength). The lens is more powerful for violet (V) than for red (R), producing images with different locations and magnifications. (b) Multiple-lens systems can partially correct chromatic aberrations, but they may require lenses of different materials and add to the expense of optical systems such as cameras.

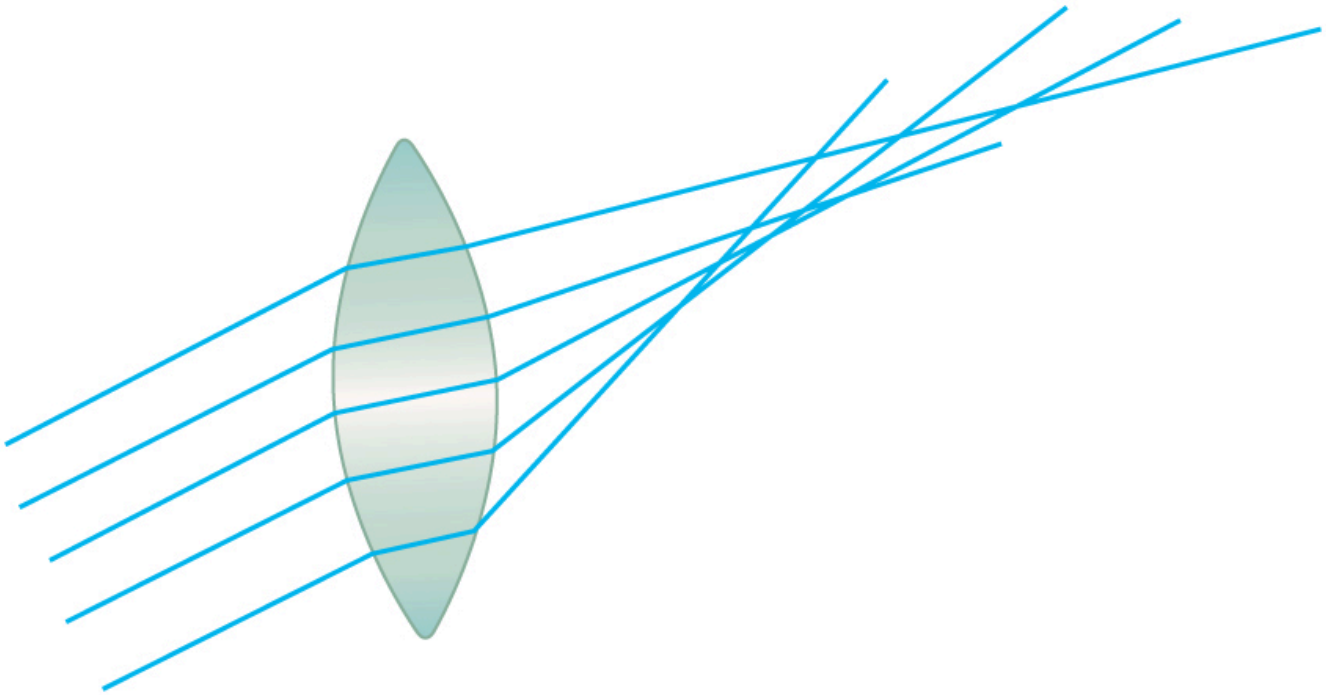


Figure 2. A coma is an aberration caused by an object that is off-center, often resulting in a pear-shaped image. The rays originate from points that are not on the optical axis and they do not converge at one common focal point.

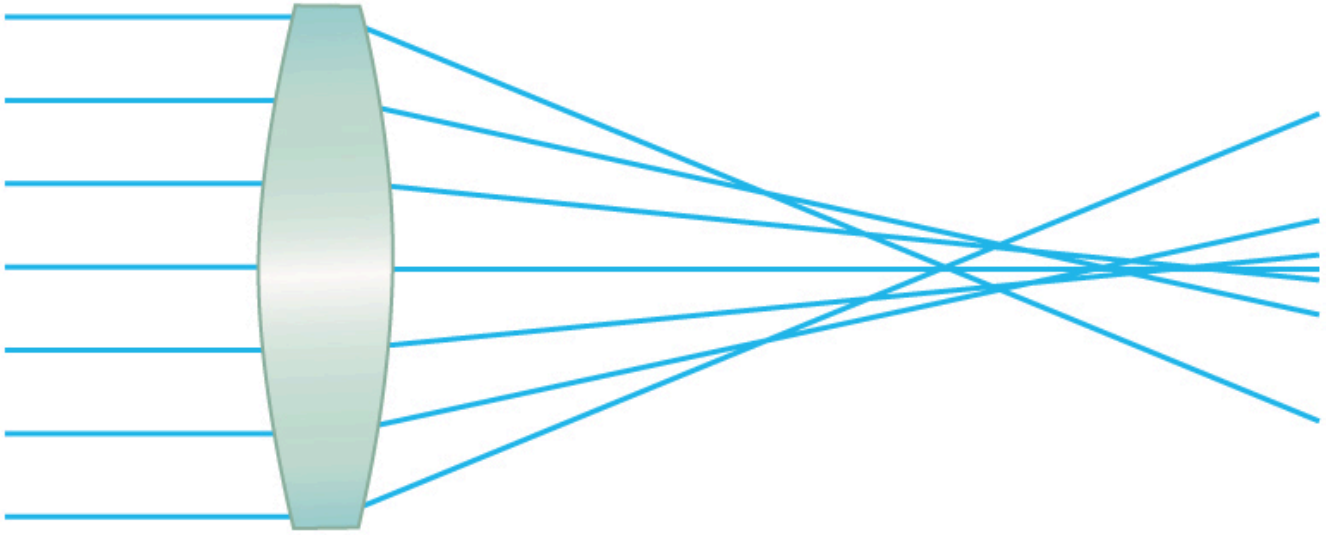


Figure 3. Spherical aberration is caused by rays focusing at different distances from the lens.

The image produced by an optical system needs to be bright enough to be discerned. It is often a challenge to obtain a sufficiently bright image. The brightness is determined by the amount of light passing through the optical system. The optical components determining the brightness are the diameter of the lens and the diameter of pupils, diaphragms or aperture stops placed in front of lenses. Optical systems often have entrance and exit pupils to specifically reduce aberrations but they inevitably reduce brightness as well. Consequently, optical systems need to strike a balance between the various components used. The iris in the eye dilates and constricts, acting as an entrance pupil. You can see objects more clearly by looking through a small hole made with your hand in the shape of a fist. Squinting, or using a small hole in a piece of paper, also will make the object sharper.

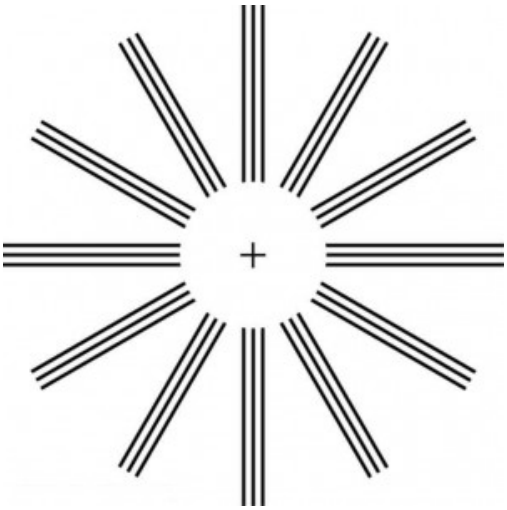


Figure 4. This chart can detect astigmatism, unevenness in the focus of the eye. Check each of your eyes separately by looking at the center cross (without spectacles if you wear them). If lines along some axes appear darker or clearer than others, you have an astigmatism.

So how are aberrations corrected? The lenses may also have specially shaped surfaces, as opposed to the simple spherical shape that is relatively easy to produce. Expensive camera lenses are large in diameter, so that they can gather more light, and need several elements to correct for various aberrations. Further, advances in materials science have resulted in lenses with a range of refractive indices—technically referred to as graded index (GRIN) lenses. Spectacles often have the ability to provide a range of focusing ability using similar techniques. GRIN lenses are particularly important at the end of optical fibers in endoscopes. Advanced computing techniques allow for a range of corrections on images after the image has been collected and certain characteristics of the optical system are known. Some of these techniques are sophisticated versions of what are available on commercial packages like Adobe Photoshop.

Section Summary

- Aberrations or image distortions can arise due to the finite thickness of optical instruments, imperfections in the optical components, and limitations on the ways in which the components are used.
- The means for correcting aberrations range from better components to computational techniques.

Conceptual Questions

1. List the various types of aberrations. What causes them and how can each be reduced?

Problems & Exercises

Integrated Concepts. (a) During laser vision correction, a brief burst of 193 nm ultraviolet light is projected onto the cornea of the patient. It makes a spot 1.00 mm in diameter and deposits 0.500 mJ of energy. Calculate the depth of the layer ablated, assuming the corneal tissue has the same properties as water and is initially at 34.0ºC. The tissue's temperature is increased to 100ºC and evaporated without further temperature increase.
(b) Does your answer imply that the shape of the cornea can be finely controlled?

Glossary

aberration: failure of rays to converge at one focus because of limitations or defects in a lens or mirror

Solutions to Problems & Exercises

(a) 0.251 μm; (b) Yes, this thickness implies that the shape of the cornea can be very finely controlled, producing normal distant vision in more than 90% of patients.

11. Wave Optics

Introduction to Wave Optics

Lumen Learning

Figure 1. The colors reflected by this compact disc vary with angle and are not caused by pigments. Colors such as these are direct evidence of the wave character of light. (credit: Infopro, Wikimedia Commons)

Examine a compact disc under white light, noting the colors observed and locations of the colors. Determine if the spectra are formed by diffraction from circular lines centered at the middle of the disc and, if so, what is their spacing. If not, determine the type of spacing. Also with the CD, explore the spectra of a few light sources, such as a candle flame, incandescent bulb, halogen light, and fluorescent light. Knowing the spacing of the rows of pits in the compact disc, estimate the maximum spacing that will allow the given number of megabytes of information to be stored.

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 2). The same is true for the colors seen in an oil slick or in the light reflected from a compact disc. These and other interesting phenomena, such as the dispersion of white light into a rainbow of colors when passed through a narrow slit, cannot be explained fully by geometric optics. In these cases, light interacts with small objects and exhibits its wave characteristics. The branch of optics that considers the behavior of light when it exhibits wave characteristics (particularly when it interacts with small objects) is called wave optics (sometimes called physical optics). It is the topic of this chapter.

Figure 2. These soap bubbles exhibit brilliant colors when exposed to sunlight. How are the colors produced if they are not pigments in the soap? (credit: Scott Robinson, Flickr)

The Wave Aspect of Light: Interference

Lumen Learning

Learning Objectives

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

- Discuss the wave character of light.
- Identify the changes when light enters a medium.

We know that visible light is the type of electromagnetic wave to which our eyes respond. Like all other electromagnetic waves, it obeys the equation $c = f\lambda$, where $c = 3 \times 10^8$ m/s is the speed of light in vacuum, f is the frequency of the electromagnetic waves, and λ is its wavelength. The range of visible wavelengths is approximately 380 to 760 nm. As is true for all waves, light travels in straight lines and acts like a ray when it interacts with objects several times as large as its wavelength. However, when it interacts with smaller objects, it displays its wave characteristics prominently. Interference is the hallmark of a wave, and in Figure 1 both the ray and wave characteristics of light can be seen. The laser beam emitted by the observatory epitomizes a ray, traveling in a straight line. However, passing a pure-wavelength beam through vertical slits with a size close to the wavelength of the beam reveals the wave character of light, as the beam spreads out horizontally into a pattern of bright and dark regions caused by systematic constructive and destructive interference. Rather than spreading out, a ray would continue traveling straight ahead after passing through slits.

Making Connections: Waves

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 (as we will see in Special Relativity) for matter waves, such as electrons scattered from a crystal.

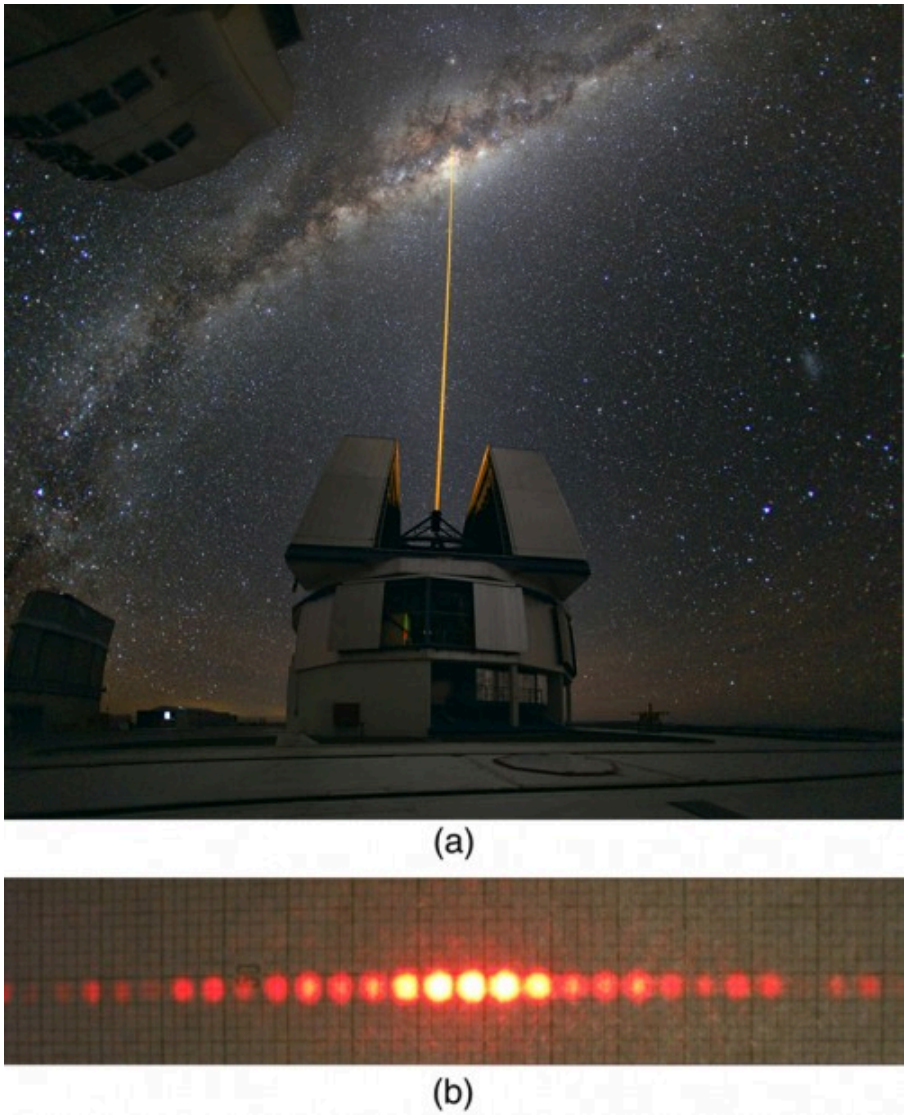


Figure 1. (a) The laser beam emitted by an observatory acts like a ray, traveling in a straight line. This laser beam is from the Paranal Observatory of the European Southern Observatory. (credit: Yuri Beletsky, European Southern Observatory) (b) A laser beam passing through a grid of vertical slits produces an interference pattern—characteristic of a wave. (credit: Shim'on and Slava Rybka, Wikimedia Commons)

Light has wave characteristics in various media as well as in a vacuum. When light goes from a vacuum to some medium, like water, its speed and wavelength change, but its frequency *f* remains the same. (We can think of light as a forced oscillation that must have the frequency of the original source.) The speed of light in a medium is

$$v = \frac{c}{n}$$

, where *n* is its index of refraction. If we divide both sides of equation $c = f\lambda$ by *n*, we get

$$\frac{c}{n} = v = \frac{f\lambda}{n}$$

$\lambda_n = \frac{\lambda}{n}$. This implies that $v = f\lambda_n$, where λ_n is the *wavelength in a medium* and that , where λ is the wavelength in vacuum and n is the medium's index of refraction. Therefore, the wavelength of light is smaller in any medium than it is in vacuum. In water, for example, which has $n = 1.333$, the range of visible wavelengths is $\frac{380\text{ nm}}{1.333}$ to $\frac{760\text{ nm}}{1.333}$, or $\lambda_n = 285$ to 570 nm. Although wavelengths change while traveling from one medium to another, colors do not, since colors are associated with frequency.

Section Summary

- Wave optics is the branch of optics that must be used when light interacts with small objects or whenever the wave characteristics of light are considered.
- Wave characteristics are those associated with interference and diffraction.
- Visible light is the type of electromagnetic wave to which our eyes respond and has a wavelength in the range of 380 to 760 nm.
- Like all EM waves, the following relationship is valid in vacuum: $c = f\lambda$, where $c = 3 \times 10^8$ m/s is the speed of light, f is the frequency of the electromagnetic wave, and λ is its wavelength in vacuum.
- The wavelength λ_n of light in a medium with index of refraction n is $\lambda_n = \frac{\lambda}{n}$. Its frequency is the same as in vacuum.

Conceptual Questions

1. What type of experimental evidence indicates that light is a wave?
2. Give an example of a wave characteristic of light that is easily observed outside the laboratory.

Problems & Exercises

1. Show that when light passes from air to water, its wavelength decreases to 0.750 times its original value.
2. Find the range of visible wavelengths of light in crown glass.
3. What is the index of refraction of a material for which the wavelength of light is 0.671 times its value in a vacuum? Identify the likely substance.
4. Analysis of an interference effect in a clear solid shows that the wavelength of light in the solid is 329 nm. Knowing this light comes from a He-Ne laser and has a wavelength of 633 nm in air, is the substance zircon or diamond?
5. What is the ratio of thicknesses of crown glass and water that would contain the same number of wavelengths of light?

Glossary

wavelength in a medium:

$$\lambda_{\mathrm{n}} =$$

, where λ is the wavelength in vacuum, and n is the index of refraction of the medium

Selected Solutions to Problems & Exercises

1.

$$\frac{1}{1.333} = 0.750$$

3. 1.49, Polystyrene

5. 0.877 glass to water

Huygens's Principle: Diffraction

Lumen Learning

Learning Objectives

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

- Discuss the propagation of transverse waves.
- Discuss Huygens's principle.
- Explain the bending of light.

Figure 1 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 wavefronts (or wave crests) as we would by looking down on the ocean waves. The side view would be a graph of the electric or magnetic field. The view from above is perhaps the most useful in developing concepts about wave optics.

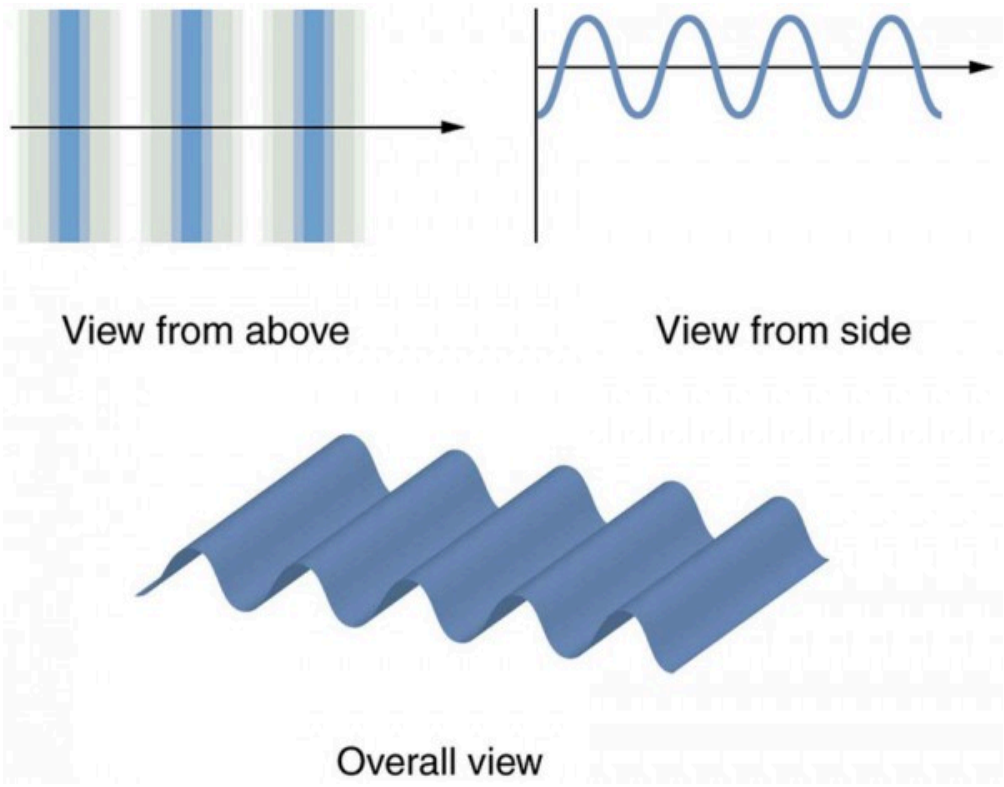


Figure 1. A transverse wave, such as an electromagnetic wave like light, as viewed from above and from the side. The direction of propagation is perpendicular to the wavefronts (or wave crests) and is represented by an arrow like a ray.

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 wavefront is a source of wavelets that spread out in the forward direction at the same speed as the wave itself. The new wavefront is a line tangent to all of the wavelets.

Figure 2 shows how Huygens's principle is applied. A wavefront is the long edge that moves, for example, the crest or the trough. Each point on the wavefront emits a semicircular wave that moves at the propagation speed v. These are drawn at a time t later, so that they have moved a distance $s = vt$. The new wavefront is a line 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. We will find it 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.

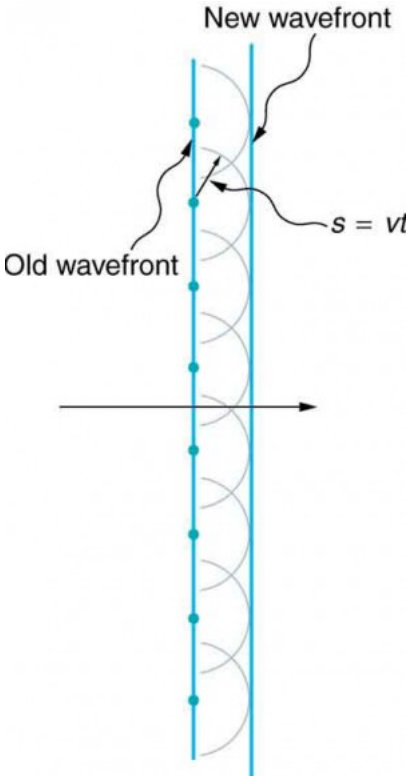


Figure 2. Huygens's principle applied to a straight wavefront. Each point on the wavefront emits a semicircular wavelet that moves a distance . The new wavefront is a line tangent to the wavelets.

Figure 3 shows how a mirror reflects an incoming wave at an angle equal to the incident angle, verifying the law of reflection. As the wavefront strikes the mirror, wavelets are first emitted from the left part of the mirror and then the right. The wavelets closer to the left have had time to travel farther, producing a wavefront traveling in the direction shown.

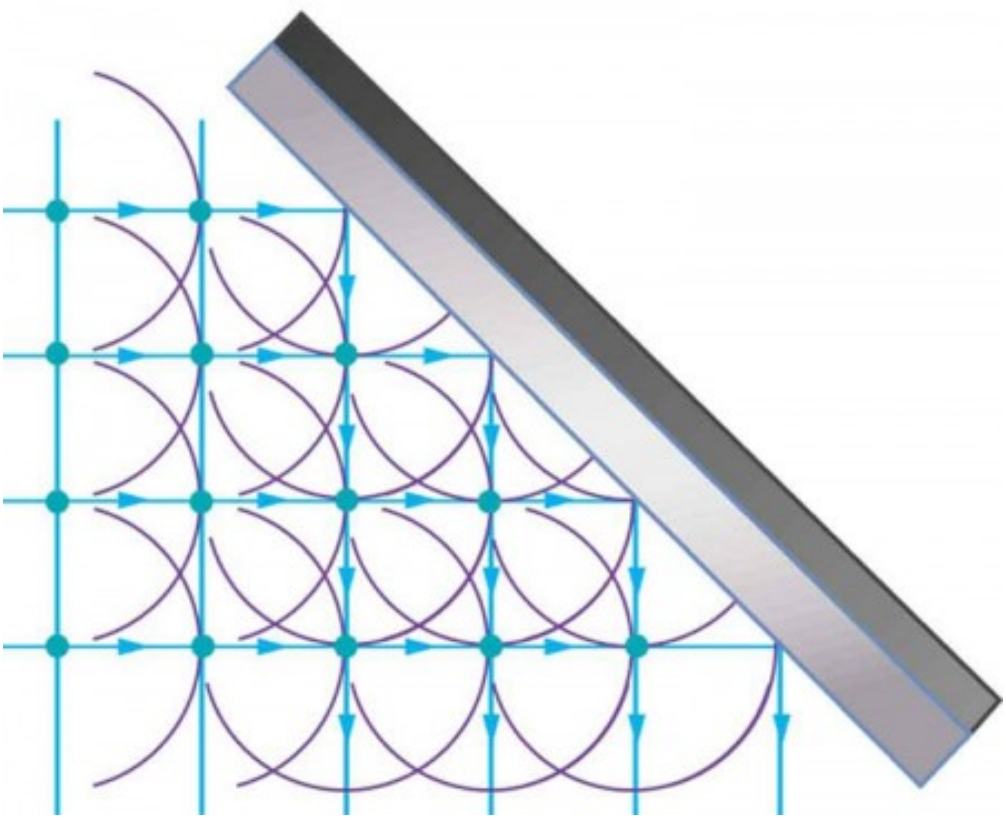


Figure 3. Huygens's principle applied to a straight wavefront striking a mirror. The wavelets shown were emitted as each point on the wavefront struck the mirror. The tangent to these wavelets shows that the new wavefront has been reflected at an angle equal to the incident angle. The direction of propagation is perpendicular to the wavefront, as shown by the downward-pointing arrows.

The law of refraction can be explained by applying Huygens's principle to a wavefront passing from one medium to another (see Figure 4). Each wavelet in the figure was emitted when the wavefront 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 wavefront 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 4, but this is left as an exercise for ambitious readers.

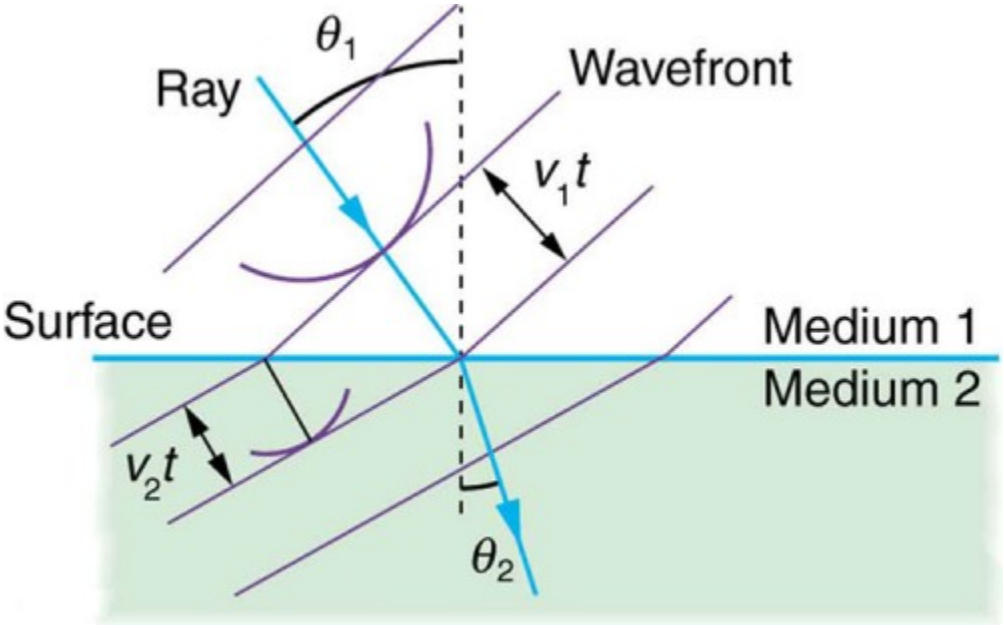


Figure 4. Huygens's principle applied to a straight wavefront traveling from one medium to another where its speed is less. The ray bends toward the perpendicular, since the wavelets have a lower speed in the second medium.

What happens when a wave passes through an opening, such as light shining through an open door into a dark room? For light, we expect to see a sharp shadow of the doorway on the floor of the room, and we expect no light to bend around corners into other parts of the room. When sound passes through a door, we expect to hear it everywhere in the room and, thus, expect that sound spreads out when passing through such an opening (see Figure 5). 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).

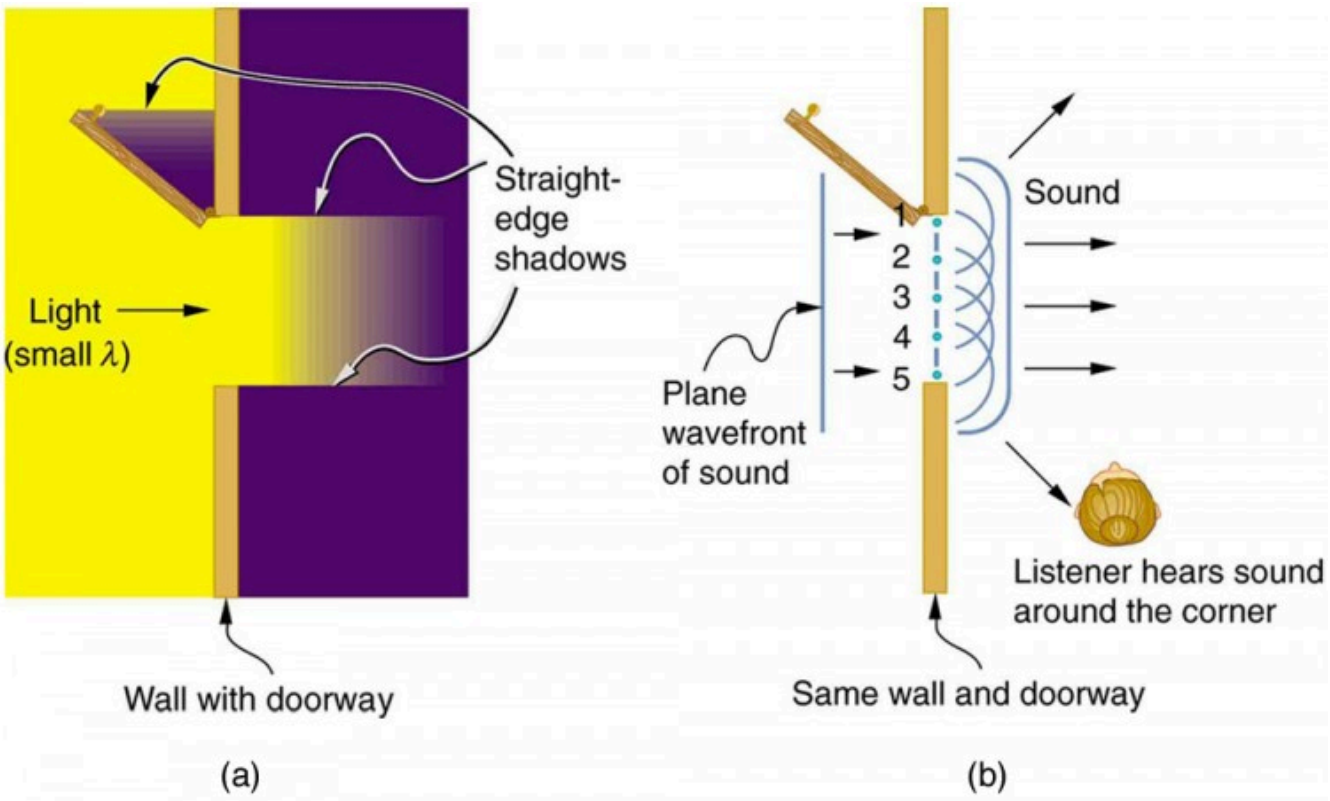


Figure 5. (a) Light passing through a doorway makes a sharp outline on the floor. Since light's wavelength is very small compared with the size of the door, it acts like a ray. (b) Sound waves bend into all parts of the room, a wave effect, because their wavelength is similar to the size of the door.

If we pass light through smaller openings, often called slits, we can use Huygens's principle to see that light bends as sound does (see Figure 6). 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 slits in Figure 7 is evidence that light is a wave.

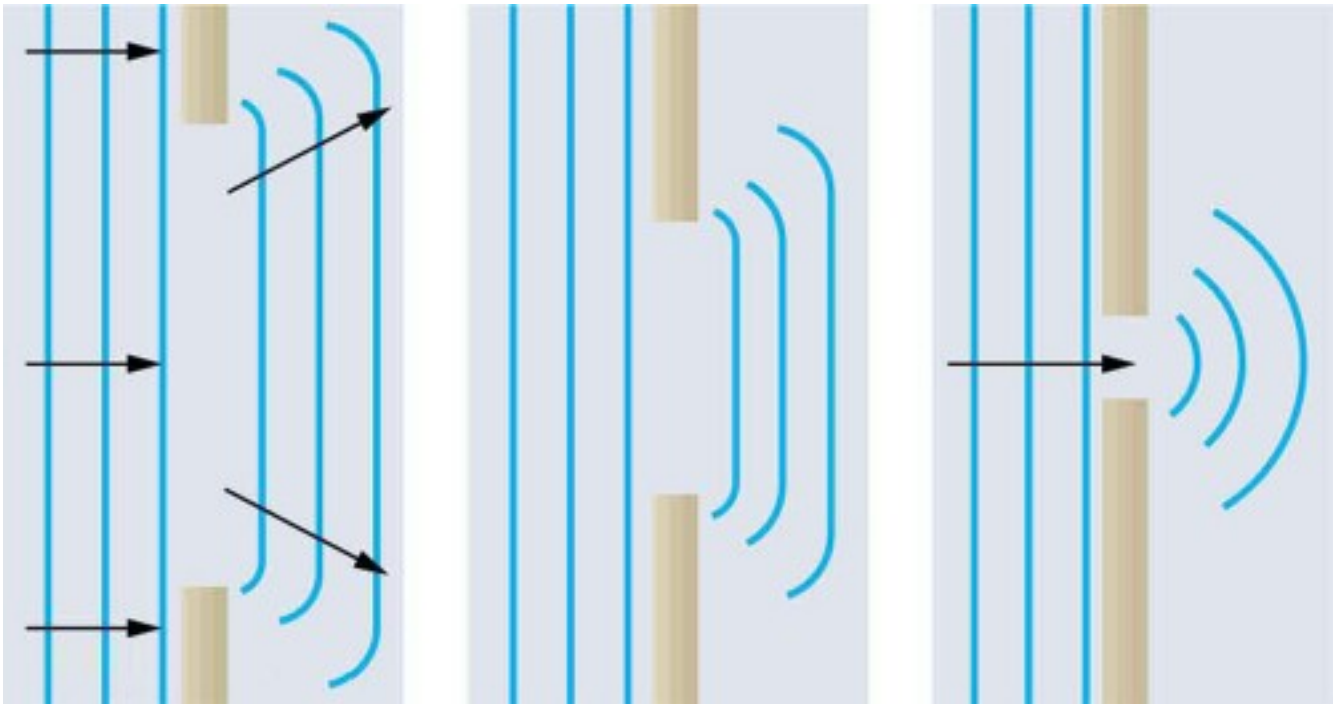


Figure 6. Huygens's principle applied to a straight wavefront striking an opening. The edges of the wavefront bend after passing through the opening, a process called diffraction. The amount of bending is more extreme for a small opening, consistent with the fact that wave characteristics are most noticeable for interactions with objects about the same size as the wavelength.

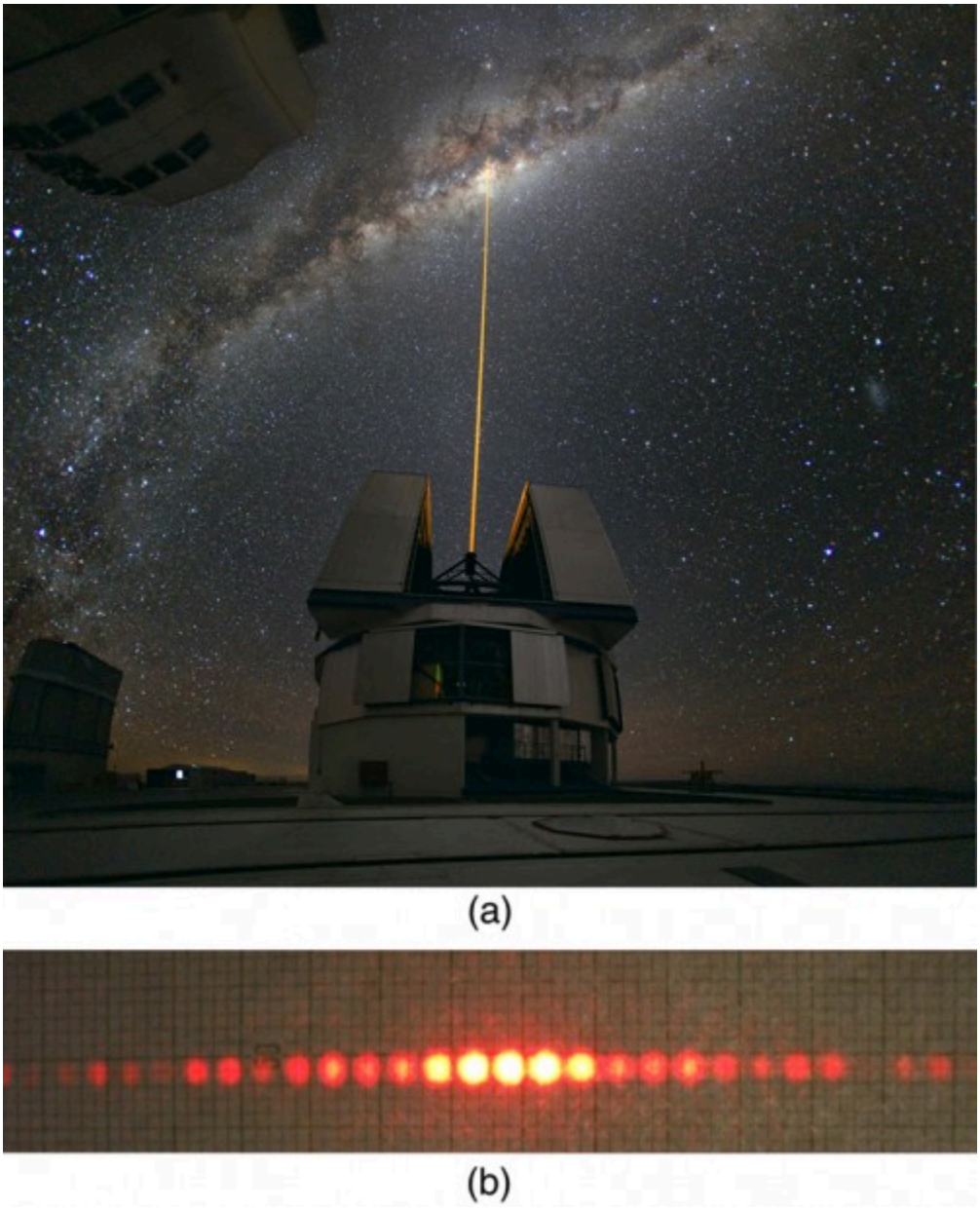


Figure 7. (a) The laser beam emitted by an observatory acts like a ray, traveling in a straight line. This laser beam is from the Paranal Observatory of the European Southern Observatory. (credit: Yuri Beletsky, European Southern Observatory) (b) A laser beam passing through a grid of vertical slits produces an interference pattern—characteristic of a wave. (credit: Shim'on and Slava Rybka, Wikimedia Commons)

Section Summary

- An accurate technique for determining how and where waves propagate is given by Huygens's principle: Every point on a wavefront is a source of wavelets that spread out in the forward direction at the same speed as the wave itself. The new wavefront is a line tangent to all of the wavelets.
- Diffraction is the bending of a wave around the edges of an opening or other obstacle.

Conceptual Questions

1. 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?
2. Under what conditions can light be modeled like a ray? Like a wave?
3. Go outside in the sunlight and observe your shadow. It has fuzzy edges even if you do not. Is this a diffraction effect? Explain.
4. Why does the wavelength of light decrease when it passes from vacuum into a medium? State which attributes change and which stay the same and, thus, require the wavelength to decrease.
5. Does Huygens's principle apply to all types of waves?

Glossary

diffraction: the bending of a wave around the edges of an opening or an obstacle

Huygens's principle: every point on a wavefront is a source of wavelets that spread out in the forward direction at the same speed as the wave itself. The new wavefront is a line tangent to all of the wavelets

Young's Double Slit Experiment

Lumen Learning

Learning Objectives

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

- Explain the phenomena of interference.
- Define constructive interference for a double slit and destructive interference for a double slit.

Although Christiaan Huygens thought that light was a wave, Isaac Newton did not. Newton felt 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 stature, his view generally prevailed. The fact that Huygens's principle worked was not considered evidence that was direct enough to prove that light is a wave. The acceptance of the wave character of light came many years later when, in 1801, the English physicist and physician Thomas Young (1773–1829) did his now-classic double slit experiment (see Figure 1).

Why do we not ordinarily observe wave behavior for light, such as observed in Young's double slit experiment? First, light must interact with something small, such as the closely spaced slits used by Young, to show pronounced wave effects. Furthermore, Young first passed light from a single source (the Sun) through a single slit to make the light somewhat coherent. By *coherent*, we mean waves are in phase or have a definite phase relationship. *Incoherent* means the waves have random phase relationships. Why did Young then pass the light through a double slit? The answer to this question is that two slits provide two coherent light sources that then interfere constructively or destructively. Young used sunlight, where each wavelength forms its own pattern, making the effect more difficult to see. We illustrate the double slit experiment with monochromatic (single λ) light to clarify the effect. Figure 2 shows the pure constructive and destructive interference of two waves having the same wavelength and amplitude.

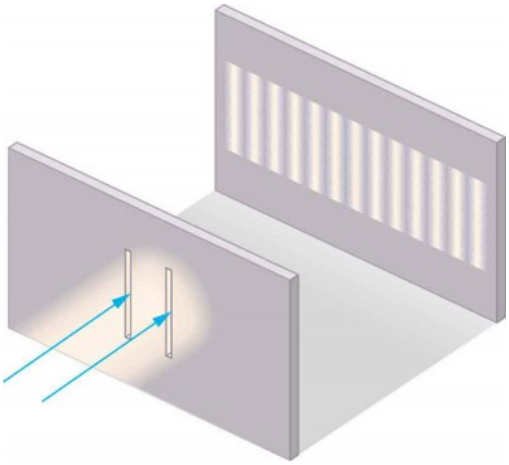


Figure 1. Young's double slit experiment. Here pure-wavelength light sent through a pair of vertical slits is diffracted into a pattern on the screen of numerous vertical lines spread out horizontally. Without diffraction and interference, the light would simply make two lines on the screen.

When light passes through narrow slits, it is diffracted into semicircular waves, as shown in Figure 3a. 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 3b. 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.

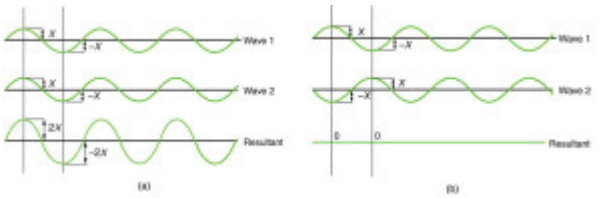


Figure 2. The amplitudes of waves add. (a) Pure constructive interference is obtained when identical waves are in phase. (b) Pure destructive interference occurs when identical waves are exactly out of phase, or shifted by half a wavelength.

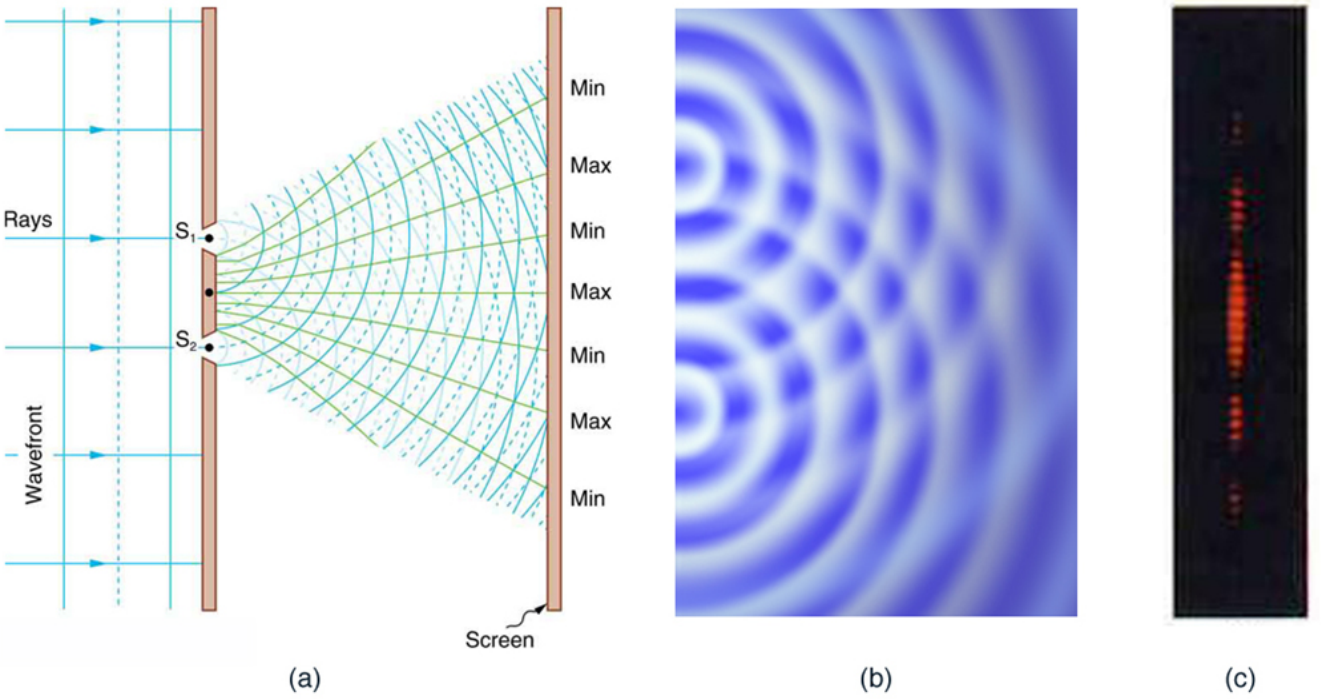


Figure 3. Double slits produce two coherent sources of waves that interfere. (a) Light spreads out (diffracts) from each slit, because the slits are narrow. These waves overlap and interfere constructively (bright lines) and destructively (dark regions). We can only see this if the light falls onto a screen and is scattered into our eyes. (b) Double slit interference pattern for water waves are nearly identical to that for light. Wave action is greatest in regions of constructive interference and least in regions of destructive interference. (c) When light that has passed through double slits falls on a screen, we see a pattern such as this. (credit: PASCO)

To understand the double slit interference pattern, we consider how two waves travel from the slits to the screen, as illustrated in Figure 4. 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 as shown in Figure 4a. If the paths differ by a whole wavelength, then the waves arrive in phase (crest to crest) at the screen, interfering constructively as shown in Figure 4b. More generally, if the paths taken by the two waves differ by any half-integral number of wavelengths [$(1/2)\lambda$, $(3/2)\lambda$, $(5/2)\lambda$, etc.], then destructive interference occurs. Similarly, if

the paths taken by the two waves differ by any integral number of wavelengths (λ, 2λ, 3λ, etc.), then constructive interference occurs.

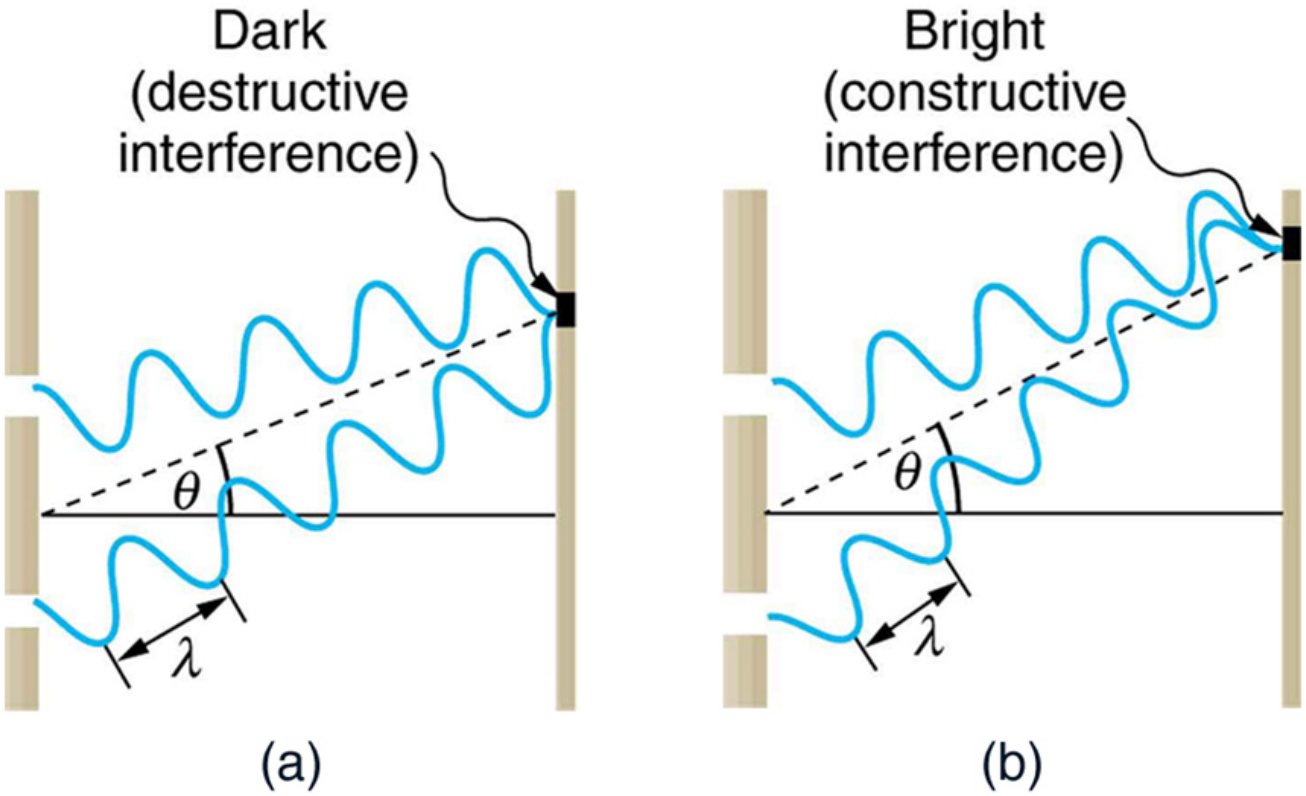


Figure 4. Waves follow different paths from the slits to a common point on a screen. (a) Destructive interference occurs here, because one path is a half wavelength longer than the other. The waves start in phase but arrive out of phase. (b) Constructive interference occurs here because one path is a whole wavelength longer than the other. The waves start out and arrive in phase.

Take-Home Experiment: Using Fingers as Slits

Look at a light, such as a street lamp or incandescent bulb, through the narrow gap between two fingers held close together. What type of pattern do you see? How does it change when you allow the fingers to move a little farther apart? Is it more distinct for a monochromatic source, such as the yellow light from a sodium vapor lamp, than for an incandescent bulb?

Figure 5 shows how to determine the path length difference 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 (see the figure) is nearly the same for each path. The difference between the paths is shown in the figure; simple trigonometry shows it to be $d \sin \theta$, where d is the distance between the slits. To obtain *constructive interference for a double slit*, the path length difference must be an integral multiple of the wavelength, or $d \sin \theta = m\lambda$, for $m = 0, 1, -1, 2, -2, \ldots$ (constructive).

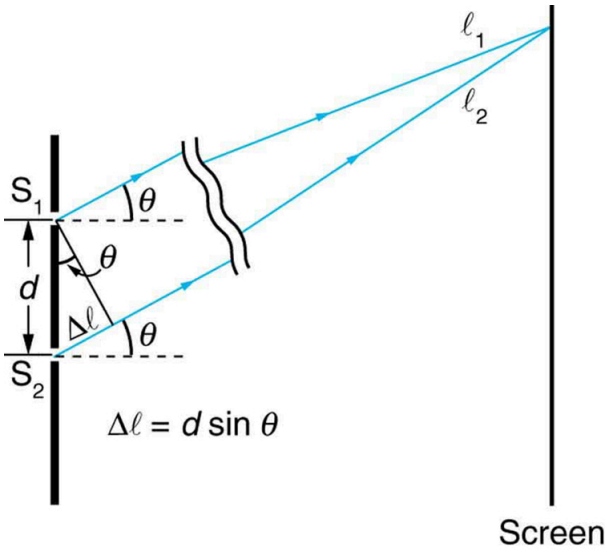


Figure 5. The paths from each slit to a common point on the screen differ by an amount dsinθ, assuming the distance to the screen is much greater than the distance between slits (not to scale here).

Similarly, to obtain *destructive interference for a double slit*, the path length difference must be a half-integral multiple of the wavelength, or

$$d \sin \theta = \left(m + \tfrac{1}{2}\right) \lambda, \text{ for } m = 0, 1, -1, 2, -2, \ldots \text{ (destructive)}$$

,

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.

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, illustrated in Figure 6. The intensity of the bright fringes falls off on either side, being brightest at the center. The closer the slits are, the more is the spreading of the bright fringes. We can see this by examining the equation $d \sin \theta = m\lambda$, for $m = 0, 1, -1, 2, -2, \ldots$.

For fixed λ and m, the smaller d is, the larger θ must be, since

$$\sin \theta = \frac{m\lambda}{d}$$

. 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.

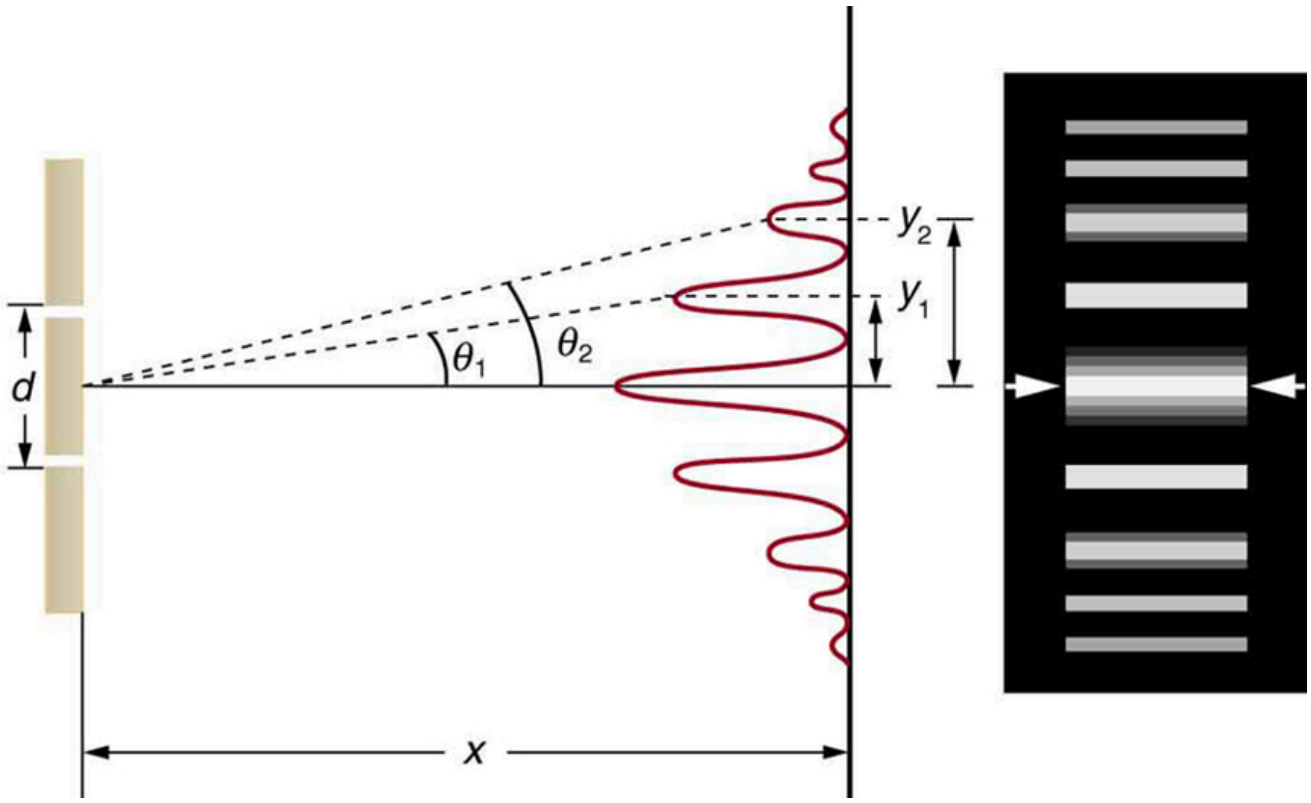


Figure 6. The interference pattern for a double slit has an intensity that falls off with angle. The photograph shows multiple bright and dark lines, or fringes, formed by light passing through a double slit.

Example 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 10.95º relative to the incident beam. What is the wavelength of the light?

Strategy

The third bright line is due to third-order constructive interference, which means that $m = 3$. We are given $d = 0.0100$ mm and $\theta = 10.95^\circ$. The wavelength can thus be found using the equation $d \sin \theta = m\lambda$ for constructive interference.

Solution

The equation is $d \sin \theta = m\lambda$. Solving for the wavelength λ gives

$$\lambda = \frac{d \sin \theta}{m}$$

.

Substituting known values yields

$$\begin{aligned} \lambda &= \frac{(0.0100 \text{ nm})(\sin 10.95^\circ)}{3} \\ &= 6.33 \times 10^{-4} \text{ nm} = 633 \text{ nm} \end{aligned}$$

Discussion

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 technique 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 2. Calculating 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 and Concept

The equation $d \sin \theta = m\lambda$ (for $m = 0, 1, -1, 2, -2, \ldots$) describes constructive interference. 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 90º. (Larger angles imply that light goes backward and does not reach the screen at all.) Let us find which m corresponds to this maximum diffraction angle.

Solution

Solving the equation $d \sin \theta = m\lambda$ for m gives

$$\lambda = \frac{d \sin \theta}{m}$$

.

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 $m = 15$.

Discussion

The number of fringes depends on the wavelength and slit separation. The number of fringes will be very large for large slit separations. However, if the slit separation becomes much greater than the wavelength, the intensity of the interference pattern changes so that the screen has two bright lines cast by the slits, as expected when light behaves like a ray. We also note that the fringes get fainter further away from the center. Consequently, not all 15 fringes may be observable.

Section Summary

- 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.

- There is constructive interference when $d \sin\theta = m\lambda$ (for $m = 0, 1, -1, 2, -2, \ldots$), 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.
- There is destructive interference when $d \sin\theta = m\lambda$ (for $m = 0, 1, -1, 2, -2, \ldots$).

Conceptual Questions

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. 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.
3. Is it possible to create a situation in which there is only destructive interference? Explain.
4. Figure 7 shows 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 slit and double slit interference. Note that the bright spots are evenly spaced. Is this a double slit or single slit characteristic? Note that some of the bright spots are dim on either side of the center. Is this a single slit or double slit characteristic? Which is smaller, the slit width or the separation between slits? Explain your responses.

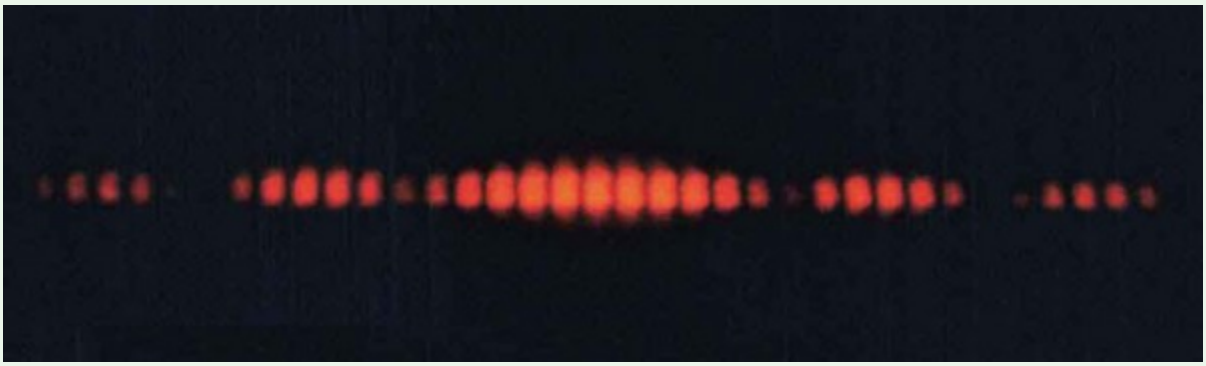


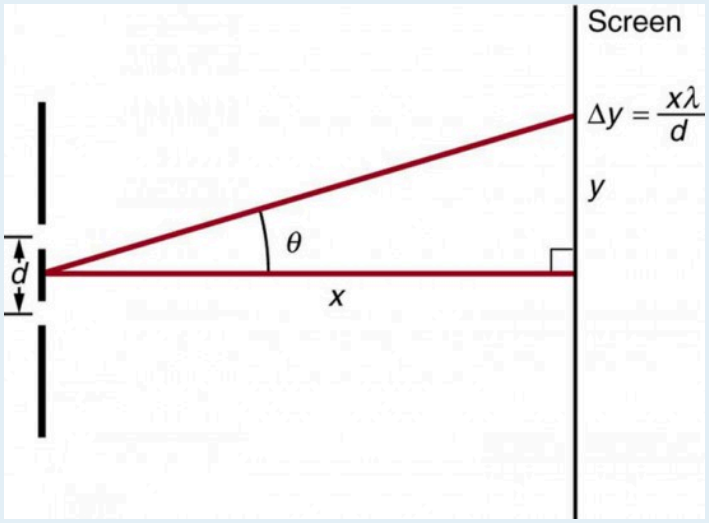
Figure 7. This double slit interference pattern also shows signs of single slit interference. (credit: PASCO)

Problems & Exercises

1. At what angle is the first-order maximum for 450-nm wavelength blue light falling on double slits separated by 0.0500 mm?
2. Calculate the angle for the third-order maximum of 580-nm wavelength yellow light falling on double slits separated by 0.100 mm.
3. What is the separation between two slits for which 610-nm orange light has its first maximum at an angle of 30.0º?
4. Find the distance between two slits that produces the first minimum for 410-nm violet light at an

angle of 45.0º.

5. Calculate the wavelength of light that has its third minimum at an angle of 30.0º when falling on double slits separated by 3.00 μm.
6. What is the wavelength of light falling on double slits separated by 2.00 μm if the third-order maximum is at an angle of 60.0º?
7. At what angle is the fourth-order maximum for the situation in Question 1?
8. What is the highest-order maximum for 400-nm light falling on double slits separated by 25.0 μm?
9. Find the largest wavelength of light falling on double slits separated by 1.20 μm for which there is a first-order maximum. Is this in the visible part of the spectrum?
10. What is the smallest separation between two slits that will produce a second-order maximum for 720-nm red light?
11. (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?
12. (a) If the first-order maximum for pure-wavelength light falling on a double slit is at an angle of 10.0º, 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?
13. Figure 8 shows 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, there will be numerous bright spots, called fringes. Show that, for small angles (where $\sin\theta \approx \theta$, with θ in radians), the distance between fringes is given by $\Delta y = \frac{x\lambda}{d}$.



$$\Delta y = \frac{x\lambda}{d}$$

Figure 8. The distance between adjacent fringes is $\Delta y = \frac{x\lambda}{d}$*, assuming the slit separation d is large compared with λ.*

14. Using the result of the problem above, 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 as in Figure 8.

15. 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 (see Figure 8).

Glossary

coherent: waves are in phase or have a definite phase relationship

constructive interference for a double slit: the path length difference must be an integral multiple of the wavelength

destructive interference for a double slit: the path length difference must be a half-integral multiple of the wavelength

incoherent: waves have random phase relationships

order: the integer m used in the equations for constructive and destructive interference for a double slit

Selected Solutions to Problems & Exercises

1. 0.516º

3. 1.22×10^{-6} m

5. 600 nm

7. 2.06º

9. 1200 nm (not visible)

11. (a) 760 nm; (b) 1520 nm

13. For small angles $\sin\theta - \tan\theta \approx \theta$ (in radians).

For two adjacent fringes we have, $d \sin\theta_m = m\lambda$ and $d \sin\theta_m + 1 = (m + 1)\lambda$

Subtracting these equations gives

```
*** QuickLaTeX cannot compile formula:
\begin{array}{}d\left(\sin{\theta }_{\text{m}+1}-\sin{\theta }_{\text{m}}\ri

*** Error message:
Missing # inserted in alignment preamble.
leading text: $\begin{array}{}
Missing $ inserted.
leading text: $\begin{array}{}d\left
Missing $ inserted.
leading text: ...\left[\left(m+1\right)-m\right]\lambda \\ d
Missing $ inserted.
```

```
leading text: ...[\left(m+1\right)-m\right]\lambda \\ d\left
Missing $ inserted.
leading text: ...\theta }_{\text{m}}\right)=\lambda \\ \text
Missing $ inserted.
leading text: ...ext{m}}\right)=\lambda \\ \text{tan}{\theta
Extra }, or forgotten $.
leading text: ...t{m}}\right)=\lambda \\ \text{tan}{\theta }
Missing } inserted.
leading text: ...frac{{y}_{\text{m}}}{x}\right)=\lambda \\ d
Extra }, or forgotten $.
leading text: ...frac{{y}_{\text{m}}}{x}\right)=\lambda \\ d
Missing } inserted.
leading text: ...frac{{y}_{\text{m}}}{x}\right)=\lambda \\ d
```

15. 450 nm

Multiple Slit Diffraction

Lumen Learning

Learning Objectives

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

- Discuss the pattern obtained from diffraction grating.
- Explain diffraction grating effects.

Figure 1. The colors reflected by this compact disc vary with angle and are not caused by pigments. Colors such as these are direct evidence of the wave character of light. (credit: Infopro, Wikimedia Commons)

An interesting thing happens if you pass light through a large number of evenly spaced parallel slits, called a *diffraction grating*. An interference pattern is created that is very similar to the one formed by a double slit (see Figure 2). A diffraction grating can be manufactured by scratching glass with a sharp tool in a number of precisely positioned parallel lines, with the untouched regions acting like slits. These can be photographically mass produced rather cheaply. Diffraction gratings work both for transmission of light, as in Figure 2, and for reflection of light, as on butterfly wings and the Australian opal in Figure 3 or the CD in Figure 1. In addition to their use as novelty items, 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 regions are narrower and brighter, while their dark regions are darker. Figure 4 shows idealized graphs demonstrating the sharper pattern. Natural diffraction gratings occur in the feathers of certain birds. 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.

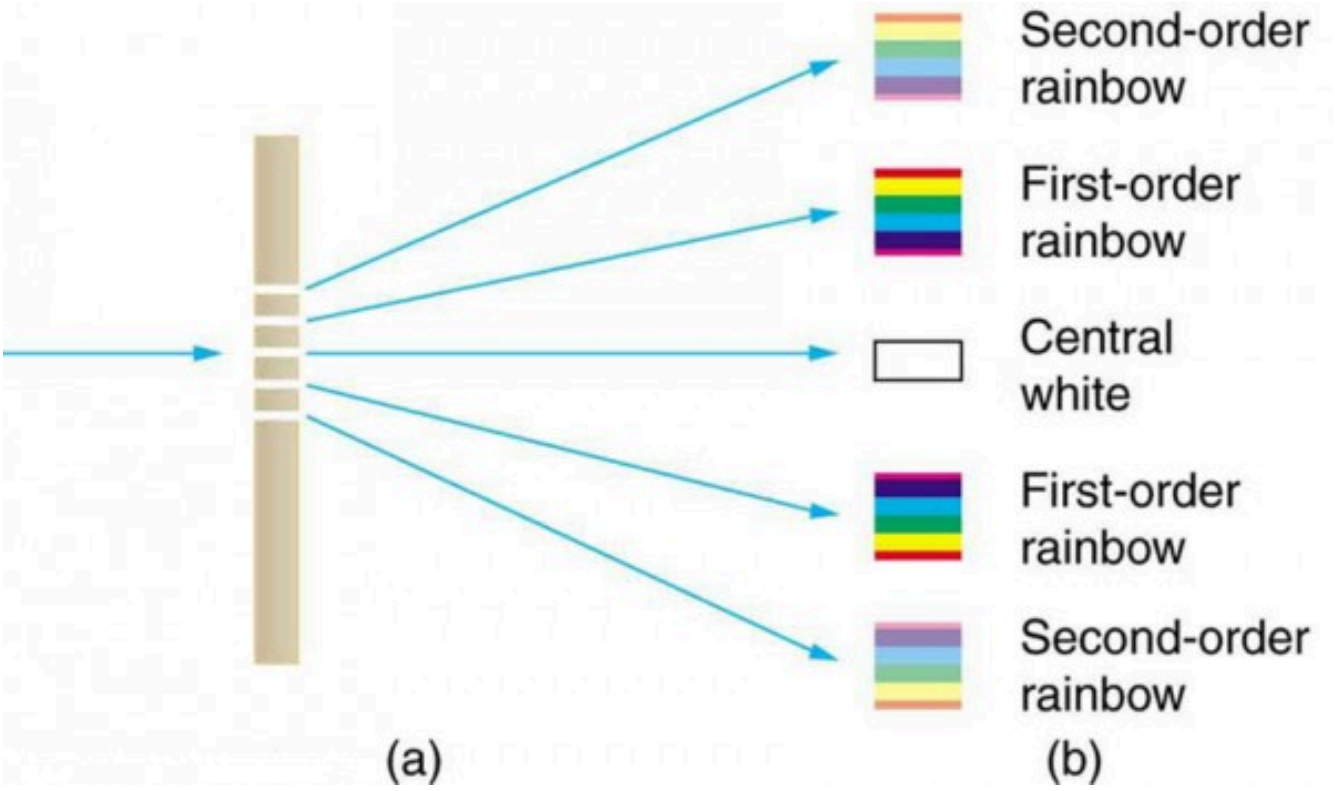


Figure 1. A diffraction grating is a large number of evenly spaced parallel slits. (a) Light passing through is diffracted in a pattern similar to a double slit, with bright regions at various angles. (b) The pattern obtained for white light incident on a grating. The central maximum is white, and the higher-order maxima disperse white light into a rainbow of colors.

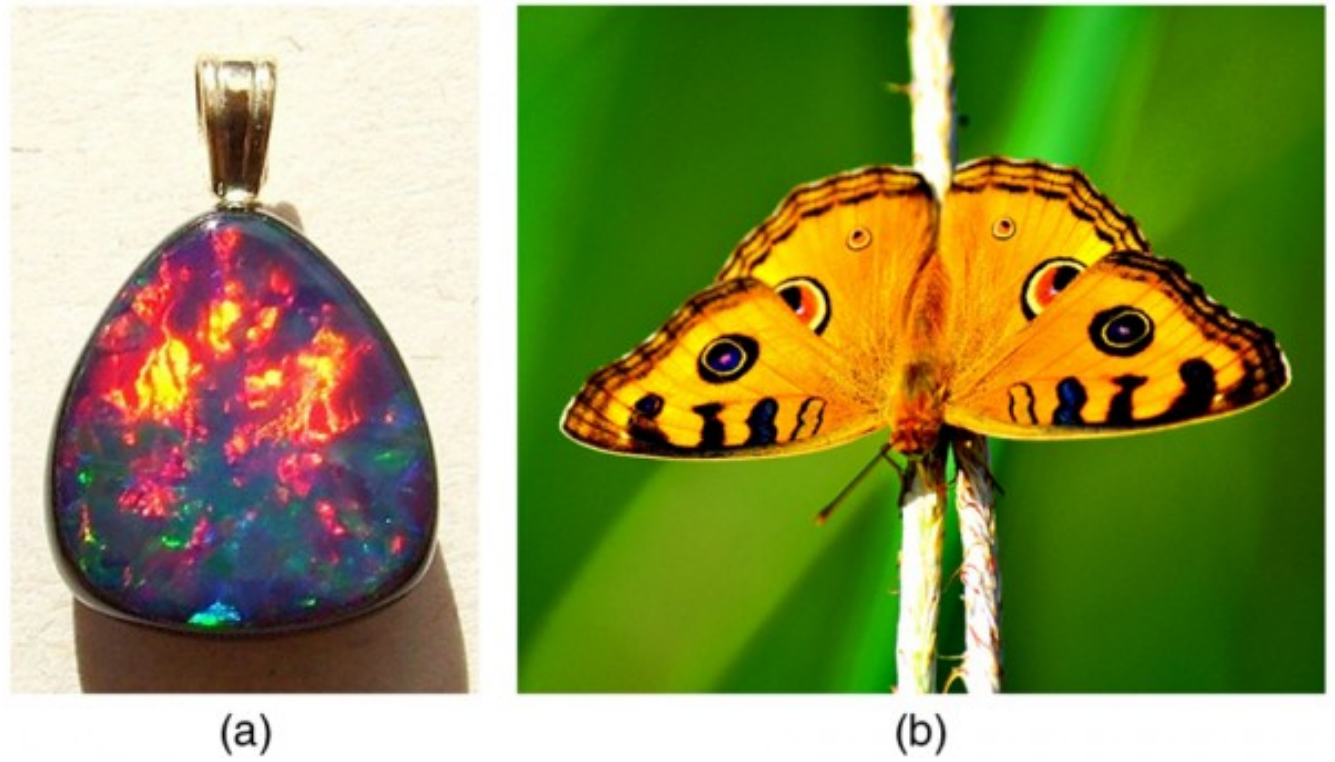


Figure 2. (a) This Australian opal and (b) the butterfly wings have rows of reflectors that act like reflection gratings, reflecting different colors at different angles. (credits: (a) Opals-On-Black.com, via Flickr (b) whologwhy, Flickr)

The analysis of a diffraction grating is very similar to that for a double slit (see Figure 5). As we know from our discussion of double slits in Young's Double Slit Experiment, light is diffracted by each slit and spreads out after passing through. Rays traveling in the same direction (at an angle θ relative to the incident direction) are shown in Figure 5. Each of these rays travels a different distance to a common point on a screen far away. The rays start in phase, and they can be in or out of phase when they reach a screen, depending on the difference in the path lengths traveled.

As seen in Figure 5, each ray travels a distance $d \sin\theta$ different from that of its neighbor, where d is the distance between slits. If this distance equals an integral number of wavelengths, the rays all arrive in phase, and constructive interference (a maximum) is obtained. Thus, the condition necessary to obtain *constructive interference for a diffraction grating* is $d \sin\theta = m\lambda$, for $m = 0, 1, -1, 2, -2, \ldots$ (constructive) where d is the distance between slits in the grating, λ is the wavelength of light, and m is the order of the maximum. Note that this is exactly the same equation as for double slits separated by d. However, the slits are usually closer in diffraction gratings than in double slits, producing fewer maxima at larger angles.

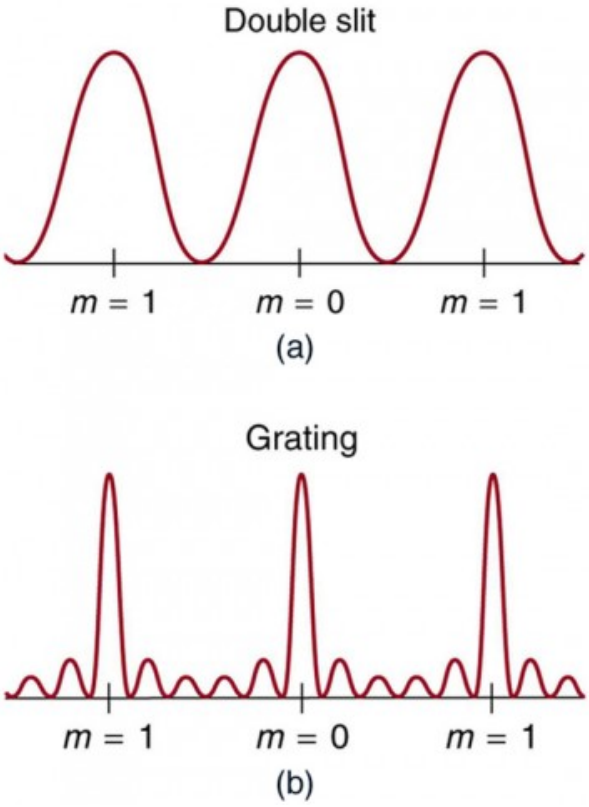


Figure 4. Idealized graphs of the intensity of light passing through a double slit (a) and a diffraction grating (b) for monochromatic light. Maxima can be produced at the same angles, but those for the diffraction grating are narrower and hence sharper. The maxima become narrower and the regions between darker as the number of slits is increased.

In Figure 5, we see a diffraction grating showing light rays from each slit traveling in the same direction. Each ray travels a different distance to reach a common point on a screen (not shown). Each ray travels a distance $d \sin\theta$ different from that of its neighbor.

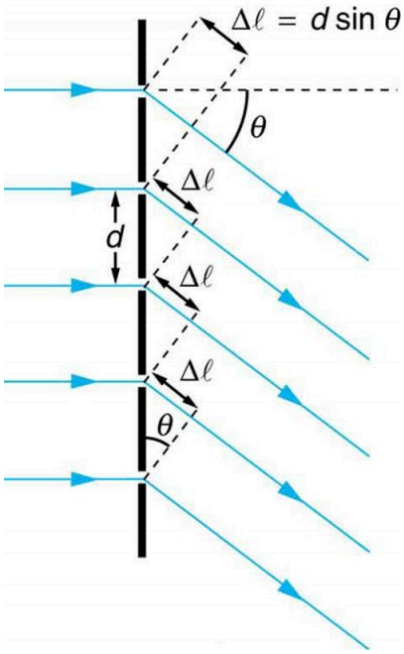


Figure 5.

Where are diffraction gratings used? 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 frequency 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 specific wavelengths for such use.

Take-Home Experiment: Rainbows on a CD

The spacing d of the grooves in a CD or DVD can be well determined by using a laser and the equation $d \sin \theta = m\lambda$, for $m = 0, 1, -1, 2, -2, \ldots$. However, we can still make a good estimate of this spacing by using white light and the rainbow of colors that comes from the interference. Reflect sunlight from a CD onto a wall and use your best judgment of the location of a strongly diffracted color to find the separation d.

Example 1. 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.

1. Find the angles for the first-order diffraction of the shortest and longest wavelengths of visible light (380 and 760 nm).
2. What is the distance between the ends of the rainbow of visible light produced on the screen for first-order interference? (See Figure 6.)

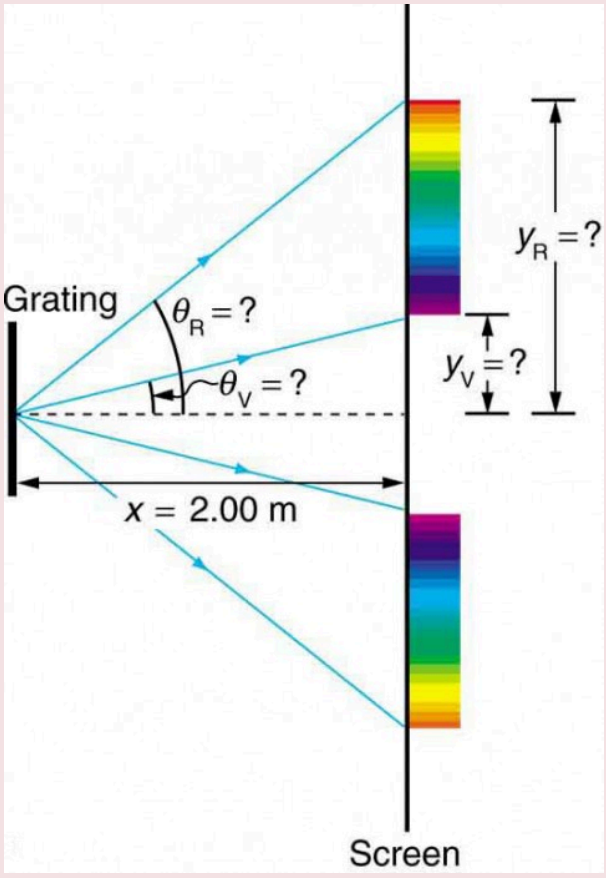


Figure 6. The diffraction grating considered in this example produces a rainbow of colors on a screen a distance from the grating. The distances along the screen are measured perpendicular to the x-direction. In other words, the rainbow pattern extends out of the page.

Strategy

The angles can be found using the equation $d \sin \theta = m\lambda$ (for $m = 0, 1, -1, 2, -2, \ldots$) once a value for the slit spacing d has been determined. Since there are 10,000 lines per centimeter, each line is separated by 1/10,000 of a centimeter. Once the angles are found, the distances along the screen can be found using simple trigonometry.

Solution for Part 1

The distance between slits is

$$d = \frac{1 \text{ cm}}{10{,}000} = 1.00 \times 10^{-4} \text{ cm}$$

or 1.00×10^{-6} 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 first order and $\lambda_V = 380$ nm $= 3.80 \times 10^{-7}$ m. Substituting these values gives

$$\sin\theta_{\mathrm{V}} = \frac{3.80 \times 10^{-7}\ \mathrm{m}}{1.00 \times 10^{-6}\ \mathrm{m}} = 0.380$$

Thus the angle θ_{V} is $\theta_{\mathrm{V}} = \sin^{-1} 0.380 = 22.33^{\circ}$.

Similarly,

$$\sin\theta_{\mathrm{R}} = \frac{7.60 \times 10^{-7}\ \mathrm{m}}{1.00 \times 10^{-6}\ \mathrm{m}}$$

Thus the angle θ_{R} is $\theta_{\mathrm{R}} = \sin^{-1} 0.760 = 49.46^{\circ}$.

Notice that in both equations, we reported the results of these intermediate calculations to four significant figures to use with the calculation in Part 2.

Solution for Part 2

The distances on the screen are labeled y_{V} and y_{R} in Figure 6. Noting that

$$\tan\theta = \frac{y}{x}$$

, we can solve for y_{V} and y_{R}. That is, $y_{\mathrm{V}} = x \tan\theta_{\mathrm{V}} = (2.00\ \mathrm{m})(\tan 22.33^{\circ}) = 0.815\ \mathrm{m}$ and $y_{\mathrm{R}} = x \tan\theta_{\mathrm{R}} = (2.00\ \mathrm{m})(\tan 49.46^{\circ}) = 2.338\ \mathrm{m}$.

The distance between them is therefore $y_{\mathrm{R}} - y_{\mathrm{V}} = 1.52\ \mathrm{m}$.

Discussion

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.

Section Summary

A diffraction grating is a large collection of evenly spaced parallel slits that produces an interference pattern similar to but sharper than that of a double slit.
There is constructive interference for a diffraction grating when d sin $\theta = m\lambda$ (for $m = 0$, 1, –1, 2, –2, …), where d is the distance between slits in the grating, λ is the wavelength of light, and m is the order of the maximum.

Conceptual Questions

1. What is the advantage of a diffraction grating over a double slit in dispersing light into a spectrum?
2. What are the advantages of a diffraction grating over a prism in dispersing light for spectral analysis?
3. Can the lines in a diffraction grating be too close together to be useful as a spectroscopic tool for

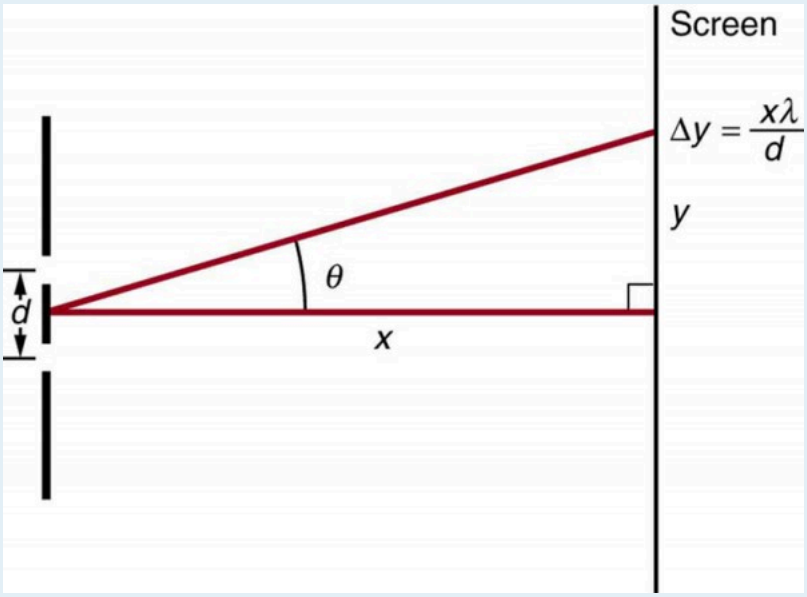
visible light? If so, what type of EM radiation would the grating be suitable for? Explain.

4. If a beam of white light passes through a diffraction grating with vertical lines, the light is dispersed into rainbow colors on the right and left. If a glass prism disperses white light to the right into a rainbow, how does the sequence of colors compare with that produced on the right by a diffraction grating?
5. Suppose pure-wavelength light falls on a diffraction grating. What happens to the interference pattern if the same light falls on a grating that has more lines per centimeter? What happens to the interference pattern if a longer-wavelength light falls on the same grating? Explain how these two effects are consistent in terms of the relationship of wavelength to the distance between slits.
6. Suppose a feather appears green but has no green pigment. Explain in terms of diffraction.
7. It is possible that there is no minimum in the interference pattern of a single slit. Explain why. Is the same true of double slits and diffraction gratings?

Problems & Exercises

1. A diffraction grating has 2000 lines per centimeter. At what angle will the first-order maximum be for 520-nm-wavelength green light?
2. Find the angle for the third-order maximum for 580-nm-wavelength yellow light falling on a diffraction grating having 1500 lines per centimeter.
3. 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 25.0º?
4. 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 60.0º?
5. Calculate the wavelength of light that has its second-order maximum at 45.0º when falling on a diffraction grating that has 5000 lines per centimeter.
6. 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 of 24.2º, 25.7º, 29.1º, and 41.0º when projected on a diffraction grating having 10,000 lines per centimeter?
7. (a) What do the four angles in the above 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.
8. What is the maximum number of lines per centimeter a diffraction grating can have and produce a complete first-order spectrum for visible light?
9. The yellow light from a sodium vapor lamp seems to be of pure wavelength, but it produces two first-order maxima at 36.093º and 36.129º when projected on a 10,000 line per centimeter diffraction grating. What are the two wavelengths to an accuracy of 0.1 nm?
10. What is the spacing between structures in a feather that acts as a reflection grating, given that they produce a first-order maximum for 525-nm light at a 30.0º angle?

11. Structures on a bird feather act like a reflection grating having 8000 lines per centimeter. What is the angle of the first-order maximum for 600-nm light?
12. An opal such as that shown in Figure 2 acts like a reflection grating with rows separated by about 8 μm. 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?
13. At what angle does a diffraction grating produces a second-order maximum for light having a first-order maximum at 20.0º?
14. Show that a diffraction grating cannot produce a second-order maximum for a given wavelength of light unless the first-order maximum is at an angle less than 30.0º.
15. If a diffraction grating produces a first-order maximum for the shortest wavelength of visible light at 30.0º, at what angle will the first-order maximum be for the longest wavelength of visible light?
16. (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?
17. (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 lines per centimeter a diffraction grating can have and produce a complete second-order spectrum for visible light?
18. A He–Ne laser beam is reflected from the surface of a CD onto a wall. The brightest spot is the reflected beam at an angle equal to the angle of incidence. However, fringes are also observed. If the wall is 1.50 m from the CD, and the first fringe is 0.600 m from the central maximum, what is the spacing of grooves on the CD?
19. The analysis shown in the figure 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?



$\Delta y = \frac{x\lambda}{d}$

Figure 6. The distance between adjacent fringes is , assuming the slit separation d is large compared with λ.

20. **Unreasonable Results.** 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?
21. **Unreasonable Results.** (a) What visible wavelength has its fourth-order maximum at an angle of 25.0º when projected on a 25,000-line-per-centimeter diffraction grating? (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?
22. **Construct Your Own Problem.** Consider a spectrometer based on a diffraction grating. Construct a problem in which you calculate the distance between two wavelengths of electromagnetic radiation in your spectrometer. Among the things to be considered are the wavelengths you wish to be able to distinguish, the number of lines per meter on the diffraction grating, and the distance from the grating to the screen or detector. Discuss the practicality of the device in terms of being able to discern between wavelengths of interest.

Glossary

constructive interference for a diffraction grating: occurs when the condition $d \sin \theta = m\lambda$ (form = 0,1,–1,2,–2, . . .) is satisfied, where d is the distance between slits in the grating, λ is the wavelength of light, and m is the order of the maximum

diffraction grating: a large number of evenly spaced parallel slits

Selected Solution to Problems & Exercises

1. 5.97º

3. 8.99×10^3

5. 707 nm

7. (a) 11.8º,12.5º,14.1º,19.2º; (b) 24.2º,25.7º,29.1º,41.0º; (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.

9. 589.1 nm and 589.6 nm

11. 28.7º

13. 43.2º

15. 90.0º

17. (a) The longest wavelength is 333.3 nm, which is not visible; (b) 333 nm (UV); (c) 6.58×10^3 cm

19. 1.13×10^{-2} m

21. (a) 42.3 nm; (b) Not a visible wavelength. The number of slits in this diffraction grating is too large. Etching in integrated circuits can be done to a resolution of 50 nm, so slit separations of 400 nm are at the limit of what we can do today. This line spacing is too small to produce diffraction of light.

Single Slit Diffraction

Lumen Learning

Learning Objectives

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

- Discuss the single slit diffraction pattern.

Light passing through a single slit forms a diffraction pattern somewhat different from those formed by double slits or diffraction gratings. Figure 1 shows a single slit diffraction pattern. Note that the central maximum is larger than those on either side, and that the intensity decreases rapidly on either side. In contrast, a diffraction grating produces evenly spaced lines that dim slowly on either side of center.

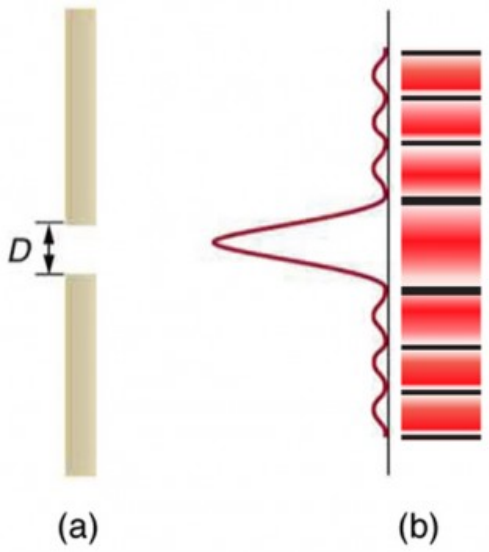


Figure 1. (a) Single slit diffraction pattern. Monochromatic light passing through a single slit has a central maximum and many smaller and dimmer maxima on either side. The central maximum is six times higher than shown. (b) The drawing shows the bright central maximum and dimmer and thinner maxima on either side.

The analysis of single slit diffraction is illustrated in Figure 2. Here we consider light coming from different parts of the *same* slit. According to Huygens's principle, every part of the wavefront in the slit emits wavelets. These are like rays that start out in phase and head in all directions. (Each ray is perpendicular to the wavefront 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 Figure 2a, they remain in phase, and a central maximum is obtained. However, when rays travel at an angle θ relative to the original direction of the beam, each travels a different distance to a common location, and they can arrive in or out of phase. In Figure 2b, 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 $\lambda/2$ farther than the one on the left, arrives out of phase, and interferes destructively. A ray from slightly above the center and one from slightly above the bottom will also cancel one another. In fact, each ray from the slit will have another to interfere destructively, and a minimum in intensity will occur at this angle. There will be another minimum at the same angle to the right of the incident direction of the light.

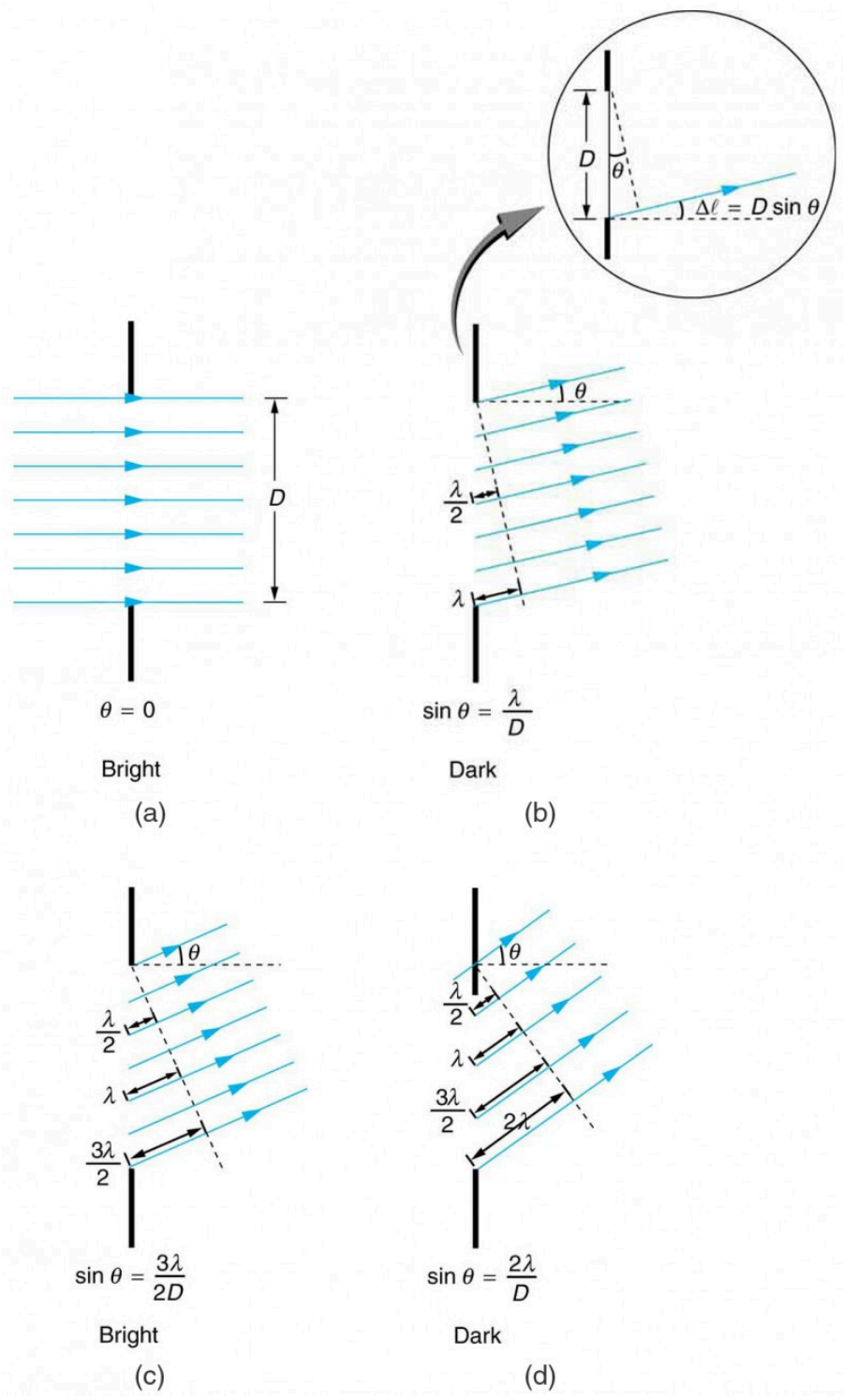


Figure 2.

In Figure 2 we see that 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 Figure 2c, 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, will also add constructively. Most rays from the slit will have another to interfere with constructively, and a maximum in intensity will occur at this angle. However, all rays do not interfere constructively for this situation, and so the maximum is not as intense as the central maximum. Finally, in Figure 2d, 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.

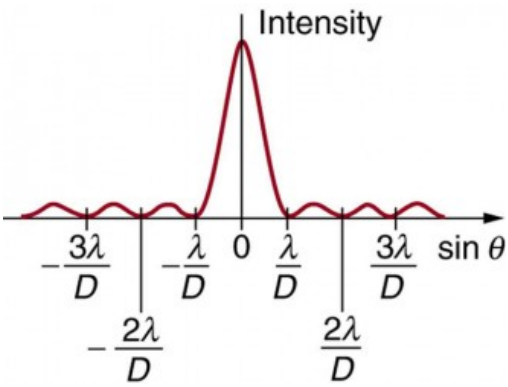


Figure 3. A graph of single slit diffraction intensity showing the central maximum to be wider and much more intense than those to the sides. In fact the central maximum is six times higher than shown here.

Thus, to obtain *destructive interference for a single slit*, $D \sin \theta = m\lambda$, for $m = 1, -1, 2, -2, 3, \ldots$ (destructive), 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 3 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 is consistent with the illustration in Figure 1b.

Example 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 45.0º relative to the incident direction of the light.

1. What is the width of the slit?
2. At what angle is the first minimum produced?

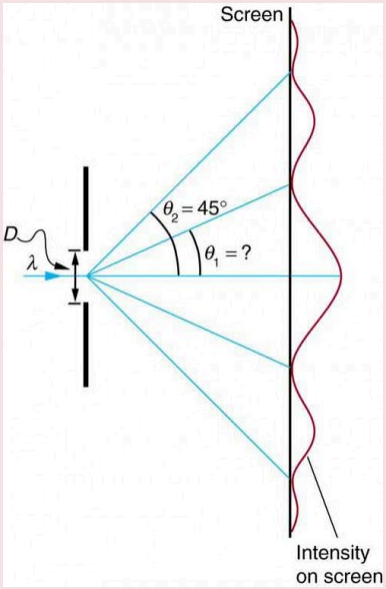


Figure 4.

A graph of the single slit diffraction pattern is analyzed in this example.

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 for Part 1

We are given that $\lambda = 550$ nm, $m = 2$, and $\theta_2 = 45.0^\circ$. Solving the equation $D \sin\theta = m\lambda$ for D and substituting known values gives

$$\begin{aligned} D &= \frac{m\lambda}{\sin\theta_2} = \frac{2(550\text{ nm})}{\sin 45.0^\circ} \\ &= \frac{1100\times10^{-9}}{0.707} \\ &= 1.56\times10^{-6} \end{aligned}$$

Solution for Part 2

Solving the equation $D \sin\theta = m\lambda$ for $\sin\theta_1$ and substituting the known values gives

$$\sin\theta_1 = \frac{m\lambda}{D} = \frac{1\left(550\times10^{-9}\text{ m}\right)}{1.56\times10^{-6}\text{ m}}$$

Thus the angle θ_1 is $\theta_1 = \sin^{-1} 0.354 = 20.7^\circ$.

Discussion

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 20.7º on either side of the original beam, for a width of about 41º. The angle between the first and

second minima is only about 24º(45.0º − 20.7º). Thus the second maximum is only about half as wide as the central maximum.

Section Summary

- A single slit produces an interference pattern characterized by a broad central maximum with narrower and dimmer maxima to the sides.
- There is destructive interference for a single slit when $D \sin \theta = m\lambda$, (form = 1,–1,2,–2,3, . . .), 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. Note that there is no $m = 0$ minimum.

Conceptual Questions

1. As the width of the slit producing a single-slit diffraction pattern is reduced, how will the diffraction pattern produced change?

Problems & Exercises

1. (a) At what angle is the first minimum for 550-nm light falling on a single slit of width 1.00 μm? (b) Will there be a second minimum?
2. (a) Calculate the angle at which a 2.00-μm-wide slit produces its first minimum for 410-nm violet light. (b) Where is the first minimum for 700-nm red light?
3. (a) How wide is a single slit that produces its first minimum for 633-nm light at an angle of 28.0º? (b) At what angle will the second minimum be?
4. (a) What is the width of a single slit that produces its first minimum at 60.0º for 600-nm light? (b) Find the wavelength of light that has its first minimum at 62.0º.
5. Find the wavelength of light that has its third minimum at an angle of 48.6º when it falls on a single slit of width 3.00 μm.
6. Calculate the wavelength of light that produces its first minimum at an angle of 36.9º when falling on a single slit of width 1.00 μm.
7. (a) Sodium vapor light averaging 589 nm in wavelength falls on a single slit of width 7.50 μm. At what angle does it produces its second minimum? (b) What is the highest-order minimum produced?
8. (a) Find the angle of the third diffraction minimum for 633-nm light falling on a slit of width 20.0 μm. (b) What slit width would place this minimum at 85.0º?
9. (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 μm. (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.

10. (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?
11. (a) If a single slit produces a first minimum at 14.5º, 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).
12. 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.)
13. **Integrated Concepts.** 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 angle to the incident direction are the boats inside the harbor most protected against wave action?
14. **Integrated Concepts.** 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?

Glossary

destructive interference for a single slit: occurs when $D \sin\theta = m\lambda$, (form=1,–1,2,–2,3, . . .), 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

Selected Solutions to Problems & Exercises

1. (a) 33.4º; (b) No

3. (a) 1.35×10^{-6} m; (b) 69.9º

5. 750 nm

7. (a) 9.04º; (b) 12

9. (a) 0.0150º; (b) 0.262 mm; (c) This distance is not easily measured by human eye, but under a microscope or magnifying glass it is quite easily measurable.

11. (a) 30.1º; (b) 48.7º; (c) No; (d) $2\theta_1 = (2)(14.5^\circ) = 29^\circ$, $\theta_2 - \theta_1 = 30.05^\circ - 14.5^\circ = 15.56^\circ$. Thus, $29^\circ \approx (2)(15.56^\circ) = 31.1^\circ$.

13. 23.6º and 53.1º

Limits of Resolution: The Rayleigh Criterion

Lumen Learning

Learning Objectives

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

- Discuss the Rayleigh criterion.

Light diffracts as it moves through space, bending around obstacles, interfering constructively and destructively. While this can be used as a spectroscopic tool—a diffraction grating disperses light according to wavelength, for example, and is used to produce spectra—diffraction also limits the detail we can obtain in images. Figure 1a shows the effect of passing light through a small circular aperture. Instead of a bright spot with sharp edges, a spot with a fuzzy edge surrounded by circles of light is obtained. 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, too.

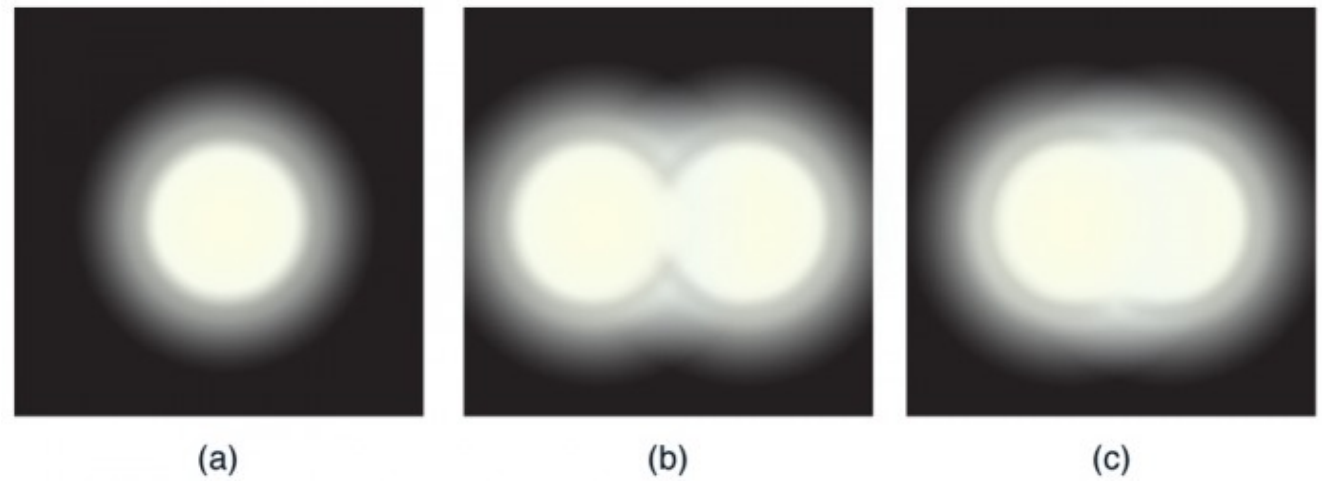


Figure 1. (a) Monochromatic light passed through a small circular aperture produces this diffraction pattern. (b) Two point light sources that are close to one another produce overlapping images because of diffraction. (c) If they are closer together, they cannot be resolved or distinguished.

How does diffraction affect the detail that can be observed when light passes through an aperture? Figure 1b 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 just barely possible to tell that there are two light sources rather than one. If they were closer together, as in Figure 1c, we could not distinguish them, thus limiting the detail or resolution we can obtain. This limit is an inescapable consequence of the wave nature of light.

There are many situations in which diffraction limits the resolution. The acuity of our vision is limited because light passes through the pupil, the circular aperture of our 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. So 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 their primary mirror.

Take-Home Experiment: Resolution of the Eye

Draw two lines on a white sheet of paper (several mm apart). How far away can you be and still distinguish the two lines? What does this tell you about the size of the eye's pupil? Can you be quantitative? (The size of an adult's pupil is discussed in Physics of the Eye.)

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) (see Figure 2a). It can be shown that, for a circular aperture of diameter *D*, the first minimum in the diffraction pattern occurs at

$$\theta = 1.22\frac{\lambda}{D}$$

(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 was developed by Lord Rayleigh in the 19th century. The *Rayleigh criterion* for 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*. See Figure 2b. The first minimum is at an angle of

$$\theta = 1.22\frac{\lambda}{D}$$

, so that two point objects are just resolvable if they are separated by the angle

$$\theta = 1.22\frac{\lambda}{D}$$

,

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.

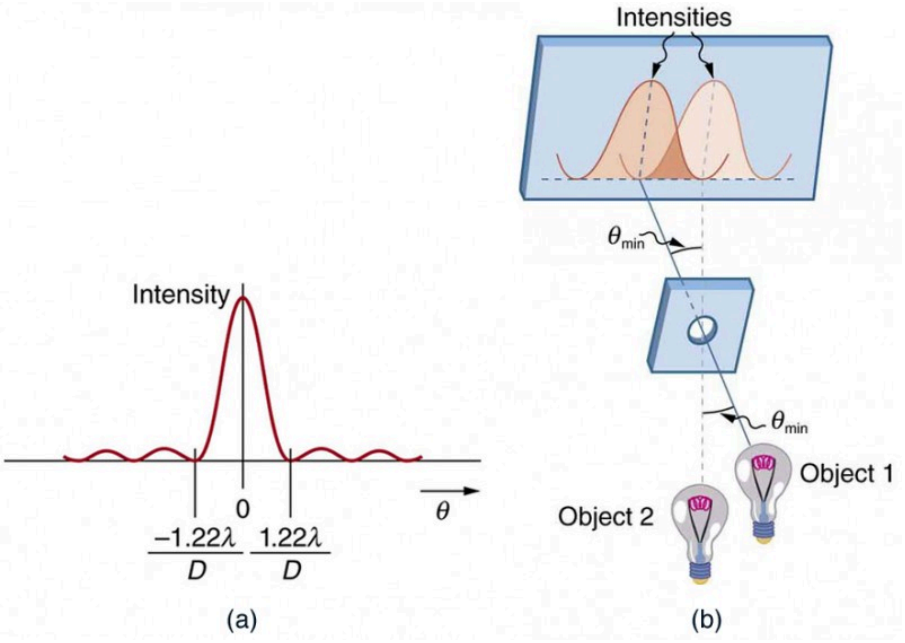


Figure 2. (a) Graph of intensity of the diffraction pattern for a circular aperture. Note that, similar to a single slit, the central maximum is wider and brighter than those to the sides. (b) Two point objects produce overlapping diffraction patterns. Shown here is the Rayleigh criterion for being just resolvable. The central maximum of one pattern lies on the first minimum of the other.

Making Connections: Limits to Knowledge

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 as with an electron microscope are used, the system is disturbed, still limiting our knowledge, much as making an electrical measurement alters a circuit. Heisenberg's uncertainty principle asserts that this limit is fundamental and inescapable, as we shall see in quantum mechanics.

Example 1. 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.

1. What is the angle between two just-resolvable point light sources (perhaps two stars)? Assume an average light wavelength of 550 nm.
2. If these two stars are at the 2 million light year 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 the equation

$$\theta = 1.22\frac{\lambda}{D}$$

gives the smallest possible angle θ between point sources, or the best obtainable resolution. Once this angle is found, the distance between stars can be calculated, since we are given how far away they are.

Solution for Part 1

The Rayleigh criterion for the minimum resolvable angle is

$$\theta = 1.22\frac{\lambda}{D}$$

.

Entering known values gives

$$\begin{aligned} \theta &= 1.22\frac{550\times10^{-9}\ \text{m}}{2.40\ \text{m}} \\ &= 2.80 \times 10^{-7}\ \text{rad} \end{aligned}$$

Solution for Part 2

The distance s between two objects a distance r away and separated by an angle θ is $s = r\theta$.

Substituting known values gives

$$\begin{aligned} s &= \left(2.0 \times 10^{6}\ \text{ly}\right)\left(2.80 \times 10^{-7}\ \text{rad}\right) \\ &= 0.56\ \text{ly} \end{aligned}$$

Discussion

The angle found in Part 1 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 non-uniformities in mirrors or aberrations in lenses that further limit resolution. However, Figure 3 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 the Earth's atmosphere.

Figure 3. These two photographs of the M82 galaxy give an idea of the observable detail using the Hubble Space Telescope compared with that using a ground-based telescope. (a) On the left is a ground-based image. (credit: Ricnun, Wikimedia Commons) (b) The photo on the right was captured by Hubble. (credit: NASA, ESA, and the Hubble Heritage Team (STScI/ AURA))

The answer in Part 2 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 5 light years in the outer parts and about 1 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 for its light to reach us.
Figure 4 shows another mirror used to observe radio waves from outer space.

Figure 4. A 305-m-diameter natural bowl 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 much longer wavelength radiation and its diffraction limit is significantly poorer than Hubble's. Arecibo is still very useful, because important information is carried by radio waves that is not carried by visible light. (credit: Tatyana Temirbulatova, Flickr)

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 the equation

$$\theta = 1.22\frac{\lambda}{D}$$

. Take, for example, a laser beam made of rays as parallel as possible (angles between rays as close to *θ* = 0º as possible) instead spreads out at an angle

$$\theta = 1.22\frac{\lambda}{D}$$

, 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 (see Figure 5). To avoid this, we can increase *D*. This is done for laser light sent to the Moon to measure its distance from the Earth. The laser beam is expanded through a telescope to make *D* much larger and *θ* smaller.

In Figure 5 we see that the beam produced by this microwave transmission antenna will spread out at a minimum angle

$$\theta = 1.22\frac{\lambda}{D}$$

due to diffraction. It is impossible to produce a near-parallel beam, because the beam has a limited diameter.

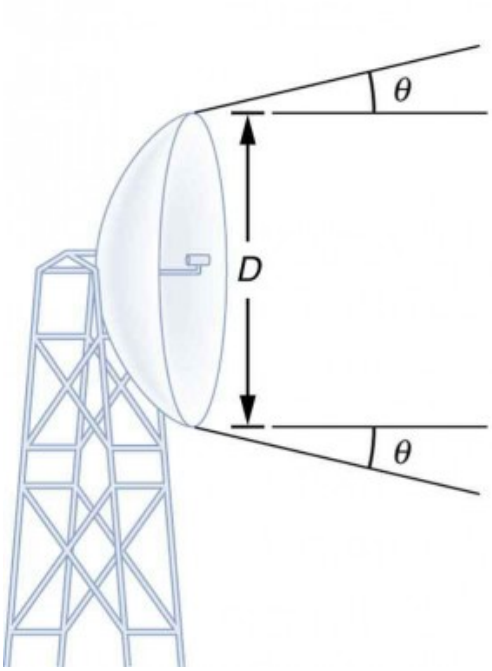


Figure 5.

In most biology laboratories, resolution is presented when the use of the microscope is introduced. The ability of a lens to produce sharp images of two closely spaced point objects is called resolution. 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. In Figure 6a we have 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 *θ* ≈ sin *θ* ≈ *θ*.

Therefore, the resolving power is

$$x = 1.22\frac{\lambda d}{D}$$

Another way to look at this is by re-examining the concept of Numerical Aperture (*NA*) discussed in Microscopes. There, *NA* is a measure of the maximum acceptance angle at which the fiber will take light and still contain it within the fiber. Figure 6b 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

$$\sin\alpha = \frac{\frac{D}{2}}{d} = \frac{D}{2d}$$

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* will be able to resolve finer details. Lenses with larger *NA* will also be 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, and so more of the diffraction modes will be collected. Thus the microscope has more information to form a clear image, and so its resolving power will be higher.

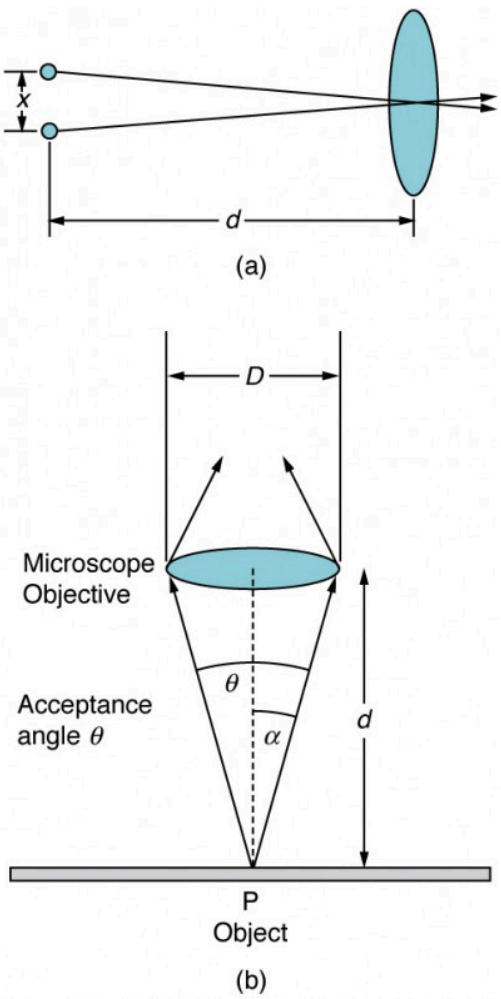


Figure 6. (a) Two points separated by at distance x and a positioned a distance d away from the objective. (credit: Infopro, Wikimedia Commons) (b) Terms and symbols used in discussion of resolving power for a lens and an object at point P. (credit: Infopro, Wikimedia Commons)

One of the consequences of diffraction is that the focal point of a beam has a finite width and intensity distribution. Consider focusing when only considering geometric optics, shown in Figure 7a. The focal point is infinitely small with a huge intensity and the capacity to incinerate most samples irrespective of the *NA* of the objective lens. For wave optics, due to diffraction, the focal point spreads to become a focal spot (see Figure 7b) 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.

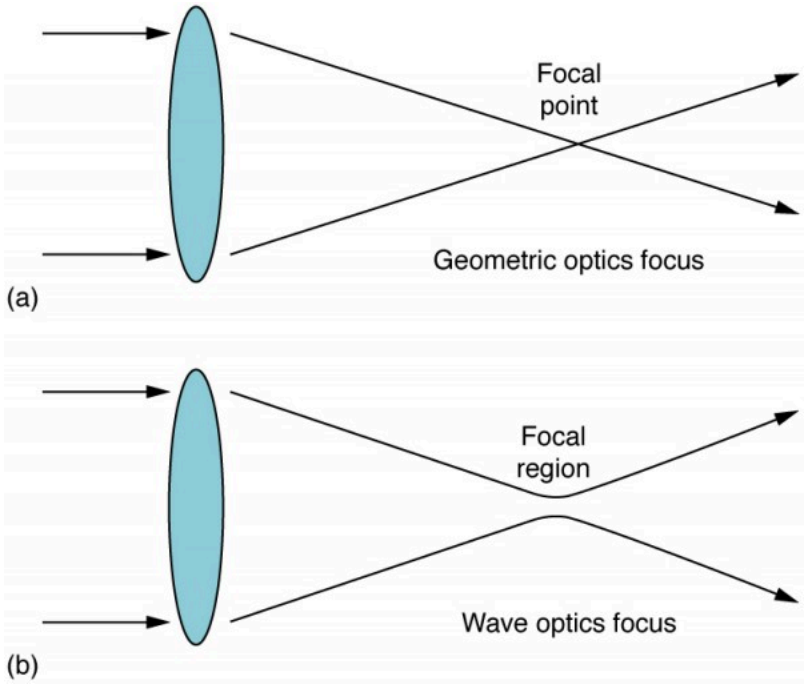


Figure 7. (a) In geometric optics, the focus is a point, but it is not physically possible to produce such a point because it implies infinite intensity. (b) In wave optics, the focus is an extended region.

Section Summary

- Diffraction limits resolution.
- For a circular aperture, lens, or mirror, 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.

$$\theta = 1.22\frac{\lambda}{D}$$

- This occurs for two point objects separated by the angle , where λ is the wavelength of light (or other electromagnetic radiation) and D is the diameter of the aperture, lens, mirror, etc. This equation also gives the angular spreading of a source of light having a diameter D.

Conceptual Questions

1. 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?

Problems & Exercises

1. The 300-m-diameter Arecibo radio telescope pictured in Figure 4 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?
2. Assuming the angular resolution found for the Hubble Telescope in Example 1, what is the smallest detail that could be observed on the Moon?
3. 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.
4. (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.)
5. 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?
6. 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?
7. 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.
8. You are told not to shoot until you see the whites of their eyes. If the eyes are separated by 6.5 cm and the diameter of your pupil is 5.0 mm, at what distance can you resolve the two eyes using light of wavelength 555 nm?
9. (a) The planet Pluto and its Moon Charon are separated by 19,600 km. Neglecting atmospheric effects, should the 5.08-m-diameter Mount Palomar telescope be able to resolve these bodies when they are 4.50×10^9 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 an Earth-based telescope. What are the reasons for this?
10. 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.
11. 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?

12. **Unreasonable Results.** An amateur astronomer wants to build a telescope with a diffraction limit that will allow him to see if there are people on the moons of Jupiter. (a) What diameter mirror is needed to be able to see 1.00 m detail on a Jovian Moon at a distance of 7.50×10^8 km from Earth? The wavelength of light averages 600 nm. (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?
13. **Construct Your Own Problem.** Consider diffraction limits for an electromagnetic wave interacting with a circular object. Construct a problem in which you calculate the limit of angular resolution with a device, using this circular object (such as a lens, mirror, or antenna) to make observations. Also calculate the limit to spatial resolution (such as the size of features observable on the Moon) for observations at a specific distance from the device. Among the things to be considered are the wavelength of electromagnetic radiation used, the size of the circular object, and the distance to the system or phenomenon being observed.

Glossary

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

Selected Solutions to Problems & Exercises

1. (a) 1.63 × 10−4 rad; (b) 326 ly

3. 1.46×10^{-5} rad

5. (a) 3.04×10^{-7} rad; (b) Diameter of 235 m

7. 5.15 cm

9. (a) Yes. Should easily be able to discern; (b) The fact that it is just barely possible to discern that these are separate bodies indicates the severity of atmospheric aberrations.

Thin Film Interference

Lumen Learning

Learning Objectives

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

- Discuss the rainbow formation by 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 noticed 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 since 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 1.

Figure 1. These soap bubbles exhibit brilliant colors when exposed to sunlight. (credit: Scott Robinson, Flickr)

What causes thin film interference? Figure 2 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). Since the ray that enters the film travels a greater distance, 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 1. 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 ray 1 and ray 2 in Figure 2 is negligible; so why should they interfere destructively and not constructively? The answer is that a phase change can occur upon reflection. The rule is as follows:

> When light reflects from a medium having an index of refraction greater than that of the medium in which it is traveling, a 180º phase change (or a λ / 2 shift) occurs.

If the film in Figure 2 is a soap bubble (essentially water with air on both sides), then there is a $\lambda/2$ shift for ray 1 and none for ray 2. Thus, when the film is very thin, the path length difference between the two rays is negligible, they are exactly out of phase, and destructive interference will occur at all wavelengths and so the soap bubble will be dark here.

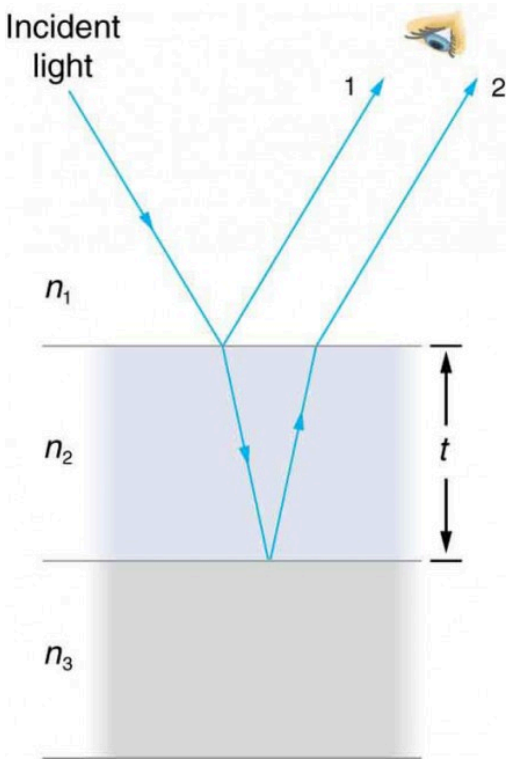


Figure 2. Light striking a thin film is partially reflected (ray 1) and partially refracted at the top surface. The refracted ray is partially reflected at the bottom surface and emerges as ray 2. These rays will interfere in a way that depends on the thickness of the film and the indices of refraction of the various media.

The thickness of the film relative to the wavelength of light is the other crucial factor in thin film interference. Ray 2 in Figure 2 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 1. Calculating Non-reflective Lens Coating Using Thin Film Interference

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 that 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? The index of refraction of glass is 1.52.

Strategy

Refer to Figure 2 and use n_1=100 for air, n_2 = 1.38, and n_3 = 1.52. Both ray 1 and ray 2 will have a $\lambda/2$ shift

upon reflection. Thus, to obtain destructive interference, ray 2 will need to travel a half wavelength farther than ray 1. For rays incident perpendicularly, the path length difference is 2*t*.

Solution

To obtain destructive interference here,

$$2t = \frac{\lambda_{n2}}{2}$$

,

where λ_{n2} is the wavelength in the film and is given by

$$\lambda_{n2} = \frac{\lambda}{n_2}$$

.

Thus,

$$2t = \frac{\lambda/n_2}{2}$$

Solving for *t* and entering known values yields

$$\begin{aligned} t &= \frac{\lambda/n_2}{4} = \frac{550 \text{ nm}/1.38}{4} \\ &= 99.6 \text{ nm} \end{aligned}$$

Discussion

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 will be reduced in intensity. These films are called non-reflective coatings; this is only an approximately correct description, though, since other wavelengths will only be partially cancelled. Non-reflective coatings are used in car windows and sunglasses.

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, respectively. That is, for rays incident perpendicularly, $2t = \lambda_n, 2\lambda_n, 3\lambda_n, \ldots$ or

$$2t = \frac{\lambda_n}{2}, \frac{3\lambda_n}{2}, \frac{5\lambda_n}{2}, \ldots$$

.

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 will observe rainbow colors of constructive interference for various wavelengths as the thickness varies.

Example 2. Soap Bubbles: More Than One Thickness can be Constructive

1. 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.
2. What three smallest thicknesses will give destructive interference?

Strategy and Concept

Use Figure 2 to visualize the bubble. Note that $n_1 = n_3 = 1.00$ for air, and $n_2 = 1.333$ for soap (equivalent to water). There is a

$$\frac{\lambda}{2}$$

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

$$\frac{\lambda_n}{2}, \frac{3\lambda_n}{2}, \text{ and } \frac{5\lambda_n}{2}$$

. 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 for Part 1

Constructive interference occurs here when

$$2t_c = \frac{\lambda_n}{2}, \frac{3\lambda_n}{2}, \text{ and } \frac{5\lambda_n}{2}, \ldots$$

.

The smallest constructive thickness t_c thus is

$$\begin{aligned} t_c &= \frac{\lambda_n}{4} = \frac{\lambda/n}{4} = \frac{650 \text{ nm}/1.333}{4} \\ &= 122 \text{ nm} \end{aligned}$$

The next thickness that gives constructive interference is $t'_c = 3\lambda_n/4$, so that $t'_c = 366$ nm.

Finally, the third thickness producing constructive interference is $t''_c \leq 5\lambda_n/4$, so that $t''_c = 610$ nm.

Solution for Part 2

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 first non-zero thickness producing destructive interference is $2t'_d = \lambda_n$.

Substituting known values gives

$$\begin{aligned} t_d &= \frac{\lambda}{2} = \frac{\lambda/n}{2} = \frac{650 \text{ nm}/1.333}{2} \\ &= 244 \text{ nm} \end{aligned}$$

Finally, the third destructive thickness is $2t''_d = 2\lambda_n$, so that

$$\begin{aligned} t_d &= \lambda_n = \frac{\lambda}{n} = \frac{650 \text{ nm}}{1.333} \\ &= 488 \text{ nm} \end{aligned}$$

Discussion

If the bubble was 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.

Another example of thin film interference can be seen when microscope slides are separated (see Figure 3). 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, and so there is a dark band 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 pure-wavelength light instead of white light is used, then bright and dark bands are obtained rather than repeating rainbow colors.

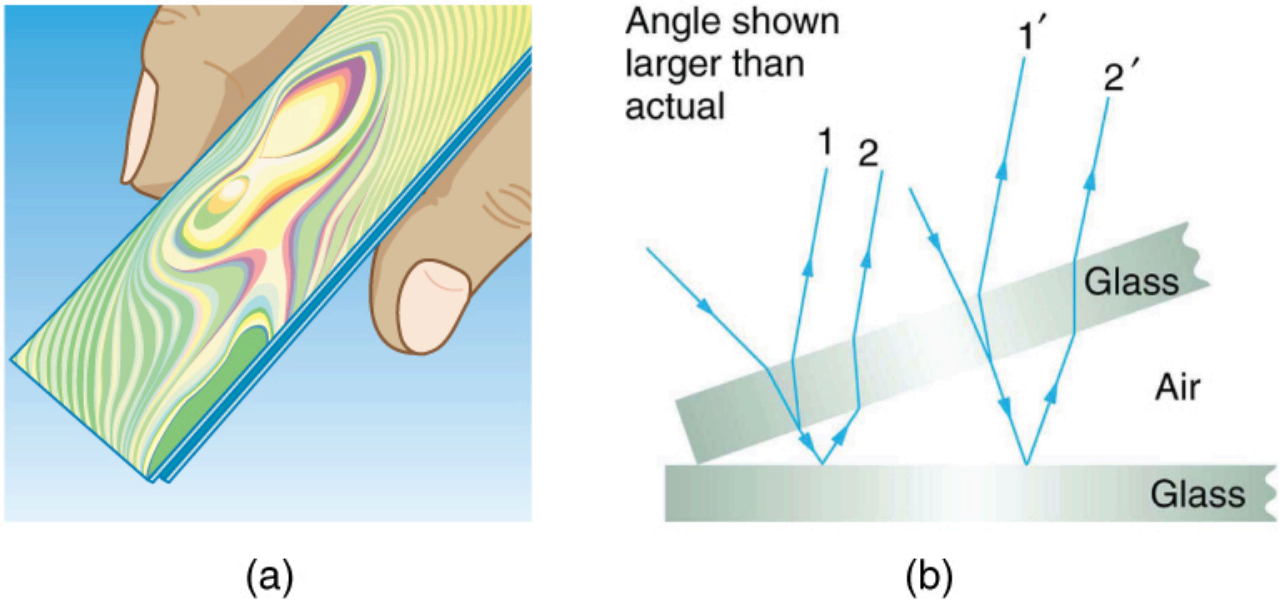


Figure 3. (a) The rainbow color bands are produced by thin film interference in the air between the two glass slides. (b) Schematic of the paths taken by rays in the wedge of air between the slides.

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 4 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 one wavelength in the distance between the lens and the blank, so that great precision can be obtained. Once the lens is perfect, there will be no rings.

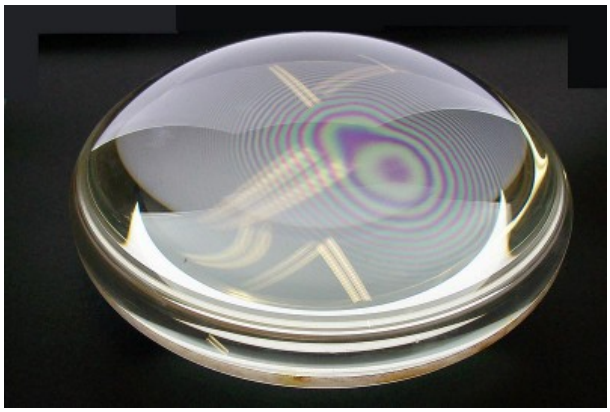


Figure 4. "Newton's rings" interference fringes are produced when two plano-convex lenses are placed together with their plane surfaces in contact. The rings are created by interference between the light reflected off the two surfaces as a result of a slight gap between them, indicating that these surfaces are not precisely plane but are slightly convex. (credit: Ulf Seifert, Wikimedia Commons)

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. Car manufacturers are offering 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. 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 New Zealand and Taiwan are using similar technologies, while the United States currency includes a thin film interference effect.

Making Connections: Take-Home Experiment—Thin Film Interference

One feature of thin film interference and diffraction gratings is that the pattern shifts as you change the angle at which you look or move your head. Find examples of thin film interference and gratings around you. Explain how the patterns change for each specific example. Find examples where the thickness changes giving rise to changing colors. If you can find two microscope slides, then try observing the effect shown in Figure 3. Try separating one end of the two slides with a hair or maybe a thin piece of paper and observe the effect.

Problem-Solving Strategies for Wave Optics

Step 1. *Examine the situation to determine that interference is involved*. Identify whether slits or thin film interference 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 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 will have constructive interference for a total shift that is an integral number of wavelengths. You will 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 90º, for example.

Section Summary

- 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.
- When light reflects from a medium having an index of refraction greater than that of the medium in which it is traveling, a 180º phase change (or a $\frac{\lambda}{2}$ shift) occurs.

Conceptual Questions

1. 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?
2. 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?
3. 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.
4. 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?
5. Answer the above question if the fluid between the two pieces of crown glass is carbon disulfide.
6. While contemplating the food value of a slice of ham, you notice a rainbow of color reflected

from its moist surface. Explain its origin.

7. 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 non-reflective 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?
8. A non-reflective coating like the one described in Example 1 works ideally for a single wavelength and for perpendicular incidence. What happens for other wavelengths and other incident directions? Be specific.
9. 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?

Problems & Exercises

1. 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?
2. 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?
3. 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.
4. 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.
5. 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?
6. 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?
7. 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 non-reflective for this wavelength?
8. (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.
9. 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.

10. Figure 3 shows two glass slides illuminated by pure-wavelength 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.
11. Figure 3 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?
12. Repeat Question 1, but take the light to be incident at a 45º angle.
13. Repeat Question 2, but take the light to be incident at a 45º angle.
14. **Unreasonable Results.** To save money on making military aircraft invisible to radar, an inventor decides to coat them with a non-reflective 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?

Glossary

thin film interference: interference between light reflected from different surfaces of a thin film

Selected Solutions to Problems & Exercises

1. 532 nm (green)
3. 83.9 nm
5. 620 nm (orange)
7. 380 nm
9. 33.9 nm
11. 4.42×10^{-5} m
13. The oil film will appear black, since the reflected light is not in the visible part of the spectrum.

Polarization

Lumen Learning

Learning Objectives

By the end of this section, you will be able

- Discuss the meaning of polarization.
- Discuss the property of optical activity of certain materials.

Polaroid sunglasses are familiar to most of us. They have a special ability to cut the glare of light reflected from water or glass (see Figure 1). Polaroids 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 1. 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: Amithshs, Wikimedia Commons)

Light is one type of electromagnetic (EM) wave. As noted earlier, EM waves are *transverse waves* consisting of varying electric and magnetic fields that oscillate perpendicular to the direction of propagation (see Figure 2). There are specific directions for the oscillations of the electric and magnetic fields. *Polarization* is the attribute that a wave's oscillations 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 2.

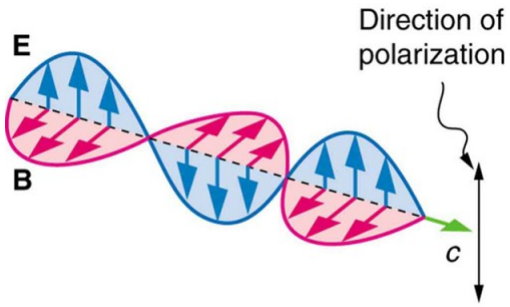


Figure 2. An EM wave, such as light, is a transverse wave. The electric and magnetic fields are perpendicular to the direction of propagation.

To examine this further, consider the transverse waves in the ropes shown in Figure 3. 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.

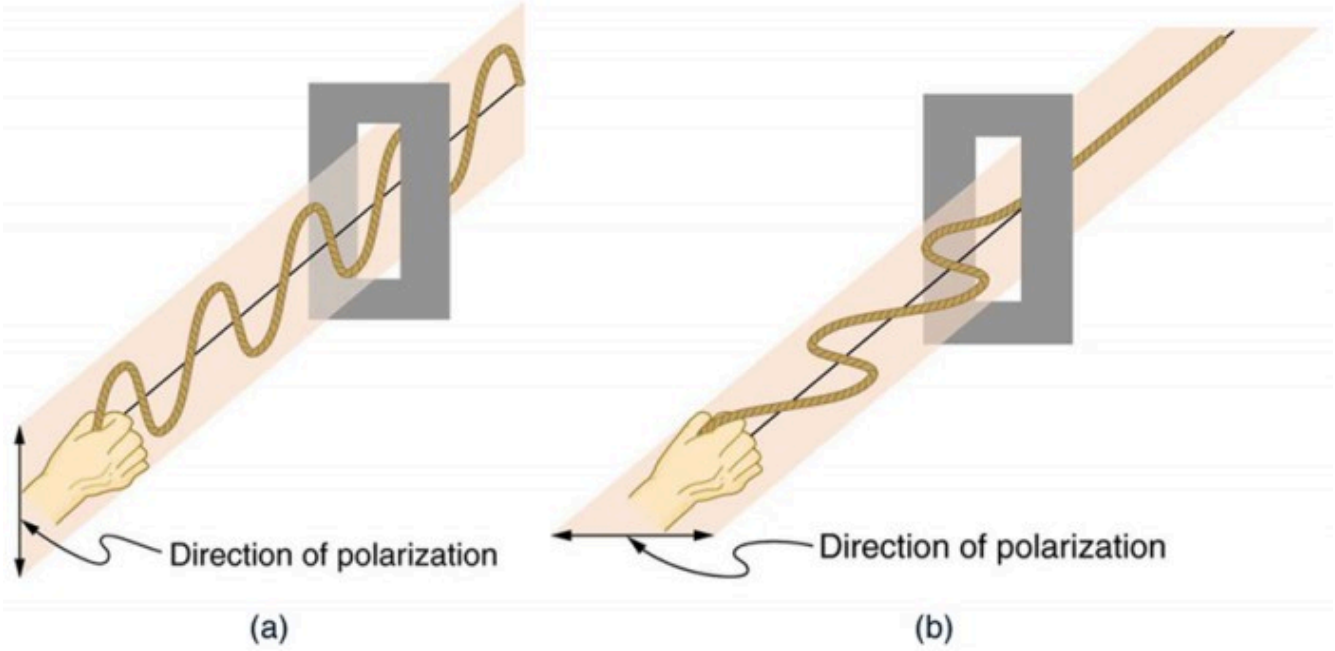


Figure 3. The transverse oscillations in one rope are in a vertical plane, and those in the other rope are in a horizontal plane. The first is said to be vertically polarized, and the other is said to be horizontally polarized. Vertical slits pass vertically polarized waves and block horizontally polarized waves.

The Sun and many other light sources produce waves that are randomly polarized (see Figure 4). Such light is said to be *unpolarized* because it is composed of many waves with all possible directions of polarization. Polaroid materials, invented by the founder of 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. Thinking 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 (see Figure 5).

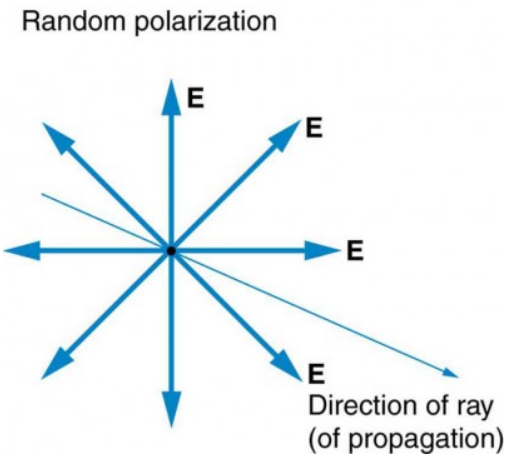


Figure 4. The slender arrow represents a ray of unpolarized light. The bold arrows represent the direction of polarization of the individual waves composing the ray. Since the light is unpolarized, the arrows point in all directions.

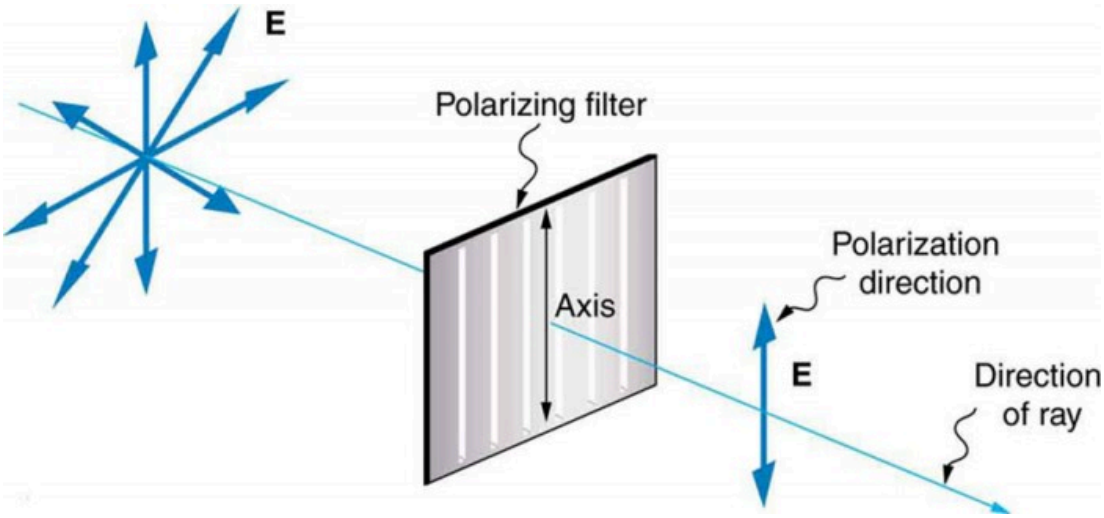


Figure 5. A polarizing filter has a polarization axis that acts as a slit passing through electric fields parallel to its direction. The direction of polarization of an EM wave is defined to be the direction of its electric field.

Figure 6 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. 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.

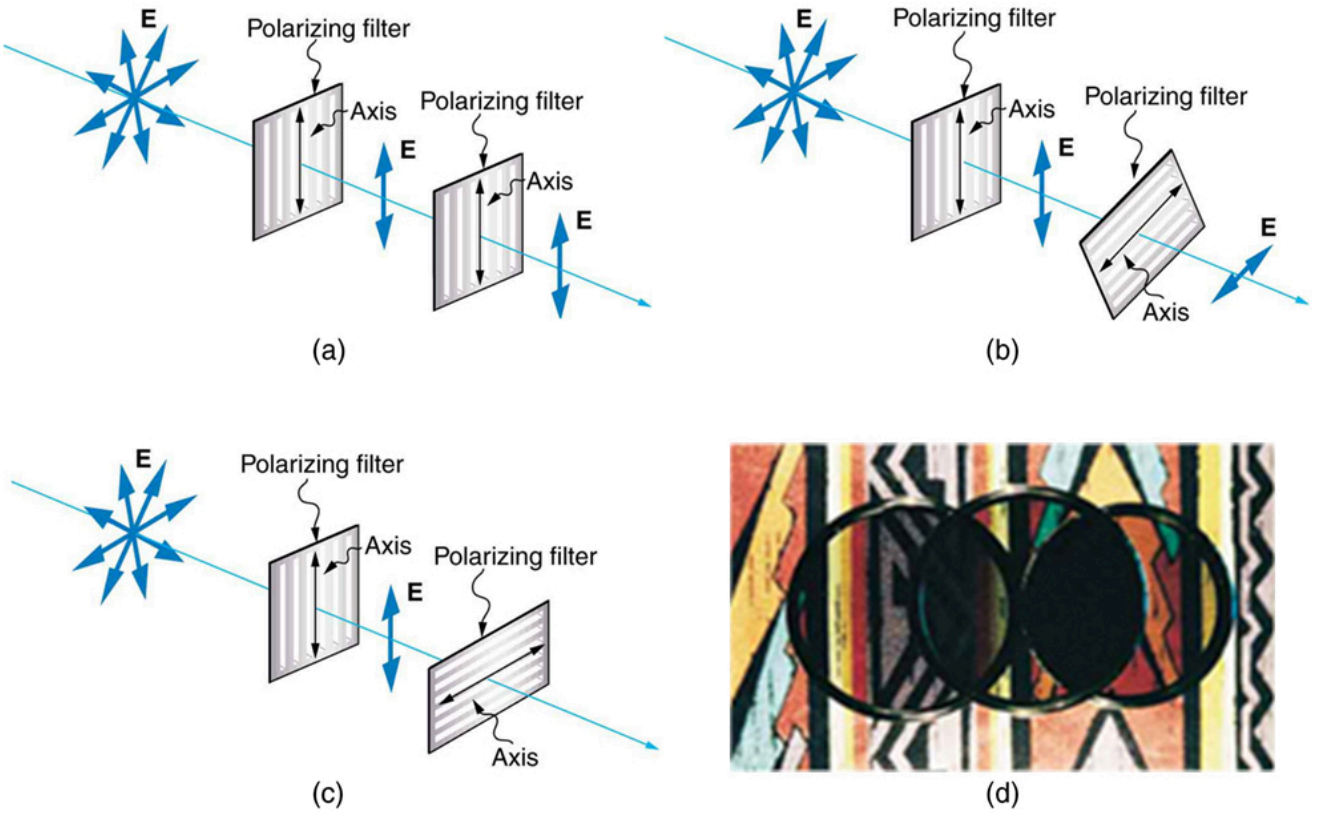


Figure 6. The effect of rotating two polarizing filters, where the first polarizes the light. (a) All of the polarized light is passed by the second polarizing filter, because its axis is parallel to the first. (b) As the second is rotated, only part of the light is passed. (c) When the second is perpendicular 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: 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 $E \cos \theta$ (see Figure 7). 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 $I = I_0 \cos^2 \theta$, where I_0 is the intensity of the polarized wave before passing through the filter. (The above equation is known as Malus's law.)

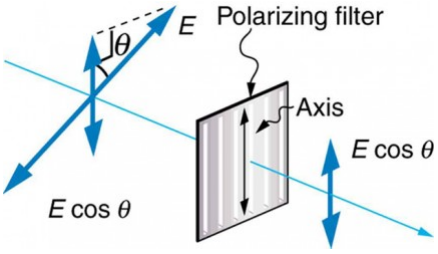


Figure 7. A polarizing filter transmits only the component of the wave parallel to its axis, , reducing the intensity of any light not polarized parallel to its axis.

Example 1. 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 $\theta = \cos^{-1} 0.3162 = 71.6^\circ$.

Discussion

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 45º, the intensity is reduced to 50% of its original value (as you will show in this section's Problems & Exercises). Note that 71.6º is 18.4º from reducing the intensity to zero, and that at an angle of 18.4º the intensity is reduced to 90.0% of its original value (as you will also show in Problems & Exercises), giving evidence of symmetry.

Polarization by Reflection

By now you can probably guess that Polaroid sunglasses cut the glare in reflected light because that light is polarized. You can check this for yourself by holding Polaroid 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 8 illustrates what happens when unpolarized light is reflected from a surface. Vertically polarized light is preferentially refracted at the surface, so that *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 would be like an arrow perpendicular to the surface and would be more likely to stick and not be reflected. Horizontal polarization is like an arrow bouncing on its side and would be more likely to be reflected. Sunglasses with vertical axes would then block more reflected light than unpolarized light from other sources.

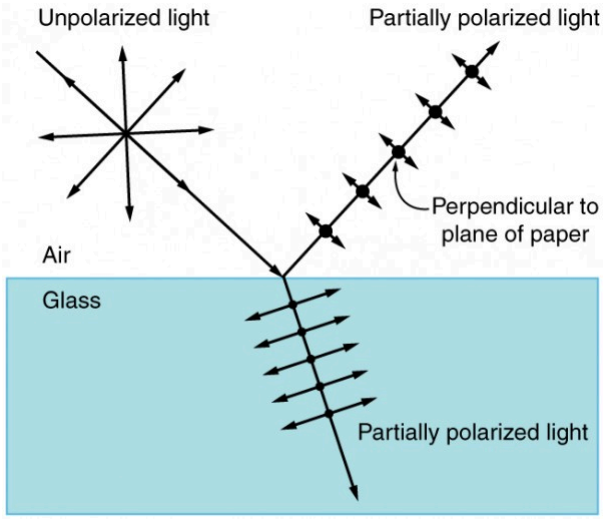


Figure 8. Polarization by reflection. Unpolarized light has equal amounts of vertical and horizontal polarization. After interaction with a surface, the vertical components are preferentially absorbed or refracted, leaving the reflected light more horizontally polarized. This is akin to arrows striking on their sides 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 a angle of reflection θ_b, given by

$$\tan\theta_b = \frac{n_2}{n_1}$$

, 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. This equation is known as *Brewster's law*, and θ_b is known as *Brewster's angle*, named after the 19th-century Scottish physicist who discovered them.

Things Great and Small: Atomic Explanation of Polarizing Filters

Polarizing filters have a polarization axis that acts as a slit. This slit passes electromagnetic 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 9.

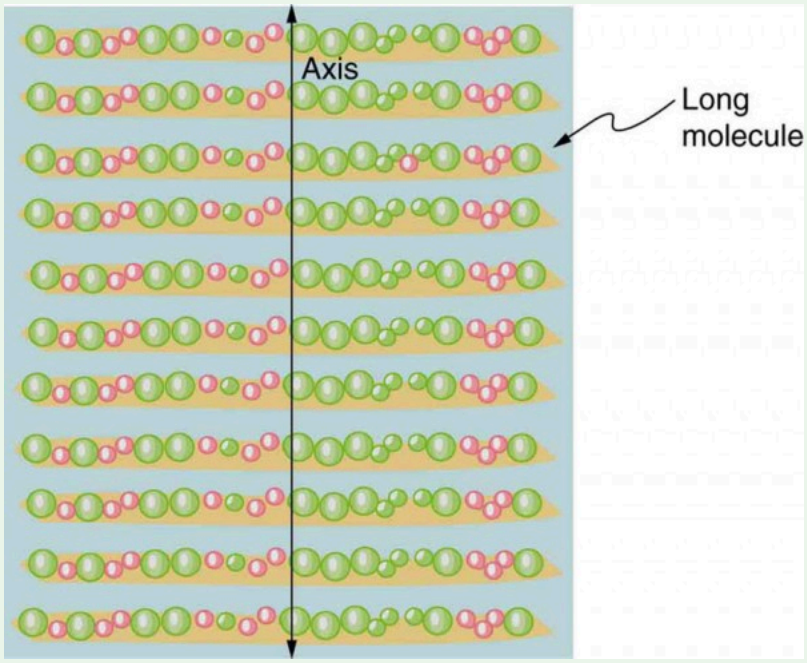


Figure 9. Long molecules are aligned perpendicular to the axis of a polarizing filter. The component of the electric field in an EM wave perpendicular to these molecules passes through the filter, while the component parallel to the molecules is absorbed.

Figure 10 illustrates how the component of the electric field parallel to the long molecules is absorbed. An electromagnetic 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 in the molecules, since electron masses are small. If the electron is forced to oscillate, it can absorb energy from the EM wave. This reduces the fields 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 will allow those fields to pass. Thus the axis of the polarizing filter is perpendicular to the length of the molecule.

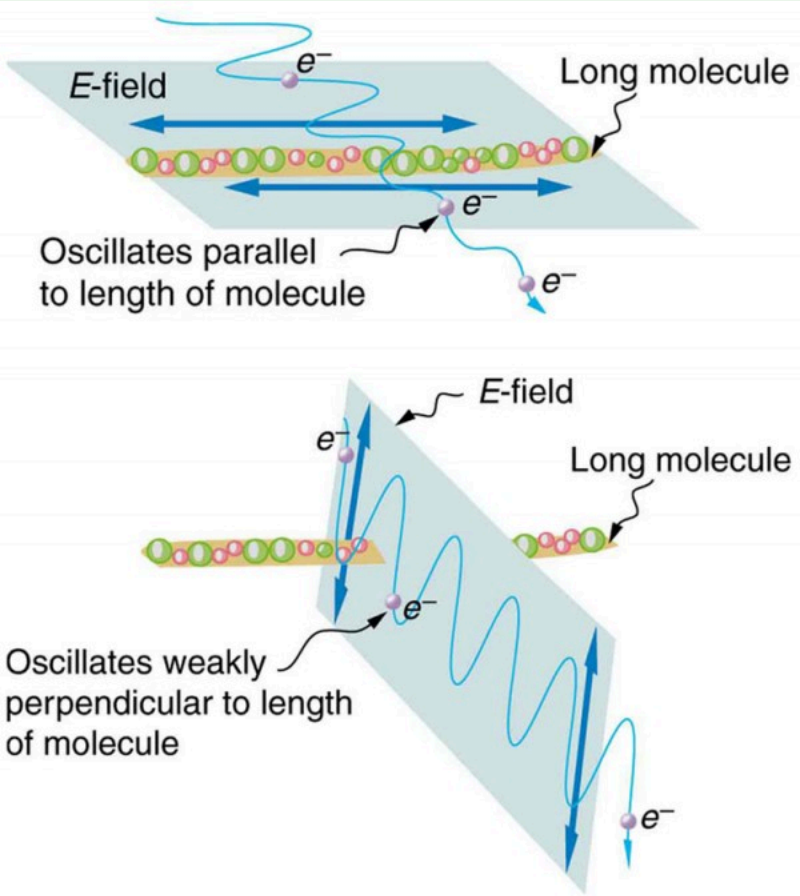


Figure 10. Artist's conception of an electron in a long molecule oscillating parallel to the molecule. The oscillation of the electron absorbs energy and reduces the intensity of the component of the EM wave that is parallel to the molecule.

Example 2. Calculating Polarization by Reflection

1. At what angle will light traveling in air be completely polarized horizontally when reflected from water?
2. 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 for Part 1

Putting the known quantities into the equation

$$\tan\theta_{\text{b}} = \frac{n_2}{n_1}$$

gives

$$\tan\theta_{\text{b}} = \frac{n_2}{n_1} = \frac{1.333}{1.00} = 1.333$$

.

Solving for the angle θ_{b} yields

$$\theta_{\text{b}} = \tan^{-1} 1.333 = 53.1^{\text{o}}.$$

Solution for Part 2

Similarly, for crown glass and air,

$$\tan\theta'_{\text{b}} = \frac{n'_2}{n_1} = \frac{1.520}{1.00} = 1.52$$

.

Thus,

$$\theta'_{\text{b}} = \tan^{-1} 1.52 = 56.7^{\text{o}}.$$

Discussion

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 not reflected is refracted into these media. So at an incident angle equal to Brewster's angle, the refracted light will be slightly polarized vertically. It will not be completely polarized vertically, because only a small fraction of the incident light is reflected, and so a significant amount of horizontally polarized light is refracted.

Polarization by Scattering

If you hold your Polaroid 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 11 helps illustrate how this happens. Since light is a transverse EM wave, it vibrates the electrons of air molecules perpendicular to the direction 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 Figure 11, 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 will only be partially polarized. Furthermore, multiple scattering can bring light to your eyes from other directions and can contain different polarizations.

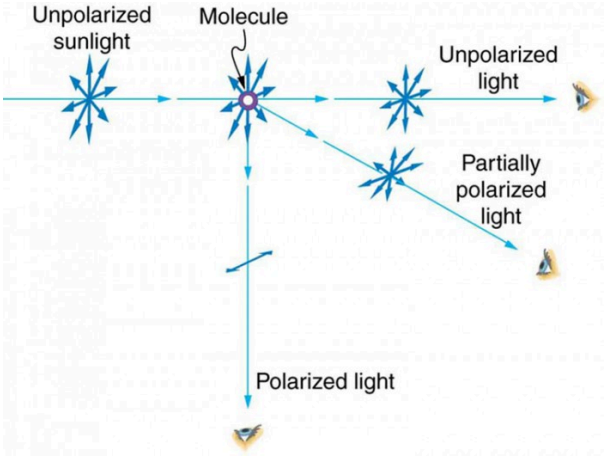


Figure 11. Polarization by scattering. Unpolarized light scattering from air molecules shakes their electrons perpendicular to the direction of the original ray. The scattered light therefore has a polarization perpendicular to the original direction and none parallel to the original direction.

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.

There is a range of optical effects used in sunglasses. Besides being Polaroid, other sunglasses have colored pigments embedded in them, while others use non-reflective or even reflective coatings. 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.

Take-Home Experiment: Polarization

Find Polaroid sunglasses and rotate one while holding the other still and look at different surfaces and objects. Explain your observations. What is the difference in angle from when you see a maximum intensity to when you see a minimum intensity? Find a reflective glass surface and do the same. At what angle does the glass need to be oriented to give minimum glare?

Liquid Crystals and Other Polarization Effects in Materials

While you are undoubtedly aware of liquid crystal displays (LCDs) found in watches, calculators, computer screens, cellphones, flat screen televisions, and other myriad 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 90º. Furthermore, this property can be turned off by the application of a voltage, as illustrated in Figure 12. 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, there is a large light 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 12 (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. One can vary the picture contrast by varying the strength of the voltage applied to the liquid crystal.

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 (see Figure 13). In addition to depending on the type of substance, the amount and direction of rotation depends on a number of 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 asymmetric 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.

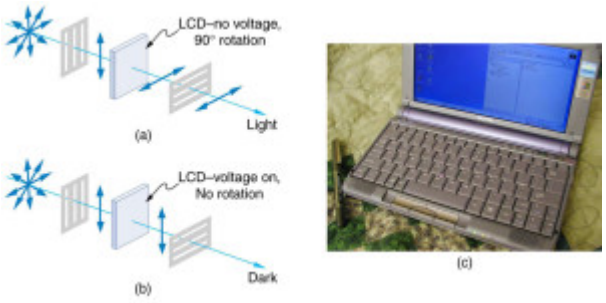


Figure 12. (a) Polarized light is rotated 90º by a liquid crystal and then passed by a polarizing filter that has its axis perpendicular to the original polarization direction. (b) When a voltage is applied to the liquid crystal, the polarized light is not rotated and is blocked by the filter, making 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: Jon Sullivan)

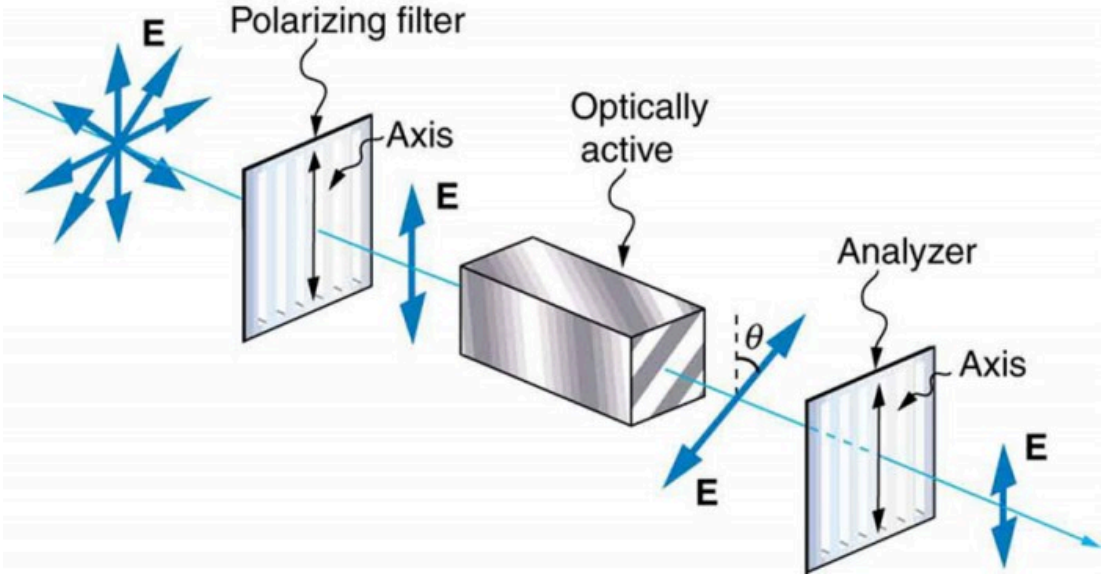


Figure 13. Optical activity is the ability of some substances to rotate the plane of polarization of light passing through them. The rotation is detected with a polarizing filter or analyzer.

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 14. It is apparent that the effect depends on wavelength as well as stress. The wavelength dependence is sometimes also used for artistic purposes.

Another interesting phenomenon associated with polarized light is the ability of some crystals to split an unpolarized beam of light into two. Such crystals are said to be *birefringent* (see Figure 15). Each of the separated rays has a specific polarization. One behaves normally and is called the ordinary ray, whereas the other does not obey Snell's law and is called the extraordinary ray. 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. The interested reader is invited to further pursue the numerous properties of materials related to polarization.

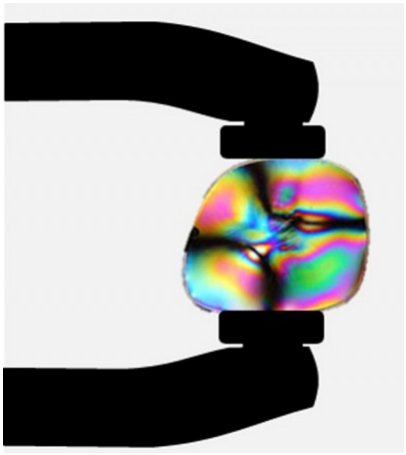


Figure 14. Optical stress analysis of a plastic lens placed between crossed polarizers. (credit: Infopro, Wikimedia Commons)

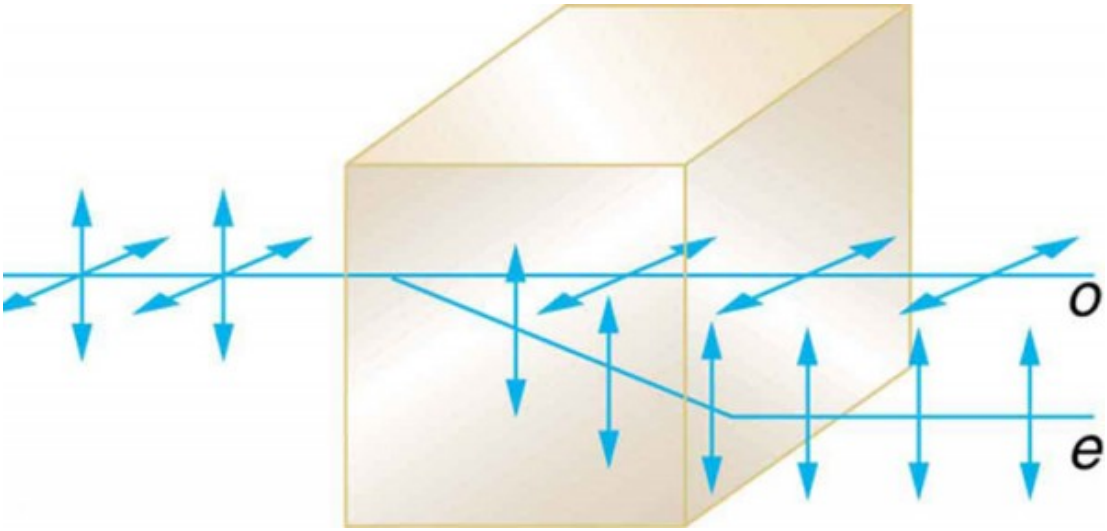


Figure 15. Birefringent materials, such as the common mineral calcite, split unpolarized beams of light into two. The ordinary ray behaves as expected, but the extraordinary ray does not obey Snell's law.

Section Summary

- Polarization is the attribute that wave oscillations have a definite direction relative to the direction of propagation of the wave.
- EM waves are transverse waves that may be polarized.
- 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.
- Light can be polarized by passing it through a polarizing filter or other polarizing material. The intensity I of polarized light after passing through a polarizing filter is $I = I_0 \cos^2 \theta$, where I_0 is the original 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 will be completely polarized at the angle of reflection θ_b, known as Brewster's angle, given by a statement known as 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.
- Polarization can also be produced by scattering.
- There are a number of types of optically active substances that rotate the direction of polarization of light passing through them.

Conceptual Questions

1. Under what circumstances is the phase of light changed by reflection? Is the phase related to polarization?
2. Can a sound wave in air be polarized? Explain.
3. 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?
4. Explain what happens to the energy carried by light that it is dimmed by passing it through two crossed polarizing filters.
5. When particles scattering light are much smaller than its wavelength, the amount of scattering is proportional to $\frac{1}{\lambda^4}$. 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?
6. Using the information given in the preceding question, explain why sunsets are red.
7. 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%?

Problems & Exercises

1. What angle is needed between the direction of polarized light and the axis of a polarizing filter to cut its intensity in half?
2. The angle between the axes of two polarizing filters is 45.0º. By how much does the second filter reduce the intensity of the light coming through the first?
3. 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 89.0º angle to the light's polarization direction?
4. 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?
5. At the end of Example 1, 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 18.4º to the direction of polarization. Verify this statement.
6. Show that if you have three polarizing filters, with the second at an angle of 45º to the first and the third at an angle of 90.0º 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.)
7. 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 $90.0^\circ - \theta$, then $I + I' = I_0$ the original intensity.

(Hint: Use the trigonometric identities cos (90.0º − θ) = sin θ and $\cos^2 \theta + \sin^2 \theta = 1$.)

8. At what angle will light reflected from diamond be completely polarized?
9. What is Brewster's angle for light traveling in water that is reflected from crown glass?
10. A scuba diver sees light reflected from the water's surface. At what angle will this light be completely polarized?
11. At what angle is light inside crown glass completely polarized when reflected from water, as in a fish tank?
12. Light reflected at 55.6º from a window is completely polarized. What is the window's index of refraction and the likely substance of which it is made?
13. (a) Light reflected at 62.5º 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?
14. 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 $\theta_b + \theta'_b = 90.0º$.
15. **Integrated Concepts.** 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?
16. **Integrated Concepts.** Suppose you put on two pairs of Polaroid sunglasses with their axes at an angle of 15.0º. 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.
17. **Integrated Concepts.** (a) On a day when the intensity of sunlight is 1.00 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 20.0º. Assuming the sunlight is unpolarized and the polarizers are 100% efficient, what is the initial rate of heating of the water in ºC/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.

Glossary

axis of a polarizing filter: the direction along which the filter passes the electric field of an EM wave

birefringent: crystals that split an unpolarized beam of light into two beams

Brewster's angle:

$$\theta_b = \tan^{-1}\left(\frac{n_2}{n_1}\right)$$

, where n_2 is the index of refraction of the medium from which the light is reflected and n_1 is the index of refraction of the medium in which the reflected light travels

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

direction of polarization: the direction parallel to the electric field for EM waves

horizontally polarized: the oscillations are in a horizontal plane

optically active: substances that rotate the plane of polarization of light passing through them

polarization: the attribute that wave oscillations have a definite direction relative to the direction of propagation of the wave

polarized: waves having the electric and magnetic field oscillations in a definite direction

reflected light that is completely polarized: light reflected at the angle of reflection θ_b, known as Brewster's angle

unpolarized: waves that are randomly polarized

vertically polarized: the oscillations are in a vertical plane

Selected Solutions to Problems & Exercises

1. 45.0º

3. 45.7 mW/m^2

5. 90.0%

7. I_0

9. 48.8º

11. 41.2º

13. (a) 1.92, not diamond (Zircon); (b) 55.2º

15. $B_2 = 0.707\ B_1$

17. (a) 2.07×10^{-2} °C/s; (b) Yes, the polarizing filters get hot because they absorb some of the lost energy from the sunlight.

Extended Topic Microscopy Enhanced by the Wave Characteristics of Light

Lumen Learning

Learning Objective

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

- Discuss the different types of microscopes.

Physics research underpins the advancement of developments in microscopy. As we gain knowledge of the wave nature of electromagnetic waves and methods to analyze and interpret signals, new microscopes that enable us to "see" more are being developed. It is the evolution and newer generation of microscopes that are described in this section.

The use of microscopes (microscopy) to observe small details is limited by the wave nature of light. Owing to the fact that light diffracts significantly around small objects, it becomes impossible to observe details significantly smaller than the wavelength of light. One rule of thumb has it that all details smaller than about λ are difficult to observe. Radar, for example, can detect the size of an aircraft, but not its individual rivets, since the wavelength of most radar is several centimeters or greater. Similarly, visible light cannot detect individual atoms, since atoms are about 0.1 nm in size and visible wavelengths range from 380 to 760 nm. Ironically, special techniques used to obtain the best possible resolution with microscopes take advantage of the same wave characteristics of light that ultimately limit the detail.

Making Connections: Waves

All attempts to observe the size and shape of objects are limited by the wavelength of the probe. Sonar and medical ultrasound are limited by the wavelength of sound they employ. We shall see that this is also true in electron microscopy, since electrons have a wavelength. Heisenberg's uncertainty principle asserts that this limit is fundamental and inescapable, as we shall see in quantum mechanics.

The most obvious method of obtaining better detail is to utilize shorter wavelengths. *Ultraviolet (UV) microscopes* have been constructed with special lenses that transmit UV rays and utilize photographic or electronic techniques to record images. The shorter UV wavelengths allow somewhat greater detail to be observed, but drawbacks, such as the hazard of UV to living tissue and the need for special detection devices and lenses (which tend to be dispersive in the UV), severely limit the use of UV microscopes. Elsewhere, we will explore practical uses of very short wavelength EM waves, such as x rays, and other short-wavelength probes, such as electrons in electron microscopes, to detect small details.

Another difficulty in microscopy is the fact that many microscopic objects do not absorb much of the light passing through them. The lack of contrast makes image interpretation very difficult. *Contrast* is the difference in intensity between objects and the background on which they are observed. Stains (such as dyes, fluorophores, etc.) are commonly employed to enhance contrast, but these tend to be application specific. More general wave interference techniques can be used to produce contrast. Figure 1 shows the passage of light through a sample. Since the indices of refraction differ, the number of wavelengths in the paths differs. Light emerging from the object is thus out of phase with light from the background and will interfere differently, producing enhanced contrast, especially if the light is coherent and monochromatic—as in laser light.

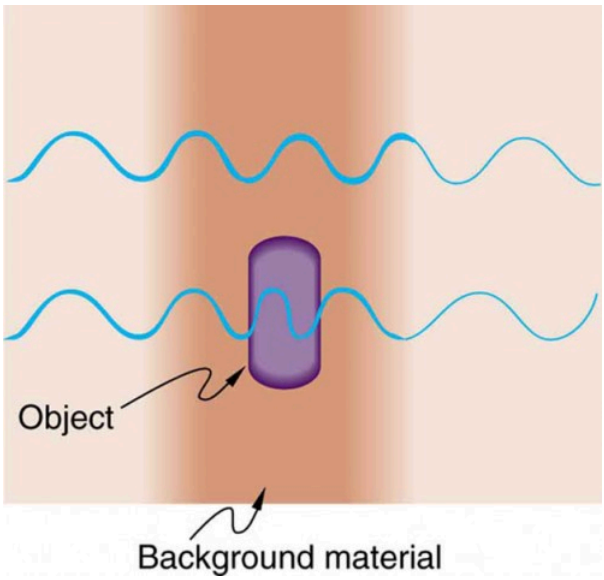


Figure 1. Light rays passing through a sample under a microscope will emerge with different phases depending on their paths. The object shown has a greater index of refraction than the background, and so the wavelength decreases as the ray passes through it. Superimposing these rays produces interference that varies with path, enhancing contrast between the object and background.

Interference microscopes enhance contrast between objects and background by superimposing a reference beam of light upon the light emerging from the sample. Since light from the background and objects differ in phase, there will be different amounts of constructive and destructive interference, producing the desired contrast in final intensity. Figure 2 shows schematically how this is done. Parallel rays of light from a source are split into two beams by a half-silvered mirror. These beams are called the object and reference beams. Each beam passes through identical optical elements, except that the object beam passes through the object we wish to observe microscopically. The light beams are recombined by another half-silvered mirror and interfere. Since the light rays passing through different parts of the object have different phases, interference will be significantly different and, hence, have greater contrast between them.

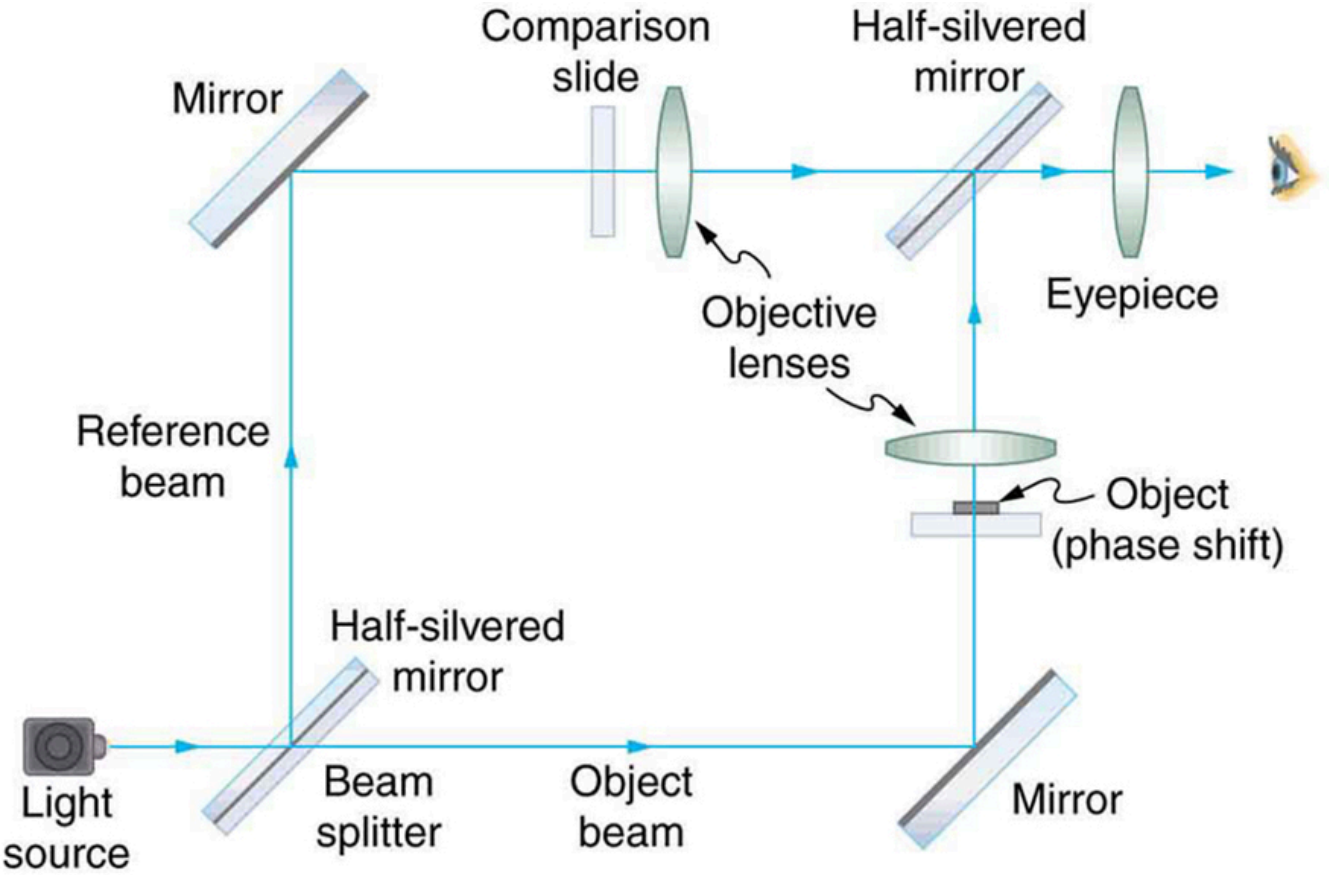


Figure 2. An interference microscope utilizes interference between the reference and object beam to enhance contrast. The two beams are split by a half-silvered mirror; the object beam is sent through the object, and the reference beam is sent through otherwise identical optical elements. The beams are recombined by another half-silvered mirror, and the interference depends on the various phases emerging from different parts of the object, enhancing contrast.

Another type of microscope utilizing wave interference and differences in phases to enhance contrast is called the *phase-contrast microscope*. While its principle is the same as the interference microscope, the phase-contrast microscope is simpler to use and construct. Its impact (and the principle upon which it is based) was so important that its developer, the Dutch physicist Frits Zernike (1888–1966), was awarded the Nobel Prize in 1953. Figure 3 shows the basic construction of a phase-contrast microscope. Phase differences between light passing through the object and background are produced by passing the rays through different parts of a phase plate (so called because it shifts the phase of the light passing through it). These two light rays are superimposed in the image plane, producing contrast due to their interference.

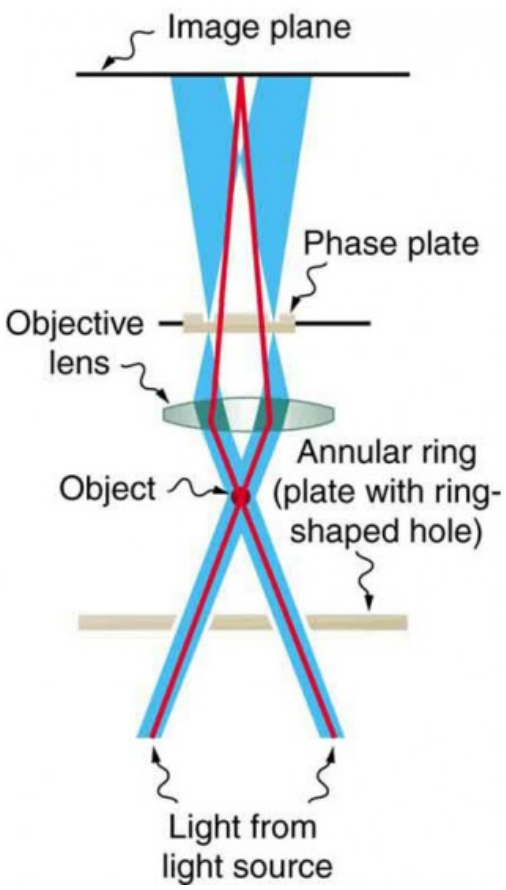


Figure 3. Simplified construction of a phase-contrast microscope. Phase differences between light passing through the object and background are produced by passing the rays through different parts of a phase plate. The light rays are superimposed in the image plane, producing contrast due to their interference.

A *polarization microscope* also enhances contrast by utilizing a wave characteristic of light. Polarization microscopes are useful for objects that are optically active or birefringent, particularly if those characteristics vary from place to place in the object. Polarized light is sent through the object and then observed through a polarizing filter that is perpendicular to the original polarization direction. Nearly transparent objects can then appear with strong color and in high contrast. Many polarization effects are wavelength dependent, producing color in the processed image. Contrast results from the action of the polarizing filter in passing only components parallel to its axis.

Apart from the UV microscope, the variations of microscopy discussed so far in this section are available as attachments to fairly standard microscopes or as slight variations. The next level of sophistication is provided by commercial *confocal microscopes*, which use the extended focal region shown in Figure 4b to obtain three-dimensional images rather than two-dimensional images.

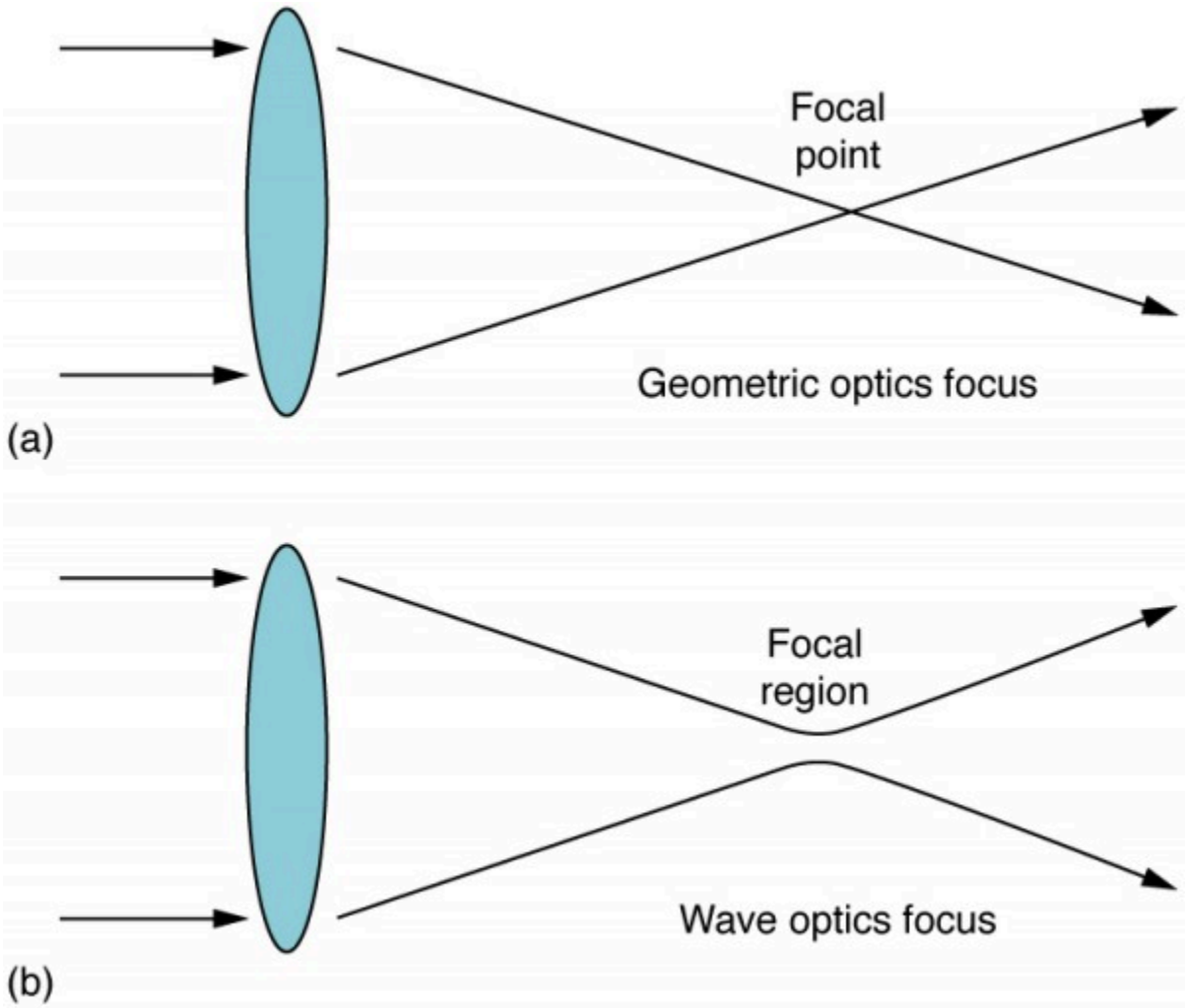


Figure 4. (a) In geometric optics, the focus is a point, but it is not physically possible to produce such a point because it implies infinite intensity. (b) In wave optics, the focus is an extended region.

Here, only a single plane or region of focus is identified; out-of-focus regions above and below this plane are subtracted out by a computer so the image quality is much better. This type of microscope makes use of fluorescence, where a laser provides the excitation light. Laser light passing through a tiny aperture called a pinhole forms an extended focal region within the specimen. The reflected light passes through the objective lens to a second pinhole and the photomultiplier detector, see Figure 5. The second pinhole is the key here and serves to block much of the light from points that are not at the focal point of the objective lens. The pinhole is conjugate (coupled) to the focal point of the lens. The second pinhole and detector are scanned, allowing reflected light from a small region or section of the extended focal region to be imaged at any one time. The out-of-focus light is excluded. Each image is stored in a computer, and a full scanned image is generated in a short time. Live cell processes can also be imaged at adequate scanning speeds allowing the imaging of three-dimensional microscopic movement. Confocal microscopy enhances images over conventional optical microscopy, especially for thicker specimens, and so has become quite popular.

The next level of sophistication is provided by microscopes attached to instruments that isolate and detect only a small wavelength band of light—monochromators and spectral analyzers. Here, the monochromatic light from a laser is scattered from the specimen. This scattered light shifts up or down

as it excites particular energy levels in the sample. The uniqueness of the observed scattered light can give detailed information about the chemical composition of a given spot on the sample with high contrast—like molecular fingerprints. Applications are in materials science, nanotechnology, and the biomedical field. Fine details in biochemical processes over time can even be detected. The ultimate in microscopy is the electron microscope—to be discussed later. Research is being conducted into the development of new prototype microscopes that can become commercially available, providing better diagnostic and research capacities.

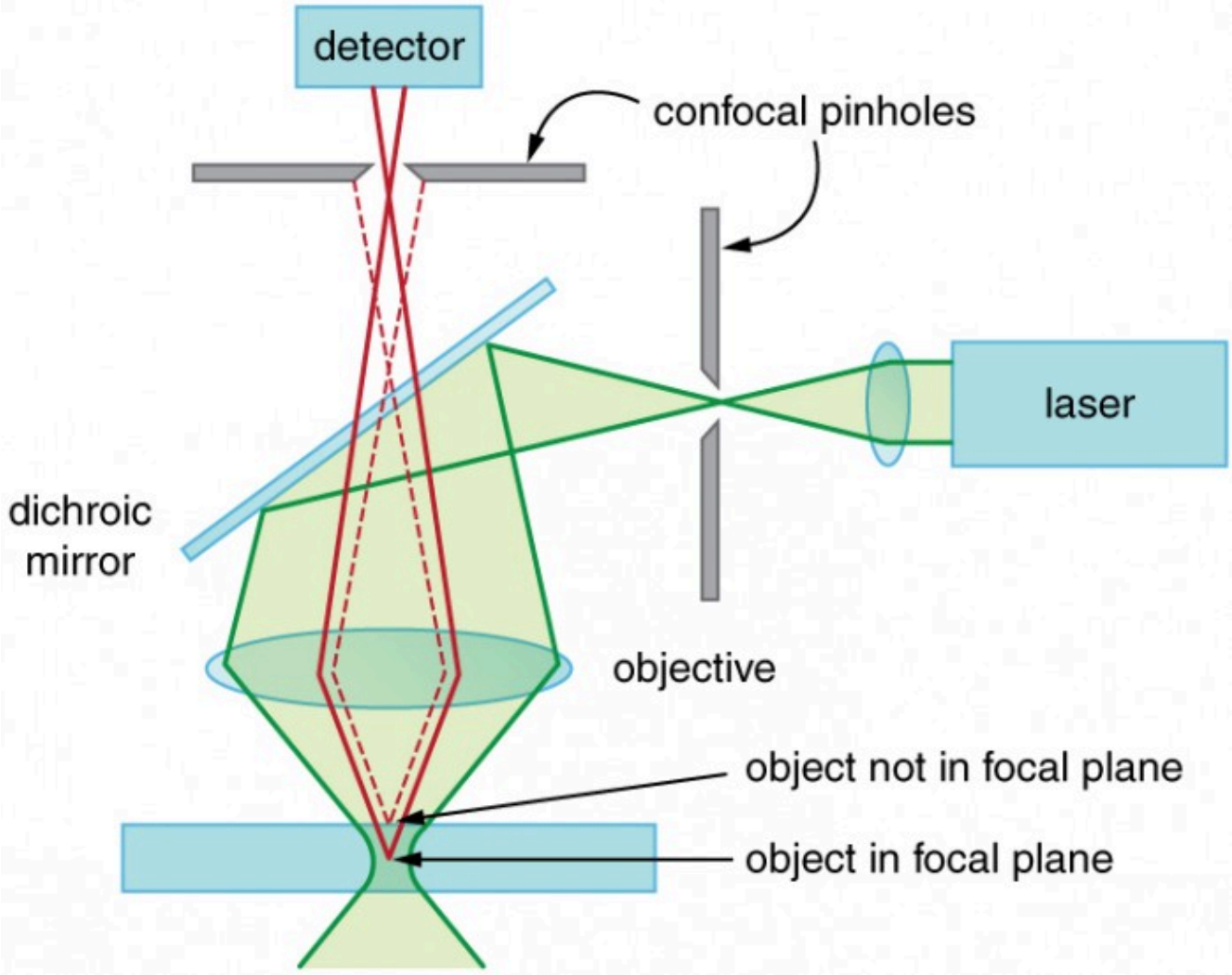


Figure 5. A confocal microscope provides three-dimensional images using pinholes and the extended depth of focus as described by wave optics. The right pinhole illuminates a tiny region of the sample in the focal plane. In-focus light rays from this tiny region pass through the dichroic mirror and the second pinhole to a detector and a computer. Out-of-focus light rays are blocked. The pinhole is scanned sideways to form an image of the entire focal plane. The pinhole can then be scanned up and down to gather images from different focal planes. The result is a three-dimensional image of the specimen.

Section Summary

- To improve microscope images, various techniques utilizing the wave characteristics of light have been developed. Many of these enhance contrast with interference effects.

Conceptual Questions

1. Explain how microscopes can use wave optics to improve contrast and why this is important.
2. A bright white light under water is collimated and directed upon a prism. What range of colors does one see emerging?

Glossary

confocal microscopes: microscopes that use the extended focal region to obtain three-dimensional images rather than two-dimensional images

contrast: the difference in intensity between objects and the background on which they are observed

interference microscopes: microscopes that enhance contrast between objects and background by superimposing a reference beam of light upon the light emerging from the sample

phase-contrast microscope: microscope utilizing wave interference and differences in phases to enhance contrast

polarization microscope: microscope that enhances contrast by utilizing a wave characteristic of light, useful for objects that are optically active

ultraviolet (UV) microscopes: microscopes constructed with special lenses that transmit UV rays and utilize photographic or electronic techniques to record images

Appendix

Appendix A. Useful Information

Lumen Learning

- Table 1, Important Constants
- Table 2, Submicroscopic Masses
- Table 3, Solar System Data
- Table 4, Metric Prefixes for Powers of Ten and Their Symbols
- Table 5, The Greek Alphabet
- Table 6, SI units
- Table 7, Selected British Units
- Table 8, Other Units
- Table 9, Useful Formulae

Table 1. Important Constants[1]

Symbol	Meaning	Best Value	Approximate Value
c	Speed of light in vacuum	2.99792458×10^{8} m/s	3.00×10^{8} m/s
G	Gravitational constant	$6.67384(80) \times 10^{-11}$ N · m^2/kg^2	6.67×10^{-11} N · m^2/kg^2
N_A	Avogadro's number	$6.02214129(27) \times 10^{23}$	6.02×10^{23}
k	Boltzmann's constant	$1.3806488(13) \times 10^{-23}$ J/K	1.38×10^{-23} J/K
R	Gas constant	8.3144621(75) J/mol · K	8.31 J/mol · K = 1.99 cal/mol · K = 0.0821 atm · L/mol · K
σ	Stefan-Boltzmann constant	$5.670373(21) \times 10^{-8}$ W/ m^2 · K	5.67×10^{-8} W/m^2 · K
k	Coulomb force constant	$8.987551788\ldots \times 10^{9}$ N · m^2/C^2	8.99×10^{9} N · m^2/C^2
q_e	Charge on electron	$-1.602176565(35) \times 10^{-19}$ C	-1.60×10^{-19} C
ε_0	Permittivity of free space	$8.854187817\ldots \times 10^{-12}$ C^2/N · m^2	8.85×10^{-12} C^2/N · m^2
μ_0	Permeability of free space	$4\pi \times 10^{-7}$ T · m/A	1.26×10^{-6} T · m/A
h	Planck's constant	$6.62606957(29) \times 10^{-34}$ J · s	6.63×10^{-34} J · s

Return to Top

Table 2. Submicroscopic Masses[2]

Symbol	Meaning	Best Value	Approximate Value
m_e	Electron mass	$9.10938291(40) \times 10^{-31}$ kg	9.11×10^{-31} kg
m_p	Proton mass	$1.672621777(74) \times 10^{-27}$ kg	1.6726×10^{-27} kg
m_n	Neutron mass	$1.674927351(74) \times 10^{-27}$ kg	1.6749×10^{-27} kg
u	Atomic mass unit	$1.660538921(73) \times 10^{-27}$ kg	1.6605×10^{-27} kg

Return to Top

1. Stated values are according to the National Institute of Standards and Technology Reference on Constants, Units, and Uncertainty, www.physics.nist.gov/cuu (accessed May 18, 2012). Values in parentheses are the uncertainties in the last digits. Numbers without uncertainties are exact as defined.
2. Stated values are according to the National Institute of Standards and Technology Reference on Constants, Units, and Uncertainty, www.physics.nist.gov/cuu (accessed May 18, 2012). Values in parentheses are the uncertainties in the last digits. Numbers without uncertainties are exact as defined.

Table 3. Solar System Data

Sun	mass	1.99×10^{30} kg
	average radius	6.96×10^{8} m
	Earth-sun distance (average)	1.496×10^{11} m
Earth	mass	5.9736×10^{24} kg
	average radius	6.376×10^{6} m
	orbital period	3.16×10^{7} s
Moon	mass	7.35×10^{22} kg
	average radius	1.74×10^{6} m
	orbital period (average)	2.36×10^{6} s
	Earth-moon distance (average)	3.84×10^{8} m

Return to Top

Table 4. Metric Prefixes for Powers of Ten and Their Symbols

Prefix	Symbol	Value	Prefix	Symbol	Value
tera	T	10^{12}	deci	d	10^{-1}
giga	G	10^{9}	centi	c	10^{-2}
mega	M	10^{6}	milli	m	10^{-3}
kilo	k	10^{3}	micro	μ	10^{-6}
hecto	h	10^{2}	nano	n	10^{-9}
deka	da	10^{1}	pico	p	10^{-12}
—	—	10^{0} (= 1)	femto	f	10^{-15}

Return to Top

Table 5. The Greek Alphabet

Alpha	A	α	Eta	H	η	Nu	N	ν
Beta	B	β	Tau	T	τ	Theta	Θ	θ
Xi	Ξ	ξ	Upsilon	Y	υ	Gamma	Γ	γ
Iota	I	ι	Omicron	O	ο	Phi	φ	ϕ
Delta	Δ	δ	Kappa	K	κ	Pi	Π	π
Chi	X	χ	Epsilon	E	ε	Lambda	Λ	λ
Rho	P	ρ	Psi	Ψ	ψ	Zeta	Z	ζ
Mu	M	μ	Sigma	Σ	σ	Omega	Ω	ω

Return to Top

Table 6. SI Units

	Entity	**Abbreviation**	**Name**
Fundamental units	Length	m	meter
	Mass	kg	kilogram
	Time	s	second
	Current	A	ampere
Supplementary unit	Angle	rad	radian
Derived units	Force	$N = kg \cdot m/s^2$	newton
	Energy	$J = kg \cdot m^2/s^2$	joule
	Power	$W = J/s$	watt
	Pressure	$Pa = N/m^2$	pascal
	Frequency	$Hz = 1/s$	hertz
	Electronic potential	$V = J/C$	volt
	Capacitance	$F = C/V$	farad
	Charge	$C = s \cdot A$	coulomb
	Resistance	$\Omega = V/A$	ohm
	Magnetic field	$T = N/(A \cdot m)$	tesla
	Nuclear decay rate	$Bq = 1/s$	becquerel

Return to Top

Table 7. Selected British Units

	1 inch (in.) = 2.54 cm (exactly)
Length	1 foot (ft) = 0.3048 m
	1 mile (mi) = 1.609 km
Force	1 pound (lb) = 4.448 N
Energy	1 British thermal unit (Btu) = 1.055×10^3 J
Power	1 horsepower (hp) = 746 W
Pressure	1 lb/in^2 = 6.895×10^3 Pa

Return to Top

Table 8. Other Units

Length	1 light year (ly) = 9.46×10^{15} m
	1 astronomical unit (au) = 1.50×10^{11} m
	1 nautical mile = 1.852 km
	1 angstrom (Å) = 10^{-10} m
Area	1 acre (ac) = 4.05×10^{3} m^2
	1 square foot (ft^2) = 9.29×10^{-2} m^2
	1 barn (b) = 10^{-28} m^2
Volume	1 liter (L) = 10^{-3} m^3
	1 U.S. gallon (gal) = 3.785×10^{-3} m^3
Mass	1 solar mass = 1.99×10^{30} kg
	1metric ton = 10^{3} kg
	1 atomic mass unit (u) = 1.6605×10^{-27} kg
Time	1 year (y) = 3.16×10^{7} s
	1 day (d) = 86,400 s
Speed	1 mile per hour (mph) = 1.609 km/h
	1 nautical mile per hour (naut) = 1.852 km/h
Angle	1 degree (°) = 1.745×10^{-2} rad
	1 minute of arc (′) = 1/60 degree
	1 second of arc (″) = 1/60 minute of arc
	1 grad = 1.571×10^{-2} rad
Energy	1 kiloton TNT (kT) = 4.2×10^{12} J
	1 kilowatt hour (kW · h) = 3.60×10^{6} J
	1 food calorie (kcal) = 4186 J
	1 calorie (cal) = 4.186 J
	1 electron volt (eV) = 1.60×10^{-19} J
Pressure	1 atmosphere (atm) = 1.013×10^{5} Pa
	1 millimeter of mercury (mm Hg) = 133.3 Pa
	1 torricelli (torr) = 1 mm Hg = 133.3 Pa
Nuclear decay rate	1 curie (Ci) = 3.70×10^{10} Bq

Return to Top

Table 9. Useful Formulae

Circumference of a circle with radius *r* or diameter *d*	$C = 2\pi r = \pi d$
Area of a circle with radius *r* or diameter *d*	$A = \pi r^2 = \frac{\pi d^2}{4}$
Area of a sphere with radius *r*	$A = 4\pi r^2$
Volume of a sphere with radius *r*	$V = \frac{4}{3}\left(\pi r^3\right)$

Return to Top

Appendix B. Glossary of Key Symbols and Notation

Lumen Learning

In this glossary, key symbols and notation are briefly defined.

Symbol	Definition
$\overline{\text{any symbol}}$	average (indicated by a bar over a symbol—e.g., $\overline{v}$ is average velocity)
°C	Celsius degree
°F	Fahrenheit degree
//	parallel
⊥	perpendicular
$\propto$	proportional to
±	plus or minus
0	zero as a subscript denotes an initial value
α	alpha rays
α	angular acceleration
α	temperature coefficient(s) of resistivity
β	beta rays
β	sound level
β	volume coefficient of expansion
β^-	electron emitted in nuclear beta decay
β^+	positron decay
γ	gamma rays
γ	surface tension
$\gamma = \frac{1}{\sqrt{1-\frac{v^2}{c^2}}}$	a constant used in relativity
Δ	change in whatever quantity follows
δ	uncertainty in whatever quantity follows
Δ*E*	change in energy between the initial and final orbits of an electron in an atom
Δ*E*	uncertainty in energy
Δ*m*	difference in mass between initial and final products
Δ*N*	number of decays that occur
Δ*p*	change in momentum
Δ*p*	uncertainty in momentum

Symbol	Definition
ΔPE_g	change in gravitational potential energy
$\Delta\theta$	rotation angle
Δs	distance traveled along a circular path
Δt	uncertainty in time
Δt_0	proper time as measured by an observer at rest relative to the process
ΔV	potential difference
Δx	uncertainty in position
ε_0	permittivity of free space
η	viscosity
θ	angle between the force vector and the displacement vector
θ	angle between two lines
θ	contact angle
θ	direction of the resultant
θ_b	Brewster's angle
θ_c	critical angle
κ	dielectric constant
λ	decay constant of a nuclide
λ	wavelength
λ_n	wavelength in a medium
μ_0	permeability of free space
μ_k	coefficient of kinetic friction
μ_s	coefficient of static friction
ν_e	electron neutrino
π^+	positive pion
π^-	negative pion
π^0	neutral pion
ρ	density
ρ_c	critical density, the density needed to just halt universal expansion
ρ_{fl}	fluid density

Symbol	Definition
$\overline{\rho}_{\mathrm{obj}}$	average density of an object
$\frac{\rho}{\rho_{\mathrm{W}}}$	specific gravity
τ	characteristic time constant for a resistance and inductance (*RL*) or resistance and capacitance (*RC*) circuit
τ	characteristic time for a resistor and capacitor (*RC*) circuit
τ	torque
Y	upsilon meson
Φ	magnetic flux
ϕ	phase angle
Ω	ohm (unit)
ω	angular velocity
A	ampere (current unit)
A	area
A	cross-sectional area
A	total number of nucleons
a	acceleration
a_{B}	Bohr radius
a_{c}	centripetal acceleration
a_{t}	tangential acceleration
AC	alternating current
AM	amplitude modulation
atm	atmosphere
B	baryon number
B	blue quark color
$\overline{B}$	antiblue (yellow) antiquark color
b	quark flavor bottom or beauty
B	bulk modulus
B	magnetic field strength

Symbol	Definition
B_{int}	electron's intrinsic magnetic field
B_{orb}	orbital magnetic field
BE	binding energy of a nucleus—it is the energy required to completely disassemble it into separate protons and neutrons
$\frac{\text{BE}}{A}$	binding energy per nucleon
Bq	becquerel—one decay per second
C	capacitance (amount of charge stored per volt)
C	coulomb (a fundamental SI unit of charge)
C_{p}	total capacitance in parallel
C_{s}	total capacitance in series
CG	center of gravity
CM	center of mass
c	quark flavor charm
c	specific heat
c	speed of light
Cal	kilocalorie
cal	calorie
COP_{hp}	heat pump's coefficient of performance
COP_{ref}	coefficient of performance for refrigerators and air conditioners
$\cos\theta$	cosine
$\cot\theta$	cotangent
$\csc\theta$	cosecant
D	diffusion constant
d	displacement
d	quark flavor down
dB	decibel
d_{i}	distance of an image from the center of a lens
d_{o}	distance of an object from the center of a lens
DC	direct current

Symbol	Definition
E	electric field strength
ε	emf (voltage) or Hall electromotive force
emf	electromotive force
E	energy of a single photon
E	nuclear reaction energy
E	relativistic total energy
E	total energy
E_0	ground state energy for hydrogen
E_0	rest energy
EC	electron capture
E_{cap}	energy stored in a capacitor
Eff	efficiency—the useful work output divided by the energy input
Eff_C	Carnot efficiency
E_{in}	energy consumed (food digested in humans)
E_{ind}	energy stored in an inductor
E_{out}	energy output
e	emissivity of an object
e^+	antielectron or positron
eV	electron volt
F	farad (unit of capacitance, a coulomb per volt)
F	focal point of a lens
$\mathbf{F}$	force
F	magnitude of a force
F	restoring force
F_B	buoyant force
F_c	centripetal force
F_i	force input
$\mathbf{F}_{net}$	net force
F_o	force output
FM	frequency modulation

Symbol	Definition
f	focal length
f	frequency
f_0	resonant frequency of a resistance, inductance, and capacitance (*RLC*) series circuit
f_0	threshold frequency for a particular material (photoelectric effect)
f_1	fundamental
f_2	first overtone
f_3	second overtone
f_B	beat frequency
f_k	magnitude of kinetic friction
f_s	magnitude of static friction
G	gravitational constant
G	green quark color
$\overline{G}$	antigreen (magenta) antiquark color
g	acceleration due to gravity
g	gluons (carrier particles for strong nuclear force)
h	change in vertical position
h	height above some reference point
h	maximum height of a projectile
h	Planck's constant
hf	photon energy
h_i	height of the image
h_o	height of the object
I	electric current
I	intensity
I	intensity of a transmitted wave
I	moment of inertia (also called rotational inertia)
I_0	intensity of a polarized wave before passing through a filter
I_{ave}	average intensity for a continuous sinusoidal electromagnetic wave
I_{rms}	average current

Symbol	Definition
J	joule
$\frac{\mathrm{J}}{\Psi}$	Joules/psi meson
K	kelvin
k	Boltzmann constant
k	force constant of a spring
K_{α}	x rays created when an electron falls into an $n = 1$ shell vacancy from the $n = 3$ shell
K_{β}	x rays created when an electron falls into an $n = 2$ shell vacancy from the $n = 3$ shell
kcal	kilocalorie
KE	translational kinetic energy
KE + PE	mechanical energy
$\mathrm{KE_e}$	kinetic energy of an ejected electron
$\mathrm{KE_{rel}}$	relativistic kinetic energy
$\mathrm{KE_{rot}}$	rotational kinetic energy
$\overline{\mathrm{KE}}$	thermal energy
kg	kilogram (a fundamental SI unit of mass)
L	angular momentum
L	liter
L	magnitude of angular momentum
L	self-inductance
ℓ	angular momentum quantum number
L_{α}	x rays created when an electron falls into an $n = 2$ shell from the $n = 3$ shell
L_{e}	electron total family number
L_{μ}	muon family total number
L_{τ}	tau family total number
L_{f}	heat of fusion
L_{f} and L_{v}	latent heat coefficients
L_{orb}	orbital angular momentum
L_{s}	heat of sublimation

Symbol	Definition
L_v	heat of vaporization
L_z	*z*-component of the angular momentum
M	angular magnification
M	mutual inductance
m	indicates metastable state
m	magnification
m	mass
m	mass of an object as measured by a person at rest relative to the object
m	meter (a fundamental SI unit of length)
m	order of interference
m	overall magnification (product of the individual magnifications)
$m\left({}^{A}\mathrm{X}\right)$	atomic mass of a nuclide
MA	mechanical advantage
m_e	magnification of the eyepiece
m_e	mass of the electron
m_ℓ	angular momentum projection quantum number
m_n	mass of a neutron
m_o	magnification of the objective lens
mol	mole
m_p	mass of a proton
m_s	spin projection quantum number
N	magnitude of the normal force
N	newton
N	normal force
N	number of neutrons
n	index of refraction
n	number of free charges per unit volume
N_A	Avogadro’s number
N_r	Reynolds number

Symbol	Definition
N · m	newton-meter (work-energy unit)
N · m	newtons times meters (SI unit of torque)
OE	other energy
P	power
P	power of a lens
P	pressure
$\mathbf{p}$	momentum
p	momentum magnitude
p	relativistic momentum
$\mathbf{p}_{\text{tot}}$	total momentum
$\mathbf{p}_{\text{tot}}$	total momentum some time later
P_{abs}	absolute pressure
P_{atm}	atmospheric pressure
P_{atm}	standard atmospheric pressure
PE	potential energy
PE_{el}	elastic potential energy
PE_{elec}	electric potential energy
PE_{s}	potential energy of a spring
P_{g}	gauge pressure
P_{in}	power consumption or input
P_{out}	useful power output going into useful work or a desired, form of energy
Q	latent heat
Q	net heat transferred into a system
Q	flow rate—volume per unit time flowing past a point
$+Q$	positive charge
$-Q$	negative charge
q	electron charge
q_{p}	charge of a proton
q	test charge
QF	quality factor

Symbol	Definition
R	activity, the rate of decay
R	radius of curvature of a spherical mirror
R	red quark color
$\overline{R}$	antired (cyan) quark color
R	resistance
R	resultant or total displacement
R	Rydberg constant
R	universal gas constant
r	distance from pivot point to the point where a force is applied
r	internal resistance
$r\perp$	perpendicular lever arm
r	radius of a nucleus
r	radius of curvature
r	resistivity
r or rad	radiation dose unit
rem	roentgen equivalent man
rad	radian
RBE	relative biological effectiveness
RC	resistor and capacitor circuit
rms	root mean square
r_n	radius of the nth H-atom orbit
R_p	total resistance of a parallel connection
R_s	total resistance of a series connection
R_s	Schwarzschild radius
S	entropy
S	intrinsic spin (intrinsic angular momentum)
S	magnitude of the intrinsic (internal) spin angular momentum
S	shear modulus
S	strangeness quantum number

Symbol	Definition
s	quark flavor strange
s	second (fundamental SI unit of time)
s	spin quantum number
$\mathbf{s}$	total displacement
$\sec\theta$	secant
$\sin\theta$	sine
s_z	z-component of spin angular momentum
T	period—time to complete one oscillation
T	temperature
T_c	critical temperature—temperature below which a material becomes a superconductor
T	tension
T	tesla (magnetic field strength B)
t	quark flavor top or truth
t	time
$t_{1/2}$	half-life—the time in which half of the original nuclei decay
$\tan\theta$	tangent
U	internal energy
u	quark flavor up
u	unified atomic mass unit
$\mathbf{u}$	velocity of an object relative to an observer
$\mathbf{u}'$	velocity relative to another observer
V	electric potential
V	terminal voltage
V	volt (unit)
V	volume
$\mathbf{v}$	relative velocity between two observers
v	speed of light in a material
$\mathbf{v}$	velocity
$\overline{\mathbf{V}}$	average fluid velocity

Symbol	Definition
$V_B - V_A$	change in potential
$\mathbf{v}_d$	drift velocity
V_p	transformer input voltage
V_{rms}	rms voltage
V_s	transformer output voltage
$\mathbf{v}_{tot}$	total velocity
v_w	propagation speed of sound or other wave
$\mathbf{v}_w$	wave velocity
W	work
W	net work done by a system
W	watt
w	weight
w_{fl}	weight of the fluid displaced by an object
W_c	total work done by all conservative forces
W_{nc}	total work done by all nonconservative forces
W_{out}	useful work output
X	amplitude
X	symbol for an element
${}^{A}_{Z}\mathrm{X}_{N}$	notation for a particular nuclide
x	deformation or displacement from equilibrium
x	displacement of a spring from its undeformed position
x	horizontal axis
X_C	capacitive reactance
X_L	inductive reactance
x_{rms}	root mean square diffusion distance
y	vertical axis
Y	elastic modulus or Young’s modulus
Z	atomic number (number of protons in a nucleus)
Z	impedance

Version History

Creation & Versioning History

NSCC Fundamentals of Heat, Light & Sound Chapter Mapping		
	Lumen Learning Physics I	**Lumen Learning Physics II**
Introduction		How to Succeed in Physics Guide
Chapter 1	I. The Nature of Science and Physics	
Chapter 2	XI. Fluid Statics	
Chapter 3	XIII. Temperature, Kinetic Theory, and Gas Laws	
Chapter 4	XIV. Heat and Heat Transfer Methods	
Chapter 5	XV. Thermodynamics	
Chapter 6	XVI. Oscillatory Motion and Waves	
Chapter 7	XVII. Introduction to the Physics of Hearing	
Chapter 8		VIII. Electromagnetic Waves
Chapter 9		IX. Geometric Optics
Chapter 10		X. Vision and Optical Instruments
Chapter 11		XI. Wave Optics
Appendix A	Appendix C. Useful Information	
Appendix B	Appendix D. Glossary of Key Symbols and Notations	

College Physics by OpenStax. Located at: http://cnx.org/contents/031da8d3-b525-429c-80cf-6c8ed997733a/College_Physics. CC BY Licence