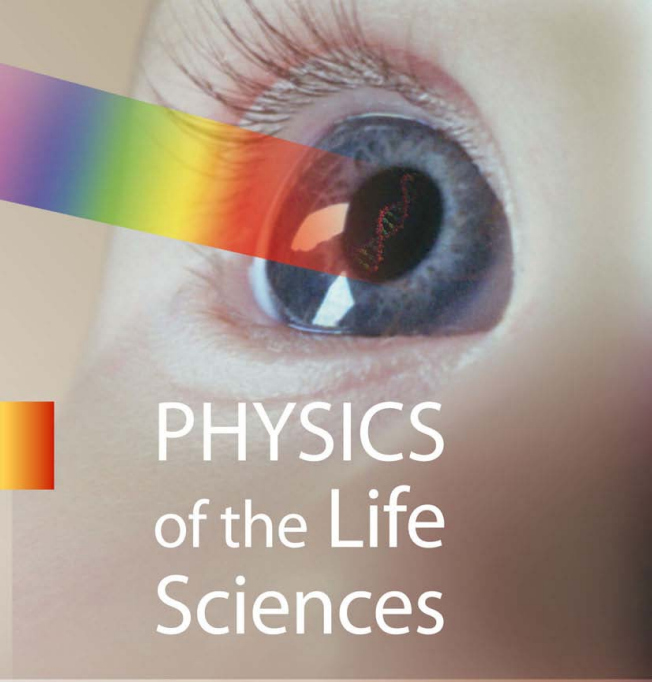


PHYSICS of the Life Sciences

Jay Newman

Springer

Physics of the Life Sciences

The image adapted for the cover was Highly Commended in the "Wellcome Trust/BBC HOW IS SCIENCE CHANGING US? 'Imagine' Photography Competition 2005" and is used by permission of the photographer, Ivan Burn.

Physics of the Life Sciences

Jay Newman

Springer

Jay Newman
Union College
Department of Physics and Astronomy
Schenectady, NY 12308, USA

ISBN: 978-0-387-77258-5 e-ISBN: 978-0-387-77259-2
DOI: 10.1007/978-0-387-77259-2

Library of Congress Control Number: 2008929543

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

About the Author

Jay Newman is the R. Gordon Gould Professor of Physics at Union College where he has taught for 30 years. While studying for his PhD in physics at New York University, he developed a keen interest in biophysics and did a three-year postdoctoral fellowship in the Biophysics Department of Johns Hopkins University. Since joining the faculty at Union College, Professor Newman has taught and developed more than 15 different courses, led student terms abroad in science research in Italy, and also spent a year at Stanford University. The experiences abroad with students stemmed from his previous stays as a Visiting Professor in Italy, once in Pavia and six times in Palermo.

His research has been on the structure, dynamics and interactions of biomolecules using laser light scattering and other physical methods. He has 60 publications, many co-authored with some of the 30 plus undergraduate students who have done research projects in his laboratory, and has received two grants from the Research Corporation and five multiyear grants from the National Science Foundation for both research and teaching.

About 15 years ago, he developed a special introductory physics course for life science students at Union College, which was the basis for this text. The idea behind the course and this book is to show the essential connections between physics and modern life sciences. Motivating this new approach to an introductory course was Professor Newman's firm belief, developed over his early training and now reinforced by almost daily news reports, that modern biology and medicine are becoming ever more quantitative and dependent on an understanding of physics fundamentals, methodology, technology, and modes of thinking. Building this bridge is the purpose and goal of *Physics of the Life Sciences*.

Preface

This textbook has its origins in a course that I began developing at Union College in the mid-1980s to teach physics to life science students in a way that would interest them and show the connections of fundamental physics to modern biology and medicine. From my own research experiences and interests in biophysics, I know that almost all areas of modern life sciences integrally involve physics in both experimental techniques and in basic understanding of process or function. However, I and many colleagues with whom I have spoken have been unhappy over the years with published attempts to direct a textbook to this audience. Most such texts are watered down engineering physics books with occasional added sections on related biology topics that are easy to skip over or assign students to read on their own.

As I set out to write this textbook, I had certain definite goals in mind. I wanted to write a book that was truly directed at life science students, one that integrated modern biology, biophysics, and medical techniques into the presentation of the material. Believing in the *less is more* credo, I chose to omit certain standard topics that are usually included in texts for this audience, while expanding on topics that have more relevance to the life and biomedical sciences. From my experience teaching to these students, I also wanted a book that would be shorter and could be fully covered in a two-semester course. Although students at Union College and comparable institutions taking this introductory course have all had some calculus, only algebra and trigonometry are used in the main body of the text. At this level, I believe that calculus adds little to the understanding of the material and can detract from focusing on the basic physical ideas. However, I have sprinkled in optional boxed calculations that do use some calculus where I felt they truly added to the discussion (averaging less than one box per chapter). These "sidebars" can be omitted without any loss of continuity.

The order of topics for this text follows a more or less traditional sequence. An exception to this is the presentation of one-dimensional mechanics through forces and energy before introducing vectors and generalizing to motion in more than one dimension. This allows students to focus on the physics concepts of kinematics, forces, and energy without being distracted by the ideas of vector analysis.

Beyond the order of topics, the presentation of material is unique in that, wherever possible, themes from biology or medicine are used to present the physics material. The material speaks to life science students. Rather than optional sections at the end of occasional chapters, life science themes are plentiful and integral to the text. The role of these topics here is more fundamental, as can be gleaned from a list of some examples.

- The early introduction of diffusion as an example of motion (full section in Chapter 2).
- The early introduction of motion in a viscous fluid as an example of one-dimensional motion, development of Hooke's law and elasticity with applications to biomaterials and viscoelasticity, protein structure, and molecular dynamics calculations (all in Chapter 3).
- Discussion of centrifugation in Chapter 5.

- Examples of rotational motion kinematics of a bacteria and of a rotary motor protein, the atomic force microscope, rotational diffusion, and cell membrane dynamics (all in Chapter 7).
- A chapter (9) on viscous fluids with discussions of blood, other complex fluids, the human circulatory system, surface tension, and capillarity.
- A chapter (11) on sound with extensive discussions on the ear and on ultrasound.
- A chapter (13) with a molecular discussion of entropy, a section on Gibbs free energy, a section on biological applications of statistical thermodynamics, and a section on biological applications of nonlinear dynamics.
- Chapters (14–15) on electric forces, fields, and energy with sections on electrophoresis, macromolecular charges in solution, modern electrophoresis methods, electrostatic applications to native and synthetic macromolecules, an introduction to capacitors entirely through a discussion of cell membranes, and sections on membrane channels and electric potential mapping of the human body: heart, muscle, and brain.
- A chapter (16) on electric current and cell membranes covering circuits through membrane models: included are sections on membrane electrical currents, an overview of nerve structure and function including measurement techniques such as patch-clamping, the electrical properties of neurons, and a second section on membrane channels with a discussion of single-channel recording.
- Chapters on electromagnetic induction and waves (18–19) that include discussion of MEG (magnetoencephalography) using SQUIDs, an entire section on NMR, and sections on magnetic resonance imaging, laser tweezers, the quantum theory of radiation concepts (revisited later), and the interaction of radiation with matter, the last a primer on spectroscopy, including absorption spectroscopy, scattering, and fluorescence.
- Four chapters (20–23) on optics include a section on optical fibers and their applications in medicine, a section on the human eye, sections on the new light microscopies (dark field, fluorescence, phase contrast, DIC, confocal and multiphoton methods), discussion of polarization in biology, including birefringence and dichroism techniques, and sections on the transmission electron microscope, scanning EM and scanning transmission EM, and x-rays and computed tomography (CT) methods.
- Three chapters (24–26) on modern physics (many of these ideas have been introduced and used throughout the book) include discussions of the scanning tunneling microscope, a section on the laser and its applications in biology and medicine, including holography. The chapter on nuclear physics and medical applications (26) includes sections on dosimetry and biological effects of radiation, radioisotopes, and nuclear medicine, and the medical imaging methods SPECT (single photon emission computer tomography) and PET (positron emission tomography).

As mentioned above, we've chosen to omit some standard topics that are either not central to the life science themes or that students find very opaque. Omitted are such topics as Kepler's laws, heat engines, induction and LR/LRC circuits, AC circuits, special relativity kinematics, particle physics, and astrophysics; Gauss's law and Ampere's law are presented in optional sections at the end of appropriate chapters.

Each chapter contains three types of learning aides for the student: open-ended questions, multiple-choice questions, and quantitative problems. In about 60 of these per chapter, we have tried to include a wide selection related to the life sciences. Complete solutions to all of the multiple choice and other problems are available to instructors. There are also a number of worked examples in the chapters, averaging over six per chapter, and about 900 photos and line drawings to illustrate concepts in the text, with many in full color.

Jay Newman
Schenectady, NY

Acknowledgments

First I'd like to thank the American Institute of Physics Press and both Maria Taylor and Elias Greenbaum for suggesting this project and providing a grant that gave me most of a year free from teaching to start writing this textbook. I've benefited greatly from collaborations with three colleagues on its development. David Peak, a former Union College colleague now at Utah State, edited portions of the manuscript and made many suggestions on the presentation of the material. Larry Brehm, formerly at IBM and now at SUNY Potsdam, contributed to the end-of-chapter problems for a number of the chapters, particularly in the mechanics and optics portions. Scott LaBrake, at Union College, checked and solved all the problems in the book, and wrote the solutions manual, as well as taught from preliminary editions of the book.

Thanks also to the many students who learned their introductory physics from preliminary versions of this text and put up with typographical errors and occasional unsolvable problems. Some of these students worked through essentially all the problems in the book helping to find errors as well. I also thank my colleagues at Union College for their interest and support in this project and for numerous discussions about physics pedagogy.

The staff at Springer, including David Packer, Anushka Hosain, and all the production team, has been most helpful in seeing this project to fruition.

Finally, I thank my extended family for their support, encouragement, and love over these many years, especially my wife, Maia.

Contents

List of Tables

1 Introduction

1. SCIENCE, PHYSICS, AND BIOLOGY

If one examines the course catalog of a large, contemporary, university in the United States for fields of instruction in science, one can find such titles as animal science, astronomy, atmospheric science, biochemistry, biology, botany, chemistry, computer science, geology, ecology, mathematical science, meteorology, physics, psychology, toxicology, and zoology, to name but a few. Each of these is a field of study in its own right consisting of many subtopics. On the other hand, a catalog from a U.S. college that existed in the early 19th century probably would show at most only two "sciences": natural history (the progenitor of geology and biology) and natural philosophy (physics and chemistry). Over the years, there has been an explosion of speciation in science, resulting in what appears at first sight to be a technological Tower of Babel.

Although the factual content of the many branches of modern science may serve to differentiate one from the other, all branches share certain common characteristics and concepts. Most important, all of the sciences share a way of thinking. Science is a search for truth predicated on the belief that there is an absolute physical reality; things aren't just figments of our imaginations. Science is based on observation. Unlike the observations of creative art or religion, for example, which tend to be private and highly personal, scientific observations are made, as best as can be done, in a public way, that is, in a way that anyone, in principle, could repeat them.

Scientific truth is couched in *models*. A model is not the thing itself, but a representation of the thing, much like a metaphor. A model is a guess about how the thing works based on a set of empirical data. (If the dataset is very large and the model appears to be especially useful, it is called a *theory*. In science, the colloquially pejorative phrase, "That's only a theory," would never be used because in science a theory is the best kind of guess one can have.) A model can be physical or pictorial or verbal. Often in science, models are mathematical. Mathematics is an incredibly economical way of expressing an idea. One equation can encapsulate tomes of empirical data. Better yet, an equation can be used to predict outcomes of experiments performed under conditions never seen before. In fact, prediction is the heart of science. Science is a relentless series of predictions designed to identify the limitations of previously established "truths." By tearing down and supplanting prior knowledge, science aspires to produce an ever-clearer picture of physical reality. In this sense, science can be said to be an insatiable pursuit of *provisional* truth.

Physics is the most elemental of all the sciences. It attempts to explain the most fundamental phenomena with the fewest assumptions and in the simplest terms. In a sense, physics strives to identify and attack the "easiest" of nature's problems. Despite its pursuit of the fundamental, however, physics has been extraordinarily successful in understanding a vast array of practically important questions such as how

J. Newman, *Physics of the Life Sciences,* DOI: 10.1007/978-0-387-77259-2_1,

to build a better steam engine, how to place a satellite in orbit, and how energy stored in atomic nuclei can be used to light cities, to cite just a few examples. Indeed, physics is the basis for a huge portion of the world's economy.

The subfields of physics bear such names as classical mechanics, thermodynamics, electricity and magnetism, optics, relativity, and quantum mechanics. Of these, classical mechanics is usually studied first because it deals with the ideas of mass, motion, force, and energy, concepts that underlie not only the other areas of physics, but also astronomy, biology, chemistry, and geology, as well as all of engineering.

Like physics, biology is a study of matter and energy. The systems of matter and energy that are of biological interest, however, are vastly more complex than those that are the focus of physics. Biology deals with *living* matter, collections of atoms and molecules that manage to harness energy to perform such extraordinary tasks as locomotion, reproduction, and computation ("thinking"). On the most primitive, microscopic level, the rules obeyed by living matter are just the fundamental laws of physics. These, as far as we can tell, are immutable. They have persisted since the origin of the universe. On a higher level of organization, however, at the level of cells and organisms, living matter obeys rules that can change. Mutation and evolution are the cornerstones of biological diversity. How the immutable, microscopic rules of physics are knit together into the macroscopic fabric of life, where matter is capable of adaptive and evolving behavior, is one of the great unsolved mysteries of contemporary scientific inquiry.

Until the 1950s or so, relatively few direct connections between physics and biology had been recognized. Up to that point, most research in biology had been descriptive, a kind of cataloging of similarities and differences. Since then, strong linkages between biology and physics have emerged. These connections have revolutionized our understanding of how life works and led to profound improvements in pharmaceuticals and clinical procedures. The impact of physics on modern biology and medical science is due, in part, to the introduction of new technologies used to study biological systems and, in part, to direct applications of physics to the detailed understanding of macromolecular processes.

Examples of new technology based on physics and used in the study of biology and medicine abound. A huge array of new microscopies (transmission electron, scanning electron, fluorescence, interference, polarization, scanning tunneling, atomic force) and spectroscopies (nuclear magnetic resonance (NMR), electron spin resonance (ESR), x-ray, neutron, and many laser-based methods such as Raman scattering) have been developed and are now routinely used to study macromolecular structure and functioning. New methods in electromagnetic sensing (e.g., superconducting quantum interference devices (SQUIDS) for measuring extremely small magnetic fields, such as those due to nerve activity, and single-membrane channel recording of electrical activities), laser and electronic instrumentation to better image events both spatially and temporally (allowing studies of extremely fast kinetics, down to 10^{-14} s, and submillimeter spatial resolutions using ultrasound, x-rays, or magnetic resonance methods), and, of course, dramatic improvements in computers, made possible by new physics, have all led to major advances in our knowledge.

In conjunction with this technological progress, has come a marked increase in the description of biological processes using fundamental physics. Detailed molecular models of the structure and functioning of many significant biological processes are now in hand. Most of this progress has been at the subcellular or single-cell level but even areas of biology involving cell–cell interactions, functioning of entire organs, developmental biology, physiology, and the ecology of plant and animal communities are now being approached with physical models and fundamental physics approaches. The rate at which new ideas in physics find application in biology is astonishing. Recent developments in nonlinear dynamics in physics, for example, have already been applied to a large variety of complex biological systems, especially in understanding how electrical activity in the heart and brain changes from health to a state of disease.

To summarize, it is fair to say that no student of today's life sciences will be adequately educated without a firm understanding of the fundamental principles of physical

science. It is to that aspect of the life scientist's education that the remainder of this book is dedicated.

2. PLAN OF THIS BOOK

Physics of the Life Sciences is designed to teach fundamental physics to students of the life sciences. Our approach is to use modern biophysical themes as much as possible to introduce the physics and to illustrate the wide variety of applications of physics in the life sciences. Indeed today's doctors, scientists, nurses, and medical and health technicians constantly use a vast array of modern technology in their work. A working knowledge of these devices and their basic functioning is a necessity. Our scientific knowledge base also is growing at an ever-increasing pace. Science is rapidly becoming interdisciplinary. Scientists from many different backgrounds, including biology, chemistry, physics, medicine, and engineering, study a vast array of diverse biological problems. What they all have in common is the use of physics and modern technology in attempts to understand particular biological phenomena. Understanding involves observing, quantifying, and developing a good model that has some predictive ability. The better our understanding of a system or phenomenon, the better is our model in making predictions about its behavior under a larger variety of conditions. As already mentioned, the best models are called theories, the pinnacles of our understanding.

This book is organized into three major parts. After an introduction and an overview of some fundamental themes in this chapter, we begin the first portion of this book, classical mechanics and thermodynamics, in Chapters 2–13. There we learn how to apply a few basic laws of motion for particles to understand the much more complex motion of real macroscopic objects and fluids. Many of the fundamental concepts we learn in the first few of those chapters are used throughout the book in our studies of a variety of biological systems and many important tools used in their study. The second major topic of study is electricity and magnetism and their synthesis in electromagnetism, found in Chapters 14–18. Aside from gravity, these are the sources of the interactions between all objects in our daily experience as well as between biological macromolecules. We introduce much of the physics through biophysical topics such as electrophoresis, biological membranes and channels, nerve conduction, and magnetic resonance imaging (MRI). After having introduced the general properties of waves in Chapter 10, and applied those ideas to sound in Chapter 11, waves are a unifying theme of the third and last major topic of this book. In Chapter 19 electromagnetic waves are discussed, which leads into light waves in optics and matter waves in quantum physics (Chapters 20–23 and 24–25); we conclude, in Chapter 26, with topics on nuclear physics, nuclear medicine, and imaging methods.

Throughout, we emphasize understanding the fundamental concepts of physics and their importance in the study of biology. To help in this, major themes and concepts are developed from specific examples and problems whenever possible. Using descriptive English to explain physical concepts can sometimes lead to confusion because many of the words used in physics have specific meanings that differ from those used in ordinary speech. Mathematics is the natural language of physics, allowing a huge body of knowledge to be expressed in compact equations. However, without an understanding and appreciation of the meanings of the variables, or letters, used in equations, readers often view them as simply a means to obtain a numerical answer to a problem by inserting values for the other letters, rather than as summaries of vast amounts of knowledge. Equations are de-emphasized in this text by keeping the most important, numbered, equations to a minimum. In addition, each chapter has a variety of nonmathematical questions at its end designed to make the reader think about key ideas in the chapter.

On the other hand, without mathematics it would be much more difficult to present a complete picture of our knowledge of science and to make predictions about

the behavior of a system. As we show, Newton's second law equation and Maxwell's four equations of electromagnetism together are equivalent to an enormous body of knowledge. Without those equations, we could not easily express the same information content in words, nor would we be able to approach the tremendous variety of problems these equations can solve. Facility with algebra and trigonometry is assumed here; an appendix is provided for readers to review some basics in algebra and trigonometry as well as in scientific notation, and a few other issues. For those readers who have had some elementary calculus, there are occasional boxed discussions that use some calculus to either derive a particular result or enhance the presentation. This material is not integral to the text and can be skipped over. Each chapter also has a variety of short-answer and open-ended problems to help in learning the material. These should be viewed as integral to the text and a fair number should be attempted to probe understanding of the material and to develop problem-solving skills that will be of benefit in all areas of a life-long education.

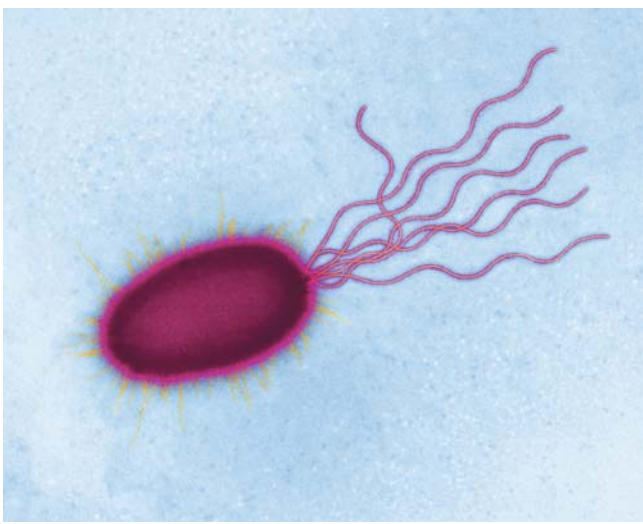
Problem solving involves some extremely useful skills, such as the ability to extract information from a written paragraph, to find the key issue or unknown, to develop solution strategies, and to be able to describe those methods and your solutions to others. Just as critical reading skills will help throughout one's life, problem-solving skills are valuable tools to have in whatever one chooses to do later in life, whether related to science or not.

A major goal of this text is for the reader to develop an appreciation of physics as a discipline that has led to tremendous advances in our civilization. We now have a basic, if incomplete, understanding of our world, ranging from the constituents of atoms to biological cells to galaxies. Although our scientific knowledge has grown explosively over the last 50 years, particularly in the life sciences, the general public's awareness and appreciation of science has declined. *Physics of the Life Sciences* hopes to show many of the interrelationships among the sciences, particularly the physical basis of our understanding of biology.

3. TWO EXAMPLES OF BIOPHYSICAL SYSTEMS: THE SINGLE CELL *E. COLI* BACTERIA AND THE HUMAN HEART

Biological systems are extremely complex, much more so than standard physical systems traditionally studied by physicists. With the tremendous growth of technological methods have come interdisciplinary laboratories and scientific collaborations with a focus on particular biological systems and questions. A glance at a list of topics discussed at various international scientific meetings with a biological focus will show the huge array of systems that are currently studied, including macromolecules, subcellular components, cells, organs, whole organisms, and even interactions between organisms. In the course of this text we show how physics and physical technologies have been applied to many of these. Here we briefly discuss two particularly important systems, one a cell and one an organ, to indicate the range of questions that have been addressed by biophysicists and other scientists.

FIGURE 1.1 *E. coli bacteria as seen using a scanning electron microscope.*



The bacterium, *Escherichia coli* (*E. coli*), is the most studied and well-characterized single-cell organism known. Discovered in 1885, these bacteria are several micrometers long rod-shaped cells (Figure 1.1), a convenient size for optical microscopy, and can be easily, cheaply, and rapidly grown in large quantities. The fact that huge numbers of these organisms can be rapidly grown has led to a number of significant biochemical discoveries including the genetic code, glycolysis, and protein synthesis regulation, and has made these organisms the powerhouse of genetic engineering. *E. coli* bacteria reappear in some of our later discussions as a prototype cell in learning some areas of physics.

Although relatively simple in structure, *E. coli* is sufficiently complex that it exhibits many of the common structures and properties of all cells. This fact also explains its widespread study and one has only to open, at random, almost any book on cell biology, biochemistry, genetics, immunology, or developmental biology, to find extensive references to *E. coli*. The bacterium is surrounded by a cell wall of several layers that shields it from its environment, the intestinal tract of humans or the solutions in a scientist's test tube. Prominent features of the bacteria are its nuclear region and its dozen or so long flagella, which it uses for propulsion. The cytoplasm, or rich broth of biomolecules outside the nuclear region, contains over one million protein molecules and roughly an equal number of other macromolecules and complexes, close to one hundred million small organic molecules including the building blocks of nucleic acids and proteins, and a similar number of small ions all suspended or dissolved in water, which make up roughly 70% of the bacteria's volume (Figure 1.2). The nuclear region contains the genetic code for the bacterium in the form of a single circular DNA molecule of nearly five million nucleotides, or building blocks, folded up into a tight structure with small special-purpose proteins. If spread out into a circle the DNA would have a diameter of about 2 mm, but in the nucleus it occupies about a 100-fold smaller size. There are also much smaller circular pieces of extranuclear DNA known as plasmids, which have become extremely important in genetic engineering. The slender flagella, about twice as long as the bacterium itself, extend out from the cell wall into the surrounding fluid, at times in coordinated helical shapes when propelling the bacteria and at other times in uncoordinated random directions.

E. coli bacteria have been used to study nearly all aspects of cellular and subcellular problems. These range from the structure and function of particular purified macromolecular components such as DNA, RNAs of different types, large numbers

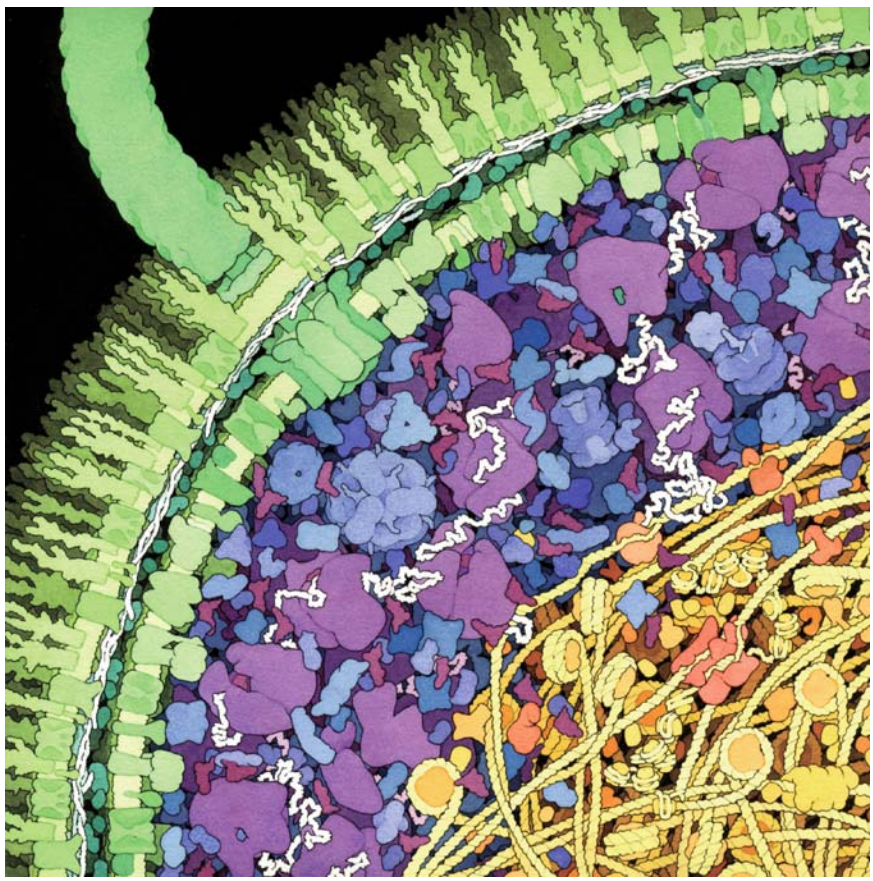


FIGURE 1.2 *Cartoon drawing of the inside of an E. coli showing the membrane and a flagellum (at the* top left, *with the rotary motor protein shown just beneath the membrane at the base of the flagellum), proteins (*middle *with other smaller molecules), and DNA/histones (*bottom*). The drawing is made to scale and according to relative concentrations.*

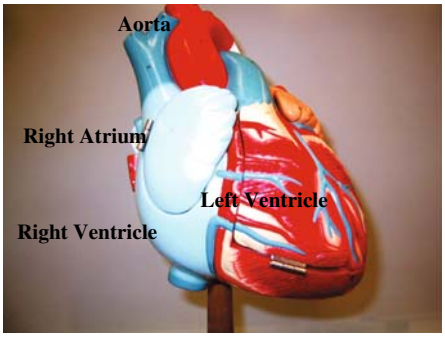


FIGURE 1.3 *Model of the human heart.*

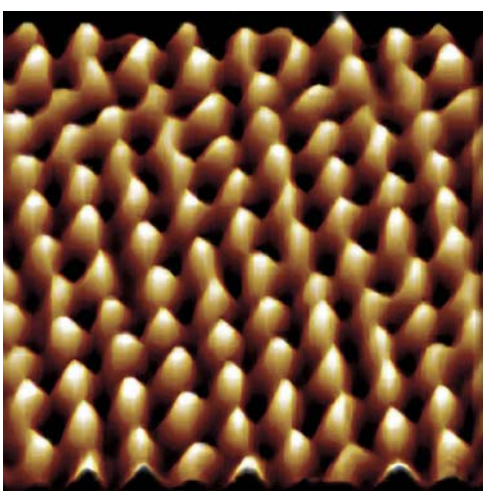
of different proteins, membranes and proteins bound to membranes, to more complex whole-cell problems such as communication with the external environment, the energy transduction mechanisms within the bacteria necessary to sustain life, and the basis of the motility of the bacteria. Clearly the cytoplasm is not just an unorganized soup of macromolecules, small organic molecules, and ions, but it is a highly organized, compartmentalized, and dynamic medium that controls the entire set of processes needed for life.

The presence of plasmid DNA in *E. coli* has led to its major role in genetic engineering. Portions of DNA from other species which, for example, code for the production of particular proteins, can be inserted, using particular enzymes, into *E. coli* plasmid DNA. Many such copies of the plasmid DNA can be grown, as bacteria reproduce every half hour under favorable conditions. Thus *E. coli* can act as a DNA factory for the production of large quantities of any portion of DNA from other organisms.

As an example of a more complex structure let's briefly consider the human heart and, specifically, the many aspects of its structure and function that involve a knowledge of fundamental physics. The heart is a multicellular organ (Figure 1.3), a structure that functions in a coherent manner to produce a cyclic process necessary for life. Adult cardiac muscle cells are one of the few types of cells in humans that are not replaced and do not divide. These permanent cells contract roughly three billion times in a typical life, providing the force necessary to circulate blood through the body.

How does the heart act as a pump? What are the electrical and chemical interactions that control the heartbeat and keep the heart functioning in a coherent manner? What is the ultimate mechanism by which cardiac muscle generates the contracting force needed to pump blood to the lungs and to the body? What are the properties of the blood and of the circulatory system external to the heart that have an impact on the heart's functioning? These are but some of the obvious questions that science has been addressing for many years. We show later in this book that the details of the answers are not completely known, but that all of these areas involve the application of a variety of physical principles. To study the flow of blood, we need an understanding of fluid flow and especially that of a complex fluid, filled with cells so thickly that it would otherwise behave as a solid if not for the elasticity of the cells. An understanding of the basic force production in muscle involves an understanding of mechanics, thermodynamics, and electrodynamics. Such phenomena as cell–cell interactions and coordinated pacemaker action of cardiac muscle cells require an understanding of electromagnetics as well as of nonlinear dynamics, a rapidly developing area of physics. Various aspects of the heart are studied using modern physical technologies including imaging and electrical recording methods in vivo, as well as other more invasive methods in animal studies. In addition, we mention the technology of the artificial heart and of heart transplants as medical areas that have associated basic science research.

FIGURE 1.4 *Atomic force microscopy image of individual oxygen atoms arrayed on a crystal.*



4. THE ATOMIC NATURE OF MATTER

One of the most profound ideas of contemporary science is that all macroscopic bodies—by which we mean bodies that can be seen with visible light—are *composite*. That is, they consist of smaller chunks of matter called atoms, whose properties are much simpler than those of the bodies in which they are found. Atoms cannot be seen with visible light, however, they can be visualized indirectly (Figure 1.4) through various forms of microscopy that don't employ light. Atoms in turn, are made of even simpler pieces of matter called electrons, protons, and neutrons, whose existence is based on much less direct—although strongly convincing—evidence. There is excellent reason to infer that protons and neutrons are also composite, made of elementary bits of matter called quarks. And that

may not be the end of it; even quarks and electrons may be composite. One speculation along these lines depicts the stuff of which they are made as extraordinarily tiny vibrations in space and time. For our purposes, trying to understand the physics of life, we need not worry about such esoteric ideas; we need to worry only about how atoms and collections of atoms behave.

It is one of the most amazing facts of nature that essentially everything in the world around us is made from fewer than 100 naturally occurring different kinds of atoms. An atom has a central nucleus composed of protons and neutrons surrounded by electrons. In atoms that are electrically neutral, the number of electrons equals the number of protons. An *element* is some material that consists of atoms, all of which contain the same number of protons. Thus, atoms with one proton are said to constitute the element hydrogen, atoms with two protons constitute the element helium, and so on. The *Periodic Table of the Elements,* first proposed in 1870 by Mendeleev, a Russian chemist, is an organization of the known elements into groupings having similar physical and chemical properties (Figure 1.5). Although atoms of a given element all have the same number of protons in their nuclei, they may have different numbers of neutrons. Two atoms with the same number of protons but different number of neutrons are said to be different *isotopes* of the same element. Different isotopes behave almost identically as far as chemical reactions are concerned, because chemical reactions involve atomic electrons only, not the atomic nuclei.

Protons and neutrons both weigh about 2000 times more than electrons. So most of the "stuff" of an atom resides in its nucleus. Nonetheless, atoms are mostly empty space. The most common isotope of hydrogen consists of one proton and one electron. Suppose we represent the proton in a hydrogen atom by the following dot: •. About how far away from this dot would the electron be on average if this dot were the actual size of the proton? Where the period next to the dot is? Maybe a centimeter? 10 centimeters? A meter? No. Actually, the electron would spend most of its time roughly 100 meters away (about the length of a football field)! The average diameter of the electron orbit in hydrogen is about 100,000 times the diameter of the proton. In atoms with more protons, the electrons spend more time nearer the nuclei, but no matter how many protons and electrons they contain, atoms are mostly empty. Despite that, it is very hard to squeeze the electrons of an atom closer to their nucleus. It is also difficult to make the electrons of two atoms interpenetrate. If that weren't true, it would be impossible for objects to have more-or-less permanent shapes and sizes.

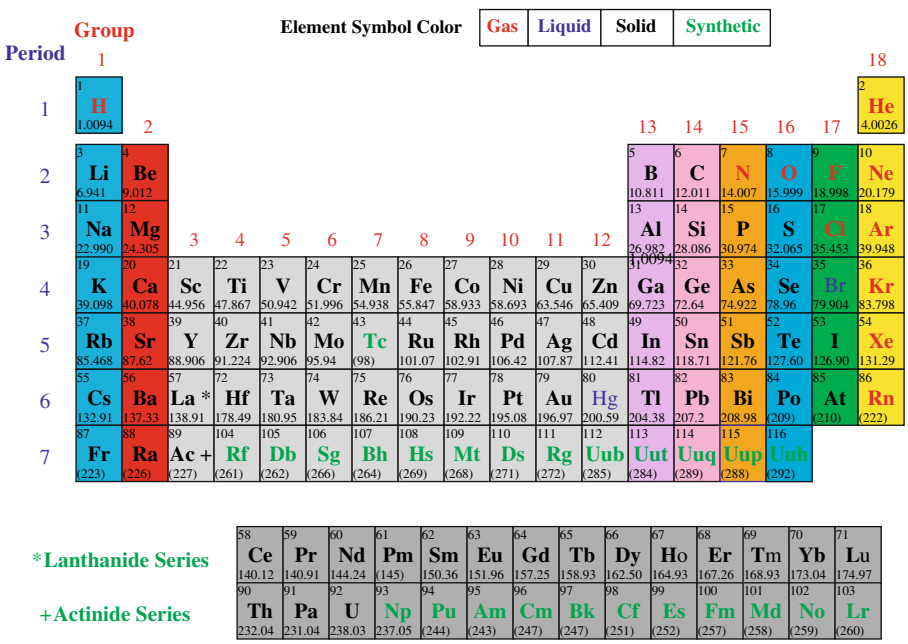


FIGURE 1.5 *Periodic Table of the Elements. See Appendix II for element names and discovery year.*

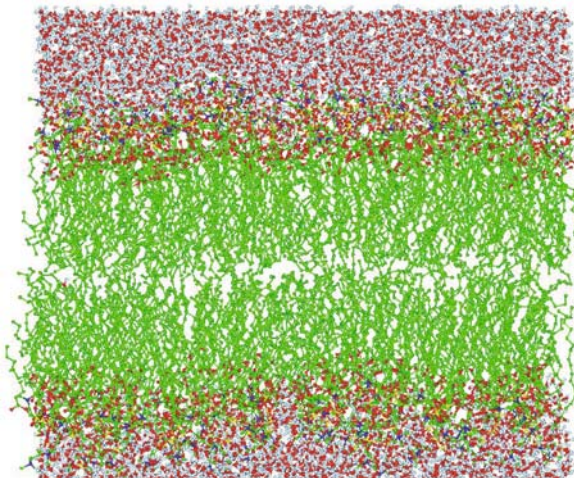


FIGURE 1.6 *Molecular model of a membrane showing disorder.*

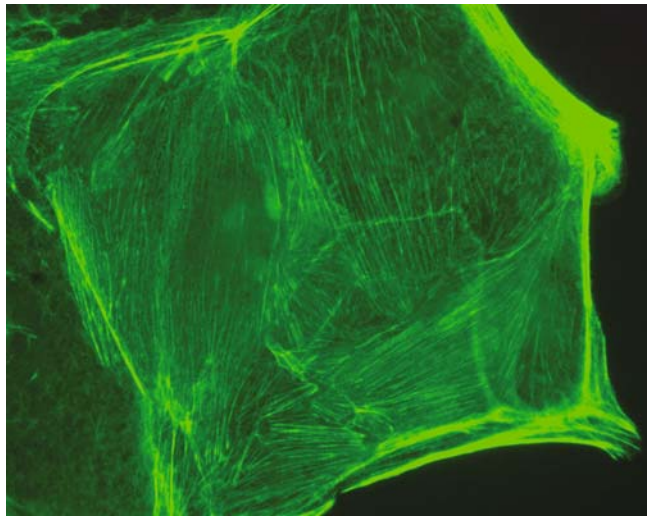


FIGURE 1.7 *Fluorescent microscopy image of the cytoplasm of a cell showing actin filament gel-like structure.*

The number of atoms in a macroscopic object may well exceed 10^{20}. The interactions of these vast swarms of atoms lead to qualitatively different states of matter. In all materials at all temperatures, the constituent atoms are in ceaseless disorganized motion. In solids, the microscopic agitation of atoms is sufficiently confined that the atoms typically do not exchange places. As a consequence, solids have an essentially permanent shape. In fluids (i.e., gases and liquids), however, atoms can pass by each other. This swapping of atomic positions produces the macroscopic phenomenon of *flow* and the microscopic phenomenon of *diffusion* or atomic mixing (which we study in Chapter 2). Fluids flow around inside closed containers and adopt shapes defined by the containers. Fluids don't have a permanent shape. Solids are characterized by the regular and enduring arrangement of their atoms, whereas fluids are characterized by atomic chaos.

Biological materials typically share features of both the solid and fluid states. For example, biological membranes that surround cells or subcellular components are basically two-dimensional highly ordered structures that also have a large degree of mobility within them (Figure 1.6). Their constituent phospholipid molecules tend to be aligned parallel to each other, but can move about within the plane of the membrane quite rapidly by diffusion. Such highly ordered, but yet fluid structures are termed *liquid crystals*. A second significant example is the gel-like nature of cellular cytoplasm (Figure 1.7). Gels have some of the properties of solids, including a rigidity, but can be greatly deformed as well. Cytoplasm is a complex material consisting of thousands of different macromolecules, including proteins, nucleic acids, phospholipids, polysaccharides, as well as smaller organic molecules and salts. Under the control of several different types of filamentous proteins that supply an internal structural rigidity, the cytoplasm can be changed back and forth between conditions that are more fluidlike and more solidlike.

5. MASS, DENSITY, AND THE SIZE OF ATOMS: EXERCISES IN ESTIMATION AND UNITS

Mass is a fundamental property of matter, about which we have more to say in Chapter 2. For now, it is sufficient to think of mass as a measure of the substance of a body. Mass can be measured by an ordinary bathroom or grocery market scale, if the body whose mass is being measured is of moderate size, and by more sophisticated scales if the body is either too large or too small. Again, we discuss how scales work

in Chapter 2. The scientific community has agreed on certain standards of measurement called "le Systéme International de Unités" or *SI units* of measurement. The SI unit of mass is the *kilogram* (kg). Table 1.1 lists the SI units for various quantities that are taken as fundamental. Table 1.2 lists commonly used prefixes designating fractions and multiples of the unit measures. A kg is roughly the mass of a rock the size of a grapefruit. A kg weighs about 2.2 pounds. It is possible to determine the masses of individual atoms with a delicate scale called a "mass spectrometer." (We discuss mass spectrometers in Chapter 17.) Compared with a rock an atom doesn't have very much mass. A rule for assessing the approximate mass of an atom is to look up the number of "atomic mass units per atom" (designated u/atom) for the atom of interest in the Periodic Table, then multiply by 1.66×10^{-27} kg/u. The number of atomic mass units per atom is essentially the average number of protons plus neutrons in all isotopes of the element in question found on Earth.

Table 1.1 *The Units of Measure Upon Which the International System of Units (SI) Is Based*

Fundamental Quantity	*SI Unit*	*Abbreviation*
Mass	kilogram	kg
Length	meter	m
Time	second	s
Electrical Current	ampere	A
Temperature	kelvin	K
Number of Atoms	mole	mol
Light Intensity	candela	cd

Table 1.2 *Commonly Used Prefixes for Power of Ten Multiples or Fractions of Base Units*

Power of Ten Mutiple	*Prefix*	*Abbreviation*
10^{-15}	femto-	f
10^{-12}	pico-	p
10^{-9}	nano-	n
10^{-6}	micro-	μ
10^{-3}	milli-	m
10^{+3}	kilo-	k
10^{+6}	mega-	M
10^{+9}	giga-	G
10^{+12}	tera-	T

ESTIMATES AND THE BIG PICTURE

In this book, we attempt to motivate and illustrate important concepts with concrete numerical examples. Detailed numerical calculations undoubtedly are of use when building a bridge or when evaluating how much of each ingredient should go into an explosive chemical reaction, however, precise values are almost never necessary to understand the essence of most physical situations. In fact, *order of magnitude estimates* (estimates that round off values to the nearest power of ten) are usually completely adequate to see why a piece of physics is the way it is. For example, suppose you would like to buy a new car whose price is on the order of $\$10^4$ and you know your bank balance is on the order of $\$10^2$. Obviously, there is little point in calculating whether the car's price is \$8,000 or \$12,000, or whether your balance is \$80 or \$120. The big picture is that in no case will you be able to pay cash for the car. In this chapter we discuss many numerical examples of sizes of different quantities. In each, please try to focus on the power of ten. The order of magnitude is the big picture.

Example 1.1 What is the mass of a typical atom of naturally occurring carbon?

Solution: The Periodic Table (see Figure 1.5) states that naturally occurring carbon has an atomic mass number of just about 12 u/atom. The average mass of an atom of carbon on Earth is therefore about (12 u/atom)$(1.66 \times 10^{-27}$ kg/u$) = 1.99 \times 10^{-26}$ kg/atom. In this calculation, note how the units of the answer are manufactured from the units of the pieces. The units are treated like algebraic quantities so that the "u's" cancel in the product (u/atom) × (kg/u) = (kg/atom)

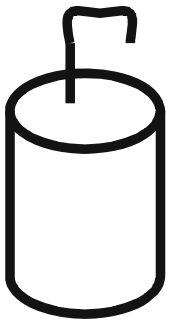


FIGURE 1.8 *A laboratory standard mass.*

We wish to demonstrate how knowledge of macroscopic properties sometimes can be converted into knowledge about atoms. Let's start with the question, how many atoms are contained in a 1 kg mass of known composition? Suppose, for example, we are told that the mass is solid gold. The Periodic Table tells us that gold has about 197 u/atom. So the mass in kg of a gold atom is 197 u/atom $\times$ 1.66 $\times$ 10^{-27} kg/u = 3.27 $\times$ 10^{-25} kg/atom. The 1 kg is some number of atoms times the mass per atom, so if we divide the latter value into 1 kg we find that 1 kg of gold contains 1 kg/3.27 $\times$ 10^{-25} kg/atom ~3 $\times$ 10^{24} atoms. (The "~" means "approximately.") That's a typical number for solids: 1 kg of a solid contains from about 10^{24} to about 10^{26} atoms.

A related question is, given the volume of a body whose mass is 1 kg, what material is the body made from? Actually, in practice one frequently measures some characteristic lengths associated with a body rather than its volume. So, to make progress on this problem it is necessary to recall that the volume of a rectangular solid (one for which each side is a rectangle) is the product of a length times a width times a height. When the solid is a cylinder with a circular cross-section, its volume is πR^2 times height, where R is the radius of the circular cross-section. And, when the solid is a sphere, its volume is $4\pi R^3/3$, where R is the radius of the sphere. (For other cases, the formulae are more complicated. We won't worry about such cases.) The SI unit of length is the *meter* (m). (A meter is a little longer than a yard.) Thus, a volume has SI units of meters cubed, m^3.

For concreteness, suppose we want to know of what a typical physics lab 1 kg mass is made. The one shown in Figure 1.8 is a cylinder with a round base 0.046 m in diameter and 0.075 m tall (ignore the hook). As the base is a circle, its volume can be calculated by the rule $V = \pi R^2 \times$ height. Remember, R is the radius of the circle so it is diameter/2. The area of the base of this mass is 1.66 $\times$ 10^{-3} m^2 and its volume is 1.25 $\times$ 10^{-4} m^3 (125 cc, or cm^3, if you are used to volume in cubic centimeters: 1 cc = 10^{-6} m^3, or 0.125 L if you prefer liters: 1 L = 10^{-3} m^3). (Please check the 1.25 $\times$ 10^{-4} m^3 result yourself.)

The next step in this little detective story is to determine the *average density* of the mass. The average density (ρ_{ave}) of a body is defined as the mass (M) of the body divided by its volume (V): $\rho_{ave} = M/V$. The average density of our lab mass is therefore (1 kg)/(1.25 $\times$ 10^{-4} m^3) = 8 $\times$ 10^3 kg/m^3. This is also a typical result. The densities of most solids are a few thousand kg/m^3. The last part in our sleuthing requires consulting what is already known about solid densities, as in Table 1.3. Inspection of such a table indicates that the density of iron (7900 kg/m^3) is quite close to 8000 kg/m^3. Of course, our lab mass could be made of a mixture of atoms (such as brass or stainless steel, e.g.) or have unseen holes inside, but if we are told it is an elemental solid (one kind of atom) with no internal cavities, then it's probably iron.

Table 1.3 *Mass Densities of Selected Materials*

Material	*Mass Density (kg/m³)*
Elemental Solids	
Aluminum	2700
Carbon (graphite)	2250
Copper	8960
Gold	19,300
Iron	7880
Lead	11,340
Lithium	534
Silicon	2420
Uranium	18,700

(*Continued*)

Table 1.3 (*Continued*)

Alloys, Molecular and Composite Solids	
Brass	8400–8700
Steel	7800–7900
Ice (at 0°C)	917
Glass	2400–2800
Hardwoods	500–700
Soft tissue organs	1030–1060
Bone	1900
Liquids	
Water (at 4°C)	1000
Blood (at 37°C)	1060
Mercury (at 0°C)	13,600
Gases	
Air	1.29
Carbon dioxide	1.98
Helium	0.18
Hydrogen	0.09
Nitrogen	1.25
Oxygen	1.43

A macroscopic measurement of density of a solid body allows us to answer the question: how far apart are atoms in the body? If we divide the density of a solid body (mass per unit volume) by the mass per atom we get atoms per unit volume. If we take the reciprocal of that, we get volume per atom. Now, if we pretend that each atom is a little cube of side L, the volume per atom is L^3. Thus, L is the cube root of the volume per atom; it is also the average distance between adjacent atoms. See Figure 1.9. For iron we have $7900\ \text{kg/m}^3/(55.8\ \text{u/atom} \times 1.66 \times 10^{-27}\ \text{kg/u}) = 8.53 \times 10^{28}$ atoms/m^3. The volume per atom in solid iron is then $(8.53 \times 10^{28}\ \text{atoms/m}^3)^{-1} = 1.16 \times 10^{-29}\ \text{m}^3$/atom, and the cube root of that, 2.26×10^{-10} m, is the average atomic spacing. The distance 10^{-10} m recurs frequently when considering atoms. You will sometimes find 1×10^{-10} m referred to as 1 ångstom = 1 Å, although in keeping with the SI conventions it is more fashionable these days to use the nanometer: 1×10^{-9} m = 1 nm. Thus, 2.26×10^{-10} m is either 2.26 Å or 0.226 nm.

The average distance between atoms in any elemental solid is roughly the same as for iron. Some other values are: lithium = 0.20 nm, carbon (graphite) = 0.21 nm, aluminum = 0.26 nm, copper = 0.23 nm, gold = 0.26 nm, and uranium = 0.28 nm. Now here is another very familiar result: it is exceedingly difficult to increase the density of a solid by squeezing it. In other words, in a solid, the atoms are crammed together about as closely as possible. This fact and the fact that the average spacing of atoms is about 0.2–0.3 nm for all elemental solids tells us the very interesting and surprising result that *all atoms are about the same size*, despite the fact that their atomic masses vary by a factor of over 200!

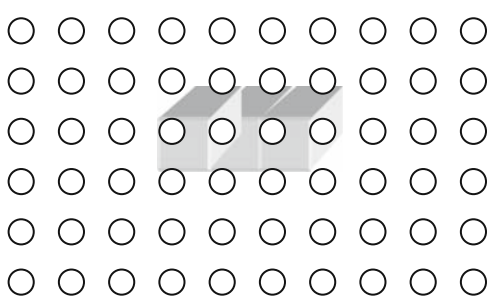


FIGURE 1.9 *A crystal with atoms arranged in a cubic array with spacing L. Each atom can be imagined to lie in a little cube with volume L^3.*

Now, you might be tempted to conclude that because liquids flow and have no permanent shape that the spacing of atoms in liquids would be a lot larger than in a solid. Let's see. A familiar elemental liquid is mercury. Its density is about 13,500 kg/m^3 and its u/atom is about 201. From these values it is straightforward to calculate that the average atomic spacing in liquid mercury is about 0.29 nm, not very different from the solids listed above. How about in water? Water is a molecular liquid. The u/molecule for water is about 18 (2 for the two hydrogen atoms and 16 for the oxygen atom). Because there are three atoms per molecule, the average mass per atom is 6 u. The density of water is about 1000 kg/m^3. Consequently, the

average atomic spacing in water is about 0.22 nm, again, more or less the same value as in solids. The remarkably different physical properties of solids and liquids arise from only very small differences in how their atoms are spaced.

What can we say about atomic spacing in gases? The most familiar gas is air, a mixture of primarily nitrogen and oxygen molecules. Let's say that the average u/molecule for air is about 29. Because nitrogen and oxygen molecules contain two atoms, the average u/atom for air is about 14.5. The density of air at room temperature and at sea level atmospheric pressure is 1.29 kg/m^3, a value that is something like 1000 times less than water. The average atomic spacing in air is about 2.7 nm, that is, about 10 times greater than in a solid or liquid. If we squeeze a quantity of air down to 1/1000 of its normal volume, it becomes a liquid; the densities of liquid oxygen and liquid nitrogen are almost exactly 1000 times that of air.

Because biological materials have properties midway between the solid and liquid states the spacing of atoms in them is about 0.2–0.3 nm. We can use this idea to assess how many atoms one might find in a typical biological cell. Cells have somewhat different sizes, but a typical cell is roughly about 20×10^{-6} m = 20 micrometers = 20 μm on a side. That is, a cell has a volume roughly about 8×10^{-15} m^3 (obtained by cubing 20 μm). If a typical atom spacing is 0.25 nm, the volume occupied by an atom is about $(0.25\text{ nm})^3 = 1.5 \times 10^{-29}$ m^3/atom. Consequently, the number of atoms per cell is about $(8 \times 10^{-15}$ m^3/cell)/$(1.5 \times 10^{-29}$ m^3/atom) $= 5 \times 10^{14}$ atoms/cell.

A cell has lots of stuff in it. All cells contain DNA, for example. Drawings of pieces of DNA in textbooks show it as a long, double helix structure. But, just how long is it? DNA consists of multiple subunits called base pairs ("C–G" and "A–T"). The number of atoms per pair is 27. Typical animal cells have about 5×10^9 pairs in their DNA. That corresponds to about 1.4×10^{11} atoms. Suppose that all of the atoms in the DNA molecule were strung end to end in a linear chain. The chain would be about $(1.4 \times 10^{11}$ atoms$) \times (2.5 \times 10^{-10}$ m/atom) = 35 m long! Of course, clumping atoms into base pairs of about 30 atoms each saves space. Even so, if the pairs were strung out in a linear chain, the DNA would still be about 1 m long. Obviously, DNA in a cell can't be a linear chain because it would burst through the cell membrane. It must be stored in a tight coil when "not in use" and only small portions must be pulled apart when transcription or replication occur. Similar conclusions can be made about other important ingredients of a cell, such as large proteins, for example.

CHAPTER SUMMARY

Each chapter has a short summary of the major concepts in the chapter. Please note that reading these summaries cannot replace a careful reading of the entire chapter.

Science progresses by developing models, the best of which are called theories. Physics, the most fundamental of the sciences, has had an increasing impact and relevance in biology as new technologies and basic understanding has developed.

As examples of physics' recent impact on biology, some aspects of bacteria, and of the human heart are discussed in Section 3.

All matter is composite, composed fundamentally of atoms, made of electrons, protons, and neutrons. We can distinguish three different states of matter: solid, liquid, and gas; but there are some common materials in biology that fall between these, such as gels or liquid crystals.

The SI unit for mass is the kilogram (kg), and another useful unit is the atomic mass unit (u) where $1\text{ u} = 1.66 \times 10^{-27}$ kg. (Mass) density, ρ, is defined as the average mass per unit volume. Using a value for the density and for the atomic weight, the typical atomic spacing (comparable to atomic size) is a few nm (10^{-9} m).

QUESTIONS

1. Discuss the difference between the density of a material and its volume. Which is an "intrinsic" property of the material not depending on the amount and which is an "extrinsic" property? Can you think of other examples of intrinsic properties of a material?
2. Why don't the elements in the periodic table have masses that are exactly integral multiples of 1 u? Think about the effect of different isotopes (elements with different numbers of neutrons) and their natural abundance.

MULTIPLE CHOICE QUESTIONS

1. Suppose the density of a solid is D and its average atomic mass is M. Which of the following represents the average spacing between atoms in the solid? (a) D/M, (b) M/D, (c) $(D/M)^{1/3}$, (d) $(M/D)^{1/3}$.
2. The atoms in a solid or liquid are said to be about the same size as the atomic spacing in the solid or liquid because (a) solids and liquids are difficult to compress, (b) atoms become much larger when they are in the gas phase, (c) atoms are in electronically excited states in the gas phase, or (d) the electrons of atoms in solids and liquids are all confined inside the respective nuclei.
3. A large protein consists of a strand of about 10,000 atoms coiled up into a ball. If the strand were pulled out into a line about how long (order of magnitude) would the strand be? (a) 10^4 m, (b) 1 m, (c) 10^{-2} m, (d) 10^{-6} m.
4. The number of gold (197 u/atom) atoms in a gold ring in the shape of a donut with a diameter of 2.0 cm and a radius of the cross-section of 2.0 mm is (estimate order of magnitude) (a) 10^{20}, (b) 10^{22}, (c) 10^{24}, (d) 10^{26}.
5. The interatomic spacing in solids and liquids is about (a) 0.2 Å, (b) 0.2 nm, (c) 0.2 pm, (d) 0.2 μm.

PROBLEMS

1. From the Periodic Table of the Elements (Figure 1.5) calculate the mass (in kg) of an atom of naturally occurring helium, oxygen, nitrogen, and phosphorus.
2. Calculate the mass (in kg) of a molecule of carbon dioxide, a molecule of water, and a molecule of the amino acid alanine ($C_3NO_2H_7$).
3. What is the average distance between silicon atoms in solid silicon?
4. What is the average intermolecular spacing of the sodium ions in a 1 M solution of NaCl? (A 1 M solution has 1 mole of NaCl molecules, or 6.02×10^{23} of them, per liter of solution.)
5. Express the density of gold in units of $\mu g/\mu m^3$ and in units of pg/nm^3.
6. In a cube of bacterial cytoplasm 100 nm on a side there are roughly 450 proteins. What is the average distance between these proteins?
7. The DNA in the *E. coli* bacteria forms a circle of about ½ mm diameter if stretched out. Roughly what is the mass of the DNA molecule? Assume the following data for the double-stranded DNA: average molecular weight of a nucleotide = 325 u; distance between pairs of bases along the DNA backbone circle = 0.34 nm. If this represents about 1% of the mass of the bacteria, what is its total mass?
8. The entire *E. coli* chromosome is replicated in 30 min. For this to occur, the double-stranded DNA must be partially unwound all along its length. Assuming the roughly 400,000 turns of the DNA double-helix unwinds starting at one end and going uniformly to the other end, what is the linear speed of the unwinding site along the DNA? Recall that each turn of the DNA helix corresponds to 10 base pairs, or to a distance of 10×0.34 nm. What is the unwinding rate in turns per minute (or revolutions per minute)? This is comparable to a high-speed centrifuge.

2

Newton's Laws of Motion for a Particle Moving in One Dimension

Living cells exchange energy and matter with their surroundings. They reproduce. Often they move about. To understand such basic aspects of life, it is essential to understand how motion is related to force and how force is related to energy. Explaining these relations for an object moving in one dimension is the goal of this and the next two chapters.

Before beginning to read and master the formal discussion of motion that follows in this chapter, however, it is very useful to remind ourselves what it feels like to move at constant velocity and to accelerate. Recall how it feels to ride in a car along a straight flat highway that has recently been resurfaced. If the car's speedometer is fixed at a constant reading you can close your eyes and not know you are moving at all, no matter how fast the speedometer says you are moving. Of course, roads aren't straight and flat for very long stretches. You feel clues that you are moving from the little bumps and turns the car makes. Riding in an elevator is probably a better example. Once the elevator gets going, only the flashing floor numbers give any hint that anything is happening, no matter how fast the elevator is traveling or whether you are going up or down. In both car and elevator examples, when you feel as if you are at rest you are moving in a straight line at a constant rate. This kind of motion is called *constant velocity*. Constant velocity feels exactly like standing still.

When the car turns or goes over a bump or speeds up or slows down, or when the elevator starts or stops, you definitely feel it. All such instances involve change in velocity. Change in velocity is called *acceleration* and acceleration can be felt. If a trinket dangles by a thread from the car's rear view mirror you can see it deflect from hanging vertically at the same instant you feel acceleration. If by some bizarre chance, you are standing on a scale as the elevator starts or stops, the scale's reading will change when you feel the acceleration.

Why you feel acceleration but not constant velocity, why acceleration causes the trinket to deflect and the scale reading to change, all require an explanation. That explanation is contained in Newton's laws of motion, discussed in this chapter. In order to understand the content of Newton's laws, we have to be able to describe motion with quantitative precision. The major goal of this chapter is to demonstrate how a body's interactions with its surroundings can explain changes in its motion. We use the term force to denote a quantitative measure of interaction. The theme of this chapter, then, is that force explains (causes) acceleration. As discussed previously, any macroscopic body is a collection of smaller, more fundamental pieces. A complete understanding of the changes in motion of a macroscopic body requires keeping track of the forces experienced by every subpiece of the body due to every other subpiece (these are called internal forces) and due to every other additional body (external forces). In this chapter, all bodies are treated as particles and all changes in motion arise from external interactions. This simplistic view allows us to develop powerful tools that can subsequently be applied to more general and more realistic behaviors. The chapter ends with a short discussion of diffusion, the random thermal motion of

J. Newman, *Physics of the Life Sciences*, DOI: 10.1007/978-0-387-77259-2_2,

small particles, to contrast this type of motion with that described in the rest of the chapter. Diffusion is an extremely important process in biology, playing a major role in our existence through, for example, gas, nutrient, and waste exchange in the blood.

1. POSITION, VELOCITY, AND ACCELERATION IN ONE DIMENSION

Until we get to Chapter 23, we are interested primarily in phenomena associated with objects that can be seen (perhaps with the aid of a microscope or telescope) with ordinary light. That doesn't narrow our interests very much. On the small end, we can certainly see inside living cells; on the large end we can see clusters of galaxies. All objects that can be seen with light are *composite*, that is, composed of smaller pieces of matter. Organisms, for example, are composed of cells; cells are composed of molecules; molecules are composed of atoms; atoms are composed of nuclei and electrons. As we show in Chapter 6, we can assign to any object a unique point called the object's *center of mass*. The motion of any object can then be thought of as consisting of two parts: motion of the center of mass and motion about the center of mass. For now, just think of the center of mass as the body's "center."

If a body moves so that all of its composite pieces do exactly what the center of mass does—for example, when the center of mass of the body moves 1 m north each composite piece also moves 1 m north—the body is said to undergo a *rigid translation* (see Figure 2.1). An object undergoing a rigid translation can be treated as a *point particle*, a mass without spatial size. Its shape and extent in space are irrelevant.

To start, imagine some object of interest moving along a straight line. The object can be microscopic (such as a protein molecule or a bacterium) or macroscopic (such as a car or even you, yourself). Motion along a line is called *one-dimensional* because only one coordinate, x, say, is needed to describe it. Here, then, x designates the location of the center of mass of a car measured from an arbitrary origin. There are two directions to go along the coordinate axis from its origin. We distinguish between them by saying one is the "positive" direction, the other the "negative." Thus, x is a signed number having units of length.

Whether it is the motion of our car or the motion of a molecule, in practice we measure one position at one time, then another position at another time, and so on, over and over. That is, in any experiment the data we collect are a sample of the motion acquired at discrete instants. This is true irrespective of what apparatus or technique we employ. For example, we (or a policeman) might use radar or sonar to identify where our car is at various moments. Such devices send out a signal and receive its echo, then another signal and its echo, on and on. Between signals we know nothing; there are gaps in the data. The same is true if we videotape a moving object. Video is really a succession of still frames (in the United States, one every thirteenth of a second). We can get detailed information about the object every frame, but nothing in between. The results, consequently, comprise a table of positions (measured with finite precision and limited accuracy) recorded at discrete sampling times. In other words, our experiment yields a finite set of position values $\{x(t_1), x(t_2), x(t_3), \ldots\}$ where $x(t_1)$ is the position measured at time t_1, $x(t_2)$ is the position at time t_2, and so on. Although we believe that our car or a bacterium moves continuously in time (i.e., the closer t_1 and t_2 are to each other, the closer $x(t_1)$ and $x(t_2)$ are to each other), the best we can do, even if (as is frequently the case) we are aided by a high-speed computer with lots of memory, is obtain a broken and punctuated approximation to its theoretical, continuously flowing motion.

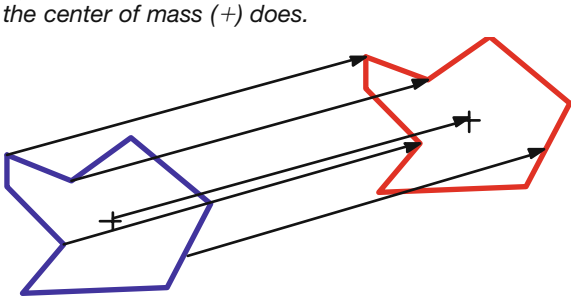


FIGURE 2.1 *An object undergoing rigid translation. All parts do what the center of mass (+) does.*

In this book we use the International Standard (SI) units in which lengths are measured in meters (m), although often we refer to small fractions of meters (e.g., cm, mm, μm, and so on) or large multiples of meters (in particular, km); see Table 2.1.

Table 2.1 *Commonly Used Units of Distance*

Name	*Abbreviation*	*Multiple of a Meter*	*Roughly Comparable to*
Meter	m	1	Length of your arm
Centimeter	cm	10^{-2}	Length of a (new) pencil eraser
Millimeter	mm	10^{-3}	Width of a pencil point
Micrometer	μm	10^{-6}	Length of a cell
Nanometer	nm	10^{-9}	Diameter of a small molecule
Kilometer	km	10^{+3}	Half a mile

A table of numbers is not usually a very useful way to characterize motion. Table 2.2 provides an example. In this table, we see the results of three different observers recording the motion of the same remote control toy model car (Figure 2.2), using the same coordinate system and the same starting time (i.e., the instant they all call $t = 0$ s), but with three different sampling rates (one every 2 s coded in blue, one every 1 s in green, and one every 0.5 s in red, respectively). (The *second*, incidentally, is the SI unit of time, often abbreviated as just s.) There is typically too much to keep track of in a table; it's hard, with tabular information, to see a "big picture."

Table 2.2 *Table of Observations on the Position of a Remote Control Toy Car as a Function of Time*

Observer #1		*Observer #2*		*Observer #3*	
Time (s)	*Position (m)*	*Time (s)*	*Position (m)*	*Time (s)*	*Position (m)*
0	0.890	0	0.890	0	0.890
				0.5	0.663
		1	0.567	1	0.567
				1.5	0.968
2	0.909	2	0.909	2	0.909
				2.5	0.633
		3	0.535	3	0.535
				3.5	1.008
4	0.700	4	0.700	4	0.700

More useful than a table is to make a plot of the data, plotting position $x(t)$ versus time t with t (the independent variable) plotted on the horizontal axis and $x(t)$ (the dependent variable) plotted on the vertical axis, as in Figure 2.3 using the same color codes for the different observers.

In the figure, we have attempted to fill in missing information by interpolating between data points (in this case, by simply "connecting the dots" with straight lines). Interpolation of Observer #1's data (in blue) gives a very crude picture of the car's motion over the interval 0 s to 4 s. Observer #2's data (in green) provides more detail and #3's (in red) even more. By interpolating, we are creating a *model* of the car's motion that will allow us to say something about where the car was at times not observed.

The word "model" is used a lot in physics. A model is a representation or an approximation of a thing, not the thing itself. Some models are better than others: for example, the blue model of the car's motion shown in Figure 2.3 is not as informative or accurate as the red model. The former model has less of a "database" to support it than does the latter. The blue model

FIGURE 2.2 *A remote controlled car whose motion we study.*

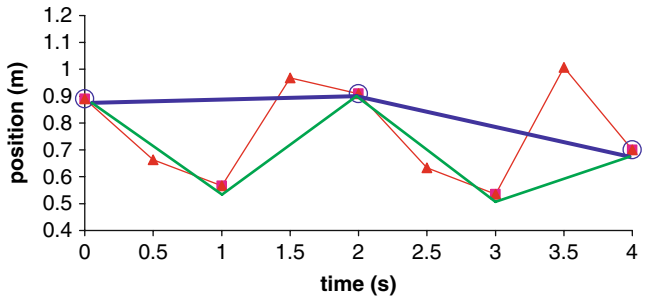


FIGURE 2.3 *The position data of Table 2.2 plotted for each observer.*

can be thought of as "provisional," a kind of first approximation. As we acquire more and more data that model is replaced by more and more sophisticated approximations.

We can imagine that if the observed sampling rate is increased so that data are taken more and more frequently, the resulting plots would more and more define a smoothly continuous curve of some sort. In fact, if we are lucky we might even be able to fit an analytic expression to the data, producing an equation model for the car's *instantaneous position*, $x(t)$, that is, an explicit relationship between position and time that would allow us to determine the car's position at any instant (not just at the times of measurement). Such analytic models are especially useful because they allow us to make predictions about events not yet witnessed.

Given a position record such as that shown in Table 2.2, or, equivalently, in Figure 2.3, we can define a number of useful quantitative tools. First, we have the notion of *distance traveled* in some time interval. The total distance traveled in any interval of time is the sum of the distances traveled during each subinterval of the motion. Furthermore, each contribution is positive, irrespective of in which direction the motion takes place. Formally, distance equals the absolute value of change in position. Thus, according to Observer #1 in Table 2.2, the total distance covered by the car in 4 s is 0.228 m, that is, from a position of +0.890 m out to +0.909 m (a distance of 0.019 m), then back to +0.700 m (an additional distance of 0.209 m). According to #2, the total distance the car travels is 1.204 m, and according to #3 the total distance is 1.938 m. Make sure you understand why.

The *average speed* over a certain time interval is the total distance traveled in that interval divided by the elapsed time. So for the three observers of Table 2.2, #1 assigns to the car's motion an average speed of 0.228 m/s = 0.057 m/s, #2, 0.301 m/s, and #3, 0.485 m/s. (Note that in calculations units are treated as algebraic quantities.)

Next, we introduce the notion of the *displacement*, Δx, in a time interval t_i to t_f ("i" implies "initial", the beginning of the interval, and "f" "final", the end of the interval). (Here, and more generally, the Greek letter Δ [capital "delta"] denotes a difference between two values.) Displacement is the *directed distance*

$$\Delta x = x(t_f) - x(t_i).$$

Displacement can be positive, negative, or zero (as opposed to distance, which is never negative), with the sign indicating the net direction of the associated motion. Thus, in the example of Table 2.2, all three observers agree that the displacement of the car, Δx, for $t_i = 0$ s to $t_f = 2$ s is +0.019 m (displacement in the + direction during this interval), for $t_i = 2$ s to $t_f = 4$ s is −0.209 m (displacement in the − direction during this interval), and for the entire interval from $t_i = 0$ s to $t_f = 4$ s is −0.190 m.

The *average velocity* $\bar{v}$ of our car is defined for a specific interval of time, $\Delta t = t_f - t_i$, as

$$\bar{v} = \frac{\Delta x}{\Delta t}. \tag{2.1}$$

Notice that this expression is different from the average speed, because it is not the distance traveled but the displacement that is in the numerator. Unlike the average speed, which is always positive, the average velocity can be positive, negative, or zero depending on whether Δx is positive (moving to the right), negative (moving to the left), or zero (either there was no motion or the object has returned to its starting point). Again, all three observers in Table 2.2 agree that the car's average velocity is +0.010 m/s from $t_i = 0$ s to $t_f = 2$ s, −0.105 m/s from $t_i = 2$ s to $t_f = 4$ s, and −0.048 m/s from $t_i = 0$ s to $t_f = 4$ s. (Contrast these results with their conclusions about average speeds over the same interval.)

Average velocity is a statement about the tendency for an object to move over a finite time interval. In between the starting time and the ending time, the object can do lots of interesting things that are not accounted for by the average velocity. Of course, as we increase our sampling rate and make our time interval smaller and smaller, less and less departure from the average motion will occur in an interval of time. This leads us to a still different (more refined) concept, namely, that of *instantaneous velocity*. Imagine starting at some generic time $t_i = t$ with our car at $x(t)$ and going to $x(t + \Delta t)$ at $t_f = t + \Delta t$, some time later. The instantaneous velocity of the car at time t, $v(t)$, is defined as

$$v = \lim_{\Delta t \to 0} \frac{\Delta x}{\Delta t}. \tag{2.2}$$

The symbol "$\lim_{\Delta t \to 0}$" is read, "in the limit as Δt approaches 0." Operationally, it means "make the sampling rate so fast that the average motion and the exact motion in the time interval Δt are indistinguishable." You can think of this as the velocity reported by a car's speedometer.

As we said before, we believe that our car moves continuously in time. Continuous, here, means that we can make a plot of position versus time without ever lifting our pencil off our paper. There are no holes or jumps in such a plot. In other words, we don't believe that our car (no matter how spiffy) is ever at $x(t)$ one instant then at a very different $x(t + \Delta t)$ an extremely short time later. Thus, despite the fact that we are making Δt exceedingly small in the denominator of Equation (2.2)—and therefore seemingly threatening to make $\Delta x/\Delta t$ exceedingly large—Δx in the numerator is also getting smaller and smaller, and the ratio of the two remains nice and finite.

Moreover, we also tacitly believe that the car's motion is *smoothly continuous*. "Smooth" means that there are no instantaneous "jerks." If the car has a nice, finite velocity $v(t)$ at time t, its velocity $v(t + \Delta t)$ is not much different a short time Δt later. As we argue in just a bit, smoothly continuous means a plot with neither holes nor sharp points (cusps).

Well, the formal definition of a velocity at an instant may be clear, but how do we actually use the definition? How, for example, do we assign a number to it? The answers to these questions depend on what information you have at the start. First, suppose another observer has taken a great deal more of the car's position data and fit a smooth curve to the data points. This smooth curve is presented to you as an accurate model of the car's motion at any time. Such a plot is shown in Figure 2.4a.

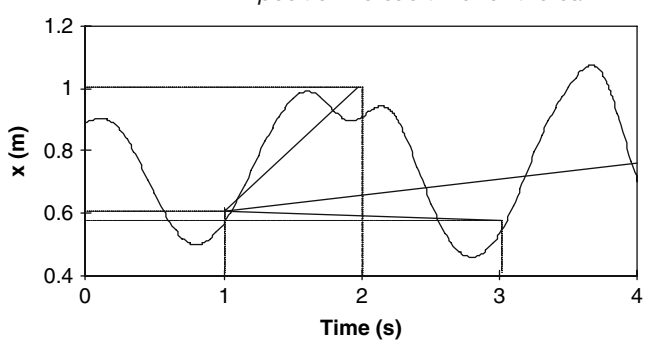
Let's try to determine, from the curve given to us, the car's instantaneous velocity at $t = 1$ s. The position at 1 s is $+0.567$ m. We take a second time, $t + \Delta t = 4$ s, say, and the corresponding position (read from Figure 2.3 or 2.4a or looked up in Table 2.2) is $+0.700$ m. We conclude that the average velocity over that interval is

$$\bar{v} = \frac{[+0.700 \text{ m}] - [+0.567 \text{ m}]}{4 \text{ s} - 1 \text{ s}} = +0.044 \text{ m/s}.$$

Note that this average velocity is the same as the slope of the line connecting the points (1 s, $+0.567$ m) and (4 s, $+0.700$ m) on the graph in Figure 2.4a (because slope is calculated by dividing rise [or fall] in the vertical direction by the corresponding run in the horizontal direction, and, in this case, that is $\Delta x/\Delta t$).

Now, let's take $t + \Delta t$ to be 3 s. Given that $x(3 \text{ s})$ is $+0.535$ m, we calculate the average velocity in this interval to be -0.017 m/s. Then, take $t + \Delta t = 2$ s. The average velocity from 1 s to 2 s is $+0.342$ m/s. Every interval we've

FIGURE 2.4a *Smooth curve of the position versus time for the car.*



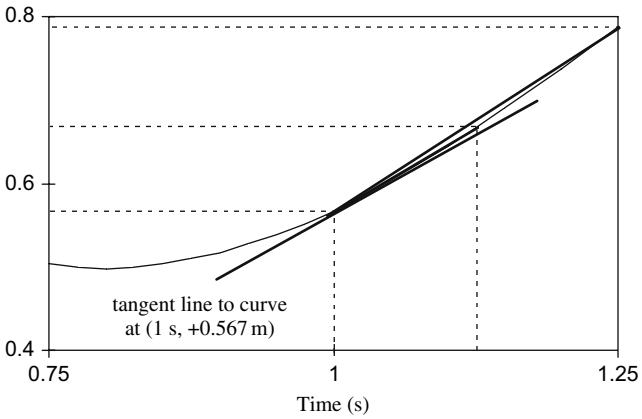


FIGURE 2.4b *Zoom-in around $t = 1$ s data from Figure 2.4a.*

picked so far has yielded quite a different average velocity. None of these can be said to be the instantaneous velocity at $t = 1$ s, because the Δts aren't very small in any of these examples. Now switch your attention to Figure 2.4b. Here the piece of the plot between $t = 0.75$ s and $t = 1.25$ s is magnified. If we take $t + \Delta t = 1.25$ s, we obtain for an average velocity about

$$\frac{[+0.79\text{ m}] - [+0.567\text{ m}]}{1.25\text{ s} - 1\text{ s}} = +0.89\text{ m/s}.$$

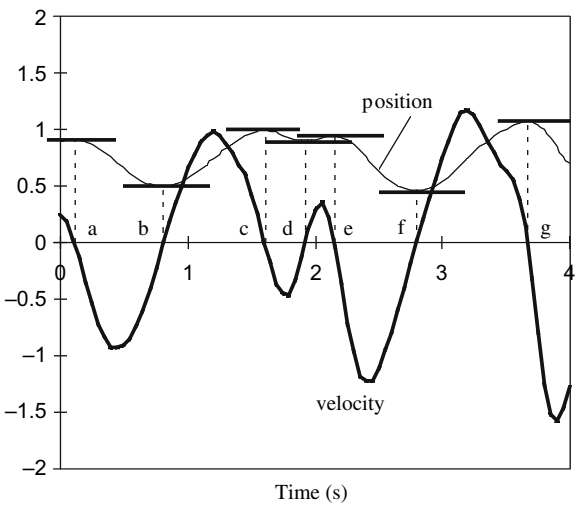
Finally, we take $t + \Delta t = 1.13$. The average velocity in this interval is $+0.82$. These last values are beginning to get closer. We're beginning to hone in on the desired velocity.

We see that the bold line connecting the point (1 s, $+0.567$ m) to the point (1.13 s, $+0.67$ m) is difficult to distinguish from the curve passing through (1 s, $+0.567$ m). If we magnify a piece of a smooth curve enough at any of its points, the curve looks progressively like a little straight line segment at that point. That line segment is called the tangent line to the curve at the point. So, in other words, the smaller and smaller we choose Δt, the closer and closer the line connecting (1 s, $+0.567$ m) to (1 s $+\Delta t$, x(1 s + Δt)) is to being the tangent line to the position versus time curve at the point of interest (i.e., [1 s, $+0.567$ m]). And, the *instantaneous velocity is the slope of the tangent line* at that point (about $+0.66$ m/s for our example).

Given a smoothly continuous position versus time graph (such as Figure 2.4a) we can make a graph of how velocity varies with time by estimating the slope of the tangent line to the curve at successive times and plotting the resulting values. We do this at some selected times and then connect our best estimates in order to obtain a smooth curve for a velocity versus time graph. In principle, one can imagine an automatic calculator that could move along the curve in Figure 2.4a continuously finding the tangent, computing its slope, and then plotting these values as we have done in Figure 2.5.

In Figure 2.5, several tangent lines to the position versus time curve (the lighter curve) are displayed. All have zero slope and the velocity graph at those corresponding times shows zero velocity. The associated instants in time correspond to "turning points," instants where the car changes direction. Between turning points the car moves continuously in one direction. Thus, from instant a to instant b the car moves toward the origin, and from instant b to instant c, the car moves away from the origin. While moving away from the origin (to more positive x-coordinates), the car's velocity is positive (the slope of the tangent line to the position versus time curve at any instant in this interval is positive) and while moving toward the origin (to less positive x-coordinates), the car's velocity is negative. Note that at the moments the car changes direction, its velocity is instantaneously equal to zero; that is, the car is instantaneously at rest.

FIGURE 2.5 *Velocity of the car obtained from its position versus time curve.*



If we had an equation for the curve in Figure 2.4a, that is, an explicit relation between x and t, we could utilize Equation (2.2) to determine an equation for how velocity varies in time. The translation of $x(t)$ into $v(t)$ is the heart of what we call calculus. These days, computers can do this translation for us.

You can see that the velocity of our car portrayed in Figure 2.5 varies in time, much as position does. Because velocity is rate of change of position, it is also useful to define rate of change of velocity. Indeed, as we show in Chapter 3, rate of change of velocity is the centerpiece of Newton's laws of dynamics.

The *average acceleration* is defined, in a similar way to the average velocity, as

$$\bar{a} = \frac{\Delta v}{\Delta t}, \tag{2.3}$$

where $\Delta v = v(t_f) - v(t_i)$. Note that the average acceleration reflects the change of the velocity with time and that in order to calculate the average acceleration from this definition, you must first have a graph of (or equations for) the velocity versus time and then obtain the ratio in Equation (2.3) for the time interval of interest. The average acceleration can be positive, negative, or zero depending on whether v is increasing (Δv is positive), decreasing (Δv is negative), or is the same at the two ends of the time interval of interest (regardless of what occurred during the interval of time). Acceleration is change in velocity per unit time, so its units are velocity units divided by time units: (m/s)/s = m/s^2, for the car example given above.

We define, analogous to instantaneous velocity, the *instantaneous acceleration* (or simply the acceleration) as

$$a = \lim_{\Delta t \to 0} \frac{\Delta v}{\Delta t}. \qquad (2.4)$$

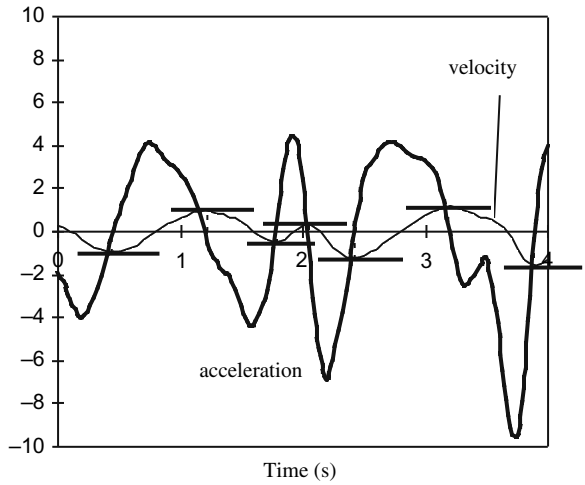


FIGURE 2.6 *Acceleration of the car obtained from the velocity data of Figure 2.5.*

Just as velocity at any instant (for motion in one dimension) is the slope of the tangent line to the position versus time curve at that instant, the acceleration at any instant (for motion in one dimension) is the slope of the tangent line to the velocity versus time curve. Thus, if we are given a plot of v versus t, we can approximate a versus t by sketching tangent lines at a number of instants, estimating the respective slopes, plotting those values, then interpolating. Starting with the velocity plot in Figure 2.5, we can then generate an acceleration plot, as in Figure 2.6. We identify several instants at which the acceleration vanishes by noting where the velocity versus time curve has tangent lines with zero slope. Note that the acceleration is not zero when the velocity is zero nor is the velocity zero when the acceleration is zero. The two quantities measure different things and it is important to keep them straight.

Previously, we said that the motion of our car (or any other object) should result in a position versus time graph that is both continuous and smooth, that is, with no holes (discontinuities) or sharp points (kinks). No holes ensures that the position doesn't abruptly change from instant to instant. No kinks ensures that the velocity doesn't abruptly change from instant to instant. The analysis of motion could continue with additional quantities, such as the time-rate-of-change of acceleration, and the time-rate-of-change of that, and so on. Remarkably, such additional quantities are unnecessary for a complete understanding of how objects move about. Newton's laws of motion, the subject of the next section, tell us that acceleration is the most complicated piece of motion analysis apparatus we need.

2. NEWTON'S FIRST LAW OF MOTION

The gist of the preceding section is that there is an intimate mathematical connection among position, velocity, and acceleration. In essence, if we know an object's position over time we can infer what its acceleration must have been; inversely, given its acceleration we can make inferences about its position. Although they are intertwined, mathematics and physics are not the same thing. In this section, we begin to probe the physical rules that underlie the mathematics of motion. Constant velocity can't be felt, but acceleration can be. What you feel when you accelerate is physics. Acceleration is the key that unlocks the secrets of much of the physical universe. *Constant velocity doesn't require an explanation, but acceleration does.*

Perhaps you are puzzled by the last sentence. Everyday experience tells us that to start a body moving we have to give it a push. When we stop pushing, the body comes to rest. In our everyday experience, rest is the natural state of things. In our everyday

experience, it is velocity that requires a cause. It took many centuries of human intellectual development before we (that is to say, Galileo, in the seventeenth century) recognized that our common experience is dominated by two phenomena that acting together obscure from us the truth about motion. One of these, gravity, makes things fall down. The other, friction, makes them stop.

It's a pity that Galileo didn't have an "air table" to play with. If he had, he wouldn't have had to work so hard to uncover the truth about motion. An air table has many holes in its top, through which jets of air can be squirted. Maybe you've seen an air table at a game arcade (or, perhaps in an introductory physics laboratory). Often hockeylike games are played on them using pucks that are levitated by the squirting air. When the air is turned off and the puck is pushed, it quickly comes to rest. Gravity makes the puck fall to the table and friction makes it stop. When the table is level and the air is on, the puck hovers in one place. The jets of air effectively cancel gravity out and render friction negligible. On a properly leveled air table once a puck is pushed it travels off at constant speed in a straight line, until it hits a sidewall. Between the initial push and when it hits the wall, no additional push is required to keep the puck going. The natural state of a body's motion is constant velocity (zero velocity, i.e., rest, is a special case). No external influence is required to keep the puck moving, however, an influence from outside is certainly required to change its velocity.

Isaac Newton, in his *Principia Mathematica* (1687), greatly extended Galileo's insight that change in motion requires cause. The first of Newton's laws is a kind of statement of faith. It says that

It is possible to find laboratories ("frames of reference") in which a body's acceleration is solely attributable to interactions between that body and other bodies.

In the laboratories of Newton's first law a body never accelerates spontaneously; every acceleration is caused by an interaction. That a body does not spontaneously accelerate is attributed to a property of all material objects called *inertia*. The frames of reference of Newton's first law are said to be *inertial frames*.

It is usually desirable to observe and describe motion in inertial laboratories, because in them every acceleration is caused by identifiable pushes and pulls and, as we show, the associated quantitative analysis is straightforward. Spontaneous accelerations observed in noninertial frames necessitate inventing fictitious causes for their explanation. For example, suppose you jump off the roof of a building (we are not recommending you do this!). You will notice that in the frame of reference you carry with you all objects—such as the building, people standing on the sidewalk below, and the Earth itself—accelerate towards the sky with exactly the same acceleration. There is no identifiable interaction that causes all of these simultaneous spontaneous accelerations. To explain them requires assigning a fictitious cause. You're carrying a poor frame of reference for doing physics, a fact that will be painfully apparent when the upward accelerating ground reaches you. People standing on the sidewalk will offer a simpler picture of what is occurring. They will say that it is you who is accelerating, and that there is an easily identifiable cause: the pull of gravity of the Earth. This situation is general: any frame of reference in which accelerations occur without cause must itself be accelerating.

There is another, perhaps more common, way to state Newton's first law, given our understanding of an inertial reference frame.

In inertial reference frames, objects traveling at constant velocity will maintain that velocity unless acted upon by an outside force; as a special case, objects at rest will remain at rest unless an outside force acts.

It's not hard for us to accept that an object at rest will remain at rest, but it is very hard to accept the fact that an object will move at constant velocity unless an outside force, one originating from another object, acts. Friction is so common in our experience that we often don't realize it is almost always present and acting to slow objects down.

Noninertial frames of reference abound. For example, while driving your car you rapidly accelerate from rest at a stoplight. A box of cookies on the seat next to you spontaneously slides toward the back of the seat and at the same time the trinket hanging from your rear view mirror also spontaneously accelerates to the rear. No object can be found that causes these accelerations. By speeding up, your car becomes an accelerated reference frame. Similarly, if you spin around on a lab stool you will observe all objects in your vicinity orbit around you in circles. Because they travel in circular paths in your reference frame, we show later that they must accelerate. But, again, no object can be identified as the cause of these accelerations. A spinning frame is noninertial.

The latter example draws attention to the following cautionary tale. As the day passes on Earth we see remarkable events in the sky. The sun rises and sets, seemingly orbiting the Earth in a circular path. Then the moon, the stars, and even the most distant galaxies do the same thing. All traveling in circles about the Earth, all, from our vantage point, therefore accelerating. To explain how all of these accelerated motions occur requires a very complicated picture of how the Earth could possibly cause them. A much simpler explanation is that the Earth is spinning: we, on the Earth, live in a noninertial frame of reference. Does that mean we have to leave the Earth in order to observe the validity of Newton's law(s)? That depends on what you want to measure. If you are doing an experiment that is completed in a few minutes and/or is confined to a small region of the Earth, the acceleration of your laboratory is probably ignorable. On the other hand, if you are interested in the motion of large volumes of air moving for hours above the Earth, for example, your acceleration will make what you see more difficult to explain. (The apparent circulation of winds around high and low pressure cells results from the acceleration of the Earth relative to the air. There is no body that can be identified as causing those circulations.)

3. FORCE IN ONE DIMENSION

The acceleration of any body is caused by interactions with other bodies. Dynamics is an exact mathematical formulation of the connection between acceleration and "interaction." How is the qualitative notion of "interaction" made mathematically precise? An interaction is a push or a pull. An interaction has a magnitude, or size, and a direction. In one dimension, say along the x-axis, there are only two choices for direction: along the positive x-axis direction or along the negative direction (right or left along the axis). We call such objects, with both a magnitude and a direction, *vector quantities*; a vector quantity in one dimension is simply a signed number measured in appropriate units. Examples of vector quantities from the first section of this chapter include position, displacement, velocity, and acceleration. Each of these has both a magnitude and a direction associated with it. On the other hand, quantities such as distance traveled or average speed do not have a direction and are called *scalar quantities*. We indicate vector quantities by placing an arrow over their symbol, for example, the acceleration vector $\vec{a}$. The simplest assumption we can make is that a physical interaction also can be represented mathematically by a vector quantity. We call such vectors *forces* and our first goal is to provide an operationally meaningful definition for force.

The definition of force we seek relies on a sequence of reasonable assumptions and their logical consequences. First, from our study of kinematics earlier in this chapter, we recall that acceleration, like force, also has a magnitude and a direction and is thus a vector quantity. Everyday experience suggests that when we push an

initially resting object in a given direction the object accelerates in that direction. So, we reasonably assume that when a body experiences a single interaction, the vector force (the cause) and the vector acceleration (the result) are parallel and that one is, at most, just a scalar multiple of the other.

Next, suppose a body experiences more than one interaction at any instant. Interactions are represented by force vectors, therefore we assume that the vector sum of the individual forces is equivalent to a single force that would yield the same acceleration. The vector sum in one dimension is simply obtained by adding the signed numbers representing the individual vectors. For example, given two acceleration vectors with magnitudes of 3 and 4 m/s^2, both pointing along the positive x-axis, the vector sum is 7 m/s^2 also along the positive x-axis, whereas if the second vector points along the negative x-axis, the vector sum of the two is $(3 - 4) = -1$ m/s^2, where the negative sign indicates that the direction is along the negative x-axis. Clearly it only makes sense to add two vectors that represent the same physical quantity, for example, accelerations. (Just as you shouldn't add "apples and oranges" because the result mixes the two kinds of fruit together and has no immediate interpretation, adding a force to a velocity doesn't make physical sense either.) Vector addition in one dimension can be generalized to add any number of vectors using simple arithmetic (Just adding positive and negative numbers). If the vectors we are adding are force vectors acting on an object, the vector sum represents the net force on the object. In particular, if a body is at rest or traveling with a constant velocity (i.e., not accelerating) the vector sum of all forces acting on the body must be zero, assuming we are in an inertial reference frame. We can exploit this quite reasonable assumption to develop a method for measuring force.

We know that all objects near the Earth fall if they are not supported. The cause of this downward acceleration is a field force. We say that the Earth is responsible for this force because it exerts a "gravitational pull" on all bodies in its vicinity. It is traditional to call the force of gravity of the Earth on any object the object's *weight*. We often measure weights by using a spring scale, such as the familiar hanging scales in a grocery store. When we place some tomatoes on a grocery scale, the tomatoes cause a spring to stretch and a needle to deflect. The deflection of the needle is taken to be a measure of the "weight" of the tomatoes. This happens primarily because the Earth somehow pulls the tomatoes down toward it and the scale somehow gets in the way and keeps the tomatoes from falling. The word more commonly used by physicists for a pull (or a push) is *force*. The force the Earth exerts on the tomatoes is called *gravity*. There's a wondrous thing about gravity: gravitational pulls exist even though the bodies involved don't touch. The Earth reaches out across empty space and pulls on the tomatoes. (Of course, the space between the Earth and the tomatoes isn't really empty: it's filled with air. But, we can get rid of the air, in a vacuum chamber, for example, and when we do we find that the pull of gravity is almost exactly the same.) Forces that exist across empty space are said to be *field forces*. In the field force picture, the Earth is viewed as creating a "gravitational force field" in the space around it. When the tomatoes are placed in the Earth's field they respond by falling toward the Earth. The scale, on the other hand, is doing something more directly to the tomatoes. It appears to stretch only when it is in direct contact with the tomatoes. The force the scale exerts on the tomatoes is an example of what is called a *contact force*. When the tomatoes hang from the scale without moving, the force down on them by the Earth is said to equal the force up on them by the scale.

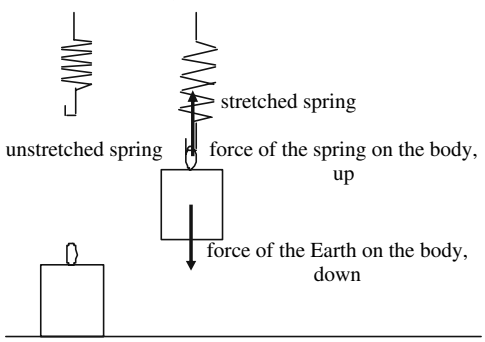


FIGURE 2.7 *Spring scale used to measure weight.*

This works because of a very useful property of springs. Suspend a simple spring from a fixed support. Attach an object to the free end of the spring and gradually lower the object until it can be let go and remain at rest. In this state of persistent rest, the object is not accelerating so the spring must be exerting an upward (contact) force on the object, balancing out the Earth's downward pull (field) on it. We note that the spring is stretched. The amount by which the spring has been stretched can be used to measure the force it is exerting. (See Figure 2.7.)

Suppose we have another object that is identical to the one that is already hanging from the spring. (We can check whether the weights of the two objects are identical by suspending them individually from the spring and noting that the stretch is the same in both cases.) Attach the second object to the end of the spring along with the first. We assume that these two bodies together are equivalent to a third body whose weight is twice that of the individuals. As long as the two hanging bodies are not too heavy (so that their combined weight does not permanently deform the spring) the new stretch is observed to be twice that when the spring is supporting just one of the bodies. In other words, the amount of stretch is directly proportional to the weight the spring supports, or, equivalently, the amount of stretch of a spring is a direct measure of how much force the spring exerts. Similarly, if we have two identical springs (two springs that stretch exactly the same amount when the same mass is suspended from each) and we hang a single weight by both springs as in Figure 2.8, we find that they each stretch by half the distance they would stretch if they each supported the full hanging weight. This should make sense because each spring is supporting half the weight with an equal upward force.

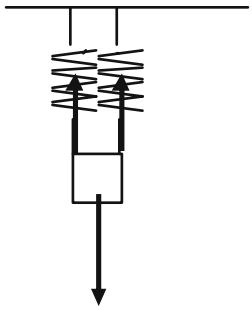


FIGURE 2.8 *Two identical spring forces each supporting half the weight of an object.*

In principle, we can imagine measuring any force on any object by replacing the force we are interested in by an appropriately calibrated, stretched spring (big stiff ones for large forces, and tiny flexible ones for small forces), keeping all other forces as before, and generating the same acceleration as when the replaced force is present. Because a spring exerts a force along its length, the direction of the spring corresponds to the direction of the replaced force and the stretch of the spring determines the force's magnitude.

4. MASS AND NEWTON'S LAW OF GRAVITY

The Earth isn't the only object that creates gravity. Every mass creates a gravitational pull on every other mass. You actually pull the tomatoes you weigh in the grocery toward you a little (and they pull you, too). It's just that the Earth's pull is so much greater than yours, you don't realize you're doing it. Mass plays two roles in producing a gravitational force. First, one mass creates a gravitational field in the space around it. Then, a second mass placed in the field of the first experiences a force due to the first's field. The two masses reciprocate in their pulls. The second makes a field of its own and the first, being in the field of the second, feels a force due to it. We say that a gravitational field has a direction—it points toward the mass making it—and a size, or *magnitude*. Let's call the magnitude of the gravitational field made by a mass M, g_M. The magnitude of the force this field produces when a mass m is placed in it is defined to be $F_{\text{of } M \text{ on } m} = mg_M$. Like mass and length, force has its own SI unit, the *newton* (N). (You don't find the newton in Table 1.1 because force is not defined as a fundamental quantity. It is expressible in terms of mass, length, and time, as we show in the next section. Because it is expressible in terms of fundamental units it is called a *derived unit*.) Gravitational field is gravitational force divided by mass, so the units of gravitational field are newtons per kilogram, N/kg.

We say that a body's weight (near the Earth) is the gravitational force the Earth exerts on that body. Thus, a mass m weighs

$$W_{\text{mass } m} = F_{\text{Earth on } m} = mg_{\text{Earth}} \tag{2.5}$$

SI units of mass (the kg), distance (the m), time (the s), and force (the N) were historically developed to be independent of the Earth's gravitational pull. Thus, a mass of 1 kg does not weigh 1 N, for example. Rather, under the SI conventions, we find that a mass of 1 kg near the Earth actually weighs about 9.8 N. Consequently, we say that the gravitational field of the Earth is about 9.8 N/kg near the Earth's surface.

Why is the condition "near the Earth's surface" important? Well, it turns out that the strength of a mass's gravitational field gets weaker the farther away one is from the mass.

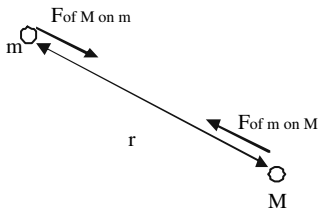


FIGURE 2.9 *Two masses attracting each other by the gravitational force.*

Very careful measurements in the laboratory show that if the centers of two uniform (i.e., no holes or irregularities), spherical masses, M and m, are separated by a distance r, then M pulls m with a gravitational force whose magnitude is given by (see Figure 2.9)

$$F_{M \text{ on } m} = G\frac{Mm}{r^2} \tag{2.6}$$

The quantity G is independent of which masses are interacting and any other physical condition. It is a so-called "universal constant" and in SI units its value is close to 6.67×10^{-11} $N\text{-}m^2/kg^2$. Equation (2.6) is known as *Newton's law of universal gravitation*. If we divide both sides of Equation (2.6) by m we get the gravitational field produced by M at a distance r from its center:

$$g_M = G\frac{M}{r^2} \tag{2.7}$$

Although Equations (2.6) and (2.7) are rigorously correct for uniform spherical masses, they can be applied to arbitrary shaped masses to obtain approximate values for gravitational forces and fields.

Example 2.1 What is the order of magnitude of the mass of the Earth?

Solution: The Earth is approximately a sphere with radius $R_E = 6.38 \times 10^6$ m (about 4000 mi) ~ 10^7 m. At the Earth's surface the r in Equation (2.7) is $r \sim 10^7$ m and we know that $g_{\text{Earth}} \sim 10$ N/kg at the surface. So, solving Equation (2.7) for M, we find $M_{\text{Earth}} \sim (10 \text{ N/kg})(10^7 \text{ m})^2/(10^{-10} \text{ N-m}^2/\text{kg}^2) \sim 10^{25}$ kg. (Make sure you see how the units work out. A careful calculation yields 5.98×10^{24} kg.) In other words, by making a laboratory measurement of G (and a measurement of R_E) it is possible to "weigh the Earth."

Example 2.2 What is the gravitational field of a typical person 1 m from the person?

Solution: The point of this example is to obtain an approximate value we can compare with the Earth's field. Thus, we treat the person as if she were a sphere of radius less than 1 m and take some typical value for mass, such as ~ 10^2 kg (remember, 1 kg weighs 2.2 pounds). One meter from the center of a 10^2 kg sphere the gravitational field due to that mass is $\sim(10^{-10}\ N\text{-}m^2/kg^2)(10^2 \text{ kg})/(1 \text{ m})^2 \sim 10^{-8}$ N/kg. Compared with the Earth's field this is a tiny value. No wonder a person weighing tomatoes doesn't affect the tomatoes very much.

Example 2.3 What is an accurate value of the Earth's gravitational field at an altitude of 300 km (about the altitude of the Space Shuttle when it is in orbit)?

Solution: Here we want to do a formal calculation to compare with 9.8 N/kg. Recall that in Equation (2.6) or (2.7) r is the distance from the center of the sphere causing the field. An "altitude" is a distance above the surface of the Earth, so that r equals R_{Earth} + 300 km. Now, a km is 1000 m, so 300 km = 3×10^5 m = 0.3×10^6 m and, therefore, $r = 6.38 \times 10^6 \text{ m} + 0.3 \times 10^6 \text{ m} = 6.68 \times 10^6$ m. Putting this value into Equation (2.7) along with $M_{\text{Earth}} = 5.98 \times 10^{24}$ kg results in a

gravitational field equal to 8.9 N/kg. In other words, where the Shuttle orbits, the Earth's gravitational pull is only about 9% less than at the Earth's surface. A Shuttle astronaut who weighs 150 pounds on Earth weighs about 137 pounds in orbit. The pull of Earth's gravity is what keeps weather and communications satellites and even the moon orbiting the Earth. The Earth's gravitational pull doesn't suddenly stop at the top of the atmosphere; it extends, in principle, "to infinity," getting weaker as r gets bigger as $1/r^2$.

The last statement may run counter to what you've heard or read about astronauts in orbit. In orbit, things are said to be "weightless." You've surely seen video of astronauts floating about aboard the Shuttle. If a 150 pound astronaut tried to step on a scale while in orbit, he wouldn't succeed in getting a reading, because the scale would float away. The resolution to the seeming contradiction that an astronaut can be apparently "weightless" and yet weigh 137 pounds requires knowing something about Newton's laws of motion, a topic we are just beginning to explore.

Thus far in this section we have been discussing the gravitational attraction of masses. Historically, in such discussions mass was referred to as gravitational mass, a property that produces gravitational fields leading to gravitational forces. We now turn to a seemingly different property of mass, inertia.

As mentioned previously, the fact that bodies are reluctant to accelerate is said to result from an intrinsic property of matter called inertia. A body's inertia can be assigned a numerical value, referred to as its mass. It is a remarkable law of nature that if two bodies experience the same net force (which we can check with calibrated springs) the ratio of the magnitudes of the resulting accelerations, a_1/a_2, has the same numerical value irrespective of what forces are acting, how the bodies were initially moving, or any other external aspect of the measurement (such as the time of day, the temperature, where the experiment is performed, and so on). With the same net force acting on each body, this ratio depends only on which two bodies' accelerations are being compared. The ratio must be directly related to an intrinsic property of the bodies. Furthermore, there is a kind of reciprocity between "heaviness" and acceleration: if body 1 feels heavier than body 2 (so that intuitively it would seem to have more mass) the ratio a_1/a_2 is less than 1, and vice versa. We define the ratio of the mass of body 2 to that of body 1 to be the numerical value of a_1/a_2 determined by exposing both to the same net force; that is,

$$\frac{m_2}{m_1} \equiv \frac{a_1}{a_2}. \tag{2.8}$$

More massive objects will experience smaller accelerations for the same force, with the accelerations inversely related to the respective masses. The unit for mass is the kilogram (kg, defined below). When used with the meter and second, the kilogram defines the SI (Système International) units (formerly known as the mks system of units). We can define the mass (m_2, say) of an object through this equation by using a standard of mass as another object ($m_1 = 1$ kg) and by measuring the accelerations of the two objects under the action of the same force (m_2 would then be just a_1/a_2 in kg).

Example 2.4 A body with mass equal to 1 kg is pulled across a leveled air table by a spring with constant stretch of 1 cm. The resulting acceleration of the 1 kg mass is observed to be 0.30 m/s^2. A second body of unknown mass is pulled by the same spring with the same constant stretch. The observed acceleration of the second mass is 0.45 m/s^2. What is the mass of the second body?

(*Continued*)

Solution: We assume that under the conditions cited, both bodies experience the same overall force due to the spring. Because the second body has a higher acceleration, we expect it has a mass less than 1 kg. We let $m_1 = 1$ kg and m_2 be the unknown mass. Then using Equation (2.8) we have

$$m_2 = (0.30 \text{ m/s}^2 / 0.45 \text{ m/s}^2) \cdot (1 \text{ kg}) \\ = 0.67 \text{ kg}.$$

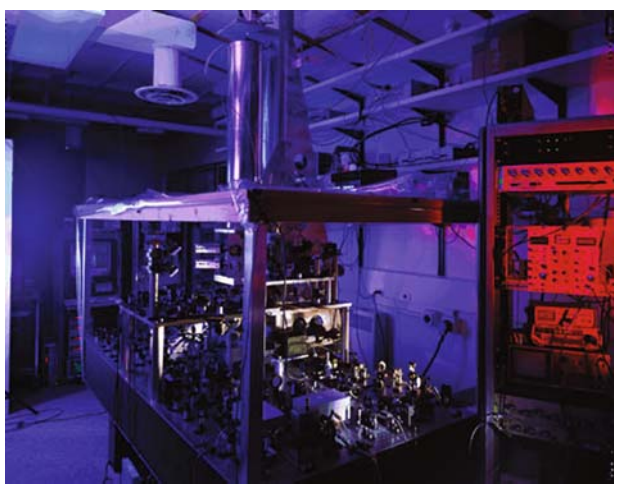
The procedure outlined above could be used, in principle, to measure the mass of any object. Of course, this is not done in practice because interactions (such as collisions) have the nasty potential for altering our standard and because the force that would impart a nice acceleration to an electron would imperceptibly perturb the motion of a kilogram. In practice, a wide range of secondary mass standards has to be used to measure unknown masses.

The standard kilogram (kg) is a platinum–iridium alloy cylinder kept at the International Bureau of Weights and Measures. Incidentally, standards for the meter and second are defined more reproducibly: the second is defined as the time needed for 9,192,631,770 vibrations of a cesium atom (a so-called atomic clock) and the meter is defined as the distance traveled by light in a vacuum in a time of 1/299,792,458 s (Figure 2.10). This, in fact, defines the speed of light in vacuum to be exactly $c = 299{,}792{,}458$ m/s. In other words, the speed of light was so well determined that in 1983 the meter was redefined so as to fix the speed of light.

Although fractions and multiples of kilograms suffice for quantifying mass in many situations, in the microworld of atoms and molecules another mass unit is more useful: the atomic mass unit (u) is defined to be exactly 1/12 of the mass of a neutral "carbon twelve" atom (an atom with 6 protons, 6 neutrons, and 6 electrons, often designated by the symbol ^{12}C). The atomic mass unit is preferred over kilograms when dealing with molecules because $1 \text{ u} = 1.66 \times 10^{-27}$ kg, and the latter is a very small and ungainly number with which to deal. The term dalton (D) is sometimes used to denote the same mass unit.

To recap this section on mass, we have discussed mass from two seemingly different approaches: gravitational mass, through Newton's law of gravity, which produces gravitational fields and forces on other masses, and inertial mass, defined through the acceleration produced by forces acting on the mass. Gravitational mass is a "static" mass with no motion required, gravitational fields and forces depending only on gravitational masses and distances. Inertial mass, on the other hand, is a "dynamic" mass, defined in terms of the acceleration response of the inertial mass to a given force of any kind. It is not necessarily apparent that these two concepts should lead to the exact same number for the mass of an object, but we have used the same symbol m for each because it has been shown that these masses have the same value to within better than 1 part in 10^{12}. This equivalence of inertial and gravitational mass has been a subject of discussion and experiment since Galileo and is still under active research.

FIGURE 2.10 *An atomic clock at NIST (National Institute of Standards and Technology) with an accuracy of about 1 s in 20 million years.*



5. NEWTON'S SECOND LAW OF MOTION IN ONE DIMENSION

Newton's first law tells us that in an inertial frame of reference a body accelerates only when it experiences a net force due to all other bodies. Equipped with the definitions of force and mass given above, the idea embodied in Newton's first law—that acceleration has a cause—can be made more precise. Thus, *Newton's second law of motion* says that

In an inertial frame of reference, the acceleration of a body of mass m, undergoing rigid translation, is given by

$$\vec{a} = \frac{\vec{F}_{\text{net on m}}}{m}, \tag{2.9}$$

where $\vec{F}_{\text{net on m}}$ *is the net external force acting on the body (i.e., the sum of all forces due to all bodies other than the mass m that push and pull on m).*

Embedded in Newton's second law are several important notions. (1) The law says that when the acceleration of a body arises from forces, the acceleration is caused by agents outside the body. A body cannot accelerate itself. Acceleration requires external force. (2) When there is a net (unbalanced) force on a body, the acceleration is in the same direction as the net force. The constant of proportionality that converts force into acceleration is the reciprocal of the body's mass. For a given force, the larger the mass, the smaller the acceleration, and vice versa. (3) Finally, as stated here, Newton's second law is applicable to a body in rigid translation, a body whose extent in space is ignorable, a point particle. For bodies that are tumbling or flexing or breaking into pieces the law of motion stated above has to be clarified and supplemented in ways we examine later.

Note that according to Equation (2.9), force has the units of mass times acceleration. Thus, in SI units one unit of force is equal to 1 kg-m/s^2. Because of the central role that force plays in describing nature, force units are given their own name. Honoring the founder of dynamics, 1 kg-m/s^2 is defined as 1 newton (1 N). (For calibration, a quarter pound hamburger with its bun, but minus the tomato and pickle, weighs about 1 N.)

Mass should be carefully distinguished from weight. *Mass is an intrinsic property of an object whereas weight is the magnitude of the force of the gravitational pull of the Earth*. If a body is in free fall, Equation (2.9) says

$$a = g = \frac{F_{\text{gravity}}}{m}, \tag{2.10}$$

where g is the magnitude of the acceleration due to gravity (9.8 m/s^2 near the Earth's surface). The force F_{gravity} is due to the pull of the Earth on the body whose mass is m. The magnitude, mg, of the gravitational force is also called the body's *weight*. A 1 kg mass thus weighs 9.8 N, because, for such a body, $F_{\text{gravity}} = 1 \text{ kg} \times 9.8 \text{ m/s}^2$. Note that weight exists whether or not the object is actually accelerating downward with acceleration g. A 1 kg body resting on a table near the surface of the Earth still weighs 9.8 N; the downward pull of the Earth on it must be canceled by an upward force of 9.8 N exerted by the table to keep it at rest. The weight of an object will vary depending on its location. For example, an object on the moon's surface weighs only about 1/6 what it does on Earth. This difference is due to the difference in the gravitational pull of the moon and has to do both with the moon's mass and radius compared to those of the Earth.

Equation (2.9) can be used to extract acceleration information from known forces or force information from known acceleration. For example, if all the forces acting on a particle of a given mass are known at every instant, the acceleration of that particle for every instant can be determined from the forces. Then, by measuring the particle's position and velocity at any one time, this dynamically inferred acceleration can be used (along with the methods we study in the next chapter) to predict the entire future motion of the particle, as well as deduce its entire past motion. Alternatively, if a complete record of a particle's motion is available, the particle's acceleration for every instant can be calculated from kinematics and forces required to produce that motion can then be determined.

Example 2.5 Television pictures are created by the collisions of a narrow beam of rapidly moving electrons with phosphor molecules on the screen of the picture tube. Suppose an electron (mass = 9.1×10^{-31} kg) in a TV is released from rest. After release it experiences a constant electrical force of 0.001 pN (where 1 pN = 1 piconewton = 10^{-12} N). What is the electron's acceleration under this force?

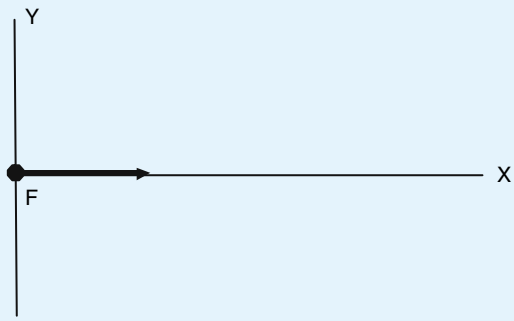


FIGURE 2.11 *An electron, initially located at the origin experiences a constant force F.*

Solution: We choose a coordinate system with the x-axis lined up along the direction of the constant force and with the origin where the electron is released (see Figure 2.11). The magnitude of the acceleration is found from Newton's second law

$$a_x = F_x/m = 0.001 \times 10^{-12}\ \text{N}/\ 9.1 \times 10^{-31}\ \text{kg} = 1.1 \times 10^{15}\ \text{m/s}^2.$$

Because the force is constant throughout this region of space, the acceleration remains constant there as well, always pointing along the x-axis. Note that gravity pulls the electron toward the Earth with an acceleration equal to about 10 m/s^2. The electrical force on the electron in this picture tube is about 10^{14} times larger than gravity! TV designers don't have to worry about gravity making their pictures sag.

Newton's second law has a wonderful range of validity and usefulness. It can be used to aim electrons to make a better TV picture. It can tell us how macromolecules vibrate and tumble in a cell when DNA is undergoing replication. It allows us to design more effective brakes to make cars safer. With it we can calculate the trajectories of planets and rocket-launched satellites to explore the bodies of our solar system. (A powerful example of such calculations is the collision of the comet Shoemaker–Levy 9 with the planet Jupiter in which the collision time was predicted with tremendous accuracy (Figure 2.12).) Newton's second law is arguably one of the central ideas of all of physics. You certainly could do less important things than practice the mantra, "Acceleration is net force over mass; acceleration is net force over mass, ..."

6. NEWTON'S THIRD LAW

According to Newton's second law, acceleration requires force from outside. Swimming fish, flying birds, and human bicyclists all accelerate because something pushes on them, according to the second law. At first, that may sound preposterous. For example, think of what it feels like to increase your speed while running. You feel strain in the muscles of your legs. Or, accelerate your car to pass on a highway. You have to push down the gas pedal. Obviously, in both cases something internal is causing the acceleration.

Well, that's not exactly correct. Suppose you are asked to exert the same strain in your legs but instead of running on a dry track you are placed on a beach with loosely packed,

dry sand. The same effort doesn't result in nearly the same acceleration. If you are placed instead on an ice rink, the same effort produces even less of an outcome. Finally, if you were put in a space suit and placed in the vacuum of space outside the Space Shuttle, moving your legs with the same strain as before would produce no acceleration at all. Clearly, moving your legs is important in producing acceleration, but what you are standing on is also important. You have to be able to push against something. That is equally true for fish and birds and accelerating cars.

The reconciliation of examples of apparent self-propulsion with Newton's second law, which says that self-propulsion is impossible, requires another law of motion:

When one body exerts a force on a second body, the second exerts a force in the opposite direction and of equal magnitude on the first; that is,

$$\vec{F}_{2 \text{ on } 1} = -\vec{F}_{1 \text{ on } 2}$$

This law, *Newton's third law of motion*, is sometimes referred to as the law of action–reaction: every "action" generates an equal and opposite "reaction." Thus, the feet of a runner do not accelerate the runner. Rather, the feet exert a force on the track, and it is the reaction force of the track back on the feet that accelerates the runner. When you run on a track a given effort leads to a certain push on the Earth; the Earth pushes back on you and that push results in your acceleration. When you run in loose sand, or on ice, you can't exert the same force on the Earth as you can by pushing on a dry track; the weaker push by you on the Earth is reciprocated with a weaker push back, and, therefore, less acceleration. In space, running doesn't result in an acceleration because there is nothing to push against and therefore nothing to push on you.

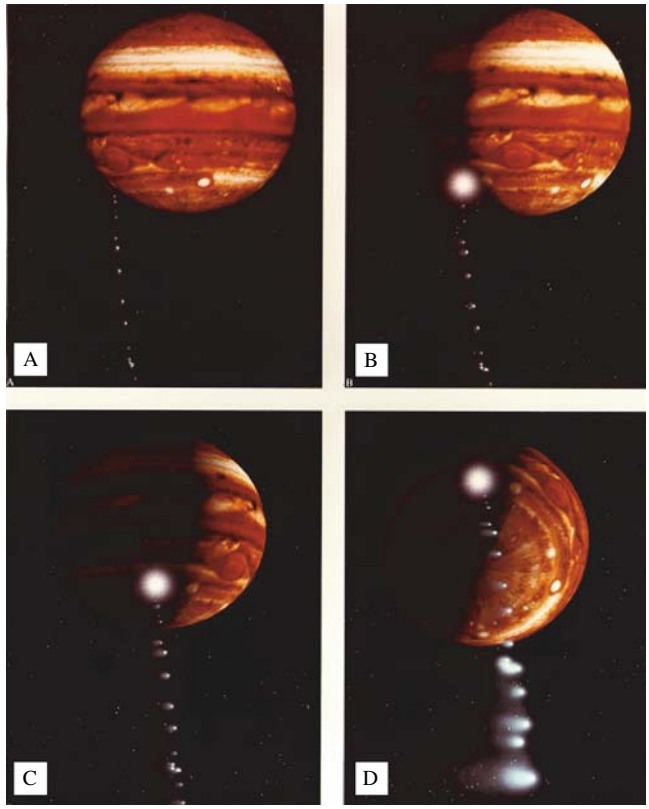


FIGURE 2.12 *Time series showing the collision of a comet with Jupiter in July 1994 as detected by the Galileo satellite probe; the comet, made from over 20 fragments, had been tracked for a year and the location and time of the impact, the first-ever observed collision of two solar system objects, had been calculated very precisely.*

Example 2.6 Newton's third law can be a source of confusion to someone who is thinking about such things for the first time. Here's an example. A young woman kicks a soccer ball 30 m downfield. But how? (Caution: The reasoning that follows contains an error! Can you spot it?) That is, Newton's third law says that the force of her foot on the ball is exactly countered by a reaction force exerted by the ball on her foot. The two are equal in magnitude and oppositely directed. The sum of two equal and opposite forces is zero, so according to Newton's second law, if there is no net force, no acceleration is possible. But, of course the ball does go downfield, so what goes on?

Solution: The wording of this problem illustrates a common pitfall in applying Newton's laws of motion. You have to be careful about identifying what is the body of interest and what are its surroundings. If we are interested in the flight of the soccer ball, then we have to keep track of the forces on the ball, and only those forces. If we are interested in the motion of the woman's foot, then we have to keep track of the forces on her foot. The foot exerts force on the ball and the ball accelerates as a result. The ball exerts a force on the foot and the foot accelerates (slows down) as a result. The two forces are equal and oppositely directed, however, they act on different bodies and each produces its own acceleration. The two don't act together on any one body and the fact that they add up to zero is irrelevant for understanding what happens to the ball.

You may be tempted, in thinking about this example, to say something like, "Well, the ball goes downfield because the woman is more powerful or more massive than the ball." Resist that temptation if you feel it creeping up on you. Keep in mind that a not very powerful nor massive 50 kg woman can easily accelerate a 1000 kg car (in neutral, with its brakes off, on a horizontal surface) by pushing it.

Example 2.7 Two ice skaters, a 90 kg father and his 40 kg daughter standing face to face and holding hands, push off from each other with a constant force of 20 N (Figure 2.13). Find their accelerations during the time they are pushing each other.

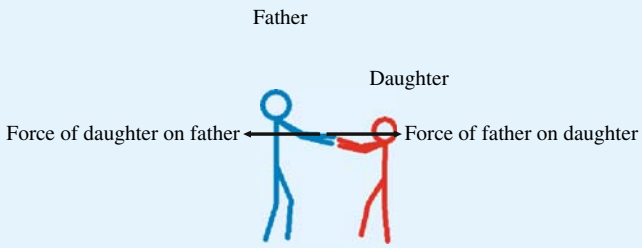


FIGURE 2.13 *Two ice skaters pushing off from each other.*

Solution: Each skater exerts a 20 N force on the other. Assuming there are no other horizontal forces acting, the man's acceleration will be $a_{man} = 20\ \text{N}/90\ \text{kg} = 0.22\ \text{m/s}^2$ to the left, whereas the girl's acceleration will be $a_{girl} = 20\ \text{N}/40\ \text{kg} = 0.5\ \text{m/s}^2$ to the right. These accelerations occur only during the time when the skaters are pushing against each other. Note that no matter which person (or both) actually takes the active role in doing the pushing, the force on each person has the same magnitude.

Example 2.8 A book lies at rest on a horizontal table. Identify all forces acting on the book and for each identify the appropriate reaction force.

Solution: The forces labeled "1" and "3" in Figure 2.14 are forces on the book. Forces "2" and "4" are exerted by the book in reaction to "1" and "3". Force "1" is the book's weight. It is due to the Earth's gravitational field. If the Earth pulls on the book, Newton's third law says that the book must pull back on the Earth with a force of equal magnitude. The reaction force to "1" is a gravitational pull exerted by the book on the Earth, and is labeled "2" in the figure. Its magnitude is the same as the book's weight. The force "3" is an upward force exerted by the table on the book because of contact between the table and the book. We know there is such a force because we know the book lies at rest, so the net force on it must be zero. When the force exerted on the book by the table is added to the force exerted on the book by the Earth, the two cancel. Clearly, the upward force of the table on the book must also have the same magnitude as the book's weight. The reaction force to "3" is a contact force, "4," exerted by the book on the table. It points down and it, too, has the same magnitude as the book's weight but it is not the book's weight. If suddenly a hole bigger than the book opened in the table below it, both "3" and "4" would suddenly disappear, but the book's weight "1" and the reaction force "2" would still exist.

So, if the force "2" is due to a gravitational pull of the book how come the Earth doesn't accelerate toward the book with an acceleration g? Newton's

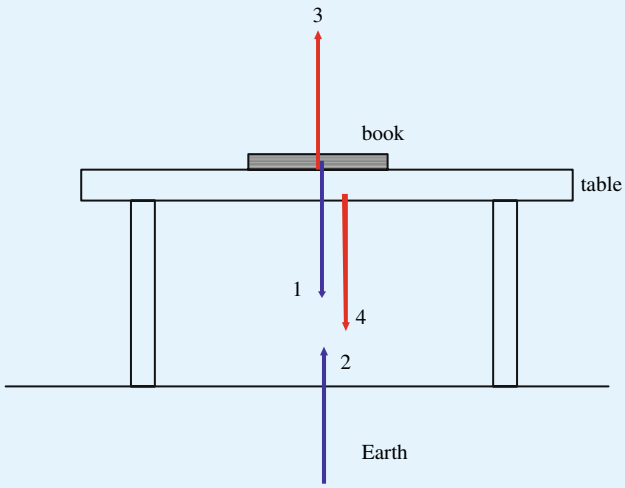


FIGURE 2.14 *Forces involved with a book on a table. Forces 1 and 3 act on the book, whereas 3 and 4, and 1 and 2 represent action–reaction pairs (see discussion of Example 2.8).*

third law says that action–reaction forces are equal, not the accelerations they produce! To find out about those, use Newton's second law: the magnitude of the Earth's acceleration is the magnitude of the force on it divided by the Earth's mass. In other words,

$$a_{\text{Earth}} = \frac{F_{\text{book on Earth}}}{M_{\text{Earth}}} = \frac{m_{\text{book}}\, g}{M_{\text{Earth}}} = \left(\frac{m_{\text{book}}}{M_{\text{Earth}}}\right) g$$

(remember, the magnitude of the force exerted by the book is equal to the book's weight) and because the ratio of the mass of the book to the mass of the Earth is on the order of 10^{-25} the book's pull on the Earth produces a negligible acceleration. Of course, if the book had a lot more mass—like that of another planet—and was as close to the Earth as the book (fortunately, the pull of gravity also depends on distance) then the acceleration of the Earth would not be negligible. But, that's another story.

7. DIFFUSION

An E. coli bacterium typically swims in a straight line for some distance, during which time its flagella undergo a coordinated helical motion driven by a rotary molecular motor located in the membrane at the flagella attachment sites (we study this molecular motor further in Section 3 in Chapter 7; see also Figure 1.2 for a cartoon sketch). In response to external stimuli of, for example, nutrient or oxygen level, the molecular motor may reverse and cause the flagella to become uncoordinated, resulting in a characteristic "twiddling" motion in which the bacterium randomly gyrates about, before finally taking off in a straight-line trajectory in some other direction. *E. coli* have been shown to respond to variations in environmental factors, being attracted to higher levels of nutrients and oxygen and repelled by poisons; this response is known as chemotaxis. If the *E. coli* are either killed or have their flagella removed they are no longer motile but they still move due to a phenomenon known as *Brownian motion*, named after Robert Brown who in 1827 noticed the random thermal motions of

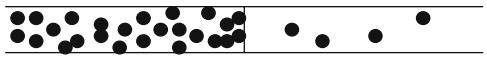


FIGURE 2.15 *Diffusion will tend to equalize the numbers of molecules in the left and right sides of the initially sharp boundary.*

suspended pollen grains under a microscope. Rapid and numerous collisions of solvent molecules with the *E. coli* produce random erratic motions. The Brownian motions of such "killed" *E. coli*, as well as the random motions of the solvent molecules themselves, are examples of a general process known as *diffusion*, which is the term for such thermally driven motions at the molecular level.

Although diffusion appears, at first glance, to be random and incapable of resulting in useful or interesting results, diffusive phenomena abound in the biological and physical world. In biology, diffusion is the process that controls both the exchange of oxygen in the hemoglobin of our red blood cells and the elimination of wastes in our kidneys. Whenever molecules move from one place to another without the expense of energy specifically earmarked for that motion, it is by diffusion; for example, diffusion controls the passive transport of molecules across a membrane and stored chemical energy is required for the process known as active transport.

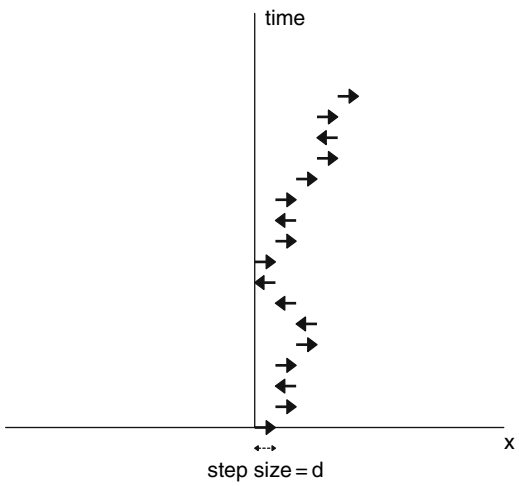
Often when there are concentration differences across macroscopic distances diffusion will play a role in reducing those differences. In these cases, even though the motion of each individual molecule may be random in direction, the collective motion that affects the local concentration of molecules can be directed. For example, in the case of one-dimensional diffusion, suppose there is a sharp spatial boundary in the concentration of some molecules as shown in Figure 2.15. Then even though any particular molecule is equally likely to move left or right, as time evolves, the variation tends to disappear because, on average, there are more molecules in the higher concentration region moving into the lower concentration region. Examples of just this type of diffusion are the oxygen and waste transport in the blood and kidneys cited above. In general when there are initial concentration variations and no active, energy-consuming processes occurring, diffusion tends to result in a uniform final state. We show the connection of this randomization process to the science of thermodynamics in Chapter 13.

The mathematics of diffusion in one dimension can be described by a related problem known as the *random walk*. Suppose that one starts at the origin and takes equal length steps in either the positive or negative x-direction with equal probability (this is also known as the drunkard's walk problem). Without regard for the details of the mathematics, it is clear that the average position of the person after many steps is still at the origin since positive or negative steps are equally likely and the average is simply computed by adding up the (plus and minus) displacements. On the other hand, it should also be clear that as time goes on, it will become more and more possible that the person will be found farther away from the origin. We can characterize this motion by calculating the average of the squares of the displacements, because these will all be positive quantities and cannot average away to zero. A calculation shows that this mean square displacement, $<(\Delta x)^2>$, is given by

$$<(\Delta x)^2> = Nd^2,$$

where N is the number of steps, d is the step size, and the brackets $< >$ indicate taking the average value (Figure 2.16).

FIGURE 2.16 *One-dimensional random walk with equal step size and time interval.*



The one-dimensional diffusion of a "killed" *E. coli* can be solved using mathematics similar to the random walk problem, but clearly the step size and number of steps do not directly apply. The analogous equation for the mean square displacement of a diffusing bacterium is given by

$$<(\Delta x)^2> = 2Dt,$$

where t is the elapsed time and D is a constant known as the diffusion coefficient, which is a property of the size and shape of the bacterium as well as of the viscosity (a measure of "stickiness") and temperature of the liquid medium in which the bacterium is found. It turns out that as this result is generalized to two (or three) spatial dimensions of

motion, the mean square displacement has an additional 2 Dt (or 4 Dt), so that in three dimensions

$$<(\Delta r)^2> = 6\,Dt. \tag{2.11}$$

The square root of the mean square displacement (known as the *root mean square* or *rms displacement*) is thus proportional to $\sqrt{t}$, a result that is very different from the linear t-dependence for a particle moving with constant velocity. Although diffusing particles may move rapidly over short times, because of their constant random changes in direction, the overall average displacements change much more slowly with time. The characteristic $\sqrt{t}$ signature of displacements in diffusion appears often in our discussions of many physical as well as biophysical phenomena. For example, we show that electrical and thermal conductivities are closely related to the diffusion of loosely bound electrons in a metal.

Example 2.9 The diffusion coefficient for sucrose in blood at 37°C is 9.6×10^{-11} m^2/s. (a) Find the average (root mean square) distance that a typical sucrose molecule moves (in three dimensions) in 1 h. (b) Now find how long it takes for a typical sucrose molecule to diffuse from the center to the outer edge of a blood capillary of diameter 8 μm.

Solution:

(a) Simple substitution finds the rms distance to be equal to

$$\sqrt{6Dt} = \sqrt{6 \cdot 9.6 \times 10^{-11}\ \text{m}^2/\text{s} \cdot 3600\ \text{s}} = 1.4 \times 10^{-3}\ \text{m}.$$

(b) This is a problem in two dimensions (in a cross-sectional plane of the capillary), so that from the above discussion, the relationship between the mean square distance and the time is $<(\Delta r)^2> = 4\,Dt$. Substituting $\Delta r = 4\ \mu\text{m} = 4 \times 10^{-6}$ m, we find that

$$t = \frac{<(\Delta r)^2>}{4\,D} = \frac{(4 \times 10^{-6}\ \text{m})^2}{4 \cdot 9.6 \times 10^{-11}\ \text{m}^2/\text{s}} = 0.04\ \text{s}.$$

Note that this answer for the time scales as the square of the capillary radius and so increases by a factor of 4 for a capillary of twice the radius. This example demonstrates why capillaries need to be so small in order to carry out efficient exchange of food and wastes between the blood and surrounding tissue.

CHAPTER SUMMARY

In one dimension, starting with the concept of displacement Δx, velocity and acceleration are defined as

$$v = \lim_{\Delta t \to 0} \frac{\Delta x}{\Delta t} \quad \textit{and} \tag{2.2}$$

$$a = \lim_{\Delta t \to 0} \frac{\Delta v}{\Delta t}, \tag{2.4}$$

where the average values over a time interval Δt are equal to these expressions without taking the limit.

The gravitational force between any two masses is given by Newton's universal law of gravity,

$$F_{M\ \text{on}\ m} = G\frac{Mm}{r^2}. \tag{2.6}$$

(*Continued*)

For a mass near the Earth's surface, this force is equal to its weight,

$$W_{\text{mass } m} = F_{\text{Earth on } m} = mg_{\text{Earth}}, \qquad (2.5)$$

with $g_{\text{Earth}} = 9.8 \text{ m/s}^2$.

Newton's second law states that

$$\vec{a} = \frac{\vec{F}_{\text{net on m}}}{m} \qquad (2.9)$$

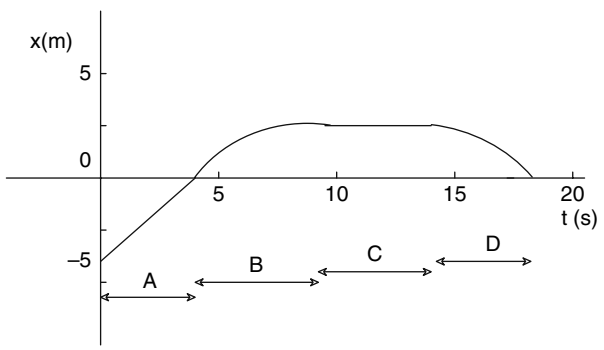
and in the absence of a net force, the acceleration must be equal to zero, a statement equivalent to Newton's first law. The third law is a statement that all forces arise from interactions between pairs of objects; the two forces (action and reaction) each act on one of the objects and are equal in magnitude, but opposite in direction.

Unlike directed motion, diffusion is a random thermal process in which the average displacement is zero, however, the mean squared displacement is given by

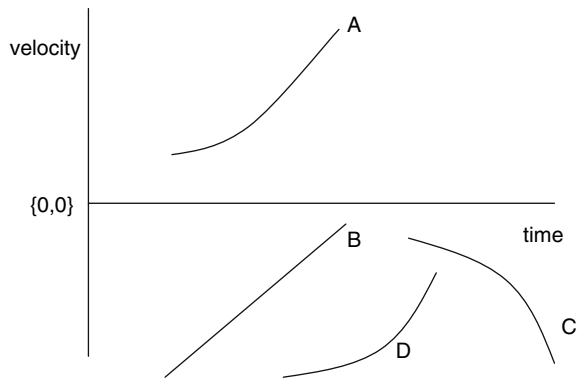
$$<(\Delta r)^2> = 6\,Dt. \qquad (2.11)$$

QUESTIONS

1. As a car moves steadily down a road, we can deduce the motion of the car by following the motion of only one piece, for example, the corner of a fender or the license plate. However, the motion of the piece only conveys complete information about the rigid structure of the car. Describe the motion through space of each of the following as a car moves forward: a tire air valve, the tip of a working windshield wiper, the top of an engine piston, and the label on a fan belt.
2. As a person runs, describe the motion through space of a wrist, a kneecap, and an elbow.
3. In the figure the position of an object is shown as a function of time. Indicate whether the velocity and acceleration in each labeled interval are positive, zero, or negative.



4. In the figure the velocity of four different objects is shown as functions of time. Indicate whether the velocity and acceleration for each labeled object are positive, zero, or negative.



5. Is the average velocity during an interval of time always equal to the sum of the initial and final velocities of the time interval divided by two? If not, give an example showing why not.
6. When an object free falls, does it travel equal distances in equal time intervals? Does its velocity increase by equal amounts in equal time intervals?
7. In each of the following situations, first identify all the forces acting on the object and then, for each force, identify the reaction force and its source:
   (a) A bird flying through the air
   (b) A horse pulling a cart
   (c) A person riding in an elevator that is accelerating upwards
   (d) A hot air balloon hovering in place
   (e) A ladder leaning against a wall.
8. A VW bug has a terrible head-on collision with an 18-wheeler truck. Which vehicle experiences the greatest force on impact? The greatest acceleration?
9. Tell whether the following pairs of forces are action–reaction pairs, and include a statement about your reasoning.

(a) The weight of a fish and the buoyant force holding it up
(b) The centripetal force on a protein molecule in a centrifuge and the force the protein exerts on the solvent surrounding it
(c) The weight of a free-fall skydiver and his frictional drag after reaching a terminal velocity
(d) The thrust on a jellyfish and the force the jellyfish exerts on the jet of water it expels
(e) The frictional force that allows you to walk and the force you exert horizontally on the Earth

10. Describe some situations in which forces act on an object but there is no motion. How can this occur?

11. What is the difference between mass and weight?

12. Which of the following situations involve field forces and which contact forces: a tug-of-war, moving paper clips around with a horseshoe magnet, riding a Ferris wheel, getting a shock when you reach for a door knob, a ball falling through the air, a train rolling on tracks, a levitated train traveling at over 340 min/h.

13. Two equal masses attract each other with a gravitational force of 18 pN. If their separation is tripled what will the gravitational force between them be?

14. A mass produces a gravitational field g at a point. If the mass is doubled and moved twice as far away from the point, what will the new gravitational field be?

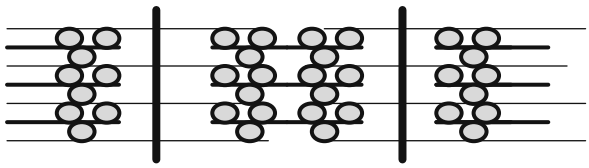
15. Discuss how you think scientists were able to determine the mass of the sun.

16. Explain why even though an astronaut in orbit around the Earth is weightless, she must exert a force in order to propel herself across the spaceship.

17. A person riding on the "whip" at an amusement park watches an ice skater coast by. The ice skater believes that she is coasting in a straight line at a constant speed. How does the person on the "whip" describe her motion? This same person believes that Newton's first law is violated for the ice skater. Why is he wrong?

18. Muscle basically consists of interdigitating thick and thin filaments that interact via cross-bridges (the "heads" of myosin molecules). Because the force a myosin head exerts on an actin thin filament is equal and opposite to the force the actin exerts back on the myosin head and thereby the thick filament, how can the muscle generate any force?

19. The detailed structure of a muscle fiber includes a series of Z-lines with actin thin filaments of opposite polarity on either side and with thick filaments not attached to the Z-lines as shown. The cross-bridge interactions tend to shorten the distance between neighboring Z-lines when a muscle contracts, but should not a given Z-line feel symmetric forces from the equivalent thin filament interaction on either side, and hence not feel a net force?



20. In each of the following cases, identify the interaction pairs of forces and draw a free-body diagram of the object in italics: (a) a *book* resting on a table; (b) a *book* resting on a table with a paperweight on top of the book; (c) a *cart* being pulled by a horse along a level road; (d) a heavy *picture* being pushed horizontally against the wall to hold it in place.

21. What causes diffusion? If a container is kept perfectly still, without any vibrations on it whatever (e.g., covered, in a draft-free room, atop a granite block mounted on shock absorbers) will diffusion occur within it?

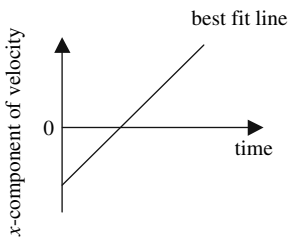
22. Why doesn't a drop of dye, when added to water, simply grow outward uniformly from the position at which it is first placed? (Or does it?) If you carefully put one drop of cream atop a mug of coffee, what happens to it? Is there any way to keep the added drop from diffusing?

MULTIPLE CHOICE QUESTIONS

1. The x-position of a particle is sampled every 0.5 s, as in the following table.

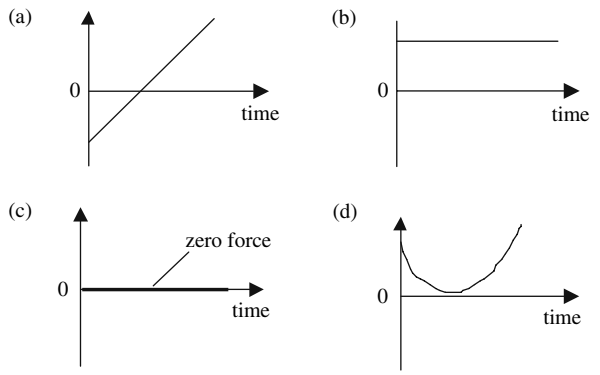
Time (s)	*x-Position (m)*
0.0	+3.0
0.5	+2.2
1.0	+3.0
1.5	+1.0
2.0	−0.5

Which one of the following must be true? (a) The x-component of the average velocity in the interval 0.0 s to 1.0 s is 0.0 m/s. (b) The average speed in the interval 0.0 s to 1.0 s is 0.0 m/s. (c) The x-component of the instantaneous velocity at 1.0 s is +3.0 m/s. (d) The x-component of the instantaneous velocity throughout the interval 1.0 s to 2.0 s is always negative.



2. The x-component of a particle's velocity is sampled every 0.5 s. The data are fit with a straight line as shown in the figure to the right. Assuming the fit is a

good approximation to the motion, which of the following best represents the x-component of the net force on the particle as a function of time?



3. A 9.8 N force causes a 1 kg mass to have an acceleration of 9.8 m/s^2. This situation is most closely related to Newton's (a) first law of motion, (b) second law of motion, (c) third law of motion, (d) law of universal gravitation.
4. A woman weighing 500 N stands in an elevator that is traveling upward. At a given instant the speed of the elevator, as well as that of the woman, is 10 m/s and both are decreasing at the rate of 2 m/s^2. At that instant, the floor of the elevator exerts a force on the woman that is (a) about 400 N, pointing up, (b) 500 N, pointing up, (c) 500 N, pointing down, (d) about 600 N, pointing up.
5. A soccer ball approaches a soccer player with a speed of 10 m/s. The player heads the ball with the net result that the ball travels off in the opposite direction with a speed of 15 m/s. The player stays more or less in place. During the time the player's head is contact with the ball the head exerts an average force of magnitude 100 N. Which one of the following is true concerning the magnitude of the average force the ball exerts on the player's head during that time? (a) It must be about zero because the head doesn't move much. (b) It's hard to say from the information given, but it certainly must be less than 100 N or else the ball wouldn't reverse direction. (c) Nothing can be said about the magnitude of the force because neither the mass of the ball nor the time of contact is given. (d) It's 100 N.
6. A bicyclist rides for 20 s along a straight line that corresponds to the $+x$-axis covering a distance of 400 m. She then turns her bike around; that takes another 20 s. Finally, she rides back to where she started (400 m in the $-x$-direction) for 40 s. The average velocity for this trip is (a) 0, (b) +3, (c) +10, (d) +15 m/s.
7. A ball is thrown directly upward. After leaving the hand the ball is observed to be at a height A and rising. A little while later, the ball is at height B and is instantaneously at rest. Later still the ball is observed to be height C and falling. All during the flight the ball is in free-fall. The acceleration of the ball (a) points up at A, is 0 at B, and points down at C; (b) points up during each portion of the flight; (c) is zero during each portion of the flight; (d) points down during each portion of the flight.
8. An object is thrown straight up. At the top of its path (a) the velocity is zero and the acceleration is zero, (b) the velocity is zero and the acceleration is equal to the weight, (c) the velocity is down and the acceleration is equal to g, (d) the velocity is zero and the acceleration is equal to g.
9. Newton's law of gravitation says that the magnitude of the gravitational force of a body of mass M on a body of mass m is GMm/r^2. The fundamental dimensions of Newton's Gravitational Force are (a) [M][L][T]$^{-2}$, (b) [M]^2[L]$^{-2}$, (c) [M][L][T]$^{-1}$, (d) [M][L]^2[T]$^{-2}$. (Here [M] represents mass, [L] length, and [T] time.)
10. Given that the Earth is about 1.5×10^{11} m from the sun and takes a year (about 3.1×10^7 s) to make one revolution around the sun, the Earth's orbital speed around the sun is (a) 4.8×10^3 m/s, (b) 2.3×10^{15} m/s, (c) 3.0×10^4 m/s, (d) 7.3×10^{14} m/s.
11. Agnes is in an elevator. Andy, sitting on the ground, observes Agnes to be traveling upward with a constant speed of 5 m/s. At one instant Agnes drops a pen from rest. Immediately after, the acceleration of the pen according to Agnes is (a) 10 m/s^2, down, (b) 0, (c) 15 m/s^2, down, (d) 5 m/s^2, up.
12. As in the previous question, Agnes is in an elevator that Andy (attached to the ground) sees traveling upward. This time Andy sees the elevator's speed increasing by 5 m/s every second. Agnes stands on a scale in the elevator and sees the reading to be 750 N. After the elevator comes to a complete stop, Agnes is still on the scale. The reading now is (a) 250 N, (b) 500 N, (c) 750 N, (d) 1000 N.
13. As I apply the brakes in my car, books on the passenger seat suddenly fly forward. That is most likely because (a) the car is not an inertial reference frame, (b) the seat supplies a forward push to make the books accelerate, (c) there is a strong gravitational field generated by the brakes, (d) there is a strong magnetic field generated by the brakes.
14. A particle of mass m_1 collides with a particle of mass m_2. All other interactions are negligible. The ratio of the acceleration of mass m_1 to the acceleration of mass m_2 at any instant during the collision (a) is small at first, then reaches a maximum value, then goes back to a small value, (b) depends on whether m_1 and m_2 stick together in the collision, (c) depends on how fast each of the particles is initially moving, (d) is always the constant value m_2/m_1.
15. A 10 kg and a 4 kg mass are acted on by the same magnitude net force (which remains constant) for the same period of time. Both masses are at rest before the force is applied. After this time, the 10 kg mass moves with a speed v_1 and the 4 kg mass moves with a speed v_2.

Which of the following is true? (a) v_1 is equal to v_2, (b) the ratio v_1/v_2 is equal to 5/2, (c) the ratio v_1/v_2 is equal to 2/5, (d) the ratio v_1/v_2 is equal to $(2/5)^2$.

16. Can an object's velocity change direction when its acceleration is constant? (a) No, this is not possible because it is always speeding up. (b) No, this is not possible because it is always speeding up or always slowing down, but it can never turn around. (c) Yes, this is possible, and a rock thrown straight up is an example. (d) Yes, this is possible, and a car that starts from rest, speeds up, slows to a stop, and then backs up is an example.

17. Can an object have increasing speed while its acceleration is decreasing? (a) No, this is impossible because of the way in which acceleration is defined. (b) No, because if acceleration is decreasing the object will be slowing down. (c) Yes, and an example would be an object falling in the absence of air friction. (d) Yes, and an example would be an object released from rest in the presence of air friction.

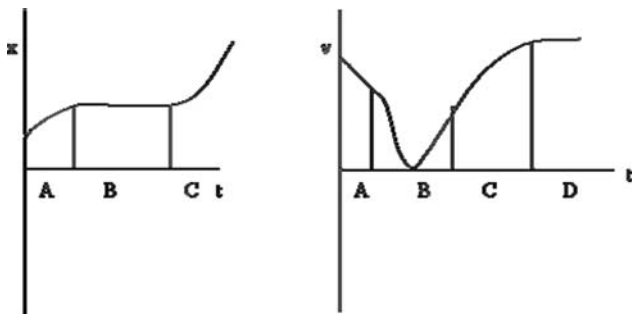
Questions 18–21 concern interpreting the two graphs below.

18. In which interval of the x versus t graph (A, B, or C) is the acceleration negative?

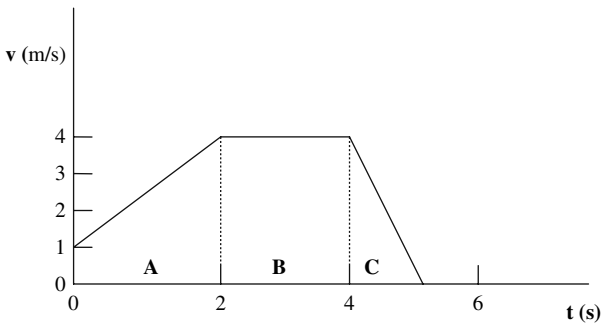
19. In which interval of the x versus t graph (A, B, or C) is the velocity constant?

20. In which interval of the v versus t graph (A, B, C, or D) is the acceleration constant but nonzero?

21. In which interval of the v versus t graph (A, B, C, or D) is the acceleration only positive?



Questions 22 and 23 refer to the following diagram.

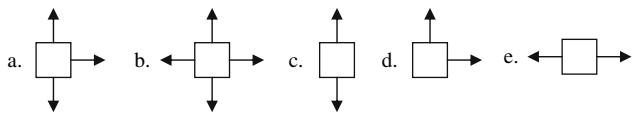


22. If the above graph is for a 4 kg object, the forces acting during each of these three intervals (A, B, C) are given (in Newtons) by (a) (6, 0, 16), (b) (−6, 0, 16), (c) (3/2, 0, −4), (d) (6, 0, −16), (e) (3/2, 0, −16).

23. If the object described by the above graph starts at the origin at $t = 0$, where will it be at $t = 4$ s? (a) $x = 11$ m, (b) $x = 13$ m, (c) $x = 8$ m, (d) $x = 4$ m, (e) $x = 22$ m.

24. A person is holding up a picture by pushing it horizontally against a vertical wall. The reaction force to the weight of the picture is (a) the normal force on the picture, (b) the pull upwards on the Earth equal to the weight, (c) the frictional force on the picture at the wall equal to the weight, (d) the frictional force on the wall by the picture, (e) the normal force on the wall by the picture.

25. Which of the following represents the correct free-body diagram for a helium (floats in air) balloon held by a string that is tied to a seat inside the passenger compartment of a train traveling to the right at a constant 60 mph?



26. A cart is being pulled along a horizontal road at constant velocity by a horse. What is the reaction force to the horse pulling on the cart? (a) the normal force of the ground on the cart, (b) the weight of the cart, (c) the friction force on the cart equal to the pull of the horse, (d) the equal backwards pull on the horse.

27. An object is thrown straight up. At the top of its path the net force acting on it is (a) greater than its weight, (b) greater than zero but less than the weight, (c) instantaneously equal to zero, (d) equal to its weight.

28. A trained seal at the circus sits on a chair and balances a physics book on its nose. On top of the book sits a basketball. Which of the objects exerts a force on the basketball? (a) the book only; (b) both the seal and the book; (c) the seal, the book, and the chair; (d) none of the above.

29. A large truck runs into a small car and pushes it 20 m before stopping. During the collision (a) the truck exerts a larger force on the car than the car exerts on the truck; (b) the truck exerts a smaller force on the car than the car exerts on the truck; (c) the truck and car exert equal forces on each other; (d) the car doesn't actually exert a force on the truck; the truck just keeps going.

30. A car weighing 10,000 N initially traveling at 30 m/s crashes into a 100 N garbage can, initially at rest, sending it flying. During the time the car is in contact with the can it exerts a force of 3000 N on the can. During the time of contact the can exerts (a) a force of 3000 N on the car, (b) a force considerably

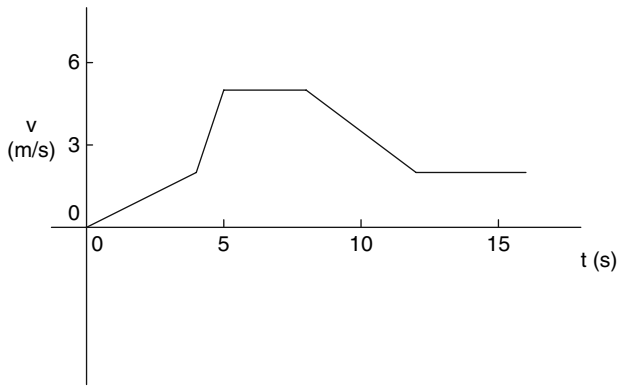
less than 3000 N on the car, (c) a force considerably greater than 3000 N on the car, (d) no force on the car.

31. As a protein diffuses in a thin long tube (effectively 1-dimensional motion) starting from $x = 0$, its average position $<x>$ and its mean square position $<x^2>$ change with time t according to (a) $<x> = <x^2> = 0$, (b) $<x> = 0$; $<x^2> \propto t^2$, (c) $<x> \propto t$; $<x^2> \propto t^2$, (d) $<x> \propto t$; $<x^2> \propto t$, (e) $<x> = 0$; $<x^2> \propto t$.

32. At a turning point in the motion of an object: (a) the velocity can be positive or negative but the acceleration must be instantaneously zero, (b) the velocity must be instantaneously zero, but the acceleration can be positive or negative, (c) both the velocity and acceleration must be instantaneously zero, (d) the velocity and acceleration must have opposite signs (i.e., one positive and the other negative), (e) none of the above is true.

PROBLEMS

1. Shown is a plot of velocity versus time for an object originally at rest at the origin. Develop the corresponding plot for acceleration.



2. (a) Using the data given, plot position versus time for $t = 0$, 4, and 8 s. Calculate the velocity for each interval [0,4] and [4,8] and determine that the average acceleration between these two time intervals is zero.

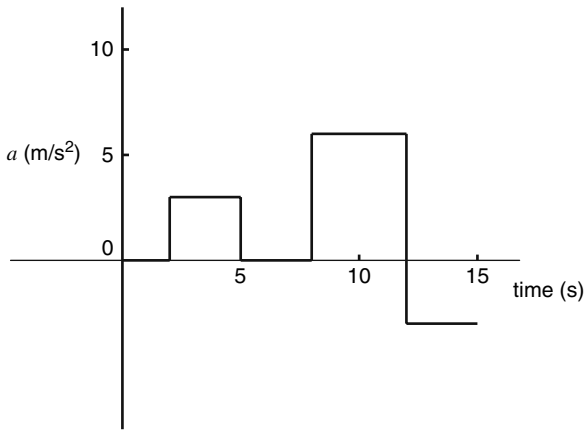
T, seconds	0	1	2	3	4	5	6	7	8
x, meters	1	7.25	9	7.75	5	2.25	1	2.75	9

(b) Now plot all nine data points. Calculate velocity again, this time for all eight time intervals from [0,1] through [7,8]. Calculate the average accelerations for the time intervals [0,2], [2,4], [4,6], [6,8] starting with the velocities just previously calculated.

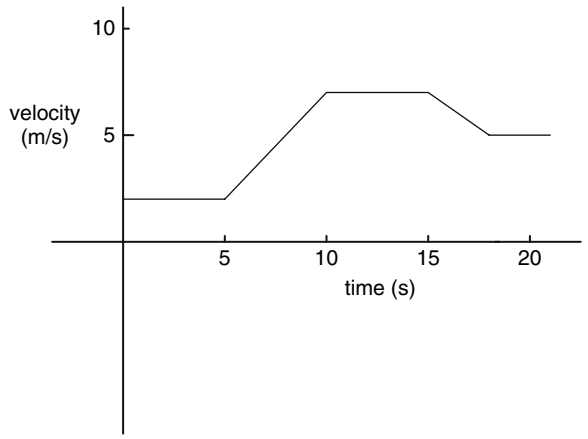
(c) Note that the given data are from the functional expression $x(t) = t^3/4 - 3t^2 + 9t + 1$. Deduce that the data describe the motion of an object that moves forward, stops and backs up, stops again, and moves forward with increasing speed.

(d) Do you see how use of 4 s time intervals misses the details of motion that is more fully described by the use of shorter time intervals? Where is the slope of the $x(t)$ curve positive? Where negative? Where zero? What is the physical meaning of the sign of the slope of the $x(t)$ curve? If the slope of the $x(t)$ curve changes sign, what does that say about the velocity and the acceleration of the object?

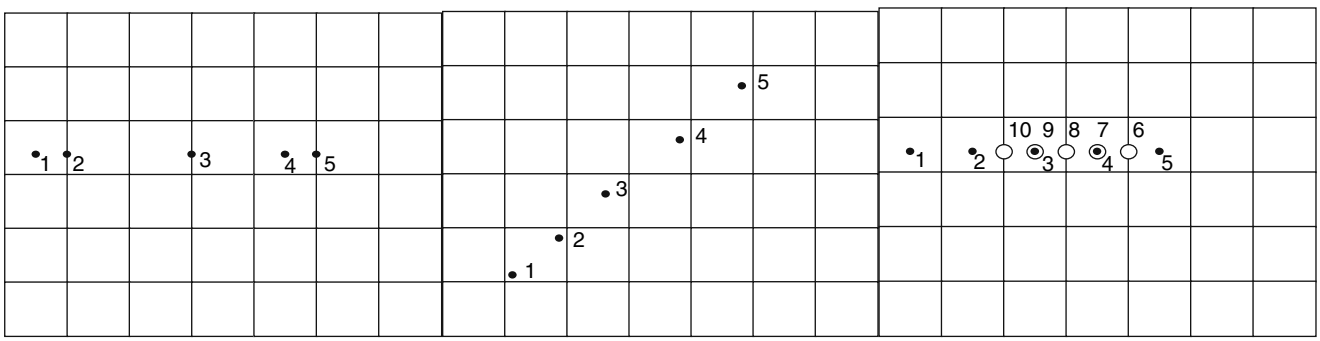
3. Shown is a plot of acceleration versus time for an object. Assuming that its initial position and initial velocity are both zero in magnitude, for how long after $t = 12$ s, must the acceleration of -3 m/s^2 persist, in order that the object be brought to rest?



4. Shown is a plot of velocity versus time for a particle starting at the origin. Sketch a plot of the acceleration corresponding to the time interval for which velocity is shown.



5. A microbiologist observes the motion of a microorganism within a slide sample. Photographic records are snapped at 5 s intervals and the successive positions of the organism are shown. Calculate the average velocities and accelerations corresponding to the appropriate 5 s intervals, assuming the grid line spacing is 25 μm, for each of the three sets of records. Such quantitative investigations of biological motion can reveal important information about the organism. We show later that the measurement of acceleration can indicate how much force certain organs of locomotion are capable of generating. If the organism moves by expelling fluid, we may be able to determine the amount of fluid ejected per unit time and its expulsion velocity.



6. A 1400 kg car accelerates uniformly from rest to 60 mph in 6 s. Find the net force needed to produce this motion.
7. A car accelerates from rest uniformly to 30 mph in 5 s, travels at a constant 30 mph for 0.3 mi, and then decelerates to rest in 6 s.
   (a) What is the average velocity for each interval and for the entire trip?
   (b) What is the displacement for each interval and for the total trip?
   (c) What is the average acceleration for the entire trip?
8. A 0.1 kg mass stretches a linear spring by 10 cm. If three identical masses are hung together from two such identical springs (as in Figure 2.8), by how much will each spring stretch?
9. A Boeing 737 jet plane lands with a speed of 60 m/s (about 135 mi/h) and can decelerate at a maximum rate of 5 m/s^2 as it comes to rest.
   (a) What is the minimum time needed before the plane will come to rest?
   (b) Could this plane land on a runway that is 2800 feet long?
10. A person throws a set of keys upward to his friend in a window 9.2 m above him. The keys are caught 3.0 s later by the friend's outstretched hand.
   (a) With what initial velocity were the keys thrown?
   (b) What was the velocity of the keys just before they were caught?
11. Suppose that a 1 kg block attached to a light rope free-falls (with acceleration g) from rest for 5 s before someone grabs the rope.
   (a) What velocity will the block have when the rope is grabbed?
   (b) In order to stop the block after an additional 5 s, what must be the constant acceleration of the block?
   (c) With what force must the rope be pulled upward to stop the block in those 5 s?
12. What is the acceleration of a 5 kg package being lowered to the ground by a light rope in which there is a tension of 25 N?
13. A truck moves through a school zone at a constant rate of 15 m/s. A police car sees the speeding truck and starts from rest just as the truck passes it. The police car accelerates at 2 m/s^2 until it reaches a maximum velocity of 20 m/s. Where do the police and the truck meet and how long does it take?
14. A person of mass 60 kg stands on top of a table located 1/2 m above the floor and then walks off the edge of the table.
   (a) Draw a free-body diagram of this situation.
   (b) During the time the person is falling to the floor, what is the upwards acceleration of the Earth as seen by the person?
   (c) As seen by the person, through what distance does the Earth move up towards her in this time?
15. The planet Pluto travels once around the sun every 248 years at a mean distance from the sun of 5890×10^6 km. Find its orbital speed around the sun (in m/s).
16. What is the gravitational field on the surface of the moon? Take the mass of the moon as 7.4×10^{22} kg and its radius as 1.74×10^6 m and calculate g as a fraction of that on the Earth's surface.
17. What is the gravitational force of the sun on Pluto with a mass of 1.5×10^{22} kg (less than the moon) and a mean distance from the sun of 5890×10^6 km?

18. Suppose your normal weight is 1200 N standing on a bathroom scale. If you stand on that same scale in an elevator in a skyscraper that is accelerating upwards at 1 m/s^2, what will the scale read?

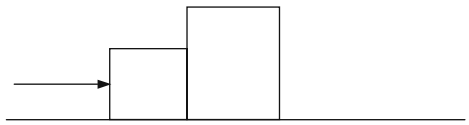
19. An eagle soaring overhead has a weight of 120 N. If the area of each wing is 1.7 m^2, find the force per unit area required to support the eagle while it soars.

20. The electron in a hydrogen atom is attracted to the proton in the nucleus with an electrical force of 8.2×10^{-8} N. What is the acceleration (magnitude and direction) of the electron? (According to classical physics this acceleration keeps the electron orbiting the nucleus.)

21. Two astronauts are out for a space walk near their shuttle. They have masses of 120 kg and 140 kg suited up in their space suits and are attached to the shuttle by umbilical cords. With both initially at rest with respect to the shuttle, if the 140 kg astronaut pushes the other one with a 20 N force for 1 s,
(a) What is the acceleration of the 120 kg astronaut during this 1 s?
(b) What is the acceleration of the 140 kg astronaut during the same 1 s?
(c) What velocity will each have after the 1 s interval with respect to the shuttle?
(d) If the umbilical is 10 m long, how long will it be before they each feel another force from the tug of the umbilical?

22. A heavy 40 kg crate sits on a shelf and is connected by a taut rope to the ceiling. If it is pushed off the shelf so that it is suspended freely find
(a) The net force on the crate.
(b) The tension force in the rope supporting the crate.
(c) If the rope is cut, what is now the net force on the crate?

23. Two heavy crates (of 10 kg and 20 kg mass) sit touching on a smooth surface of ice as shown. If a 20 N force pushes on the 10 kg crate as shown:
(a) What is the acceleration of both blocks?
(b) What is the net force on the 20 kg block?
(c) What force does the 20 kg block exert on the 10 kg block?
(d) What is the origin of the force in part c?
(e) Repeat the problem if the two blocks are physically interchanged (in parts (b) and (c) interchange the two masses as well) and the same force pushes the 20 kg block.



24. A 0.01 g water strider, an insect that can "walk on water," propels itself with its six legs to travel along at 0.5 m/s.
(a) What vertical force must the surface tension of water provide to each foot?
(b) If the insect is able to travel at constant velocity by overcoming a total resistive force from the water of 10^{-6} N, find the horizontal force from the water on each leg as the bug "walks."

25. A single nonmotile cell is confined to a thin capillary tube so that it essentially undergoes one-dimensional diffusion with a diffusion coefficient of 10^{-9} cm^2/s. Find (a) the time it takes for the cell to diffuse a distance of 1 cm (express your answer in hours), and (b) the rms distance the cell will travel in 1 s (expressed in μm). Why don't your answers to (a) and (b) scale linearly so that 3600 s/h multiplied by the answer to (b) would give a 1 cm distance?

26. As cells crawl along a surface in tissue culture their cytoplasm is observed to undergo "retrograde" flow in the direction opposite to the motion of the leading edge of the cell. When this motion is studied by imaging the cell in a microscope and making a movie of the motion, a feature in the cytoplasm is observed to travel a distance of 1.1 μm in 25 s. What is the speed of this retrograde flow?

3

Applications of Newton's Laws of Motion in One Dimension

Newton's laws of motion are a very powerful tool that allows the study of a vast array of problems dealing with the motion of all the objects of our daily lives. Valid over an enormous range of distances, speeds, and masses, Newton's laws only lose their predictive power in the microworld or when objects travel at extremely high speeds, much higher than we are capable of propelling ordinary objects (except in particle accelerators). In this chapter we continue our study of one-dimensional motion in three "case studies" of interesting example applications. The goal here is to see the power of Newton's laws as well as to learn some interesting ideas about various types of motion along a single direction. We gain some valuable insights and tools so that when we generalize to study the motion of objects in the real three-dimensional world we are well prepared for that undertaking. The case studies in this chapter include motion when the net force is constant (we study the local gravitational force near the Earth), one-dimensional motion of an object in a fluid (where we show that there are frictional forces that vary with time), and the oscillatory motion of an object attached to a spring. After learning something about springs, we next consider the deformation of an elastic solid and the phenomenon of viscoelasticity. This is a topic of special interest in the study of structural biomolecules such as bone and blood vessels. We conclude the chapter with a discussion of the structure and dynamics of macromolecules, specifically illustrating how to apply Newton's second law to the difficult problem of determining the molecular motions (here in one dimension) of the constituent atoms of a protein.

1. THE CONSTANT FORCE

Very frequently in dealing with mechanics problems, we know the forces acting on an object and want to predict its future motion, or perhaps even learn of its past motion. For example, the gravitational forces acting on the planets can be calculated extremely accurately from information on their positions relative to the sun, and these forces then, using Newton's second law, predict their accelerations. Knowing the position and velocity of a planet at some time, together with its acceleration, allows scientists to calculate the trajectories of the planets extremely accurately into the distant future. In principle, from knowledge of the acting net force, Newton's second law provides the acceleration of an object as a function of time; from that one can extract information about velocity, and then from that, position. The general case of this kind of problem requires sophisticated mathematical tools (called solutions to "differential equations"). But, there is one special case—in which the net force on an object is a constant, producing a constant acceleration for extended periods—that can be treated easily and whose solution shows us how the more general case works. As we have seen, the gravitational force on a mass m is given by $F = mg$, where g is the constant

J. Newman, *Physics of the Life Sciences*, DOI: 10.1007/978-0-387-77259-2_3,

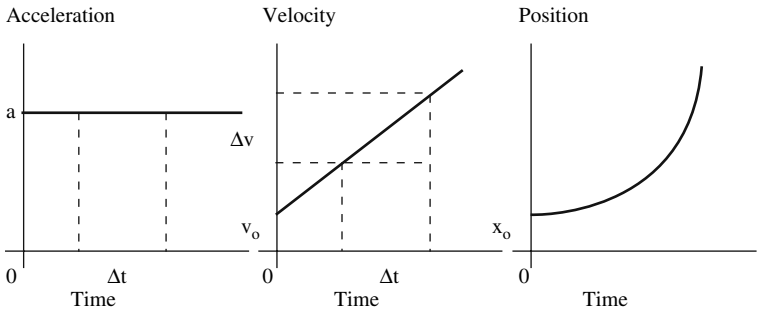


FIGURE 3.1 *Time-dependence of variables for constant acceleration.*

free-fall acceleration due to gravity. The situation is shown in Figure 3.1 (left), where the acceleration is some unchanging value, $g = 9.8\ \text{m/s}^2$.

Because acceleration at any instant is the slope of the tangent line to the velocity versus time graph at that instant, a constant acceleration means that the tangent line to the velocity curve has the same slope all along the curve. The only way that can be true is if the velocity versus time curve is itself a straight line with slope equal to the constant a. Knowing a doesn't tell us everything about the velocity v, however, only that in any time interval Δt the velocity changes by the amount $\Delta v = a\Delta t$. On the other hand, if the value of the velocity is known at a particular moment, then the velocity is determined at every instant for which a remains the acceleration. In Figure 3.1(center), the velocity is specified as being v_0 at $t = 0$, for example. When that is the case, velocity depends on time in the following explicit way.

$$\Delta v = v(t) - v_0 = a\Delta t = a\cdot(t - 0) = at,$$

so that

$$v(t) = v_0 + at. \tag{3.1}$$

The latter relation says that once a and v_0 are specified, just plug into Equation (3.1) a value of time and the velocity at that time is automatically determined.

There's another way of understanding how to go from acceleration to velocity. In the last chapter we said that there is a graphical interpretation of acceleration: at any instant, it is the slope of the tangent line to the velocity versus time graph. There is another graphical interpretation when we go the other way, from acceleration to velocity. Note that by drawing vertical lines from the times at the ends of the time interval Δt, we construct a rectangle on the acceleration versus time graph (Figure 3.1-left), the base of which is Δt and the height of which is a. Because $\Delta v = a\Delta t$, we can interpret Δv as the area under the acceleration versus time graph in the associated interval Δt. Now, in general, even when acceleration is not constant, we know that $\Delta v = \bar{a}\Delta t$. So, extrapolating from the constant acceleration case, we assign the average acceleration a graphical interpretation:

$$\bar{a} = \frac{(\text{the area under the acceleration versus time graph in the interval } \Delta t)}{\Delta t}.$$

We make use of this idea in just a moment.

Having determined velocity as a function of time, we can determine position, x, at any time for the special case of constant acceleration as well. First, note that velocity at a given instant is the slope of the tangent line to the position versus time graph at that instant. Velocity is constantly changing, therefore the slope of the x versus t tangent line is also constantly changing. Thus, position versus time has a curved graph (as in Figure 3.1-right). What is its shape? Well, first we know that $\Delta x = \bar{v}\Delta t$. Arguing by analogy with the acceleration–velocity situation, we state that the average velocity in an interval Δt is the area under the velocity versus time graph divided by Δt. Let's say that t_i is the first instant of Δt and t_f is the last, and that $v(t_i) = v_i$ and $v(t_f) = v_f$. The shape under the velocity versus time graph defined by vertical lines drawn from the ends of Δt is a trapezoid, in particular, a right triangle sitting on top of a rectangle. See Figure 3.2.

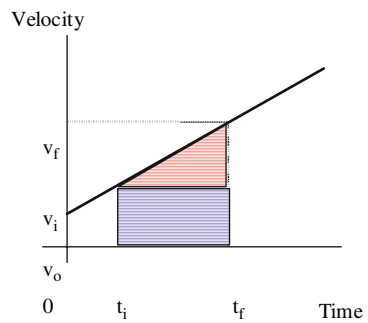


FIGURE 3.2 *Finding the average velocity during an interval of time.*

The area of the rectangle is $v_i\Delta t$ and the area of the triangle is $(v_f - v_i)\Delta t/2$ (one half base times height), so adding the two together gives the area under the graph as $(v_i + v_f)\Delta t/2$. (Remember, this is only true for the special case of constant acceleration.) As a result, we conclude that when acceleration is constant

$$\bar{v} = \frac{1}{2}(v_i + v_f).$$

Now, suppose that $t_i = 0$ and $t_f = t$ (a general value of time after $t = 0$); then

$$\bar{v} = \frac{v_o + v(t)}{2}.$$

Combining this with the definition of average velocity as

$$\bar{v} = \frac{\Delta x}{\Delta t} = \frac{x(t) - x(0)}{t}$$

and writing $x(0)$ as x_o, we find

$$x(t) = x_o + \frac{v_o + v(t)}{2}t.$$

After substituting from Equation (3.1) for $v(t)$, we have

$$x(t) = x_o + v_o t + \frac{1}{2}at^2. \tag{3.2}$$

(The curve of x versus t is consequently a parabola when acceleration is constant.) Equations (3.1) and (3.2) represent useful relations for the velocity and position as functions of time, respectively, of an object undergoing motion with a constant acceleration. With a bit of algebra one can solve for t in Equation (3.1) and substitute into Equation (3.2) in order to eliminate time and have a third (although not independent) relation between the other variables,

$$v^2 = v_o^2 + 2a(x - x_o), \tag{3.3}$$

where x and v are evaluated at the same time. Table 3.1 summarizes these three relations, which have been derived exclusively from definitions in the special case of constant acceleration. We show later that constant acceleration arises from a situation in which the object experiences a constant force, and although this is often not true, it represents the simplest case and can sometimes also be a useful approximation to the motion. But, note that these three relations are true in the general case of nonconstant acceleration, as long as we replace a by $\bar{a}$ wherever it appears. Of course, in the general case we have to be able to calculate $\bar{a}$ (the area under the acceleration versus time graph from 0 to t, divided by t) to make these relations useful.

It is very straightforward and elegant to derive Equations (3.1) and (3.2) directly from the definitions of acceleration and velocity using calculus. The definitions of velocity and acceleration, rewritten using derivative notation, are

$$v = \mathrm{d}x/\mathrm{d}t$$

and

$$a = \mathrm{d}v/\mathrm{d}t.$$

Starting from the definition of a, after multiplying by dt and integrating both sides of the equation, we can write

$$\int_{v_o}^{v} \mathrm{d}v' = \int_0^t a\mathrm{d}t'.$$

Integrating leads to Equation (3.1) because the acceleration is assumed constant and can be factored out from the integral. (If, in fact, the acceleration is not constant but is a known function of time then this integral expression can be solved for more complex cases of nonconstant acceleration.) Then, inserting the definition of v into Equation (3.1), multiplying again by dt and integrating, we have

$$\int_{x_0}^{x} \mathrm{d}x' = \int_0^t v_0 \mathrm{d}t' + \int_0^t at'\mathrm{d}t'$$

that integrates to give Equation (3.2), because both v_o and a are assumed constant. By the same algebraic elimination of t as in the text we arrive at the third relation, Equation (3.3).

Table 3.1 *One-Dimensional Kinematic Relations for Constant Acceleration Motion*

	Equation	*Variables*
1.	$v(t) = v_0 + at$	v, a, t
2.	$x(t) = x_0 + v_0 t + \frac{1}{2}at^2$	x, a, t
3.	$v^2 = v_0^2 + 2\,a\,(x - x_0)$	v, a, x

Example 3.1 An *E. coli* bacterium travels a total distance of 100 μm along a straight line from one position of rest to another. For a brief time during this trip it accelerates from rest at a constant acceleration to a speed of 20 μm/s and for another brief time near the end, it decelerates (with the same magnitude of acceleration, but oppositely directed) coming to rest after the total distance traveled. If the total time for the trip is 5.4 s, find the time during which the bacterium accelerates, the time during which it decelerates, its acceleration, and the fraction of the distance traveled at constant velocity.

Solution: Here is an example of a problem in which acceleration is not constant throughout. On the other hand, the total trip can be divided up into three different phases, where in each acceleration is constant: (1) an acceleration from rest ($v_0 = 0$, acceleration $= +a$), (2) a constant velocity portion (velocity = a constant, acceleration = 0), and (3) a deceleration to rest (velocity = the same constant as in the previous phase, acceleration $= -a$). We can separately write expressions for the distances traveled in each portion and add them up to total the 100 μm distance. Writing the distances and respective times as d_1, t_1, d_2, t_2, and d_3, t_3, we have (using Equation (3.2) of Table 3.1)

$$d_1 = \frac{1}{2}at_1^2,$$
$$d_2 = vt_2,$$
$$d_3 = vt_3 + \frac{1}{2}(-a)t_3^2 = vt_3 - \frac{1}{2}at_3^2,$$

where v is the constant velocity of the middle portion of the trip and a is the magnitude of the constant acceleration and deceleration. Before adding these, we note that the times t_1 and t_3 are equal because we can write, according to Equation (3.1) in Table 3.1, expressions for the velocity in the first and third intervals

$$v = 0 + at_1 \text{ and } 0 = v + (-a)t_3.$$

Then, using the fact that $t_3 = t_1$ and $t_{tot} = t_2 + 2t_1$, we have

$$d_{tot} = \frac{1}{2}at_1^2 + vt_2 + vt_1 - \frac{1}{2}at_1^2 = vt_2 + vt_1 = v(t_{tot} - t_1).$$

Because d_{tot}, v, and t_{tot} are given, we can solve this for t_1 to find

$$t_1 = t_3 = t_{tot} - \frac{d_{tot}}{v} = 5.4 - \frac{100}{20} = 0.4 \text{ s}.$$

To then find the acceleration, we can use the velocity expressions to write $v = at_1$, for example, and find that

$$a = \frac{v}{t_1} = \frac{20}{0.4} = 50 \text{ μm/s}^2.$$

Finally, because the time traveled at constant velocity is $t_2 = 5.4 - 2(0.4) = 4.6$ s, the distance traveled at constant velocity is $d_2 = vt_2 = 20\ (4.6) = 92$ μm, representing 92% of the distance traveled in the interval.

3.1.1. FREE-FALL: AN EXAMPLE OF CONSTANT ACCELERATION

A common situation in which there is a constant acceleration is in "free-fall." An object released near the Earth's surface falls under the influence of gravity at a constant acceleration equal to $a = 9.8\ \text{m/s}^2$, as long as air resistance is negligible. We represent the magnitude of this free-fall acceleration by the symbol g. Actually, the acceleration any body experiences due to gravity decreases with increasing height from the Earth's surface, but since it decreases only by about 1.5% at an altitude of 50 km (about 30 miles) we can almost always treat it as a constant. We can analyze one-dimensional free-fall situations without any new mathematical developments, because we already have all the necessary relations among position, velocity, acceleration, and time for one-dimensional motion in Table 3.1. It is usual in free-fall problems to take a coordinate system in which x is horizontal and y is vertical (with "up" being the positive direction). In the next two examples we treat vertical motion only. For these examples, translate the quantities in Table 3.1 by replacing x by y, and a by $-g$.

Example 3.2 A tennis ball is thrown upwards with an initial speed of 12 m/s. Find how high it will rise and how long it will take to return to its starting height.

Solution: The tennis ball rises until its velocity is momentarily zero. As it rises it is uniformly slowed by the downward pull of gravity that acts continuously. Even at the moment its velocity has become zero, the ball still has the same constant downward acceleration. After coming to momentary rest, the ball continues to accelerate downward, its speed continuously increasing.

Knowing that the highest point is characterized by a zero velocity for an instant, we can find the maximum height the tennis ball reaches directly by using Equation (3.3) in Table 3.1. That's because we know initial and final velocities and the acceleration; only the displacement is unknown. We don't, for this part of the problem, have to deal with time. We write Equation (3.3) in the form

$$v_{\text{top}}^2 = v_0^2 + 2(-g)(y_{\text{top}} - y_o) = v_0^2 - 2gH,$$

where v_0 is the given initial velocity and H is the maximum height. We find

$$0 = (12\ \text{m/s})^2 - 2 \cdot (9.8\ \text{m/s}^2)H, \quad \text{or} \quad H = 7.3\ \text{m}.$$

The second part of the problem requires time information. One way of finding it is to write the equation for the displacement of the ball and set it equal to zero

$$y - y_o = v_0 t_{\text{round trip}} + \frac{1}{2}(-g)t_{\text{round trip}}^2 = 0.$$

In using this and any of the kinematic equations, we must be careful about signs: the upward initial velocity is positive and downward acceleration is negative. To solve for the desired quantity $t_{\text{round trip}}$ requires solving a quadratic equation, although in this case a simple one. Whenever one solves a quadratic relation there are always two solutions. Which is the appropriate one for the problem at hand requires some additional physical reasoning. For this example, we find that either $t_{\text{round trip}} = 0$ (one time at which the ball is indeed at $y = y_o$), or

$$t_{\text{round trip}} = \frac{2v_0}{g} = \frac{2 \cdot 12\ \text{m/s}}{9.8\ \text{m/s}^2} = 2.4\ \text{s}.$$

Of course, the 0 s solution is physically trivial, and not the one of interest here.

(*Continued*)

An alternative solution, incidentally, involves finding the separate times for the ball to go up and down. The first time can be found from $v = v_0 - gt$, with $v = 0$ at the top. We have

$$t_{\text{up}} = \frac{v_0}{g},$$

which we note is half of our previous answer. This result demonstrates that the time for the ball to go up is equal to the time for it to return down, a result that we might have assumed true from the symmetry of the motion.

Example 3.3 A ball is dropped from a height of 20 m to the ground below. A second ball is thrown downward with a speed of 10 m/s after waiting some time Δt after releasing the first ball from rest. To have both balls hit the ground at the same time, how long should time Δt be?

Solution: We take the same coordinate conventions as in the previous example: up is positive, down is negative. Also, we take $y = 0$ to be on the ground. (We could have set $y = 0$ anywhere, such as at the launch point of both balls, for example. It doesn't matter where you choose your coordinate origin, but once having done so, you have to remember to systematically keep it there in all your calculations.) The first ball is dropped from rest (so its initial velocity is zero) and travels downward with an acceleration $-g$. After time t_1 it is at position y_1 given by

$$y_1 = (+\,20\,\text{m}) + \frac{1}{2}(-g)t_1^2.$$

The $+20$ m in this equation represents y_1 at $t_1 = 0$. With $y_1 = 0$ m (i.e., just about to hit the ground), we find that

$$t_1 = \sqrt{\frac{2 \cdot 20\,\text{m}}{9.8\,\text{m/s}^2}} = 2.0\ \text{s}.$$

Again, because we are solving a quadratic equation to find t_1, there are two times when the ball could be at $y = 0$ and be traveling under free-fall conditions: in this case, ± 2.0 s. The negative solution corresponds to a time 2.0 s before when it is at $+20$ m with a velocity of zero. That is, if we launched the ball from $y = 0$ with the correct upward velocity, in 2.0 s it would be at $+20$ m up and just ready to fall back down. Because we want the time elapsed after the ball is already at $+20$ m, we choose the positive solution.

For the first Δt seconds of the first ball's flight, the second ball is at $y_2 = +20$ m. Then, abruptly, it is thrown downward and, once in free-fall, falls according to the equation

$$y_2 = (+20\,\text{m}) + (-10\,\text{m/s})t_2 + \frac{1}{2}(-g)t_2^2,$$

where -10 m/s is the downward initial velocity of the ball. In this equation, t_2 is the time for the second ball to fall from $+20$ m to a position y_2; t_2 is zero when t_1 is Δt, and, in general, $t_2 = t_1 - \Delta t$. We want the second ball to be at $y = 0$ at the same instant the first ball is there. Substituting for y_2 $(= 0)$ and g, we obtain the quadratic equation

$$0 = 20 - 10t_2 - 4.9t_2^2.$$

Writing this in the standard form, $ax^2 + bx + c = 0$, that is,

$$4.9t_2^2 + 10t_2 - 20 = 0,$$

we find two possible solutions (see the appendix on solving quadratic equations):

$$t_2 = \frac{-10 \pm \sqrt{(10)^2 - 4 \cdot 4.9 \cdot (-20)}}{2 \cdot 4.9} = 1.2 \text{ s or } -3.3 \text{ s}.$$

(Trace through the units of the numbers in the latter expression—i.e., what are the units of the "10," the "20," and the "4.9"—to assure yourself that the final answer really does have units of s.) As in the first ball case, there are two possible times for the second ball to be at $y = 0$: one is positive, the time elapsed from its release; the other is negative, a time before the moment declared to be $t_2 = 0$ in this problem. Clearly, we want the positive solution. The time that the second ball takes to reach the ground is 1.2 s. Therefore the person needs to wait a time $\Delta t = (2.0 \text{ s} - 1.2 \text{ s}) = 0.8 \text{ s}$ after dropping the first ball before throwing the second one.

2. MOTION IN A VISCOUS FLUID

Up until now we have assumed that all motion has occurred in the absence of any frictional forces to slow objects down. In this section we relax that assumption to include frictional forces in the important case of motion in a fluid, being either a liquid or a gas. We show that in some cases our assumption has been realistic, whereas in other cases it has been a poor one. The nature of the frictional drag forces on macroscopic and microscopic objects leads to very different kinds of motion, considered below in separate discussions.

2.1. FORCES ON A MACROSCOPIC OBJECT IN A FLUID

Macroscopic objects immersed in a fluid (liquid or gas) experience two forces in addition to their weight (microscopic objects are discussed later in this section). There is a *buoyant force* that always acts vertically upward and a *drag* (or *frictional*) *force* directed opposite to the velocity of the object (Figure 3.3). If the object is sinking in the fluid then the frictional force also points upward, but if the object is rising, the frictional force will then be downward. For now, we treat the buoyant force $\boldsymbol{F}_\mathbf{B}$ as a small constant correction, returning to a more detailed consideration in a discussion of fluids in Chapter 8. In the rest of this section we investigate the drag force and its effect on motion.

An object moving in a fluid is surrounded by a thin layer of fluid, known as a boundary layer, that moves along with it. If immersed in the fluid, as the object moves, it must push fluid away and around itself to move forward and this motion causes the fluid in the immediate vicinity of the object to flow. We can distinguish two limiting types of flow based on the fluid properties, namely the fluid density ρ and *viscosity* (or "stickiness") η, as well as the size L and speed v of the object. The quantity that determines the flow behavior of an object in a fluid is the Reynolds number $\Re$, given by the dimensionless ratio

$$\Re = \frac{L\rho v}{\eta}. \tag{3.4}$$

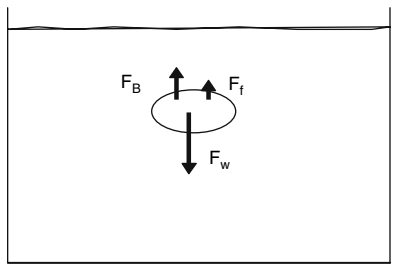


FIGURE 3.3 *Forces on a macroscopic object submerged and falling in a fluid.*

Representative values for $\Re$ are given in Table 3.2. Note that the fluid properties in the Reynolds number are both intrinsic properties of the fluid. Two volumes of a given fluid will always have the same density (introduced briefly in Section 5 of Chapter 1) and viscosity, regardless of their size or shape, as long as they are at identical environmental conditions (such as temperature and pressure). Density is discussed further in Section 8.1. In general, fluids that pour very slowly such as molasses, maple syrup, and the like are very viscous (have large viscosities) whereas fluids such as water or alcohols, or especially gases such as air, have low viscosities. We also study fluid viscosity further in Chapter 9.

Table 3.2 *Typical Reynolds Numbers for Some Moving Objects in a Fluid*

Situation	*Reynolds Number*
Person swimming	1,000,000
Large flying bird	100,000
Flying mosquito	100
Swimming bacteria	0.0001

For $\Re$ values much larger than 1, the fluid flow near the object is *turbulent* (chaotic, swirling flow), as seen, for example, in fast flowing water near a waterfall (Figure 3.4). Such fluid flow around an object leads to a "wake" (much like that produced by a motor boat speeding across a lake) and results in frictional forces reapplied to the object by the fluid that tend to slow the object. In this case the magnitude of such a frictional, or drag, force is often proportional to the square of the object's speed and can be written as

$$F_f = \frac{1}{2} C\rho A v^2, \quad (\Re \gg 1) \tag{3.5}$$

where C is a drag coefficient with a value typically near 1.0 (but which may vary with velocity, something that we ignore), ρ is the fluid density, and A is the effective cross-sectional area perpendicular to the velocity v.

FIGURE 3.4 *Fast flowing turbulent water in the Andes.*

FIGURE 3.5 *A crouching skier minimizes drag forces.*

If the object is not spherical, then different orientations can present different effective areas A leading to different frictional forces; for example, a thin rod oriented with its axis along the flow velocity presents the minimum effective area, and so the minimum drag force occurs leading to the rod's most rapid flow. This can be demonstrated by dropping two similar large flat rocks into a lake or pool of water. If the speeds of the two rocks are compared when dropping one held vertically on edge and the other held flattened side horizontally, the rock dropped on edge will fall at a much faster rate, due to the decreased drag. The notion of an effective area is used by skiers, bikers, and skydivers, for example, to minimize frictional drag. In each case, the person can reduce the drag force of the air that is slowing them down by huddling over and wearing tight-fitting clothing so as to minimize their effective cross-sectional area (Figure 3.5).

The drag force is also proportional to the fluid density. Comparing water and air, the two most common fluids in biology, the drag force for the same object at a given velocity is over 800 times more in water than in air. Streamlined shapes of fish and aquatic animals developed in order to reduce drag forces involved in swimming to minimize the expenditure of energy required for locomotion. Similarly, the aerodynamic design of birds and other flying animals reduces drag in air. Frictional forces in air are only apparent at high speeds because of the relatively low density compared to liquids.

The other limiting type of fluid flow, when $\Re$ is much smaller than 1, is called *laminar* flow and is an orderly smooth flow around an object, such as seen in the streamlined nonturbulent flow of water over a rock in a stream or the smooth flow around a kayak (Figure 3.6). In this case the magnitude of the drag is linearly proportional to the relative speed of the object and fluid v; thus

$$F_f = fv \quad (\Re \ll 1) \tag{3.6}$$

FIGURE 3.6 *Two kayakers. Which water flow is near-laminar, and which turbulent?*

FIGURE 3.7 *Skydivers reach a terminal velocity in free-fall.*

where f is a coefficient of friction that depends on the size and shape of the object and the viscosity of the fluid. If the object is spherical, then the coefficient of friction is given by Stokes' law as

$$f = 6\pi\eta r, \tag{3.7}$$

where r is the radius of the object and η is the fluid viscosity.

Looking back to Equation (3.5), we see that when $\Re \gg 1$ the density of the fluid is important, but the viscosity of the fluid does not enter. In this regime, objects are able to coast along at relatively large velocities after having accelerated to the point where the driving force is balanced by the drag force. In strong contrast, in the regime of $\Re \ll 1$, viscous forces dominate and objects move very slowly and are not able to "drift" for appreciable distances at all; as soon as an external or propulsion force stops, motion ceases abruptly. To give some idea of when these limits occur, for spherical objects with a density close to that of water, like most biological objects, the radius must be smaller than about 40 μm in air, or about 150 μm in water, for the frictional drag to be described by Equation (3.7) when the object is falling under its own weight.

Let's return to our macroscopic object in a fluid that we started to consider in this section under the influence of gravity, buoyancy, and frictional forces. Adding these three forces acting on the object and writing Newton's second law, we find that

$$mg - F_B - F_f = ma. \tag{3.8}$$

If we imagine releasing the immersed object from rest, it will initially fall (assuming the buoyant force is less than the weight; otherwise it will rise just as a bubble rises to the surface). However, because the frictional force grows as the velocity increases, eventually the net force on the object will become zero and the particle will have zero acceleration. In this case we can set $a = 0$ in Equation (3.8) and solve for the constant velocity at which the object will continue to fall, known as the *terminal velocity*, v_{term} (see Figure 3.7). Depending upon the value of the Reynolds number, and thus whether the flow is turbulent or laminar, we will find two different relations for the terminal velocity.

Usually for free-fall objects in air (but not for microscopic objects, which do not fall rapidly, nor for highly streamlined objects), the flow will be turbulent, the Reynolds number large, and the terminal velocity given by substituting Equation (3.5) into Equation (3.8) with $a = 0$ to find

$$v_{\text{term}} = \sqrt{\frac{2(mg - F_B)}{C\rho A}}. \tag{3.9}$$

On the other hand, for objects that are streamlined so turbulence is minimized or when the Reynolds number is small, the flow is laminar, the frictional force is linear in the velocity, and the terminal velocity will be given by substituting Equation (3.6) into Equation (3.8) and setting $a = 0$ to find

$$v_{\text{term}} = \frac{mg - F_B}{f}. \tag{3.10}$$

How does an object approach its terminal velocity in the case of a linear frictional force?

From Equation (3.8) with F_f given by Equation (3.6), we have

$$mg - F_B - fv = ma = m\,\mathrm{d}v/\mathrm{d}t.$$

The solution to this equation is given by

$$\begin{aligned} v(t) &= v_{\text{term}}(1 - \mathrm{e}^{-(f/m)t}) \\ &= v_{\text{term}}(1 - \mathrm{e}^{-t/\tau}), \end{aligned}$$

with v_{term} given by Equation (3.10) and $\tau = m/f$. This can be checked by direct substitution. (Try it!) The result shows that the terminal velocity is approached exponentially, so that when $t = \tau$, $v(t = \tau) = v_{\text{term}}(1 - \mathrm{e}^{-1}) = v_{\text{term}}(0.63)$, and when $t = 2\tau$, $v(t = 2\tau) = v_{\text{term}}(1 - \mathrm{e}^{-2}) = v_{\text{term}}(0.86)$, and so on.

The time τ is called the time constant and is the time to reach 63% of the terminal velocity from rest. In the case of our *E. coli* bacterium the time constant is a very short time of about 1 μs.

Examples of terminal velocities of various objects are given in Table 3.3.

Table 3.3 *Terminal Velocities of Various Objects Falling Under Gravity in Different Fluids*

Situation	*Terminal Velocity (m/s)*
Sky diver	100 (225 mph)
Person sinking in water	1
Pollen (~0.04 mm diameter) in air	0.05
Algae spores (same diameter as pollen) in water	0.00005

2.2. FORCES ON A MICROSCOPIC OBJECT IN A FLUID

Microscopic objects behave quite differently in fluids. Constant collisions of fluid molecules with a microscopic object buffet the object about in a random path, overcoming its weight so that it does not settle under gravity but remains suspended in the fluid. This type of random diffusive motion is known as *Brownian motion*. In this case the particle weight and buoyant forces are unimportant and the motion is entirely governed by collision forces that result in diffusion, discussed in Chapter 2.

A particularly interesting case of microscopic objects' motion in a fluid is the motility of organisms: self-propulsion using some mechanism to generate *thrust*. Thrust is a propulsive force that can be produced by somehow pushing back on the surrounding fluid to generate, by Newton's third law, a forward-directed force. It can be generated by squirting fluid backwards as done by clams and jellyfish, for example, or by pushing backwards on the fluid using tentacles, fins, or arms and legs in the case of our swimming (we consider this further in Chapter 6).

Consider the motion of a swimming microorganism such as an *E. coli* bacterium. When the bacterial flagella that are used to generate thrust stop rotating, the viscous forces are so great that motion ceases nearly instantaneously (within a millionth of a second). The bacterium swims by using a set of coordinated rotating flagella to propel itself at speeds of tens of micrometers per second. As long as the flagella rotate in a co-ordinated manner, the dominant forces are simply thrust and frictional forces that balance rapidly to result in constant velocity motion. The typical bacterial motion consists of linear propulsion at a terminal velocity for some distance followed by periods of "twiddling," or uncoordinated rotation of flagella when the rotary motors powering the flagella reverse for short times. In a uniform environment, the bacterium takes off in a random direction again when its flagella come together to produce a coordinated thrust. Investigators have shown that bacteria can sense variations in chemicals (nutrients, oxygen, poisons) and that this results in longer straight line swimming toward or away from chemicals in a process known as *chemotaxis*. The origin of the chemical detection scheme used by bacteria remains unclear.

3. HOOKE'S LAW AND OSCILLATIONS

In this section we study the properties of springs and the motion of a mass attached to a spring. This may seem to be a very specific application of the physics we have learned and you may wonder why it is worthy of an entirely separate section. Linear springs, those exerting a force linearly proportional to the extent of their stretch from equilibrium, can be used to model the interactions between atoms and molecules fairly well near their equilibrium positions. In other words, under some circumstances we can picture the atoms in molecules as being held together by springs rather than by complex electromagnetic forces. The properties of springs and the motion they produce is therefore of importance not only in problems dealing directly with springs, but also in the

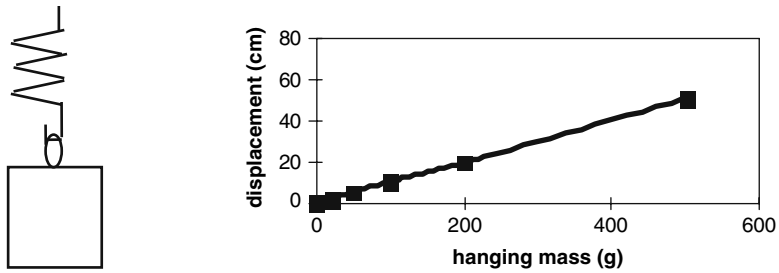


FIGURE 3.8 *Hanging mass on spring and plotted results.*

much larger context of all types of linear forces. This notion is used repeatedly in this book in modeling many different phenomena.

Let's do an experiment with a spring. We support the spring from above and stretch it by hanging different masses on the bottom end, as shown in Figure 3.8. Recording the position of a mark on the bottom of the spring for each hanging mass, we obtain the data shown in the first two columns of Table 3.4. The third column is then obtained by calculating the differences in position of the spring mark with and without the hanging weight to obtain the displacement from the starting position with no hanging weight.

Table 3.4 *Data for Hanging Mass on a Spring*

Hanging Mass (g)	*Position (cm)*	*Displacement (cm)*
0	22.5	0
20	24.6	2.1
50	27.9	5.4
100	33.2	10.7
200	42.4	19.9
500	73.6	51.1

The displacement versus hanging mass data are plotted in Figure 3.8 and are seen to be linear. In each case, the hanging weight is in equilibrium, supported by an equal upward force due to the spring. From our data we could conclude that (at least over a limited range of stretch of the spring) the force that the spring exerts on the hanging mass is proportional to its displacement x from its unstretched equilibrium length, or

$$F = -kx, \tag{3.11}$$

where the constant of proportionality k is called the spring constant. The negative sign indicates that the spring force is a restoring force; if the spring is stretched, the spring force tends to pull it back to a shorter length, whereas if compressed to a shorter length, the spring force tends to restore it to its longer equilibrium length. Equation (3.11) is known as *Hooke's law* and correctly describes the spring force for small displacements. Using it we can determine the spring constant of our spring from a calculation of the slope of the line in Figure 3.8. We first find directly from the graph that $\Delta x/\Delta m = 0.1$ cm/g $= 1.0$ m/kg. From this we can then calculate that $k = \Delta F/\Delta x = \Delta mg/\Delta x = g/(\Delta x/\Delta m) = 10$ N/m, using a value of $g = 10$ m/s^2. With any constant external force F_{ext}, continuously applied to maintain a stretched (as in our hanging weight experiment) or compressed length for the spring, the spring responds with an equal but opposite force according to Equation (3.11) (Figure 3.9). In general, gravity need not play any role as the following discussion shows.

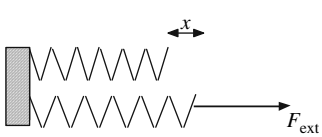


FIGURE 3.9 *An external force stretches a linear spring by distance* x *while the spring pulls back in the opposite direction with a force of equal magnitude,* $F = kx = F_{ext}$.

Let's now take the same spring from our static equilibrium experiment with a hanging weight and clamp it in a horizontal orientation between a fixed wall and a mass $m = 1$ kg lying on a horizontal frictionless surface (an air track, e.g.) as shown

in Figure 3.10. If we pull the mass to the right, stretching the spring 10 cm, and then release it from rest, we can record its position as it moves under the influence of the spring force (note that you don't need tremendous skill to record these data; there are automatic recording schemes that can do these rapid measurements for you). Figure 3.11 shows a plot of these data. By inspection the position versus time graph looks like a cosine function, with repeated oscillatory motion of the mass on the spring. Let's now investigate this situation further to try to explain the observed motion.

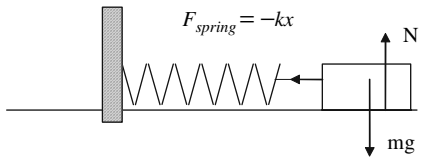


FIGURE 3.10 *A block attached to a spring sliding on a frictionless surface.*

If we consider the forces acting on our oscillating mass we first see that the vertical forces are the weight and the normal force, equal and opposite resulting in no net vertical force and therefore no vertical acceleration or velocity; the mass stays in contact with the surface. This discussion anticipates our generalization to two-dimensional situations later in Chapter 5, but it's clear that the motion here is only horizontal. The only horizontal force is that due to the spring and so, according to Newton's second law, we set that force equal to the product of the mass and the horizontal acceleration a. Unlike our previous Newton's law problems, the applied net force is now a function of position

$$F_{\text{net}} = -kx = ma, \tag{3.12}$$

so that the acceleration of the mass is proportional to its distance from the equilibrium, unstretched, position of the spring, taken as $x = 0$. The acceleration is not a constant, but varies with the displacement from equilibrium!

When we release the mass from rest at $x = A = 10$ cm, the initial acceleration of the mass is given, from Equation (3.12), by $a = -(k/m)A = -(10\ \text{N/m}/1\ \text{kg})(0.1\ \text{m}) = -1.0\ \text{m/s}^2$, where the negative sign indicates that the acceleration is in a direction opposite to the displacement and hence will tend to restore the mass to $x = 0$. The mass gains an increasing velocity back toward $x = 0$, all the while decreasing its acceleration as it approaches $x = 0$. At $x = 0$ the mass momentarily has no acceleration, but it has gained a velocity along the negative x-axis and so continues past $x = 0$. Once x is negative, the spring has been compressed and responds with a force directed back toward the origin, along the positive x-axis. This net force results in an acceleration also directed along the positive x-axis (in agreement with Equation (3.12) with negative x values so that $a > 0$), that acts to decrease the speed of the mass. Because the motion is symmetric about the origin, the velocity of the mass as it passes the origin turns out to be just enough to have the mass, as it slows down, reach the position $x = -10$ cm. At this point the mass momentarily stops, but is acted on by a maximal positive acceleration equal to $+(kA/m) = 1.0\ \text{m/s}^2$. The next phase of the motion, from $x = -10$ cm to $x = +10$ cm is the mirror image of the above description. The motion continues, with the mass oscillating back and forth between the limits of $x = \pm A = \pm 10$ cm, where A is known as the *amplitude*, or maximum distance from the origin.

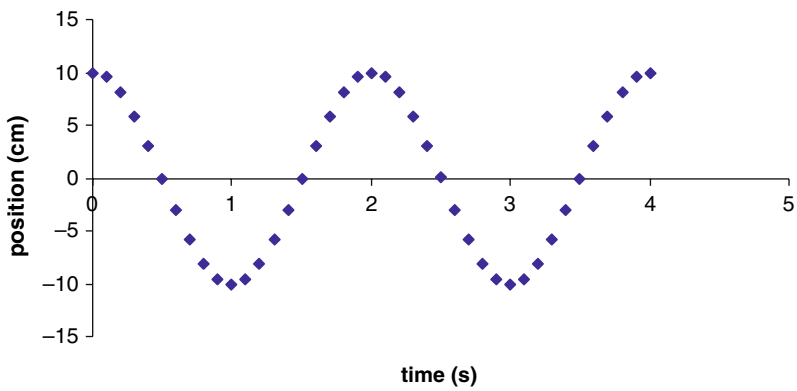


FIGURE 3.11 *Data for the position of the mass attached to a horizontal spring versus time.*

Example 3.4 A horizontal linear spring with a spring constant of 10 N/m is stretched a distance of 10 cm and a 2 kg block resting on a frictionless surface is attached. When the block is released, find its acceleration. What is its acceleration after moving 10 cm? After moving 20 cm?

Solution: The horizontal force on the block when released is entirely due to the spring and is given by Hooke's law as $F = (10\ \text{N/m})(0.1\ \text{m}) = 1\ \text{N}$, so that the initial acceleration is then $a = F/m = 0.5\ \text{m/s}^2$. After traveling 10 cm, the block is at the equilibrium position of the spring and will momentarily feel no force because the spring is unstretched. Therefore the acceleration is also zero at that instant, even though the block is sliding with some velocity. In fact the block has reached its maximum velocity at that point because in the next instant the spring becomes compressed and begins to push back on the block and to decelerate it. After traveling another 10 cm, for a total of 20 cm along the surface, the spring is fully compressed, and by symmetry the block will have been slowed to have zero velocity at that instant. Although the velocity is momentarily zero, the spring force is maximal and equal to 1 N in the opposite direction to the initial force, producing a maximal acceleration of $0.5\ \text{m/s}^2$ in that same direction.

We have qualitatively explained the oscillatory motion we observed, but can we provide a quantitative explanation as well? Using a computer, we can curve fit the data in Figure 3.11 to a cosine function,

$$x = A\cos\left(\frac{2\pi t}{T}\right), \tag{3.13}$$

where A, known as the amplitude, is the maximum value that x reaches and T, known as the period, is the repeat time of the oscillating cosine function. Reading these values directly off the graph shows that $A = 10$ cm and $T = 2$ s for these data, so that the mass on the spring has a position that oscillates around the origin according to $x = 10\cos(\pi t)$, with t measured in seconds and x in cm.

The motion of the mass on a spring is an example of a more general type of *cyclic* or *periodic motion* that repeats itself with a regular time interval. Spring motion is also known as an *oscillatory motion* because it is a back-and-forth periodic motion like that of a pendulum, as contrasted with, for example, the periodic motion of the Earth around the sun each year. The oscillatory motion of a mass on a spring represented by Equation (3.13) is known as *simple harmonic motion*. The term harmonic comes from the mathematical definition of the sine and cosine as harmonic functions. It is an ideal limit, because it represents oscillatory motion that persists forever with the same amplitude. In Chapter 10, we return to this problem and give more realistic models to describe oscillatory motion. In the rest of this section we pursue the ideal motion of a mass on a horizontal spring and see what more we can learn about simple harmonic motion.

Remembering back to the beginning of the last chapter, we can use those techniques to analyze Figure 3.11 for the velocity of the mass oscillating on the horizontal spring. Recall that we need to compute the slope of the smooth curve extrapolated through the datapoints as a function of time in order to plot the velocity of the mass as a function of time. This can be done using the help of a computer to find the results shown in Figure 3.12. We noted above that the position versus time data looked like a cosine curve; these new results look similar in that they oscillate with the same period of $T = 2$ s but they are both shifted over in time and have a larger amplitude. After a bit of thought we can recognize the shape of the curve to be the negative of a sine curve, or a sine curve that has been shifted over by half of its period, and can be represented as

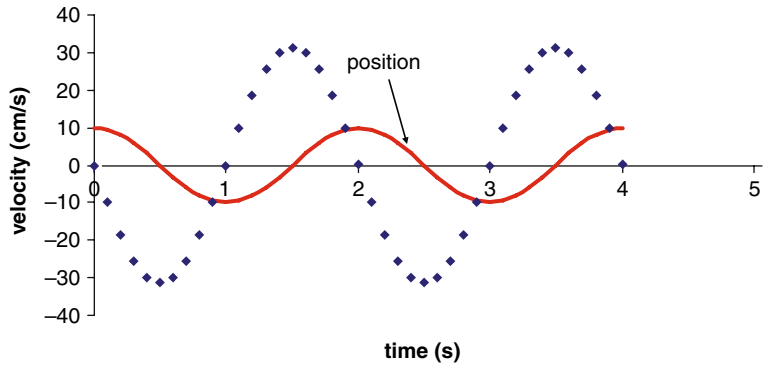


FIGURE 3.12 *Velocity versus time data* (blue) *obtained for the mass on a horizontal spring from the running slope of the position versus time graph shown in red.*

$$v = -v_{\max}\sin\left(\frac{2\pi t}{T}\right), \tag{3.14}$$

where $v_{\max}$ is the amplitude of the curve or, with $T = 2$ s, as $v = v_{\max}\sin(\pi t)$. A more complete analysis (see the box on the next page) shows the connection between the amplitudes of the position and velocity graphs

$$v_{\max} = \left(\frac{2\pi}{T}\right)A, \tag{3.15}$$

or in this case $v_{\max} = \pi A = (3.14)(10\text{ cm})/\text{s} = 31.4$ cm/s, in agreement with the plotted values.

To generate a plot of the acceleration of the mass as a function of time we repeat our computer slope calculation, this time plotting the slope of the velocity versus time graph to obtain Figure 3.13. Note that the plot is the negative of a cosine graph, the period is the same 2 s, and that the amplitude is even larger. We can write the functional form of the graph as

$$a = -a_{\max}\cos\left(\frac{2\pi t}{T}\right), \tag{3.16}$$

and in this case with $T = 2$ s, $a = -a_{\max}\cos(\pi t)$. The analysis in the box below using calculus shows that

$$a_{\max} = \left(\frac{2\pi}{T}\right)^2 A = v_{\max}\left(\frac{2\pi}{T}\right), \tag{3.17}$$

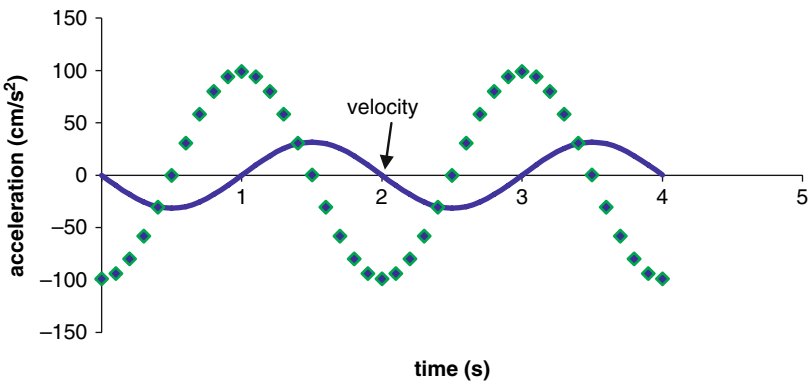


FIGURE 3.13 *Acceleration versus time* (green) *for the mass on a horizontal spring example, obtained from the running slope of the velocity versus time graph shown in blue.*

or in this case with $T = 2$ s, $a_{max} = \pi^2 A = (3.14)^2(10\text{ cm})/\text{s}^2 = 98.6\text{ cm/s}^2$, in agreement with the graph. Note the important result from substituting Equation (3.17) into Equation (3.16) that

$$a = -\left(\frac{2\pi}{T}\right)^2 A\cos\left(\frac{2\pi t}{T}\right) = -\left(\frac{2\pi}{T}\right)^2 x \tag{3.18}$$

so that the acceleration and position of the mass are proportional, in agreement with Equation (3.12) which states that $a = -(k/m)x$.

From our set of three graphs for the position, velocity, and acceleration of the mass undergoing simple harmonic motion we see that the velocity and acceleration changes are out of phase with the position. In particular, the velocity varies as sin $2\pi t/T$ rather than cos $2\pi t/T$. This is expected because, for example, when the particle is at its amplitude at $t = 0$, $T/2$, T, . . . , with its largest acceleration in the opposite direction due to the maximal restoring force, the velocity vanishes instantaneously. Similarly, while at the equilibrium position where the force and acceleration vanish instantaneously, the particle has its maximum velocity in either direction.

At this point in our discussion of simple harmonic motion we can answer the important question: what determines the period of oscillation of a mass on a spring? If we compare Equations (3.12) and (3.18), we can write that

$$a = -\frac{k}{m}x = -\left(\frac{2\pi}{T}\right)^2 x. \tag{3.19}$$

Equating the coefficients of x in this expression we can solve for the period to find

$$T = 2\pi\sqrt{\frac{m}{k}}. \tag{3.20}$$

We see that the period of oscillation is proportional to the square root of the oscillating mass and inversely proportional to the square root of the spring constant. The larger the mass is, the larger the period, and the stiffer the spring is, the shorter the period. These observations should make intuitive sense. What is not so intuitive is that the period is independent of the amplitude of the oscillation. No matter what amplitude we give to the mass on the spring when we start the motion, the period will be the same. This is true as long as Hooke's law is obeyed, the so-called "linear response" of the spring. For large amplitudes, nonlinear forces will act and the period will no longer be independent of the amplitude.

We can check the prediction for the period based on Equation (3.20), by comparing its calculated value with the value read off the plot of about 2 s. From our experimentally determined spring constant of 10 N/m measured using hanging weights and with a value of 1 kg for the mass used in the oscillation experiment, we find a predicted value of

$$T = 2\pi\sqrt{\frac{m}{k}} = 2\pi\sqrt{\frac{1}{10}} = 1.99\text{ s},$$

in good agreement with the experimental period.

A useful parameter to introduce here is the *frequency f*, or number of oscillations per second (measured in hertz, Hz, or oscillations/s),

We can solve for the displacement of the mass on a spring directly from Newton's second law, Equation (3.12), by using the definition of a

$$ma = m\,\text{d}^2x/\text{d}t^2 = -kx, \text{ or}$$

$$\frac{\text{d}^2x}{\text{d}t^2} + \frac{k}{m}x = 0.$$

This is an example of a differential equation for $x(t)$, an equation with derivatives and functions of x, which is to be solved for $x(t)$. This equation of motion for the mass on a spring states that the second derivative of x is proportional to $-x$. Trying a solution of the form

$$x = \text{A}\cos(\omega t),$$

and substituting this into the above equation and differentiating twice, we find that

$$(-\omega^2 A)\cos(\omega t) + \frac{k}{m}A\cos(\omega t) = 0.$$

In order for this equation to hold we must have

$$\omega^2 = k/m \text{ or } \omega = \sqrt{\frac{k}{m}}.$$

We can also find an expression for the velocity of the mass, Equations (3.14) and (3.15), or (3.23), by differentiating the equation for x,

$$v = \text{d}x/\text{d}t = -A\omega\sin(\omega t).$$

which is simply related to the period because one oscillation occurs in a time T, so that

$$f = \frac{1}{T} = \frac{1}{2\pi}\sqrt{\frac{k}{m}}. \tag{3.21}$$

The larger the frequency of oscillation is, meaning the greater the number of oscillations per second, the shorter the period. Another parameter worth introducing is the *angular frequency* (or *angular velocity*) ω, measured in rad/s, and related to the frequency and period through

$$\omega = 2\pi f = \frac{2\pi}{T}. \tag{3.22}$$

The angular frequency is introduced here more for convenience because the factor $(2\pi/T)$ appears in many of the above equations. Rewriting some of the above equations in terms of the angular frequency we have the following collection:

$$x = A\cos(\omega t); \quad v = -A\omega\sin(\omega t); \quad a = -\omega^2 x; \quad \omega = \sqrt{\frac{k}{m}} \tag{3.23}$$

Example 3.5 A 0.1 kg mass is attached to a linear vertical spring and set into oscillation. If 0.25 s is the shortest time for the mass to travel from its highest to its lowest point, find the period, the frequency, and the angular frequency of the simple harmonic motion.

Solution: The trajectory from highest to lowest point is half of a full cycle of the motion so that the period would be 0.5 s. The frequency is then equal to 2 Hz, because two full cycles occur in 1 s. The angular frequency is equal to $\omega = 2\pi f = 4\pi$ rad/s or 12.6 rad/s.

Example 3.6 A 0.5 kg mass is hung from a spring, stretching it a distance of 0.1 m. If the mass is then pulled down a further distance of 5 cm and released, find (a) the period of the oscillations, (b) the maximum height the mass reaches from its release point, (c) the maximum acceleration the mass experiences, and (d) the maximum velocity of the block.

Solution: (a) According to Equation (3.20), the period of the motion depends only on m and k. From knowing that the 0.5 kg mass initially stretches the spring by 0.1 m, we can compute the spring constant to be $k = F/x = (0.5\text{ kg})(9.8\text{ m/s}^2)/0.1\text{ m} = 49\text{ N/m}$. On substitution into Equation (3.20), we find that

$$T = 2\pi\sqrt{\frac{0.5}{49}} = 0.63\text{ s}.$$

(b) The mass oscillates with an amplitude of 5 cm around the equilibrium position (the initial suspension height). Therefore, the mass rises at most 10 cm above its starting point where it again reaches its amplitude but above the equilibrium position. (c) As the mass oscillates, its acceleration is given by $F_x = ma_x = -kx$. At first glance you might wonder why we have seemingly neglected the weight of the hanging mass. This was intentional because in stretching the

(*Continued*)

spring 0.1 m when first connected, the spring supports the weight, allowing the mass to stay suspended at equilibrium. When the spring is further stretched, it supplies the additional force $-kx$, where x is the displacement from the equilibrium point. The maximum acceleration thus occurs when x is at its minimum value of -5 cm, measured from the equilibrium position, or at the starting position. This acceleration is given by $a_x = -kx/m = -(49\ \text{N/m})(-0.05\ \text{m})/0.5\ \text{kg} = 4.9\ \text{m/s}^2$. At the topmost point of its oscillation 10 cm above the starting point, the mass has this same value of acceleration but directed downward. (d) According to Equation (3.23), the maximum velocity is given by ωA, because the sine has a maximum value of 1. To find ω, we note from Equation (3.22) that

$$\omega = \frac{2\pi}{T} = \frac{2\pi}{0.63\,\text{s}} = 10\ \text{rad/s}.$$

Then the maximum velocity of the mass is given by ωA = (10 rad/s)(0.05 m) = 0.5 m/s.

Example 3.7 The two hydrogen atoms in the hydrogen molecule H_2 oscillate about the center of mass of the molecule with a natural vibrational frequency of 1.25×10^{14} Hz. What is the spring constant of the effective spring equivalent to the bonding forces in the molecule? You will need to know that the effective mass of H_2 for motion about the center of mass is 1/2 the mass of a hydrogen atom.

Solution: We know that the angular frequency of oscillation is related to the spring constant and the mass through Equation (3.23). Using a hydrogen mass of $1\ \text{u} = 1.67 \times 10^{-27}$ kg, we can solve for k in Equation (3.23) to find

$$k = \omega^2 m = (2\pi f)^2 m = 4\pi^2 f^2 m = 520\ \text{N/m}.$$

This is a typical value for the effective spring constant of a single covalent bond. Weaker ionic bonds, such as in NaCl, have smaller spring constants of about 100 N/m, and double or triple bonds have stiffer spring constants, with values up to several thousand N/m.

4. FORCES ON SOLIDS AND THEIR ELASTIC RESPONSE; BIOMATERIALS AND VISCOELASTICITY

4.1. ELASTIC RESPONSE OF SOLIDS

Just how strong is the force that holds ordinary objects together? To get a rough idea we can perform a pulling experiment. For example, let's attach a weight to the end of a copper wire hanging vertically that is 1 m long and has a cross-sectional area of 10^{-6} m^2. If we add 5 kg to the end of the wire, it will stretch by about 5×10^{-4} m (i.e., by about 0.5 mm). If we add 10 kg (i.e., double the added force), the stretch is about 10^{-3} m (double the stretch). A shorter wire of the same cross-sectional area doesn't stretch as much. A 0.1 m long wire (one tenth as long as the original) that has 10 kg added only stretches by about 10^{-4} m (one tenth as much as the original). A thicker 1 m long wire also doesn't stretch as much as the original. A wire that has a cross-sectional area of 10^{-5} m^2 (ten times the cross-sectional area of the original) and 10 kg added stretches by about 10^{-4} m (one tenth as much as the original). These results can be summarized (see Figure 3.14) by saying that the amount a copper wire stretches when a force is applied to its ends is: (1) proportional to the applied force, (2) proportional to the original length,

and (3) inversely proportional to the cross-sectional area. Interestingly, if we remove the added weight, the wire returns to its original length. Such a stretch with a return to the original form is called an *elastic deformation*. Of course, all of these observations are invalid if too much weight is added. If too much weight is added, the wire can permanently stretch (*plastic deformation*) or even break.

The rule for elastic deformation that we have written in words can be written in equation form:

$$\frac{F}{A} = Y\frac{\Delta L}{L}, \tag{3.24}$$

where F is the applied force, A is the cross-sectional area, L is the original length, and ΔL is the stretch. The constant of proportionality is called *Young's modulus*. It is a number with units N/m^2 that measures the strength of a material. Materials with larger Ys are more difficult to pull apart than materials with smaller Ys. The left-hand side of the equation, F/A, is called the applied *stress* (in this case tensile stress) measured in N/m^2, or pascal (1 Pa = 1 N/m^2), whereas the ratio $\Delta L/L$ is the resulting (dimensionless) *strain* produced.

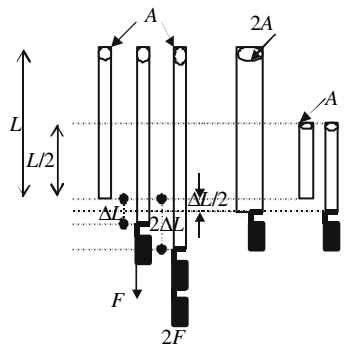


FIGURE 3.14 *Results from elastic deformation of a wire. Left three: Wire of length L stretches a distance ΔL with force F and 2ΔL with 2F. Thicker wire with 2A cross-sectional area is only stretched ΔL/2 by force F. Right two: Wire of length L/2 and area A is only stretched by ΔL/2 by force F.*

Example 3.8 Given the data in the preceding paragraph, estimate Young's modulus for copper.

Solution: When 10 kg is added to the 1 m wire of cross-sectional area 10^{-6} m^2 it stretches by about 10^{-3} m. The weight of a mass of 10 kg is (10 kg) $\times$ g ~ (10 kg) $\times$ (10 N/kg), about 100 N. Thus, 100 N/10^{-6} m^2 ~ Y (10^{-3} m)/(1 m). Solving for Y, $Y \sim 10^{11}$ N/m^2.

The elasticity of solids is due to the fundamental atomic nature of materials. Individual atoms and molecules in a solid are bound to each other by electromagnetic forces that, to a reasonable approximation, can be treated as a set of stiff connecting springs. For small deformations this is a very good model of a solid and we can imagine that the shape changes in a solid are due to small compressions or expansions of the set of springs keeping the solid intact. This model of a solid held together by effective springs can give rise to the entire set of properties of the solid, including its thermal and electrical properties, although we do not study these here, as well as its structural properties discussed below.

When a solid, which is not free to translate or rotate, is subject to external forces it will deform. If the solid were perfectly rigid, there would be no response, or deformation, whatsoever. All real solids, however, are deformable, and it is this phenomenon that we wish to study. In biology there are a number of structural solids whose properties are fundamental to the life processes of the organism. These include bone; soft tissue such as cartilage, skin, and blood vessels; shells; wood; and many others, including artificial medical implants.

Imagine putting the femur bone (the long bone of the thigh) under tension by exerting forces on either end along the long axis of the bone and away from the bone's center (Figure 3.15). If we were to gradually increase the magnitude of the tensile force, just as we described above for the copper wire, and measure the length of the bone as a function of the applied force we would be able to plot the graph shown in Figure 3.16. For relatively small applied forces, the bone stretches by small proportional amounts in the linear portion of the graph; of course, for a bone the applied forces needed to produce a significant length change are very large as we quantitatively work out shortly. If the applied force is removed, the femur returns to its original length because it is *elastic*, just like a spring. As the applied forces get somewhat larger, the bone response is no longer linear, but even beyond this linear limit on removal of the force the bone still returns to its original length. In this range of forces, the effective springs defining the

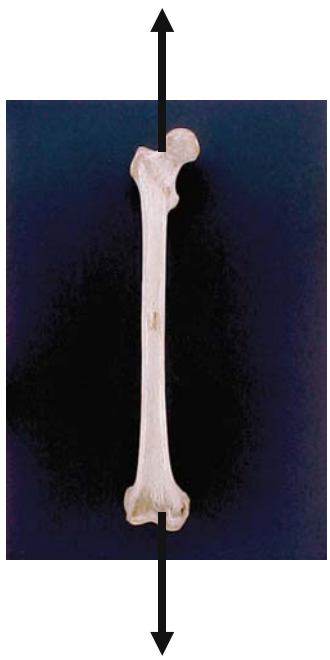


FIGURE 3.15 *The femur under tension.*

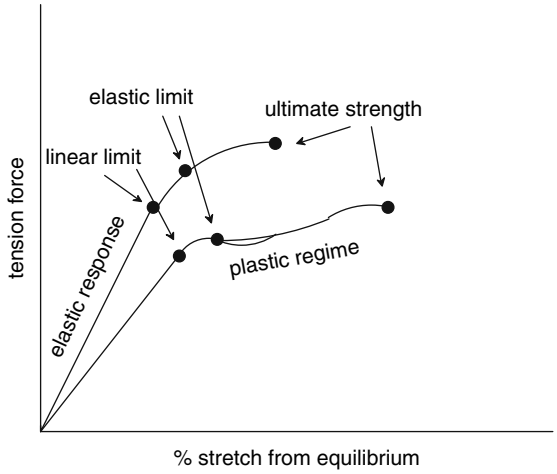


FIGURE 3.16 *The upper curve is the typical response of bone and the lower curve shows the plastic behavior of some other materials.*

internal structure of the bone become nonlinear but we are still in the elastic regime. As the applied force is further increased the femur will break, or *fracture*, at a certain value, known as its *ultimate strength*. In the adult human femur this happens when stretched by about 3%. For other materials, such as metals, glasses, and some polymers, beyond a certain applied force, the *elastic limit* is reached and the material enters a *plastic regime* in which it is permanently deformed even when the forces are removed. For bone, the plastic regime does not exist but every solid material will have a qualitatively similar force–elongation curve with linear and nonlinear regimes, an elastic limit, and ultimate strength, if not plastic and viscoelastic (see below) regimes.

The linear elastic regime is described by Equation (3.24). The same expression also applies to the case when the applied forces tend to compress the bone, as, for example, when standing upright, although the Young's modulus for compression is roughly 1/3 that of the modulus for tension. This difference is due to the anisotropic nature of bone and leads to greater strain for the same stress on compression over that on tension. In addition the ultimate strength of bone is over 25% greater for compression than for tension. If we rewrite Equation (3.24) in the form

$$F = \frac{YA}{L_0}\Delta L,$$

and note that the restoring force the solid exerts is $F_{\text{restore}} = -F$, then we see that solids also obey Hooke's law with an effective spring constant

$$k = \frac{YA}{L_0}. \tag{3.25}$$

Example 3.9 Estimate Young's modulus for compression of bone from the following data. The femur of an 85 kg person has an effective cross-sectional area of about 6 cm^2 and a length of about 0.5 m. When the person lifts a 100 kg mass, careful measurements show that the femur compresses by about 0.04 mm. Also, if the ultimate compressive strength of the femur is 1.7×10^8 Pa, find the maximum weight that the femur can support.

Solution: The 100 kg mass is assumed to be carried equally by both legs, so that the load on each leg is a force of (50 kg) (9.8) = 490 N. The added stress is then found to be $F/A = 8.2 \times 10^5$ Pa, which results in a strain of (0.04 mm)/(0.5 m) = 8×10^{-5}. Young's modulus is then found as the ratio of the stress to strain, or $Y = (8.2 \times 10^5)/(8 \times 10^{-5}) = 10^{10}$ Pa.

From the ultimate compressive strength, we find that the maximum weight that the femur can support is F = (ultimate strength)(area) = $(1.7 \times 10^8)(6 \times 10^{-4}) = 10^5$ N. This enormous weight implies that normally the femur will not fracture under compressive forces. We show below, however, that it is much more common for large bones to fracture under bending or twisting.

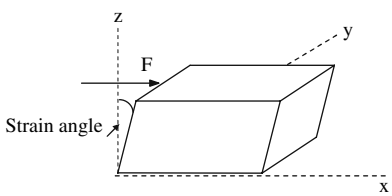


FIGURE 3.17 *A solid undergoing shear deformation due to the shearing force* F.

An implicit assumption here is that the solid is uniform and isotropic throughout (Y does not depend on direction). Although not considered here, in cases where the material is anisotropic (some crystals, e.g.), Young's modulus may differ in each of three orthogonal directions, x, y, and z, and there will be three different expressions for Equation (3.24) along the x-, y-, and z-axes with three different normal stresses, strains, and Young's moduli.

If the applied force is not normal to the surface, but parallel to the surface (Figure 3.17), the type of stress applied is called a *shear stress*. In this case the

response of the material is a *shear strain* deformation in which the solid distorts to different extents along the direction parallel to the surface. In the linear case, this distortion results in a constant strain angle as shown in the figure. Again the stress and strain are proportional with, in this case, the proportionality constant known as the *shear modulus*. Once again, we remark that if the material is anisotropic there will be various shear stresses, strains, and moduli. In this case there are six possible shear stresses, because for a force applied along the x- (or y- or z-) axis, there are two possible independent planes of orientation, the xy or xz (or four others) (see Figure 3.17 where the shear stress is along the x-axis and the strain angle is shown for shear of the xy planes). Corresponding to these six shear stresses there are six shear strains and six shear moduli.

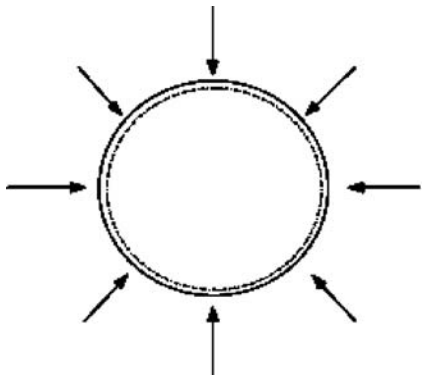


FIGURE 3.18 *An object immersed in a fluid has a pressure (force/area) acting on it from the fluid normal to every surface.*

These six shear strains and the three normal strains mentioned above for tensile stresses together form a 9-component, 3×3 array, called the strain tensor. The mathematics of tensor analysis allow one to write relations between the stress and strain tensors that describe all of the elastic moduli and to set up any problem in the linear deformation of solids, most of which then need to be solved numerically by a computer. This type of analysis is used, for example, by mechanical and civil engineers in construction projects using steel or concrete beams or by bioengineers designing artificial limbs.

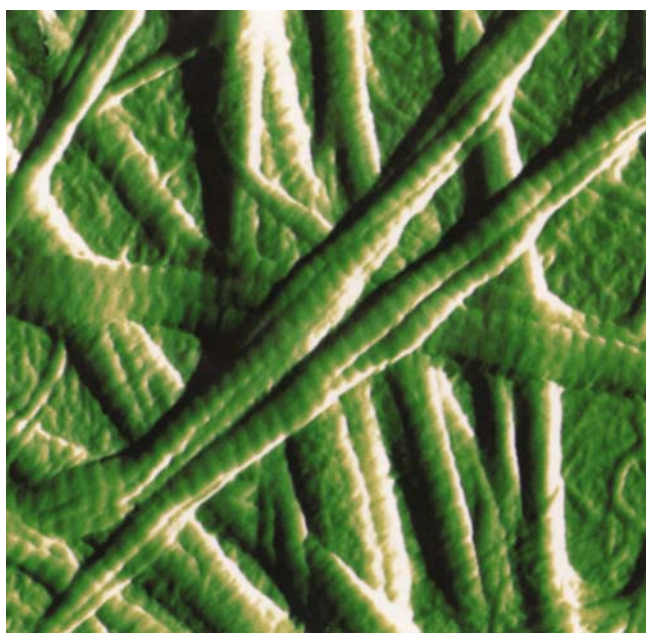
A related type of stress–strain relation is for torsion, or twist around some axis of rotation. This is a particularly prevalent type of stress for bone and most leg fractures are torsional fractures. For example, skiers are particularly susceptible to this type of fracture because bone is weak under torsion and, as we show in Chapter 7, long skis make it easy to twist the leg bones.

One last type of stress–strain relation should be mentioned here. When an object is immersed in a fluid, the fluid exerts a force normal to the surface of the object everywhere (Figure 3.18). This force per unit area is called the pressure. We consider pressure in much more detail in Chapter 8. In this case the analogous quantity to the strain is a small fractional change in the volume of the object and the proportionality constant between the pressure and the strain is known as the bulk modulus.

4.2. BIOMATERIAL STRENGTH

In the world of biomaterials, there are certain motifs that recur both in structural proteins and, on a larger scale, in bone, tissue, and muscle. On a microscopic scale, most proteins involved in providing structural strength are organized into filaments. Notable examples include actin and myosin (the major muscle proteins), collagen (a major component of bone and connective tissue), tubulin (the major protein in microtubules which provide a cellular framework), and the keratins (a class of proteins found mostly in vertebrate horn, hoof, hair, and skin). Within this motif there are variations, but a key point is the elastic nature of the structures formed.

FIGURE 3.19 *Image of collagen fibers.*



Collagen is the most abundant protein in mammals and is fundamentally a stiff triple helix that associates into bundles. In connective tissue these collagen fibers are cross-linked together in a network by a protein called elastin. Elastin is perhaps the most elastic of known proteins and is responsible for the high elasticity of skin and blood vessels. In tendon, collagen bundles associate in a repetitive pattern of filamentous structures, as shown in Figure 3.19, in association with water, polysaccharides, and other proteins. In bone, a very dense and specialized form of connective tissue,

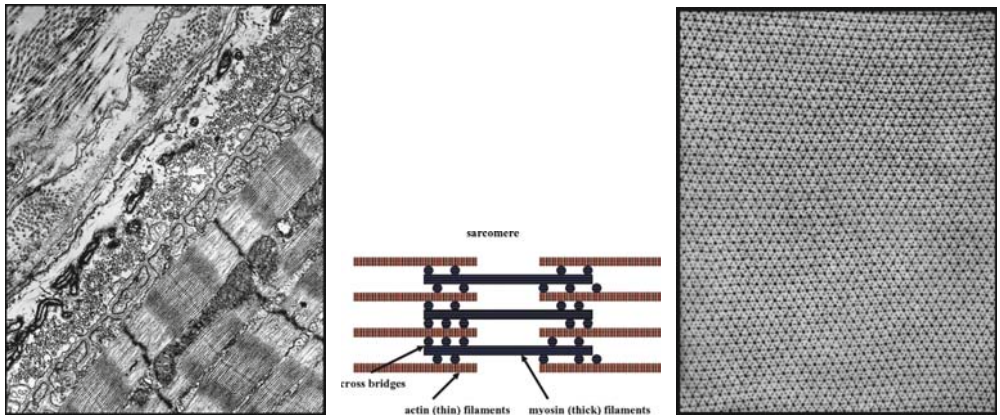
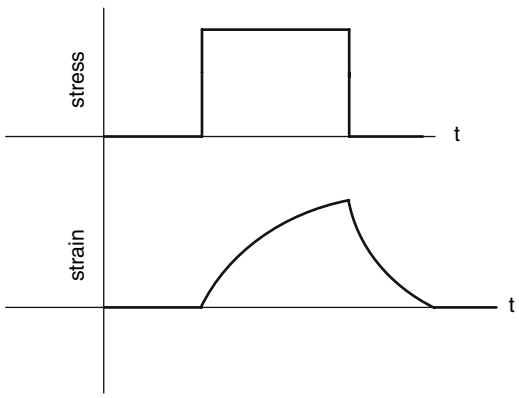


FIGURE 3.20 *Arrays of muscle filaments in a muscle fiber bundle or fibril:* (left) *microscope image of longitudinal array showing thick and thin filaments overlapping in lower right portion of photo; (center) schematic for the interpretation of the left image. Thick filaments show up darker in the microscope image;* (right) *cross-sectional view through a muscle fiber bundle showing the thick and thin filaments in a hexagonal array.*

solid deposits of minerals are present in addition. Collagen filaments are very effective at resisting tensile stresses, and the mineral deposits in bone resist compressive stresses. The composite material bone has tensile and compressive moduli nearly equal to that of aluminum.

Actin is a small (~5 nm) globular protein that self-associates to form long filaments, known as F-actin, in the cytoplasm of cells and with other associated proteins in the form of thin filaments in muscle. In cells, the process of actin self-association, or polymerization, has been shown to provide sufficient force to change the shape of cells and actin is known to be intimately involved in generating force for cellular locomotion. Myosin, which has a rodlike "tail" end and two globular "heads," forms the thick filaments of muscle by the ordered aggregation of the tail portions of the myosin together with other proteins. In muscle, these two filamentous structures, the thin and thick filaments, interdigitate in a regular hexagonal array in a muscle fibril (Figure 3.20). These two sets of independent filaments interact with each other via the "cross-bridges," or heads of myosin. In a complex chain of chemical and structural events that is only partly understood in detail, the myosin heads attach to specific sites on the actin molecules and, using the energy released by the hydrolysis of ATP, undergo a structural change that forces the thin and thick filaments to slide relative to one another, thus shortening the muscle fibril. In a muscle, these myofibrils, each about 1 μm diameter, are themselves organized in a series of regular arrays. All muscles generate tension forces by shortening their overall length. Our bodies use sets of pairwise antagonistic muscles to allow us to move our limbs about our joints in various directions. The composite structure of muscle and tendon or bone is a second motif in structural proteins: the overall structure consists of subunits that are in turn made up of many similarly organized smaller subunits.

FIGURE 3.21 *The phenomenon of creep.*



4.3. VISCOELASTICITY

Most biomaterials do not obey a linear relationship between applied stress and strain nor can they be analyzed solely in terms of their elasticity. Biomaterials also have a viscous component to their response to an external stress, a phenomenon known as *viscoelasticity*. What are the characteristics of viscoelasticity? If a constant stress is applied to a viscoelastic material for a fixed time interval, the characteristic strain response is shown in Figure 3.21. This phenomenon is called

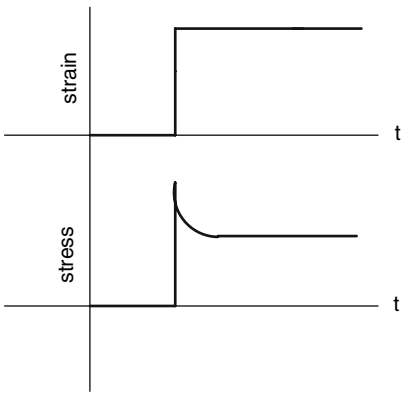


FIGURE 3.22 *Stress relaxation. There is no connection here to yoga.*

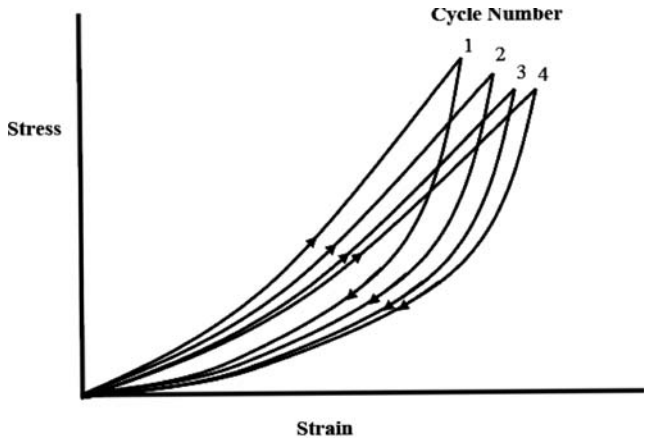


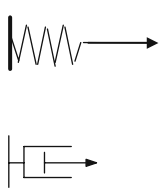
FIGURE 3.23 *Hysteresis shown for the deformation of ligament/tendon where the graphs for application and removal of stress do not superimpose. Also, with repeated stress, the hysteresis curve also shifts to larger strains. Both of these are viscoelastic phenomena.*

creep, the slow progressive deformation under constant stress. A related property, known as *stress relaxation*, is illustrated in Figure 3.22 where the material is held at constant strain, clamping its length, and the stress is found to relax over time.

If a material is subjected to a cycle of applied stress and removal of the applied stress, the stress–strain relationship for viscoelastic materials is not reversible, as shown in Figure 3.23 for ligament/tendon, and the material exhibits *hysteresis*, or irreversible behavior. This irreversibility means that the stress–strain path on elongation is different from the path on relaxation back to the original unstressed position. Viscoelasticity should be distinguished from plasticity, mentioned earlier in connection with nonreversible deformations at high stress, in that viscoelastic materials return to their original shape after applied stresses are removed, but only after some time has elapsed. Nearly all biomaterials exhibit some degree of creep, stress relaxation, and hysteresis, but to different extents and with different characteristic times involved.

In order to characterize viscoelastic materials, two types of mechanical experiments can be done. In one case transient constant stresses or strains are applied and the response of the material is investigated. Usually either creep or stress relaxation is studied in this method. In the other case cyclic, or dynamic, stresses or strains are applied and the time-dependent response of the material is investigated as a function of the frequency of deformation. By examining the frequency dependence of both the elastic and viscous moduli, separately, as functions of frequency, this method often can lead to models for the molecular origin of the viscoelastic behavior. We mention here that other nonmechanical types of characterization, such as ultrasonic and spectroscopic methods, can be used to study the elastic properties of materials as well. Also, in recent years a new type of microscopy (atomic force microscopy; see Chapter 7) has been used to measure variations in the elastic modulus of bone with a spatial resolution of about 50 nm and has shown a strong correlation between the elastic and structural properties of bone. In addition to characterizing natural biomaterials, viscoelastic measurements are also performed on various implant and prosthesis materials.

Models of viscoelastic behavior usually use various combinations of simple elastic springs and simple viscous dashpots (Figure 3.24), representing the ideal viscous behavior of a simple fluid in which the stress is proportional not to the strain, but to the time rate of change of the strain as we show when we discuss viscous fluids in Chapter 9. A dashpot is a mechanical element, pictured as a piston, with a frictional force between the piston and outer walls of the cylinder that



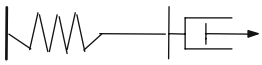
spring: F = –kx

elastic solid: stress = (elastic modulus) (strain)

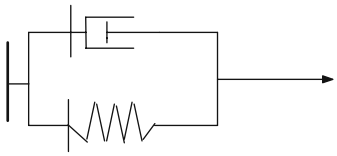
dashpot: F = –fv

viscous fluid: stress = (viscosity) (strain rate)

FIGURE 3.24 *Linear springs and dashpots and their analogy with elastic solids and viscous fluids.*



Maxwell model-in series



Kelvin-Voigt model -in parallel

FIGURE 3.25 *Two simple arrangements of linear springs and dashpots used to model the viscoelastic properties of materials.*

depends on the velocity of the piston. These elements (ideal springs and dashpots) can be connected in various ways (series, parallel, or combinations: Figure 3.25) in order to model different types of viscoelastic behavior. When connected in series (the Maxwell model) the spring, under an applied stress, will deform instantly whereas the dashpot will deform continuously while the stress is applied. This model is often used to describe *viscoelastic fluids* because those materials will flow while the stress is applied. When connected in parallel (the Kelvin–Voigt model) the spring will limit the deformation of the dashpot under the applied stress, and this model is often used to describe *viscoelastic solids*, those materials that are more solidlike in their behavior. In some respects this type of analysis is very similar to electrical circuit analysis with various electrical elements connected together, a topic that we discuss in more detail later on.

5. STRUCTURE AND MOLECULAR DYNAMICS OF PROTEINS

Biomolecules are biologically significant molecules that are usually quite large and are therefore also termed macromolecules. They have an enormously complex and rich variety of structures but are made from simpler structural building blocks. For example, proteins are all made from the 20 or so different amino acids, each of which is a relatively small well-defined structure. Human cells manufacture on the order of 60,000 different proteins with the structure of each uniquely related to its function. Most proteins, for example, have a very simple *primary structure*, simply a single linear string of amino acids forming the backbone of the protein. The sequence of amino acids along the backbone is unique for each different protein and sometimes a single amino acid substitution, through an error in protein manufacture by the cell or through genetic engineering, can result in a defective protein that leads to a specific disease. A prime example of this is sickle-cell anemia, a crippling disease that causes red blood cells to deform and clog capillaries and which is caused by a single incorrect amino acid in the hemoglobin molecule.

The primary structure of a protein contains all the information necessary for the protein to spontaneously fold and attain a unique overall conformation, or three-dimensional structure. Scientists have discovered this by unfolding proteins through gentle heating until they have lost all ordered structure and then cooling the proteins to watch them spontaneous refold to form the completely native and functional structure. Different categories of structural motifs have been discovered as more and more proteins have had their detailed structures determined. There are various types of helical structures in which the amino acids are arranged through ordered repeating hydrogen bonds to form helices of different detailed structures. The *α-helix* (Figure 3.26, left) is a common example, although there are many other types of known helices naturally occurring in proteins.

Another structural motif is the *β-pleated sheet* structure (Figure 3.26, right) in which portions of the backbone, either contiguous or separated, associate side to side to form a structural sheet. These locally organized regions of a protein make up what is termed the *secondary structure* of the protein. Some proteins consist entirely of a single motif; for example, there is a class of elongated proteins called the fibrous proteins that are helical in structure and include the important structural macromolecules of actin, myosin, collagen, and the keratins. Others are termed globular and can have regions with different structural motifs (Figure 3.27 left), but yet with a unique overall three-dimensional structure, held together by a variety of weaker bonds and known as the *tertiary structure*.

Still other proteins are composites, consisting of several independent protein subunits that are then more loosely associated together as, for example, in hemoglobin (Figure 3.27, right) with four such structural domains and an iron atom bound at the juncture of the subunits. The structural relationship between the subunits of such composites is known as the *quaternary structure*. Thus, we see that the overall structure of a protein involves strong co-valent bonding along the

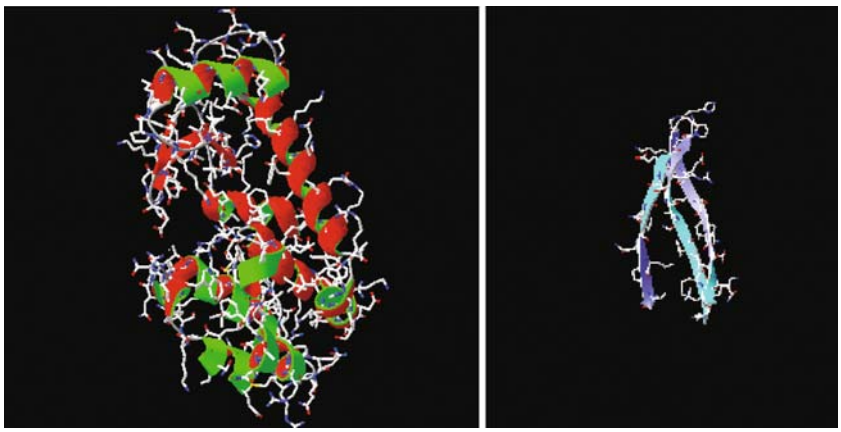


FIGURE 3.26 (left) α-*Helix and* (right) β-*sheet, with ribbons showing folding.*

backbone, weaker local bonding determining the local conformation, and perhaps weaker still bonding between more distant portions of the backbone to provide the overall stability of the protein.

Recently there has been rapid growth in our understanding of not only the structural motifs available to macromolecules (proteins in particular) but also of the design algorithms or strategies nature uses to produce these motifs. This knowledge has led to major advances in the *protein folding problem*: how a linear macromolecule rapidly undergoes a structural transition to find its native three-dimensional conformation out of the huge number of total possible conformations. There have been some successful projects to design from scratch new small proteins—proteins that do not exist in nature—with well-defined characteristics. This will clearly be an exciting area of future research.

Until now in our discussion we have stressed the structure of macromolecules and the figures that have been used to illustrate the ideas have, by necessity, been static structures. This limitation of the printed page and of molecular model representations has hampered much thinking in biophysical research. Only in recent years has the importance of the dynamics of macromolecules been completely acknowledged by scientists.

Atoms and small molecules constantly undergo very rapid and random thermal motions. The extent of these motions depends on the local environment and the interactions with neighboring atoms and molecules. Even in the solid state, atoms and small molecules execute small vibrational motions about their equilibrium locations. Larger macromolecules or even microscopic objects also undergo random thermal motions, known as Brownian motion or diffusion (refer to Section 7 of Chapter 2). Not only does the entire macromolecule move about due to the solvent collisions, but there are typically also small structural changes rapidly occurring; there are internal motions of different portions of the macromolecule with respect to each other, larger motions of those portions of the macromolecule less tightly bound. Therefore all of the static molecular model representations of structures represent time-average structures. One must keep the notion of dynamical motions clearly in mind because many important functions of proteins involve not only a structural role but a time-dependent dynamical role as well. In some cases the

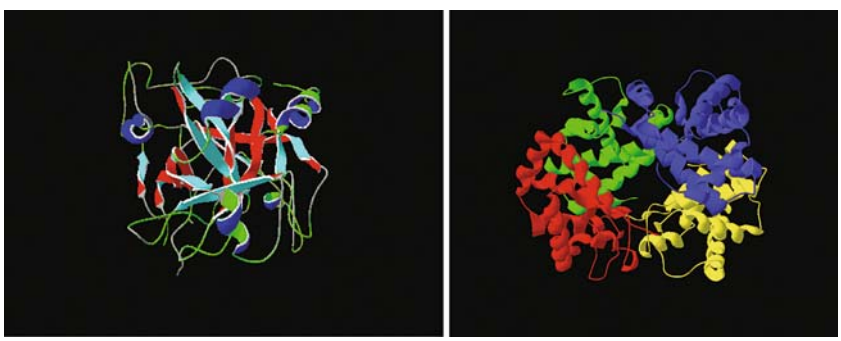


FIGURE 3.27 *Computer models of* (left) *a generic globular protein with helical,* β*-sheet* (arrows), *and random coil* (thin line) *components, and* (right) *hemoglobin with its four identical subunits specifically arranged.*

binding of a small molecule or ion to a macromolecule may cause a large conformational change to occur. Even in these cases it must be kept clearly in mind that the initial and final conformations are not frozen structures.

Molecular dynamics treats each atom in a molecule as a point particle with forces acting on it both from external sources and from other atoms within the molecule. These calculations were not possible until the advent of computers to not only perform the huge number of repetitive calculations, but also to keep track of all the position and interaction variables. Early studies focused on simple liquids in the 1960s, followed by studies of more complex liquids in the 1970s. (Water is a prime example of a complex liquid because it forms a variety of structures from H-bonding.) Dynamical simulations of biological molecules began in the late 1970s with studies on small proteins.

Those first studies started a revolution in our thinking about the structural dynamics of macromolecules. Previously, biological macromolecules were often pictured as rigid structures, in part because our main source of information on their structure came from high-resolution x-ray diffraction studies that gave ball and stick models based on the average positions of the atoms in the macromolecule. These static pictures of biomolecules set the image of structural models. Computer simulations now show a remarkable degree of motion in macromolecules, with portions of the structure having rapid, large amplitude motion, particularly for surface, but also internal, regions. Indeed movies of the motions of macromolecules have been made illustrating the extent of typical movement. Simulations have become a major tool in the study of proteins and have been used to help narrow down (or "refine") the possible detailed structures determined by other physical methods.

The basis for molecular dynamics calculations is the solution of the equations of motion for each atom in the protein. One begins at some arbitrary moment of time with a set of coordinates for each atom based on information from other techniques, most notably x-ray diffraction (see Chapter 23) and nuclear magnetic resonance (NMR; see Chapter 18). Some assumptions are made about the interactions between the atoms so that a set of forces, F_{ij}, can be computed, where F_{ij} is the force on the ith atom due to the jth atom. Then we can write a set of Newton's second law equations, one for each atom, of the form (here we illustrate the method in one dimension; it is relatively easy to generalize this to two or three dimensions as we show in Chapter 5)

$$m_i a_i = \sum_j F_{ij}, \tag{3.26}$$

where the left-hand side of the equation is for the ith particle and the summation notation Σ is used to indicate a sum over all the other atoms labeled j (excluding the term $i = j$) to give the net force on the ith atom. With a given set of forces between the atoms, once the accelerations are determined, they are used to solve for the velocities and positions of all the atoms at the next instant (after some very short time). Then a new set of forces is calculated based on the new positions of the atoms and new accelerations are used to compute the new velocities and positions. This process is repeated countless times to generate a movie of the structure of the macromolecule as a function of time. We show an example of how this is done just below, but the time steps used must be very short indeed and so the number of calculations required is enormous. As a rough rule of thumb, each picosecond (10^{-12} s) of simulation time requires about 1 h of supercomputing time, although this is constantly decreasing as computers are improved.

One method for performing the calculations is to divide time into steps, Δt, of very short duration (~10^{-15} s) and to write difference equations using time as a discrete variable rather than a continuous variable. To do this we can use a modified form of Equation (3.2), the one-dimensional kinematic equation for the position as a function of time for constant acceleration, and write the following difference equations as approximations in one dimension.

$$x_i(t + \Delta t) = x_i(t) + v_i(t)\Delta t + a_i(t)\frac{\Delta t^2}{2} \tag{3.27a}$$

and

$$x_i\,(t - \Delta t) = x_i\,(t) - v_i\,(t)\Delta t + a_i\,(t)\frac{\Delta t^2}{2}. \tag{3.27b}$$

Note carefully the signs in Equation (3.27b). If we add these two equations together, eliminating the velocity term, and substitute for $a_i(t)$ using Equation (3.26) we have

$$x_i\,(t + \Delta t) = 2x_i\,(t) - x_i\,(t - \Delta t) + \frac{\sum_j F_{ij}\,(t)}{m_i}\Delta t^2. \tag{3.28}$$

Equation (3.28) allows us to solve for the position of the *i*th atom at some later time if we know its position at the present and one preceding step in time as well as the current forces acting on it. In a similar way, if the velocities of the atoms as a function of time are of interest, we can subtract Equation (3.27b) from (3.27a) to find an expression for the velocity of the *i*th atom,

$$v_i\,(t) = \frac{x_i\,(t + \Delta t) - x_i\,(t - \Delta t)}{2\Delta t}. \tag{3.29}$$

These algorithms can be used to follow the positions and velocities of each atom at successive times, remembering that the forces F_{ij}, assumed to be constant during the time interval Δt, are re-evaluated after each interval of time because they are dependent on the positions of the atoms that evolve as the calculation is performed (Figure 3.28). To initiate a calculation one needs a set of initial coordinates, usually obtained from other independent information on the structure of the protein or system, as well as either a second set of initial coordinates at a slightly different time or equivalently a set of initial velocities for each atom. Often calculations are initiated using zero for the initial velocities in the so-called zero-temperature limit, and the system is allowed to evolve for some time, reaching an "equilibrium distribution" of velocities. The following very simple example illustrates the major features of a molecular dynamics calculation in one dimension.

Example 3.10 Suppose that there are three atoms of mass m along the x-axis located at $x = 0$, 1, and 2 at time 0 and that were at $x = 0.01$, 1.01, and 1.99 at a time $t = -0.01$, with x and t measured in nm and ps, respectively (see Figure 3.29). Using time steps of 0.01 ps, calculate the positions of each particle for the first three steps of motion assuming the following forces are acting.

$$F_{2\text{ on }1} = \frac{100m}{x_{1,2}^2};\; F_{3\text{ on }2} = -\frac{100m}{x_{2,3}^2};\; F_{3\text{ on}1} = \frac{400m}{x_{1,3}^2}.$$

Solution: Setting up the equations to iteratively (repeated calculations, updated each time) solve for the positions of the three atoms, we have

$$x_1\,(t + \Delta t) = 2x_1\,(t) - x_1\,(t - \Delta t) + \frac{F_{2\text{ on }1}\,(t) + F_{3\text{ on }1}\,(t)}{m}(\Delta t)^2,$$

$$x_2\,(t + \Delta t) = 2\,x_2\,(t) - x_2\,(t - \Delta t) + \frac{F_{1\text{ on }2}\,(t) + F_{3\text{ on }2}\,(t)}{m}(\Delta t)^2,$$

$$x_3\,(t + \Delta t) = 2x_3(t) - x_3\,(t - \Delta t) + \frac{F_{1\text{ on }3}\,(t) + F_{2\text{ on }3}\,(t)}{m}(\Delta t)^2,$$

(*Continued*)

where we need to be careful about the signs; for example, $F_{2 \text{ on } 1} = -F_{1 \text{ on } 2}$ according to Newton's third law. The table below shows the needed calculations for each iteration. Starting with values for x_i for $t = -0.01$ and 0 ps, we first calculate F_{ij} at time 0 and use these forces in the three equations above to get x_i values at $t = 0.01$ ps. Then these new position values are used to calculate the F_{ij} at $t = 0.01$ ps and the process is continued, remembering each time to increment the time by Δt. (Why is 0.01 ps a short enough time interval in this example?)

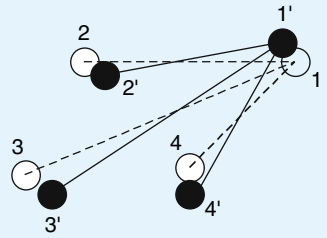


FIGURE 3.28 *A four-atom molecule shown at two times separated by one step in a molecular dynamics calculation. The "primed" numbers show the change in position after time Δt, and the lines show the changes in separation distances from atom 1, reflecting changes in interaction forces between atoms.*

Time (in ps)	*−0.01*	*0*	*0.01*	*0.02*	*0.03*	*0.04*	*0.05*
x_1 (in nm)	0.01	0.0	0.01	0.04	0.09	0.17	0.29
x_2 (in nm)	1.01	1.0	0.97	0.92	0.85	0.76	0.64
x_3 (in nm)	1.99	2.0	2.01	2.02	2.03	2.04	2.04
$F_{2 \text{ on } 1}/m = -F_{1 \text{ on } 2}/m$	—	100	109	129	173	287	—
$F_{3 \text{ on } 1}/m = -F_{1 \text{ on } 3}/m$	—	100	100	102	106	114	—
$F_{3 \text{ on } 2}/m = -F_{2 \text{ on } 3}/m$	—	−100	−93	−83	−72	−61	—

From the table it is clear that particle 1 feels a positive force from both other particles and is accelerating toward the right whereas particle 2 feels a force toward the left from both other particles and is accelerating toward the left. Particle 3 is initially fairly stable in its position, roughly balanced in the short term by oppositely directed forces from the other two particles. Clearly, if nothing else, this example illustrates the need for a high-speed computer to follow the motion over longer times even in one dimension with few particles.

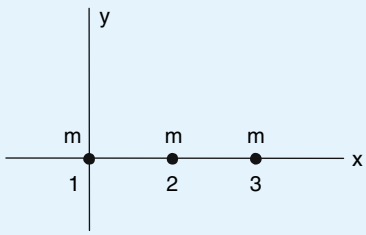


FIGURE 3.29 *Three atoms of mass m for a simple molecular dynamics calculation.*

Of course the crux of any molecular dynamics calculation is to correctly account for all of the forces between atoms, including forces between covalently bonded atoms, longer-range forces between nonbonded atoms, and even forces at the surfaces of the protein between solvent molecules and surface atoms. It has been a triumph of molecular dynamics that such calculations have yielded an understanding of the motions of biomolecules on a subnanosecond ($<10^{-9}$ s) time scale. Limitations of computing power have restricted longer time dynamics to approximations of specific interesting portions of a molecule where some active mechanism is known to occur, such as a molecular hinge or a local conformational change on binding a small molecule. Large-scale slow structural changes still await future studies for further understanding from molecular dynamics simulations.

CHAPTER SUMMARY

Three important examples of one-dimensional motion are discussed in this chapter. The first case is constant acceleration (most notably, free-fall) for which a set of three equations is developed in Table 3.1:

$$v(t) = v_0 + at,$$

$$x(t) = x_0 + v_0 t + \frac{1}{2} at^2,$$

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

When an object moves through a fluid and a (non-negligible) frictional force is present, the object will reach a terminal velocity when the increasing frictional force (proportional to v^2 in Equation (3.5) when the Reynolds number $\Re$ is much greater than 1, or to v in Equation (3.6) when $\Re$ is much smaller than 1) causes the net force on the object to vanish.

Linear springs obey Hooke's Law,

$$F = -kx, \qquad (3.11)$$

where k is called the spring constant. A mass oscillating on a linear spring will have a position, a velocity, and an acceleration given by

$$x = A\cos(\omega t); \quad v = -A\omega \sin(\omega t);$$
$$a = -\omega^2 x; \quad \omega = \sqrt{\frac{k}{m}} \qquad (3.23)$$

and will oscillate with a period T, given by

$$T = 2\pi\sqrt{\frac{m}{k}}. \qquad (3.20)$$

When a solid undergoes a stress (F/A), its linear response is a proportional stress ($\Delta L/L$) according to

$$\frac{F}{A} = Y\frac{\Delta L}{L_o}, \qquad (3.24)$$

where Y is the elastic or Young's modulus. Depending on the relative orientation of the applied forces and the surfaces of the solid, the applied stress can produce a stretch, compression, shear, twist, or pressure.

Molecular dynamics is the simulation of molecular motions by the solution of Newton's second law in an iterative stepwise calculation in which the time steps are very short and the forces on and positions of all the atoms in the molecule need to be recalculated at every time step of the calculation.

QUESTIONS

1. A ball is thrown straight up in the air. What is its velocity at its highest point? What is its acceleration at that point?
2. If a ball is thrown upward with a speed of 6 m/s, what is its velocity when it returns to that height? What is its acceleration at that time?
3. If gravity always acts downward, why does it take the same time for a ball to travel upward as it does for it to return to the same height?
4. If the velocity of a particle is a constant, what does the graph of displacement versus time look like?
5. If the acceleration of a particle increases linearly from zero at time zero to a at time t, what is the average acceleration in that interval?
6. Give some examples of laminar flow of a fluid? What are some examples of turbulent flow?
7. How will the Reynolds number defined in Equation (3.4) change under the following conditions?
   (a) With increasing flow for a given object in a fluid
   (b) For larger objects in the same fluid and at the same flow velocity
   (c) A given object flowing at the same velocity in a variety of fluids with increasing viscosity (usually as the viscosity increases, e.g., by adding a viscous liquid to water in increasing amounts, the density will increase as well but by a much smaller factor).
8. Why is the terminal velocity at large Reynolds number independent of viscosity whereas the value at small Reynolds number is independent of the density? Explain in terms of the relative importance of these two parameters.
9. Why does it take some time for a skydiver to reach a terminal velocity after jumping from a plane?
10. Can you think of an example in which the buoyant force on an object is greater than its weight? Will there be a terminal velocity in that case and, if so, describe it.
11. Explain in words why the force of gravity can be ignored in writing the net force on a mass attached to a vertical spring when measuring displacements from the equilibrium position.
12. A mass oscillates on a vertical spring around its equilibrium position with an amplitude A. Where is

the speed of the mass greatest? Least? Where is the magnitude of the acceleration of the mass greatest? Least?

13. If a mass hanging from a spring has a period of oscillation of 2 s, what will the period be when a second identical spring is also attached to support half the weight?
14. Suppose two identical masses are each suspended from identical springs. If the first is pulled down a distance D and the second is pulled down a distance $2D$, which will complete one oscillation faster? Which will have the greatest maximum speed?
15. A mass oscillates on a vertical spring around its equilibrium position with a period T and an amplitude A. If a second identical mass is added to the first and the amplitude is doubled, what is the new period of oscillation?
16. A mass on a spring oscillates according to the equation $x(t) = 0.035 \cos(0.6\, t)$ (in SI units). What are the period, frequency, angular frequency, and amplitude of the motion?
17. A science museum director wishes to set up 4 spring-mass systems to oscillate with periods that are ratios of each other. Suppose that she wants four oscillators with periods in the ratio 1:2:4:8. She can only find two different types of springs with spring constants that differ by a factor of 4 (k and $4k$) and has only seven masses, one of mass m and six of mass $4m$. Can she do it, and if so, how?
18. Using values that are representative of typical products, compute the spring constant for a suspension spring of an automobile, of a dump truck, and of a grocery scale.
19. Based on the typical values for Y and d_o in solids, what is a typical force acting between atoms in a solid?
20. Springs supply a force that is described by Hooke's law. Because of this a simple, but useful, model to describe the forces between atoms is to imagine that they are connected by microscopic springs. Discuss this picture based on your general knowledge of how a spring pushes or pulls.
21. You have a choice in using steel rods for reinforcing a supporting beam to minimize any compression. One option is to use a rod of length L and radius r and the other is to use two rods of length L and radius $0.6\, r$. Which option will work better?
22. Which column can support a greater weight for a given compression: one with a cross-sectional radius of 5 cm and a length of 50 cm or one of the same material but with a 7.5 cm radius and a 100 cm length?
23. State clearly the difference among the linear limit, the elastic limit, and the ultimate strength of a material.
24. What is the difference between stress and strain? Which one causes the other? Give some examples of stresses and strains.
25. A shock absorber of an automobile functions as a dashpot. Is such a dashpot connected in parallel or in series with the suspension spring? Explain how the car behavior supports your answer.
26. Carefully explain in your own words what it means to solve molecular dynamics problems iteratively.
27. In a molecular dynamics calculation for a protein of 40,000 molecular weight, with a mass composition of 50% carbon, 7% hydrogen, 23% oxygen, 16% nitrogen, and 1% sulfur, and using time steps of 0.1 ps, calculate the total number of iterative calculations needed to follow the dynamics for 10 ns. In the calculation, an average of 10 water molecules per amino acid (with an average of 140 for the molecular weight of an amino acid in the protein) are considered to interact with the protein and each water molecule is treated as a single source of interactions. (Hint: You will need to compute the total number of atoms in the protein and the number of solvent molecules to include in the calculations.)

MULTIPLE CHOICE QUESTIONS

Questions 1–3 refer to a ball dropped from rest and falling vertically under the influence of gravity.

1. The ratio of the distance it falls in a 1 s interval after 4 s to the distance it falls in the next 1 s interval after 5 s is (a) 9/11, (b) 36/25, (c) 25/16, (d) 36/16.
2. The ratio of the ball's velocity at 5 s to that at 4 s after being released is (a) 25/16, (b) 5/4, (c) $\sqrt{\frac{5}{4}}$, (d) 1.
3. The ratio of the ball's acceleration at 5 s to that at 4 s after being released is (a) 5/4, (b) 1, (c) $\sqrt{\frac{5}{4}}$, (d) 25/16.
4. Which of the following is not true of an object in one-dimensional free-fall? (a) the velocity is always zero its highest point, (b) the velocity and acceleration are oppositely directed while moving upwards, (c) the acceleration is not zero at its highest point, (d) the average speed and average velocity are always the same because the motion is one-dimensional.
5. A ball is thrown vertically downward. Taking $g = 10\ \text{m/s}^2$, if in the first second it travels a distance of 7 m, at the end of 2 s it will have traveled a total distance of (a) 14 m, (b) 20 m, (c) 24 m, (d) 32 m.
6. A constant horizontal force is exerted on a cart that is initially at rest on a frictionless horizontal track. The force acts for a time t during which the cart moves a distance d. If the force is halved and applied to the same cart for twice the time, the cart will move a distance (a) d, (b) 2d, (c) 4d, (d) d/2.
7. The frictional force on a small steel ball falling through water is due to (a) buoyancy, (b) viscosity, (c) turbulent flow, (d) thrust.

8. For an object immersed in a fluid, the larger the Reynolds number is the (a) larger the viscosity of the fluid, (b) smaller the density of the fluid, (c) slower the object will fall, (d) none of the above.
9. Laminar flow is characterized by (a) wakes, (b) vortices, (c) streamlines, (d) chaotic flow.
10. For objects with a density near that of water to have a frictional force proportional to both their velocity and to their radius while falling in water, they must have a radius (a) above 1 mm, (b) below 150 μm, (c) between 150 μm and 1 mm, (d) it is not possible.
11. A mass hangs from an ideal spring. When the mass is set into oscillation with an amplitude of 1 cm its frequency is 10 Hz. When the amplitude is increased to 2 cm the new frequency will be (a) 5 Hz, (b) 7 Hz, (c) 10 Hz, (d) 20 Hz.
12. A 50 g mass attached to a spring oscillates vertically with a period of 0.80 s. If the spring and mass are placed on a horizontal surface with negligible friction and the mass is set into motion with the same amplitude as in the vertical case it will (a) oscillate about an equilibrium point that is the same distance from the fixed end of the spring and with the same period, (b) oscillate about an equilibrium point that is closer to the fixed end of the spring and with the same period, (c) oscillate about an equilibrium point that is the same distance from the fixed end of the spring and with a longer period, (d) oscillate about an equilibrium point that is closer to the fixed end of the spring and with a longer period.
13. A 50 g mass attached to a long spring is lifted 1.5 cm and dropped from rest. The resulting frequency is measured to be 1.25 Hz. The 50 g mass is then lifted 3.0 cm and dropped from rest. The resulting frequency is measured to be (a) 0.63, (b) 0.88, (c) 1.25, (d) 2.50 Hz.
14. A 0.5 kg mass oscillates about the equilibrium position on a vertical spring with spring constant 10 N/m. Where is its equilibrium position measured from the unstretched spring position (without the hanging mass)? (a) 0.05 m, (b) 0, (c) 0.49 m, (d) 5 m, (e) none of these.
15. A 10 N mass stretches a vertical spring by 10 cm. When set into oscillation, the time for the mass to travel from its highest to its lowest position is equal to (take $g = 10\ m/s^2$): (a) 0.31 s, (b) 0.63 s, (c) 0.99 s, (d) 1.99 s, (e) none of these.
16. The frequency of harmonic motion of a 1 kg mass attached to a simple spring is 1 Hz. The spring constant (a) is 1 N, (b) is 2π kg/m, (c) is $4\pi^2$ N/m, (d) cannot be determined from the information given.
17. Sedimentation of spheres of the same material but different radii in a liquid at low is a phenomenon where larger spheres beat smaller ones to the bottom of a container. This effect is due to the fact that (a) the larger spheres have smaller buoyant forces on them, (b) the pressure difference between the top and bottom of a larger sphere is greater than the pressure difference between the top and bottom of a smaller sphere, (c) larger spheres always beat smaller spheres because their gravitational acceleration is larger, (d) the terminal velocity of a sphere is proportional to its radius squared.
18. As a skydiver jumps out of an airplane, her
    (a) Vertical velocity decreases and vertical acceleration increases.
    (b) Vertical velocity decreases and vertical acceleration decreases.
    (c) Vertical velocity increases and vertical acceleration increases.
    (d) Vertical velocity increases and vertical acceleration decreases.
    (e) Vertical velocity increases and vertical acceleration remains constant.
19. Two cylindrical artificial bones are made of the same material and length, one with twice the radius as the other. When the two have the same tension force applied, the larger bone stretches by what factor compared to the smaller bone? (a) 2, (b) 0.25, (c) 0.5, (d) 4, (e) 1.
20. Given two rods made of the same material, one with twice the radius of the other and also with twice the length, if the same weight is suspended from each of the rods when held vertically, the longer rod will stretch (a) the same as, (b) twice, (c) half, (d) four times as much as the shorter rod.

PROBLEMS

1. A rural bus travels a straight line route of 20 km total distance. It makes a total of 5 stops along the route, each for exactly 2 min. If its average velocity in each driving interval is 45 km/h,
    (a) What is the total time for the round-trip route?
    (b) What is the average velocity for the one-way trip?
2. A ball is dropped from the Sears tower in Chicago with a height of 1454 ft (443 m). At what speed (in m/s and in mph) will it hit the ground, neglecting air resistance?
3. A truck travels on a straight road at 20 km/h for 60 km. It then continues in the same direction for another 50 km at 40 km/h. What is the average velocity of the truck during this 110 km trip?
4. The driver of a blue car, moving at a speed of 80 km/h, suddenly realizes that she is about to rear-end a red car, moving at a speed of 60 km/h. To avoid a collision, what is the maximum speed the blue car can have just as it reaches the red car?
5. A jumbo jet must reach a speed of 290 km/h on the runway for takeoff. What is the least constant acceleration needed for takeoff from a runway that is 3.30 km long?
6. In a car accident, a car initially traveling at 30 min/h (13.4 m/s) hits a tree and comes to rest in a distance

of 3 m. What was the deceleration of the car? How many *g*s is this?

7. The fastest sustained runner is the pronghorn antelope, capable of running at 55 min/h for 1/2 mile. How long does it take this antelope to run the 1/2 mile?

8. To bring your truck to rest, you first require a certain reaction time to begin braking; then the truck slows under the constant braking deceleration. Suppose that the total distance covered by your truck during these two phases is 39.7 m when the truck's initial velocity is 16.7 m/s, and 17 m when the truck's initial velocity is 10 m/s. What are your reaction time and deceleration of the truck?

9. At the instant a traffic light turns green, a car starts with a constant acceleration of 1.3 m/s^2. At the same instant a truck, traveling with a constant speed of 7.0 m/s, overtakes and passes the car.
(a) How far beyond the traffic signal will the car overtake the truck?
(b) What will the velocity of the car be at that instant?

10. Dropped from rest at the top of a 30 m tall building, a ball passes a window that is 1 m tall and has its lower ledge at a height of 8 m from the ground.
(a) How long will the ball take to pass by the window?
(b) What will be its speed when it reaches the bottom ledge of the window?

11. A 0.2 kg ball is thrown vertically downward at 8 m/s from the top of a 10 m tall cliff. (Neglect air resistance.)
(a) Find the velocity with which the ball hits the ground.
(b) How long does the ball take to hit the ground from the instant it is thrown?
(c) If the ball rebounds upward with a velocity of 10 m/s find the maximum height it will reach.

12. A ball is dropped from the top of a 45 m tall building. A second ball is thrown down after a 1 s pause. With what minimum initial speed should it be thrown to reach the ground first?

13. A rock is dropped from a cliff 60 m high (neglect air friction).
(a) How long does it take for the rock to hit the ground?
(b) Find the velocity and acceleration of the rock just before hitting the ground.

14. A ball, dropped from a cliff over the ocean, hits the water in 4.0 s.
(a) How high is the cliff?
(b) If a second ball is thrown from the same cliff and hits the water in 5.0 s, what was its initial velocity (magnitude and direction, please)?

15. An automobile driver traveling at 60 mph approaches a town that has a posted limit of 30 mph. Our driver dutifully applies the brakes, exactly 100 yards before the town limit, imparting a deceleration of −5 mph/s. Nonetheless, a police officer stops him. Our driver admits that he might have been going a bit fast outside of town but insists that he was always going at or below the town's speed limit while within its boundary. Is his claim correct?

16. A person throws a ball straight upward with an initial velocity of 15 m/s while standing on the edge of a cliff that is 100 m high. The ball rises to some height and then falls back down in such a way that it lands at the base of the cliff.
(a) Determine the time it takes for the ball to reach its maximum height and the maximum height above the cliff.
(b) How long does it take to reach the base of the cliff, and what is its velocity just before it strikes the ground?

17. A cartoon coyote comes up with a brilliant scheme to get lunch for himself by dropping a 500 kg boulder on a passing animated roadrunner. Unfortunately, when he cuts the rope holding the boulder in place, the rope becomes tangled around his ankle, and drags him toward the edge of the cliff. If the coyote's mass is 30 kg and his frantic clawing at the ground produces a force of 120 N resisting being dragged off the cliff, what is his acceleration toward the cliff?

18. In a device known as an Atwood machine, two masses (m_1 and m_2) are connected by a massless rope over a frictionless pulley.
(a) What is the acceleration of each mass if $m_1 = 10$ kg and $m_2 = 20$ kg?
(b) What is the tension in the cord?

19. A 5 kg block sits at rest on a frictionless horizontal surface.
(a) If a constant 15 N force pushes the block to the right, find the speed of the block after the force has been applied for 5 s.
(b) Suppose that in part (a) there is a constant frictional drag force of 5 N acting on the block when pushed by the same 15 N force. Draw a carefully labeled free-body diagram of the block, and find the acceleration of the block (magnitude and direction, please).
(c) Suppose a second block of mass 2 kg is placed on top of the 5 kg block in part (b) which is still being pushed by the 15 N force to the right and has the 5 N frictional drag force acting on it. Reconsider part (b) and find the net horizontal force (magnitude and direction, please) that must act on the 2 kg block in order for it to stay at rest on top of the 5 kg block. What is the origin of this force? (Hint: First consider the two blocks as one to find their acceleration.)

20. A microorganism is within a water droplet atop a microscope slide that measures 24 × 76 mm. The organism is swimming at 0.5 mm/s at precisely the middle of the slide and parallel to the slide's long axis, that is, parallel to its length. At that moment,

someone picks up one end of the slide and the tilt induces the water droplet to begin to move in the same direction in which the organism is swimming. If the water droplet picks up speed at 1 mm/s^2, how long is it until the organism goes over the edge of the slide?

21. A certain car ad once boasted of zero to 60 mph in 6 s, and 60 to zero in 3 s. What distances would be covered by this car during the respective positive and negative accelerations? Assume constant acceleration values for each case.

22. A zoo animal paces back and forth across the front of its cage a span of 8 m. A zoo attendant counts 1 min for a dozen round trips of the animal. Assuming that the creature spends as much time speeding up as slowing down and never travels at constant speed (i.e., it speeds up to the middle of the cage, whereupon it begins to slow down), how fast is the animal moving right at the middle of the cage? Assume, of course, that both accelerations are constant.

23. Fleas are notorious jumpers, reaching heights of nearly 20 cm, roughly 130 times their own height. Assuming that the flea acquires its initial velocity in leaving the ground over a distance of half its height, find the average acceleration the flea must have to reach a height of 20 cm. Express your answer as a multiple of g.

24. Common terns hover in a stationary position over the ocean watching for a tasty fish. When they see one, they immediately stop their wings and simply free-fall into the ocean to catch the fish. Calculate how long a fish near the surface has to move away after the instant a tern sees it from a height of 3 m above the surface.

25. Repeat Problem 2 above, but now include air resistance. Assume a ball of 3 cm radius with an average density of 4400 kg/m^3, a density of air of 1.3 kg/m^3, and a value of C = 1.

26. Estimate the terminal velocity of a skydiver with a closed parachute. Take values from the previous problem and assume the diver has a mass of 75 kg and an effective cross-sectional area of 0.4 m^2. If the terminal speed with an open parachute is 18 km/h, find the effective area of the parachute. The buoyant force is negligible.

27. Block #1 is attached to a horizontal spring and slides on a frictionless horizontal surface. Block #2 has the same mass as #1 and also sits on the same frictionless surface. It is attached to a spring with three times the stiffness of the other one. If both blocks have the same amplitude of motion find the ratio of the following quantities (#2/#1): the periods of the motion, the angular frequencies, the maximum velocities, the maximum accelerations, and the maximum displacements.

28. Attached to a spring on a frictionless table top, a 1 kg mass is observed to undergo horizontal simple harmonic motion with a period of 2.5 s after stretching the spring. The spring is then held vertically and a 0.2 kg mass is attached and gently lowered to its equilibrium position.
(a) Find the distance the spring is stretched.
(b) If the spring is then stretched an additional 5 cm and released, find the period of the subsequent motion.
(c) What is the maximum acceleration of the 0.2 kg mass?
(d) What is its maximum velocity?

29. A 0.8 kg mass attached to a vertical spring undergoes simple harmonic motion with a frequency of 0.5 Hz.
(a) What is the period of the motion and the spring constant?
(b) If the amplitude of oscillation is 10 cm and the mass starts at its lowest point at time zero, write the equation describing the displacement of the mass as a function of time and find the position of the mass at 1, 2, 1.5 s, and at 1.25 s.
(c) Write the equation for the speed of the mass as a function of time and find its speed at the times given in part (b)? (Be careful to check that you have the correct starting speed at time 0.)

30. Find the natural frequency of vibration of the salt molecule NaCl given its effective mass of 13.9 atomic mass units and a spring constant of 100 N/m.

31. In the dangerous sport of bungee-jumping, a thrill-seeker jumps from a great height with an elastic cord attached to the jumper's ankles. Consider a 70 kg jumper leaping from a bridge 226 m high. Suppose further, that instead of using a specifically designed cord, the jumper uses a 9.00 mm diameter nylon mountain climber's rope with an effective force constant k = 4900 N/m.
(a) What is the length of rope needed to stop the jumper 10 m above the ground?
(b) What is the maximum force that the rope will exert on the daredevil?
(c) Expressing this maximum force in terms of the weight of the jumper, did the jumper make a wise choice to use the mountain climber's rope?

32. A 70 kg daredevil stretches a steel cable between two poles 20 m apart. He then walks along the cable, loses his balance, and falls where he luckily lands in a safety net, which acts like a spring with spring constant k = 1750 N/m. If his speed when he strikes the net is 10 m/s, what is the amplitude of the oscillation as he bounces up and down?

33. A 2 kg mass attached to a vertically held spring is observed to oscillate with a period of 1.5 s.
(a) Find the spring constant.
(b) If the amplitude of the oscillation is 10 cm, find the magnitude of the maximum acceleration of the mass and state where in the oscillation of the mass this maximum acceleration occurs.

(c) If the hanging mass is doubled and the amplitude is halved, find the magnitude of the maximum velocity of the new mass on the same spring and state where in the oscillation of the mass this maximum velocity occurs.

34. A 0.5 kg mass is attached to a spring with a spring constant of 8.0 N/m and vibrates with an amplitude of 10 cm.
(a) What are the maximum values for the magnitudes of the speed and of the acceleration?
(b) What are the speed and the acceleration when the mass is 6 cm from the equilibrium position?
(c) What is the time it takes the mass to move from $x = 0$ to 8 cm?
(d) What is the period of the motion?
(e) What are the displacement, velocity, and acceleration as functions of time?

35. In an experiment to investigate Hooke's law with springs, weights are hung on a spring; the spring stretches to different lengths as shown in the table below.
(a) Make a graph of the applied force versus the stretch of the spring and if the data are linear obtain the slope of the best fit line. What does this slope represent?
(b) If the spring is stretched 102 cm, what force does the spring exert on the suspended weight?

F (N)	2	4	6	8	10	12	14	16	18
x (mm)	15	32	49	64	79	98	112	126	149

36. A rod-shaped bacterium (with an equivalent spherical radius of 0.5 μm) rotates its flagella at 100 revolutions per second to propel itself at a uniform velocity of 100 μm/s. Calculate the thrust (propulsive force) generated by the flagella, assuming the only other force is a frictional one given by Stokes' law. Note that this speed is extremely fast, namely about 50 body lengths per second. Show that the equivalent speed for a human would be about 200 mph. Take $\eta = 10^{-3}$ in SI units, the value for water.

37. A 10 g inflated balloon falls at a constant velocity. What is the buoyant force acting on the balloon? (The frictional force can be neglected here.)

38. A 75 kg person falls from the second floor of a building and lands directly on one knee with his body otherwise vertical.
(a) If the fall is from a height of 10 m, find the velocity on impact with the ground.
(b) If it takes 5 ms for the person to come to rest, find the average force acting during the collision.
(c) Using the data of Example 3.9, will the femur break?

39. Bone has a larger Young's modulus for stretch (1.6×10^{10} N/m^2) than for compression (0.94×10^{10} N/m^2). By how much is each femur, or thigh bone, of the legs compressed when a weightlifter lifts 2200 N? Take the dimensions of the femur to be 0.6 m long and have an average radius of 0.01 m.

40. A medieval knight is "racked," stretching his body with a force of 1200 N. Using the data in the previous problem, by how much will the knight's femur bones be stretched?

41. Four concrete columns, each 50 cm in diameter and 3 m tall, support a total weight of 5×10^4 N. Find the distance that each column has been compressed by the weight of the load. (Use an elastic modulus of 20×10^9 N/m^2 for concrete). Find the effective spring constant for a column and then find the period of small amplitude oscillations assuming an effective spring constant equal to the sum of the values for the four columns, and neglecting the weight of the columns. We show later that such natural oscillations at the corresponding frequency make such structures susceptible to absorbing energy from external sources (such as wind, earthquakes, etc.) leading to larger amplitude vibrations and possible damage.

42. Steel pillars support a pier extending out into the ocean from the beach. If the pillars are solid 10×10 cm steel (Young's modulus $= 2.0 \times 10^{11}$ N/m^2) and are 4 m long, find the distance each is compressed if each pillar supports a weight of 2000 N.

43. A guitar is being restrung with a string having a diameter of 1.4 mm and a length of 0.82 m when no tension is applied. If the string has a Young's modulus of 1.4×10^{11} N/m^2 and is tightened by wrapping it three times around a peg with a 2.5 mm diameter, find the tension in the string.

44. Fill in the steps to derive Equations (3.27), (3.28), and (3.29) in Section 5 of the chapter.

4

Work and Energy in One Dimension

In this chapter we introduce work, kinetic energy, the energy associated with motion, and provide a general framework for appreciating the concept of energy and its usefulness in all areas of science. We present these ideas for one-dimensional motion, the theme of the previous two chapters, leaving the generalization to more than one dimension for the next chapter. A major goal of this chapter is to appreciate the extremely important and general conservation of energy principle. It is used again and again in future discussions of various other forms of energy, including electrical, magnetic, and eventually their synthesis in electromagnetic energy, as well as various types of chemical and nuclear energy. In addition, later we study the science of thermodynamics dealing with energy and its flow in bulk matter. The conservation of energy principle is perhaps the most important and fundamental principle of all science.

Our discussion of forces and the laws of motion thus far is entirely sufficient to be able to describe the motion of most inanimate objects: planets, moons, and satellites, or projectiles, and sliding and rolling objects (with some additional ideas needed here). In fact with some added mathematics, only the generalization of these laws to three dimensions and a knowledge of forces is needed, no matter how complex and interesting the motion may be. A simple example illustrates, however, that for living organisms force alone will not provide a sufficient framework to understand their behavior. When you lift a heavy weight and hold it in the air you get tired even though you are not doing any work (we show that doing work, as defined in physics, requires a displacement). This simple observation implies that another concept, the source of forces, is needed to understand living organisms as well as some dynamic inanimate systems. Your muscles require energy to function and provide a force. We need to develop an appreciation of energy as the source of force and here we begin this development.

1. WORK

When a constant net force F acts on an object of mass m originally at rest, the object experiences an acceleration F/m, and its velocity increases. The longer the net force acts, and correspondingly the greater the distance it acts over, the faster the object is made to move. From our knowledge of Newton's laws and kinematics, we can calculate the velocity of the object as a function of time to be

$$v = \frac{F}{m}t, \tag{4.1}$$

or as a function of the distance the object travels x, we can calculate the velocity to be

$$v = \sqrt{2ax} = \sqrt{2\frac{F}{m}x}. \tag{4.2}$$

J. Newman, *Physics of the Life Sciences*, DOI: 10.1007/978-0-387-77259-2_4,

In this and the next section we learn a different way of describing what has occurred in this example. In words, we say that the net force has done work on the object and in doing so has increased the energy of motion, or kinetic energy, of the object. Let's first carefully define work and kinetic energy and then derive a theorem that is very general indeed and is the motivation for this alternative description.

We develop the definition of work in this chapter with the case of one-dimensional motion in which a constant force F acts on an object, originally at rest, along the x-axis. The work done on an object by the constant force when the object has undergone a displacement Δx is defined to be

$$W_F = F\Delta x. \qquad \text{(constant force along } x \text{ direction).} \qquad (4.3)$$

Suppose our object is a sled being pulled by a rope along a horizontal surface. If the rope is held horizontally then the work done by a tension force of $T = 20$ N along the rope in pulling the sled a distance $L = 5$ m is given by Equation (4.3) as $W = TL = (20\text{ N})(5\text{ m}) = 100$ N-m. The SI unit for work is the N-m which is called the *joule* (J; 1 N-m = 1 J).

Example 4.1 A group of campers is having a tug of war in which five of them pull on a heavy rope toward the left and five others pull toward the right. Suppose that each camper on the left pulls toward the left with an average force of 220 N and each of the campers on the right pulls with an average force of only 210 N. During the time when the rope moves a distance of 3 m to the left, how much work does each camper do and what is the net work done by all ten of them?

Solution: Each camper on the left does an amount of work equal to (220 N)(3 m) = 660 J, whereas each camper on the right does an amount of work equal to $-(210\text{ N})(3\text{ m}) = -630$ J. Note that this work is negative because the campers on the right, while pulling to the right, have displacements to the left. The net amount of work done by all is then $W = 5(660) - 5(630) = 150$ J. Clearly this could be found as well by computing the net force on the rope (=50 N) and multiplying it by the displacement.

The above definition of work is in conflict with our colloquial usage of the word work. If the campers on the right had pulled a bit harder in the example, the rope might have not moved at all and no work would have been done, despite a great deal of effort exerted by all. While a hiker carrying a heavy backpack is standing still she does no work, although we would commonly say that she is doing work, using up energy, and will get tired even standing in place. Indeed extra energy is being used to support the weight of the backpack, but the only work done is internal work within the muscles of the hiker. Without any displacement of the backpack or any displacement of the tug-of-war rope, no work is done according to our definition (Figure 4.1). This example shows that some care is needed in calculating the work done by a force.

FIGURE 4.1 *A hiker does no work in supporting a backpack.*

The above definition and discussion are fine as long as the forces acting on the object are constant, but we have already seen two examples of forces that are not constant and for which Equation (4.3) does not apply. The frictional force in a fluid is dependent on the velocity and changes as the object accelerates, whereas the spring force changes continually in magnitude and periodically in direction as well. In order to modify Equation (4.3) to be able to calculate the work done by a variable force, we must use a "divide and conquer" strategy. From a graph of F versus x, we divide the region of interest along the x-axis of width Δx into small

displacement intervals, each of width δx as shown in Figure 4.2. In each of the intervals we replace the varying force with its average value and calculate the work for that displacement interval using Equation (4.3), so that the contribution to the work from that small displacement interval δx is

$$\Delta W = F_{\text{ave}}\,\delta x. \tag{4.4}$$

As can be seen in Figure 4.2, ΔW represents the area contained in the rectangle with height F_{ave} and width δx; this area is also nearly equal to the actual area under the curve representing F for that interval of δx and becomes more closely equal to the actual area as the width of the interval δx gets smaller and the number of such intervals grows. These contributions to the work from a total displacement of Δx add up to the total work given by

$$W_F = \Sigma \Delta W = \Sigma F_{\text{ave}} \int \delta x, \qquad \text{(force along } x\text{-direction)}, \tag{4.5}$$

where the sums are over each of the intervals. Thus, the graphical interpretation of the work done in a displacement Δx is the area under the curve representing F versus x and bounded by two vertical lines at the beginning and end of the displacement interval.

In cases where the curve representing the force as a function of distance is actually either a straight line or a simple curve, it may be easy to calculate the area directly. For example, in the case of a spring force, $F = -kx$, the graph is linear (Figure 4.3) and the area under the line can be directly calculated as in the following example.

Example 4.2 Using Figure 4.3, calculate the work done in stretching a spring from x_1 to x_2.

Solution: To stretch the spring we can use an external force equal and opposite to the spring force, given itself by Hooke's law as $F_{\text{spring}} = -kx$. The work done by the external force will be positive because the force and displacement are in the same direction, whereas the work done by the spring will be equal in magnitude but negative. The area between the diagonal line in Figure 4.3 representing $F_{\text{ext}} = kx$ and the x-axis in the figure is equal to the work done by the external force. We can calculate this simply by finding the area of the large triangle with apex at the origin and base extending to x_2 and subtracting the area of the smaller triangle at the apex with base reaching x_1. The area of a triangle is given by 1/2 base $\times$ height, so we have only to take half of the product of the base (x_2 or x_1) times the height (kx_2 or kx_1) to obtain a net work of

$$W_{\text{ext}} = \frac{1}{2}k(x_2^2 - x_1^2).$$

Note that the work done by the spring is just the negative of this

$$W_{\text{spring}} = -\frac{1}{2}k(x_2^2 - x_1^2). \tag{4.6}$$

(*Continued*)

As readers who have had some calculus and have seen some integration should recognize, the discussion leading up to the general definition of work in one dimension, Equation (4.5), is a prelude to defining work as an integral. All that is missing is taking the usual limit as the size of the intervals δx approach zero resulting in the following integral for the work done by the force F directed along the displacement,

$$W_F = \int F\,\mathrm{d}x.$$

The graphical interpretation of this integral is, in fact, the area under the curve represented by the integrand F when plotted against x between the limits of integration, namely the displacement interval, as shown in Figure 4.2. As an example application of using this general definition for work, we calculate the work done by the spring force on an attached mass, $F = -kx$, as the spring changes its position from x_1 to x_2. We find

$$\begin{aligned} W_{\text{spring}} &= \int_{x_1}^{x_2} (-kx)\,\mathrm{d}x \\ &= -\frac{1}{2}k(x_2^2 - x_1^2), \end{aligned}$$

as found in Example 4.2. With more complicated forces, the method of Example 4.2 does not work and integration must be used.

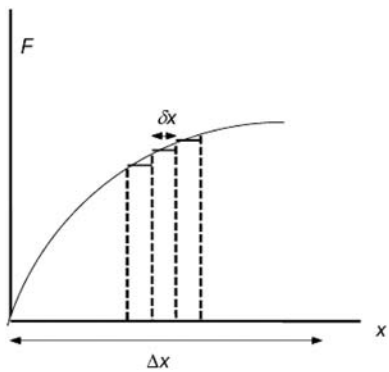


FIGURE 4.2 *Divide-and-conquer strategy for calculating the work done by a varying force.*

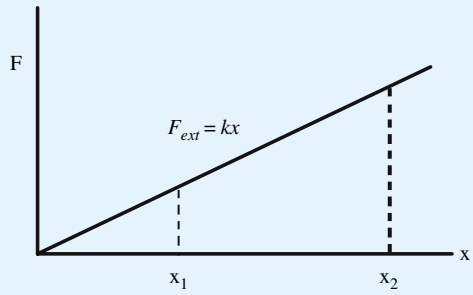


FIGURE 4.3 *External force stretching a linear spring versus displacement.*

2. KINETIC ENERGY AND THE WORK–ENERGY THEOREM

At the beginning of the last section we used Newton's laws and kinematics to analyze the motion of an object with a constant net force acting on it in order to find the velocity of the object as both a function of time and of its position. In this section we reconsider that problem using our knowledge of work. Recall that we were considering an object that experienced a constant net force, F_{net}, acting along the x-axis. Let the object of mass m have a velocity of v_1 when it is located at position x_1 and move, under the influence of F_{net}, to position x_2, where it has a velocity v_2. Then we have, because the acceleration $a = F_{\text{net}}/m$ = constant, from one of the kinematic relations valid for constant acceleration,

$$v_2^2 = v_1^2 + 2\left(\frac{F_{\text{net}}}{m}\right)(x_2 - x_1). \tag{4.7}$$

We can also calculate the work done by the constant net force to be

$$W_{\text{net}} = F_{\text{net}}\,(x_2 - x_1). \tag{4.8}$$

Substituting for $F_{\text{net}}(x_2 - x_1)$ from Equation (4.8) into Equation (4.7), and solving for W, we have

$$W_{\text{net}} = \frac{1}{2}\,mv_2^2 - \frac{1}{2}\,mv_1^2. \tag{4.9}$$

The expression $1/2\ mv^2$ is defined as the translational *kinetic energy* KE of the mass

$$\text{KE} = \frac{1}{2}\,mv^2. \tag{4.10}$$

Kinetic energy is also measured in joules, where 1 J equals 1 kg-m^2/s^2. You can "feel" 1 J if you drop a 1 kg mass 10 cm onto your outstretched palm. The stinging sensation that results is "equivalent to" about 1 J.

Example 4.3 What is the kinetic energy of a 1 ton car traveling at 75 miles/h?

Solution: A ton is a weight of 2000 pounds. One kilogram weighs 2.2 pounds so 1 ton equals 2000 pounds/2.2 pounds/kg = 910 kg. One mile is about 1600 m, so 75 miles/h is about 1.2×10^5 m/h. One hour is 3600 s, so 75 miles/h = $1.2 \times$

10^5 m/h $\times$ 1/3600 h/s = 33 m/s. Then the kinetic energy of the car is (1/2)(910 kg)(33 m/s)^2 = 5 $\times$ 10^5 J. If 1 J produces a sting, imagine the feeling you would experience if 500,000 J were deposited on you.

Finally, we can rewrite Equation (4.9) in terms of kinetic energy as

$$W_{\text{net}} = \text{KE}_2 - \text{KE}_1 = \Delta\text{KE}. \tag{4.11}$$

Equation (4.11) is known as the *work–energy theorem*. It states that the net work done on an object is equal to the change in its kinetic energy. If the net work done on the object is positive, its kinetic energy will increase, whereas if the net work done is negative, the object's kinetic energy will decrease.

What is the distinction between kinetic energy and work? Clearly, from Equation (4.11), they are both measured in the same units, joules. Furthermore, these two quantities can exchange back and forth; work done on an object can change its kinetic energy by either speeding it up or slowing it down, and the kinetic energy of an object can also be used to do work on another object with which it interacts. In the next section we introduce other forms of energy, associated with an object's position due to interactions with other objects, that can also be used to perform work and can also be changed by performing work. Thus, we can think of energy, in general, as the ability to do work, the energy itself being stored either in the motion or the external interactions of the object.

Example 4.4 Using the work–KE theorem, estimate the height to which a person can jump from rest. Make some reasonable assumptions as needed.

Solution: Once a person leaves the ground, he is completely governed by free-fall. Therefore, the key to a good standing high jump is to attain the fastest initial vertical velocity on leaving the ground. This initial velocity is governed by the acceleration obtained as the legs are stretched and push against the ground (Figure 4.4). Putting these ideas together, and assuming that a constant net upward force F is exerted on the person during the contact portion of the jump (the force on the person from the ground is actually $mg + F$; why?), we can write that

$$Fy = \frac{1}{2}mv_0^2,$$

where y is the distance over which the force F acts (the distance from a crouched to extended leg position), m is the mass of the person, and v_0 is the initial velocity on leaving the ground. Here we've assumed that the starting KE is zero when in a crouched position and $1/2mv_0^2$ is the KE when just leaving the ground. If the upward force from the ground varies, then think of F as its average value and

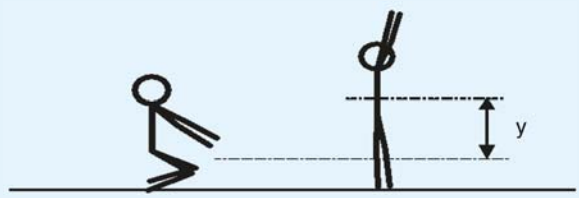


FIGURE 4.4 *Standing high jump showing the upward acceleration phase.*

(*Continued*)

FIGURE 4.5 *In a good high jump, the person's center of mass actually goes under the bar.*

everything else follows correctly. The height h that a person can jump is then given from the kinematic relation that $v^2 = v_0^2 - 2gh$ (=0 at the highest point), so that

$$h = \frac{v_0^2}{2g}.$$

Substituting for v_0 from the work–KE expression, we find that

$$h = \frac{Fy}{mg}.$$

The distance y can be estimated to be at most about 1/3 the height of a person (from a deep crouching position to full extension). Therefore, the maximum height a person can jump is limited by the force that he can exert. We can estimate this to be about the weight of the person, so that $h \approx y$, implying that a person can raise his center of mass about 1/3 of his height. For a 6 foot tall person with center of mass 3 feet above the ground, the center of mass can be raised to about 5 feet. Based on this analysis by swinging arms and legs, this person is limited to a standing high jump of about 5 feet. Modern running high-jumpers can achieve much higher jumps because they are both running and also able to arch their bodies over the bar while their center of mass, a sort of average coordinate that we study in Chapter 6, actually goes below the high bar (see Figure 4.5).

3. POTENTIAL ENERGY AND THE CONSERVATION OF ENERGY

Just as the energy associated with an object's motion can be used to do work, so too can the energy of interaction of an object with other objects by virtue of its location. This type of energy is known as *potential energy*. There are many types of potential energies, each due to a specific type of position-dependent interaction energy. In this section we learn about gravitational potential energy, due to the gravitational interaction between an object and the Earth, and about elastic potential energy (potential energy of a spring), due to the Hookean forces within an object that are ultimately related to internal molecular interactions. In the course of this book we show other forms of potential energy including thermal, electric, magnetic, chemical, and nuclear. We show that, within an isolated system, although energy can be converted from one of these forms to another, the total energy of the system remains constant.

Consider a crate of mass m resting on the edge of a table, a height h above the floor. If the crate falls from the table, gravity will do work on the crate increasing its kinetic energy. After falling to the floor, the work done by gravity will be

$$W_{\text{grav}} = mgh. \qquad \text{(falling through height } h\text{)}. \tag{4.12}$$

Applying the work–KE theorem, we could calculate the kinetic energy of the crate just before hitting the floor as KE = mgh. Of course in the next instant the crate hits the floor and there are very strong upward forces that act to quickly stop the crate, doing negative work on the crate so that its ultimate kinetic energy is zero.

To then lift the crate back up and place it on the table again requires positive work to be done by an outside force. During the lifting of the crate, both gravity and the external lifting force act. One way to lift the crate is to very slowly raise it at constant velocity with an equal and opposite force to its weight (as allowed by Newton's first law), starting and stopping with just a slight extra appropriate nudge. In this case the work done by the outside force ($W_{\text{ext}} = mgh$) and the work done by gravity ($W_{\text{grav}} = -mgh$, negative because of the opposite directions of the downward force of gravity and the upward displacement) are just equal and opposite, so that the net work is zero. This makes sense because the starting and ending kinetic energies are both zero, so that there is also no change in kinetic energy. The work–KE theorem then says that the net work done must be zero.

In fact, regardless of the manner in which the outside force is applied and regardless of the path of the crate in reaching the tabletop, the net amount of work done must be zero because there is no change in kinetic energy. To lift the crate the outside force must be at least equal to mg. If the outside force is greater than mg, there will be a net upward force that will accelerate the crate upward. In order to have the crate end up at rest on the table, the outside force must then be less than mg for some portion of the trip so that during this time the net force is downward and the crate is slowed down. In any case, because the kinetic energy change is zero, the net work done by the two applied forces must add to zero and so the work done by the external force to lift the crate back up on the table must always be $W_{\text{ext}} = -W_{mg} = mgh$, the same as in Equation (4.12).

Example 4.5 Suppose that a 3 kg package is lifted vertically from the ground and tossed onto a counter 2 m off the ground (Figure 4.6). Imagine that for the first meter a force equal to twice the weight of the package is exerted, and then the person lets go of the package tossing it up to just reach the counter. Find the work done on the package by the person and by gravity and find the maximum speed of the package.

Solution: The work done by the person is simply the product of the force, 2 mg = 2(3)(9.8) N, and the distance of 1 m over which the force acts. We find that W_F = 59 J. Similarly, the work done by the gravitational force is the product of mg and the net displacement, 2 m, with a minus sign inserted because the weight and displacement are oppositely directed. We have that $W_{\text{grav}} = -59$ J. What is the significance, if any, of that fact that these are equal in magnitude? If the same force were to be exerted by the person over a shorter distance, doing less net work, the package would not reach the counter height. On the other hand, if a larger upward

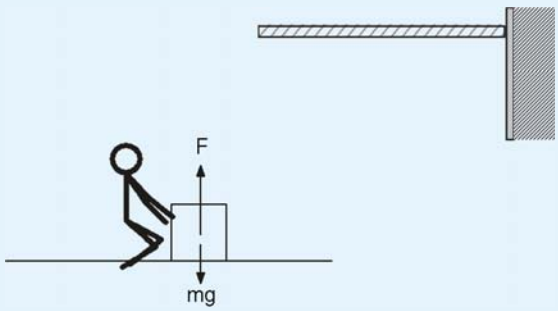


FIGURE 4.6 *Lifting a heavy package to then toss it up to a shelf.*

(*Continued*)

force were exerted, then the package would rise above the counter level and fall back down, arriving on the counter with some net speed. Our particular conditions have the package just reaching the counter. To find the maximum speed of the package, we first note that this must occur just when the package is released (why?). We can find this speed by using the work–KE theorem, noting that the net work done in the first 1 m is $W_F + W_{\text{grav}} = 59 - 59/2 = 30$ J, because half the work of gravity is done in that 1 m. Equating this work with the change in kinetic energy from zero (the package is assumed to start at rest on the ground), we have

$$W_{\text{net}} = 30\,\text{J} = \frac{1}{2}mv^2 = \frac{1}{2}(3)v^2,$$

so that the maximum speed is

$$v = \sqrt{\frac{2(30)}{3}} = 4.5 \text{ m/s}.$$

We define the *gravitational potential energy* at height y, relative to some reference level ($y = 0$) to be

$$\text{PE}_{\text{grav}} = mgy. \tag{4.13}$$

When an object changes its height from y_1 to y_2 in the presence of gravity, there is a corresponding change in PE_{grav}, where $\Delta\text{PE}_{\text{grav}} = \text{PE}_{\text{grav, final}} - \text{PE}_{\text{grav, initial}} = mg(y_2 - y_1)$, equal in magnitude to the work done by gravity. As we have just seen, when $(y_2 - y_1) > 0$, corresponding to an increase in height, the work done by gravity is negative whereas the $\Delta\text{PE}_{\text{grav}}$ is positive; similarly when $(y_2 - y_1) < 0$, corresponding to a decrease in height, the work done by gravity is positive and the $\Delta\text{PE}_{\text{grav}}$ is negative. Thus we can write

$$W_{\text{grav}} = -\Delta\text{PE}_{\text{grav}}, \tag{4.14}$$

which states that the work done by gravity is equal to the negative of the change in gravitational potential energy.

If gravity is the only force acting, starting with the work–energy theorem, Equation (4.11), we can substitute Equation (4.14) for the work to find

$$\Delta\text{KE} = \text{KE}_2 - \text{KE}_1 = -\Delta\text{PE}_{\text{grav}} = -(\text{PE}_{\text{grav}_2} - \text{PE}_{\text{grav}_1}), \tag{4.15}$$

or, rearranging Equation (4.15), we find

$$(\text{KE} + \text{PE}_{\text{grav}})_1 = (\text{KE} + \text{PE}_{\text{grav}})_2. \tag{4.16}$$

Each side of this equation represents the total mechanical energy, $E = \text{KE} + \text{PE}_{\text{grav}}$, of the object at a fixed position. The positions 1 and 2 are completely arbitrary, therefore we can conclude that

Mechanical energy remains a constant of the motion,

$$E = KE + PE_{\text{grav}} = \textit{constant}. \tag{4.17}$$

This is the principle of conservation of mechanical energy.

As we have seen, the KE and PE individually are not conserved but may transform from one to the other; however, the sum of the kinetic and gravitational potential energies remains constant at the value of the total mechanical energy.

The choice of reference point for gravitational potential energy is totally arbitrary; only differences in potential energy matter in Equation (4.16), as is readily seen in the form of Equation (4.15). When the total mechanical energy is given, however, as in Equation (4.17), its value implicitly depends on a reference position for potential energy.

In Example 4.2 we found that the work done by a spring, with a spring constant k, in stretching from x_1 to x_2 is given by

$$W_{\text{spring}} = -\frac{1}{2}k(x_2^2 - x_1^2). \tag{4.18}$$

In a similar manner to the gravitational case, we introduce the spring potential energy function as the negative of the corresponding work,

$$\text{PE}_{\text{spring}} = \frac{1}{2}kx^2. \tag{4.19}$$

If a mass m is attached to the end of the spring, then following a similar procedure as that used to get Equation (4.16), we find that if the spring force is the only force acting (suppose the spring and the motion of the mass are horizontal so that gravity can be ignored)

$$(\text{KE} + \text{PE}_{\text{spring}})_1 = (\text{KE} + \text{PE}_{\text{spring}})_2. \tag{4.20}$$

We see that in the work–energy theorem, the work done by each force that can be associated with a potential energy can be replaced by the negative of its potential energy change. Generalizing this result, we can write that the total mechanical energy, defined as the sum of the kinetic and all potential energies (gravitational, spring, and any others), will be a constant of the motion if all the forces acting can be associated with a potential energy

$$\text{E} = \text{KE} + \text{PE}_{\text{grav}} + \text{PE}_{\text{spring}} + \text{PE}_{\text{other}} = \text{constant}. \tag{4.21}$$

Later in Chapter 15 we add electrical potential energy to our list and in Chapter 17 we add a magnetic energy term as well. We also show in Chapter 5 that the frictional force cannot be associated with a potential energy and that when friction acts within a system, there is always a loss of mechanical energy to thermal energy.

Example 4.6 A spring is held vertically and a 0.1 kg mass is placed on it, compressing it by 4 cm. The mass is then pulled down a further 5 cm and released giving it an initial velocity of 1 m/s downward. Find the maximum compression of the spring relative to its unstretched length. What is the maximum velocity of the mass and where does it occur? What is its maximum acceleration and where does it happen?

Solution: Refer back to Example 3.6 for a somewhat simpler related problem solved using force considerations only. We first find the spring constant by noting that the 0.1 kg mass compresses the spring by 0.04 m at which point it is in equilibrium with its weight balanced by the upward spring force. This means that $mg = kx_0$, so that

$$k = \frac{mg}{x_0} = \frac{0.1 \times 9.8}{0.04} = 25 \text{ N/m}.$$

This initial compression of the spring balances the weight of the mass and for the subsequent motion we can ignore the gravitational potential energy changes. Once

(*Continued*)

the mass is pushed down an additional distance y_0 and given an initial velocity v_0, we can write down the initial energy relative to the equilibrium position as

$$E = \frac{1}{2}ky_0^2 + \frac{1}{2}mv_0^2,$$

where y_0 is the initial displacement from the equilibrium position, itself 4 cm below the origin, as shown in Figure 4.7. Even though the height of the mass changes as it moves, we still do not include the gravitational potential energy because the weight of the mass has been removed from the problem by measuring displacements from the equilibrium position (see just below).

Mechanical energy is conserved therefore the spring will have an equal energy at all points in its motion, and in particular at its amplitude A, at which point its kinetic energy will vanish. At that point we can write the total energy as

$$E = \frac{1}{2}kA^2 = \frac{1}{2}ky_0^2 + \frac{1}{2}mv_0^2.$$

Solving for A, we find

$$A = \sqrt{y_0^2 + \frac{mv_0^2}{k}} = \sqrt{0.05^2 + \frac{0.1 \cdot 1^2}{25}} = 0.08 \text{ m}.$$

The maximum compression of the spring is then the initial 4 cm and an additional 8 cm, for a total of 12 cm.

Alternatively, we could refer the potential energy to the point $x = 0$ in which case we would write that the total energy is given by $E = \frac{1}{2}kx^2 - mgx + \frac{1}{2}mv^2$, including gravitational PE as well, and then set its initial value equal to its value at the amplitude where there is no KE, but both forms of PE. This can be solved for the amplitude as well, but the mathematics involves solving a quadratic equation and is omitted here. The result in this case is found directly to be 12 cm from the origin, in agreement with the calculation above. You should verify this.

As the spring relaxes and the mass rises, its maximum speed will occur at the equilibrium position where all of the spring's potential energy is converted to kinetic energy. We can find this speed by writing

$$E = \frac{1}{2}mv_{\text{max}}^2 = \frac{1}{2}kA^2,$$

so that using our amplitude, we find

$$v_{\text{max}} = \sqrt{\frac{k}{m}}A = 1.3 \text{ m/s}.$$

Because the mass is not attached to the spring it will actually fly off the spring on its way up as the spring decelerates; if it were attached to the spring it would continue to oscillate. The maximum acceleration occurs at the initial amplitude position where the spring force is greatest and has a magnitude, from Hooke's law, of

$$a_{\text{max}} = \frac{kA}{m} = 20 \text{ m/s}^2.$$

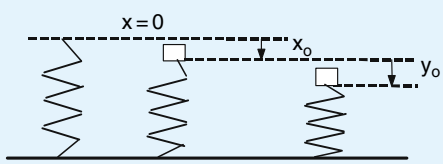


FIGURE 4.7 *Spring arrangement for Example 4.6.*

Before we leave this section dealing with conservation of energy, let's consider two biological energy aspects: energy considerations from the perspective of the Earth and from that of a single biological cell.

The ultimate energy source for life on Earth is the sun, delivering about 5×10^{24} J/year with about half of this getting absorbed by the surface of the Earth. Estimates of the total fraction of this energy actually captured by photosynthetic plants, both terrestrial and marine, are about 0.1%. Recent estimates of human energy consumption give a rate of about 5×10^{20} J/year (with nearly 90% coming from fossil fuels), which amounts to about 1/10 of the energy captured by plants on the Earth. Reserves of fossil fuels on the Earth are estimated to be about 4×10^{23} J, with an additional 2.5×10^{24} J in radioactive nuclear fuels. Although human consumption appears to be only a small fraction of the energy available, it is becoming increasingly clear that the persistent use of fossil fuels is having an effect of the fraction of the solar energy that is trapped within the Earth's atmosphere, causing a global warming. We return to a discussion of this "greenhouse effect" at the end of Chapter 13.

Energy considerations in biological cells are centered around the ATP (adenosine triphosphate) molecule. ATP stores chemical energy from the oxidation of foodstuffs (small sugar molecules) that themselves were ultimately produced using solar energy whether they originated from plants or animals. This formation of ATP from ADP (adenosine diphosphate) and inorganic phosphate occurs in a series of highly efficient coupled reactions catalyzed by the enzyme ATP synthase (F1-ATPase), a very interesting molecule further discussed in Section 3 of Chapter 7. The high-energy phosphate bond, with an energy roughly twice that of a hydrogen bond, is the source of most of the cellular energy, and therefore, of the energy used by the human body. Each of us uses between about 50 and 75 kg of ATP each day, approximately the weight of a person. When exercising strenuously, the rate of usage can approach 0.5 kg/min. Clearly our bodies do not contain that much ATP. It is constantly synthesized with each F1-ATPase molecule capable of generating about 300 ATP molecules per second. Each ATP molecule in the human body is recycled over 1000 times per day in order to generate sufficient energy to sustain life.

4. FORCES FROM ENERGY

At the beginning of this chapter we pointed out that many situations can be analyzed using energy concepts as well as force concepts. Are there advantages of introducing these new ideas on energy? There is a clear need for energy concepts to understand the production of forces in living or inanimate dynamical systems that generate mechanical energy from chemical or other energy forms. These notions are developed over the course of this book in various ways as we learn more physics. At this point, we have seen how to generate a potential energy function from knowledge of the forces acting on an object. The reverse is also true; it is also possible to find the forces acting on an object from knowledge of the potential energy function. As we have seen, energy is a scalar quantity, whereas force is a vector quantity, in general having x-, y-, and z-components as we study in Chapter 5, and so it is often easier to deal with energy first and then, if needed, to calculate the forces involved from the potential energy function. In this section we learn how this can be done.

We have seen in Equation (4.14) that the work done by gravity can be expressed as a change in a gravitational potential energy function. When forces other than gravity are present, often other potential energy functions can be defined as functions of displacement, similar to Equation (4.13), as, for example, we have seen for springs with Equation (4.19). Forces for which this can be done are called *conservative forces* and are characterized by the fact that the work they do when acting on an object only depends on the displacement of the object and not on its actual path, trajectory, or velocity. Generalizing Equation (4.14) to any conservative force

$$W = F_x \Delta x = -\Delta \text{PE}, \tag{4.22}$$

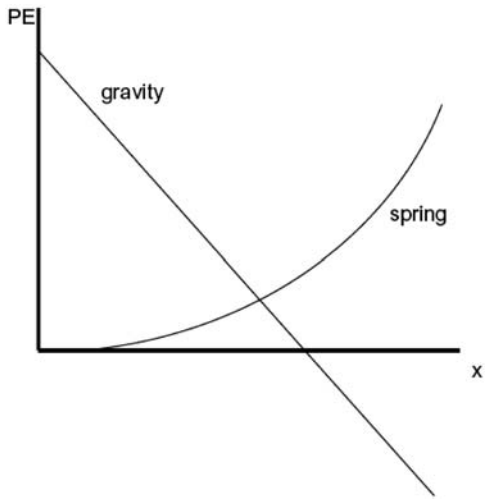


FIGURE 4.8 *Potential energy functions for gravity and springs.*

we see that the x-component of the force can be found from knowing how the potential energy changes in the x-direction

$$F_x = -\frac{\Delta \text{PE}}{\Delta x}. \tag{4.23}$$

Although this has been written for the case when the force is constant, it can also be written for forces that vary from point to point. The conclusion is that the potential energy function, which is just a scalar, contains all the information of the force, itself a vector quantity. Although in the case of one-dimensional motion, this does not seem to be a huge advantage, we show that the potential energy function contains all the information needed to calculate the force in three dimensions as well. For this reason alone, it should be clear that using energy concepts will often make it simpler to understand the motion of objects.

From Equation (4.23), it is clear that if the PE is increasing as x increases, the force in the x-direction will be negative, or tending to drive the system toward lower potential energy. On the other hand if the PE is increasing as x decreases, the force will be in the positive direction tending again to drive the system toward lower potential energy. Similarly, if the PE decreases as x increases, the force will be in the positive direction, whereas if the PE decreases in the negative direction, the force will now be in the negative direction. In all cases the force is such as to drive the system toward lower potential energy. We show just below that at a minimum in the potential energy versus x graph, where the slope is zero, there is no force acting in the x direction, and such a point is an equilibrium point. This picture allows us to consider the PE versus x graph as a sort of "slide" along which a particle always tends to move downhill in potential energy.

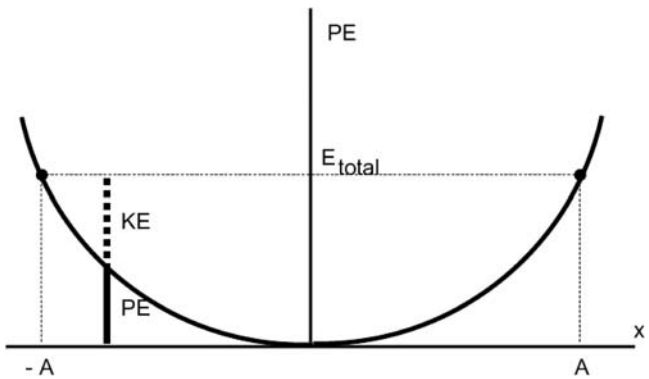
Not every force, however, can be found from a potential energy function. The frictional force is a prime example of a *nonconservative force* because the work done by this force depends on other factors than just the displacement of the object, such as its velocity or its actual trajectory. In the development of conservation of mechanical energy in the previous section, if there is a frictional force acting then the total mechanical energy E will no longer be a constant. Starting from the work–KE theorem, it is straightforward to show that the work done by the friction force is equal to the change in mechanical energy of the system

$$W_f = \Delta E = \Delta \text{KE} + \Delta \text{PE}, \tag{4.24}$$

where ΔPE represents the total change in potential energy from all conservative forces. The lost mechanical energy shows up as other forms of energy, most notably in the form of thermal energy in slightly warming the object and its environment.

Potential energy functions depend on the position of an object. A very useful way to represent potential energies is through the use of graphs. Figure 4.8 shows two examples of such graphs, one for the gravitational potential energy function and the other for the spring potential energy function. In the case of gravity, the potential energy is linear in the height, whereas for springs the potential energy function is quadratic in the displacement of the mass from equilibrium. Given an object with a certain total mechanical energy, in the absence of nonconservative forces, the kinetic and potential energies must add up to a constant total.

FIGURE 4.9 *Potential energy function for a spring, showing the turning points corresponding to the amplitude of oscillation. At any location between the turning points the total energy is divided between PE and KE as shown by the vertical bars.*



In the graphs of spring potential energy versus position in Figure 4.9, a point where the constant total energy intersects the potential energy function defines a point where the energy is totally potential and, hence, a point at which there is no kinetic energy. At such a *turning point* of the motion, the velocity is zero and the object cannot be found beyond the turning point where the total energy lies below the potential energy curve. If there are two turning points then the region between them defines a domain in which the particle is trapped and must

oscillate, constantly exchanging kinetic for potential energy and vice versa. If there is only one turning point, then an object will continue its motion unbounded.

In Figure 4.9, the point $x = 0$ where the potential energy is zero represents the position where the kinetic energy is a maximum because the total energy is all kinetic energy at that point. From our discussion of springs you will remember that as a mass on a spring oscillates it has its maximum speed as it passes through the equilibrium point. As the mass oscillates it constantly exchanges kinetic energy for potential energy and back again.

Remembering Equation (4.23), the negative of the slope of a graph of *PE* versus x will be the force on the object in the x-direction. Thus, the steeper the graph, the stronger the force and a positive slope (the curve for $x > 0$ in Figure 4.9) corresponds to a force in the negative direction, whereas a negative slope (the curve for $x < 0$ in the figure) indicates a positive force. These directions should make sense to you based on the motion of a mass on the spring. Those points that have zero slope are points where there is no force acting and are called points of *equilibrium*. We can distinguish three types of equilibrium: *stable*, *neutral*, and *unstable*. These are distinguished by what happens if the object is slightly displaced from the equilibrium position. For a point of stable equilibrium, there will be a restoring force tending to maintain the equilibrium. These points are graphically represented by zero-slope points in a potential valley or trough as in Figure 4.9. To either side of the equilibrium point, the sign of the force determined from Equation (4.23) produces a restoring force as shown in Figure 4.10A. Thus a mental picture of a small ball rolling on the potential energy curve will give a good idea of the nature of the forces. The steeper the walls are, the stronger the restoring force. In the case of neutral equilibrium (Figure 4.10B), there is no force over an interval so that a small displacement still results in no force acting. When an object is in unstable equilibrium (Figure 4.10C), a small displacement will result in a large force that tends to sweep the object farther away from the equilibrium point. In this case the graphical picture is an equilibrium point at the top of a hill so that the sign of the force is such as to produce an unstable equilibrium.

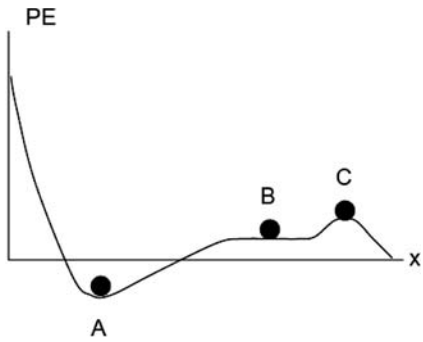


FIGURE 4.10 *A potential energy function showing points of stable (A), neutral (B), and unstable (C) equilibrium.*

Example 4.7 Figure 4.11 shows several additional examples of one-dimensional potential energy functions for a point mass. Examine these figures carefully and for each indicate: (a) the turning points, if any, depending on the total energy of the particle (E_1 through E_5); (b) the equilibrium points and their type; (c) the motion expected for different total energies of the particle.

Solution: A: The particle, in this case, cannot have a total energy, $E = \text{KE} + \text{PE}$, less than zero, because $\text{KE} \geq 0$ always and the potential baseline everywhere is at $\text{PE} = 0$. If the particle has an energy less than the barrier height ($0 < E < E_2$), and is initially found close to the origin, then the particle will have turning points at $x = 0$ and at the barrier and will be trapped, bouncing back and forth between $x = 0$ and the barrier. The steep walls give a very large force

$$\left(F = -\frac{\Delta \text{PE}}{\Delta x} = -\text{slope}\right)$$

when the particle hits them, simply turning it around and trapping it. There are no equilibrium points because the particle cannot be at rest (except for the uninteresting case when $E = 0$). If the particle is initially outside the barrier wall and traveling toward $x = 0$, it will rebound off the barrier and travel forever out toward larger x values unbounded, never returning. This model potential is useful for representing a trapped particle in the simplest potential. For energies $> E_2$, the particle will not be bound, but will slow down when passing over the barrier, because the KE will decrease when the $\text{PE} = E_2$ at the barrier.

(*Continued*)

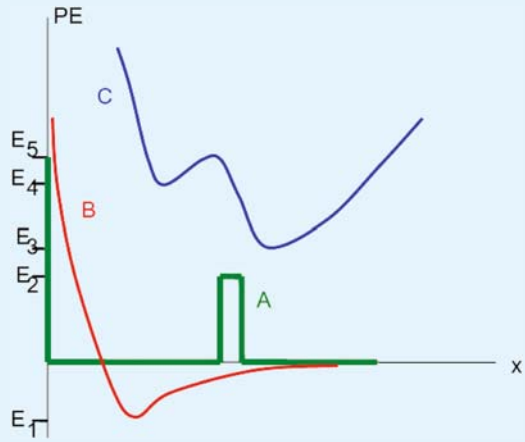


FIGURE 4.11 *Three different potential energy functions for a point mass: (A) is usually known as a barrier potential, (B) is a typical interatomic potential, and (C) illustrates a potential with two minima.*

B: In this case the lowest energy possible for the particle is E_1 and for particle energies within the range $E_1 < E < 0$, the particle will be trapped in the energy "well" and bounce back and forth between two turning points defined by the particular energy of the particle. The collisions of the particle with the potential near $x = 0$ will be harder (greater force) because the walls are steeper. There is a stable equilibrium at the bottom of the well. If the particle has an energy $E > 0$, then it will not be trapped and will, if headed toward $x = 0$, rebound from the potential wall and travel off freely. This example is a common one for an electron in an atom or an atom in a molecule, representing a single stable situation for negative energies with positive energies indicating an ionized electron or dissociated molecule.

C: In this case the lowest energy possible is E_3 and a particle with energy between $E_3 < E < E_4$ will clearly be trapped within the deeper well and have two turning points and one stable equilibrium position at the bottom of the well. If the particle energy exceeds E_4, but is less than E_5, the particle could be trapped in either well depending on its initial location. In either case there are two turning points and stable equilibria at the well bottoms. With an energy greater than E_5, the particle is still trapped in the overall well but is now free to roam over a larger range of x values. This example is quite a common one in chemistry and might represent the potential seen by one molecule in its interactions with another one. A molecule trapped in the higher-energy well might, with some "help" from an enzyme, be able to overcome the energy barrier presented by the middle hump (a point of unstable equilibrium) and then find the lower energy minimum. In a different context, this potential might also be used to represent the energies of different conformations of a macromolecule with two possible stable states of different energies. Because of their common use in describing atomic and molecular interactions, it is important to be comfortable with such graphs and to know how to interpret their major features.

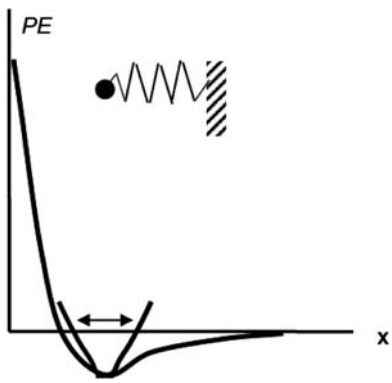


FIGURE 4.12 *A general potential energy function with a spring potential approximation near the equilibrium point.*

There is a special reason for emphasizing springs and the potential energy they store. It is shown in the box that near the minimum of any potential energy curve, the potential energy can be well represented by a quadratic function of the displacement from equilibrium, just the relation that holds for springs. Given this fact, we are justified in using the pictorial representation that an object trapped near a minimum in a potential well is, in fact, attached to a linear spring (see Figure 4.12). This is an often-used representation for the forces on atoms or molecules near their equilibrium

positions. We return to this theme later in the book after we learn a bit more in Chapter 10 about oscillations and the more realistic cases when damping (or frictional) forces are present.

5. POWER

Often when work is done on or by an object, the rate at which the work is done, and the consequent rate at which energy is transferred, is of interest. When a brick wall is built, the total work to lift and assemble all the bricks can be calculated, but the rate at which the wall is built is also of separate interest, particularly to the workers. When we expend energy doing work with our muscles, there is a maximal rate at which we can do work based on our bodies' limited ability to generate tension, just as there is a maximum rate at which cars can accelerate. Similarly our hearts have a maximal rate at which they can do work pumping blood through our bodies. Toasters and electric heaters give off heat, or thermal energy, at a rate that we later see how to calculate. All of these rates are controlled by the appropriate variables of the particular problem.

The rate at which work is done is known as the power P where

$$P = \frac{\Delta W}{\Delta t}. \tag{4.25}$$

If a constant force is acting then, using the definition of work in Equation (4.3), we can write that power is given by

$$P = F\frac{\Delta x}{\Delta t} = Fv. \tag{4.26}$$

If the force and velocity are in the same direction, either both positive or both negative, then the power is positive and, if there is only the one force acting, the velocity will increase in magnitude as will the kinetic energy. If the force is acting in the opposite direction to the velocity, then the power is negative and the velocity will decrease in magnitude as will the kinetic energy. Units for power are given by 1 J/s = 1 watt (W). The watt is familiar from its use in electrical power, indicating the rate at which energy is given off by light bulbs. Also, those of you who receive bills for electric power might recognize the common unit of energy used as the kW-hr, a product of a power measured in kW and a time measured in hours.

Any reasonably behaved mathematical function $U(x)$ can be written as a series, expanded about some point x_0,

$$U(x) = U(x_0) + \left.\frac{dU}{dx}\right|_{x0}(x - x_0) + \frac{1}{2}\left.\frac{d^2U}{dx^2}\right|_{x_0}(x - x_0)^2 + \ldots$$

If $U(x)$ represents any potential energy function and x_0 is a position of a stable energy minimum, then the slope dU/dx at position x_0 is equal to zero. Furthermore, the value of $U(x_0)$ is arbitrary and can be taken as zero. For small displacements from equilibrium the remaining quadratic term in the series dominates and if we let the second derivative of U with respect to x evaluated at x_0, a constant, be renamed k, we have

$$U(x) = \frac{1}{2}k\,(x - x_0)^2.$$

With $(x - x_0)$ being the displacement from the equilibrium position, this is precisely the expression for the potential energy of a spring when stretched a distance $(x - x_0)$ from its equilibrium length. Graphically this implies that near the minimum of any (mathematically well-behaved) potential energy curve, we can approximate the curve as a parabola as shown in Figure 4.12. Thus for small displacements about the stable equilibrium position, all objects feel a springlike restoring force.

Example 4.8 Let's try to calculate the wind power possible to tap using high-efficiency windmills (Figure 4.13). Assume a wind speed of 10 m/s (about 20 mph) and a windmill with rotor blades of 45 m diameter.

Solution: To calculate the maximum power possible, we need to calculate the kinetic energy of the wind intercepted by the rotor blades of the windmill. We take the density of air from Table 1.3 as $\rho = 1.29\ \text{kg/m}^3$. Then the KE = 1/2 mv^2 = 1/2 $(\rho V)v^2$, where V is the volume of air. We can calculate the volume of air intercepting the rotor blade cross-sectional area A per second by imagining a cylinder of air with the diameter of the blades and a length given by $(v)(1\ \text{s})$, the distance traveled in 1 s. Then we can write that, first assuming all this energy is collected by the windmill, $P = \Delta W/\Delta t = \Delta KE/\Delta t$ = 1/2 $\rho A v^3$. Substituting in numbers, we find that P = 1/2 $(1.29)(\pi\ 45^2/4)(10)^3 = 1.0 \times 10^6$ W. Typical efficiencies of modern

(*Continued*)

FIGURE 4.13 *0.75 Megawatt generating windmills in Minnesota.*

windmills are greater than 40%. This means that roughly 40% of the wind energy is converted into electrical energy. Note that the power has a large dependence of wind velocity, proportional to v^3, so that an increase in wind speed of 10% translates into an increase in power by a factor of $(1.1)^3 = 1.33$, or a 33% increase. Good location of windmills is therefore extremely important.

CHAPTER SUMMARY

In one dimension, the work done by a constant force acting along the same direction as the displacement is

$$W_F = F\Delta x. \qquad (4.3)$$

The net work done on an object is equal to the change in its kinetic energy, KE,

$$W_{\text{net}} = \text{KE}_2 - \text{KE}_1 = \Delta\text{KE}, \qquad (4.11)$$

where

$$\text{KE} = \frac{1}{2}mv^2. \qquad (4.10)$$

Work done by conservative forces on an object can be related to a potential energy function PE through

$$W = F_x\Delta x = \ -\ \Delta\text{PE}, \qquad (4.22)$$

(*Continued*)

so that, in turn, the force acting on the object can be determined from that potential energy function from

$$F_x = -\frac{\Delta \text{PE}}{\Delta x}. \quad (4.23)$$

Two examples are gravitational and spring potential energy, given by

$$\text{PE}_{\text{grav}} = mgy. \quad (4.13)$$

and

$$\text{PE}_{\text{spring}} = \frac{1}{2}kx^2. \quad (4.19)$$

In the absence of any dissipative forces, such as friction, the total mechanical energy E is conserved:

$$E = \text{KE} + \text{PE}_{\text{grav}} + \text{PE}_{\text{spring}} + \text{PE}_{\text{other}} = \text{constant}. \quad (4.21)$$

Power P is the time rate of change at which work is done,

$$P = \frac{\Delta W}{\Delta t}, \quad (4.25)$$

and can also be written as

$$P = F\frac{\Delta x}{\Delta t} = Fv. \quad (4.26)$$

QUESTIONS

1. Give some examples that contrast the "physics" definition of work with the colloquial usage of work. In particular, give some examples where no work is done (according to our physics definition) whereas in ordinary speech one would say that work was done.
2. Can work be done on an object without moving it? Give an example to illustrate your answer.
3. Conservation of energy would seem to imply that holding a heavy weight at rest, doing no work, should not require any energy. What is wrong with this argument?
4. A heavy crate sitting on the ground is lifted vertically onto a table, then pushed horizontally across the table, and then lowered vertically to the ground. Fill out the following table with your answers for whether the work done by the external force and by gravity are positive, negative, or zero for each part of the motion.

Portion of Motion	*Gravity*	*External Force*
Vertical lift		
Horizontal slide		
Vertical lowering		

5. In slowly compressing a vertical spring a distance d, a mass placed on top of the spring will compress the spring until it reaches equilibrium with mg balanced by a spring force equal to kd, so that $d = mg/k$. On the other hand, the initial potential energy of the mass mgd is converted into spring potential energy $1/2kd^2$ when the mass is released from rest, so that $d = 2mg/k$. What is wrong with the above reasoning and which is the correct result? (Hint: Think of what happens in actually doing each of the two different experiments.)
6. In our discussions the location of the zero of gravitational potential energy is arbitrary but the zero of spring potential energy is not. Why is this the case? When the location of zero gravitational potential energy is shifted by a distance y_0, the gravitational potential energy at some location changes by mgy_0, an arbitrary constant. What would happen if the location of zero spring potential energy were shifted by a distance x_0 from its proper location?
7. Two springs with spring constants that differ by a factor of two are stretched (a) by the same amount, and (b) with the same force. Compare the force exerted and stretch of the two springs for each situation.
8. Describe, in words, the types of energy a mass on a spring has at various points on its potential energy curve shown in Figure 4.9.
9. Explain how the motion of a marble rolling in a bowl is similar to the motion of a mass on a spring. Think in terms of potential energy diagrams.
10. Check that the units on both sides of Equation (4.23), relating energy to force, agree. Why is there a minus sign in the equation?
11. Two students are solving a physics problem having to do with finding the velocity of a ball when it reaches the ground after being dropped out of a ten-story

building. One chooses the zero of gravitational potential energy to be on the ground, and the other chooses it to be at the tenth floor of the building. Can they both get the same answer?

12. Two workmen are stacking heavy cinder blocks from the ground to a raised pallet. If one of them stacks 100 of the blocks in 20 min and the other stacks 100 of them in 30 min, which one has done more work? Which one has the greater power output?

13. Two joggers run up stairs, starting out together, but one runs up 4 flights in 15 s and stops and the other runs up 12 flights in a minute. Which has done more work? Over the first 15 s, which has the greater power output? Over the minute interval, which has the greater average power?

14. Which laser emits the most energy: a continuous laser with a power level of 10^{-2} W, or a pulsed laser emitting a series of 10^{-12} s duration pulses every 10^{-2} s with each pulse having a power of 10^{7} W?

MULTIPLE CHOICE QUESTIONS

1. A 1 kg mass initially compresses a vertical spring by 0.1 m. The mass is not attached to the spring and, after being released from rest, it leaves the spring and eventually reaches a maximum height above its starting point of 0.5 m. There is no friction during this motion. The change in the mass's mechanical energy during this process (a) must be about +5 J, (b) must be zero, (c) must be about −5 J, (d) cannot be calculated because the spring constant is not given.

2. The fundamental SI dimensions of work are (a) MLT^{-1}, (b) MLT^{-2}, (c) ML^2T^{-1}, (d) ML^2T^{-2}.

3. A 75 kg hiker carries a 25 kg backpack up a mountain trail with an average inclination angle of 5° over a distance of 3 km. The total work done by the hiker is about (a) 260 kJ, (b) 65 kJ, (c) 3000 kJ, (d) −260 kJ.

4. A lead ball weighing 10 N falls 0.8 m from rest into a bucket of sand. The ball stops after making a crater 0.2 m deep. According to the work–energy theorem the work done by the sand on the ball in bringing it to rest is (a) −10 J, (b) −2 J, (c) 0 J, (d) +10 J.

5. A 5 kg block is accelerated from rest by a constant force of 10 N over a distance of 1 m on a frictionless horizontal surface. The block then slides at a constant speed for 2 m before hitting a spring with a spring constant of 10 N/m. The work done by the spring in bringing the block to rest momentarily before returning it in the reverse direction is (a) 10 J, (b) 20 J, (c) −20 J, (d) −10 J.

6. A mass m is lowered gently onto a vertical spring of length L with spring constant k until it just touches the spring. Let y be the distance the spring is compressed and v be the velocity of the mass. When the mass is released from rest, the equation for conservation of energy is (a) $1/2\ mv^2 + 1/2\ ky^2 + mgy = mgL$, (b) $1/2\ mv^2 + 1/2\ k(L - y)^2 + mgy = mgL$, (c) $1/2\ mv^2 + 1/2\ ky^2 + mg(L - y) = mgL$, (d) $1/2\ mv^2 + 1/2\ ky^2 + mgy = 0$.

7. A mass M rests on top of a vertical spring with spring constant k. If a second mass m is stuck to mass M, the maximum distance the spring is further compressed is given by (a) mg/k, (b) $mg/2k$, (c) $2mg/k$, (d) $(m + M)g/k$.

8. Two identical springs with 5 N/m spring constants are both attached to the same 2 kg mass as shown. If the mass is pulled down slightly and released, it will oscillate with a period of

(a) $2\pi\sqrt{\frac{2}{5}}$,

(b) $2\pi\sqrt{\frac{2}{10}}$,

(c) $2\pi\sqrt{\frac{2}{2.5}}$,

(d) $4\pi\sqrt{\frac{2}{5}}$.

9. A mass weighing 10 N is initially held at rest on a vertical spring that is compressed by 0.1 m. When released, the mass accelerates upward, leaves the spring and eventually reaches a height of 0.9 m above its starting height. The work done by the spring on the mass is (a) −10 J, (b) +1 J, (c) +9 J, (d) +10 J.

10. In the absence of friction, when an object in neutral equilibrium is given a small momentary push, it will (a) return to its equilibrium position, (b) stop at a new equilibrium location, (c) move at a constant velocity until the potential changes, (d) depends on the object and type of potential energy function.

11. A bricklayer is building a wall. If the 0.5 kg bricks are all identical with a 0.1 m height and he builds a stack 10 blocks tall and 10 blocks wide in 1 h, his power output is (a) 3.75 W, (b) 0.063 W, (c) 0.076 W, (d) 0.069 W. (Take $g = 10\ m/s^2$.)

12. A girl pulling a sled exerts a 20 N force horizontally for 10 s. How much power does she generate in watts while moving the sled 20 m? (a) 10, (b) 20, (c) 30, (d) 40.

13. A block slides a distance d down a frictionless inclined plane, with inclination angle θ, changing its height by a displacement H. The work done by gravity is equal to (a) $mgH \sin\theta$, (b) mgH, (c) $-mgH$, (d) mgd (e) $-mgd$.

PROBLEMS

1. In mowing a lawn, a boy pushes a lawn mower a total distance of 350 m over the grass with a force of 90 N directed along the horizontal. How much work is done by the boy? If this work were the only expenditure of energy by the boy, how many such lawns

would he have to mow to use the energy of a 200 cal candy bar? (use 1 calorie = 4200 J)

2. As a bacterium swims through water it propels itself with its flagella so as to overcome the frictional drag forces and move at, more or less, constant velocity of 100 μm/s for periods of time. If the frictional drag force on a bacterium is 0.1 μN, how much work does the bacterium do in 1 s of sustained velocity.

3. A 100 N crate sits on the ground and is attached to one end of a rope passing over a frictionless light pulley. If someone pulls down on the rope with a constant force of 110 N lifting the crate a distance of 3 m, find
(a) The work done by the person
(b) The work done by gravity
(c) The increase in potential energy of the crate
(d) The velocity of the crate after rising 3 m.

4. An elevator car weighing 8000 N in a tall office building is lifted by a steel cable attached to the elevator motor. It travels from ground level to the 50th floor, a distance of 200 m in 75 s. Ignore the brief time during which the elevator accelerates or decelerates.
(a) How much work is done by the motor in lifting the elevator?
(b) At what rate is this work done?
(c) Answer the previous parts for the downward nonstop trip.

5. A ball is thrown downward from the roof of a 24 m tall building with an initial speed of 5 m/s.
(a) Use energy principles to find the speed with which the ball hits the ground.
(b) Find the time it took for the ball to reach the ground.
(c) If the ball were thrown upwards from the roof with the same speed repeat the calculations for parts (a) and (b).

6. A boy throws a 0.1 kg ball from a height of 1.2 m to land on the roof of a building 8 m high.
(a) What is the potential energy of the ball on the roof relative to its starting point? Relative to the ground?
(b) What is the minimum kinetic energy the ball had to be given to reach the roof?
(c) If the ball falls off the roof, find its kinetic energy just before hitting the ground.

7. Water leaves a garden hose held vertically with a velocity of 5 m/s. If the hose is held at a height of 2 m, find the speed with which the water hits the ground.

8. How much mechanical work is done by a 2 cm long × 0.2 mm diameter muscle fiber that shortens by 20% during a sustained contraction generating an average stress of 38×10^4 N/m^2?

9. A 65 kg rock climber scales a 200 m vertical wall in 10 min. Find the work done by gravity on the hiker. If the hiker consumed oxygen at a rate of 2 L/min, corresponding to an internal energy production of 4×10^4 J/min, what fraction of the hiker's energy was used to climb the wall? (This fraction is termed the hiker's efficiency.)

10. In throwing a 0.5 kg lacrosse ball from rest, the lacrosse stick exerts an average force of 500 N along a distance of 1.2 m before the ball leaves the net.
(a) How much work was done on the ball by the stick?
(b) With what velocity does the ball leave the lacrosse stick?

11. A weight lifter "snatches" a 1200 N weight by exerting a 1400 N average force for the first meter off the ground, then relaxing his grip and "getting under" the bar to catch it and give it a final upward push.
(a) How much work is done in the first 1 m of lifting by the man? By gravity?
(b) What velocity will the weight attain after the one meter lift?
(c) If the man essentially exerts no force starting at 1 m height, how much farther will the bar rise and how long will it take to rise to that height? During that brief time he will finalize his position to "get under" the bar and then push it to full arm extension.
(d) How much additional work must he do to raise the weight to 2.4 m, the height of his full arm extension?

12. A 5 N/m horizontal spring is compressed 0.1 m and a 0.1 kg mass is attached. The mass glides on a frictionless horizontal surface. What is the maximum speed of the mass as it oscillates?

13. A 2 kg block slides back and forth on a frictionless horizontal surface bouncing between two identical springs with $k = 5$ N/m. If the maximum compression of a spring is 0.15 m, find the gliding velocity of the block between collisions with the springs.

14. A 0.2 kg mass is dropped 0.5 m onto a vertical spring with a 10 N/m spring constant and sticks to it.
(a) What speed does the mass have as it hits the spring?
(b) Find the equilibrium position of the mass relative to the original position of the top of the spring as it oscillates.
(c) Find the maximum compression of the spring.
(d) What is the maximum speed of the mass as it oscillates on the spring?

15. A 20 N/m vertical spring is stretched 5 cm when a mass is attached. If the same mass is set into oscillation after stretching the spring an additional 10 cm find
(a) The mass
(b) The maximum kinetic energy of the mass
(c) The maximum speed of the mass and where it occurs relative to the original unstretched position of the spring

16. The power stroke of the myosin protein on an actin filament that generates tension in a muscle appears to be a 10 nm displacement generated by about a 1 pN force. Each power stroke is accompanied by the

splitting of one ATP molecule which releases about 4.9×10^{-20} J.

(a) How much work is done by one myosin in a single power stroke?

(b) What is the efficiency of the process; that is, what fraction of the ATP-generated energy does useful work?

17. A powerful pulsed laser emits a series of brief ns (10^{-9} s) pulses of light, one per ms (10^{-3} s). If each pulse has a power of 10^{10} W, calculate the energy per pulse and the average power of the laser over a second.

18. A bricklayer is building a garden wall 1.0 m tall out of bricks that are 10 cm tall, 30 cm long, and weigh 10 N each. If the wall is 3 m long

(a) How much work must be done to build the wall if all the bricks start out at ground level?

(b) If he works for two hours and then takes a one hour lunch followed by a two hour rest, and then returns to finish the wall in two more hours, what is the average power he uses to build the wall over his seven hour day? Over his actual four hour construction time?

19. Two kids, Jimmy and Sally, ride on sleds on a frozen pond at the same speed. When they are 30 m from a log in the ice, Sally drags her foot to slow her sled down at a constant deceleration while Jimmy continues at constant velocity. Jimmy reaches the log in 5 s and Sally's sled comes to a stop right at the log in 10 s.

(a) What is the initial velocity of both sleds?

(b) What is the acceleration of Sally's sled?

(c) If Sally plus sled have a combined mass of 50 kg, what is the drag force that Sally's foot applies?

(d) How much work was done by Sally's foot in bringing the sled to rest?

5

Motion, Forces, and Energy in More Than One Dimension

In the previous three chapters we have detailed the kinematics of one-dimensional motion, forces producing the motion, their dynamical connections via Newton's laws, and the important concept of energy. Having built up an arsenal of tools for the description and prediction of motion in one dimension, we need just one more added tool in order to generalize to the study of kinematics and dynamics in two or three dimensions. Although we obviously live in a three-dimensional world, it is very useful to study two-dimensional motion, which can describe any motion confined to a plane, for example, free-fall near the Earth's surface—but now with horizontal motion thrown in—or circular motion, or the local motions of a membrane protein confined to a cell surface. We limit most of our discussion to two-dimensional motion, but the extension to three dimensions is clear.

The missing mathematical tool that we need to complete this agenda is vector algebra and is the opening subject of this chapter. With knowledge of vectors, the goal of this chapter is to see how to generalize our fundamental results so far for one-dimensional motion so that we can apply them to more realistic situations. Both kinematical and dynamical problems are studied as well as the generalizations of work and energy to more than one dimension. Frictional forces are not only extremely common, but often play either a crucial role or provide an ultimate limit to mechanical motion, as we show. Both static and kinetic contact friction are discussed and their role in some problems where one object slides over a surface is illustrated. Circular motion is one type of regular motion in a plane and we examine the dynamics of such motion with applications to the important experimental technique of centrifugation. We return to circular motion as the theme of Chapter 7 on aspects of rotational motion.

1. VECTOR ALGEBRA

Vectors are mathematical representations for quantities that have not only a magnitude, or amount, but also have a direction. Quantities without directionality, such as time, speed, mass, energy, and temperature, are called *scalars*. These are totally defined by an amount, given by a number and units. Vector quantities, including position, displacement, velocity, acceleration, and force also require some specification of their direction. This can be done graphically by representing vector quantities as arrows (the pointed ends known as "heads" and the other ends as "tails") with their lengths drawn to scale according to the amount of the quantity and pointing in the proper orientation. Thus, for example, the length of a drawn displacement vector might scale according to the rule 1 cm = 100 km, and the length of a drawn velocity vector might follow 1 cm = 100 km/hr.

Vector analysis originates in how displacements behave. For specificity, suppose we view a single *E. coli* bacterium under a microscope and record its two-dimensional

J. Newman, *Physics of the Life Sciences*, DOI: 10.1007/978-0-387-77259-2_5,

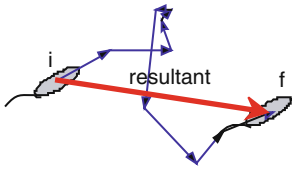


FIGURE 5.1 *Series of equal time displacement vectors for an E. coli with the resultant displacement.*

position at various times. From these, we can construct a series of displacement vectors, each of which starts at the initial position and ends at the final position for that time interval (Figure 5.1). Each vector is labeled using a special symbol, for example, $\vec{A}$, to indicate that it is a vector. In this text we use an arrow over a letter to indicate that it is a vector quantity. You should discipline yourself to do the same when solving problems. Vectors, like scalars, can be added, subtracted, and multiplied, but precisely how these operations are done is different from the way they work for scalars (ordinary numbers). Failure to distinguish between scalars and vectors can lead to unnecessary calculational problems. Our first task is to learn how to add and subtract vector quantities.

If the bacterium depicted in Figure 5.1 had been made to move along a straight line, its sequence of displacement vectors for individual time intervals would all lie along that line. "Adding" those displacements together would simply require adding algebraic quantities (with plus and minus signs for positive or negative displacements as we have been doing) in order to find the net displacement over the entire time interval. In vector language, the individual displacements would be connected together head to tail and the net displacement, known as the resultant or vector sum, would be an arrow (of the correct length) with its tail at the tail of the very first displacement and its head at the head of the very last.

In the two-dimensional case of Figure 5.1, the net displacement is arrived at in a similar way: the head of each individual displacement vector is connected to the tail of the next in the sequence and the resultant (i.e., net) displacement is a vector with its tail at the tail of the first vector and its head at the head of the final vector.

Well, this head-to-tail construction for adding displacement vectors is fine for a sequence of displacements, but how do we add two velocity or two force vectors together, situations where sequence has no meaning? We need to generalize the graphical construction rule to permit the addition of two vectors even if they aren't originally connected in the correct head-to-tail way. We do so by defining *vector equality.* Two vectors are said to be equal if they have the same length and point in the same direction. To check whether that is true, imagine translating one vector rigidly (no rotating as you go, please) until its tail coincides with the tail of the other. If the two heads also coincide, the two vectors are equal. This is shown in Figure 5.2.

With the notion of vector equality, any two vectors representing the same quantity can be added. Recall that you cannot add a velocity vector to a force vector; they are like apples and oranges. Translate one rigidly until its head is at the tail of the second. The resultant is a vector of the same kind whose tail is at the tail of the first and whose head is at the head of the second. Which is the first vector and which the second? It doesn't matter. The order of the vector addition does not affect the result, as is illustrated in Figure 5.3.

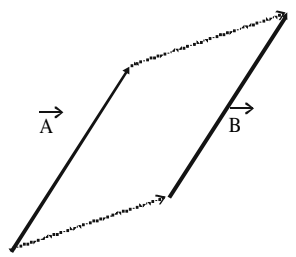


FIGURE 5.2 *When $\vec{A}$ is rigidly translated until its tail coincides with the tail of $\vec{B}$, the heads of the two vectors coincide also: therefore $\vec{A} = \vec{B}$.*

Example 5.1 Graphically add three vectors, all with tails at the origin of some coordinate system but with heads at different points in the *x*–*y* plane:

$\vec{A}$ has head at (3,0), $\vec{B}$ has head at (2,4), and $\vec{C}$ has head at (−4,2), where the notation means (*x, y*) (see Figure 5.4).

Solution: Each vector is first graphed according to its (*x, y*) coordinates using a given scale. Then vectors $\vec{B}$ and $\vec{C}$ are moved so that their tails sit at the head of the previous vector.

We then can read the resultant from the graph by reading the (*x, y*) coordinates of its head to be $\vec{A}+\vec{B}+\vec{C}=(1,6)$. We can also measure the magnitude of the resultant directly by measuring its length and using the scale of the diagram to find a magnitude of about 6.1. The direction of the resultant is found using a protractor to be about 80° above the *x*-axis. Notice that the *x*- and *y*-coordinates of the head of the resultant are equal to the sum of the

separate x- and y-coordinates of the heads of the three vectors. We show why this is so below.

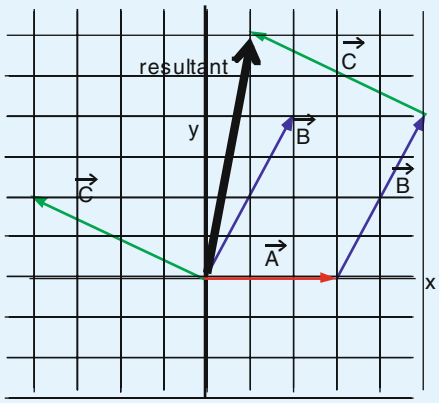


FIGURE 5.4 *Graphical addition of three vectors.*

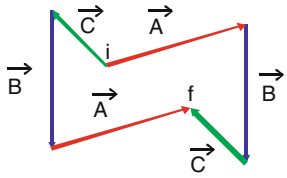


FIGURE 5.3 *Graphical addition of three vectors, showing that the order of addition doesn't matter. Starting at point i, the sum* $\vec{A} + \vec{B} + \vec{C}$ *ends up at point f, regardless of the order of addition.*

For calculational purposes it is often very useful to refer vectors to some underlying coordinate system. Figure 5.5 shows a vector $\vec{A}$ with its tail attached to the origin of a Cartesian (i.e., x–y) coordinate system. The figure shows two other vectors $\vec{A}_x$ and $\vec{A}_y$, also with tails at the origin. The latter vectors are constructed as follows. From the head of $\vec{A}$ draw a line parallel to the y-axis; where that line intersects the x-axis is the head of $\vec{A}_x$; draw a second line from the head of $\vec{A}$ parallel to the x-axis; where that line intersects the y-axis is the head of $\vec{A}_y$. The vector $\vec{A}_x$ is called the "component vector of $\vec{A}$ in the x-direction" and $\vec{A}_y$ the "component vector of $\vec{A}$ in the y-direction." Now, by rigidly translating either $\vec{A}_x$ or $\vec{A}_y$ it is easy to see that $\vec{A} = \vec{A}_x + \vec{A}_y$.

The promised calculational advantage to "decomposing" a vector into its coordinate components requires one more idea: multiplying a vector by a scalar. When a vector is multiplied by a scalar the result is a new vector pointed parallel (or antiparallel) to the first and with length equal to the first vector's length times the magnitude of the scalar. Figure 5.6 shows examples. The vector $2\vec{A}$ is twice as long as $\vec{A}$ and points in the same direction. The vector $-0.5\,\vec{A}$ is half as long as $\vec{A}$ and points in the opposite direction. (The reason multiplying a vector by a negative number produces a vector in the opposite direction is this: we require that when we add $\vec{A}$ and $-\vec{A}$ the result is a vector of zero length; the only way that can be is if when the tail of $-\vec{A}$ is attached to the head of $\vec{A}$, in the head-to-tail addition method, the head of $-\vec{A}$ is back at the tail of $\vec{A}$. Then the resultant's head and tail are at the same place and, as required, it has no length. In other words, $-\vec{A}$ is the same size as $\vec{A}$ but antiparallel to it.)

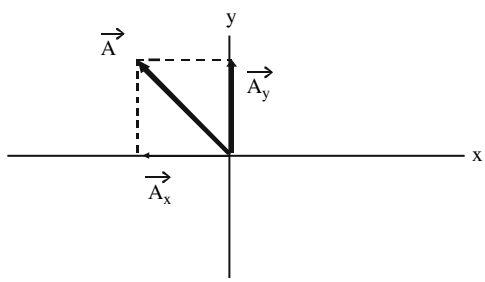
We write the two-dimensional vector $\vec{A}$ as the *ordered pair*

$$\vec{A} = (A_x, A_y),$$

where the (signed) numbers A_x and A_y are called the *x- and y-components of the vector $\vec{A}$*. A three-dimensional vector is written as an ordered *triple*. A vector is not simply a number; that's why we use the arrow symbol. A vector is a set of numbers, from which its magnitude and direction information can be extracted. Because the (x, y) components of a vector are perpendicular to each other, the magnitude of a vector (denoted by putting the vector symbol inside a pair of vertical lines) can be obtained from them by Pythagoras' theorem; for example,

$$|\vec{A}| = \sqrt{A_x^2 + A_y^2}.$$

FIGURE 5.5 *The x- and y-component vectors of the vector* $\vec{A}$.



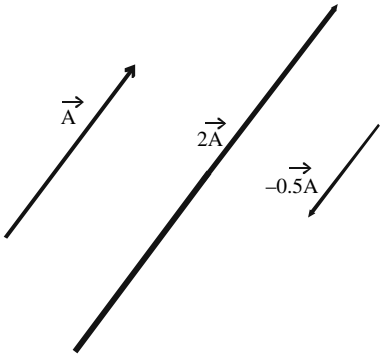


FIGURE 5.6 *The result of multiplying a vector by a scalar.*

The direction of a vector can be deduced by using a little trigonometry: let θ be the angle the vector makes with the x-axis; then

$$\cos(\theta) = \frac{A_x}{|\vec{A}|}.$$

Because the cosine can have the same value for more than one angle you have to draw a picture to get the orientation of the angle right (i.e., whether θ is above the x-axis or below it). The component notation makes vector addition much easier and more accurate than drawing head-to-tail pictures. The rule is this: when two vectors given in component notation are added, the x-component of the resultant is the sum of the x-components of the two vectors you started with and the y-component of the resultant is the sum of the y-components (and, if necessary, the z-component is the sum of the z-components). Note how this rule makes the calculation in Example 5.1 so much easier.

Example 5.2 Calculate analytically the resultant of the two vectors $\vec{A} = (0, 6)$ and $\vec{B} = (5, 0)$.

Solution: $\vec{C} = \vec{A} + \vec{B} = (5,6)$. This sum or ordered pair completely specifies the resultant vector. If we wanted to express the resultant in terms of its magnitude and direction we could do so by writing

$$|\vec{C}| = \sqrt{5^2 + 6^2} = 7.8.$$

The direction of the resultant (sketched in Figure 5.7) is found using

$$\cos\theta = \frac{C_x}{|\vec{C}|} = 5/7.8 = 0.641,$$

so that $\theta = 50.2°$, above the x-axis.

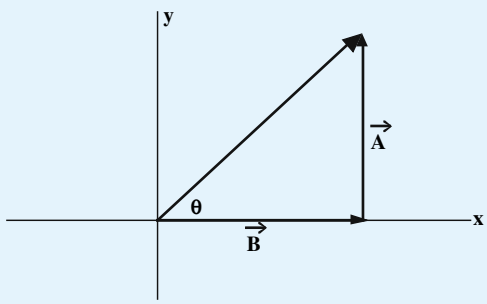


FIGURE 5.7 *Analytical vector addition.*

Example 5.3 Given $\vec{A} = (5, 2)$ and $\vec{B} = (-3, -5)$, express $\vec{C} = \vec{A} + \vec{B}$ in terms of (a) ordered pair notation, and (b) magnitude and direction.

Solution: (a) Adding separately the x- and y-components of the two vectors we find that

$$\vec{C} = ([5 - 3], [2 - 5]) = (2, -3).$$

(b) We then find that the magnitude of $\vec{C}$ is given by

$$|\vec{C}| = \sqrt{2^2 + (-3^2)} = \sqrt{13} = 3.6$$

and the direction, is given by

$$\cos^{-1}\left(\frac{2}{3.6}\right) = 56.3°$$

but, this time below the x-axis. (Draw a sketch to make sure you see why.)

This procedure can clearly be generalized to add together any number of vectors that lie in the x–y plane. First find each of the vector's components along the x- and y-axes, separately add the x- and y-components algebraically, and then finally combine the two remaining vectors using trigonometry. Table 5.1 summarizes this procedure.

Table 5.1 *Steps in Component Method of Vector Addition/Subtraction*

1. Make a rough sketch of the vectors, if not given.
2. Find the x-, y- (and z-) components of each vector, if not given order pair notation.
3. Perform the algebraic $+/-$/or multiplication by a scalar separately to each component, finding the x-, y- (and z-) components of the resultant.
4. If needed, combine the components of the resultant, using the Pythagorean theorem and trigonometry, to find the magnitude and direction of the resultant.

Example 5.4 Given the three vectors:

$\vec{A} = (2, -3, 1)$, $\vec{B} = (-5, 0, 2)$, and $\vec{C} = (0, 4, 1)$, find
a) $\vec{A} + \vec{B}$, *b)* $\vec{C} - \vec{A}$, and *c)* $\vec{A} + 2\vec{B} - \vec{C}$.

Solution: For each part we add the appropriate components of each vector separately to find:

a) $\dot{A} + \dot{B} = ([2 - 5], -3, [1 + 2]) = (-3, -3, 3)$.

b) $\dot{C} - \dot{A} = ([0 - 2], [4 + 3], [1 - 1]) = (-2,7,0)$.

c) $\dot{A} + 2\dot{B} - \dot{C} = ([2 - 2 \cdot 5], [-3 - 4], [1 + 2 \cdot 2 - 1]) = (-8, -7,4)$.

Can you draw the vectors involved in this example? Can you find their magnitudes?

2. KINEMATICS

With these properties of vectors and methods for vector addition, we are now in a position to generalize our discussion of kinematics to two- (or three-) dimensional motion. Start by identifying a reference point and establish a Cartesian coordinate system with its origin at this point. It doesn't matter where the origin is or how the axes are oriented, although some choices may make life simpler than others. We come back to how to choose smart systems in a moment. The rigidly translating object whose motion we wish to describe has a position vector $\vec{r}$ with tail at the origin and head at the point (x, y, z). Thus, we can write $\vec{r} = (x, y, z)$. As the

object moves x, y, and z change in time. The velocity of the object is another vector: $\vec{v} = (v_x, v_y, v_z)$. The components of $\vec{v}$ and the components of $\vec{r}$ are related just as in the one-dimensional case:

$$v_x = \lim_{\Delta t \to 0} \frac{\Delta x}{\Delta t}, \qquad v_y = \lim_{\Delta t \to 0} \frac{\Delta y}{\Delta t}, \qquad v_z = \lim_{\Delta t \to 0} \frac{\Delta z}{\Delta t}.$$

Similarly, the object may be accelerating. Acceleration is yet another vector: $\vec{a} = (a_x, a_y, a_z)$. The components of acceleration and the components of velocity are related analogously to the one-dimensional case:

$$a_x = \lim_{\Delta t \to 0} \frac{\Delta v_x}{\Delta t}, \qquad a_y = \lim_{\Delta t \to 0} \frac{\Delta v_y}{\Delta t}, \qquad a_z = \lim_{\Delta t \to 0} \frac{\Delta v_z}{\Delta t}.$$

When the object we are interested in is confined to move along a line, position, velocity, and acceleration are all along the same line. When the object is free to move in space, position, velocity, and acceleration can all point in different directions. This fact makes dealing with two- or three-dimensional motion more subtle. But, the preceding equations point out a very useful simplification: the x- (respectively, y-, z-) component of velocity only changes due to the x- (respectively, y-, z-) component of acceleration, and the x- (respectively, y-, z-) component of position only changes due to the x- (respectively, y-, z-) component of velocity.

SPECIAL CASE I: CONSTANT FORCE—FREE-FALL—PROJECTILE MOTION

In Chapter 3 we discussed the motion of an object in free-fall near the Earth where the motion was purely vertical. Such motion results when the initial velocity of the object has no horizontal component. Gravity is a purely vertical force resulting in a constant vertical acceleration; as we just argued, a vertical acceleration can only produce changes in the vertical component of velocity. So if there is no horizontal motion to start with, gravity can't produce any. But suppose the object is moving with some initial horizontal component of velocity; what does gravity do then? It can only change the vertical component of velocity, so the horizontal component remains unchanged during the object's flight. This result may surprise you: the horizontal and vertical components of an object's motion while it is in free-fall are completely independent of each other. Thus, for example, if an object is dropped from rest at a certain height off the ground at the same instant a second object is thrown from the same place with a large horizontally directed velocity, the two will strike the ground at exactly the same time! Both of these objects leave their starting point with zero vertical velocities. The time an object is in free-fall depends only on the vertical distance it has to travel and its initial vertical velocity, and because of how they start out (both start at the same point with no vertical velocity), both of these objects travel the same vertical distance in the same time.

Because the acceleration due to gravity is vertical and because the horizontal component of velocity in a free-fall situation cannot change, it is smart to orient our coordinate system as follows: one axis vertical (that will cause the acceleration to have a single component, along this vertical axis) and one axis in the direction of the horizontal component of velocity (that will cause the velocity and position vectors to have only two components, one vertical and one along this horizontal axis). It is usual to call the horizontal axis x and the vertical axis y (with up as positive). The z-axis is irrelevant; free-fall motion is at most two-dimensional.

In this coordinate system, the acceleration is the constant vector $\vec{a} = (0, -g)$. We can use the results tabulated in Table 3.1 to fill in how the velocity and position vectors vary in time, because $a_x = 0$ and $a_y = -g$, both constant. Just replace x by y for the vertical component of motion. We find that under free-fall $\vec{v} = (v_{0x}, [v_{0y} - gt])$ and $\vec{r} = ([x_0 + v_{0x}\, t], [y_0 + v_{0y}\, t - \frac{1}{2} g t^2\,])$.

Example 5.5 A cat running initially horizontally with a speed of 1.6 m/s runs horizontally right off a table 0.80 m high. Find (a) how long the cat is in the air, (b) how far it travels horizontally before it lands, and (c) its velocity just before hitting the ground.

Solution: The cat's motion is two-dimensional with the only acceleration due to gravity once it leaves the table. We take our coordinate axes to point in the usual way with the origin at the point the cat leaves the table. The vertical, or y, motion can be described by

$$y = 0 + 0 - \frac{1}{2} g t^2,$$

because the cat leaves from the origin and has no initial y-velocity. Substituting $y = -0.80$ m we find the time the cat is in the air is

$$t = \sqrt{-\frac{2y}{g}} = \sqrt{\frac{2 \cdot 0.80 \text{ m}}{9.8 \text{ m/s}^2}} = 0.4 \text{ s}.$$

During this time the cat's horizontal velocity remains constant (we neglect air resistance), so that the cat has traveled a horizontal distance given by

$$x = v_{0x} t = 1.6 \text{ m/s } (0.4 \text{ s}) = 0.64 \text{ m}.$$

To find the velocity of the cat as it is about to land, we need first to find its y-velocity just as it hits the ground, since we already know the x-velocity has remained constant. We find that because the initial y-velocity is zero,

$$v_y = -gt = -9.8 \text{ m/s}^2 \cdot 0.4 \text{ s} = -3.9 \text{ m/s}.$$

The cat's velocity just before hitting the ground can then be expressed as either

$$\vec{v} = (1.6, -3.9) \text{ (m/s)}$$

or by

$$v = \sqrt{1.6^2 + 3.9^2} = 4.2 \text{ m/s} \quad \text{at} \quad \theta = \cos^{-1}\left(\frac{1.6}{4.2}\right) = 68°,$$

where the angle is measured below the horizontal.

Example 5.6 A football is kicked with a speed of 40 m/s at an angle of 40° above the ground. Find (a) its velocity after 1 s, (b) the maximum height it reaches and its speed at that point, (c) the time for it to hit the ground.

Solution: We take the origin on the ground at the point the ball is kicked. (a) The initial velocity of the football has both horizontal ((40 m/s)cos(40°) = 30.6 m/s) and vertical ((40 m/s)sin(40°) = 25.7 m/s) components. Because there is only a

(*Continued*)

vertical acceleration, the horizontal component remains constant and the vertical component is governed by

$$v_y = v_{0y} - gt.$$

Therefore, after 1 s the y-component of velocity is $v_y = 25.7\ \text{m/s} - 9.8\ \text{m/s}^2\,(1\ \text{s}) = +15.9\ \text{m/s}$. We can then write the football's velocity at 1 s as $\vec{v} = (30.6, 15.9)\ \text{m/s}$.

(b) What characterizes the position of maximum height is that the y-velocity is instantaneously zero. We can solve for this height most simply by using the equation

$$v_y^2 = v_{0y}^2 - 2gy,$$

(that's Equation (3) of Table 3.1 with y substituted for x), then setting v_y to be zero and solving for the maximum height y_{max},

$$y_{\text{max}} = \frac{v_{0y}^2}{2g} = 33.7\ \text{m}.$$

An alternative method is to first find the time to reach this position (using $v_y = v_{oy} - gt = 0$), and then substitute this time into the equation for y (just below). Try it. At this position the football has only a horizontal velocity, the same as its initial horizontal velocity of 30.6 m/s.

(c) To find the time the football was in the air, we can take the equation for y

$$y = v_{0y}\,t - \frac{1}{2}\,gt^2,$$

and set $y = 0$ to solve for the times when the football is on the ground. As in our previous free-fall examples, there are two times when the ball is at $y = 0$: $t = 0$ (when it started out) and $t = \dfrac{2v_{0y}}{g} = 5.2\ \text{s}$.

SPECIAL CASE II: UNIFORM CIRCULAR MOTION

The special case of circular motion deserves our consideration because of the many important instances of such motion. Figure 5.8 shows a particle executing circular motion. (It is useful to put a reference point at the center of the circle and reckon all positions relative to it.) Velocity is a vector and vectors have both magnitude and direction. In circular motion the direction of the velocity vector (always tangent to the circle about which the particle travels) is constantly changing. Thus, even if the magnitude of the velocity remains constant (the case of so-called uniform circular motion), *there must be a nonzero acceleration*, because acceleration is the time rate of change of velocity.

What is the nature of this acceleration? We consider here the case of *uniform circular motion* where the speed of the particle traversing the circle remains constant. (See Figure 5.9, with the magnitudes

$$|\vec{v}_1| = |\vec{v}_2| = v.)$$

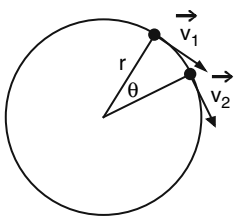


FIGURE 5.8 *A particle, shown at two different times, traveling in a circle.*

We examine the particle at two instants of time separated by the interval $\Delta t = t_2 - t_1$. In that time interval the particle has traveled a distance equal to $v\Delta t$ along the circle and has traveled through an angle θ.

In the same time interval, the particle's velocity vector, while maintaining a constant length, has also rotated through the same angle θ. (Do you see why? Hint: the velocity vectors are rigidly attached at right angles to their respective position vectors.). The triangle formed from $\vec{v}_1$ and $\vec{v}_2$ and their difference $\Delta\vec{v}$ and the one formed from the two position vectors $\vec{r}_1$ and $\vec{r}_2$ and the corresponding displacement vector $\Delta\vec{r}$ are similar, as seen in Figure 5.9. (Both are isosceles and both have the same included angle θ.) Because these triangles are similar, we can write

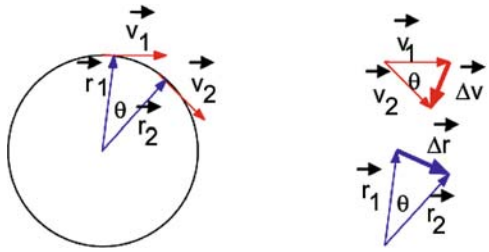


FIGURE 5.9 *Similar triangles formed from position and velocity vectors.*

$$\frac{|\Delta\vec{v}|}{|\vec{v}_1|} = \frac{|\Delta\vec{r}|}{|\vec{r}_1|},$$

or, since the magnitude of the velocity is v and the magnitude of the position vector is r,

$$|\Delta\vec{v}| = \frac{v}{r}|\Delta\vec{r}|.$$

Now, divide both sides of the latter equation by Δt and take the limit as Δt goes to zero. The left-hand side becomes the magnitude of the acceleration vector at any instant. The quantity

$$|\Delta\vec{r}|/\Delta t$$

on the right-hand side approaches the magnitude of the velocity vector at any instant; that is, it approaches the value v. Figure 5.9 suggests that as θ becomes smaller and smaller (as Δt approaches zero), the acceleration vector points more and more in toward the center of the circle, perpendicular to the velocity vector, which is always tangent to the circle. This acceleration is called *centripetal* (from a Greek word meaning "center-seeking") and we can express its magnitude as

$$a_{\text{cent}} = \lim_{\Delta t \to 0} \frac{|\Delta\vec{v}\,|}{\Delta t} = \frac{v^2}{r}. \tag{5.1}$$

Because the centripetal acceleration lies along the radius of the circle at any point it is sometimes referred to as the *radial acceleration.* As the particle travels around the circle at constant speed, it carries with it a velocity vector pointing tangent to the circle and an acceleration vector pointing radially inward. The velocity vector always has the same magnitude (in uniform circular motion) but its direction is constantly changing; the same is true for the centripetal acceleration. Because the acceleration direction is changing all the time, circular motion is not an example of constant acceleration, and the particle's position at any instant cannot be obtained by using kinematic equations for constant acceleration such as those found in Table 3.1.

Example 5.7 A protein molecule is spinning in an ultracentrifuge at 80,000 rpm at a fixed distance of 5 cm from the axis of rotation. Find the centripetal acceleration it experiences and express it in terms of a number of *g*s.

Solution: The protein travels in a circular trajectory of radius $r = 0.05$ m so that its velocity is

$$v = 8 \times 10^4 \frac{\text{rev}}{\text{min}} \cdot \frac{1\text{min}}{60\text{ s}} \cdot \frac{2\pi r}{\text{rev}} = 420 \text{ m/s}.$$

(*Continued*)

Then, from Equation (5.1), the centripetal acceleration of the protein is

$$a_{\text{cent}} = \frac{v^2}{r} = 3.5 \times 10^6 \text{ m/s}^2.$$

This acceleration is $3.5 \times 10^6/9.8 = 360{,}000$ times that of gravity, which we call "360,000 *g*s."

3. DYNAMICS

With the aid of vector analysis it is straightforward to generalize Newton's laws of motion and the ideas of work and energy to more complex situations in two or three dimensions. In this section we first show how vector equations make this generalization formally transparent, and then develop some problem-solving strategies to help in applying these ideas to understand a large variety of problems involving the translational motion of objects.

Newton's first and third laws require no further modification in leaping from one to two or three dimensions. The first law singles out a special single direction because objects traveling at constant velocity do so along a fixed direction. Similarly, the third law tells us that if an object exerts a force on a second object, this second object reciprocates with an equal but opposite reaction force acting back on the first object; these pairs of action–reaction forces are necessarily co-linear and in that sense the third law is a one-dimensional statement. We show below how the third law can be applied in studying the motion of various objects in more than one dimension.

Newton's second law, the key equation that relates the interactions acting on a body to the consequent motion, is a vector equation stating that the net vector force acting on an object divided by the mass of the object (a scalar) is equal to the vector acceleration:

$$\frac{\vec{F}_{\text{net}}}{m} = \vec{a}$$

Back in Chapter 2 when we introduced Newton's second law in one dimension (see Equation (2.9)), we anticipated this section by writing it in vector form even though vector algebra was not needed. Now that we understand how to combine vectors, we can simply add all the external forces acting on a body to obtain the vector resultant or net force. According to Newton's second law this net force divided by the mass of the object is equal to the vector acceleration that the object experiences. Let's see how to apply these ideas to a first example.

PROBLEM-SOLVING STRATEGY

1. The first step is to make a rough sketch of the problem, if there is not already one supplied as part of the problem, and to identify the object(s) whose motion is to be studied, if that is not clear.
2. The second step is to identify all the forces acting on the object (and only on that object) by constructing a carefully labeled external force diagram (such a diagram is sometimes known as a free-body diagram), a crucial step in solving the problem.
3. From the external force diagram, with a set of chosen coordinate axes, the next step is to write down the equations of motion, the component Newton's second law equations, being very careful to use appropriate labeling and to write down the x- and y-components in separate equations.
4. Once the equations of motion are obtained, solve for the unknowns of the problem, by performing the required algebra.
5. Whenever possible, check your results in limiting cases or in simplified circumstances.

Example 5.8 Let's return to the father and daughter ice skaters of Example 2.7. Suppose that the father skates backwards and holds his daughter's arms up at a 30° angle. Find the girl's acceleration, ignoring whatever friction there might be between her skates and the ice, if the man pulls with a force of 30 N.

Solution: We start with a diagram for the girl.

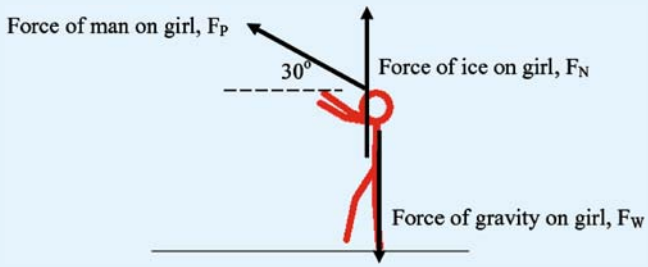


FIGURE 5.10 *Sketch for Example 5.8.*

There are three forces acting on the girl. In addition to her weight (the force of the gravitational pull of the Earth), there is the vertical force from the ice in contact with her skates, and there is the pull of her father on her directed upward at an angle of 30° with respect to the horizontal. To find the acceleration of the girl requires adding all forces on her as vectors. Each of the forces shown above has a vertical component and the man also exerts a force with a horizontal component. To add the vectors we can first add the vector components separately in the vertical and horizontal directions (with the proper signs included!).

We know that because the child glides in steady contact with the ice, she experiences no acceleration in the vertical direction and so the net force on her in the vertical direction must be zero. Adding together all components of force in the vertical direction (we take up to be positive, down negative) leads to

$$F_N + F_P \sin 30 - F_w = 0.$$

This equation is not needed to solve for the acceleration of the child, which is in the horizontal direction only, but it might be a useful part of a full analysis of the problem. For example, we can solve for the upward force exerted on the girl by the ice if we wanted to:

$$F_N = mg - F_P \sin 30 = 40 \text{ kg} \cdot 9.8 \text{ m/s}^2 - 30 \text{ N} \cdot 0.5 = 377 \text{ N}.$$

Clearly this force is reduced from the force the ice would have exerted in the absence of the father's upward pull F_P, which would have exactly equaled the child's weight. The father, in this case, is helping the ice support the girl's weight. So here's a question. Can the father actually lift the girl off the ice by applying a sufficiently large force at the same 30° angle? (Answer: Yes, but only if

$$F_p = \frac{mg}{\sin 30} = 2mg = 784 \text{ N}.$$

Do you see why? That would be a pretty strong father, because 784 N is 176 pounds!)

Back to the original question. In the horizontal direction there is only one component of force on the girl, the horizontal component of her father's pull: $F_P \cos 30$. According to Newton's second law, we have

$$a = \frac{F_P \cos 30}{m},$$

so that after substituting numbers for F_P and m, we find that $a = 0.65 \text{ m/s}^2$ pointing to the left.

Having completed our first multidimensional problem we note that there is a definite strategy in solving problems of this type and an awareness of the steps involved can be a great help in approaching new problems.

We conclude this section with three example problems (see box on page 106).

Example 5.9 A piano of mass 100 kg slides down a smooth (frictionless) ramp 5 m long inclined 15° with the horizontal. If the piano starts from rest, what is its speed at the bottom of the ramp?

Solution: We start with a rough sketch of the situation and an external force diagram.

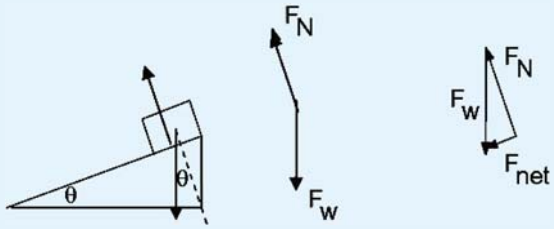


FIGURE 5.11 *Sketch, external force diagram, and net force acting.*

Only two forces act on the piano, gravity and the upward normal force of the ramp. Because the piano stays on the ramp, there is no motion (in particular, no acceleration) perpendicular to the ramp and so the net force perpendicular to the ramp must be zero. Thus the normal force must exactly cancel the component of the weight perpendicular to the ramp. The acute angle between the normal and the weight is the same as the ramp's angle of inclination (you should prove this), therefore we can write that

$$F_N = F_w \cos\theta,$$

where θ is 15°. The remaining component of the weight is the only unbalanced force and it produces a net acceleration down the ramp according to Newton's second law

$$a = \frac{F_{\text{net}}}{m} = \frac{F_w \sin\theta}{m}.$$

Because $F_w = mg$, we have that the piano's acceleration down the ramp is

$$a = g \sin\theta = 9.8 \sin 15 = 2.54 \text{ m/s}^2.$$

The form of this result should make sense because if $\theta = 0$, there is no acceleration and if $\theta = 90°$, we have free-fall. To find the velocity of the piano at the bottom of the ramp, assuming it starts from rest, we use the one-dimensional kinematic equation relating velocity, acceleration, and distance (Table 3.1) to find

$$v^2 = 2ax,$$

so that the velocity after traveling 5 m down the ramp is $v = \sqrt{2ax} = \sqrt{2 \cdot 2.54 \cdot 5} = 5.0$ m/s.

Example 5.10 Two crates of 30 kg and 20 kg mass are connected by a light (massless) rope while being pulled along a smooth (frictionless) floor by a horizontal force of 40 N applied to the heavier crate. Find the acceleration of each crate and the tension in the rope.

Solution: As usual, we begin by making a rough sketch and external force diagrams for each separate component of the system that has mass. The rope, having negligible mass, is not considered as an object, but simply as a means to transmit force.

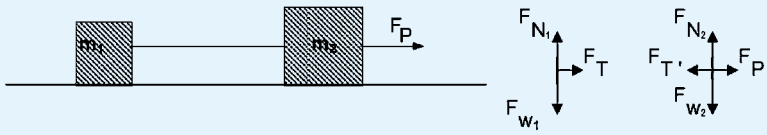


FIGURE 5.12 *Sketch and external force diagram for Example 5.10.*

The motion is one-dimensional and we really only need concern ourselves with writing Newton's second law for motion along the floor. We first note that F_T and $F_{T'}$ are, by Newton's third law, equal in magnitude. Writing one equation for each crate, we then have

$$F_P - F_T = m_2\,\mathrm{a} \quad \text{and} \quad F_T = m_1\,\mathrm{a},$$

where $m_1 = 20$ kg and $m_2 = 30$ kg, and we have explicitly used the fact that the two crates move together with the same acceleration as long as the rope is taut. Eliminating the tension force, we have that

$$a = \frac{F_P}{m_1 + m_2} = \frac{40\ \mathrm{N}}{50\ \mathrm{kg}} = 0.8\ \mathrm{m/s^2}.$$

Finally, the tension can be found by substituting into either of the Newton's second law expressions above to find

$$F_T = 20\ \mathrm{kg} \cdot 0.8\ \mathrm{m/s^2} = 16\ \mathrm{N}.$$

Note that the net force on each block is different, being ($F_P - F_T = 40 - 16 = 24$ N) on m_2 and 16 N on m_1. Treated as one composite object, the two crates have a total external force equal to the 40 N applied force and a total mass of 50 kg. The ratio of the net force to the total mass also gives the solution to this problem for the acceleration of either block.

Example 5.11 In a recently developed cell motility assay, a single myosin protein molecule can be seen to move along an actin protein filament stuck to the bottom of a petri dish. Actin and myosin are the major constituents of muscle and myosin can be pictured as a small molecular motor that uses chemical energy to produce mechanical force and subsequent motion. The force generated by a single myosin molecule has been measured to be about 5 pN (1 pN = 10^{-12} N). Idealize the situation to consider only the myosin molecule and the actin filament, ignoring the bathing fluid, and analyze the motion using Newton's laws. (Actually, to visualize the myosin molecule in a microscope, a ~1 μm radius plastic sphere is first chemically attached.)

(*Continued*)

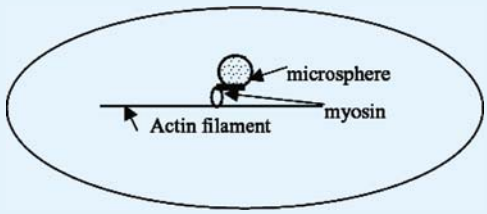


FIGURE 5.13 *Schematic of an actin filament and myosin molecule with plastic sphere attached. The drawing is not to scale; the microsphere is actually relatively much larger than the myosin whose head rotates to generate a force allowing it to move along the actin filament.*

Solution: We first need to compute the masses involved. If we assume that the density of the sphere is close to that of water ($\rho = 1000\ \text{kg/m}^3$), we can calculate the mass of the sphere to be $m = \rho\left(\frac{4}{3}\pi r^3\right) = 4 \times 10^{-15}$ kg (myosin's mass of 450 kD $= [4.5 \times 10^5][1.66 \times 10^{-27}] = 7.5 \times 10^{-22}$ kg is negligible compared to this). The actin filament is stuck to the petri dish and does not move. Given the force exerted by the myosin molecule on the actin, an equal and opposite force propels the (myosin + sphere) along the actin with an acceleration given by:

$$a_{\text{myo}} = \frac{F}{m} = \frac{5 \times 10^{-12}}{4 \times 10^{-15}} = 1.3 \times 10^3\ \text{m/s}^2.$$

If myosin with its plastic sphere accelerated at the rate found, then in 1 ms it should move a distance of about 0.5 mm (using $x = \frac{1}{2}at^2$). Direct measurements of the displacement show discrete steps of about 10 nm that occur in a single clock cycle of ATP hydrolysis, roughly 1 ms. Clearly our idealized problem has omitted the interactions with the surrounding solvent. These forces play a major role in determining the motion and account for the large discrepancy in calculated displacement.

4. WORK AND ENERGY

Work and energy are scalar quantities; therefore, at first glance, you might guess that in the "generalization theme" of this chapter to motion in more than one dimension these quantities are unaffected. This is not quite the case because, for example, as we saw in our one-dimensional analysis back in the previous chapter, work involves the product of a force and a displacement, both of which are vector quantities themselves. In this section we learn the general definition of work and kinetic energy. With these definitions, the work–energy theorem and conservation of energy law we learned in the previous chapter need no modifications but allow us to study a much broader array of multidimensional problems.

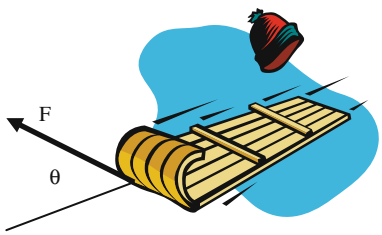


FIGURE 5.14 *A sled being pulled along the ice by a force F. Only the component of F along the ground does any work as the sled moves along.*

Let's return to the example at the beginning of the previous chapter of a sled of mass m being pulled along an icy (frictionless) surface by a constant force acting along a rope. If the rope is held at an angle θ above the horizontal (Figure 5.14), then the tension can be written as the vector sum of the horizontal x- and vertical y-components. The y-component of the tension, being vertical, cannot contribute to the motion along the x-direction. Its effect is to reduce the normal force of the

ground on the sled. Therefore, we need to modify our definition of the work done by a constant force, $W = F\Delta x$, in this more general case where the force is not necessarily along the direction of motion, because only the x-component of the tension will produce an acceleration along the x-direction. A more general definition of work, valid for all constant forces regardless of their direction, is

$$W_F = F_x \Delta x, \quad \text{(constant force)}, \tag{5.2}$$

or, in terms of the angle θ between the applied force and the displacement,

$$W_F = F\Delta x \cos\theta. \quad \text{(constant force)}. \tag{5.3}$$

Generalizing this to the case when a variable force acts on an object we can write that the work is given by

$$W_F = \sum \Delta W = \sum [F \cos\theta]_{\text{ave}} \, \delta x, \quad \text{(general definition)}, \tag{5.4}$$

where we have inserted cos θ into Equation (4.5). If several forces act on an object we simply add up the individual (scalar) contributions to the work, keeping track of their sign.

We defined kinetic energy as KE = 1/2 mv^2 for motion along one dimension. In more than one dimension there will be components of velocity along the different coordinate axes directions and the kinetic energy remains as originally defined as long as we remember that the square of the net velocity is given by $v^2 = v_x^2 + v_y^2$ $(+v_z^2)$ in two (or three) dimensions. The potential energy expressions we introduced in the previous chapter also are unaffected by the jump to higher dimensions because gravity and spring forces are basically one-dimensional, involving only vertical distances or the stretched distance along the spring axis, respectively.

In Section 5 of the previous chapter we introduced power as the rate at which work is done and derived an expression for it in the one-dimensional case, $P = Fv$. Because both force and velocity are vectors in two and three dimensions, we need to see how to generalize this expression for power as well. We saw in Equation (5.2) that for a constant force, it is only the component of force along the displacement that contributes to the work done by that force. Since the velocity is in the same direction as the displacement (from its definition as)

$$\vec{v} = \frac{\Delta \vec{r}}{\Delta t},$$

the power generated by a force can be written as

$$P = F_v v, \tag{5.5}$$

where it is only the component of the force along the velocity that does any work.

Given these modifications of the definitions of work and kinetic energy, the work–energy theorem and conservation of energy stand as presented in the previous chapter. This section concludes with two examples.

Example 5.12 Let's reanalyze Example 5.9 using energy ideas. Recall that a piano of 100 kg mass is sliding down a frictionless ramp 5 m long inclined at an angle of 15° starting from rest and the problem is to find its speed at the bottom.

(*Continued*)

Solution: In the example we found the velocity using Newton's laws and kinematics equations. Let's solve this problem in two ways using energy ideas: first using the work–energy theorem and second using conservation of energy. First, only the component of the weight acting down the inclined plane contributes to the work; this component is given by $mg \sin\theta$ (see Figure 5.15) and acts over a distance of $h/\sin\theta$. Therefore, the work done by gravity is simply equal to the product

$$W = mg \sin\theta \left(\frac{h}{\sin\theta}\right) = mgh.$$

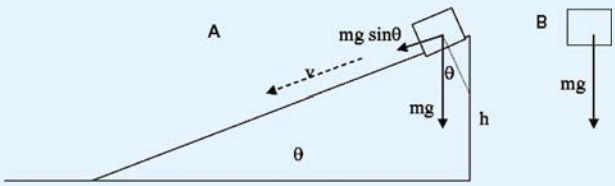


FIGURE 5.15 *A: A block sliding down an inclined plane; and B: the same block falling vertically.*

Setting this work equal to the change in kinetic energy (initially zero) we have that

$$mgh = 1/2\ mv^2,$$

so that we find the speed at the bottom to be

$$v = \sqrt{2gh} = 5.0 \text{ m/s}.$$

Second, using conservation of energy ideas, the piano starts from rest at height h with a total initial energy given by

$$E_i = mgh,$$

and ends up at the bottom of the ramp with only kinetic energy, because $h = 0$ at the bottom. Therefore, because total energy is conserved, we can write an identical equation as several lines ago, that $mgh = 1/2\ mv^2$, to find the same numerical result for the speed as well.

Which method is easier? Since we know the form for the gravitational potential energy (that it only depends on the height h) it was simpler to keep track of the total constant energy. Note that if the piano were to fall vertically through height h, the speed at the bottom would be the same (but of course the piano would be so much the worse!). The work done by gravity is given by the product of mg and h, and does not depend on the path taken by the object but simply on its weight and overall height change. On the other hand, the ramp is useful to steer the velocity of the piano.

Example 5.13 In a loop-the-loop roller coaster ride (Figure 5.16) the car of mass m starts from rest at point A at a height H. The loop-the-loop has a height of $H/3$. Assuming no friction, find: (a) the speed of the roller coaster car at point B at the top of the loop-the-loop and (b) the speed of the car at point C.

Solution: (a) The initial mechanical energy of the roller coaster at point A is completely gravitational potential energy mgH relative to a zero of potential

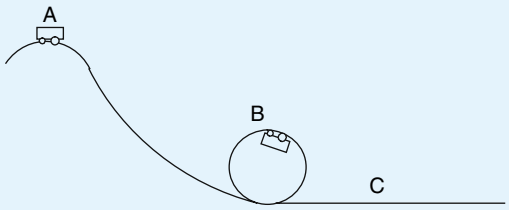


FIGURE 5.16 *A loop-the-loop roller coaster, showing the car at the start and upside down near the top of the loop.*

energy at the bottom. Because we are assuming that there is no friction, mechanical energy is conserved and the mechanical energy at point B must also be equal to mgH. But the energy at point B is actually partly gravitational potential and partly kinetic so that we can write

$$E_A = mgH = E_B = mg\frac{H}{3} + \frac{1}{2}mv_B^2,$$

where we have used the fact that the roller coaster is at a height of $H/3$ at B and has a velocity v_B. Solving this equation for the speed at B, we find that

$$v_B = \sqrt{\frac{2(mgh - \frac{1}{3}mgh)}{m}} = \sqrt{\frac{4\,gH}{3}}.$$

(b) At point C, there is no potential energy, so that the full initial mechanical energy is transformed into kinetic energy and we have

$$E_A = mgh = E_C = \frac{1}{2}mv_C^2.$$

Solving this for v_C, we have that

$$v_C = \sqrt{2\,gH},$$

an expression that should look somewhat familiar to you. This result tells us that the speed of the roller coaster at C is the same as it would be if the car just fell vertically through height H. Of course, the track has provided a softer "landing" for the car and steered it so it is traveling horizontally instead of falling, but the speed of the car is given by the free-fall result.

5. CONTACT FRICTIONAL FORCES

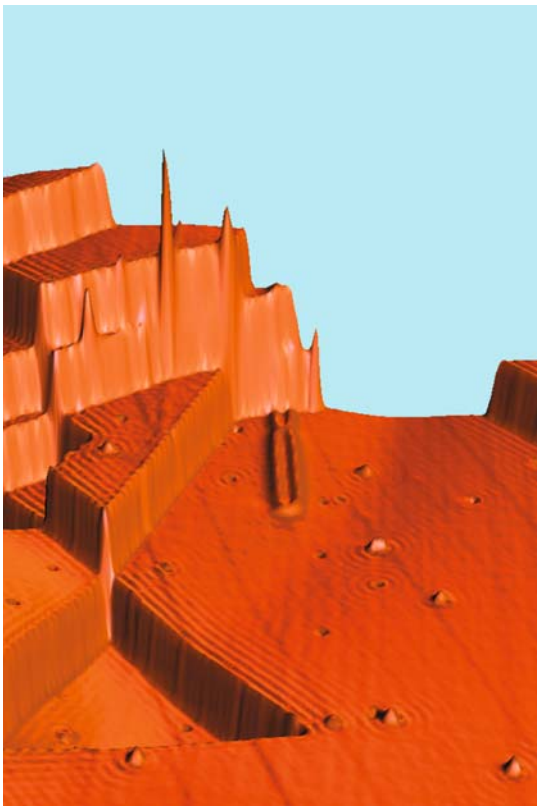
Up until now in our discussion of mechanics we have basically ignored one of the most common forces of our everyday experience, friction. Only in our discussion of the motion of an object in a fluid did we consider the resistive force of friction. In this section we discuss the common frictional force acting between two solid objects in contact with each other, as, for example, a book on a table surface. Only under certain very unusual circumstances can these contact frictional forces be neglected, circumstances such as motion on a smooth ice surface or on a cushion of air. Usually contact friction will be an important force and, in fact, friction is an essential force for most types of motion. Without it we would not be able to walk, automobiles would not be able to move, and even machinery would not be able to function (Figure 5.17).

FIGURE 5.17 *Friction is essential to lots of activities, human and otherwise.*

Imagine two solid objects sliding relative to each other, such as a block sliding on a table surface. Friction is the contact force acting parallel to the surface of contact (as contrasted with the normal force which is also a contact force but is directed perpendicular to the contact surface). It is produced by electromagnetic interactions between the molecules at the contacting surfaces of the two objects. On a microscopic scale, these surfaces are rough and irregular (Figure 5.18). Molecules at microscopic contact points bond together and as the block slides along the table these bonds constantly are broken and reform, thus slowing the block. Such a frictional force between moving objects is always in a direction to slow the motion and is called *sliding friction* or *kinetic friction.* The frictional force on a block is directed opposite to its velocity. It is found that although the frictional force depends on the nature of the material surfaces, surprisingly, it does not depend on the contact area (to a good approximation). The kinetic friction is proportional to the normal force F_N, and can be written as

$$F_{kfr} = \mu_k F_N, \tag{5.6}$$

FIGURE 5.18 *Microscopic irregularities on a smooth copper surface. The stripes are about 1.5 nm apart.*



where μ_k is the coefficient of kinetic friction, which depends on the two material surfaces. This is clearly not a vector equation because F_{kfr} is parallel and F_N is perpendicular to the surface. Often a point of confusion for the student, this equation should make sense in terms of magnitudes because the larger F_N is, the more contact between microscopically irregular surfaces and the greater the frictional force. This equation is an empirical approximate one and the coefficient of kinetic friction depends on the degree of smoothness of the surfaces, as well as on whether they are wet or lubricated. Clearly the relation is not a general law, but a useful approximate relation that emphasizes the fact that only the normal force and the nature of the materials are factors in determining the kinetic friction. The area of contact does not enter the equation, so that in our example of a block on a table, the block will experience the same frictional force sliding on any of its surfaces, regardless of their size. We are not able to calculate the kinetic friction from the fundamental principles of electromagnetic interactions.

Example 5.14 Let's again reconsider the problem of Example 5.9 in which a 100 kg piano slides down a ramp inclined at 15° with the horizontal, but suppose now that we include friction. If the coefficient of sliding friction is 0.2, find the acceleration of the piano down the ramp and its velocity at the bottom after sliding 5 m from rest.

Solution: Using the external force diagram, we can write Newton's second law for the two orthogonal directions, along the ramp and perpendicular to it. We have that

$$F_w \sin\theta - F_{kfr} = ma,$$

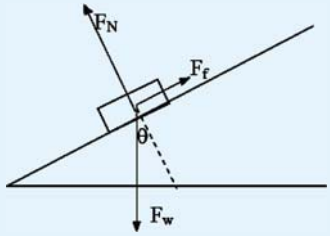


FIGURE 5.19 *External force diagram for Example 5.14.*

where a is the acceleration down the ramp and

$$F_N - F_w \cos\theta = 0.$$

In the second equation, the acceleration is zero because the piano only accelerates along the ramp and not perpendicular to it. We also need the relation for the kinetic friction force

$$F_{kfr} = \mu_k F_N.$$

Solving for F_N from the second equation to find $F_N = mg\cos\theta = 100 \cdot 9.8 \cdot \cos 15 = 950\ N$ we then find that $F_f = \mu_k F_N = 0.2 \cdot 950 = 190$ N. Substitution into the first equation then allows us to solve for the acceleration

$$a = (mg \sin\theta - F_{kfr})/m = (100 \cdot 9.8 \cdot \sin 15 - 190)/100 = 0.64 \text{ m/s}^2,$$

compared to the value of 2.54 m/s^2 found in the absence of friction in Example 5.9. Using this value for the acceleration, we can find the velocity of the piano after sliding 5 m to the bottom of the ramp:

$$\text{v} = \sqrt{2ax} = \sqrt{2 \cdot 0.64 \cdot 5} = 2.5 \text{ m/s}$$

compared to the value of 5 m/s found in the absence of friction.

We can also solve for the velocity of the piano at the bottom of the ramp using work–energy ideas. The total initial mechanical energy of the piano is entirely gravitational potential energy at the top of the ramp and is given by $E_i = mgh$, where $h = 5 \sin 15 = 1.29$ m. Similarly at the bottom of the ramp the total final mechanical energy is kinetic energy given by $E_f = 1/2\ mv^2$. Now, unlike the situation in the absence of friction for which mechanical energy is conserved, in the presence of friction the initial mechanical energy is reduced by the work of friction, which is negative, resulting in a decreased final mechanical energy. The work done by friction is always negative because frictional forces always act in a direction opposing the motion and therefore are directed opposite to the displacement. The work of friction is found from $W_{fr} = -F_f x = -\mu_k F_N x = -0.2(950)(5) = -950$ J. Our energy equation is given by $W_{fr} = E_f - E_i$. This is the work–kinetic energy theorem, where each mechanical energy term on the right is given by the sum of the kinetic and potential energies at one time of the problem. In our case we have that -950 J $= 1/2\ mv^2 - mgh$, and substituting in for the mass and height of the ramp we can solve for the final velocity of the piano at the bottom of the ramp, obtaining the same value as above.

When two objects are in contact, but at rest with respect to each other, there are also molecular bonds that form between contact points. Just sitting at rest does not result in any net force along a direction parallel to the contact surface; if there were, this force would spontaneously make the object accelerate. But if we try to push a block with a force directed along the table surface, the molecular bonds supply a frictional force in the opposite direction, opposing the impending motion. This type of friction is called static friction and arises in response to an applied force that would otherwise result in motion. As long as there is no motion, the static friction force is always as large as it has to be to cause a net balance of all forces on the block.

Imagine applying a force to our block on the table, starting with a small force and increasing its strength gradually. Until the molecular bonds are ruptured to allow motion, the static friction is exactly equal and opposite to the applied force and there is no motion. Once a threshold applied force is exceeded, the bonds then rupture and motion occurs. It is found that this maximum static friction force depends solely on the nature of the two surface materials and the normal force but not on the surface area, and is given by

$$F_{\text{sfr, max}} = \mu_s F_N, \tag{5.7}$$

where μ_s is the coefficient of static friction. Although this equation looks very similar to Equation (5.6), you need to keep the differences clearly in mind. This equation is for the maximum static friction force and holds only for impending motion. In general, the static friction force will be less than, or at most equal to $\mu_s F_N$:

$$F_{\text{sfr}} \leq \mu_s F_N. \tag{5.8}$$

Note again that these equations are not vector equations, but simply relations between magnitudes of forces, because the frictional forces are parallel to the contact surfaces, whereas the normal forces are perpendicular to those same surfaces. It is almost always the case that μ_s is greater than μ_k, a fact that agrees with our experience: it is easier to keep a heavy crate moving than it is to start its motion. Table 5.2 gives some values for coefficients of friction.

Table 5.2 *Static and Kinetic Coefficients of Friction*[*]

Object and Surface	μ_s	μ_k
Steel on steel (dry)	0.7	0.6
Steel on ice	0.03	0.02
Metal on metal (lubricated)	0.15	0.07
Rubber on concrete (dry)	1.0	0.9
Human joints (lubricated with synovial fluid)	0.005	0.005

[*]Values are approximate and vary greatly with the surface conditions.

Example 5.15 Two identical blocks of 20 kg mass are attached by a light cord going over a frictionless pulley at the edge of a tabletop with one block on the tabletop whereas the other is free to fall vertically as shown in Figure 5.20. The coefficients of static and kinetic friction between the table and the one block are 0.6 and 0.4, respectively. Analyze the motion to decide if the blocks move and, if so, find their acceleration to the left or right and the tension in the cord.

Solution: The external force diagrams for the two blocks are first drawn in Figure 5.20, being careful to label them appropriately.

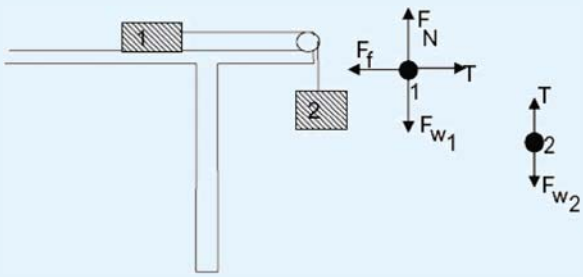


FIGURE 5.20 *Sketch and external force diagrams for Example 5.15.*

From the diagram we can write down the set of Newton's second law equations governing the motion:

$$F_N - F_{w_1} = 0$$

for the vertical forces on the block on the table because there is no vertical acceleration for block 1, and assuming for the moment that there is motion,

$$T - F_{kfr} = m_1 a \quad \text{and} \quad F_{w_2} - T = m_2 a$$

for the forces along the direction of motion for each block. Implicit in these last two equations is the fact that if motion occurs, the block on the table will move to the right, the rope will remain taut, and the tension force acting on each block is the same. To see whether the block on the table in fact moves, we need to find the maximum static friction force acting to the left and compare it to the net force pulling the block to the right when there is no motion; this force is just equal to the hanging weight $m_2 g$. That the tension force in the rope equals the weight of the hanging block when no motion occurs follows from the last equation above with $a = 0$. Using the first equation for the normal force, we can find that the maximum static friction force, given by $\mu_s F_N$, is equal to

$$F_{\text{sfr, max}} = \mu_s m_1 g = 120 \text{ N}.$$

Comparing this with the much larger value of $m_2 g = 196$ N implies that the block must move to the right; the maximum static friction is not enough to cancel the pull of the tension force to the right. Because the blocks do move, the appropriate frictional force is due to sliding friction. Returning to our equations and eliminating the tension from the two $F = ma$ equations (this can be done most easily by separately adding the left- and right-hand sides of the equations), we have that

$$m_2 g - F_{kfr} = (m_1 + m_2)a.$$

Substituting $\mu_k F_N$ for the friction force, we find that

$$m_2 g - \mu_k F_{N_1} = (m_1 + m_2)a$$

Finally, substituting $m_1 g$ for the normal force, and solving for a, we find

(*Continued*)

$$a = \frac{m_2 g - \mu_k m_1 g}{m_1 + m_2} = 2.9 \text{ m/s}^2.$$

We can find the tension in the cord by substituting this result for a into either of the $F = ma$ equations that have the tension force in them, resulting in a value of $T = 140$ N.

We intuitively believe that as the two surfaces in contact with each other are made smoother, the frictional force between them should decrease, and this is often the case. However, as two surfaces are made ultrasmooth, so that, even on a microscopic scale a substantial portion of the surfaces are in close contact, the frictional forces dramatically increase. This is due to the large increase in molecular bonds, or microwelds, that then form. At a microscopic level, computing the strength of the forces between surfaces is a formidable problem.

6. CIRCULAR MOTION DYNAMICS

Recall from earlier in this chapter that a particle traveling in a circle at constant speed has an acceleration directed toward the center of the circle, known as the centripetal acceleration. In order for a particle to travel in uniform circular motion a net force must be applied to it in the direction of the centripetal acceleration. This force, known as the *centripetal force*, might be supplied, for example, by a tension force due to a cord attached to the particle that is being swung in a circular trajectory. In the case of a car traveling along a circular exit ramp of a highway, the centripetal force is supplied by friction between the tires and the road (Figure 5.21). The term centripetal force is used for the net "center-directed" force, regardless of its origin, and is not a new type of force. An object traveling in uniform circular motion satisfies Newton's second law, but with an acceleration that is specifically equal to the centripetal acceleration

$$F_{\text{net}} = ma_{\text{cent}} = \frac{mv^2}{r}. \tag{5.9}$$

In uniform circular motion the net force must point toward the center of the circle. The key to analyzing uniform circular motion is to draw a careful external force diagram and to substitute the net inward radial force into Equation (5.9). Two examples should help to illustrate this method.

FIGURE 5.21 *A car going around a turn has a centripetal force F supplied by the tires.*

Example 5.16 A Ferris wheel of radius 20 m is rotating at 1.5 revolutions per minute. Find the forces exerted on an 80 kg man by his seat when he is at the top or at the bottom of the wheel as it rotates.

Solution: When he is at the top or bottom of his motion, the only two forces that act on the man are gravity and the upward push of the seat, as shown in the external force diagrams in Figure 5.22.

According to Newton's second law, at the top of the Ferris wheel we must have that

$$F_w - F_N = ma = m\frac{v^2}{r}$$

FIGURE 5.22 *Ferris wheel and external force diagram of man at top and bottom points of the circular path.*

where v and r are the velocity and radius of the circular Ferris wheel, and m is the man's mass. Because 1.5 rpm, for a 20 m radius wheel, gives a linear velocity of

$$v = \frac{1.5 \cdot 2\pi r}{60 \text{ s}} = 3.1 \text{ m/s}$$

we can solve for the normal force to find

$$F_N = m\left(g - \frac{v^2}{r}\right) = 80 \cdot \left(9.8 - \frac{3.1^2}{20}\right) = 746 \text{ N}.$$

At the bottom of the Ferris wheel, the external force diagram looks the same, but the normal force must be larger than the man's weight, because it must produce a net force in toward the center of the wheel. In this case, we write that

$$F_N - F_w = m\frac{v^2}{r},$$

so that at the bottom, the normal force is given by

$$F_N = m\left(g + \frac{v^2}{r}\right) = 80 \cdot \left(9.8 + \frac{3.1^2}{20}\right) = 822 \text{ N}.$$

The seat in which the man sits must supply this variable force in order to keep him orbiting in circular motion. At other points along the circular trajectory, the seat must supply the necessary centripetal force at an appropriate angle to the vertical. For example, at the two points along the axle height, the seat must supply the entire horizontal centripetal force as well as a vertical force to balance the man's weight as indicated in the sketch below.

Example 5.17 A car travels at constant speed around a circular highway exit ramp with a 200 m radius banked at a 3° angle. The roadway is sloped (see Figure 5.23a) so that when a car travels at a particular speed, the horizontal component of the normal is sufficient to provide the needed centripetal acceleration without any friction. What is the speed for which the exit ramp is designed?

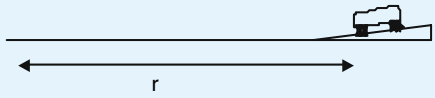


FIGURE 5.23A *Car on a banked circular highway exit ramp of radius r.*

Solution: The external force diagram for the car is complicated with five forces acting on the car (Figure 5.23b). In addition to its weight and the normal force there are frictional forces in two directions as well as a power driving force propelling the car forward. The forward propulsion force balances the rear frictional force (F_{f2}) so that the car travels at a constant speed.

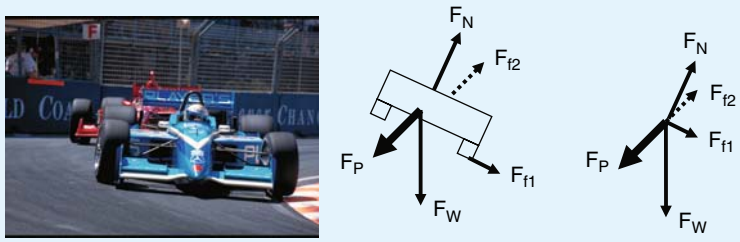


FIGURE 5.23B *External force diagram of a car on a banked roadway.*

In terms of the centripetal force needed to keep the car in its circular path, only the horizontal components of both the normal force and the sideways directed frictional force contribute. The banked road is designed so that a car traveling at the designated speed needs no sideways directed friction to travel the exit ramp. At that speed (and only that speed) the frictional force F_{f1} can be set equal to zero and we have that the horizontal component of the normal is equal to the centripetal force

$$F_N \sin 3° = m\frac{v^2}{r}.$$

The normal force has a vertical component just equal to the weight of the car, or

$$F_N \cos 3° = mg.$$

Eliminating the normal force from these two equations, we find that

$$\tan 3° = \frac{v^2}{rg},$$

so that the speed for which the road was designed is given by

$$v = \sqrt{rg \tan 3^\circ} = 10.1 \text{ m/s}^2 \quad \text{or} \quad 23 \text{ mph}.$$

Cars going around the exit ramp at higher speeds (needing greater centripetal acceleration) must have a frictional force whose inward horizontal component also contributes to the centripetal force needed to keep the car in its circular path or the car will veer outward. Similarly, cars traveling at a slower speed will need a frictional force directed radially outward to reduce the total centripetal force to the corresponding value of mv^2/r or the car will veer radially inward off the circular roadbed.

There is often some confusion about whether an object traveling in a circle has a force on it that is directed radially outward, often termed a "centrifugal force." This "force" seems to arise naturally from our experience. Clothes spinning in a clothes dryer fly outward against the drum; we "feel" an outward force on us as we sit in a car that makes a sharp inward turn; as we ride a roller coaster "around the world" we feel squashed down in our seats. These "centrifugal forces" are not caused by a real push or pull; they are not real forces. They are caused by trying to understand or by experiencing nature from an accelerating, or noninertial, frame of reference.

In reality, when sitting in a moving car, we tend to keep going in a straight line unless we are pulled to travel with the accelerating car as it makes a turn. An object dropped out the car window as the car makes a sharp turn will not fly radially outward as if it had a "centrifugal force" on it, but will move along a tangent to the initial path of the car, in accord with Newton's first law. Once dropped out of the window, there are no longer any horizontal forces acting and the object will maintain a constant horizontal velocity, disregarding any air friction, while accelerating (falling) vertically to the ground. Our bodies also follow Newton's laws and need a force to make them turn with the car. This force is supplied by a friction force between the seat and our bodies to keep us moving with the car as it turns; we seem to "feel" an outward directed force only because our body must supply the force needed to keep our upper torso sitting upright as the seat pulls us along with the car as it turns.

If a particle is traveling in a circle but also changing its speed then, in addition to a real centripetal acceleration, there will be an acceleration directed tangentially, along the velocity vector. In this case of nonuniform circular motion, the two components of acceleration, centripetal and tangential, vary as the particle moves along the circle (Figure 5.24). The centripetal acceleration is always directed toward the center of the circle and equal to v^2/r, but now also varies in magnitude as v changes. A nonzero tangential acceleration as a result of a tangentially applied force will result in a varying speed of the particle, and a consequently varying centripetal force.

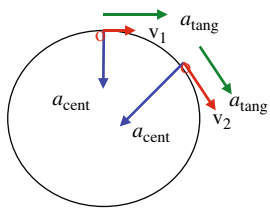


FIGURE 5.24 *An object in nonuniform circular motion. As the speed increases due to a tangential acceleration, so does the centripetal acceleration.*

Example 5.18 Consider a car exiting a highway on a circular 3° banked exit ramp (Figure 5.25). If the car enters the ramp at 65 mph and slows to 35 mph at the end of the quarter-circular 200 m radius ramp with a constant deceleration, find the magnitude of the net acceleration of the car at the beginning and end of the ramp.

(*Continued*)

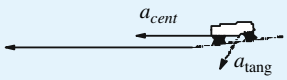


FIGURE 5.25 *The components of acceleration of a decelerating car traveling along a banked roadway, with velocity opposite in direction to a_{tang}, both oriented perpendicular to the page.*

Solution: First, we find the constant tangential acceleration of the car. Using the one-dimensional kinematics equation when the acceleration is constant, $v^2 = v_0^2 + 2ax$ with $x = \pi r/2$ corresponding to one-quarter of a circle, and converting the velocities to m/s (1 mph = 0.447 m/s), we have

$$a = \frac{v^2 - v_0^2}{2x} = -0.96 \text{ m/s}^2$$

where the negative sign indicates a deceleration. Although this tangential acceleration is constant on the ramp, the centripetal acceleration varies as the speed varies, from

$$a_{\text{cent}} = \frac{v^2}{r} = 4.2 \text{ m/s}^2$$

at 65 mph to $a_{\text{cent}} = 1.2$ m/s^2 at 35 mph. Therefore, combining these two orthogonal accelerations using the Pythagorean theorem, we find the net acceleration at each location: $a_{\text{net}} = \sqrt{a_{\text{cent}}^2 + a_{\text{tang}}^2} = 4.3$ m/s^2 at the start and 1.5 m/s^2 at the end of the ramp.

7. CENTRIFUGATION

Sedimentation refers to the process by which particles in a fluid settle to the bottom under the influence of gravity. Microscopic particles or macromolecules that normally remain in suspension due to thermal collisions with solvent molecules can be made to sediment under the influence of additional external forces. A number of types of external forces have been used to speed up sedimentation, including electrical and magnetic forces. Here we discuss the most common method used, centrifugation, to artificially increase gravity in order to sediment suspended objects.

Let us imagine a centrifuge tube, containing a solution of proteins, spinning about a vertical axis in a centrifuge (Figure 5.26).

The path of a protein is basically circular as the centrifuge tube spins, with a very small drift velocity v_r outward (or radial) toward the bottom of the tube. The protein does not fall vertically because of its microscopic size and the collisional forces from the solvent that keep it suspended. If we analyze the horizontal forces acting on the protein, treating it as a particle, there are two forces that provide the net centripetal force required to produce circular motion. These are the buoyant and frictional forces acting in response to the protein's slow drift velocity. The frictional force has a magnitude $F_f = fv$, taken from Equation (3.6), and acts in the direction opposite to the drift velocity or toward the center of the (nearly) circular trajectory. Arising from the increasing pressure in the solvent with increasing depth in the tube, the buoyant force also points toward the center of the circle. This pressure variation is due to the fluid deeper in the tube (closer to the tube bottom) having to support the fluid farther out near the top of the tube and maintain its circular motion. At any instant the fluid near the bottom of the tube, if not constrained by the tube would fly off tangentially. The tube bottom is being driven

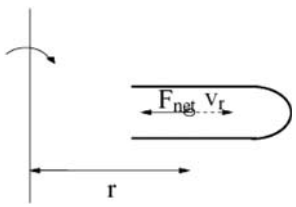


FIGURE 5.26 *A side view of a horizontal centrifuge tube showing a sedimenting molecule with a net inward force acting on it.*

against the fluid to provide the centripetal force to steer it around in a circle, just as in a clothes dryer the walls push the clothes radially inward to keep them traveling around in a circular path within the dryer. Fluid farther up the tube is given its centripetal force by the fluid nearer the bottom of the tube. We show in Chapter 8 that this pressure variation gives rise to a buoyant force.

Newton's second law in the radial direction is then

$$fv_r + F_B = ma_{\text{cent}}, \tag{5.10}$$

where the buoyant force is, as we study in more detail in Chapter 8, equal to the effective weight of the displaced fluid. In this case the weight of the displaced fluid is not due to an acceleration g downward, but rather to a_{cent} directed outward along the centrifuge tube, so that $F_B = m_0 a_{\text{cent}}$, with m_0 equal to the mass of the displaced water. Remember that v_r is the speed of the protein as it moves radially outward along the length of the tube. Substituting the expression for the buoyant force and solving for the ratio of the protein sedimentation velocity to its acceleration, known as the sedimentation coefficient s, we find

$$s = \frac{v_r}{a_{\text{cent}}} = \frac{(m - m_0)}{f}. \tag{5.11}$$

The sedimentation coefficient has units of seconds, from the ratio of a velocity to an acceleration, but because typical values are on the order of 10^{-13} s, we define the Svedberg (S), with 1 S = 10^{-13} s, and use it as a fundamental unit for sedimentation coefficients. Table 5.3 lists some sedimentation coefficients of biological materials, together with the times required to sediment them at various accelerations measured in multiples of g. Sedimentation coefficients are seen to depend on the particle mass, frictional properties and also the fluid density (through the term m_0), and are often used to characterize macromolecules; indeed many are named simply by their sedimentation coefficients such as the 30 S and 50 S ribosomes.

Today's ultracentrifuges routinely attain rotational speeds of over 75,000 rpm, representing accelerations of several million g's. Spinning solutions at these speeds allows the "pelleting" of even soluble proteins at the bottom of the centrifuge tube after hours of spinning. Every laboratory that studies biomolecules or cells is equipped with centrifuges for the preparation, and often for the characterization, of materials. Figure 5.27 shows a typical ultracentrifuge and a "rotor" that is used to hold the sample tubes. The figure also shows the results of an accident in which the extremely high energies involved in spinning the rotor at high speeds led to the destruction of a centrifuge.

How is the sedimentation coefficient determined experimentally?

From the definition of s we can write,

$$v_r = dr/dt = s\, a_{\text{cent}}.$$

Now,

$$a_{\text{cent}} = \frac{v^2}{r}$$

where v means the speed of the protein as it orbits around its circular path (and not the radial drift speed). We can write

$$v = \frac{2\pi r}{T},$$

where T is the time to complete one revolution. Thus, $dr/r = s\omega^2 dt$, where $\omega = 2\pi/T$, which can be integrated from r_0 (at time t_0) to r (at time t) to yield

$$\ln[r(t)] = \ln[r_0(t_0)] + \omega^2 s(t - t_0).$$

This equation is the basis for determining the sedimentation coefficient from a series of measurements of the boundary between the solution and the pure solvent, $r(t)$, as it moves down the centrifuge tube. A plot of $\ln[r(t)]$ as a function of $(t - t_0)$ should be a straight line with a slope of $\omega^2 s$, and because ω is known s can be found.

TABLE 5.3 *Typical Sedimentation Coefficients, Accelerations, and Corresponding Approximate Times Needed to Spin Down a Sample in a Centrifuge Tube*

Sample	*Sed. Coeff (S)*	*No. gs to Pellet*	*Time to Pellet*
Whole cells	10^6	100	10 min
Cell nuclei	10^5	700	10 min
Mitochondria	10^4	7000	10 min
Ribosomes	30, 50 S	100,000	2 h
Soluble proteins	1 – 5 (globular) 5 – 20 (elongated)	500,000	hours

FIGURE 5.27 (upper) *Modern ultracentrifuge;* (middle) *fixed angle rotor;* (bottom) *centrifuge remains after accident in which a rotor exploded while spinning.*

CHAPTER SUMMARY

This chapter generalizes our description and analysis of motion to more than one spatial dimension. The kinematical equations of Table 3.1 are generalized in a straightforward way using vector analysis, so that, for example, for free-fall along the vertical y-direction:

$$\vec{a} = (0, -g),$$

$$\vec{v} = (v_{0x}, [v_{0y} - gt]),$$

$$\text{and } \vec{r} = ([x_0 + v_{0x}\,t], [y_0 + v_{0y}\,t - \frac{1}{2}\,gt^2\,]),$$

where the parenthesis notation $\vec{A}_x = (A_x, A_y)$ indicates the x- and y-components of the vector $\vec{A}$.

An object moving in a circle has a centripetal acceleration given in magnitude by

$$a_{\text{cent}} = \frac{v^2}{r}, \qquad (5.1)$$

and a centripetal force acting on it given, in magnitude, by

$$F_{\text{net}} = ma_{\text{cent}} = \frac{mv^2}{r}. \qquad (5.9)$$

Newton's second law is generalized using vector analysis and can be written in a transparent form $\vec{F}_{\text{net}} = m\vec{a}$, meaning that only the net force in a particular direction, say x, will act to produce an acceleration along the x-direction.

The work done by a force pointing in any direction on an object moving along the x-axis is defined by

$$W_F = \Sigma \Delta W = \Sigma [F\cos\theta]_{\text{ave}} \delta x, \qquad (5.4)$$

where $F\cos\theta$ gives the component of the force along the x-direction and the summation allows for a variable force to be considered constant over short intervals of distance δx (see the discussion of Equation (4.5) as well).

Friction can be empirically described in the two cases of sliding (kinetic) motion and of static impending motion by

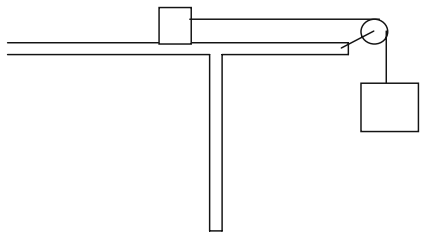
$$F_{kfr} = \mu_k F_N, \qquad (5.6)$$

$$F_{sfr} \leq \mu_s F_N. \qquad (5.8)$$

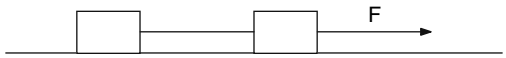
It is particularly important to see the examples worked out in this chapter and to practice doing problems in order to appreciate the awesome power of Newton's laws coupled with vector analysis.

QUESTIONS

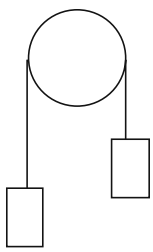
1. If we chose to orient our x- and y-axes at 45° to the vertical rather than horizontal and vertical, write down the corresponding equations of motion (analogous to Table 3.1) along the x- and y-axes.
2. Name as many physical quantities as you can that you believe to be vector in nature. Now name as many that you believe to be scalar in nature. Compare your lists with those of your classmates. Attempt to resolve any differences by challenging each other's reasoning and supporting evidence. (Just in case you overlook them, consider in your list: temperature, weight, volume of an object, and density.)
3. A vector quantity has both magnitude and direction. If the measurement of a particular physical property requires the use of signed numbers (i.e., both positive and negative numbers) is the property necessarily a vector?
4. Describe a coordinate system useful for detailing the position of an object within the field of view of a microscope.
5. Is time a vector? If so, what is its direction? What does time measure? Is there any meaning to "negative time"? Could you tell if something were moving backwards in time? If all time everywhere slowed down or speeded up, would there be any way to detect it?
6. Can you add a vector and a scalar (in any way that is useful or makes physical sense)? What about multiplying together a scalar and a vector? What about multiplying two vectors together? What sort of possible complications or ambiguities might arise with such operations?
7. Show how you can add three vectors together, all of which have the same magnitude, and end up with a zero result. Can this sort of "vector addition to zero" work with any number of vectors?
8. What is the relationship between a vector and a coordinate system? Between a vector and the number line? What properties or values of a vector depend on the coordinate system used to express it?
9. Give examples of two objects that have different positions but undergo identical displacements. (Hint: Think of a group of choreographed stage dancers.)
10. Does a vector of zero magnitude have a direction?
11. Compare the driving patterns of a single typical day for a local delivery truck, and a long-distance freight truckdriver. Compare instantaneous velocity, average velocity, presence or absence of acceleration (constant velocity or not), net displacement, and distance logged on the odometer.
12. Two marbles sitting on a tabletop are flicked off, one just falling vertically and the other shot out horizontally off the table. Which one hits the ground first?
13. The string on a yoyo breaks while doing an "around the world" just as the yoyo is at the top of its orbit. What happens?
14. If a protein in a centrifuge feels a centripetal force directed in toward the axis of rotation, why does it slowly migrate radially outward and not radially inward?
15. Sketch (nonartistic) external force diagrams for each of the following, showing all the forces acting on the object.
    (a) A high jumper clearing the highbar
    (b) A canoe being paddled along
    (c) A boy riding on an escalator
    (d) A jet airplane cruising at a constant speed
    (e) A lead weight sinking in the ocean
16. Two blocks, one sitting on a table and the other heavier one hanging over its edge, are connected by a light string as shown in the figure. Which force makes the block on the table move, the tension in the string or the weight of the hanging block? Are these two forces equal?



17. Two blocks of equal mass sit on a tabletop and are connected by a light string. A second string is pulled with a force F as shown in the figure. If you draw an external force diagram and do some thinking you will see that the tension force that pulls the left block is $F/2$. Why does the right block of equal mass need a force F to pull it at the same acceleration?

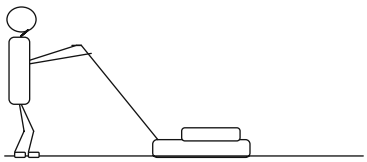


18. Two blocks, each of mass m and connected by a light string, hang over a frictionless pulley at rest as shown in the figure. Why do the blocks remain at rest even though there is a net downward force due to gravity of $2\,mg$?



19. A man is mowing his lawn by pushing on the handle of a push lawnmower (see the figure). Why is the upward normal force on the mower from the ground

not equal to the weight of the mower? To what is it equal?



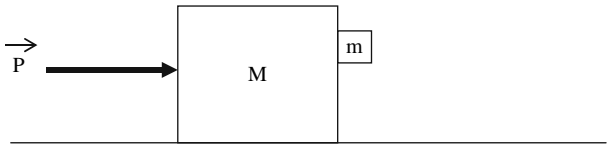
20. Why is it usually easier to keep a heavy object moving than to start it moving from rest?

21. Is the work done by friction always negative? Give an example to illustrate your answer.

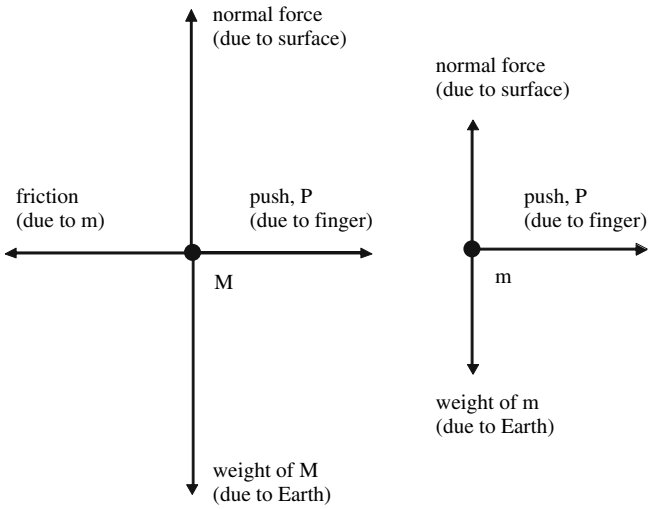
22. It is an apparent paradox that making two surfaces smoother and smoother will eventually increase the frictional force between them. Why is this true?

23. Explain clearly in words why the work done by friction cannot be expressed as a difference in a potential energy function, as can be done, for example, for the work done by gravity.

24. A large cube of mass M is accelerated across a level frictionless surface by a finger applying a constant horizontal push $\vec{P}$. A small cube of mass m is held in place on the front face of the large cube by static friction, as shown in the figure.



A student is asked to draw external force (free body) diagrams for the two masses in this problem. Each force is given a descriptive name and the object that causes each force is identified. The student's diagrams are shown below. Please make whatever alterations are necessary to make these diagrams correct. You may add or delete forces, change the sizes of the forces shown (so that accelerations are qualitatively correct), and change the labels to more accurately identify what the force is and from where it comes.



25. For the previous question, which of the following, if any, are true? Circle the letter of any true statement.

(a) The surface over which the large cube slides exerts a force on the large cube parallel to the surface with magnitude equal to P.

(b) The surface over which the large cube slides exerts a force on the large cube parallel to the surface with magnitude equal to Mg.

(c) The surface over which the large cube slides exerts an upward vertical force on the large cube with magnitude equal to P.

(d) The surface over which the large cube slides exerts an upward vertical force on the large cube with magnitude equal to Mg.

(e) The surface over which the large cube slides exerts an upward vertical force on the large cube with magnitude equal to $(M + m)g$.

(f) The large cube exerts a force in the horizontal direction on the small cube with magnitude equal to P.

(g) The large cube exerts a force in the horizontal direction on the small cube with magnitude less than P.

(h) The large cube exerts a force in the horizontal direction on the small cube with magnitude greater than P.

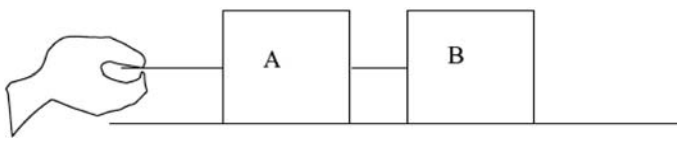
(i) The large cube exerts a force in the horizontal direction on the small cube with magnitude equal to mg.

(j) The large cube exerts an upward vertical force on the small cube with magnitude equal to P.

(k) The large cube exerts an upward vertical force on the small cube with magnitude equal to mg.

26. Is it possible to have a heavy crate slide up an inclined plane and have it come to rest at its highest point without sliding back down? Why or why not? If possible, what conditions would have to be met for this to happen?

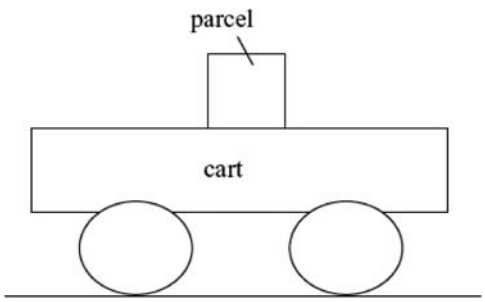
27. Two blocks, each weighing 10 N and connected by massless strings, are pulled across a horizontal table at constant speed, as shown in the figure. The force of kinetic friction on each block is 5 N. Draw an external force diagram for block A. In the diagram label each force, identify what body causes it, and make sure the forces have the correct relative magnitudes.



28. According to the definition of work, the work done by an external force in moving a heavy crate along a horizontal surface should be the same whether the force

is pulling upward at a 45° angle or pushing downward at a 45° angle. In practice it is easier to pull the crate than to push it at 45°. Why is this so?

29. A cart carries a parcel as shown in the figure to the right. The parcel is not lashed down. The mass of the parcel is M and the mass of the cart is 5 M. The cart is traveling to the right and is slowing down. As the cart slows, the parcel doesn't slip over the surface of the cart. Draw external force diagrams for the parcel and for the cart, labeling each force and the body that is responsible for the force. The relative sizes of the forces should be qualitatively correct.



30. Two blocks are attached by a light string with one resting on a rough table and the other hanging over the edge via a frictionless pulley as shown in the figure for Question 16 above. If the blocks are initially at rest, the tension force is equal to the hanging weight. As the weight of the hanging block is increased, eventually the blocks will move. At that point is the tension in the string more, the same, or less than the weight of the hanging block? If the hanging weight were continuously increased, would the tension force change gradually or abruptly when the blocks move?

31. In the Ferris wheel example (Example 5.16), can the normal force of the seat on the man ever be zero? If so, find an equation for the required velocity of the man for this to occur.

32. Why is a high-speed curved roadway banked? If a car goes around such a curve with too rapid a velocity, in which direction must a frictional force act on the tires of the car to keep it on the road? If a car goes around such a curve too slowly in which direction must the frictional force act?

33. Why do you feel a "centrifugal force" directed radially outward when you ride in a car and make a sharp inward turn? Is this a real force? What is the origin of the centripetal force on the car? On you in the car?

34. A girl does an around-the-world with a yo-yo. Which of the following vectors for the yo-yo are along the string direction: the velocity, the centripetal acceleration, the displacement for a one-half revolution, and the tangential acceleration?

35. For an object undergoing circular motion, assuming all other variables to be constant, fill in the blanks with "increases", "decreases", or "remains the same":
(a) As the object speeds up, the magnitude of the centripetal acceleration _____.
(b) When the object has a constant negative tangential acceleration, the centripetal acceleration magnitude _____.
(c) When the object has no tangential acceleration, the centripetal acceleration magnitude _____.

36. For an object in circular motion, state whether the following are true or false.
(a) The velocity is always perpendicular to the centripetal acceleration.
(b) With the circle center as origin, the displacement is always perpendicular to the velocity.
(c) Because the velocity is not constant, there is always a tangential acceleration.
(d) The net acceleration can never point outside the circular orbit.

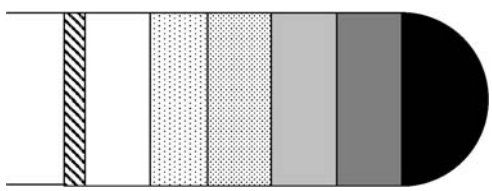
37. Small enough particles will not sediment in a glass of water even if their density is greater than that of water. Why don't all particles that are denser than water, regardless of size, sediment?

38. Which will sediment faster in a centrifuge: a 30 S ribosome spinning at 10^5 g's or a 50 S ribosome spinning at 50,000 g's?

39. A particle is traveling in uniform circular motion about a circle of radius r with speed v. Write a vector expression for its acceleration at any point in terms of its angle from the x-axis, which goes through the circle center. Use ordered pair notation.

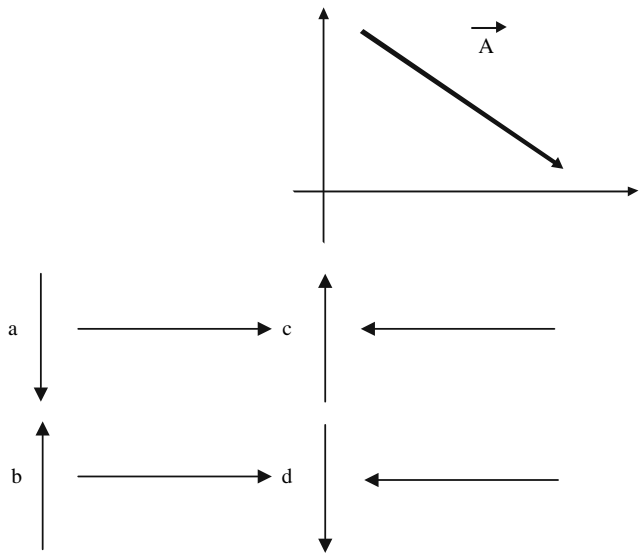
40. If, as a particle executes uniform circular motion in the x–y plane, the particle also has a constant speed along the z-axis, describe its trajectory in words and write a vector equation for its position using order triplet notation.

41. One variation of centrifugation uses a solvent mixture (typically an aqueous sucrose solution of varying concentration) with an increasing density with depth along the centrifuge tube. The sample to be studied is layered on the top of the tube and the tube is spun so that it lies horizontally (in a swinging bucket rotor; see the figure). Known as density gradient centrifugation, what do you expect to happen if the density range includes the density of the sample macromolecules? (Hint: Consider Equation (5.10).)



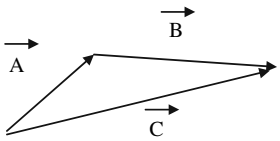
MULTIPLE CHOICE QUESTIONS

1. The figure shows a vector $\vec{A}$ and two coordinate axes. The components of the vector $(-1)\vec{A}$ along these axes are most likely



2. One force vector has components (5 N, −3 N) and a second has components (−2 N, 2 N). These forces produce a net force with scalar components (a) (−10 N, −6 N), (b) (7 N, −5 N), (c) (−5 N, 7 N), (d) (3 N, −1 N).

3. Vectors $\vec{A}$, $\vec{B}$, and $\vec{C}$ are related to each other as shown. The magnitude $A = 3$ and the magnitude $B = 4$. The magnitude C must be between (a) 1 and 7, (b) 5 and 7, (c) −7 and 1, (d) 1 and 5.



4. The magnitude of the force vector with components (5 N, −5 N) is (a) 0.0, (b) 5.0, (c) 7.1, (d) 25.0 N.

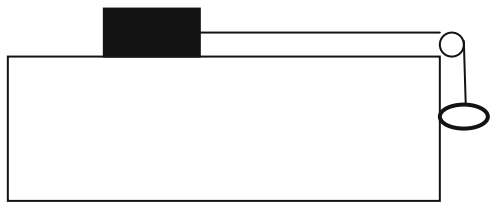
5. A force $\vec{F}_1$ has x-component +5 N, and y-component +2 N. A second force $\vec{F}_2$ has x-component −3 N, and y-component −3 N. The x- and y-components, respectively, of $\vec{F}_1 - \vec{F}_2$ are (a) 8 N, 5 N, (b) 7 N, −6 N, (c) 3 N, −1 N, (d) 3.16 N, 18.4° below the positive x-axis.

6. The velocity of a particle at one instant has an x-component of +30 m/s and a y-component of −40 m/s. Given that the instantaneous speed is the magnitude of the instantaneous velocity, what is the particle's instantaneous speed? (a) 10 m/s, (b) 50 m/s, (c) 70 m/s, (d) 2,500 m/s.

7. A girl is riding on the outer edge of a merry-go-round with a streamer pulling a rubber ball attached by a string. If the string breaks, as seen by someone on the ground the ball will (a) fall vertically down, (b) fly radially outward from the merry-go-round, falling vertically as it goes, (c) fall vertically while traveling tangentially forward from the merry-go-round, (d) fall vertically while traveling tangentially backward from the merry-go-round.

8. A ball is attached to a string and spun in a circle in a horizontal plane. The physical forces acting on the ball include its (a) weight and the centrifugal force, (b) weight and the tension force, (c) weight and the centripetal force, (d) weight and the force of the hand holding the string.

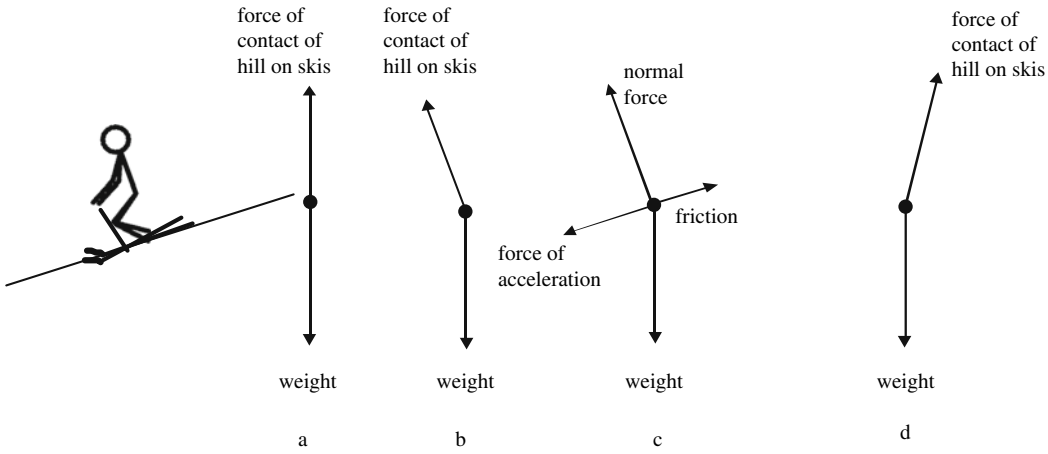
9. A 1000 kg block sits on a frictionless table, connected by a massless rope over a frictionless pulley to a 0.01 kg washer hanging off the edge of a table. The magnitude of the acceleration of the washer will be (a) 0 m/s^2, (b) 0 m/s^2 $< a <$ 9.8 m/s^2, (c) 9.8 m/s^2, (d) $a >$ 9.8 m/s^2.



10. In the previous question, the magnitude of the acceleration of the block will be (a) 0 m/s^2, (b) 0 m/s^2 $< a <$ 9.8 m/s^2, (c) 9.8 m/s^2, (d) $a >$ 9.8 m/s^2.

11. A wrench is dropped from rest from the top of the mast of a sailboat traveling (forward) at 10 m/s in still water. Ignoring air resistance and assuming the mast is vertical, the wrench hits the deck (a) directly next to the mast, (b) some distance away from the mast to the rear of the boat, (c) some distance away from the mast to the front of the boat, (d) in an unpredictable place because there is insufficient information.

12. A skier skis in a straight line down a hill gradually picking up speed as he goes. Which of the following could plausibly be an external force diagram for the skier during this motion? Assume air resistance is negligible.



Questions 13 and 14 refer to a block pulled up an inclined plane, with inclination angle θ, by a constant force F applied at an angle Φ measured from the inclined plane.

13. The work done by the force F in sliding the block a distance d along the incline is (a) $Fd \sin \Phi$, (b) $Fd \cos \theta$, (c) $Fd \sin \theta$, (d) $Fd \cos \Phi$.

14. The magnitude of the work done by gravity for the same motion is given by (a) $mgd \sin \theta$, (b) $mgd \sin \Phi$, (c) mgd, (d) $mgd \cos \theta$.

15. Suppose a ball is thrown with an initial velocity of 8 m/s at a 60° angle above the horizontal and stays in the air for 1.1 s. How far (in m) will it travel in the horizontal direction? (a) 7.6, (b) 8.8, (c) 4.4, (d) 5.9, (e) 10.3.

16. A yo-yo with mass M is spun in a loop-the-loop of radius R at a constant speed v. The tension in the string is T. What is the centripetal force on the yo-yo when at the bottom of its trajectory? (a) $T - Mg$, (b) $Mg - T$, (c) $Mg + T$, (d) $T - Mg + Mv^2/R$, (e) none of these.

17. Two identical blocks of mass m are tied together (by a light cord) and pulled up a rough inclined plane at constant speed by a pulling force F directed along the incline and applied to the upper block. Which of the following statements is true?
(a) The work done by F is zero because the blocks move at constant speed.
(b) The total friction force must equal F because the blocks move at constant speed.
(c) The tension in the cord is F because the two blocks are identical.
(d) The work done by F is equal in magnitude to the work done by gravity plus the work done by friction.
(e) None of the above is true.

18. For the previous problem, a free-body diagram of the lower block would include all of the following forces except
(a) mg down
(b) T up along the incline
(c) F up along the incline
(d) Friction down along the incline
(e) Normal force perpendicular up from the surface

19. Two identical blocks of mass m are tied together by a light cord. One sits on a horizontal frictionless surface and the other one hangs over a frictionless light pulley and is held in place. When released from rest, the hanging block falls a distance d. Which of the following is a true statement?
(a) The tension in the rope is equal to mg.
(b) The work done by gravity on the hanging mass is equal to the gain in KE of the block on the frictionless surface.
(c) The work done by the tension in the cord equals the gain in KE of both blocks.
(d) The tension in the rope plus the normal force on the block on the horizontal surface adds up to mg.
(e) The work done by gravity on the hanging block is equal to the gain in KE of both blocks.

20. A ball attached to a string is spun around in a horizontal circle. If the string is cut quickly at an instant of time, the ball's initial velocity points
(a) Radially outward because the ball felt a centrifugal force
(b) Radially inward because the string exerted a centripetal force
(c) Vertically downward because of its weight
(d) Tangentially because of Newton's first law
(e) Somewhere between radially outward and tangentially depending on its speed

21. In a frictionless roller coaster, if the car starts from rest at a height equal to twice that of the loop-the-loop portion, the speed at the top of the loop (point A) can be found by (a) equating the initial potential energy to the kinetic energy at point A, (b) by equating the initial kinetic energy to the sum of the potential and kinetic energy at point A, (c) by equating half the initial potential energy to the kinetic energy at point A,

(d) by equating half the initial potential energy to the sum of the potential and kinetic energies at point A.

22. The net work done by all the forces in sliding a crate from rest up an inclined plane, coming to rest at the top (a) is always zero because there is no change in kinetic energy, (b) is nonzero and depends on the height of the plane, (c) is nonzero but depends on the path up the incline as well as its height because the work done by friction depends on the path, (d) is nonzero but depends on the details of the force applied by the person as well as the factors of part (c), (e) none of the above.

23. In the loop-the-loop demonstration in which a small cart rolls around the looped track, the condition that needs to be satisfied for the cart to just get around the loop is (a) the starting potential energy must equal that at the top of the loop, (b) the kinetic energy at the top of the loop is just equal to zero, (c) the normal force at the top is just equal to zero, (d) the kinetic energy at the top is just equal to the weight of the cart, (e) none of the above.

24. A block of mass m slides down an inclined plane (angle of inclination θ) a distance d along the plane. If the block slides down at constant velocity, the work done by friction is given by (a) $mg \sin \theta\, d$, (b) mgd, (c) $-mgd$, (d) $-mg \sin \theta\, d$, (e) cannot be determined from what is given.

25. A block is given a push up an inclined plane. During its round-trip motion the frictional force is (a) always directed upward, (b) always directed downward, (c) directed upward till it reaches its maximum height and then directed downward, (d) directed downward until it reaches its maximum height and then directed upward.

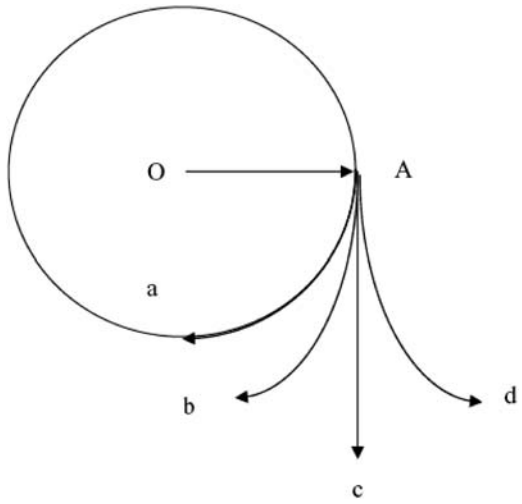
26. Macroscopic friction is caused by microscopic forces between atoms arising primarily from their (a) gravitational, (b) electrical, (c) strong nuclear, (d) weak nuclear interactions.

27. While trying to slide a heavy piano along a rough floor, just before there is any motion (a) the friction force is equal to μ_kN, (b) the friction force is less than μ_kN, (c) the friction force is a maximum, (d) the friction force is less than μ_sN, where N is the normal force.

28. A block of mass m slides down a distance d along an inclined plane with inclination angle θ from rest, starting at height h, with y its vertical coordinate and v its velocity. If the coefficient of kinetic friction is μ_k, when the block is at height y the work–energy theorem is of the form (a) $1/2\, mv^2 + mg(y - h) = \mu_k mg \cos \theta\, d$, (b) $1/2\, mv^2 + mg(y - h) = -\mu_k mg \cos \theta\, d$, (c) $1/2\, mv^2 + mgy = \mu_k mg \cos \theta\, d$, (d) $1/2\, mv^2 + mg(h - y) = -\mu_k mg \cos \theta\, d$.

29. As two identical very smooth plane metal surfaces are polished more and more and then put into tight microscopic contact (a) the friction force is increased, (b) the normal force is reduced, (c) the friction force is unchanged, (d) the friction force is reduced.

30. A small mass, attached to a thread, orbits in a circle around a fixed point O on a horizontal frictionless surface. When viewed from above, as shown to the right, the mass orbits in a clockwise sense. At point A, the thread suddenly breaks. Which of the paths displayed is the one that the mass most likely travels along after the break?



31. A skier of mass M skis along an irregularly shaped, rough slope from point A to point B. The total distance along the slope from A to B is D and the magnitude of the vertical drop from A to B is H. The skier's kinetic energies at A and B are equal. The work done by friction during this trip (a) must be exactly $-MgH$, (b) must be exactly $+$MgH, (c) must be exactly $-MgD$, (d) cannot be calculated because the shape of the slope and the coefficient of kinetic friction are not given.

32. A bug is on the rim of a spinning CD that is rotating counterclockwise viewed from above. The radius of the CD is A and the time it takes for one complete revolution is T. There is a fixed (x-, y-) coordinate system (doesn't rotate with the CD) with its origin at the center of the CD. At $t = 0$ the bug's position in this coordinate system is (A, 0). At $t = T/4$, the x-component of the bug's velocity is (a) $-2\pi A/T$, (b) $4\pi^2 A/T^2$, (c) $\pi A/(2T)$, (d) zero.

Questions 33 and 34 refer to: A particle executes uniform circular motion around a circle of radius equal to 1 m with a speed of 2 m/s.

33. The period of the motion is (a) 2π, (b) 2, (c) π, (d) 1 s.

34. The acceleration of the particle is (a) zero, (b) 2 m/s, pointing toward the center of the circle, (c) constant, with a magnitude of 4 m/s^2, (d) 4 m/s^2, pointing toward the center of the circle.

35. As a car exits from a highway slowing down as it goes clockwise on a circular exit ramp, the net acceleration on the car is directed (a) towards the rear of the car, (b) towards the center of the circular exit

ramp, (c) at some angle between the rear of the car and the center of the circular ramp, (d) at some angle between the center of the circular ramp and the forward direction.

36. A satellite revolves around the Earth in a circular orbit at a constant speed. Which one of the following statements is true? (a) Its acceleration is zero because its speed is constant. (b) Its acceleration is zero because its velocity is constant. (c) Its acceleration and its velocity are both not constant. (d) Its velocity is not constant but its acceleration is a nonzero constant.

37. The Space Shuttle orbits the Earth in a circular orbit at an altitude of 300 km. The Shuttle's mass is 10^6 kg. The period of the orbit is about 5000 s. The radius of the Earth is 6.4×10^3 km and its mass is 6×10^{24} kg. The acceleration of the Shuttle is (a) zero because its speed is constant, (b) about 0.01 m/s^2, (c) about 10 m/s^2, (d) about 8×10^3 m/s^2.

38. The forces responsible for pelleting a protein in an ultracentrifuge are (a) its weight and buoyant force, (b) its buoyant and frictional forces, (c) its weight and frictional force, (d) its weight and centrifugal force.

39. A centrifuge tube is completely filled with water and has a very small bubble (initially stuck) at the bottom of the tube. As the tube is spun in the centrifuge, the bubble will (a) stay at the bottom, (b) steadily rise in the tube at a constant speed, (c) rapidly accelerate to the top of the tube, (d) it's impossible to say given the large variety of factors involved.

40. As a centrifuge rotor accelerates from rest to its final speed, a protein accelerating in a centrifuge tube inside the rotor has an acceleration (a) radially outward, (b) radially inward, (c) tangentially in the direction of the velocity, (d) at some intermediate angle between the inward radial direction and the tangent direction of part (c), (e) none of the above.

PROBLEMS

1. A chessboard consists of 64 squares. Shown numbered are the successive positions of a rook ("castle") for one particular game. Two possible labeling schemes for the squares are shown in (b) and (c); each using ordered integer pairs. Using each of the labeling schemes, list the successive positions of the rook and from the positions determine the displacement vectors that indicate the successive movements of the rook throughout the game. Note that the displacement vector sets should be the same for the two labeling schemes, although the position labels differ between the two.

Comment on any physical meaning to the instantaneous and average velocities of the chess piece.

a

	7						
	6	4			5		
2		3					
1							

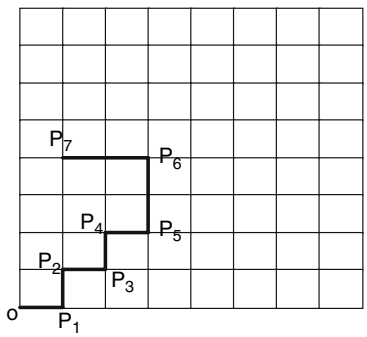
b

1,3							
1,2	2,2						
1,1	2,1	3,1					

c

			–1,1	0,1	1,1		
		–2,0	–1,0	0,0	1,0		
			–1,–1				

2. A bug walks along a chessboard, following the boundaries of the squares. For each of the points P_i, what is the distance traveled from the starting corner, and what is the displacement vector? Express your vector answers in both Cartesian (x, y) and polar (r, θ) coordinates. Assume the board squares measure 25 mm along an edge.

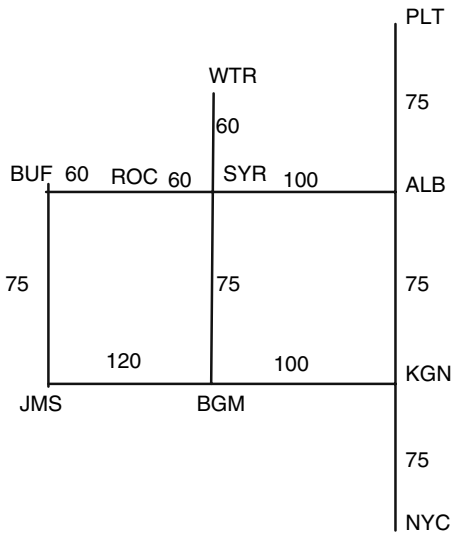


3. Refer to the simplified, albeit tortured, map showing some of the major cities of New York State. The questions refer to the distance traveled and displacement for various trips among the cities shown. Assume travel between consecutive cities "as the crow flies."

(a) Calculate and compare the net displacements for the following pairs of trips.
BGM to WTR via BUF versus ALB to PLT via WTR
KGN to SYR versus NYC to BGM
WTR to ROC versus WTR to PLT

(b) What is the distance traveled and the net displacement for each of these trips:
BGM to JMS to BUF to ROC to WTR to SYR
ALB to SYR to BGM to KGN to BGM

(c) Name two different trips that have the same displacement.



4. Suppose that a swimmer can maintain a stroke that gives her a 3 mph speed in a pool. If she sets out straight across a river that flows with a 1 mph current, she will be carried downstream with the current at the same time as her stroke carries her across. For a river 176 yards (1/10 mile) wide, figure out how far downstream she will end up, assuming that throughout her river crossing she maintains the same stroke that moves her along at 3 mph in the pool. How can our swimmer get directly across the river? (See the next problem as well.)

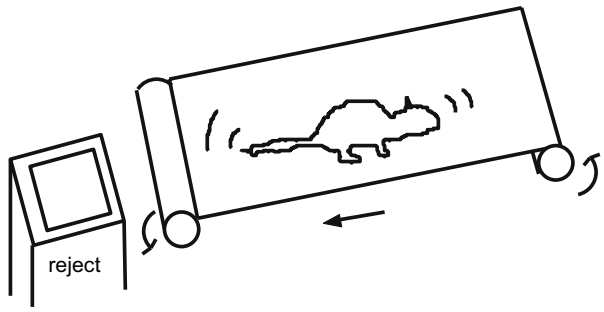
5. Fisherman Joe has a boat with a motor that has two speeds: on and off. When the motor is on, the boat will do 2 mph in an otherwise quiet pond. On an expedition, Joe finds himself at the shore of a river with a 7 mph current.

(a) Why can't Joe move directly across the river with his boat?

(b) Suppose Joe turbocharges his motor so it will move through still water at 12 mph. If he directs the motor appropriately, he can now get directly across the river. Where must he point the boat for a direct traverse?

(c) How long will it take him to thus cross a quarter mile wide section of the river?

6. Professor Igor is attempting to breed a superrat. On an endurance test, one of his prize specimens maintains an apparent stationary position at the middle of a treadmill that is 1 m long and 30 cm wide, while the track moves back at 2 m/sec. After several minutes, the rat quits, turns, and begins to crawl to the side edge of the track. With what constant speed must our exhausted hero travel if he is not to be carried to the end of the track and dropped into the reject bin?



7. Consider these four vectors, represented by order pairs:

$$\vec{A}_1 = (2, 0); \vec{A}_2 = (0, 1); \vec{B}_1 = (3, 0); \vec{B}_2 = (0, 4).$$

(a) Of these four vectors, which are perpendicular to each other? Which are parallel?

(b) Calculate each of the following and represent the results, to scale, on graph paper as well.

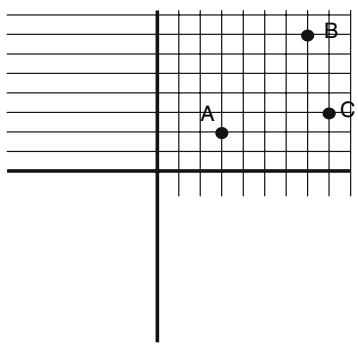
$$\vec{A}_1 + \vec{A}_2 \quad \vec{B}_1 + \vec{B}_2 \quad \vec{A}_1 + \vec{B}_1 \quad \vec{A}_2 + \vec{B}_2 \quad \vec{A}_1 + \vec{A}_2 + \vec{B}_1 + \vec{B}_2.$$

Note: For problems using compass bearings, note that degree headings are customarily measured clockwise from North. E (east) is 90°, for example. Also, the directions NE, NW, SE, and SW are oriented exactly 45° from the appropriate main bearings (N, S, E, and W).

8. Determine how far and in what direction a hiker ends up after the following treks:

- 2 mi N, then 1 mi E.
- 1 mi E, then 2 mi N.
- 2 mi NE, then 1 mi E.
- 2 mi NE, then 1 mi W.
- 2 mi N, then 2 mi W, then 2 mi S. What about the return trip (i.e., the same hike backwards)?
- 4 mi S, then 3 mi E.
- 3 mi NW, then 3 mi NE.

9. The figure shows three points within a rectangular coordinate grid. The coordinates of each point can also be considered as a vector, representing the displacement from the origin O to the respective point. Thus, for example, the coordinates of A also represent the vector OA.



(a) Represent each of the point-vectors with ordered pairs.
(b) Find the distance between: points A and B; points A and C; and points B and C.
(c) Determine the following vectors. OA + OB; OB + OC; OA + OB + OC; 2OB, OA + 2OB, 3OB, −OB.
(d) Determine the difference of vectors OA –OC according to: OA −OC = OA + (−OC); that is, first determine –OC, given OC and perform the indicated addition. Show the results of each of the steps graphically.

10. Calculate each of the following operations on the given vectors, here represented with ordered pairs. Then show on graph paper the representations of the given vectors and the vector result of the requested operation, all to scale.

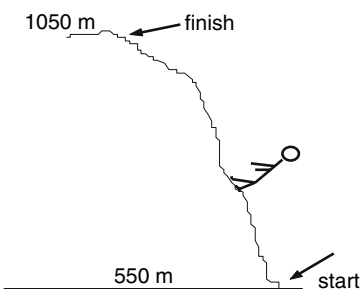
$\vec{A} = (2, 3)$; $\vec{B} = (5, 6)$; $\vec{C} = (2, -1)$; $\vec{D} = (6.5, 2.5)$; $\vec{E} = (-4, -2)$.

(a) $-\vec{A}$; (b) $\vec{A}+\vec{B}$; (c) $\vec{A}-\vec{B}$; (d) $\vec{A}+\vec{C}$; (e) $\vec{A}+\vec{D}$; (f) $\vec{D}-\vec{B}$; (g) $\vec{C}+\vec{E}$; (h) $\vec{C}-\vec{E}$; (i) $-\vec{C}-\vec{E}$.

11. A popular exercise in orienteering is to be given a compass heading and a distance to be hiked, at the successful completion of which is another set of instructions containing yet another heading and distance, and so on. At the end of it all, the hiker has hopefully arrived safe and sound back at camp.

Construct a map, to scale, that shows the path of travel for a hiker who successfully completes the following course.
100 yd, N
150 yd, E
60 yd, NW

12. A mountain climb begins at 550 m above sea level and finishes atop a 1050 m high peak. The average incline is 75° above the horizontal. Give the horizontal and vertical components of the hikers' displacements. What vector represents the sum of the horizontal and vertical displacements?



13. A ball is thrown horizontally from the roof of a 25 m tall building with a speed of 20 m/s.
(a) With what velocity will it land (magnitude and direction, please)?
(b) How long will it be in the air?
(c) How far from the building's ground floor will it land?
(d) What is its acceleration just before it hits the ground?

14. Three identical balls are thrown off a building, all with the same initial velocity. One of the balls is thrown horizontally, the second ball is thrown at some angle above the horizontal, and the third is thrown at some angle below the horizontal. Rank the speeds of the balls as they reach the ground.

15. A Northrop B-2 Stealth bomber is flying horizontally over level ground, with a speed of 300 m/s at an altitude of 10.6 km (35,000 feet).
(a) Neglecting air resistance, how far will a bomb travel horizontally between its release and its impact on the ground?
(b) If the bomber flies straight ahead at the constant speed above, where will the bomber be when the bomb hits the ground?

16. A cartoon coyote sets out to capture the elusive roadrunner by wearing a pair of *Acme* jet-powered roller skates, which provide a constant horizontal acceleration of 10 m/s^2. The coyote starts off at rest 100 m from the edge of a cliff at the instant the roadrunner zips past him in the direction of the cliff.
(a) If the roadrunner moves with constant speed, what is the minimum speed the roadrunner must have in order to reach the cliff before the coyote?
(b) At the edge of the cliff the roadrunner escapes by making a sudden turn, and the coyote continues straight off the cliff. If the cliff is 200 m above the ground, where does the coyote land, assuming that his skates remain horizontal and continue to work while in flight?
(c) What are the components of the coyote's impact velocity?

17. A cartoon coyote chasing an animated roadrunner fails to make it around a tight corner, and runs directly off the edge of a 100 m cliff at a horizontal speed of 20 m/s. How far from the base of the cliff does he land, and how much time does the road runner spend in flight?

18. A game of *Battleship*™. An enemy ship is on the left side of a mountain located in the middle of the ocean and this ship has the ability to maneuver within 1 mi (1600 m) of the 800 m tall mountain. A gun located on the deck of the enemy ship can fire projectiles with an initial speed of 650 mph (=289 m/s) at angles between 0° (horizontally from the ship) and 90° (directly overhead of the ship.) You are stationed on a ship on the right side of the mountain and you can maneuver your ship from the shoreline located 500 m from the middle of the mountain to any larger distance. At what distance(s) from the rightmost shoreline can you maneuver your ship so that you will not be hit by the enemies' projectiles?

19. A cartoon coyote comes up with a brilliant scheme to get lunch for himself by dropping a 500 kg boulder on a passing animated roadrunner. Unfortunately, when he cuts the rope holding the boulder in place, the rope becomes tangled around his ankle, and drags him off the edge of the cliff. If the coyote's mass is 30 kg and his frantic clawing at the ground produces a force of 120 N resisting being dragged off of the cliff, what is his acceleration toward the cliff?

20. Two balls are thrown off the roof of a 25 m tall building. One is dropped from rest and then 1 s later the second is thrown outward with a velocity with horizontal and downward components of 10 and 15 m/s, respectively.
(a) Which ball hits the ground first?
(b) With what velocity does each ball hit the ground?
(c) Which ball travels the greater displacement?

21. A lacrosse goalie clears the ball by throwing it downfield at a speed of 10 m/s at a 35° angle above the ground.
(a) How long will it be in the air? (Assume the ball leaves the goalie's stick at ground level.)
(b) How far will it go before hitting the ground, assuming no one is there to catch it?
(c) At what point will it have its minimum speed?
(d) With what velocity (magnitude and direction) will it hit the ground?
(e) If someone catches the ball on its way down at a height of 1.0 m, with what velocity will the ball hit the net of the lacrosse stick?

22. A 30 kg penguin slides down the side of a glacier that has a constant slope of 50°. What is the acceleration of the penguin and what is the normal force it feels?

23. The largest rope lariat ever spun used a 100 foot long rope with a loop of 95 feet spun in a circle. What is the centripetal acceleration of a point on the rope spun at 60 rpm?

24. The fastest a manmade device ever spun is 4500 miles per hour achieved by a 6 inch fiber rod spun about one end in a vacuum. What is the centripetal acceleration of a point on the rim of this rod in terms of *g*'s?

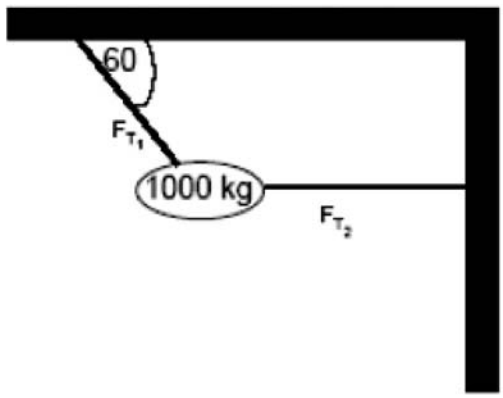
25. A 20 cm radius wheel is turning at the rate of 5 rpm (revolutions per minute). Find (a) the speed of a point on the rim, (b) the centripetal acceleration of a point on the rim, and (c) the time for one revolution.

26. A block sitting on smooth ice is tied to a 1 m cord and spun in a horizontal circle at constant speed. If the block is revolving at 15 revolutions in 1 min and the cord is cut, find the magnitude and direction of the block's velocity just then.

27. A 100 kg sled is slid across a smooth ice field by a group of four dogs tied to the sled pulling with a 350 N force along a rope at an angle of 20° above the horizontal.
(a) If the sled travels at a constant speed, find the drag force on the sled.
(b) Find the work done by the dogs after pulling the sled for 1 km.

28. *The Pumpkin on the Nott*: The Nott Memorial is a 16-sided Victorian building and national historic landmark located in Schenectady, NY. The Nott Memorial is topped with an approximately hemispherical dome 89 feet in diameter. Suppose that the dome is frictionless when wet. Somehow an individual has balanced a pumpkin at the top of the dome at an angle of $\theta_i = 0°$ with the vertical. Suppose that on a rainy night, a gust of wind starts the pumpkin sliding from rest. It loses contact with the dome when the line from the center of the hemispherical dome to the pumpkin makes a certain angle with respect to the vertical. At what angle does this happen?

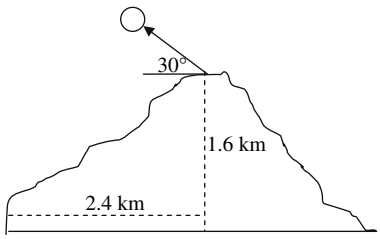
29. *Raiders of the Last Exam*: In order to prevent cheating, a diabolical physics professor has booby-trapped her office where the exam answers are kept. A 1000 kg mass is suspended by a 4 m rope from the ceiling, and pulled to one side of the room where a second rope holds it. The rope holding the mass makes a 60° angle with respect to the horizontal. When a student attempts to open the file cabinet containing the answers, the weight will be released to swing back and forth in front of the cabinet, crushing anyone foolish enough to stand in front of it.



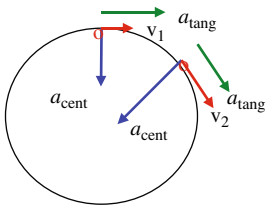
(a) What is the tension in the two ropes before the trap is sprung?
(b) What is the maximum velocity of the swinging mass?
(c) The student is quicker than expected, and jumps back before the mass hits. If the student ducks back in to grab the answers just after the mass passes, how much time does she have to get them before the mass returns?

30. In the movie *Volcano*, solid chunks of rock, called *lava bombs* were ejected from the growing volcano. Consider a volcano, shown below, with a lava bomb being ejected.
(a) What would the magnitude of the initial velocity, at the top of the volcano, have to be in order for a *lava bomb* to land at the base of the volcano?
(b) What would be the time of flight of this projectile from the top of the volcano to its base?

(c) What is the final velocity of the *lava bomb* just before it hits the ground at the base?
(d) What is the acceleration of the *lava bomb* just before it hits the ground at the base?

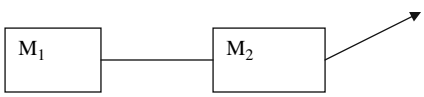


31. Suppose that a Lockheed C-5A Galaxy (the largest aircraft in the world with some specifications shown below) sits on a runway at an airport waiting for take-off clearance. When given clearance, the pilots apply full power to the plane's engines and accelerate down the runway.

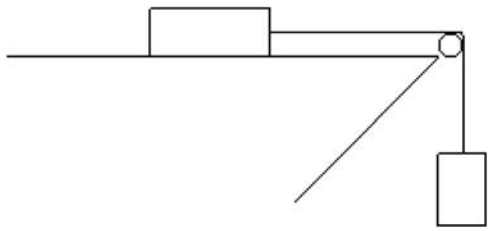


(a) If the plane takes off when its velocity reaches 195 mi/hr and not before, what is the acceleration of the plane in order to take off at the indicated speed if the plane has to be airborne in 9800 feet? (Hint: 1600 m = 1.0 mi and 1 hr = 3600 s.)
(b) How long does it take before this plane becomes airborne (takes off)?
(c) What force (magnitude and direction) would be required to support a C5-A fully loaded with fuel horizontally in flight?

32. Two blocks, with masses M_1 and M_2, are connected by a light horizontal cord and pulled by a second cord with a force F at an angle θ with respect to the horizontal so that the blocks slide along a horizontal surface at a constant speed.
(a) Draw a carefully labeled free-body diagram for each block showing all forces.
(b) From your labeled diagram in part (a), write equations describing the motion.
(c) If $M_1 = 1.0$ kg, $M_2 = 2.0$ kg, and the applied force is 10 N at an angle of 30°, find the coefficient of kinetic friction between the blocks and the horizontal surface.



33. Two blocks are connected by a light cord. One block, of mass 4 kg, sits on a horizontal table with static and kinetic coefficients of friction of 0.6 and 0.4, respectively, whereas the other block, of 2 kg mass, hangs over a frictionless light pulley as in the figure. The blocks are released from rest.
(a) Draw a carefully labeled free-body diagram for both blocks and, by using Newton's laws, show that they do not move.



(b) If the two blocks are exchanged, so that the 4 kg is now the hanging block and the 2 kg sits on the table, find their acceleration now.
(c) In words state why the tension in the cord is equal to the weight of the hanging block in part (a) but not in part (b)
(d) What is the minimum mass that one needs to add to the 2 kg block in part (b) for it to remain at rest when released?

34. A 75 kg crate is being pulled up a 5 m long (frictionless) ramp inclined at a 30° angle from the horizontal by a force of 500 N at an angle of 15° above the ramp.
(a) What is the acceleration of the crate?
(b) What will be its velocity at the top of the ramp if it starts from rest?
(c) How much work is done to get the crate up the ramp by pulling the rope?
(d) How much work is done by gravity over the 5 m ramp?
(e) Using the work–energy theorem redo part (b).

35. Suppose the toy car in the frictionless loop-the-loop example of 6.13 starts from a height of 1.2 m and the loop itself has a height of 0.25 m.
(a) Find the speed of the car at the top of the loop.
(b) How fast will it be going at the bottom of the loop on the way up? On the way down?
(c) Find the minimum height that the car must start from to just get over the top of the loop. (Hint: The speed at the top cannot be zero or the car, traveling in a circle, would not reach there. The minimum speed required at the top is such as to have a centripetal acceleration at the top just equal to g as the car leaves the track at the top.)

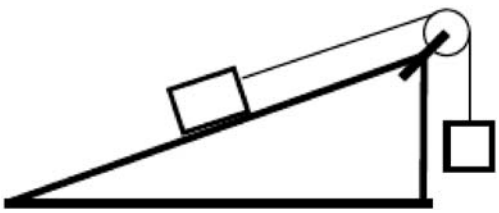
36. In a loop-the-loop roller coaster (see Figure 5.16) if a car of 500 kg mass starts essentially at rest from the top of a 15 m tall hill find
(a) Its speed when traveling vertically on the 7 m diameter frictionless loop

(b) Its velocity when leaving the loop at ground level
(c) The net force on the car when at the position in part (a)

37. A roller coaster car, with a mass of 500 kg, crests a 20 m high hill while moving at a speed of 10 m/s. It then rolls down the other side, all the way to ground level, before climbing a second hill.
(a) What is the speed of the car when it is 10 m up the second hill?
(b) What is the maximum possible height of the second hill?
(c) If the car is subject to a frictional force that causes it to lose 8000 J of energy, what is the maximum height of the second hill?

38. A block of mass 12 kg slides from rest down a frictionless 35° incline and is stopped by a strong spring with stiffness constant $k = 3.00 \times 10^4$ N/m. The block slides 3.0 m from the point of release to a point where it comes to rest against the spring. When the block comes to rest, how far has the spring been compressed?

39. Two blocks are connected by a light string with one block of 5 kg mass sitting on a frictionless 30° inclined plane and the second block of 8 kg mass hangs from the string which runs over a frictionless light pulley as shown.



(a) Find the acceleration of the block on the plane.
(b) Find its velocity after traveling 2 m along the plane from rest. Do this two ways: using your answer to part (a) and using energy principles.

40. In the previous problem if the block on the incline is 5 kg as before,
(a) Find the hanging mass needed so that the 5 kg mass is in equilibrium.
(b) Find the hanging mass needed so that the 5 kg mass slides down the 2 m distance along the plane in 2 s.

41. In a pinball game with marbles, a 10 N/m spring is compressed 3.0 cm releasing a 50 *g* marble from rest. If the marble needs to travel 60 cm up a 3° incline before entering the scoring zone of the game table, will it make it? If not, how much must the spring be compressed so that it will?

42. A crate is pushed along the ground at constant velocity for a distance of 5 m. If the friction force is 5 N, how much net work is done on the crate? How much work is done by the friction force? By the external pushing force? By gravity?

43. A 2 kg box is pushed 3 m up a 30° incline at constant velocity by a 20 N force directed along the surface of the incline with a coefficient of kinetic friction of 0.6. What net work is done on the box? How much work is done by the pushing force? By gravity? By the friction force? By the normal force? Check that your answers are consistent.

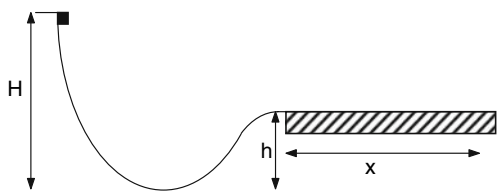
44. A 110 kg upright piano is being pulled by a light rope angled at a 20° angle below the horizontal. If the tension in the rope is 30 N and the coefficient of kinetic friction is 0.3 find:
(a) The normal force on the piano.
(b) The friction force.
(c) The acceleration of the piano.
(d) Why is this a poor method to move a heavy piano?

45. A 20 kg wheelbarrow held at a 30° angle is being pushed along the ground by a force *F* at a constant velocity. If the coefficient of kinetic friction is 0.4 find:
(a) The net force acting on the wheelbarrow.
(b) An expression for the normal force in terms of *f*.
(c) A numerical value for *F*.
(d) Why is the normal force greater than the weight of the wheelbarrow?
(e) Would it be easier to pull the wheelbarrow at the same angle at constant velocity?

46. Two heavy crates sit on the floor, the 3 kg one on top of the 10 kg one.
(a) What is the normal force from the floor on the 10 kg block?
(b) What is the normal force acting on the top crate?
(c) If the bottom crate is pushed horizontally with a 10 N force along the smooth floor and the coefficients of static and kinetic friction between the crates are 0.6 and 0.4, what is the acceleration of the bottom crate and the top crate?
(d) Find the maximum horizontal force that can be applied to the bottom crate without the top crate slipping. (Hint: First find the maximum static friction force and the resulting acceleration of the top crate.)

47. A 0.1 kg block is given an initial velocity of 5 m/s up an inclined plane at a 30° angle, travels up the plane, and then returns back to the bottom. The coefficient of friction between the block and plane is 0.4. Find
(a) The work done by gravity for the entire trip
(b) The work done by the friction force for the entire trip
(c) The net change in kinetic energy of the block

48. A small mass *m* slides down a frictionless ramp from rest as shown in the figure below and then enters a region where the coefficient of friction is 0.5. Where does the mass stop? Find an expression for *x* in terms of the given parameters.



49. Two blocks are connected by a light cord. One block, of 4 kg mass, sits on a horizontal plane with static and kinetic coefficients of friction of 0.6 and 0.4, and the other block, of 2 kg mass, hangs over a frictionless light pulley as in Figure 5.20. If the blocks are released from rest:
(a) Show that the blocks do not move.
(b) What minimum additional force would be needed to pull down on the 2 kg block to produce motion?
(c) If the two blocks are exchanged, find their acceleration now.
(d) What is the minimum mass that one needs to add to the 2 kg block in part (c) for it to remain at rest when released?

50. A 2.5 kg block sits on an inclined plane with a 30° inclination. A light cord attached to the block passes up over a light frictionless pulley at the top of the plane and is tied to a second 2.5 kg mass freely hanging vertically. The coefficients of static and kinetic friction between the block and the plane are 0.5 and 0.3. When released from rest find:
(a) The acceleration of the blocks.
(b) The tension in the string.
(c) Explain why the tension supporting the hanging block is not equal to its weight.
(d) Find the time for the block on the inclined plane to travel 0.5 m up the plane.
(e) Find the minimum angle of inclination at which the block on the plane will remain at rest.

51. The eruption of the Mt. St. Helens volcano on May 18, 1980 triggered a huge avalanche of snow down its slopes, estimated at 96 billion cubic feet. The maximum speed of the avalanche was clocked at 250 mph. Estimate the average force (in N) exerted on the land at the base of the mountain assuming that all the snow was traveling at this speed and stopped in 5 s. Take the density of snow to be half that of water.

52. Two blocks sit on an inclined plane with a 30° inclination angle. If the blocks are connected by a light rope with the 5 kg block above and the 3 kg block below, find the acceleration of the blocks if the coefficients of sliding and static friction are 0.3 and 0.5, respectively. Does the order of the blocks matter?

53. In the previous problem find the maximum angle at which the blocks do not slide down the plane. Does the order of the blocks matter now?

54. Two identical springs with $k = 5$ N/m, are separated by 2 m, with a small coefficient of kinetic friction $\mu_k = 0.02$ acting on the horizontal surface between them. If a 0.1 kg block starts out being released from one of the springs after compressing it by 0.2 m, find the final position of the block, tracing its trajectory. (For simplicity, assume that you can ignore friction for the portion of the motion when the blocks are in contact with the springs.)

55. Two blocks are attached by a light cord with each block sitting on a different inclined plane as shown.

If the angles of inclination are 30° and 60°, the respective masses are 10 kg and 6 kg, and the coefficients of sliding and static friction are 0.3 and 0.5, do the masses move, and if so in which direction and with what acceleration?

56. A roller coaster rises and falls on a semicircular portion of track that has a radius of curvature of 20 m. How fast can the roller coaster travel so that a 60 kg man will not leave his seat at the top? (Hint: Find the threshold condition—what changes—when the man is just about to leave his seat.)

57. In a circus performance, a stuntman is riding a bicycle in a loop-the-loop. Assuming that the loop is a circle of radius $R = 2.7$ m, what is the least speed that the performer can have at the top of the loop in order to remain in contact with the track? Does your result depend on the mass of the performer?

58. The so-called *ROTOR* is an amusement park thrill ride. Riders enter the ride, which is a large hollow cylinder that is rotated rapidly around its central axis, and stand against a wall. As the ride starts, the riders, wall, and floor move in unison. At a predetermined speed, the floor falls away but the riders remain pinned to the wall. If the coefficient of static friction between the riders clothing and the wall is 0.40, and the radius of the ride is $R = 3.0$ m, what is the minimum speed needed so that the riders do not fall when the floor drops? What is the magnitude of the centripetal force on the rider if the rider has a mass of 60 kg?

59. The moon travels around the Earth in a nearly circular orbit of radius 3.84×10^8 m with a period of 27.3 days.
(a) What is the speed of the moon in orbit relative to the Earth?
(b) What is the centripetal acceleration of the moon based on its orbital period?

60. An exit ramp off a highway has a radius of curvature of 150 m and is banked at a 4° angle. For what speed is the ramp designed?

61. A circular gear of 5 cm radius starts from rest and accelerates to 60 rpm in 10 s.
(a) What is the (assumed constant) tangential acceleration of a point on the rim of the gear?
(b) What is the centripetal acceleration after 5 s? After 10 s?

62. A yo-yo is spun in a vertical circle ("around the world") of radius 40 cm. Find the difference in the string tension

at the top and bottom in terms of the weight of the yo-yo. (Ignore the spinning of the yo-yo around its own axis.) (Hint: No work is done on the yo-yo as it circles (why?), so conservation of energy can be applied.)

63. In an amusement park ride, people stand against the outer wall of a large spinning drum and after the drum rotates beyond a certain speed, the floor falls away, leaving the people suspended against the wall. If the radius of the ride is 12 m, and the coefficients of static and kinetic friction are 0.4 and 0.2, how fast must the drum spin so that no one will fall. Find both the velocity of the people and the rpm of the drum.

64. A heavy 20 kg crate is pushed with a force of 50 N down a ramp making an angle of 30° with the horizontal. The crate is pushed down the incline with the force directed at an angle of 30° below the surface of the ramp. The coefficient of kinetic friction is 0.3 and the coefficient of static friction is 0.6.

(a) Draw a free-body diagram for the crate, carefully labeling each force with an appropriate symbol and clearly showing the direction of each force. (Read the problem carefully.)

(b) Write down—but do not solve—the equations from Newton's second law for motion along and perpendicular to the ramp

(c) Now solve your equations from part (b) to find the acceleration of the crate.

(d) If you stop pushing, does the crate slide down the plane? (Show your work in answering this.) If so, find the acceleration.

65. A block of mass *M* is attached to a light cord and spun clockwise in a vertical circle at constant speed. At the top of the circle the tension in the cord is equal to three times the weight of the block.

(a) Draw a free-body diagram for the block at the top of the circle

(b) If the radius of the circle is 0.75 m and the mass of the block is 2 kg, find the speed of the block

(c) If the cord were to break when the block is at a point along a horizontal diameter while moving upward, describe in words the trajectory of the block and calculate the maximum height the block will reach.

66. A 50,000 N truck exits a highway at 50 mph onto a 2.5° banked exit ramp which makes a semicircle of 250 m radius, slowing at constant deceleration to 20 mph by the end of the ramp.

(a) What is the tangential acceleration of the truck on the ramp?

(b) What is the net acceleration at the beginning and end of the ramp?

(c) What is the required frictional force on each of the truck's eight tires to keep it traveling on the road at the beginning and end of the ramp?

67. A Ferris wheel of 15 m radius rotates at 2 rpm. Find the normal force from the seat on a 50 kg boy when he just passes a point at the height of the wheel axis on the way up.

68. How long will it take for a 5 S protein to completely spin to the bottom of a 5 cm centrifuge tube filled with solution when spun at 500,000 *g*s? Express your answer in hours.

69. Find the centripetal force acting on a 42 kDalton (1 Dalton = 1 g/mole) protein molecule spinning at 50,000 rpm and located a distance of 8 cm from the axis of rotation. If the protein has a net radial force directed inward, toward the axis of rotation, why does it slowly migrate outward toward the bottom of the centrifuge tube? Explain this as carefully as you can.

70. Calculate the sedimentation coefficient in water for

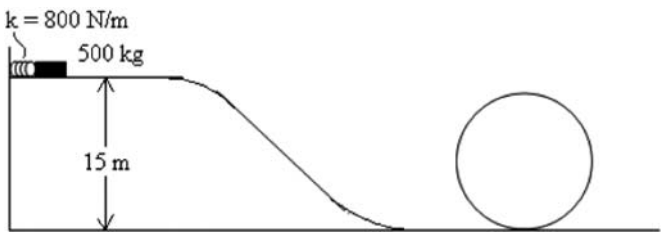
(a) A spherical cell of 3 μm radius.

(b) A spherical (or globular) macromolecule of 3 nm radius. Assume that the density of each is 1.05 g/cm^3.

(c) Calculate the number of *g*'s required in a centrifuge if the cells are to be sedimented through 2 cm in 5 min.

(d) Similarly, calculate the number of *g*'s required to sediment the macromolecule at a rate of 1 mm/h.

71. An amusement park thrill ride consists of a cart with some riders (of total mass 500 kg) that is set in motion by a large spring with spring constant 800 N/m. The cart travels along the flat horizontal section of track that is located 15 m above the ground and then down the ramp toward the loop-the-loop which has an unknown diameter. The entire track is frictionless.



(a) If the spring is initially compressed by 3 m, what is the speed of the cart as it leaves the spring?

(b) What is the speed of the cart at the bottom of the ramp before the loop-the-loop?

(c) If the speed of the cart at the top of the loop-the-loop is 8.55 m/s, what is the diameter of the loop?

(d) How much work was done by gravity on the cart as it traveled from the bottom of the loop-the-loop to the top?

(e) Suppose that the horizontal section of the track at the top were not frictionless, but that a frictional force was present, with a coefficient of kinetic friction of 0.20. What would the speed of the cart be as it left the spring?

6 Momentum

In this chapter we begin our study of more realistic systems in which the objects are no longer point particles but have extension in space. Up until now we've generally limited ourselves to the dynamics of point masses, first in one dimension and then generalized to two and three dimensions. Indeed, not all of the problems we studied were limited to point masses, but the object's size and shape were not relevant in the problem and so were not considered. For such objects we've learned how to describe and predict translational motion using Newton's laws, some of the complications due to frictional forces, and the important concept of energy. In general we can divide the motion of real extended bodies into two parts: translational motion, described by following a particular average coordinate of the object, known as its center of mass as it moves about, and all other motions with respect to this point. This chapter focuses on translational motion of systems, or collections of objects, and the following chapter deals with rotational motion.

We begin this chapter by introducing the important concept of momentum. As we've seen, all forces come in pairwise interactions. When studying the interactions between different objects, it turns out that we can re-formulate Newton's second law in terms of momentum. If the system we are studying is "isolated"—meaning that it does not interact with the outside world—then our reformulation is particularly simple and leads to a new fundamental law, the law of conservation of momentum. After seeing this for a system of two particles, we next define and learn how to compute the center of mass of a system, that special average point of a system at which all its mass appears to be concentrated in order to explain the net translational motion of the system. The last section of the chapter shows how to reformulate the dynamics of translational motion of any system in terms of the center of mass momentum. Here we also see the general formulation of conservation of momentum.

1. MOMENTUM

Thus far in our discussions of dynamics we have focused on forces as the origin of motion according to Newton's laws. There is a very useful alternative approach based on momentum that we wish to develop in this chapter. Very often this alternative approach is to be preferred because it does not hinge on the specific forces or interactions between objects, which are usually unknown or only incompletely understood. In this section, we first introduce momentum, the basic quantity used in this approach, for a particle. Then we reformulate Newton's second law using momentum and show how this leads to the conservation of momentum principle for a collection of particles. Later in this chapter we generalize this approach to arbitrary collections of extended objects.

An object of mass m traveling at velocity $\vec{v}$ has a linear *momentum* (or just momentum) $\vec{p}$, given by

$$\vec{p} = m\vec{v}. \tag{6.1}$$

J. Newman, *Physics of the Life Sciences*, DOI: 10.1007/978-0-387-77259-2_6,

Newton's second law for an object can be written in terms of momentum as

$$\vec{F}_{\text{net}} = \frac{d\vec{p}}{dt}.$$

Using the definition of $\vec{p}$ (Equation (6.1)) and the product rule for derivatives, we can write this as

$$\vec{F}_{\text{net}} = \frac{d(m\vec{v})}{dt} = m\frac{d\vec{v}}{dt} + \vec{v}\frac{dm}{dt}.$$

In the case when the mass is not changing the last term vanishes and using the definition of acceleration we get the usual form of $\vec{F}_{\text{net}} = m\vec{a}$. In cases where the mass is changing (e.g., a rocket ejecting substantial amounts of fuel), the full expression is needed and this form of Newton's second law is the correct expression.

Note that momentum is a vector quantity, defined as the product of the mass, an intrinsic property of the object, and its velocity, a quantity depending on its motion. It has units of kg-m/s, which have no other special name. Clearly, based on Newton's first law, a particle with no net force on it will maintain a constant momentum. When the particle feels a net force, due to some interaction, its momentum will change with time. Also clearly, based on Newton's second law, the larger the interaction (force) acting on the particle, the greater will be the change in its momentum.

How does the momentum of a particle contrast with its velocity? First, we note that both of these quantities are vectors, in fact with the same direction. If we compare two particles of different mass traveling at the same velocity, the one with larger mass will also have proportionally larger momentum. For example, a truck with four times the mass of a car, both traveling at the same speed along a highway, has four times the momentum of the car, in accord with our colloquial usage of the word momentum. On the other hand, if the same truck is traveling at only 1/4 the velocity of the car, then both vehicles have the same momentum.

How does the momentum of a particle contrast with its kinetic energy? Now, note that these are very different quantities, with kinetic energy a scalar and momentum a vector. A particle with a given mass will have its momentum doubled if its velocity doubles, but will have its kinetic energy quadrupled in that case. Kinetic energy is produced by doing work on a particle, as we've seen in the work–kinetic energy theorem. How is momentum produced? Well, clearly they are related, but the direct answer is that momentum is produced by forces acting on the particle as we now show.

Newton's second law for an object can be written in terms of its momentum by noting that $m\vec{a}$ is defined as

$$m = \lim_{\Delta t \to 0} \frac{\Delta \vec{v}}{\Delta t},$$

and because m is constant, we can further write that

$$m\vec{a} = \lim_{\Delta t \to 0} \frac{\Delta mv}{\Delta t} = \lim_{\Delta t \to 0} \frac{\Delta \vec{p}}{\Delta t}.$$

We therefore find that

$$\vec{F}_{\text{net}} = \lim_{\Delta t \to 0} \frac{\Delta \vec{p}}{\Delta t}. \tag{6.2}$$

This is actually the form that Newton proposed for the second law and is more general than the form $\vec{F}_{\text{net}} = m\vec{a}$, because it allows for cases in which the mass of an object may change with time. Such a situation might arise when mass is either being added or removed from the object over time (see the boxed discussion). For example, a rocket burns fuel and decreases its mass by ejecting the waste gases or the mass of a boat that is drifting by a pier may suddenly increase when you jump into it. In both of these cases our original form of $F = ma$ does not apply because the mass is changing.

Example 6.1 An *E. coli* bacterium of mass $m = 6 \times 10^{-16}$ kg is initially swimming at a constant velocity of 8 μm/s toward the east. One ms later it is found to be swimming at 10 μm/s toward the north. Find the change in the *E. coli*'s momentum and the average external force acting on the bacteria during the 1 ms time interval.

Solution: It is very tempting to write that the change in the bacterium's momentum is the product of its mass and the change in its speed $(10 - 8) = 2$ μm/s. This temptation must be strongly resisted because it is the change in the velocity vector that is appropriate and this is not a one-dimensional problem. Figure 6.1 shows a vector diagram for the initial and final momenta and the change in momentum of the bacterium. From the figure it is clear that the change in momentum is found from the hypotenuse of the triangle formed so that

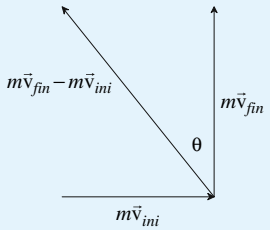


FIGURE 6.1 *Vector subtraction for Example 6.1.*

$$\Delta p = m\sqrt{v_{\text{ini}}^2 + v_{\text{fin}}^2} = 6 \times 10^{-16}\sqrt{(8 \times 10^{-6})^2 + (10 \times 10^{-6})^2}$$

$$= 7.7 \times 10^{-27}\ \text{kg} \cdot \text{m/s}$$

The direction of this momentum change is given by

$$\theta = \tan^{-1}\left(\frac{8}{10}\right) = 39^{\circ},$$

where the angle θ is measured west of north as shown in the figure.

The average force acting over this interval of time is then given by Equation (6.2) (without the limit) and is found to be

$$\vec{F} = \frac{\Delta\vec{p}}{\Delta t} = \frac{7.7 \times 10^{-27}}{10^{-3}} = 7.7 \times 10^{-24}\,N$$

in the same direction as the momentum change.

Example 6.2 The fastest passenger elevator in the world (in a 70-story building in Yokohama, Japan) attains a maximum speed of 12.5 m/s (28 mph) taking passengers from the ground to the top floor in 40 s. Find the maximum change in your momentum if you were to ride in this elevator. What is the net change in your momentum for the entire trip?

Solution: The maximum change in your momentum would occur during the acceleration or deceleration phase of the ride. Assuming your mass to be 80 kg, during the acceleration phase your momentum would increase from zero to $p =$ (80 kg) (12.5 m/s) = 1000 kg m/s, so that your maximum change in momentum would just be 1000 kg m/s. For the entire trip to the 70th floor your net change in momentum is zero because both your starting and ending momentum are zero.

Suppose that two otherwise isolated point particles undergo a collision. We would like to understand what occurs and be able to predict the outcome. When far enough apart, the two particles move independently and do not interact. They will each have some momentum and if they are to collide must be moving along a line connecting them; let's call this the x-axis and we see that this problem for two-point particles is really one-dimensional. Momentum is a particularly useful concept in this situation, as we show. Suppose that particle #1 has momentum p_1 and particle #2 has momentum p_2, both directed along the x-axis. For them to collide they must be moving toward each other, but they might both be moving in the same direction with one

particle "catching up" to the other, so let's label the momenta as both positive for this discussion.

If we write Equation (6.2) for each of the particles we have

$$\vec{F}_{2\text{ on }1} = \frac{\Delta\vec{p}_1}{\Delta t} \quad \text{and} \quad \vec{F}_{1\text{ on }2} = \frac{\Delta\vec{p}_2}{\Delta t}, \tag{6.3}$$

where the only force on each particle is from the other one. These forces need not be contact forces acting only during a (macroscopic) contact between the two particles, but can also be long-range forces acting over long distances. Now, using Newton's third law, we know that these two forces are reaction-pair forces and are always equal and opposite to each other. We can conclude then that because the vector sum of the two forces always adds to zero, we must have at all times that

$$\frac{\Delta\vec{p}_1}{\Delta t} + \frac{\Delta\vec{p}_2}{\Delta t} = 0 \quad \text{or} \quad \frac{\Delta(\vec{p}_1 + \vec{p}_2)}{\Delta t} = 0 \quad \text{or} \quad \Delta(\vec{p}_1 + p_2) = 0. \tag{6.4}$$

For this to be true it must be that the *net momentum of the two particles remains constant with time*. We say that, in this situation, *momentum is conserved*. We have specifically written these last few steps using vectors and in a general way to show the power of the law of conservation of momentum, even though our current example is one-dimensional. All we have used in this derivation are Newton's laws (specifically the second law written in terms of momentum and the third law) and the fact that the particles were otherwise isolated, not interacting with any other objects. Thus, we really have proven that any two isolated objects, not necessarily point particles, that collide will have a total momentum that remains constant (Figure 6.2). Furthermore, even if there are external forces acting on the two particles, as long as there are no external forces acting along the direction of their motion, momentum will still be conserved. For example, momentum will be conserved for horizontal (frictionless) motion of two colliding objects even though gravity may act vertically. We show this in a couple of examples just below.

What does this tell us about the interactions between the two particles and the outcome of the collision? The beauty of this formulation is that the outcome is independent of the interactions; we do not need to know anything about the details of the interaction in order to predict something about the outcome. All we need to know is contained in Equations (6.3) and (6.4). During the collision the two objects will exert equal and opposite forces on each other for some period of time. If the collision involves short-range forces, so that the collision time Δt is short, then the product of the (typically) large force on one particle from the other and the short collision time is called the impulse,

$$\text{Impulse} = F\Delta t = \Delta p = p_{\text{final}} - p_{\text{initial}}. \tag{6.5}$$

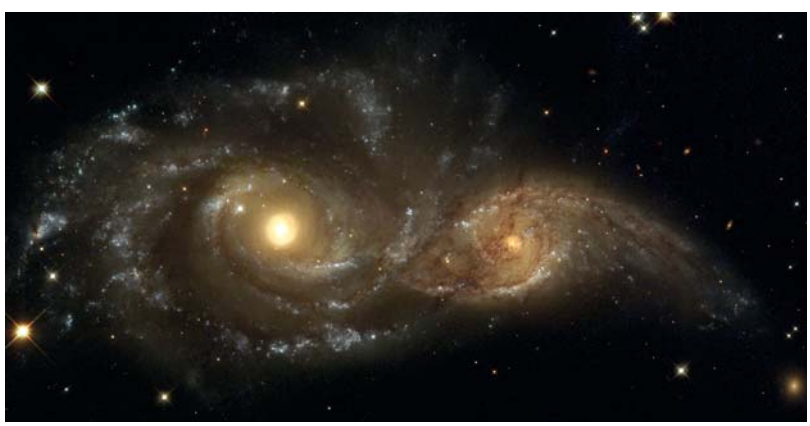


FIGURE 6.2 *Even two colliding galaxies conserve momentum.*

The impulse represents the "net effect" of a collision between two objects. It lumps together the acting force and its duration into a single parameter that is able to predict the change in momentum of the particle due to the collision. Figure 6.3 shows a plot of a typical interaction force on a particle as a function of time during a collision. The impulse represents the area under this curve, equal to the average force acting multiplied by the duration Δt.

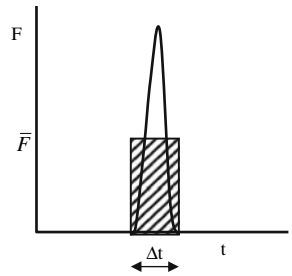


FIGURE 6.3 *Typical force acting on a particle during a collision. Usually the force is large and short-lived. The area under the curve equals the impulse, which is also equal to the product of the average force and the collision duration because the area under the rectangle equals that under the force curve.*

Suppose, for example, the two objects are identical, with the same mass m, and are traveling toward each other at the same speed v. Then, although each object has momentum mv, the net momentum before the collision is, in fact, zero. Do you see why? (Remember that momentum is a vector!) In this case, conservation of momentum predicts that the final momentum must be zero as well. There are two possible final situations for which the final momentum can be zero. In one case the two particles stick together and come to rest, whereas in the other case they bounce off each other and go off in opposite directions with the same magnitude of momentum that they had, and thus at the same speed. Although both of these situations conserve momentum, they differ in whether they conserve kinetic energy. The two particles that stick together and come to rest clearly have lost their kinetic energy, giving it up to other forms of energy such as sound and heat, because we know that ultimately energy must be conserved.

In more complex situations with two unequal mass objects traveling at different speeds, the algebra becomes a bit more involved and the possible outcomes will depend on whether kinetic energy is conserved. We do not dwell on these situations in detail, but simply point out that conservation of momentum offers a major additional tool in their study. In the third section of this chapter we generalize our formulation of conservation of momentum to more complex systems. A few examples should help you to appreciate the power of this new conservation law.

Example 6.3 A 60 kg boy dives horizontally with a speed of 2 m/s from a 100 kg rowboat at rest in a lake. Ignoring the frictional forces of the water, what is the recoil velocity of the boat?

Solution: Since there are no external horizontal forces acting (we have ignored the frictional resistance force of the water here), momentum is conserved as the boy dives off the boat. Because the initial momentum of the (boy + boat) system is zero, the total momentum immediately after the boy dives off the boat must also be zero so that the boy and the boat must have equal, but oppositely directed, momenta. Note that it is the momenta that must be equal and opposite, not the velocities. In equation form

$$\vec{P}_{\text{ini}} = \vec{P}_{\text{fin}}, \quad \text{with} \quad \vec{P}_{\text{ini}} = 0 \quad \text{and} \quad \vec{P}_{\text{fin}} = \vec{P}_{\text{boy}} + \vec{P}_{\text{boat}}.$$

We therefore have that

$$0 = (60 \text{ kg})\,(2 \text{ m/s}) + (100 \text{ kg})\, v_{\text{boat}},$$

so that the boat's recoil velocity is found to be 1.2 m/s in the direction opposite to the boy's velocity.

Example 6.4 Two ice skaters, both traveling at a speed of 5 m/s and heading straight toward each other, collide and lock arms together. If their masses are 80 g and 50 kg, find the velocity with which they move together after the collision.

(*Continued*)

Solution: There are no horizontal external forces acting, so therefore momentum is conserved and we know that the sum of the skaters' two initial momenta is equal to their combined final momentum. Initially their momentum is $P_{ini} = (80\text{ kg})\ (5\text{ m/s}) - (50\text{ kg})\ (5\text{ m/s}) = 150$ kg-m/s in the direction the 80 kg skater is traveling. When they lock together, their combined mass is 130 kg and we must have that

$$P_{fin} = (130\text{ kg})\ v_{fin} = P_{ini} = 150\text{ kg-m/s},$$

so that $v_{fin} = 1.2$ m/s in the direction the heavier skater was traveling.

In Example 6.3 we ignored the fluid medium and its frictional force. The surrounding fluid medium is often of primary importance. Let's turn our attention to the problem of animal locomotion and, in particular, the motion of sea creatures such as the squid or jellyfish. These creatures, and indeed all animals that swim or fly through a fluid medium, move by virtue of reaction forces provided by the surrounding fluid medium. The jellyfish propels itself by jet propulsion, ejecting a volume of water in a jet that provides a thrust force in the opposite direction. Fish and birds generate thrust in a more continuous fashion by pushing back on the fluid medium with fins or wings (Figure 6.4). In any case, we can analyze such locomotion in either of two ways: a difficult method using the detailed reaction forces or much more easily using momentum.

Let's discuss the jet propulsion of a jellyfish in order to derive an expression for the thrust propelling it. We can model the jellyfish as a balloon that fills with water and then collapses driving water out in a jet (Figure 6.5). Let the initial mass of water contained within the balloon be m_0 and suppose that the collapse results in a uniform rate of decrease of the mass, $\Delta m/\Delta t$. Then the rate at which momentum is ejected from the balloon will be

$$\frac{\Delta p}{\Delta t} = \frac{\Delta m}{\Delta t}\, v,$$

where we assume a constant velocity for the jet of water expelled. By Newton's second law, the rate of momentum ejection provides a net force, known in this context as the thrust. If we take the initial volume of the jellyfish to be that of a 0.1 m radius sphere filled with water of density $\rho = 1000\text{ kg/m}^3$, then $m_0 = (volume)\ (density) = 4/3\pi r^3\rho = 4.2$ kg of water. If this water is ejected in 1 s through a 1 cm radius circular

FIGURE 6.4 *Bird, rocket, or fish, propulsion is by thrust, a reaction force, that conserves momentum.*

aperture, then we can first calculate the velocity of water flow. This is found by assuming that the total volume of water flows out in a cylindrical jet whose length is proportional to the velocity; that is, $\rho AL = m_0$, where $L = vt$. Knowing the mass m_0, density, cross-sectional area $A = \pi(.01\text{ m})^2$, and time $t = 1$ s, we can calculate the water velocity to be $v = 13$ m/s first, and then

$$F = \frac{\Delta m}{\Delta t} v$$

to find about 55 N of thrust is generated. This is actually a greater force than the weight of the initial water contained in the balloon. A similar analysis can explain the thrust of a rocket or that of a bird, but a realistic analysis will be more complex because of the nonconstant velocities involved in the problem.

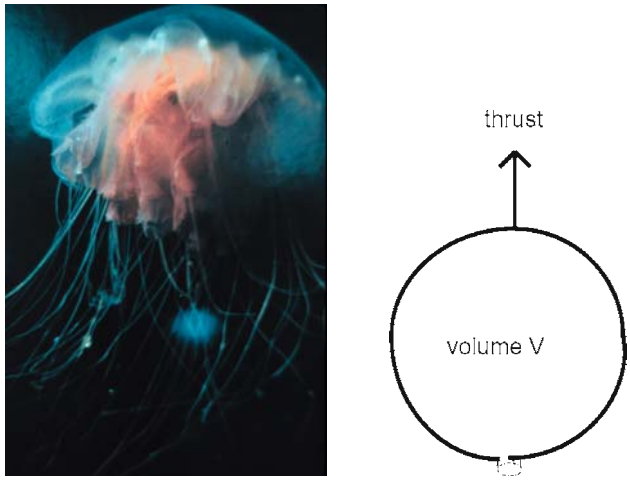


FIGURE 6.5 *Model of a jellyfish as a balloon.*

2. CENTER OF MASS

The simplest systems are composed of a single point particle introduced in the previous chapters. Here we begin our systematic study of increasingly more complicated systems of two particles, of many particles, or of a single extended object such as a person. For any system there is a well-defined point, the *center of mass*, at which the entire mass of the system can be considered to be concentrated in order to understand its translational motion.

A rough analogy to finding the center of mass can be made to locating the population center of the United States. Rather than weighting locations by their mass, they are weighted, in this case, by their local populations. This two-dimensional problem on a map could be attacked in a number of approximate ways, one of which we illustrate. Using census figures for the state populations and choosing some appropriate location as the population center within each state (e.g., by specifying latitude and longitude of its largest city) one could find the U.S. population center by separately averaging the latitude and longitude of the states, weighting each by its population. Thus California, Texas, and New York, together with more than one third of the U.S. population, dominate in the calculations and we expect to find a population center somewhere in the Midwest, even though the Midwest population is not particularly large. This example illustrates the notion of weighting locations by a local property or characteristic.

To introduce the definition of center of mass consider two particles of masses m_1 and m_2 attached by a light (massless) rod of length L as shown in Figure 6.6. If this system were tossed into the air it would translate and rotate about before landing on the ground. One special point, the center of mass of the system, would travel in the same trajectory as a single particle of mass $(m_1 + m_2)$ launched with the same initial velocity (we show this in the next section). Qualitatively this point can be imagined to be determined by finding the balance point along the rod. That is, imagine moving your finger along the rod until you can balance the rod with its masses on either end. That point is also the center of mass. For example, if $m_1 = m_2$ the center of mass would be located in the center of the rod at a distance of $L/2$ from either end. If $m_1 > m_2$, then the balance point would be closer to m_1, but how much closer?

Because the balance point in Figure 6.6 will be closer to the more massive particle, we want to define the center of mass as an average position of the two particles, with more massive particles counting more in the averaging process. We therefore define the center of mass along one dimension, x_{cm}, relative to an arbitrary origin, as

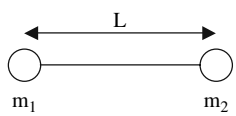


FIGURE 6.6 *Two masses separated by a light rod.*

$$x_{cm} = \left(\frac{m_1}{m_1 + m_2}\right)x_1 + \left(\frac{m_2}{m_1 + m_2}\right)x_2, \tag{6.6}$$

where x_1 and x_2 are the x-coordinates of masses m_1 and m_2, respectively.

Example 6.5 Find the center of mass of the Earth–moon system given that the mean radius of the Earth is 6.37×10^6 m, the mean radius of the moon is 1.74×10^6 m, the Earth–moon mean separation distance is 3.82×10^8 m, and that the Earth is 81.5 times more massive than the moon.

Solution: The Earth–moon separation is so much larger than the radius of either; therefore we can treat both bodies as point masses for the purposes of this calculation. With an origin at the center of the Earth (see Figure 6.7), we can write

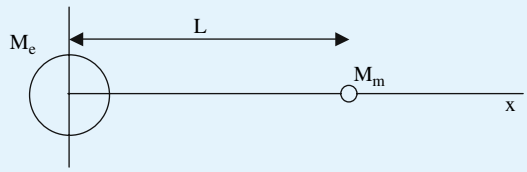


FIGURE 6.7 *The Earth–moon system.*

$$x_{cm} = \frac{M_e(0) + M_m(L)}{M_e + M_m} = \frac{1}{1 + \dfrac{M_e}{M_m}}L = 0.012\,L = 4.63 \times 10^6 \text{ m}.$$

Thus, the center of mass of the Earth–moon system actually lies within the Earth.

We can generalize this definition in a straightforward way to systems of more than two particles by simply adding terms for additional point masses in both the numerator and denominator in Equation (6.6). However, with more than two particles, the system need not be one-dimensional if the masses are not co-linear. In this case we can also define the y- (and z-, if needed) components of the center of mass in a similar way and combine them by writing $\vec{r}_{cm} = (x_{cm}, y_{cm}, [z_{cm}])$. Using the summation notation that Σ indicates to sum over all particles in the system, we can write (with a similar equation for z_{cm})

$$x_{cm} = \sum\left(\frac{m_i}{M}\right)x_i \quad \text{and} \quad y_{cm} = \sum\left(\frac{m_i}{M}\right)y_i, \tag{6.7}$$

where M is the total mass of the system. The subscript i denotes a particular numbered particle and the summation sign indicates that i is to be varied from number 1 to the total number of particles in the system while performing the additions indicated.

Example 6.6 Find the center of mass of a water molecule using the following data (Figure 6.8): radius of O = 0.14 nm, radius of H = 0.12 nm, bond length of O–H bond = 0.097 nm, and H–H angle subtended at O = 104.5°.

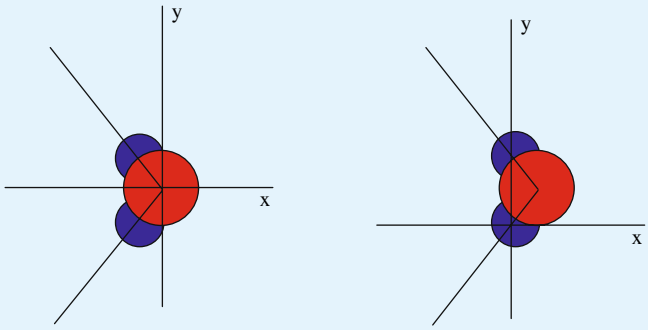


FIGURE 6.8 *Space-filling models of a water molecule with two different coordinate system origins.*

Solution: We solve this problem in two different ways using two different coordinate origins to see that the answer is independent of the chosen origin.

(1): In the first solution we set the origin on the O center and use the axes shown on the left. In this case the atoms have their centers located at: O (0,0); H ($-0.097 \cos 52.3°, \pm 0.097 \sin 52.3°$) = ($-0.059, \pm 0.077$), where the O–H bond is the center-to-center distance and we have found the x- and y-components of the H centers. We solve for the x- and y-coordinates of the center of mass (taking the masses of O and H as 16 and 1) by writing

$$x_{\text{cm}} = \frac{16(0) + 1(-0.059) + 1(-0.059)}{16 + 1 + 1} = -0.0066 \text{ nm},$$

and

$$y_{\text{cm}} = \frac{16(0) + 1(0.077) + 1(-0.077)}{18} = 0,$$

where the zero value for y_{cm} should be expected from the fact that the two H atoms are symmetrically situated above and below the x-axis.

(2): Using the coordinates shown on the right in the figure the atoms have their centers located at: O ($0.097 \cos 52.3°, 0.097 \sin 52.3°$) = (0.059, 0.077), H (0, 0) and H ($0, 2 \cdot 0.097 \sin 52.3°$) = (0, 0.15). Using the same basic relations we write

$$x_{\text{cm}} = \frac{16(0.059) + 1(0) + 1(0)}{18} = 0.053 \text{ nm},$$

and

$$y_{\text{cm}} = \frac{16(0.077) + 1(0) + 1(0.15)}{18} = 0.077 \text{ nm}.$$

Although these two answers appear at first glance to be different, the shift in origins must be accounted for in comparing them. The origin on the right is located at the point ($-0.059, -0.077$) with respect to the origin on the left and if we compare the actual spatial location of the center of mass in both parts (by, e.g., adding the origin coordinates on the right with respect to those on the left

(*Continued*)

to the answers in part 2), we find identical results, indicating a unique spatial location for the center of mass.

A different approach to the problem might be to first recognize that by symmetry the y_{cm} must lie along the x-axis using the left set of coordinates in Figure 6.8 and then to set up the problem as a two-mass system with the total H mass located at $x = -0.059$ nm and the O mass at the origin. Try it.

In the case of a solid extended object, if it is uniform throughout and has some symmetry we can often determine its center of mass by inspection, based on the notion of a balance point that was qualitatively introduced with Figure 6.6. For example, a uniform solid rod will have its balance point, or center of mass, at its geometric center. Even if an object has multiple parts, each of which is uniform throughout and has some symmetry, we can reduce the problem to finding the center of mass of a collection of particles, one for each part of the object with the mass of each part located at the center of mass of that part. This is illustrated in the following example.

Example 6.7 The solid objects shown in the three figures below are all made from the same uniform material and have the same thickness. In part (c) there is a small hole in the larger circular plate. Find the center of mass of each object using the coordinate system shown. Take $R = 0.1$ m and $L = 0.05$ m.

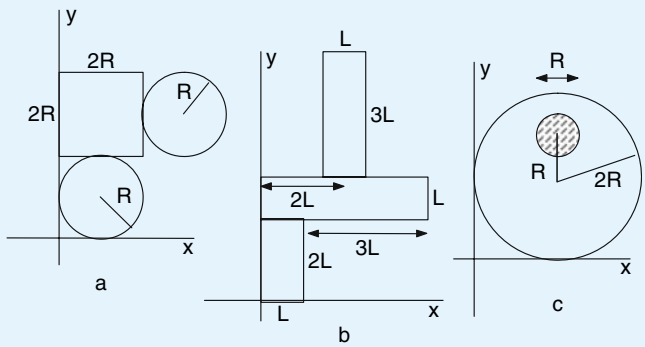


FIGURE 6.9 *Uniform solid objects for Example 6.7.*

Solution: Because all the objects are made of a uniform material and have the same thickness, their masses are simply proportional to their areas. This is true because the mass m is equal to the product of the density ρ of the material, its thickness t, and its area A, or

$$m = \rho t A.$$

Because both the density and thickness are constants, $m \propto A$, and furthermore, because in Equation (6.6) only the ratio of masses appears, we do not need to know the thickness or density of the materials as they cancel and do not appear in the final result. In what follows we therefore set the proportionality constant simply equal to 1 and numerically equate masses and areas.

We proceed by replacing each regular shape with a point mass having the same total mass as that portion of the entire object and located at its center of mass (these are found by inspection because the shapes are highly symmetric). Each problem then reduces to a set of point masses, all located in the same plane. In part (c) we use a trick: let the hole have a negative mass according to its size and superimpose the larger solid circular plate with the

smaller circular plate of "negative" mass, thus canceling the mass within the hole region!

In (a) we have the following three objects: $M = 4R^2 = 0.04$ located at $(R, 3R) = (0.1, 0.3)$; $M = \pi R^2 = 0.031$ at $(R, R) = (0.1, 0.1)$; and $M = \pi R^2 = 0.031$ at $(3R, 3R) = (0.3, 0.3)$. We then find that

$$x_{\text{cm}} = \frac{4R^2(R) + \pi R^2(R + 3R)}{4R^2 + 2\pi R^2} = 0.16 \text{ m},$$

and

$$y_{\text{cm}} = \frac{4R^2\,(3R) + \pi R^2\,(R + 3R)}{4R^2 + 2\pi R^2} = 0.24 \text{ m}.$$

In (b) we have three point masses: $M = 2L^2 = 0.005$ at $(L/2, L) = (0.025, 0.05)$; $M = 4L^2 = 0.01$ at $(2L, 2.5L) = (0.1, 0.125)$; and $M = 3L^2 = 0.0075$ at $(2L, 4.5L) = (0.1, 0.225)$. The center of mass is given by

$$x_{\text{cm}} = \frac{2L^2\,(L/2) + 4L^2\,(2L) + 3L^2\,(2L)}{9L^2} = 0.083 \text{ m},$$

and

$$y_{\text{cm}} = \frac{2L^2(L) + 4L^2\,(2.5L) + 3L^2(4.5L)}{9L^2} = 0.14 \text{ m}.$$

In (c), using the trick mentioned above, we have two point masses: $M = \pi(2R)^2 = 0.13$ at $(2R, 2R) = (0.2, 0.2)$; and $M = -\pi(R/2)^2 = -0.0079$ at $(2R, 3R) = (0.2, 0.3)$. Using the same method we find

$$x_{\text{cm}} = \frac{4\pi R^2\,(2R) \;-\; \pi\,\dfrac{R^2}{4}\,(2R)}{3.75\pi R^2} = 0.2 \text{ m},$$

as expected, and

$$y_{\text{cm}} = \frac{4\pi R^2\,(2R) \;-\; \pi\,\dfrac{R^2}{4}\,(3R)}{3.75\pi R^2} = 0.19 \text{ m}.$$

This last number seems reasonable because with a hole cut out we expect y_{cm} to be somewhat less than $2R = 0.2$ m.

You should go through each answer, following all the steps, and see that the center of mass position makes qualitative sense.

In the general case of nonuniform and/or nonsymmetric objects, the center of mass can always be found experimentally by finding the balance point along three mutually perpendicular axes, if possible, or by suspending the object separately from three different points, drawing vertical lines from those points, and looking for the intersection of the three lines (Figure 6.10). The reason why this latter method works becomes clearer after we have discussed rotational motion, but the center of mass must lie suspended vertically under the suspension point. Alternatively, one can use more complex mathematics to calculate the center of mass position. We show a more

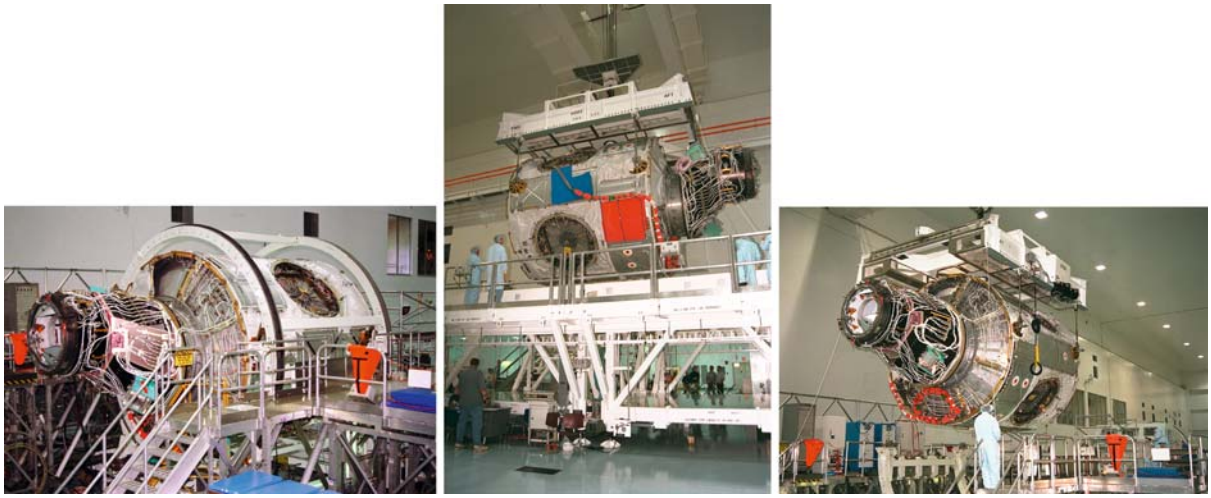


FIGURE 6.10 *Node I, a part of the International Space Station, being readied* (left) *and having its center of mass determined by suspending it from above.*

direct experimental approach to finding the center of mass in the next section. There we show that the center of mass translates about in space as if all external forces act directly on the entire mass of the system located at its center of mass.

3. CENTER OF MASS MOTION: NEWTON'S SECOND LAW AND CONSERVATION OF MOMENTUM

In the last section we learned how, in principle, to find the center of mass of any object, and in practice, to find that point for a collection of particle masses or symmetric objects. Here we show that the translational motion of a system of particles or an extended object is fully described by knowledge of the center of mass motion. The main goals of this section are to generalize Newton's second law for a particle to a very similar result for the center of mass of a system and to generalize the law of conservation of momentum.

The derivation of the generalization of Newton's second law to a system of extended objects is straightforward using some calculus (see box on the next page), but otherwise is cumbersome. The resulting *Newton's second law for a system* is

$$\vec{F}_{\text{net,ext}} = \Sigma \vec{F}_{\text{ext}} = M\vec{a}_{\text{cm}}, \tag{6.8}$$

where the sum Σ is over all of the external forces acting on the system, M is the total mass of the system (assumed constant; a system that does not exchange mass with its surroundings is known as a *closed system*), and $\vec{a}_{\text{cm}}$ is the acceleration of the center of mass. In this expression the only forces that produce an acceleration of the center of mass are forces exerted on the system by objects that are external to the system, so-called *external forces*. All of the *internal forces* between the particles of the system cancel pairwise because they are equal and opposite according to Newton's third law. This equation also applies to extended objects because they can be considered to be built up from particles. We conclude that the translational motion of a system can be completely described by replacing the entire system by a point mass with total mass M located at the system's center of mass, $\vec{r}_{\text{cm}}$, with only external forces acting (Figure 6.11).

FIGURE 6.11 (top) *A good high jumper has a center of mass that actually goes under the bar in well-defined two-dimensional free-fall motion while his flexible body goes over it.* (bottom) *An unmanned Titan rocket explodes shortly after takeoff in 1998. Despite fragmenting into many pieces the center of mass continues in a well-defined trajectory (see Example 6.8).*

As a byproduct of the derivation of Equation (6.8), we show (in the box) that the total momentum of the system, the vector sum of the individual particle momenta, is equal to the momentum of the center of mass, or the product of the total mass M and the center of mass velocity $\vec{v}_{cm}$,

$$\vec{P}_{\text{cm}} = M\vec{v}_{cm} = \Sigma\vec{p}_i. \tag{6.9}$$

Thus an alternative way to write Equation (6.8), in terms of the center of mass momentum, is

$$\vec{F}_{\text{net, ext}} = \lim_{\Delta t \to 0} \frac{\Delta \vec{P}_{\text{cm}}}{\Delta t}. \tag{6.10}$$

We see that the center of mass moves as if all the mass of the system is located there and experiences the net external force on the entire system. So, no matter whether the system is composed of a single extended object (such as the high jumper of Figure 6.11) or many independent parts (such as the exploded rocket in that figure), the center of mass of the system moves in a well-defined trajectory based on the total mass and the net external force on the system.

Written in this form we can deduce a very important consequence:

In the absence of a net external force on a system, the center of mass momentum, or total momentum of the system, does not change with time, and is said to be conserved.

This is a statement of the *principle of conservation of momentum*, a very powerful and general result, which holds for all isolated systems, those with no net external forces applied. It is a fundamental principle that holds on every scale of distance: on the atomic or nuclear scale as well as on the scale of the size of the universe. We saw a preliminary version of this in the first section of this chapter for collisions between two particles, but the principle is much more general than we saw there.

Conservation of momentum is the second of a handful of conservation laws that we study in this book. We have already learned the conservation of energy principle and seen its tremendous value as a tool in understanding motion. Later on we demonstrate its value in all other areas of physics that we study. Energy and momentum conservation are two of the cornerstones of physics. Because the momentum of an isolated system is constant, if we compute the total momentum at any time, its value at any other time will be the same vector, namely the same value and in the same direction. Just as with energy conservation, we can use our knowledge of the situation at one instant of time to find the total momentum, which will remain constant as long as there are no external forces acting. On the other hand, unlike energy conservation, momentum is a vector and therefore a direction as well as a magnitude is fixed in time.

We note that the kinetic energy of a particle $\text{KE} = \frac{1}{2}mv^2$ can in fact be rewritten in terms of the particle's momentum in place of its velocity. Using the definition of the magnitude of the momentum $p = mv$, we have that $\text{KE} = p^2/2m$. You need to keep in mind that although the kinetic energy, a scalar, can be written in terms of the square of the particle's momentum, the conservation laws of energy and of momentum are two different laws that keep different quantities constant. For a closed system (one with no exchange of mass with its surroundings) with no external forces acting, the total, or center of mass, momentum will be conserved as will the total mechanical energy. However, the kinetic energy of the system may change because it can be exchanged for potential

A derivation of Equation (6.8): Starting from a rewriting of Equation (6.6) for the x-component of the center of mass

$$M\vec{x}_{\text{cm}} = \Sigma m_i\vec{x}_i,$$

we can differentiate both sides of the equation with respect to time to find

$$M\vec{v}_{\text{cm}} = \Sigma m_i\vec{v}_i,$$

or

$$\vec{P}_{\text{cm}} = \Sigma\vec{p}_i,$$

where $\vec{P}_{\text{cm}}$ and $\vec{p}_{\text{i}}$ are the momentum of the center of mass and the individual particles. Thus we see that the center of mass momentum is equal to the total momentum of the system of particles. If we further differentiate this equation with respect to time we have

$$M\vec{a}_{\text{cm}} = \text{d}\vec{P}_{\text{cm}}/\text{d}t = \Sigma\text{d}\vec{p}_i/\text{d}t = \Sigma\vec{F}_{\text{i, net}},$$

where we have used Newton's second law for each particle, assumed that the total mass of the system is constant, and $\vec{F}_{\text{i,net}}$ is the net force on the ith particle. To complete the derivation of Equation (6.8), we note that the forces on particle i are of two types: external, arising from objects outside the system, and internal, arising from other particles in the system. These latter internal forces cancel pairwise in the summation because a force on particle 2 from particle 3 is equal and opposite to the force on particle 3 from particle 2 and all possible pairs of forces will be summed. The final result is Equation (6.8). Newton's law can be generalized still further to a system of extended objects with no change to Equation (6.8).

energy within the system. In the first section of this chapter we saw a few examples of the application of momentum conservation to the collision between two objects. Two quite different examples should help to provide an appreciation for the power of the conservation of momentum principle.

Example 6.8 A rocket of mass M explodes into three pieces at the top of its trajectory where it had been traveling horizontally at a speed $v = 10$ m/s at the moment of the explosion. If one fragment of mass $0.25\,M$ falls vertically at a speed of $v_1 = 1.2$ m/s, a second fragment of mass $0.5\,M$ continues in the original direction, and the third fragment exits in the forward direction at a 45° angle above the horizontal (see Figure 6.12), find the final velocities of the second and third fragments. Also compare the initial and final kinetic energies to see how much was lost or gained.

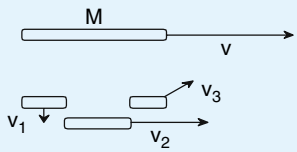


FIGURE 6.12 *The rocket before* (top) *and just after* (bottom) *the explosion of Example 6.8.*

Solution: Although the rocket is not an isolated system, the forces in the explosion are assumed to be so much greater than the weight of the rocket that we can neglect gravity at the moment of the explosion. This situation is very similar to that of any collision in which two objects interact very strongly for a very short time as, for example, when a tennis racket hits a ball. In all such cases we can neglect gravity during the collision and treat the system as isolated. Therefore the initial momentum of the rocket $P_{\text{ini}} = Mv$ in the horizontal direction must be conserved during the explosion and the sum of the momenta of the three fragments must add up to exactly this same P_{ini} value. Using vector addition, we can write that conservation of momentum in the horizontal direction implies

$$Mv = \frac{1}{2}Mv_2 + \frac{1}{4}Mv_3 \cos 45^\circ,$$

where the velocities are labeled as in the figure. Note that the first fragment falls vertically and does not contribute to this equation for the horizontal momenta. Conservation of momentum in the vertical direction gives a second equation

$$0 = \frac{1}{4}Mv_1 - \frac{1}{4}Mv_3 \sin 45^\circ.$$

Substituting that $v_1 = 1.2$ m/s, we find first, from the second equation above after canceling the common factor M, that

$$v_3 = \frac{1.2}{\sin 45^\circ} = 1.7 \text{ m/s},$$

and then, on substitution into the first equation above, that

$$10 = 0.5v_2 + 0.25 \cdot 1.7 \cos 45^\circ,$$

so that $v_2 = 19$ m/s.

The initial kinetic energy is $\text{KE}_i = \frac{1}{2}mv^2 = 50$ M, and the total final kinetic energy is $\text{KE}_f = \frac{1}{2}(M/4)(1.2)^2 + \frac{1}{2}(M/2)(19)^2 + \frac{1}{2}(M/4)(1.7)^2 = 90.8\,M$, both measured in J with M in kg. The kinetic energy has increased by over 80% with the

excess coming from the chemical energy released in the explosion. Kinetic energy alone is not conserved in this example, but momentum is. On the other hand, the general principle of conservation of energy is obeyed with the total mechanical, chemical, and other sources of energy remaining constant for the rocket.

In the above example we've seen how in an explosion we can use conservation of momentum to learn about the motion of the final pieces. Similarly in a collision between objects we can use conservation of momentum during the collision to learn about the final motions of the objects after collision. For microscopic objects that interact during a collision, of say atoms, the forces are all conservative and the collisions, aside from conserving momentum, tend to be elastic, conserving energy as well. In most cases of macroscopic objects colliding, the collisions tend to be inelastic, so that energy is lost (or gained in the explosion of the last example) even though momentum is conserved during the collision. The next two examples illustrate some of these possibilities.

Example 6.9 A hockey puck of 0.5 kg mass traveling at a speed of 5 m/s collides with an identical stationary puck in a glancing (not head-on) collision. If the first puck is deflected by 30° and travels with a final speed of 3 m/s, find the final velocity of the puck that was hit if it moves off at a 45° angle as shown. Ignore any friction between the ice and pucks.

Solution: Because there are no external horizontal forces acting, momentum is conserved. With the initial direction of motion chosen as the x-axis, the initial momentum is only

$$p_{ix} = mv_0 = (0.5)(5) = 2.5 \text{ kg m/s}.$$

After the collision, both pucks have x momentum (components of their momentum vectors) that must add up to the initial momentum as

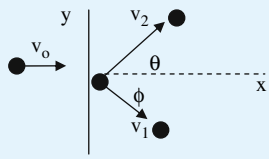
$$2.5 \text{ kg m/s} = (0.5 \text{ kg})(3 \cos 30 \text{ m/s}) + (0.5 \text{ kg})(v \cos 45),$$

where v is the final velocity of the second puck. Solving for v we find $v = 3.4$ m/s. Notice that energy is not conserved in this collision because the initial KE = 1/2 $(0.5)(5)^2 = 6.25$ J, whereas the sum of the final KE = 1/2 $(0.5)(3)^2$ + 1/2 $(0.5)(3.4)^2 = 5.1$, amounting to a loss of about 18% of the initial KE.

Example 6.10 Suppose one proton moving with a speed v_0 collides with a second proton initially at rest. If one of the protons emerges at a given angle ϕ from the incident direction, find the speeds of both after the collision and the angle θ

(*Continued*)

at which the second proton emerges from the collision. Work this out in general and then take $v_0 = 10^6$ m/s and $\phi = 30°$.



Solution: We are searching for three unknown quantities, and so require three independent equations. We work this problem out without substituting in numbers so that we can learn about the general case. These equations can be obtained from conservation of momentum (two equations, one for the incident direction, say x and one for the direction perpendicular to that, say y) and conservation of energy.

Conservation of momentum in the x-direction gives

$$p_{0x} = mv_0 = p_{fx} = mv_2 \cos\theta + mv_1 \cos\phi, \tag{1}$$

where m is the proton mass, and v_1 and v_2 are the protons' final velocities. Conservation of momentum in the y-direction gives

$$p_{0y} = 0 = mv_1 \sin\phi - mv_2 \sin\theta. \tag{2}$$

Energy conservation gives us the equation

$$\frac{1}{2}mv_0^2 = \frac{1}{2}mv_1^2 + \frac{1}{2}mv_2^2. \tag{3}$$

We now have our three equations in three unknowns—this was the physics part of the problem—and the remainder of the problem is to solve for them algebraically. This is a bit complicated, so follow closely. First we can cancel all the m's in all three equations and if we then solve for $v_2\cos\theta$ and $v_2\sin\theta$ in Equations (1) and (2) we have

$$v_2 \cos\theta = v_0 - v_1 \cos\phi,$$

and

$$v_2 \sin\theta = v_1 \sin\phi. \tag{4}$$

We can then square each of these and add them together to find, using $\sin^2\theta + \cos^2\theta = 1$, that

$$v_2^2 = (v_0 - v_1 \cos\phi^2) + (v_1 \sin\phi)^2,$$

but from Equation (3), after canceling 1/2 m from each term, we have that

$$v_0^2 = v_1^2 + v_2^2 = v_1^2 + (v_0 - v_1 \cos\phi)^2 + (v_1 \sin\phi)^2.$$

Expanding out the terms in parentheses and combining again we have

$$v_0^2 = 2\,v_1^2 + v_0^2 - 2\,v_0\,v_1 \cos\phi.$$

Simplifying this, we have

$$2v_1 (v_1 - v_0 \cos\phi) = 0,$$

which has the solutions $v_1 = 0$ or $v_1 = v_0\cos\phi$. The solution $v_1 = 0$ gives $v_2 = \pm v_0$ indicating a head-on solution in which one proton stops and the other goes on in the forward direction (we must reject the negative solution for v_2 as unphysical.) The other solution, from Equation (3), gives $v_2 = \pm v_0\sin\phi$. In that case to find ϕ, after substitution for v_1 in Equations (4) we have that

$$v_2 \cos\theta = v_0 - v_0 \cos^2\phi = v_0(1 - \cos^2\phi) = v_0 \sin^2\phi$$

and

$$v_2 \sin\theta = v_0 \sin\phi \cos\phi.$$

Dividing these equations we find that

$$\tan\theta = 1/\tan\phi.$$

Therefore, given an angle θ for the first proton, the second emerges such that $\theta + \phi = 90°$. In our case, if $v_0 = 10^6$ and $\phi = 30°$, we find $v_1 = 8.7 \times 10^5$ m/s, $v_2 = 5.0 \times 10^5$ m/s and $\theta = 60°$. You can check these by direct substitution into Equations (1)–(3), after canceling m.

In this chapter we have learned how to describe the translational motion of a system of extended objects using the center of mass and momentum conservation. In general such systems will have two other types of motion: overall rotational motion and internal motions. Internal motions include all relative motions of portions of the system other than overall rotational tumbling, including shape changes as well as vibrational motions. We come back to this topic much later in the book in discussions on the structure of matter. Rotational motion is taken up in detail in the next chapter.

CHAPTER SUMMARY

The momentum of a particle of mass m is defined as

$$\vec{p} = m\vec{v}. \qquad (6.1)$$

Using this definition, we can write Newton's second law for the particle in terms of its momentum as

$$\vec{F}_{\text{net}} = \lim_{\Delta t \to 0} \frac{\Delta \vec{p}}{\Delta t}. \qquad (6.2)$$

If two particles interact in the absence of any external forces, then their total momentum is conserved, meaning that it will remain a constant in time.

A useful concept in discussing collisions is the impulse, defined by the product of the collision force and its duration, and shown to equal the change in momentum of the object:

$$\text{Impulse} = F\Delta t = \Delta p = p_{\text{final}} - p_{\text{initial}}. \qquad (6.5)$$

For a collection of masses m_i, each located at (x_i, y_i), with total mass M, we define the center of mass to be located at the point

$$x_{\text{cm}} = \Sigma\left(\frac{m_i}{M}\right)x_i \quad \text{and} \quad y_{\text{cm}} = \Sigma\left(\frac{m_i}{M}\right)y_i. \qquad (6.7)$$

Then for a system of such masses, Newton's second law can be shown to be

$$\vec{F}_{\text{net, ext}} = \Sigma\vec{F}_{\text{ext}} = M\vec{a}_{\text{cm}}, \qquad (6.8)$$

or written in term of momentum, as

$$\vec{F}_{\text{net,ext}} = \lim_{\Delta t \to 0} \frac{\Delta \vec{P}_{\text{cm}}}{\Delta t}. \qquad (6.10)$$

In the case of an isolated system, with no external forces acting, the center of mass momentum, equal to the total momentum of the system, is conserved: $\vec{P}_{\text{total}} = \text{constant}$. This is a vector equation and, in general, stands for the three independent equations for which each component (x, y, and z) of momentum remains constant.

QUESTIONS

1. What are the differences and similarities between momentum and velocity? Between momentum and kinetic energy?
2. Is it possible for the center of mass of a solid object to lie physically outside the object? Give an example or two to support your assertion.
3. For uniform (constant density) objects, is it true that the center of mass must lie along a symmetry axis, if there is one? Give some examples.
4. Explain, in your own words, why only external forces result in a change in the center of mass momentum of a system of interacting particles or an extended object.
5. Carefully define an isolated system. Give some examples and explain why it is that momentum is only conserved for an isolated system.
6. Is a rocket traveling in outer space an example of an isolated system? If so, how can the rocket change its momentum if it is to be conserved?
7. Two identical twins of equal mass are ice skating toward each other at the same speed. What happens when they collide? What happened to their momentum?
8. In a collision of a tennis ball with a racket, why should the tension in the strings of the racket be made as large as possible?
9. When a collision between two objects occurs and there is a net change in the momentum of one object there are very large forces acting for a very short time. The product of the average force on the object during the collision and the duration of the collision is called the *impulse*. If a tennis ball of mass m and velocity v bounces off a wall and rebounds with the same speed, what is the impulse on the ball? Why does a new tennis ball bounce higher than an older tennis ball when dropped from the same height?

MULTIPLE CHOICE QUESTIONS

1. A 3 kg mass has position coordinates (−2, 2 m) and a 1 kg mass has position coordinates (3, 0 m). The center of mass of this system has coordinates (a) (1, 2 m), (b) (−3, 6 m), (c) (−0.75, 1.5 m) (d) (0, 0 m).
2. A 2 kg mass is at $x = 0$ m, $y = +2$ m, and a 3 kg mass is at $x = 2$ m, $y = 0$ m. The x- and y-coordinates, respectively, of the center of mass of this system are (a) +6/5 m, +4/5 m, (b) +2/5 m, +2/5 m, (c) 0, 0 m, (d) +2 m, +2 m.
3. A 5 kg bowling ball with a center of mass velocity of 4 m/s strikes the padded end of the bowling lane and comes to rest in 0.01 s. The average force exerted on the ball is (a) 400 N, (b) 2000 N, (c) zero, (d) 500 N.

Questions 4 and 5 refer to a car weighing 900 N that is heading north at 14 m/s. It makes a sharp turn and heads west at 18 m/s. During the turn, a good luck charm hanging from the rear view mirror is angled from the vertical for a total of 5 s.

4. The magnitude of the average force on the car during this time is (a) 720 N, (b) 73 N, (c) 420 N, (d) 210 N.
5. The direction of the average force on the car during this time is (a) 38° S of W, (b) 38° N of W, (c) 52° S of W, (d) 52° N of W.
6. An 80 kg man and a 40 kg girl are skating on smooth level ice. Initially, they are in contact and at rest. The man pushes the girl away from him with a force of 30 N. Immediately after they are no longer in contact the girl's speed is 2 m/s. At the same instant the man's speed (a) must be zero, (b) must be 2 m/s also, (c) must be 1 m/s, (d) depends on how much force the girl exerts on the man.
7. A 40 kg boy is standing on a 5 kg skateboard at rest. If he jumps off with a horizontal velocity of 1 m/s, neglecting friction the recoil velocity of the skateboard is (a) 1 m/s, (b) 0.1 m/s, (c) 0.03 m/s, (d) 8 m/s.
8. A 50 kg astronaut in orbit can give a 10 kg wrench a speed of 10 m/s by throwing it. The speed the astronaut will recoil with after doing so will be (a) 0 m/s, (b) 2 m/s, (c) 10 m/s, (d) 50 m/s.
9. In a head-on collision between a seagull and a jet airplane
   (a) The momentum of the airplane is exactly conserved.
   (b) The total kinetic energy is exactly conserved.
   (c) The magnitude of the change in momentum of the seagull divided by the collision time equals the magnitude of the average force on the jet.
   (d) The total momentum is zero.
   (e) None of the above is true.
10. A 0.1 kg meter stick has two masses attached: 0.3 kg at 20 cm and a 0.4 kg at 100 cm. The center of mass of the system lies at the following indicator on the meter stick. (a) 57.5 cm, (b) 63.8 cm, (c) 65.7 cm, (d) 70 cm, (e) none of the above.

PROBLEMS

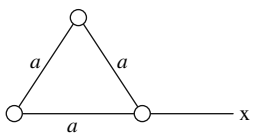
1. Find the center of mass of the following sets of point masses.
   (a) A 2 kg mass at $x = 5$ cm and a 5 kg mass at $x = -2$ cm
   (b) A 1 kg mass at $y = 0$ and a 4 kg mass at $y = 10$ cm
   (c) Three small objects each of the same mass m, located at the following points (0,0), (0,10 cm), (10 cm, 0)
   (d) Point mass m at (0,0), point mass $3m$ at (0, 5 cm), point mass $5m$ at (5 cm, 0) and point mass m at (5, 5 cm)
2. Using Table 6.1, find the center of mass of
   (a) A person standing upright with hands at sides
   (b) An outstretched arm and an arm bent upward at the elbow by a right angle
   (c) A person bent over so that there is a right angle between her straight legs and upper body/head and between her upper body and straight arms

Table 6.1 *Distances and Masses of Portions of the Typical Human Body (Expressed as % of Total Height and Mass)*

Hinge Points (from floor)	*Center of Mass (from floor)*	*Mass*
Neck 91.2	Head 93.5	6.9
Shoulder 81.2	Trunk/neck 71.1	46.1
Elbow 67.2	Upper arms 76.0	6.6
Hip 52.1	Lower arms 55.3	4.2
Wrist 46.2	Hands 43.1	1.7
Knee 28.5	Upper legs 42.5	21.5
Ankle 4.0	Lower legs 18.2	9.6
	Feet 1.8	3.4

3. From the text discussion, you know that the center of mass can be found through "balancing methods", that is, suspending an object from a point. This procedure indicates that for three equal masses situated at the vertices of an equilateral triangle, the center of mass will be at the intersection of the three angle bisectors of the triangle. From elementary geometry theorems, it is known that the three angle bisector segments intersect at a point that is 2/3 of a segment length away from its angle vertex. Calculate the center of mass for the mass arrangement shown and compare its position to the intersection of the angle bisectors.

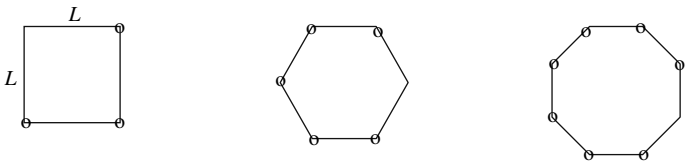
Note that because the height of the triangle is a 3/2, the method is a physical manifestation of the theorem that the bisectors of angles of an equilateral triangle intersect at the center of mass of the triangle (usually called the "centroid" by mathematicians). This is true whether the physical triangle is constructed of sides only, of similar and uniform cross-section, or if the triangle is a uniform plate.



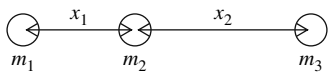
4. Calculate the center of mass for three equal masses situated at the vertices of a 3-4-5 right triangle.

5. Calculate the center of mass for the arrangement of three masses also situated at the vertices of a 3-4-5 right triangle, but where the masses are in the ratio 3:4:5, with the largest opposite the hypotenuse and the smallest opposite the shortest side. Compare the result with the previous problem.

6. By symmetry, the center of mass of a uniform regular polygon is at its center. Similarly, this is true for an arrangement of equal masses situated at the vertices of such a polygon. Where is the center of mass for each of the arrangements shown, where only a subset of the vertices of the polygon is occupied by masses? Make sure you arrange the coordinate framework to take advantage of any remaining symmetry. (Hint: Missing masses can be represented as $M - M = 0$ mass, so that the sum of a negative mass can be added to the situation with a full complement of masses at the vertices.)



7. Consider the three spherical masses shown. How far to the right of m_2 should m_3 be so that the center of mass of the entire arrangement is located exactly at the position of m_2?

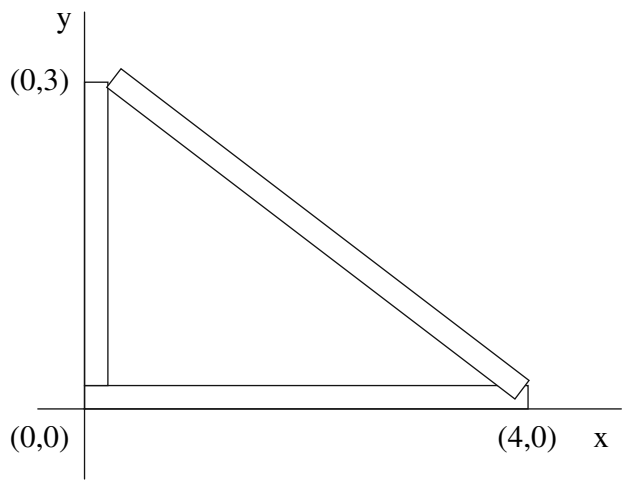


8. Consider a uniform linear arrangement of ten masses, that is, with equal spacing between, ranging from 1 to 10 kg, each 1 kg more than the previous. Where is the center of mass of the assemblage?

9. Three uniform spheres of radii R, $2R$, and $3R$ lie in contact with each other from left to right in the order given with their centers along the x-axis. Remembering that the volume of a sphere is given by $(4/3)\pi r^3$, find the position of the center of mass of the three spheres as measured from the left edge of the smallest sphere.

10. Find the center of mass of a screwdriver with the following characteristics: a wooden cylindrical handle (density of wood = 0.5×10^3 kg/m^3; cylinder length and diameter = 10 and 2 cm) and a steel cylindrical rod (density of steel = 7.8×10^3 kg/m^3; 15 cm long and 0.5 cm in diameter, with an additional 3 cm flat uniformly tapered head with a triangular cross-section).

11. Three uniform rods (identical except for their lengths) form the right triangle shown with coordinates measured in meters.



(a) Replace each rod by an equivalent point mass at its proper location, assuming a 1 m rod has a mass of 1 kg. Show these in a figure.
(b) Solve for the x- and y-coordinates of the center of mass using the given coordinate system.

12. A ball of 0.5 kg mass is dropped from rest at a height of 1 m. What is its momentum as it hits the ground?

13. A 0.1 kg ball bounces perpendicularly off a wall with the same speed of 5 m/s that it hit the wall.
(a) What is the change in momentum of the ball when it hits the wall?
(b) If the collision took 5 ms, what average force was exerted on the ball?
(c) Did the wall change its momentum, and if so, why doesn't it move?

14. Tennis pros can often serve the ball at speeds in excess of 125 mph. High-speed photography shows that the racket and ball make contact for about 4 ms. Find the average force that must be exerted on the ball to serve it at 125 mph. Use a mass of 0.05 kg for the tennis ball.

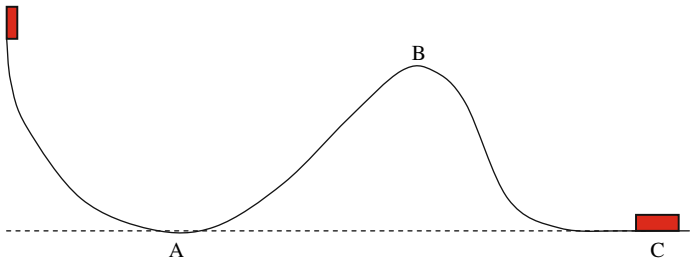
15. A rocket used for fireworks explodes just when it reaches its highest point in a vertical trajectory. It initially bursts into three fragments with masses of m, $3m$, and $4m$, each of these to explode slightly later. If the $4m$ fragment falls vertically with an initial velocity of 8 m/s, and the $3m$ fragment is ejected with a velocity of 10 m/s at an angle of 30° above the horizontal, find the velocity of the third fragment.

16. A 5 kg crate initially at rest is pushed along a frictionless horizontal surface by a 10 N force directed at an angle of 30° above the horizontal.
(a) Find the velocity of the crate after 5 s.
(b) If at this time (after the 5 s) the applied force is removed and the crate travels up a 30° incline with a coefficient of kinetic friction of 0.2, use the work energy theorem to find how far along the incline the crate travels before coming to rest.
(c) If the same process is repeated as in part (a) but this time after removing the applied force at $t = 5$ s, the crate collides with a horizontal spring 2 s later, compressing it a distance of 50 cm. Find the spring constant.
(d) If the crate in part (c) travels back along the same path after leaving the spring, and then collides and sticks to a similar 2 kg crate at rest, find the final velocity of the two crates after the collision.

17. A 10,000 kg railroad car traveling at a speed of 24 m/s strikes a 1200 kg automobile initially at rest on the track. Assume that the auto sticks to the railroad car after the collision.
(a) What is the speed of the auto–railroad car system immediately after the collision?
(b) What is the percentage loss in kinetic energy of the auto–railroad car system as a result of the collision?

After the collision the auto–railroad car system slides along the track with a coefficient of kinetic friction equal to 0.9.
(c) What is the frictional force?
(d) How much work will be done by the frictional force during the time the system takes to come to a halt?

18. A 200 kg roller coaster car falls on a circular portion of a frictionless track starting from rest from a height of 10 m, equal to the radius of the track, before reaching a second hill of 8 m height (point B).



(a) What is the centripetal force (magnitude and direction, please) on the car when it is at point A at the bottom of the circular track. Use energy ideas to get the velocity.
(b) Find the speed of the car at point B.
(c) If the car collides and locks together with a second car of mass 300 kg at point C, find the final speed of both cars.

19. A hockey puck traveling at 1.2 m/s collides with a second stationary equal mass puck and, after the collision, moves with a speed of 0.8 m/s deflected by an angle of 30°. Find the velocity (magnitude and direction) of the other puck after the collision. Also, find the fraction of the initial energy lost in the collision.

20. Alpha particles are routinely accelerated using a particle accelerator and are directed with the use of magnets into targets composed of various elements. A famous experiment called Rutherford's experiment has a beam of alpha particles incident on a target of gold. An alpha particle (a helium nucleus) is accelerated to a certain speed and makes an elastic head-on collision with a stationary gold nucleus. What percentage of its original kinetic energy is transferred to the gold nucleus?

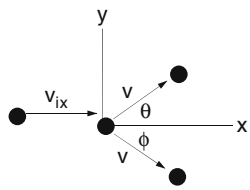
21. A ballistic pendulum is used to study the principles of momentum and energy. Suppose that a steel ball of mass $m = 50$ g traveling with an initial velocity V undergoes an inelastic collision with a stationary pendulum arm of length $R_{cm} = 30.5$ cm of mass $M = 250$ g. After the collision the center of mass of the ball and pendulum arm rises from its lowest point thought a height Δh_{cm}, where it momentarily comes to rest at an angle $\theta = 27°$.

(a) Write an equation that governs the momentum of the ball and pendulum arm during the collision and solve this for the initial velocity of the ball.
(b) After the collision, mechanical energy is conserved. Write an equation that shows conservation of mechanical energy immediately after the collision to the point where the pendulum arm and ball come to rest momentarily at the angle θ. Solve this equation for the velocity of the ball and pendulum arm after the collision. Express your answer in terms of R_{cm} and θ and you may ignore any rotational motion of the arm.
(c) Using the equations that you have written in parts (a) and (b) what is the expression for and the value of the initial velocity of the ball?
(d) What fraction of the initial kinetic energy of the ball has been lost in the collision?

22. An automobile has a mass of 2300 kg and a velocity of 16.0 m/s. It makes a rear-end collision with a stationary car whose mass is 1800 kg. The cars lock bumpers and skid off together with their wheels locked.
(a) What is the velocity of the center of mass of the two-car system?
(b) What is the velocity of the two cars just after the collision?
(c) What is the change in total kinetic energy during the collision?
(d) What is the magnitude of the impulse experienced by the 2300 kg car?
(e) If the duration of the collision is 0.100 s, what is the magnitude of the average force experienced by the 2300 kg car?
(f) What is the magnitude of the average force experienced by the 1800 kg car?

23. A 0.01 kg bullet traveling at 300 m/s ricochets off a stationary steel block of 2 kg mass. The bullet is deflected by 5° and travels at 250 m/s after the collision. Find the velocity (magnitude and direction) of the block after the collision.

24. A 10 g projectile is fired at 500 m/s into a 1 kg block sitting on a frictionless surface. The projectile lodges in the center of the block, and both move off together.
(a) What is the final velocity of the block after the collision?
(b) The block slides along the frictionless surface some distance and then encounters a ramp, which slopes up at an angle of 60°. What distance does the block travel along the surface of the ramp before coming to a stop?
(c) If the coefficient of friction between the block and the ramp is $\mu_k = 0.2$, how far does the block slide up the ramp before stopping?



25. A proton moving with an initial velocity v_{ix} in the x-direction, as shown in the figure, collides elastically with another proton that is initially at rest. If the two protons have equal speeds after the collision, what is the speed of each proton after the collision in terms of v_{ix}, and what are the directions of the velocity vector after the collision?

7 Rotational Motion

Once the translational motion of an object is accounted for, all the other motions of the object can best be described in the stationary reference frame of the center of mass. A reasonable image to keep in mind is to imagine following a seagull in a helicopter that tracks its translational motion. If you took a video of the seagull you would see quite different motion than you would from the ground. The seagull would appear always ahead of you but would rotate and change its "shape" as it flapped its wings (e.g., see the film *Winged Migration*). You've probably seen such wildlife videos that can track animals and "subtract" their translational motion leaving only the other collective motions about their "centers": lions seemingly running "in place" as the scenery flies by. In physics, we've already shown how to account for the translational motion of the center of mass. Aside from a possible constant velocity drift in the absence of any forces, motion of the center of mass is caused by external forces acting on the object. We now turn to the other motions about the center of mass as viewed from a reference frame fixed to the center of mass.

These collective motions are of two types: coherent and incoherent. *Coherent* motions are those overall rotations or vibrations that occur within a solid in which the constituent particles making up the object interact with each other in a coordinated fashion. If the solid is rigid (with all the internal distances between constituent parts fixed) the only collective motion will be an overall rotation about the center of mass. For such a rigid body, a complete description of its motion includes the translational motion of the center of mass and the rotational motion about the center of mass. Because this nice separation of the problem can be made, we first present the description, or kinematics, of pure rotational motion of a rigid body about a fixed axis, the *axis of rotation*. In this case all points of an object rotate in circles about some fixed point on the axis of rotation. This type of motion occurs, for example, when a door is opened, or for the wheels of a stationary bicycle, or when you lift an object by rotating your forearm about a stationary elbow. Even if the solid is not rigid, its collective coherent motions can be described as a rigid body rotation (of the average-shaped body) as well as other coherent internal motions that can change the object's shape.

We next introduce the energy associated with rotational motion and the rotational analog of mass, known as the moment of inertia. We show that well-placed and directed forces can produce rotational motion and we introduce the notion of *torque*, the rotational analog of a force. For pure rotational motion there is an equation that is the rotational analog of Newton's second law that can describe the dynamics of motion. Continuing with rotational analog quantities we introduce angular momentum, the rotational analog of (linear or translational) momentum and learn a new fundamental conservation law of angular momentum. Key in following the presentation of our understanding of rotational motion is to keep in mind the strong analogy with what we have already learned. A preview glance at Table 7.2 below shows that the important new concepts in this chapter all have direct analogs with equations we have already studied.

One of the new and revolutionary types of microscopy, atomic force microscopy, is discussed as an application of the material in this chapter. The technique allows

J. Newman, *Physics of the Life Sciences*, DOI: 10.1007/978-0-387-77259-2_7,

extremely high resolution maps of the microscopic surface topography, or structure, of materials and has been used extensively to study biological molecules and cells.

After briefly considering the effects of diffusion on the rotational motion of macromolecules, the chapter concludes with a study of the special case of objects in static equilibrium. This is an important simplification of Newton's laws and provides a powerful method of analyzing equilibrium situations.

The other category of collective motions is known as *incoherent*. These are random motions of the atoms of the material, about the equilibrium positions in a solid or with no fixed average position in a fluid. Constituent fluid particles move about much more independently, in the ideal case not interacting with their neighbors at all. In Chapters 8 and 9 we discuss the flow of ideal fluids as well as some of the complications that occur in complex fluids in which there are strong interactions between constituents. Later in Chapters 12 and 13, we study the subject of thermodynamics concerned with describing the fundamental thermal properties of macroscopic systems. We show that collective incoherent internal motions of an object give rise to an internal energy that is responsible for its temperature.

1. ROTATIONAL KINEMATICS

A rigid body—one with a fixed shape—has motions that are limited to pure translation and pure rotation about its center of mass. Combinations of these can give rise to motions that appear more complex, such as rolling, but which can be simplified to pure rotations in a reference frame fixed to the center of mass. Since we've already learned how to handle translational motion in the previous chapter, here we first take up the problem of pure rotational motion about a fixed axis of rotation, leaving their synthesis leading to general rigid body motions for a discussion later in the chapter.

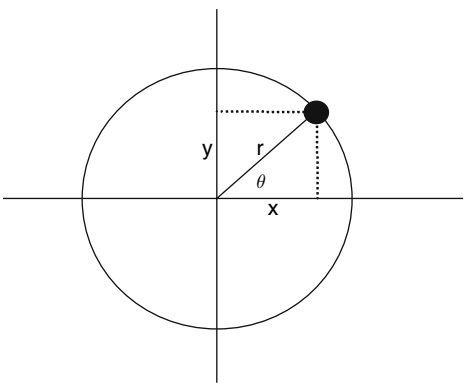
Consider the motion of a point particle on the circumference of a circle, as shown in Figure 7.1. In order to describe its position and motion we could use its x- and y-coordinates or, better, its r and θ polar coordinates. These latter coordinates are preferred because r is constant if the particle remains on the circle and so in polar coordinates there is really only one variable θ, whereas both x and y change as the particle moves on the circle. To describe the motion of the particle on the circle we could use its x- and y-components of velocity, both of which would continuously change, or, even better, we could use the θ-component of velocity known as the angular velocity ω, whose average value is defined as

$$\overline{\omega} = \frac{\Delta\theta}{\Delta t}. \tag{7.1}$$

In this expression, the particle has moved between two angular positions in a time Δt, where the angular displacement $\Delta\theta$ must be measured in radian units and not in degrees.

The fundamental unit of angular measure is the radian, because it is defined as the ratio of the arc length s to the radius r as $\theta = s/r$, and is a pure number with no units. This definition of an angle in radians leads to the fact that there are 2π radians in one revolution around a circle, given the circumference of $s = 2\pi r$. In one complete revolution there are also 360° and thus the radian is equal to $360°/2\pi$ or about 57.3°. The unit for ω is, according to Equation (7.1), the radian per second (rad/s); despite the fact that the radian unit is a pure number and we could write the units for angular velocity as 1/s, it is useful to retain the term "rad" in the numerator. (Note: Most pocket calculators can do calculations using either radians or degrees and because rad must be used here, some care must be taken when first using a new calculator.)

FIGURE 7.1 *A particle executing circular motion.*



If our particle travels in a circle at a constant speed, executing uniform circular motion, then the instantaneous value of ω is constant and equal to the average value. Notice that because ω is an angular variable there are really only two possible directions of travel: clockwise or counterclockwise around the circle. Just as the sign (+ or −) for a linear quantity depends on the coordinate system,

we are free here to label the sign of the angular velocity in an arbitrary way, as long as we are self-consistent within the context of any particular discussion.

In the more general case, when the particle does not travel at a constant speed, the angular velocity will vary and we need to introduce the concept of the instantaneous angular velocity, defined in a similar way to v, as

$$\omega = \lim_{\Delta t \to 0} \frac{\Delta \theta}{\Delta t}. \tag{7.2}$$

The distance traveled by the particle along the circumference Δs is proportional to the angular displacement $\Delta\theta$ from the relation $s = r\theta$, therefore we have that $\Delta s/\Delta t = r\,\Delta\theta/\Delta t$ so that

$$v = r\omega, \tag{7.3}$$

where r is the radius of the circle. We also need to introduce the concept of an angular acceleration α in order to account for a changing ω in analogy with our introduction of a linear acceleration a to describe changes with time in the linear velocity v. We define the average and instantaneous angular acceleration in direct analogy with their linear counterparts as

$$\overline{\alpha} = \frac{\Delta \omega}{\Delta t}; \quad \alpha = \lim_{\Delta t \to 0} \frac{\Delta \omega}{\Delta t}. \tag{7.4}$$

Again, because Equation (7.3) implies the change in the magnitude of the linear velocity is proportional to the change in the angular velocity, we have a relationship between the linear and angular accelerations,

$$a_{\text{tang}} = r\alpha. \tag{7.5}$$

The linear acceleration in Equation (7.5) is called the tangential acceleration and is directed parallel (or antiparallel) to the tangential velocity. It is the tangential acceleration that is responsible for changing the speed of the particle executing circular motion. Don't confuse the tangential acceleration with the centripetal acceleration,

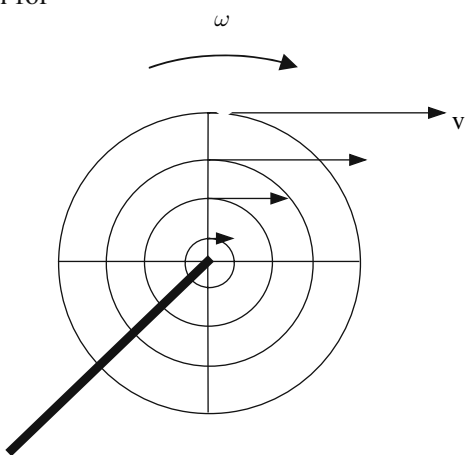
$$a_{\text{cent}} = \frac{v^2}{r},$$

discussed earlier in connection with circular motion in Chapter 5. Even in the case of uniform circular motion, where the speed and ω are constant and so $a_{\text{tang}} = 0$, there is a nonzero radially directed centripetal acceleration required to steer the object around the circle. If a_{tang} is not equal to zero the particle's speed will change as it travels in circular motion and it will have both tangential and radial components of acceleration.

It is useful to rewrite the expression for the centripetal acceleration in another equivalent form. For extended objects such as a wheel, the velocity of different parts of the wheel will be different, depending on their distance from the axis of rotation (Figure 7.2). For this reason it is more useful to rewrite the expression for the centripetal acceleration in terms of ω using Equation (7.3)

$$a_{\text{cent}} = \frac{v^2}{r} = \omega^2 r. \tag{7.6}$$

FIGURE 7.2 *A rotating wheel with its increasing velocity with increasing distance from the axis of rotation.*



Having introduced the angular variables, θ, ω, and α, needed to describe rotational motion, we are now in a position to derive a set of equations among these variables in the case of constant angular acceleration as we did in Chapter 3 when the linear acceleration was constant (see Table 3.1). Because we have the proportionality of s and θ, v and ω, as well as a and α, we can proceed by simply dividing each of the linear variables in the kinematic relations of Table 3.1 by the radius of the circle r to arrive at a set of kinematic equations for the angular variables:

$$\omega(t) = \omega_o + \alpha t; \tag{7.7}$$

$$\theta(t) = \theta_o + \omega_o t + \frac{1}{2}\alpha t^2; \tag{7.8}$$

$$\omega^2 = \omega_o^2 + 2\alpha(\theta - \theta_o). \tag{7.9}$$

These three equations serve as a basis for describing pure rotational motion with constant angular acceleration just as their linear counterparts were used in Chapter 3.

Example 7.1 A stationary exercise bicycle wheel starts from rest and accelerates at a rate of 2 rad/s^2 for 5 s, after which the speed is maintained for 60 s. Find the angular speed during the 60 s interval and the total number of revolutions the wheel turns in the first 65 s.

Solution: Using Equation (7.7) we find that the angular speed is given by

$$\omega = 0 + \alpha t = 2 \cdot 5 = 10 \text{ rad/s}.$$

In the first 5 s, the wheel rotates through

$$\theta = \frac{1}{2}\alpha t^2 = \frac{1}{2} \cdot 2 \cdot 5^2 = 25 \text{ rad},$$

and in the next 60 s the wheel rotates through an additional

$$\theta = \omega_o t = 10 \cdot 60 = 600 \text{ rad},$$

because there is no acceleration during this interval of time. The 625 rad total corresponds to

$$\theta = \frac{625 \text{ rad}}{2\pi \text{ rad/rev}} = 99.5 \text{ rev}.$$

The key to solving this problem was to divide the total time interval into two portions, only one of which had an acceleration.

Example 7.2 Helical bacterial flagella drive E. coli at constant speed when they rotate around counterclockwise (CCW, as viewed from behind the bacterium) at a uniform angular velocity, appearing much like a corkscrew (Figure 7.3). From time to time the flagella motor reverses to clockwise (CW) rotations, causing the flagella to disorganize themselves and the bacterium to tumble, before switching

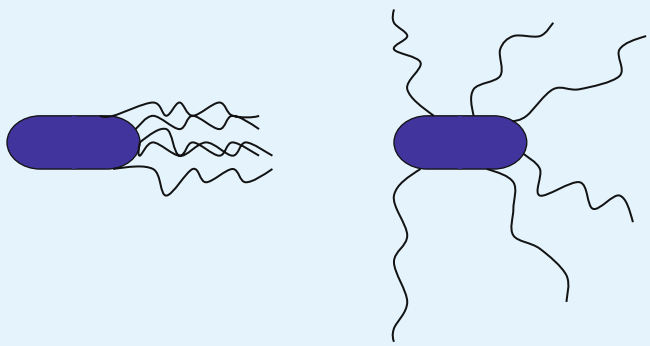


FIGURE 7.3 (left) *Coordinated flagella lead to swimming;* (right) *disordered flagella lead to "twiddling".*

again to CCW rotation so the bacterium swims off in a different direction. Suppose that the flagella "motor" rotates with a frequency of 4 Hz (4 rotations per second) when in either the CCW or CW state. If the flagellum spends 98% of its time in the CCW state and takes 5 ms to reverse its rotation (with the CCW to CW transition occurring on average every 5 s), find the average angular acceleration during a CCW to CW transition and the net angular rotation in a 10 s interval.

Solution: We are given that in a 5 ms interval, the flagellum reverses its rotation from an angular velocity of $\omega_o = -2\pi(4)$ rad/s to $\omega = 2\pi(4)$ rad/s. Therefore using the equation $\omega(t) = \omega_o + \alpha t$, we find that the average angular acceleration is

$$\overline{\alpha} = \frac{\omega - \omega_o}{t} = \frac{2\pi(4 - (-4))}{5 \times 10^{-3}} = 1.0 \times 10^4 \text{ rad/s}^2.$$

In an average 10 s interval, because the flagellum spends 98% of its time in the CCW state, it will spend only 0.2 s in the CW state, or 0.1 s in each of two CW intervals since the transitions occur every 5 s on average. For the 98% of the time in a CCW state, the flagella rotate at 8π rad/s producing a net rotation of $\theta = \omega t = 8\pi(9.8) = 246$ rad. During each of the other 0.1 s, there will be an acceleration to the CW state during 5 ms, a stay of 90 ms in that state, and an acceleration back to the CCW state for 5 ms (for a total of 100 ms = 0.1 s). We need to compute the net rotation during this time and multiply by 2 for the two such 0.1 s intervals. But by symmetry, the two 5 ms intervals will produce exactly opposite net rotations, canceling their contributions, and leaving only the $\theta = \omega t = -8\pi(.09) = -2.3$ rad contribution for each of the two intervals. Adding up the angular contributions, we have

$$\theta_{\text{net}} = 246 - 2(2.3) = 241 \text{ rad} = \frac{241}{2\pi} \text{ rev.} = 38.4 \text{ rev.}$$

Such studies on bacteria flagella have led to an increased understanding of the detailed energy sources and molecular interactions necessary for motility.

2. ROTATIONAL ENERGY

Now that we have a set of rotational variables to describe the kinematics of rotational motion, we take up the description of rotational dynamics of a rigid body. We begin our study of rotational dynamics in this section with a discussion of the rotational kinetic energy for an object with a fixed axis of rotation. Recall that in this case all parts of the object rotate about this axis in circular motion. To begin, consider a single particle when it is moving in a circle (Figure 7.4). Using our expression for kinetic energy, $KE = {}^1/_2 mv^2$, and the fact that the velocity can be written in terms of the angular velocity and the radius of the circle, $v = \omega r$, we can write that $KE = {}^1/_2 m(\omega r)^2$. Defining

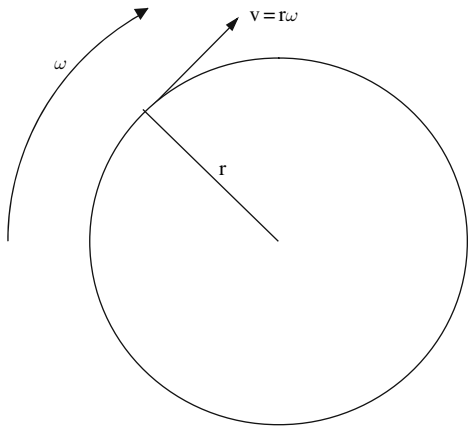


FIGURE 7.4 *A particle traveling in a circle.*

$$I = mr^2, \qquad \text{(for a single particle)}, \qquad (7.10)$$

where I is called the *moment of inertia* (for reasons that will become clear), we can write the kinetic energy of our particle as

$$KE = \frac{1}{2} I\omega^2. \qquad (7.11)$$

Note that this equation has a form similar to that of translational kinetic energy if we make a correspondence between the angular velocity ω and the linear velocity v and between the moment of inertia I and the mass m. We call the type of KE in Equation (7.11) rotational kinetic energy and discuss it further below after generalizing to the rotational motion of an extended rigid object. Note that if the particle travels in a uniform circular motion, its rotational kinetic energy is a constant, but if there is a tangential force acting on it as well as a centripetal force, then there will be a tangential acceleration and the angular velocity will change as will the rotational kinetic energy.

Example 7.3 A 25 kg girl riding on the outer edge of a large merry-go-round with a 10 m diameter has a (rotational) kinetic energy of 20 J. Find the girl's moment of inertia relative to the axis of rotation and find the number of revolutions the merry-go-round makes per minute.

Solution: The girl's moment of inertia, calculated as if she were a point mass, is given as

$$I = mr^2 = 25\ (5)^2 = 625\ \text{kg-m}^2.$$

To proceed we first calculate the angular velocity of the girl (and merry-go-round) using the expression for the rotational kinetic energy, Equation (7.11), so that

$$\omega = \sqrt{\frac{2(KE)}{I}}.$$

We find that $\omega = 0.25$ rad/s so that in 60 s the girl has gone around an angle $\theta = \omega t$ of 15 rad, corresponding to 15 rad/2π = 2.7 rev.

FIGURE 7.5 *Physics on a merry-go-round.*

Although this example does not deal with a point mass, we simplified our analysis to that case. Next, we want to generalize our discussion to extended rigid bodies. As a simple model, consider a rigid collection of point masses m_i attached together by "massless" rods with a fixed axis of rotation. If the assembly rotates about this axis, each will circle about a common central axis at some radius r_i. Using the summation convention, we can write the total kinetic energy of the collection as the sum of the kinetic energies of all the particles

$$\mathrm{KE} = \sum \mathrm{KE}_i = \frac{1}{2}\sum m_i v_i^2, \tag{7.12}$$

and, because $v_i = r_i\omega$, we have that

$$\mathrm{KE} = \frac{1}{2}\sum (m_i r_i^2)\omega^2, \tag{7.13}$$

where the summation is only over the different masses at their corresponding perpendicular distances from the axis of rotation, because the angular velocity of all the particles is the same. We define the summation to be the moment of inertia of the assembly of masses about the axis of rotation

$$I = \sum (m_i r_i^2), \tag{7.14}$$

so that the total kinetic energy may still be written in the simple form of Equation (7.11).

To analyze the more realistic model of a rotating extended rigid body, rather than a collection of particles, we can follow a procedure where we divide the object up into small elements of mass m_i, each at corresponding distances r_i from the axis of rotation (Figure 7.6). We can first ask how it is that a single force, applied to the rigid body at a localized point, can make the entire body rotate. When the external force is applied, internal forces that keep the object rigid do the appropriate amount of work on each localized mass element to maintain the shape of the body as it rotates. These internal forces are actually transmitted by electromagnetic interactions but for our purposes can be imagined to be transmitted via very stiff springs between molecules, the ultimate mass elements. At this point one can imagine how the spring properties can affect the overall rigidity of the solid and give rise to changes in the shape of an object when external forces act on it. When the applied forces become larger, our rigid body will eventually become deformed. We have already briefly studied the deformation of solids in Chapter 3, where we introduced the Young's modulus as well as the shear and bulk modulus to describe different deformations.

The division of a solid body, rotating about a fixed axis, into small mass elements allows a similar argument as in the case of discrete point masses and leads to an identical expression for the kinetic energy in Equation (7.11). The moments of inertia of various rigid objects with some symmetry are shown in Table 7.1. (See the boxed discussion for a calculation of one case.) Note that in every case the moment of inertia is equal to the product of the total mass and the square of the pertinent dimension apart from a numerical factor that depends on the geometry as well as the axis of rotation.

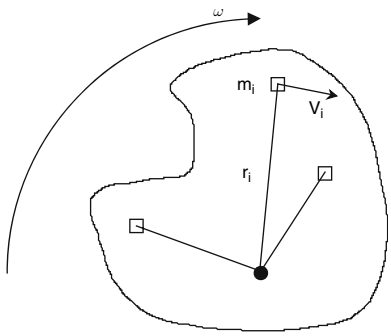


FIGURE 7.6 *A rigid body with discrete mass elements undergoing an overall rotation.*

For a continuous rigid body the definition of the moment of inertia given in Equation (7.14) needs to be rewritten as

$$I = \int r^2 \mathrm{d}m,$$

where r is the perpendicular distance of the differential element of mass from the axis of rotation. If the object has a constant mass density ρ then this can be rewritten as

$$I = \rho \int r^2 \mathrm{d}V,$$

where $\mathrm{d}V$ is the volume element containing mass $\mathrm{d}m$. As an example we calculate the moment of inertia of a right circular cylinder of radius R and length L about its axis (see Figure 7.7). We divide the cylinder into volume elements that are cylindrical shells of radius r, length L, and thickness dr. All of the mass in this shell has the same r and therefore the same I. The volume of the shell is $\mathrm{d}V = 2\pi rL\mathrm{d}r$, so that the integral becomes

$$I = \rho \int_0^R 2\pi rL \cdot r^2 \mathrm{d}r = 2\pi\rho L \int_0^R r^3 \,\mathrm{d}r.$$

This expression integrates to yield $I = 2\pi\rho R^4 L/4$, and can be rewritten in terms of the total mass of the cylinder, $M = \rho\pi R^2 L$, as $I = \frac{1}{2} MR^2$, giving the expression in the table.

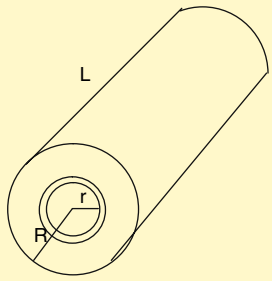
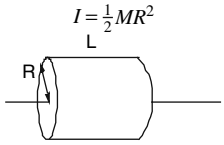
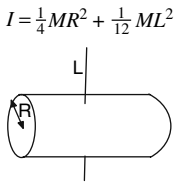
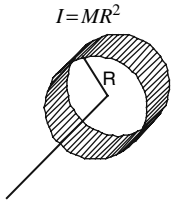
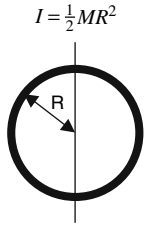
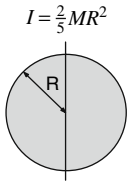
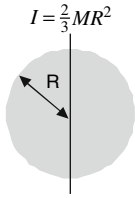
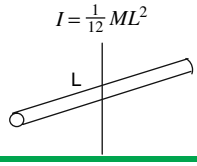
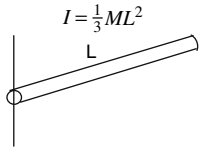


FIGURE 7.7 *Construction used to calculate the moment of inertia of a rod.*

Table 7.1 *Moments of Inertia of Various Symmetrical Objects*

SOLID CYLINDER about symmetry axis 	**SOLID CYLINDER about central diameter** 
HOOP about symmetry axis 	**HOOP about any diameter** 
SPHERE about any diameter 	**SPHERICAL SHELL about any diameter** 
LONG ROD about perpendicular axis at center 	**LONG ROD about perpendicular axis at end** 

Example 7.4 Calculate the moment of inertia of the gadget shown in Figure 7.8. The small masses are attached by a light rigid rod and pivot about the left end of the rod. Use a value of $m = 1.5$ kg and $d = 0.2$ m. If the assembly were to pivot about its midpoint, find the moment of inertia about this axis as well.

Solution: Using Equation (7.14), we simply add up the individual contributions to the moment of inertia. With the pivot point at the left end, we find

$$I = m(2d)^2 + 2m(3d)^2 + 3m(5d)^2 = 97md^2 = 5.8\ kg \cdot m^2,$$

and with the pivot point at the middle of the assembly, we find

$$I = m\left(\frac{d}{2}\right)^2 + 2m\left(\frac{d}{2}\right)^2 + 3m(2.5d)^2 = 19.5md^2 = 1.2\ kg \cdot m^2.$$

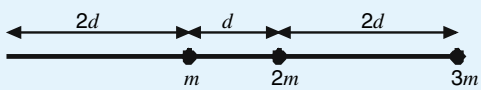


FIGURE 7.8 *Gadget of Example 7.4 with three point masses attached by "massless" rods.*

It should make intuitive sense, after a moment's thought that I should be smaller in the second case, because the masses are traveling in smaller radii circles. From Equation (7.11), for the same angular velocity in both situations, we expect there to be less kinetic energy in the second case, in agreement with the smaller I.

Example 7.5 Find the moment of inertia of the object shown in Figure 7.9 when pivoted about its symmetry axis. The cylinder has a mass M, radius r, and length L, whereas the hoop has a mass $M/10$ and radius $3r$. Use $M = 0.1$ kg, $r = 5$ cm, and $L = 25$ cm.

Solution: The moment of inertia of the hoop is simply the product of its mass and the square of its radius since all its mass lies at the same radius. The moment of inertia of the cylinder cannot be found so simply because its mass is distributed over varying distances from the axis of rotation. Using Table 7.1 we look up its moment of inertia and then write the total moment of inertia as the sum of the hoop's and the cylinder's as

$$I = \frac{1}{2}Mr^2 + \frac{M}{10}(3r)^2 = 1.4Mr^2 = 3.5 \times 10^{-4}\, kg \cdot m^2.$$

Note that the length L of the cylinder is not in the answer, only its total mass and radius.

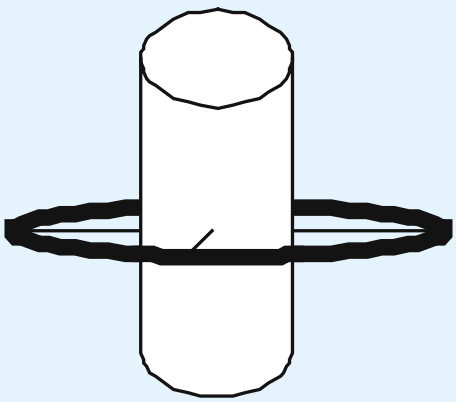
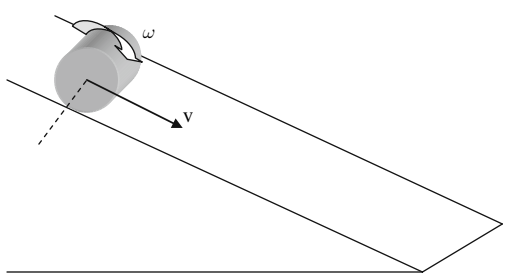


FIGURE 7.9 *A solid cylinder and a hoop connected by light rods both rotating about their common symmetry axis.*

Now that we have introduced moment of inertia and rotational kinetic energy for an extended object that is rotating about a fixed axis of rotation, we are in a position to generalize these ideas to the case of rolling motion. A wheel or other symmetric object that rolls can be shown to have a total kinetic energy that consists of two parts: the translational kinetic energy of the center of mass plus the pure rotational kinetic energy of the object about a fixed horizontal axis through its center of mass (see Figure 7.10). When all the forces acting on a system of rigid bodies are conservative so that the work done by those forces can be expressed as a potential energy difference, we can write the conservation

FIGURE 7.10 *A cylinder rolling down an inclined plane has a total kinetic energy equal to the sum of its center of mass translational KE and its rotational KE about the center of mass axis.*



of energy equation for the system, composed of translating, rotating, or rolling symmetric rigid bodies, as

$$\frac{1}{2}mv^2 + \frac{1}{2}I\omega^2 + PE = E = constant. \qquad (7.15)$$

(Conservation of Mechanical Energy)

The sum is the constant total mechanical energy of the system with the first two terms representing the total translational kinetic energy of the center of mass of all objects in the system and the pure rotational motion about the center of mass of each rotating object.

If there are also nonconservative forces present, such as friction, then the right-hand side of Equation (7.15) will no longer be a constant but will decrease with time because of the work of friction, just as in the translational motion situations we studied in Chapter 5. A few examples help us to see how to apply conservation of energy principles in order to study problems with rotational motion.

Example 7.6 An empty bucket of 1 kg mass, attached by a light cord over the pulley for a water well, is released from rest at the top of the well. If the pulley assembly is a 15 cm uniform cylinder of 10 kg mass free to rotate without any friction, find the speed of the bucket as it hits the water 12 m below.

Solution: The initial energy of the bucket–pulley system can be taken as pure gravitational potential energy, measured with respect to a zero level at the water surface. When the bucket just reaches the water the final energy is the sum of kinetic energy of the bucket (translational KE) and pulley (rotational KE). With no frictional forces present, the initial and final energies are equal and we can write

$$mgh = \frac{1}{2}mv^2 + \frac{1}{2}I\omega^2,$$

where m is the mass of the bucket, v its velocity as it hits the water, ω the angular velocity of the pulley as the bucket hits the water, and I the moment of inertia of the pulley, given by $I = \frac{1}{2}Mr^2$, where M is the pulley mass and r is its radius.

The bucket's velocity and the pulley's angular velocity are related by $v = \omega r$ because the cord is wrapped around the pulley at radius r and does not slip, so that we can rewrite our energy equation as

$$mgh = \frac{1}{2}mv^2 + \frac{1}{2}\left(\frac{1}{2}Mr^2\right)\omega^2 = \frac{1}{2}\left(m + \frac{1}{2}M\right)v^2.$$

Solving for the bucket's velocity

$$v = \sqrt{\frac{2mgh}{\left(m + \frac{1}{2}M\right)}}.$$

Substitution of numbers results in

$$v = \sqrt{\frac{2 \cdot 9.8 \cdot 12}{1 + 0.5 \cdot 10}} = 6.3 \text{ m/s}.$$

Note that this result is independent of the radius of the pulley. If the pulley were massless, then we would find that

$$v = \sqrt{\frac{2mgh}{m}} = \sqrt{2gh} = 15.3 \text{ m/s},$$

FIGURE 7.11 *A bucket suspended from a pulley and falling into a well.*

considerably faster. Because energy is conserved in both cases, why does the bucket have a much larger KE with a massless pulley and where does the missing energy go in the original problem? The smaller translational KE of the bucket is due to the relatively large rotational KE of the pulley just before the bucket hits the water.

Example 7.7 Suppose that a hoop and a cylinder, with the same radius and mass, both roll down an inclined plane, with an inclination angle θ, from rest at a height H without slipping. With what velocity does each arrive at the bottom and which will arrive first?

Solution: We can use the conservation of energy principle to solve this problem. The initial energy of each object is the same gravitational potential energy $E_i = mgH$. After rolling down the incline, the final energy of each is purely kinetic, equal to the sum of the center of mass translational KE and the rotational KE, $E_f = {}^1/_2\, mv^2 + {}^1/_2\, I\omega^2$, where I is the moment of inertia of each object about an axis through its center. Because there is no slipping, there is no loss of energy due to friction and the total energy of each object is conserved. We then can write that $E_i = E_f$ or

$$mgH = \frac{1}{2}\, mv^2 + \frac{1}{2}\, I\omega^2.$$

Since there is no slipping, we can also relate the center of mass velocity v to the angular velocity of each object through the same relation $v = r\omega$, where r is the radius of the hoop or cylinder. Then, on substituting for ω $(= v/r)$, we have that

$$v^2 = \frac{2mgH}{(m + I\,/\,r^2)}.$$

(*Continued*)

Looking up in Table 7.1 that $I_{hoop} = mr^2$ (it's easy to see why this is so, because all the mass of the hoop lies the same distance r from its center) and $I_{cylinder} = {}^1/_2\, mr^2$, we have that

$$v_{hoop} = \sqrt{\frac{2mgH}{(m+m)}} = \sqrt{gH}$$

$$v_{cylinder} = \sqrt{\frac{2mgH}{(m+m/_2)}} = \sqrt{\frac{4gh}{3}},$$

Note carefully that the final velocities do not depend on the radius of the object, the mass, or on the inclination angle. From our expressions it is clear that the cylinder arrives at the bottom of the incline with a faster speed. Also, rotational kinematics tells us that $\Delta t = \Delta\theta/\omega_{average}$ and because they both start from rest and accelerate uniformly, $\omega_{average} = \omega_{final}/2$, so that the object with the greatest final angular velocity will reach the bottom fastest. With equal radii, the cylinder clearly wins the race. Without rotational motion, both of these objects would take the same time to slide down the incline, arriving with the same speed. The rotational motion takes up some of the translational kinetic energy into rotational kinetic energy about the center of mass. The object with the greater moment of inertia gains the greater rotational kinetic energy and therefore loses the most translational energy and loses the race as well!

3. TORQUE AND ROTATIONAL DYNAMICS OF A RIGID BODY

We turn now to the mechanism by which rotational motion is produced. In order to have an object translate from rest, we require a net force to act. But a force, no matter how large, is not necessarily able to make an object at rest rotate. Consider the example shown in Figure 7.12a in which a door is to be opened. Pushing on the hinged side of the door with F_1, no matter how hard, will not open the door; similarly, pushing on the edge of the opened door toward the hinges with F_2 will also not result in any rotation of the door. Thus, it is clear that a force will not produce rotational motion unless it is well-placed and well-directed.

To clarify what is meant by well-placed and well-directed, consider the same door with force F_3 applied, also shown in a top view in Figure 7.12b. The force acts in the horizontal plane at an angle θ with respect to the horizontal position vector from the

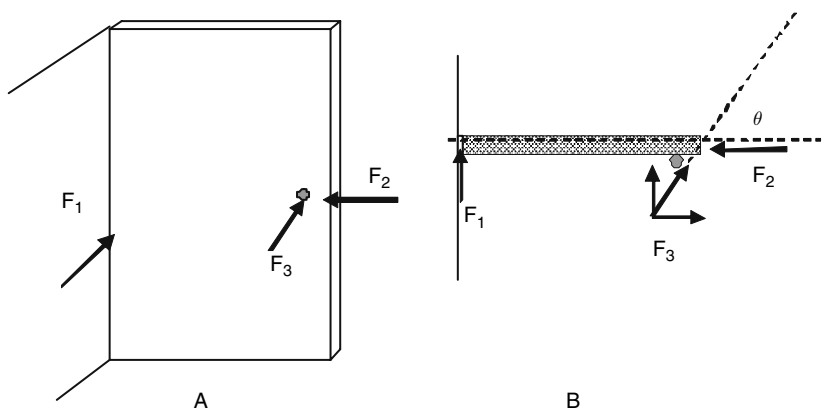


FIGURE 7.12 *(**A**) Door, hinged at the* left, *pushed more or less effectively in different directions and at various locations. (**B**) A* top *view of the door.*

axis of rotation to the point of application. If we imagine taking the components of this force along and perpendicular to the position vector, it is clear that only the perpendicular component will result in rotation of the door. The component parallel to the position vector, the so-called radial component, itself will not result in rotation of the door no matter how large it is. Furthermore, if the same force is exerted on the door at a closer distance to the hinge, it is less effective in rotating the door. In the limit of applying the force directly on the hinge, no rotation at all will occur no matter in what direction the force is aimed.

Having shown the need for care in defining the quantity that "drives" objects to rotate, let's first look at the work–energy theorem in the case of rotational motion. From Chapter 4 we know that the net work done by external forces on an object is equal to the change in its kinetic energy. If the object is confined to rotate about a fixed axis of rotation, any change in its kinetic energy must be in its rotational kinetic energy and, in that case, we can write

$$W_{\text{net, ext}} = \Delta KE_{\text{rot}} = \Delta\frac{1}{2}I\omega^2, \tag{7.16}$$

where I is the total moment of inertia of the object with respect to the axis of rotation. Imagine a short interval of time Δt, during which a net force does an amount of work ΔW to produce a change in angular velocity from ω to $\omega + \Delta\omega$. In this case we can write the work–energy theorem as

$$\Delta W_{\text{net, ext}} = \frac{1}{2}I(\omega + \Delta\omega)^2 - \frac{1}{2}I\omega^2. \tag{7.17}$$

Expanding the term in brackets, we can rewrite this as

$$\Delta W_{\text{net, ext}} = I\omega\ \Delta\omega + \frac{1}{2}I(\Delta\omega)^2.$$

Because we are interested in taking the limit as the time interval approaches zero, in which case so does $\Delta\omega$, we neglect the second term on the right (which will be much smaller than the first) and rewrite our expression as

$$\Delta W_{\text{net, ext}} = I\omega\frac{\Delta\omega}{\Delta t}\Delta t.$$

Writing $\omega\Delta t = \Delta\theta$ and $\Delta\omega/\Delta t = \alpha$, which are correct in the limit as Δt approaches zero, we have

$$\Delta W_{\text{net, ext}} = I\alpha\ \Delta\theta. \tag{7.18}$$

Now for our case of pure rotational motion we know that all points of the object rotate in circles. From the general definition of work, $\Delta W = (F_{\text{net,ext}})_x\Delta x$, a net external force acting on a particle that is traveling in a circle will do an amount of work given by

$$\Delta W = F_\perp r\Delta\theta,$$

where $F_\perp$ is the component of the net applied force that acts along the tangential displacement direction and $\Delta x = s = r\Delta\theta$ is the distance over which the force acts (see Figure 7.13). If we define the rotational analog of the force, known as the torque, or as the moment of the force, to be

$$\tau = F_\perp r, \tag{7.19}$$

we obtain the analog expression for the work done for pure rotational motion

$$\Delta W = \tau_{\text{net, ext}}\,\Delta\theta. \tag{7.20}$$

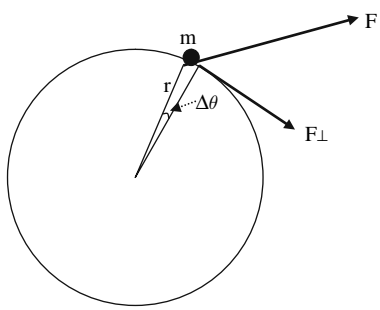


FIGURE 7.13 *Torque on a particle in circular motion.*

The units for torque are N-m, the same units as those for work or energy. This follows directly from Equation (7.19), or from Equation (7.20) since $\Delta\theta$ is dimensionless, but because torque and energy are different concepts, we never write a torque in units of joules, but always use N-m. Before discussing torques in more detail, let's first introduce the rotational analogue equation to Newton's second law.

On comparing Equations (7.18) and (7.20), we see that

$$\tau_{\text{net,ext}} = I\alpha, \tag{7.21}$$

which is the rotational analog of Newton's second law. Note that τ, I, and α are the rotational analogs of F, m, and a, respectively. The moment of inertia, and not the mass, enters into the rotational version of Newton's second law, therefore not only the mass of the system, but also its distribution from the axis of rotation is important in determining the response of the system to an applied torque.

With these results in hand let's first examine Newton's second law, in both rotational and translational forms, for the simplest case of the rotational motion of a single particle of mass m. Suppose the particle is located a distance r from the axis of rotation, attached to the center of the circle by a light rod, and is set in rotational motion by a force F acting as shown in Figure 7.13. In that case $I = mr^2$ and we have from Equation (7.21) that

$$\tau = mr^2\alpha.$$

From Equation (7.19), the torque on the mass is given by

$$\tau = rF_{\perp},$$

so that the torque depends on three factors: the magnitude of the applied force, where it is applied, and its orientation with respect to r, a line perpendicular from the axis of rotation to the point of application of the force. Only the perpendicular component of F contributes to the torque's ability to make the particle rotate around the circle and we have

$$rF_{\perp} = mr^2\alpha.$$

The outward radial component of the force must be more than balanced by a large inward radial force supplied by the light rod that is required to keep the mass traveling in a circle. The net inward radial force is then the centripetal force.

An alternate description of the rotational motion can be given by analyzing the tangential forces and accelerations. The tangential component of the force produces a tangential acceleration. Newton's second law in the tangential direction lets us write that

$$F_{\perp} = ma_{\text{tang}} = m\alpha r,$$

in agreement with the previous equation. Although tangential forces and accelerations can be used in the simplest rotational problems, the first approach uses the natural variables to describe rotational motion, angular acceleration and torque. With more than a single particle in the system, if the distances from the axis of rotation are different for the particles, in general it will be much more difficult to analyze the problem in terms of linear variables and much easier in terms of rotational variables. This is also true for extended real objects that are not treated as particles. Two example problems illustrating the application of the rotational form of Newton's second law help to make this discussion more concrete.

Example 7.8 Let's reconsider the problem of opening a door as discussed at the beginning of this section. Suppose the door is uniform and has a mass $m = 10$ kg, a height $h = 2.5$ m, and a width $w = 1$ m. The moment of inertia of a uniform rectangular slab (with dimensions $w \times h$) about a vertical axis of rotation along one of the edges is given by $I = \frac{1}{3}mw^2$, independent of h (those of you who know some calculus might try to derive this following the method used in the boxed calculation in Section 1). Suppose that the door is pushed with a steady horizontal force $F = 5$ N acting at the edge of the door and directed at a constant 30° angle from the normal to the door as it is opened (The force changes direction as the door is opened to keep the angle with respect to the door constant at 30° as shown in Figure 7.14). Find the angular acceleration of the door and the time for it to swing fully open, rotating a total of 90°.

Solution: A steady torque acts to push the door, thereby producing a constant angular acceleration. The torque is given by

$$\tau = F_{\perp} r = F\cos 30\, w = 5 \cos 30 = 4.3\ N \cdot m,$$

where the perpendicular component of F is obtained from the figure and the distance r equals w in this case. The constant angular acceleration of the door is given by

$$\alpha = \frac{\tau}{I} = \frac{4.3}{\frac{1}{3}(10)(1)^2} = 1.3\ \text{rad/s}^2.$$

Using this acceleration, we can find the time for the door to swing by 90° ($= \pi/2$ rad) angle to be

$$\theta = \tfrac{1}{2}\alpha t^2 \text{ or } t = \sqrt{\frac{2\theta}{\alpha}} = \sqrt{\frac{2 \cdot (3.14/2)}{1.3}} = 1.6\ \text{s}.$$

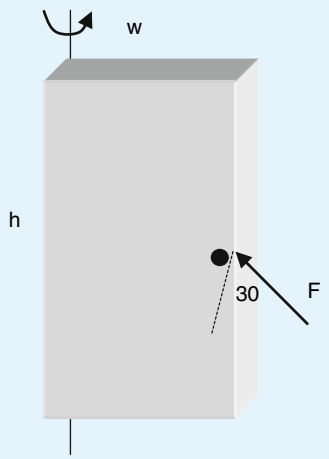


FIGURE 7.14 *How long does it take to open a door?*

Example 7.9 An ultracentrifuge is spinning at a speed of 80,000 rpm. The rotor that spins with the sample can be roughly approximated as a uniform cylinder of 10 cm radius and 8 kg mass, spinning about its symmetry axis (so that, from Table 7.1, $I = \frac{1}{2}mr^2$). In order to stop the rotor in under 30 s from when the

(*Continued*)

motor is turned off, find the minimum braking torque that must be applied. If no braking torque is applied, the rotor will stop in 30 min. Find the frictional torque that is present under normal spinning conditions.

Solution: We first need to find the minimum angular acceleration needed to stop the rotor in under 30 s. Using

$$\omega = \omega_0 + \alpha t,$$

we find that to stop the rotor requires an angular acceleration of

$$\alpha = -\frac{\omega_o}{t}.$$

Because ω_o is given as 80,000 rpm, we first must convert it to rad/s, $\omega_o = (80{,}000)2\pi/60 = 8.38 \times 10^3$ rad/s, where the factor 2π converts revolutions to radians and the factor 60 converts from minutes to seconds. Then we find that with $t = 30$ s,

$$\alpha = -\frac{\omega_o}{t} = \frac{8.38 \times 10^3}{30} = -279 \text{ rad/s}^2.$$

The braking torque must have a magnitude of at least

$$\tau = I\alpha = \tfrac{1}{2} mr^2\alpha = \tfrac{1}{2} 8(0.1)^2(279) = 11.2\, N \cdot m.$$

In the absence of a braking torque, we recalculate the angular acceleration using $t = 30$ min to find an acceleration 60 times smaller, $\alpha = -4.7$ rad/s^2, so that the normal frictional torque has a magnitude 60 times smaller as well, or $\tau = 0.19$ N-m.

Figure 7.15 shows that the expression for the torque can be written in two equivalent ways:

$$\tau = rF_\perp = r(F \sin\theta), \tag{7.22}$$

or, by regrouping terms,

$$\tau = (r \sin\theta)F = r_\perp F. \tag{7.23}$$

We see that the torque can be calculated by either taking the product of r, the distance from the axis of rotation to the point of application of the force, and the component of the force perpendicular to r, or by taking the product of F and the component of r perpendicular to F, known as the *moment* (or *lever*) *arm*. The moment arm is the perpendicular distance from the axis of rotation to the *line of action* of the force (the line along which the force is applied). Two additional examples clarify the calculation of torques and their use in rotational motion problems.

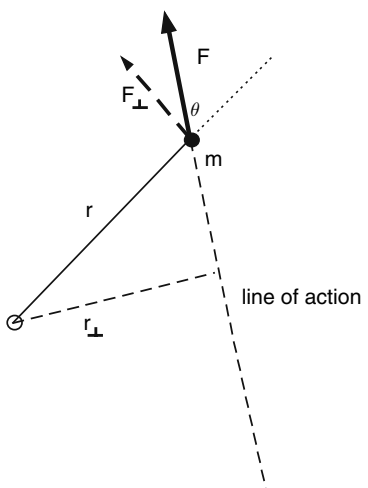


FIGURE 7.15 *Equivalent definitions of torque:* $\tau = rF_\perp = r_\perp F$.

Example 7.10 Calculate the forces that the biceps muscle and the upper arm bone (humerus) exert on a person's forearm when supporting a weight as shown in Figure 7.16 without any movement. The forces acting on the forearm include

FIGURE 7.16 *A person's arm supporting a weight and the force diagram for Example 7.10.*

its weight mg, the weight of the object held in the hand Mg, the pull of the biceps muscle F_{biceps}, and the humerus connection at the elbow socket, F_{hum}. Take the weight to be 20 N, the length L of the (uniform) forearm as 40 cm, its mass as 2 kg, with the biceps connecting $d = 4$ cm from the elbow pivot point, and assume that the arm is held at 40° with respect to the vertical.

Solution: To calculate the two unknown forces, we must realize that the net force and net torque on the forearm must both be zero because the weight is held at rest. This example anticipates the subject of statics, which we take up in Section 7 below. If we add up the net force and set it equal to zero and set the net torque equal to zero as well, we will obtain two independent equations that will allow us to solve for the two unknown forces. Only three of the four forces produce a torque about the elbow because the force from the humerus acts at the elbow joint and has zero lever arm. The torque equation is

$$\tau_{\text{net}} = Mg(L \sin 40) + mg\left(\frac{L}{2} \sin 40\right) - F_{\text{biceps}}(d \sin 40) = 0.$$

The lever arm distances were obtained from the distances along the forearm from the elbow pivot point by taking the horizontal components, those perpendicular to the vertical forces. We can then solve for the biceps force directly; canceling the common term sin 40,

$$F_{\text{biceps}} = \left(Mg + \frac{mg}{2}\right)L/d = \frac{(20 + 2 \cdot 9.8/2)0.4}{0.04} = 300\ N.$$

The force from the humerus can be obtained by summing the forces on the forearm to zero

$$Mg + mg + F_{\text{hum}} - F_{\text{biceps}} = 0,$$

to find that $F_{\text{hum}} = 260$ N. Note that to lift a relatively small 20 N weight requires very large forces on the bones and muscles of the body.

Example 7.11 Find the net torque about both the left end (A) and the center (B) of the uniform rod shown in Figure 7.17 with the set of external forces shown. Use the following values $F_1 = 30$ N, $F_2 = 20$ N, $Mg = 20$ N, $F_3 = 10$ N, $F_4 = 15$ N, a rod length $L = 40$ cm, with F_2 acting at $L/3$ from the right end.

(*Continued*)

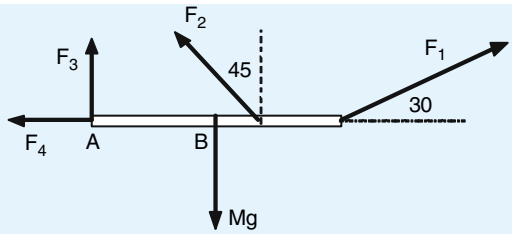


FIGURE 7.17 *A set of forces acting on a uniform rod.*

Solution: We first note that both F_3 and F_4 do not produce a torque about the left end (A) of the rod. Adding the torques from the other three forces about the left end, we find a net torque

$$\tau_{\text{net, A}} = Mg\frac{L}{2} - F_2 \cos 45\,\frac{2L}{3} - F_1 \sin 30\, L.$$

In this expression we took torques tending to rotate the rod clockwise about the left end as positive and used the perpendicular components of the forces in the expressions for the torques. Substituting in the numbers, we find

$$\tau_{\text{net, A}} = 20(0.4/2) - 20(\cos 45)(2/3)(0.4) - 30(\sin 30)(0.4) = -5.8\ N\cdot m,$$

where the negative sign indicates that the net torque would produce a counterclockwise rotation about the left end.

Repeating this procedure taking torques about the center (B), note that F_4 and Mg do not produce any torque and we have

$$\tau_{\text{net, B}} = F_3\frac{L}{2} - F_2\cos 45\,\frac{L}{6} - F_1\sin 30\,\frac{L}{2}.$$

Substituting in numbers, we find

$$\tau_{\text{net, B}} = 10(0.2) - 20(\cos 45)\frac{0.4}{6} - 30(\sin 30)\frac{0.4}{2} = -1.9\ N\cdot m.$$

Why do we get these two different results when taking torques about two different points with the same set of forces acting? First, note that calculating torques explicitly depends on the reference point. In fact, because the net torque is not equal to zero, the rod is not in equilibrium. So although the torque calculations are both correct, let's discuss which one we would use to describe the motion of the rod. The actual motion of the rod can be separated into a translation of the center of mass due to the nonzero net force (you should check that there is both an upward and leftward net force) and a rotation about the center of mass. If we were to move with the center of mass then we would see a pure rotation of the rod about its center and then the net torque about the center would be equal to the moment of inertia of the rod about its center times its angular acceleration. We could then find this angular acceleration and combine it with the translational acceleration of the center of mass to describe the overall motion of the rod.

As an example of rotational motion in an important biological macromolecule, let's discuss some of what is known about the world's smallest rotary motor, an enzymatic protein, F1-ATPase, which helps in the efficient production of ATP in cells. Discovered in 1956, this protein is found in virtually identical form in species ranging from bacteria to mammalian cells. Figure 7.18 shows both a schematic drawing and a ribbon model of the protein structure. The central γ subunit acts as a shaft able to rotate within the array of alternating α and β subunits arranged in a circle. This protein is a reversible rotary motor. Normally, when driven to rotate at very high rotational speed of several thousand

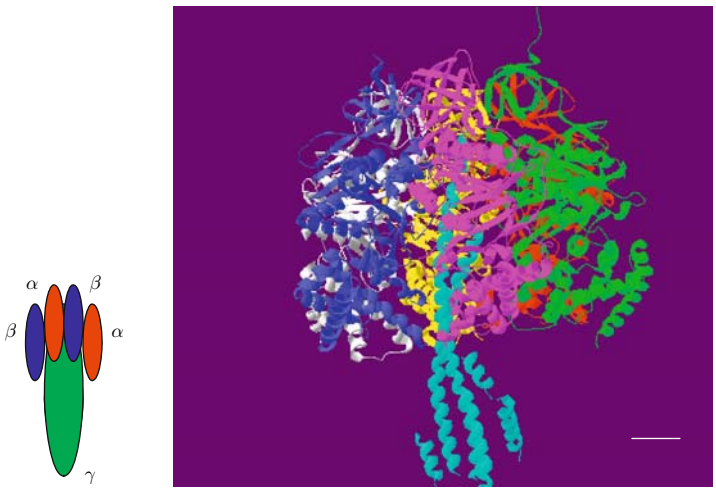


FIGURE 7.18 (left) *Schematic of F1-ATPase, the world's smallest rotary motor with three pairs of alternating α and β subunits and the γ subunit shaft* (right) *molecular model with the γ subunit shaft in light blue (scale bar = 2 nm).*

revolutions per minute by energy from a proton (or hydrogen ion) membrane pump, it acts as an enzyme helping to generate huge amounts of ATP daily. When the protein is supplied with ATP, it can run in reverse, causing the γ subunit shaft to rotate just like a motor.

Recently biophysicists were able to attach a rodlike molecule to the γ subunit, and measure the torque generated by the rotary motor in turning this attached rod. The measurement was done using laser tweezers, discussed in Chapter 19. In fact, they were able to lower the ATP concentration sufficiently so that individual step rotations of 120° of the shaft were observed. The individual torque measured for each step rotation was 44 pN-nm, where the incredibly small units used are those appropriate for the small force and step size involved. These researchers then calculated the work done by this rotary motor in each step rotation. Using Equation (7.20) and a step rotation angle of $\Delta\theta = 120° = (2\pi/3)$, they found that $\Delta W = (2\pi/3)(44 \text{ pN-nm}) = 92 \text{ pN-nm} = 92 \times 10^{-21}$ J. This value is very close to the energy liberated by one ATP molecule when it is hydrolyzed to ADP. Thus, this smallest of all rotary motors is nearly 100% efficient in converting energy into rotational work. It remains to be seen to what future applications our knowledge of this protein will lead.

4. ANGULAR MOMENTUM

In our discussion of momentum in Chapter 6 we were able to rewrite Newton's second law for a system, $\vec{F}_{\text{net, ext}} = m\vec{a}_{\text{cm}}$, in terms of its total, or center of mass, momentum so that the net external force was equal to the rate of change of the total linear momentum

$$\vec{F}_{\text{net, ext}} = \lim_{\Delta t \to 0} \frac{\Delta \vec{P}_{\text{total}}}{\Delta t}. \tag{7.24}$$

Recall also that in the absence of a net external force, this equation leads to the powerful conservation of momentum principle. In this section we analyze rotational motion in an analogous manner and introduce the important new quantity angular momentum and the principle of conservation of angular momentum, our third fundamental conservation principle. (Energy, linear momentum, . . . are you counting? There are not many more in this book.)

If you had to guess how angular momentum L should be defined, based on the rotational analog quantities to those defining the linear momentum, it is hoped that you would come up with the expression

$$L = I\omega. \tag{7.25}$$

Because linear momentum is defined as $\vec{p} = m\vec{v}$ and the rotational analogs to m and v are I and ω, respectively, this would be the natural candidate. Of course, such an intuitive guess needs to be corroborated, but this is a correct expression. You may have noticed that

in our analogy, Equation (7.25) has omitted vector signs on L and ω, where they might be expected. This is intentional on our part. It turns out that the vector nature of the rotational variables is subtle and is not needed in our basic discussions of rotational motion.

In the case of a particle of mass m constrained to rotate in a circular orbit, from the expression for its moment of inertia about the axis of rotation $I = mr^2$ we can write an alternative expression for the angular momentum of such a particle as

$$L = mr^2\omega = rm(r\omega) = r(mv) = rp. \tag{7.26}$$

For a system of particles or an extended body rotating about a fixed axis of rotation, an argument similar to the one given for the moment of inertia shows that the total angular momentum can also be written as

$$L_{\text{total}} = \sum r_i\, p_{i,\perp}, \tag{7.27}$$

where the sum is over the mass elements of the system and r_i and $p_{i,\perp}$ are the distances (measured from the axis of rotation) and components of momenta perpendicular to r_i (or tangential to the circular trajectories for pure rotational motion).

Now that we have defined angular momentum, we turn to the rotational equation corresponding to Equation (7.24). By analogy we should guess that this is

$$\tau_{\text{net, ext}} = \lim_{\Delta t \to 0} \frac{\Delta L_{\text{total}}}{\Delta t}, \tag{7.28}$$

where again we omit vector signs. This can most easily be seen by noting that the time rate of change of L in Equation (7.28) can be written as

$$\lim_{\Delta t \to 0} \frac{\Delta L}{\Delta t} = \lim_{\Delta t \to 0} I \frac{\Delta \omega}{\Delta t} = I\alpha,$$

using the definition of angular acceleration. Substituting for $I\alpha$ from Equation (7.21) then yields Equation (7.28). We reach an important conclusion from this equation.

In the absence of a net external torque on a system, its total angular momentum remains constant. This is a statement of the principle of conservation of angular momentum.

Along with conservation of energy and of (linear) momentum, it is one of the fundamental conservation laws in nature. For an extended body undergoing pure rotational motion conservation of angular momentum has the simple form

$$I\omega = \textit{constant}, \quad \text{(isolated system)}, \tag{7.29}$$

where the constant is the value of L_{total} at any instant of time. The following example illustrates the application of conservation of angular momentum.

Example 7.12 An ice skater begins a spin by rotating at an angular velocity of 2 rad/s with both arms and one leg outstretched as in Figure 7.19. At that time her moment of inertia is 0.5 kg-m^2. She then brings her arms up over her head and her legs together, reducing her moment of inertia by 0.2 kg-m^2. At what angular velocity will she then spin?

Solution: Because there are no acting external torques (any friction is ignored here), angular momentum is conserved and we can write that

$$I_{ini}\,\omega_{ini} = I_{fin}\omega_{fin}.$$

FIGURE 7.19 *An ice skater uses angular momentum conservation.*

In this case the skater's moment of inertia has decreased and so her angular velocity will increase.

We find

$$0.5(2) = 0.3\omega_{\text{fin}},$$

so that her angular velocity becomes 3.3 rad/s. The same principle controls the rotational motion of a ballerina or a diver as they change their moment of inertia by controlling their body configuration.

We summarize the rotational motion equations of this chapter in Table 7.2, indicating the corresponding equations for translational motion. In the last examples of this section we integrate the concepts presented so far in solving two more complex rotational motion problems.

Table 7.2 *Kinematic and Dynamic Equations for Rotational and Translational Motion*

Applicability	*Rotational*	*Translational*	*Relations Between Variables*
$\alpha(a) = \text{constant}$	$\omega = \omega_o + \alpha t$	$v = v_o + at$	$s = r\theta$
$\alpha(a) = \text{constant}$	$\omega^2 = \omega_o^2 + 2\alpha(\theta - \theta_o)$	$v^2 = v_o^2 + 2a(x - x_o)$	$v = \omega r$
$\alpha(a) = \text{constant}$	$\theta = \theta_o + \omega_o t + \frac{1}{2}\alpha t^2$	$x = x_o + v_o t + \frac{1}{2}at^2$	$a_{\text{tang}} = r\alpha$
General	$\text{KE} = \frac{1}{2}I\omega^2$	$\text{KE} = \frac{1}{2}mv^2$	$I = \sum m_i r_i^2$
General	$\tau_{\text{net,ext}} = I\alpha$	$F_{\text{net,ext}} = ma$	$\tau = rF_\perp = r_\perp F$
General	$\tau_{\text{net,ext}} = \frac{\Delta L_{\text{total}}}{\Delta t}$	$F_{\text{net,ext}} = \frac{\Delta P_{\text{total}}}{\Delta t}$	$L = I\omega = rp_\perp$
General	$\tau_{\text{net,ext}} = 0 \Rightarrow L_{\text{total}} = \text{constant}$	$F_{\text{net,ext}} = 0 \Rightarrow P_{\text{total}} = \text{constant}$	

Example 7.13 A hoop of mass 2 kg and radius 0.5 m has two spokes the length of a diameter, each of mass 0.1 kg. The hoop is made to rotate from rest by a light cord attached to a 0.02 m diameter shaft which is threaded over a frictionless pulley, and attached to a 10 kg weight (as shown in Figure 7.20). Find the angular velocity of the hoop after the 10 kg weight has fallen a distance of 1 m.

Solution: The tension in the cord supplies a torque to rotate the hoop–spoke assembly at an increasing velocity. We solve this problem in two ways: using torques and angular accelerations and using energy concepts.

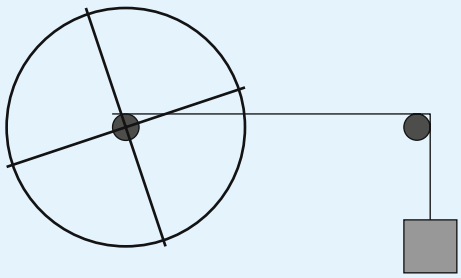


FIGURE 7.20 *A hoop being turned by a cord tied to a hanging weight.*

Using the first method, we first find the torque acting on the hoop and the moment of inertia of the rotating assembly so that we can substitute them into Newton's second law for rotations in order to find the angular acceleration. We have

$$\tau = Tr,$$

where T is the tension in the rope and r is the shaft radius. The cord tension is the only force that produces a torque on the hoop. The total moment of inertia is that of the hoop (MR^2, with M and R the mass and radius of the hoop) and that of the two spokes (see Table 7.1 for I for a rod rotating about an axis through its midpoint)

$$I = MR^2 + 2\left(\frac{1}{12}\, m\,(2R)^2\right),$$

where m is the mass of each spoke of length $2R$. To proceed, we first need to find the tension T which is not equal to the hanging weight. An independent equation for T can be obtained from the equation of motion for the hanging mass m'

$$m'g - T = m'a,$$

where $m'g - T$ is the net force on the hanging mass and a is its linear acceleration. Solving for T, multiplying by r to find the torque, and inserting this into Newton's second law for rotations along with the expressions for I and α we have

$$\tau = m'(g - a)r = \left(MR^2 + \frac{2}{3}\, mR^2\right)\left(\frac{a}{r}\right).$$

Here we have substituted $\alpha = a/r$ because the cord unwinds with a linear acceleration proportional to the angular acceleration of the shaft. We can solve this expression for the acceleration to find

$$a = \frac{m'gr^2}{MR^2 + \frac{2}{3}mR^2 + m'r^2} = 0\ .019\ \text{m/s}^2.$$

Now that we have a value for a we can solve for the angular velocity of the hoop. After the hanging weight has fallen 1 m, its velocity (and that of a point on the shaft) will be given from

$$v^2 = 2ax,$$

as

$$v = \sqrt{2ax} = 0.19\ \text{m/s},$$

so that the angular velocity of the hoop–spoke assembly will be

$$\omega = \frac{v}{r} = 19\ \text{rad/s},$$

where we divide by the radius of the shaft because v is the velocity of a point on the shaft.

An alternate solution, in this case much more elegant and straightforward, uses energy conservation. We simply write expressions for the initial and final total energies:

$$E_{\text{ini}} = m'gh,$$

where h is the 1 m height and the initial energy is all gravitational potential energy of the hanging mass, and

$$E_{\text{fin}} = \frac{1}{2}\ I\omega^2 + \frac{1}{2}\ m'v^2,$$

where the final energy is all kinetic, rotational, and translational. Equating these energy expressions because there is no loss of energy due to friction, inserting the above expression for I, and substituting $v = r\omega$ for the velocity of the hanging weight, we have

$$m'gh = \frac{1}{2}\left(MR^2 + 2\,\frac{1}{12}\,m(2R)^2\right)\omega^2 + \frac{1}{2}\,m'(r\omega)^2.$$

Solving this for ω, we find

$$\omega = \sqrt{\frac{2m'gh}{\left[MR^2 + \frac{2}{3}mR^2 + m'r^2\right]}} = 19\ \text{rad/s}.$$

Notice the beautiful simplicity of the conservation of energy approach!

Example 7.14 A 5 m radius merry-go-round with nearly frictionless bearings and a moment of inertia of 2,500 kg-m^2 is turning at 2 rpm when the motor is turned off. If there were 10 children of 30 kg average mass initially out at the edge of the carousel and they all move into the center and huddle 1 m from the

(*Continued*)

FIGURE 7.21 *Angular momentum conservation on a physics carousel.*

axis of rotation, find the angular velocity of the carousel. If then the brakes are applied, find the torque required to stop the carousel in 10 s.

Solution: Before the brakes are applied there are no external torques acting on the carousel (friction is absent in the bearings) so that we know angular momentum is conserved. Using this guiding principle, we can first write expressions for the initial and final angular momentum and then equate them to solve for the final rotational velocity. We have

$$L_{\text{ini}} = I_{\text{ini}}\,\omega_{\text{ini}},$$

where $I_{\text{ini}} = I_{\text{carousel}} + I_{\text{children}} = 2500 + 10(30)(5)^2 = 10^4\ \text{kg-m}^2$, treating the children as point masses located at the edge of the carousel, and $\omega_{\text{ini}} = 2\ \text{rpm} = 2(2\pi)/60 = 0.2\ \text{rad/s}$. Similarly the final angular momentum is given by an identical expression with $I_{\text{fin}} = 2500 + 10(30)(1)^2 = 2800\ \text{kg-m}^2$. Using conservation of angular momentum, we then can write

$$L_{\text{ini}} = 10^4(0.2) = L_{\text{fin}} = 2800(\omega_{\text{fin}}),$$

so that the final angular velocity is $\omega = 0.71\ \text{rad/s}$. Now, when the brakes are applied, the frictional torque will produce an angular deceleration given by

$$\alpha = \frac{\tau}{I}.$$

But the angular acceleration required to stop the carousel in 10 s can be computed from kinematics to be

$$\alpha = \frac{\Delta\omega}{\Delta t} = \frac{-0.71}{10} = -0.071\ rad/s^2.$$

Substituting this value into the previous equation and solving for the frictional torque gives

$$\tau = I\alpha = 2800(-0.071) = -200\ N\cdot m.$$

5. ATOMIC FORCE MICROSCOPY

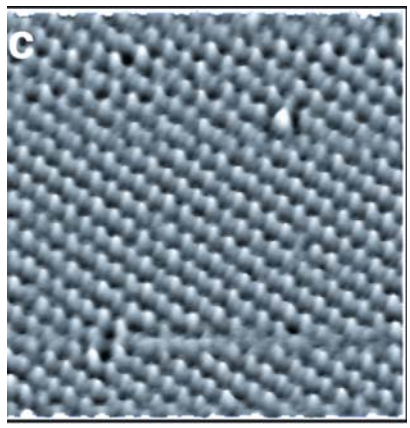


FIGURE 7.22 *Atomic resolution of a mica surface by AFM.*

As an application of the material of this chapter, we consider the functioning of the atomic force microscope (AFM), invented in 1986 by Gerd Binnig, who also invented the scanning tunneling microscope (see Chapter 24) and shared the Nobel Prize in 1986 for its discovery. The AFM provides images of the surface topography of samples with atomic resolution (see Figure 7.22). It is basically a very simple instrument that uses a fine tip attached to a cantilever (a device having a "beam" extending beyond its support, like a diving board; see Figure 7.23) and is raster-scanned (in a particular $x - y$ pattern) across, while in contact with, the surface to be studied. As the tip encounters small surface height changes, the cantilever is deflected proportionately due to the torque acting on it, and the height information can be recorded as a function of the $x - y$ position of the tip. This information can later be displayed in a topographical map of the surface with atomic resolution. This simple and amazing technique works because the effective springs acting between molecules on the surface are stiffer than the effective cantilever spring as we discuss below.

In one method to provide extremely sensitive information about the position of the cantilever, a laser beam is reflected from the cantilever surface onto a position-sensitive optical detector. The detector has several segments and the relative intensities recorded on the different portions of its surface allow a very sensitive measure of the laser beam deflection. By using a relatively long distance between the cantilever and the detector, a small angular deflection of the laser beam will result in a relatively large linear displacement (Figure 7.24). This scheme is called an optical lever arrangement and can be used to measure deflections corresponding to height changes of 0.01 nm (about 10% of the size of a hydrogen atom!).

How is a macroscopic tip able to measure the surface height with subatomic resolution? The essential conditions are to have an effective spring constant for the cantilever that is much smaller than the effective spring constant that holds the surface atoms together and to have the tip apply a very small (10^{-7} to 10^{-11} N) force on the surface so that the effective contact area is extremely small. In that way the cantilever will not distort the surface of the material, but will itself bend under the contact torque from an area of atomic dimensions on the material surface. Effective interatomic spring constants are on the order of 10 N/m, whereas the effective spring constant of a small piece of household aluminum foil can be made to be at least ten times smaller. Cantilevers used in AFM are usually microfabricated silicon made with integrated tips or with glued diamond tips with effective spring constants of 0.1–1.0 N/m.

The most common mode for imaging biological samples is the constant force mode. In this scheme a feedback mechanism varies the sample height so that the contact forces (or torques, because the lever arm distance is constant) can be kept small and constant. In this case, the small variations in sample height are tracked to

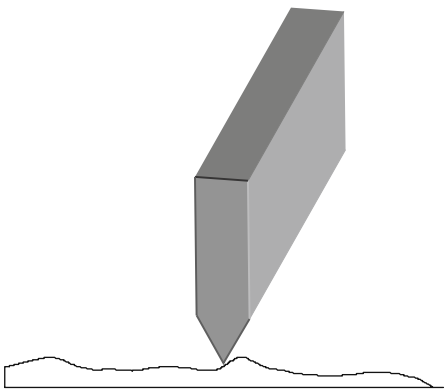


FIGURE 7.23 *The cantilever: the heart of the atomic force microscope.*

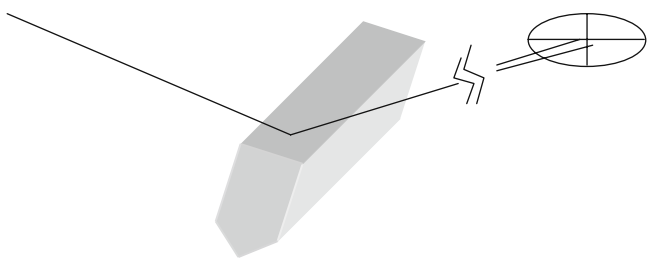


FIGURE 7.24 *An optical lever arrangement to measure small displacements of the cantilevered tip.*

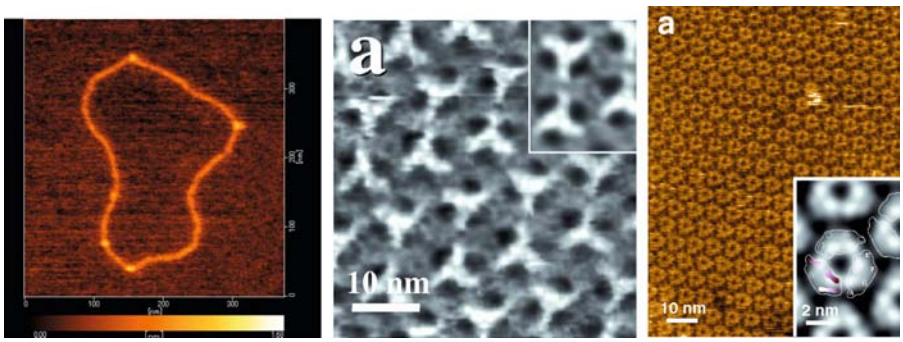


FIGURE 7.25 *AFM images of* (left) *plasmid DNA,* (center) *an E. coli membrane protein crystal,* (right) *purple membrane (bacterial light-sensitive proteins containing an analog of rhodopsin) with high-resolution inset.*

produce an image of the sample topography. Direct monitoring of cantilever deflection without feedback varying of the sample-to-cantilever height is usually not used since the larger forces occurring with large cantilever deflections can damage the surface. Biological samples are supported on a substrate, such as glass for thicker samples or cleaved mica that is flat to atomic dimensions, for thinner specimens. Figure 7.25 shows extremely high resolution images of several biological samples.

Example 7.15 A microfabricated integrated tip and cantilever for an AFM has an effective spring constant of 0.1 N/m. An optical deflection scheme is used to measure the deflection of the tip at the end of a 100 μm cantilever. A laser beam is reflected from the top surface of the tip and detected by a sensor 2 m away from the tip. Using the relation $s = r\theta$, a small angular deflection of the tip results in a relatively large deflection of the laser beam due to the large lever arm distance r. If the detector senses a 0.1 mm beam displacement from the "neutral," noncontact position, calculate the applied contact force the tip exerts on the sample surface.

Solution: A 0.1 mm beam displacement with a 2 m lever arm implies an angular rotation of the tip corresponding to

$$\theta = \frac{s}{r} = \frac{0.0001}{2} = 5 \times 10^{-5}\,\text{rad}.$$

The corresponding displacement of the tip, which has only a 100 μm lever arm is $s = r\theta = 100 \times 10^{-6} \cdot 5 \times 10^{-5} = 5 \times 10^{-9}\,m = 5$ nm. Using the force constant of the assembly, assuming Hooke's law applies, the applied force acting on the sample is $F = kx = 0.1 \cdot 5 \times 10^{-9} = 5 \times 10^{-10}\,N = 500\,pN$. To appreciate how small this force is, note that it is only about 100 times the force generated by a single myosin molecule interacting with an actin filament.

A wide variety of different biological samples have been studied using AFM. Included in these are nucleic acids, under physiological conditions so that dynamic processes of DNA–protein interactions can be studied as they occur (in so-called "real-time,"), biological membranes, in which individual lipids can be distinguished, cell surfaces, arrays and crystals of proteins, and even isolated proteins. Great care must be exercised to rule out artifacts in the images due to tip structure effects, scan speed artifacts, lateral forces on the tip due to frictional drag as the tip is scanned, and other problems, but the quality and the reliability of the images are steadily improving.

6. ROTATIONAL DIFFUSION; CELL MEMBRANE DYNAMICS

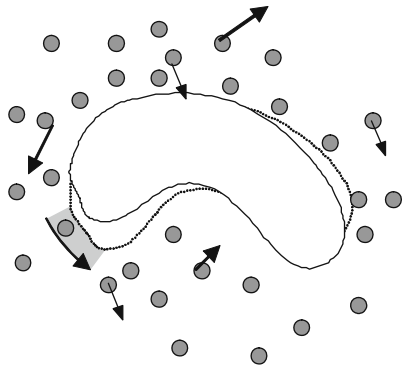


FIGURE 7.26 *Cartoon of a macromolecule undergoing rotational diffusion due to random collisions with solvent molecules.*

In our discussion of diffusion in Chapter 2 we learned that the translational random motion of macromolecules and microscopic objects is due to constant thermal collisions with the background fluid. Under the influence of numerous collisions with the fluid, there also will be rotational motion about the center of mass occurring due to random (in both direction and magnitude) torques acting on the molecule (see Figure 7.26). Just as in the case of translational motion, where there is a frictional force acting that is proportional to the velocity (see Equation (3.6)), there will be a frictional torque acting that, to a good approximation, is proportional to the angular velocity of the molecule

$$\tau_f = -f_R \omega. \tag{7.30}$$

Even if the molecule is spherical in shape, it may be asymmetric in other ways such as its electrical or optical properties, and these properties may allow one to distinguish different orientations. For an isolated spherical molecule of radius r, Perrin showed that the *rotational frictional coefficient*, which is the proportionality constant f_R between the frictional torque and the angular velocity, is

$$f_R = 8\pi\eta r^3, \tag{7.31}$$

where η is the fluid viscosity or "stickiness" that we study in detail in Chapter 9. In general, the rotational frictional coefficients for a few other simple shapes, such as ellipsoids or rods, have been calculated and the common result is a third-order dependence on the largest spatial dimension. This large dependence on size can be used to determine molecular dimensions very precisely (see the example below).

Rotational diffusion of an object can be characterized by the time it takes for the object to "randomize" its orientation or lose its "memory" of its initial orientation. This time is known as the *rotational relaxation time* t_R, and clearly is related to the rotational frictional coefficient, where the greater the friction, the slower the tumbling of the object and the longer its rotational relaxation time will be. Characteristic rotational relaxation times for small molecules are very fast, from ps to ns ($10^{-12} - 10^{-9}$ s), whereas larger macromolecules may have time constants of 10^{-3} s or longer. The *rotational diffusion coefficient* D_R has units of 1/s and is related to the rotational relaxation time ($D_R = 1/2t_R$). We can relate D_R to the rotational frictional coefficient through the general relation

$$D_R = \frac{k_B T}{f_R}, \tag{7.32}$$

where k_B is the Boltzmann constant and T is the absolute temperature. We show later in Chapter 12 that k_BT represents an average thermal energy from the collisions of all the solvent molecules. We see that the higher the temperature is, the larger D_R and the shorter t_R; the greater the rotational friction is, the smaller D_R and the longer t_R. These should make intuitive sense to you.

Example 7.16 A spherical virus, with electrical properties that allow one to distinguish its orientation, is in a water solution at 20°C (293 K). By studying the time-dependence of its interaction with light, the rotational diffusion time is measured to be 0.2 ms. Calculate the effective radius of the virus. Use a value of 0.001 (SI units) for the viscosity of water.

(*Continued*)

FIGURE 7.27 *A lipid, the structural unit of biological membranes, with polar head and nonpolar tail.*

Solution: From our discussion we know that the rotational time constant is related to the rotational diffusion coefficient by

$$t_R = \frac{1}{2D_R}$$

and is further related to the sphere radius using Equations (7.31) and (7.32). Substituting for these, we find that

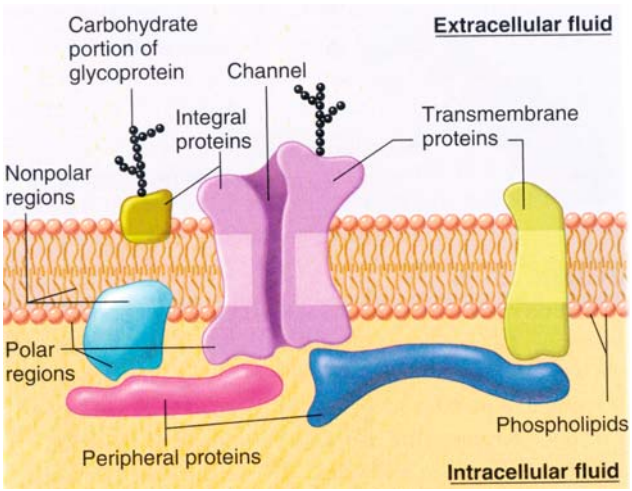
$$t_R = \frac{8\pi\eta r^3}{2k_B T}.$$

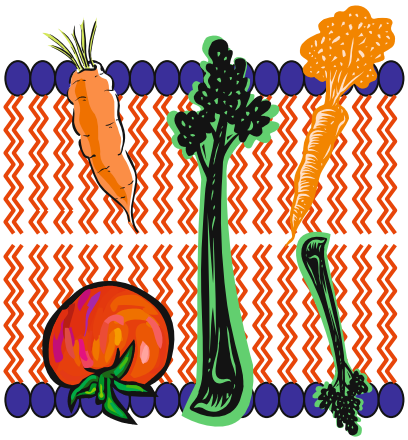
Solving for the sphere radius r, we have

$$r = \left(\frac{k_B T t_R}{4\pi\eta}\right)^{1/3} = \left(\frac{1.38 \times 10^{-23} \cdot 293 \cdot 2 \times 10^{-4}}{4\pi \cdot 0.001}\right)^{1/3} = 85\ nm.$$

One interesting area of biophysical research that involves rotational diffusion is the study of cellular membrane dynamics. Membranes are made up of a variety of lipid molecules that have electrically charged head groups and linear hydrocarbon tail portions (Figure 7.27) and serve as a boundary for cells and other organelles. The charged head group is highly attracted to polar water molecules (hydrophilic) whereas the tail groups are repelled by water molecules (hydrophobic). Biological membranes are bilayers, composed of two layers of lipid molecules arranged with the hydrophobic tails inside the membrane and with the hydrophilic head groups on the outer surface in contact with the water-based fluid inside and outside the cell (Figure 7.28). Synthetic bilayers can be made from purified lipid molecules, but natural biological membranes contain large numbers of proteins in addition to other smaller molecules. Membrane proteins are classified according to their association as either integral or peripheral. Integral proteins are those that are tightly bound to the membrane, some of them even spanning across the full width of the membrane. These latter proteins are important in allowing small molecules and proteins to cross the membrane barrier through channels, or molecule-specific pores. (We study the electrical properties of membranes in Chapters 15 and 16.) Peripheral proteins are more loosely bound to one of the surfaces of the membrane and can be dissociated by changes in pH or ionic concentrations.

FIGURE 7.28 *Two cartoons of a cell membrane showing the phospholipid bilayer and a typical collection of associated (geometric or vegetable) proteins.*





In the 1970s it was first discovered that the individual lipid molecules in a membrane, as well as the embedded proteins, are quite fluid, diffusing about on the two-dimensional surface of the membrane at rates of several micrometers per second. Up until that time membranes were viewed as static structures but measurements in the 1970s showed that lipids actually can not only diffuse about in their own monolayer (two-dimensional translational diffusion) but even, in rare events, "translocate" from one monolayer to the other (by "flipping" in a rotational diffusion "event"). A model of biological membranes known as the *fluid-mosaic model* was developed to describe this dynamic structure and modified versions of it are still useful today. Proteins in the membrane are confined (to various degrees) in

different domains or regions of the membrane. Many proteins and other macromolecules bind to specific cellular receptor proteins on the membrane. Often these will first bind to the membrane surface through nonspecific binding and then diffuse on the two-dimensional membrane surface until a specific receptor is found. Two-dimensional diffusion greatly speeds the binding kinetics over three-dimensional diffusion and is responsible for faster molecular recognition rates.

7. STATIC EQUILIBRIUM

An object that has both a constant linear momentum and a constant angular momentum is said to be in equilibrium

$$\vec{p} = constant \quad \text{as well as} \quad L = constant \quad \text{at equilibrium.}$$

This definition clearly includes the special cases when $\vec{p} = 0$ *and* $L = 0$ the object is at rest. According to Newton's second law, at equilibrium we must therefore have that

$$\begin{aligned} F_{x,\,\text{net}} &= 0. \\ F_{y,\,\text{net}} &= 0. \end{aligned} \tag{7.33}$$

In addition to a third similar equation if the problem involves three dimensions, $F_{z,\,\text{net}} = 0$, the rotational form of Newton's second law leads to another condition that follows from the constancy of the angular momentum (see Equation (7.28)), namely

$$\tau_{\text{net}} = 0, \tag{7.34}$$

where all of the torques are computed using the same arbitrary axis of rotation. If both $\vec{p}$ and L are zero, then the object is in static equilibrium, whereas if $\vec{p}$ and L are nonzero constants, it is in dynamic equilibrium. An example of dynamic equilibrium might be the (dynamic) balancing of an automobile wheel and tire so that it turns at constant angular velocity without any wobble (due to torques) acting. Simple static balancing of the wheel and tire at rest does not always reveal whether a wheel will wobble when rotating.

In this section we focus on the conditions for static equilibrium and some example applications. Our world is full of examples of objects in static equilibrium. All manmade fixed structures on Earth, including buildings, bridges, tunnels, and so on, are in static equilibrium. Gravity plays a key role in most statics problems. Although gravity acts on all portions of an extended object, for purposes of calculating the torque due to gravity acting on such an object we can consider the weight to act at a single point, known as the *center of gravity*. For us the center of gravity is identical to the center of mass, a distinction only occurring when the object is large enough that the value of g varies over the dimensions of the object. You might want to refer back to the discussion on center of mass in Chapter 6 to review its calculation.

In the rest of this section we consider three static equilibrium situations and see how to analyze the forces acting in each situation. The procedures used in these problems are similar in each case.

Example 7.17 A two-section ladder leans against a wall at a 70° angle from the ground and a man slowly climbs up the ladder as shown in Figure 7.29. Each of the sections of the ladder is 6 m long with the bottom section weighing 60 N and the top section weighing 40 N. With the ladder opened so that it is 8 m in total

(*Continued*)

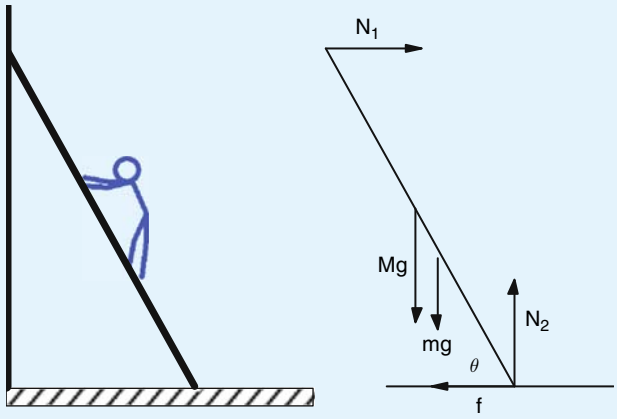


FIGURE 7.29 *Ladder leaning against a wall and the external force diagram.*

length, find all the forces acting on the ladder when the 70 kg man has climbed halfway up. Assume that there is no friction between the ladder and wall and that the coefficient of static friction between the ladder and ground is 0.65.

Solution: From a sketch of the situation, we construct an external force diagram showing all the forces on the ladder, the object of interest. There are five forces acting as shown in Figure 7.29: two normal forces, one friction force, and the weights of the man and ladder. Next we need to determine where the forces act, if not already clear. Aside from the forces acting at the top and bottom of the ladder, we are told that the man stands at its midpoint. We need to find the center of mass of the ladder which is made from two uniform 6 m sections that overlap by 4 m. The sketch (Figure 7.30) shows the needed information to calculate the center of mass of the ladder. We find its position, measured from the bottom of the ladder to be

$$x_{\text{cm}} = \frac{40(5) + 60(3)}{100} = 3.8\ m.$$

(Alternatively we could have treated the two ladder sections separately and used separate weights for each ladder section acting at their respective centers.) Now we are in a position to calculate the three unknown forces on the ladder, N_1, N_2, and f, when the man is at its midpoint.

Balancing vertical and horizontal forces, we can write

$$N_2 = (M + m)g,$$

and

$$N_1 = f.$$

From the first of these we can find $N_2 = 790$ N, but another equation is needed to proceed further. An independent equation can be obtained by summing the torques about any point and setting them equal to zero. To simplify this equation, we choose the bottom of the ladder as this point. Doing so eliminates f and N_2 from the torque equation because these forces have zero lever arm. We can write

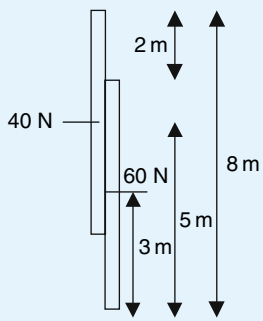


FIGURE 7.30 *Ladder dimensions.*

$$N_1(L\sin\theta) - Mg\left(\frac{L}{2}\cos\theta\right) - mg(x_{cm}\cos\theta) = 0,$$

where L is the length of the ladder and the appropriate sin or cos factors are introduced to find the lever arms of the three forces. Solving this for N_1, we have

$$N_1 = \frac{\left(mgx_{cm} + Mg\dfrac{L}{2}\right)\cos\theta}{L\sin\theta} = \frac{(100\cdot 3.8 + 70\cdot 9.8\cdot 4)\cos 70}{8\sin 70} = 140\ N.$$

Note that the friction force is equal in magnitude to this same 140 N value, far less than its maximum value of $\mu_s N_2 = 510$ N.

The steps used in the last example are appropriate for analyzing all static equilibrium problems and are summarized in Table 7.3. As you read the following two additional examples, note that they are approached using the outline in the Table.

Table 7.3 *Method to Solve Static Equilibrium Problems in Mechanics*

Step	*Procedure*
1.	Draw an external force diagram roughly to scale and carefully label all of the forces on the object of interest and distances.
2.	Determine which are the known and unknown quantities.
3.	Write the appropriate equations using $\vec{F}_{net} = 0$, one for each relevant spatial dimension in the problem.
4.	Write the appropriate equations using $\tau_{net} = 0$ about a convenient axis of rotation until sufficient independent equations are obtained to solve for the unknown quantities.
5.	Solve the set of algebraic equations for the unknowns.

Example 7.18 Consider the situation when a person is exercising with a dumbbell held in one arm outstretched horizontally as shown in Figure 7.31. The forces involved are the weights of the arm and dumbbell, the pull of the deltoid muscle F_M at an angle of 20° from the humerus bone acting at point A, and the force of the shoulder joint F_J acting at the axis of the shoulder joint, point O. If the arm is treated as uniform and weighs 50 N and the dumbbell weighs 75 N, find the force exerted by both the muscle and joint to hold the dumbbell in position. Take point A to be ¼ of the distance from the shoulder joint to the dumbbell.

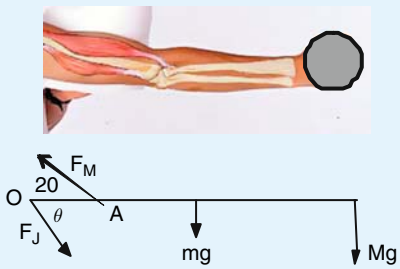


FIGURE 7.31 *An outstretched arm supporting a dumbbell* (above) *with equivalent forces drawn for analysis* (below).

Solution: On first glance it might be surprising that the shoulder joint exerts a downward force. To see why this must be the case, we can imagine taking torques about the center of mass where the arm weight acts. Then both the dumbbell and the muscle force will act to produce a clockwise rotation about the center

(*Continued*)

of mass and the shoulder joint must supply a torque tending to produce a counterclockwise rotation, hence a downward force. There are three unknown quantities in this problem: F_M, F_J, and θ. We can obtain two equations by writing

$$\sum F_{\text{horiz}} = F_J \cos\theta - F_M \cos 20 = 0$$

and

$$\sum F_{\text{vert}} = F_M \sin 20 - F_J \sin\theta - mg - Mg = 0.$$

To proceed further we can write an additional equation, taking torques about point O,

$$\sum \tau_o = mg(L/2) + MgL - F_M (L/4) \sin 20 = 0.$$

This last equation can be directly solved for F_M to find

$$F_M = (mg/2 + Mg)\left(\frac{4}{\sin 20}\right) = 1200 \text{ N}.$$

Substituting this back into the force balance equations, we can write

$$F_J \sin\theta = F_M \sin 20 - mg - Mg = 280\, N$$

and

$$F_J \cos\theta = F_M \cos 20 = 1100\, N.$$

Solving first for θ, we find, by dividing one equation by the other,

$$\tan\theta = 0.25 \quad or \quad \theta = 14^\circ,$$

and then by substituting into either force balance equation,

$$F_J = 1100\, N.$$

Note the relatively large forces needed to support a modest weight. These large forces make muscles and joints very susceptible to injury.

Example 7.19 Suppose that a 50 N uniform crate at rest is pushed with a horizontal force of 30 N applied at the top of the crate with dimensions as shown in Figure 7.32. If the coefficient of static friction is 0.7, will the crate slide along the surface or pivot at point O? If it will pivot, find the minimum applied force that will make the crate pivot about O.

Solution: For the crate to slide, the external force F must be greater than the maximum static friction force given by $\mu_s N$. The normal force is equal to the weight, although if the crate is about to pivot, the normal force will act at point O and not through the center of mass. In either case, we find the friction force to equal 35 N, more than the external force F, and so the crate will not slide. For

the crate to pivot about point O, the torque produced by F must overcome that of the weight of the crate (note again that when the crate is about to pivot, the normal force must act at point O and therefore produce no torque; similarly the friction force produces no torque). We can therefore write

$$\tau_{\text{net}} = FL - Mg(L/2).$$

But this will be a positive quantity so that the crate will, in fact pivot about point O. We find the minimum force needed to pivot about O by noting that in that case $\tau_{\text{net}} = 0$, so that $F = Mg/2 = 25$ N.

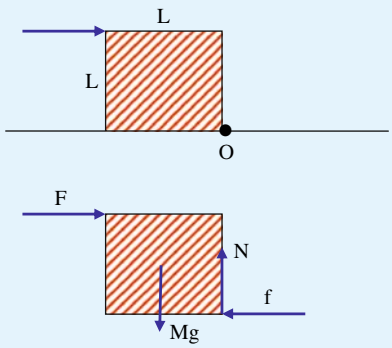


FIGURE 7.32 *A heavy crate about to pivot (?) or slide (?) along a rough surface with external force diagram below.*

CHAPTER SUMMARY

Table 7.2 provides a useful summary table of rotational kinematical and dynamical equations and a comparison with corresponding linear equations.

The rotational kinematical equations analogous to those studied for linear motion (with θ, ω, and α equaling the rotational angle, velocity, and acceleration, respectively) are

$$\omega(t) = \omega_o + \alpha t; \tag{7.7}$$

$$\theta(t) = \theta_o + \omega_o t + \frac{1}{2}\alpha t^2; \tag{7.8}$$

$$\omega^2 = \omega_o^2 + 2\alpha(\theta - \theta_o). \tag{7.9}$$

Rotational kinetic energy is defined as

$$\text{KE} = \frac{1}{2} I\omega^2, \tag{7.11}$$

where the moment of inertia, I, is given by

$$I = \sum(m_i r_i^2). \tag{7.14}$$

When the forces acting on an object are conservative, so that the work they do can be expressed in terms of a PE, conservation of energy can be expressed as

$$\frac{1}{2}mv^2 + \frac{1}{2}I\omega^2 + \text{PE} = E = \textit{constant}. \tag{7.15}$$

The torque produced by a force F acting on an object can be calculated in either of two equivalent ways:

$$\tau = rF_{\perp} = r(F\sin\theta), \tag{7.22}$$

or

$$\tau = (r\sin\theta)F = r_{\perp}F. \tag{7.23}$$

When a net external torque produces a rotation of an object about a fixed axis, the amount of work done is given by

$$\Delta W = \tau_{\text{net, ext}}\,\Delta\theta. \tag{7.20}$$

The angular momentum of a system can be written as either

$$L = I_{\text{total}}\omega \quad \textit{or} \quad L = \sum r_i p_{i,\perp}. \tag{7.25/7.27}$$

Newton's second law has a rotational form which can be written in two forms, analogous to $F = ma$ and to $F = dp/dt$:

$$\tau_{\text{net, ext}} = I\alpha, \tag{7.21}$$

(*Continued*)

or

$$\tau_{\text{net, ext}} = \lim_{\Delta t \to 0} \frac{\Delta L_{\text{total}}}{\Delta t}. \qquad (7.28)$$

From the previous equation, we see that if there is no net external torque acting on a system then the total angular momentum of the system is conserved (remains a constant in time).

Atomic force microscopy is an imaging technique that uses a cantilevered microfabricated tip that is scanned over a surface to produce an atomic image of the surface. The contact torque bends the cantilever and a feedback system moves the sample height in order to maintain a constant deflection as the cantilever is scanned over the surface.

Diffusion of asymmetric molecules results in not only translational diffusion of the center of mass, but also rotational diffusion about the center of mass. The rotational diffusion coefficient D_R, similar to the translational diffusion coefficient discussed in Chapter 2, describes the time it takes for a molecule to rotate, or tumble, in solution. The corresponding rotational times can be very fast (ps to ns) for small molecules and much slower (~ms) for larger macromolecules.

In the special case when there is no motion of a system, said to be in static equilibrium, then the net force and torque on the system must both equal zero:

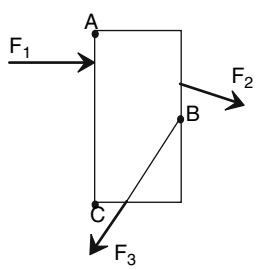
$$F_{x,\text{ net}} = 0,$$
$$F_{y,\text{ net}} = 0, \qquad (7.33)$$

and also

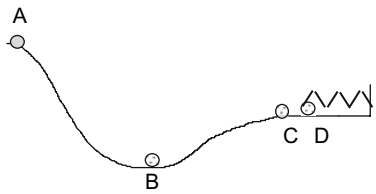
$$\tau_{\text{net}} = 0. \qquad (7.34)$$

QUESTIONS

1. Describe the possible overall motions of a slinky thrown into the air. How does the motion depend on the initial conditions as it is released?
2. Compare angular velocity as measured in units of rad/s and rev/min (rpm).
3. A piece of gum is stuck to the tire of a bicycle. As a girl starts to ride the bike from rest, does the gum have an angular velocity? A tangential velocity? An angular acceleration? A tangential acceleration? Answer these four questions again when the girl now coasts along at a constant translational velocity.
4. Explain why as a potter's wheel spins, the clay pot being made tends to expand radially outward.
5. If an object has a constant angular velocity, does it undergo any acceleration?
6. Picture two horses, side by side, on a merry-go-round. Which has the greater angular velocity? The greater linear velocity?
7. What is the difference between average angular acceleration and instantaneous angular acceleration? Can you give an example where they are not equal?
8. Explain why rotational velocity rather than linear velocity is the natural variable to use when describing pure rotational motion. Illustrate your argument with an example.
9. In the following examples state which of the two objects has the larger moment of inertia (or are they the same), measured about their symmetry axis:
   (a) Two balls of equal mass and radius: one solid and one hollow
   (b) Two solid cylinders with the same mass and radius: one twice as long as the other
   (c) A solid cylinder of mass M and radius $2R$, or one of mass $2M$ and radius R
10. Explain in words. The expression for the moment of inertia I of a long rod (see Table 7.1) depends on the axis about which the rotation occurs. Why is the numerical value of I for any such rod four times greater for rotation about the end as compared with the middle?
11. Discuss the definitions of "line of action" and "lever arm" in Figure 7.15. For a given force and pivot point, how should the force be oriented to maximize the torque applied to an object?
12. Discuss the equivalence of the two expressions for the torque $\tau = F_\perp r = r_\perp F$, carefully defining the terms.
13. In the following diagram state whether each force would produce a clockwise or counterclockwise rotation about each of the three labeled pivot points.



14. A ball rolls from rest down a steep incline, up a smaller hill and collides with and compresses a spring located at the hilltop as shown. Describe the different types of energy associated with the ball at each labeled point.



15. Define angular momentum in words giving an example to illustrate your definition.

16. Discuss the differences and similarities of the conservation of momentum and the conservation of angular momentum. Can you find examples where one and not the other quantity is conserved?

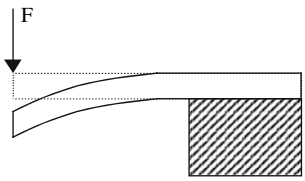
17. In the game "crack the whip", a number of participants join hands and run along the ground (or skate along the ice). Usually this human chain begins to "whip," with the trailing end of the line beginning to fishtail. Those at the end are soon flung free. Explain the mechanics of motion considering angular and linear speed.

18. Consider an individual riding in an automobile while using a lap belt but no shoulder harness. Suppose the car is brought to stop in a frontal collision. Describe the subsequent motion of the body and the potential for bodily injury in terms of inertia and angular momentum.

19. If one attempts to carry and transport an object such as a ladder, a large storm window, or a sheet of plywood, one finds that it is relatively easy to do so if the load is lifted at the proper point. Attempting to lift and carry at other points is difficult if not impossible. Discuss this matter considering torque and center of gravity.

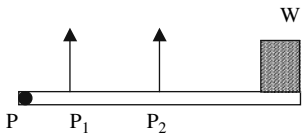
20. Why should a house painter, when shifting a raised ladder sideways, attempt to keep the ladder as close to vertical as possible?

21. Applying what you know about the nature of interatomic forces, explain why any force F, however small, applied to the end of a cantilevered bar as shown, must result in some amount of sag to the bar.



22. Consider the lever with fulcrum P and weight W as shown. The lever arm is pinned at P but is free to pivot.

(a) If you had to hold this lever arm horizontal by exerting an upward pull to counteract the downward force of the weight W, which would be easier, a pull at P_1 or at P_2?



265

(b) Imagine the lever represents your forearm in a horizontal position with a handheld weight W and elbow at P. Does your biceps function as if it works at P_1 or at P_2? Discuss the relative merits of arm design and function in light of your answer.

23. A centrifuge is a laboratory appliance useful for separating dissolved solid particles from liquid. When a sample is placed within a chamber, a dummy sample of approximately equal mass should always be placed into the chamber diametrically opposite. Explain the reason for this procedure.

MULTIPLE CHOICE QUESTIONS

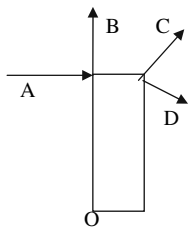
Questions 1–3 refer to a CD, with its information stored starting at an inner radius of 2.2 cm and out to an outer radius of 5.7 cm, that spins at 5500 rpm.

1. What is the fastest velocity at which information can be read? (a) 1970 m/s, (b) 33 m/s, (c) 256 m/s, (d) 3300 m/s.

2. If it takes 4 s to get up to speed from rest, what is the angular acceleration of the CD? (a) 1400 rad/s^2, (b) 23 rad/s^2, (c) 144 rad/s^2, (d) 8600 rad/s^2.

3. In getting up to speed from rest (see previous question), the CD makes (a) 730, (b) 1750, (c) 1150, (d) 180 revolutions.

4. The moment of inertia of a 20 cm uniform rod of 2.4 kg mass rotating perpendicular to its long axis about the rod center is (a) 0.008 kg-m^2, (b) 0.032 kg-m^2, (c) 80 kg-m^2, (d) 4.0 kg-m^2.

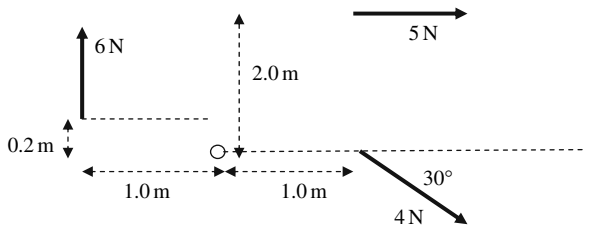


5. In the above diagram, which of the equal magnitude forces produces the largest torque about point O?
(a) A
(b) B
(c) C
(d) D

6. As a particle traveling in a circle speeds up at a constant rate, its net acceleration (a) increases and points more and more toward the tangential direction, (b) increases and points more and more toward the inward radial direction, (c) increases and points more

and more toward the outward radial direction, (d) decreases and points more and more toward the inward radial direction, (e) none of the above.

7. The net torque exerted by the forces shown about point O is

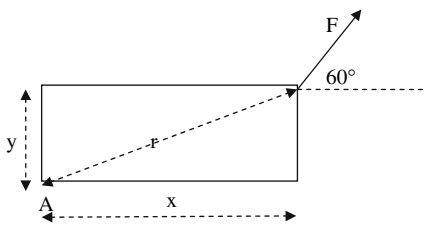


(a) 15 N-m, (b) 18 N-m, (c) 6 N-m, (d) 14 N-m, (e) none of the above.

8. A 5 N force acts alone to slow down a 2 m radius uniform rotating circular platform. The force acts at the edge of the platform and is directed at 30° to the outward radial direction. The applied torque is equal to (a) 8.7 Nm, (b) 5 Nm, (c) 0, (d) 10 Nm.

9. If the platform in the last question has a mass of 20 kg, the angular deceleration of the platform is (a) 0.25 rad/s^2, (b) 0, (c) 0.125 rad/s^2, (d) 0.5 rad/s^2.

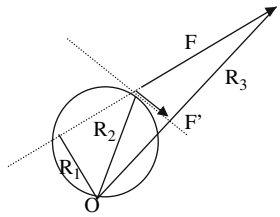
10. From the diagram, the magnitude of the torque of force F about point A, the lower left corner, is given by
(a) Fr
(b) Fx
(c) Fy
(d) Fr sin 60
(e) None of these



11. A particle is speeding up while traveling clockwise in a vertical circle. When it is at the 3 o'clock position, its net acceleration might point (a) toward 9 o'clock, (b) vertically downward, (c) toward noon, (d) toward 6 o'clock, (e) none of these are possible correct choices.

12. Which of the following uses correct logic? A cylinder of mass $2M$ and radius R has a race down an incline with a hoop of mass M and radius R. The winner is
(a) The cylinder, because it has the larger mass and therefore will accelerate faster
(b) The cylinder, because its mass is distributed throughout its volume and not all concentrated at the radius so it will travel faster even though it has twice the mass
(c) The hoop because it has the smaller mass and will therefore have less rotational and more translational speed compared to the cylinder
(d) The hoop because it has the larger moment of inertia and for the same torque will have the larger angular acceleration

13. Consider two equal mass cylinders rolling with the same translational velocity. The first cylinder (radius R) is hollow and has a moment of inertia about its rotational axis of MR^2, and the second cylinder (radius r) is solid and has a moment of inertia about its axis of $0.5\,Mr^2$. What is the ratio of the hollow cylinder's angular momentum to that of the solid cylinder? (a) $r^2/2R^2$, (b) $2R^2/r^2$, (c) $r/2R$, (d) $2R/r$.



14. A force F acts on a circular disk as shown. F' is the tangential component of the force at the point of application. The torque that the force F produces about point O is given by (a) $F'\,R_2$, (b) FR_3, (c) FR_1, (d) FR_2.

15. An Atwood machine is a real pulley mounted on a real shaft used to help lift a heavy weight by attaching it to another weight by a rope strung over the pulley. Once the weights leave the ground (a) the sum of the kinetic and gravitational potential energies of the two weights is constant, (b) the sum of the kinetic and gravitational potential energies of the two weights increases with time, (c) the sum of the kinetic and gravitational potential energies of the two weights and the rotational kinetic energy of the pulley decreases with time, (d) the sum of the kinetic and gravitational potential energies of the two weights and the rotational kinetic energy of the pulley is constant.

16. Two point masses, each 5 kg, lie at either end of a light rod of length 2 m. What is the moment of inertia of the system about the left end of the rod (in kg-m^2)? (a) 10, (b) 5, (c) 40, (d) 20, (e) none of the above.

17. A cylinder with 2 kg mass and 0.01 kgm^2 moment of inertia ($I = \frac{1}{2}\,MR^2$ for a cylinder) is rolling down an inclined plane with 30° inclination. At a point where its center of mass velocity is 1.0 m/s and its height from the ground is 0.1 m, what is its total mechanical energy (with respect to the ground)? (a) 1.5 J, (b) 1.96 J, (c) 2.96 J, (d) 3.46 J, (e) none of the above.

18. A 3 kg point mass is at the end of a light 2 m rod hanging vertically and hinged at the other end. If a 5 N force is exerted at the midpoint of the rod at a 45° angle below the horizontal, the initial angular acceleration of the mass is (a) 0.83 rad/s^2, (b) 0.42 rad/s^2, (c) 0.29 rad/s^2, (d) 0.59 rad/s^2.

19. What physics principle does a high diver use in executing a dive? (a) Conservation of momentum, (b) conservation of angular momentum, (c) conservation of moment of inertia, (d) conservation of torque.

20. An isolated horizontal circular platform is spinning on a frictionless axle with a person standing at its edge. If the person walks halfway in toward the center of the platform, the principle that allows you to find the new angular velocity is (a) conservation of energy, (b) conservation of momentum, (c) conservation of angular momentum, (d) conservation of torque.

21. An ice skater is spinning with her hands overhead and legs straight, with a moment of inertia of 0.3 kg-m^2, so that she has a 2 s rotational period. She extends her arms sideways, increasing her moment of inertia to 0.4 kg-m^2. Her final rotational period is (a) 2.7 s, (b) 1.5 s, (c) 2.3 s, (d) 1.7 s.

22. A spinning ice skater pulls in her outstretched arms. What happens to her angular momentum about the axis of rotation? It (a) does not change, (b) increases, (c) decreases, (d) changes but it is impossible to tell which way.

23. Answer each of these by choosing Yes or No.
(a) If the net force on a rigid body is zero, can it have an angular acceleration? Yes or No
(b) If the net torque on a rigid body is zero, can it have a linear acceleration? Yes or No
(c) Is angular momentum necessarily conserved in part (a)? Yes or No
(d) Is angular momentum necessarily conserved in part b)? Yes or No

24. Effective interatomic spring constants are on the order of (a) 0.1 N/m, (b) 10 N/m, (c) 10^3 N/m, (d) 10^5 N/m.

25. A uniform ladder is leaning against a rough vertical wall. The ladder makes an angle θ with the horizontal ground. Which of the following statements is false? (a) The weight of the ladder equals the normal force at the ground; (b) the normal force at the wall equals the frictional force at the ground; (c) the torque produced about the contact point with the wall by the weight, by the frictional force and by the normal force at the ground must all add to zero; (d) the weight of the ladder can be considered to act at the center of the ladder.

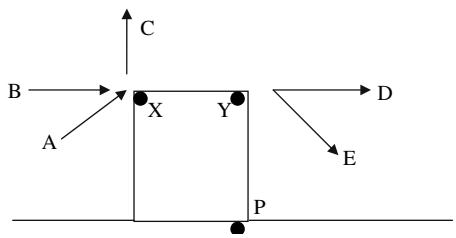
26. A ladder is leaning against a wall with a man standing at its midpoint. Which of the following is a false statement? (a) The net vertical force on the ladder is zero; (b) the net horizontal force on the ladder is zero; (c) the net torque about the bottom of the ladder is zero; (d) the net torque about the top of the ladder is zero; (e) none of the above is false.

27. In an Atwood machine (two unequal masses hung over a real pulley), the tension in the string attached to both masses is not the same because (a) the two masses are not the same, (b) the pulley has a nonzero moment of inertia, (c) there is friction in the pulley's bearings, (d) the acceleration of the two masses is different, (e) none of the above.

28. A uniform plank is used as a seesaw, but the fulcrum is not placed at the center, but at a point 1/3 of the length from one end. If the plank has a mass M, what mass must be placed at the end of the shorter side in order to keep it balanced? (a) M, (b) $2M$, (c) $M/2$, (d) $M/3$.

29. A 10 kg child and a 20 kg child sit balanced on the ends of a teeter-totter. The teeter-totter is a uniform plank of mass 5 kg which is placed on a fulcrum. Suppose now that each child moves halfway in toward the fulcrum. Will the teeter-totter remain in balance? (a) No, the end with the heavier child will go down. (b) No, the end with the lighter child will go down. (c) Yes, the teeter-totter will remain in balance. (d) It is not possible to tell from the information given.

30. Suppose you wish to tip a large packing crate so that you can put a hand truck under it. Assume the crate does not slide along the floor and that it tips about point P. Should you push or pull on the crate (or does it matter?) And where should you apply your force in order to use the smallest force to tip the crate? (a) Push on point X in the direction A. (b) Push on point X in the direction B. (c) Pull up on point X in the direction C. (d) Pull on Point Y in the direction E. (e) More than one of these choices will work equally well.



PROBLEMS

1. In 1986 the *Voyager* plane was the first to circumnavigate the Earth without refueling, taking just over 9 days to travel 24,987 miles around the Earth over both poles. Find its average speed and average angular velocity.

2. (a) How long does a centrifuge take to get up to a rotational speed of 80,000 rpm from rest with an acceleration of 40 rad/s^2?
(b) When shut down after spinning at that speed for 2 h, the rotor slows at a rate of 4 rad/s^2 without the brakes being applied. How long does it take to come to a stop after the two hour spin?
(c) Find the total number of revolutions the rotor has spun during the entire centrifugation run.

3. If the rotor of the previous problem is modeled as a uniform cylinder of 20 kg mass and 25 cm diameter, find its kinetic energy when spinning at its top speed.

4. A motor rotor turns at 1800 rpm. What are the angular and linear velocities of a point on the motor winding 20 cm away from the rotor axis?

5. A certain car tire is guaranteed to give 40,000 miles (64,400 km) of use.
(a) If the tire radius is 25 cm, how many revolutions does this amount of use correspond to?
(b) What is the angular speed for the tires (radius 25 cm) of a car traveling at 100 km/hr?

6. The world's largest clock face has a radius of 15.4 m. If that is the length of the minute hand find the linear speed of the tip of the minute hand.

7. What is the linear speed of the minute hand of a wristwatch if a 1 cm length is assumed?

8. A maple seed wing pair falling to the ground whirls around at 3 revolutions per second.
(a) If the overall length of the wing pair is 6 cm, what is the horizontal linear speed of the wing tip?
(b) Assume that because of air resistance, the wing pair falls at a constant speed of 80 cm/s. What is the total linear speed of a wing tip during fall?

9. A large tree is blown to and fro in a strong wind. The tree swings through an arc of 12° taking one second to swing from one way to the other. If a bird's nest is 24 m above the ground in the tree's branches, what is the average speed of the nest with respect to the ground as it moves back and forth with the tree?

10. Because of the rotation of the Earth, a person standing at the equator is moving through space at considerable speed with respect to another who stands at either pole. Compute this speed, considering the Earth as a sphere.

11. A compressor motor for a cooling system, responding to a thermostatic control, turns on and is brought up to its operating speed of 1200 rpm is 1.4 s.
(a) What is the angular acceleration of the motor shaft?
(b) If the motor assembly of the previous part has a mass of 9 kg and is modeled as a solid cylinder of radius 20 cm, what is the angular momentum of the motor at operating speed?

12. A dormant bacterium responds to stimulus and begins to move, via rotary motion of its flagellum.
(a) If it takes 2.5 ms to attain the normal rotational speed of 4 Hz, what is the angular acceleration of its flagellum?
(b) Suppose in response to an environmental stimulus, the rate of rotation of a bacterium flagellum decreases from 4 Hz to 3 Hz. The change is observed to occur slowly, over a 15 s interval. What is the angular acceleration of the flagellum?

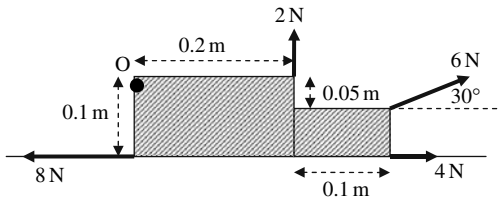
13. A recording tape is wound up onto a take-up spool. In order to achieve sound fidelity during playback, the tape movement must be such that its linear speed is constant throughout time. From start to finish, the spool radius ranges from 1 cm to 0.5 cm. If the linear speed of the tape is 5 cm/s, what are the rotational speeds of the take-up spool at the beginning and at the end of play?

14. A clock escapement wheel oscillates back and forth with each swing in a single direction amounting to 1/8 turn and taking 1/2 s. What is the average angular speed of the wheel during each swing?

15. A 50 cm outer diameter tire on a bicycle has a 0.05 kg piece of chewing gum stuck to its edge.
(a) If the bike starts from rest and attains a linear speed of 6 m/s in 30 s by a uniform acceleration, what is the angular acceleration of the gum?
(b) How many revolutions did the wheel make in that time?
(c) What were the tangential and radial components of the gum's acceleration at the end of the 30 s?
(d) How large must the force from the tire on the gum have been for it to remain stuck on the tire during the entire acceleration?

16. A honeybee flaps its wings about 200 times per second. Assume a wing is 0.7 cm in length and swings through an arc of 100°. What is the average speed of a wing tip during flight of the bee?

17. Find the net torque on the object shown about the pivot point O. (Hint: Look at the two components of the 6N force separately.)



18. A space station consisting of a ring with radius $R = 20$ m and mass $M = 100{,}000$ kg is spun around its center at a rate of $\omega = 0.7\ \text{s}^{-1}$ in order to produce artificial gravity. The moment of inertia is $I = MR^2$ for a ring.
(a) What is the centripetal acceleration of a point on the outside of the ring?
(b) What is the linear velocity of a point on the outside of the ring?
(c) What is the kinetic energy associated with the rotation?
A spherical asteroid of mass 50,000 kg and radius 15 m collides with the station at a speed of $v =$ 10 m/s and lodges in the center of the ring.
(d) What is the linear velocity of the combined station and asteroid after the collision?
(e) What is the rotational velocity of the combined station and asteroid after the collision, assuming that the asteroid was not initially rotating?

19. As a publicity stunt, a toy company constructs the world's largest yo-yo, consisting of a sphere 4 m in diameter with a mass of 1000 kg, with a steel cable wrapped around the middle of the sphere. They demonstrate it by dropping it off the George Washington Bridge, in New York City, using a crane to hold the free end of the cable. The yo-yo rolls down the cable without slipping.
(a) When the yo-yo has fallen a distance of 10 m, how many radians has the sphere turned through?

(b) What is the angular velocity of the spinning sphere when it has fallen a distance of 10 m?
(c) What is the linear velocity of the center of the sphere?

20. A cylinder of moment of inertia I_1 rotates about a vertical frictionless axle with angular velocity ω_i. A second cylinder that has moment of inertia I_2 and initially not rotating is dropped onto the first cylinder. Because friction exists between the two surfaces of the cylinders, they eventually reach the same final angular speed, ω_f.
(a) What is the expression for the magnitude of ω_f?
(b) Show that the kinetic energy of the system decreases in this interaction and calculate the ratio of the final rotational energy to the initial rotational energy.
(c) Why does the kinetic energy of the system decrease?

21. *The Pumpkin on the Nott revisited*! Suppose that the Nott Memorial is topped with an approximately hemispherical dome of radius $R = 89$ feet. Somehow an individual has balanced a spherical pumpkin at the top of the dome at an angle of $\theta_i = 0°$ with the vertical. Suppose that a gust of wind starts the pumpkin rolling from rest. It loses contact with the dome when the line from the center of the hemispherical dome to the pumpkin makes a certain angle with respect to the vertical. At what angle does this happen? Compare your results with those of Chapter 5, problem 28.

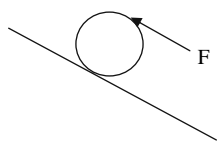
22. During most of its lifetime a star maintains an equilibrium size in which the inward force of gravity on each atom is balanced by an outward pressure force due to the heat of nuclear reactions in its core. After all of the hydrogen "fuel" is consumed by nuclear fusion, the pressure force drops and the star undergoes a *gravitational collapse* until it becomes a neutron star. In a neutron star, the electrons and protons are squeezed together by gravity until they fuse into neutrons. Neutron stars spin very rapidly and emit intense radio pulses, one pulse per rotation.
(a) Our sun has a mass $M = 2 \times 10^{30}$ kg and radius $R = 3.5 \times 10^8$ m and rotates once every 27 days. What is the initial magnitude of the angular velocity of the Sun?
(b) Suppose that the sun after undergoing gravitational collapse, forms a pulsar that is observed to emit radio pulses every 0.1 s. What is the magnitude of the angular velocity of the pulsar? (The sun would not actually form a neutron star as it is well below the minimum mass limit of 4 solar masses.)
(c) If the sun does not lose any mass in the collapse, what is the radius of the neutron star after the collapse? (Hint: Consider the sun before and after the collapse to be a solid sphere with moment of inertia $I_{star} = \frac{2}{5} MR^2$.)
(d) Is there a change in kinetic energy of the collapsing sun? If your answer is yes, how much work did gravity do in collapsing the sun and why is work done collapsing the sun? If your answer is no, then explain why gravity does no work in collapsing the sun.

23. A rotating space ship has a mass of 1,000,000 kg, most of it due to a large cylindrical tank of water (radius 10 m) on the central axis of the ship (with outer hull radius 20 m). While making its way to Alpha Centauri, the ship spins about this axis to generate the illusion of gravity.
(a) Initially the rotation rate is set so that the centripetal acceleration of a person just inside the outer hull is equal to the normal acceleration of gravity on the surface of the Earth. What is the linear speed of a point just inside the outer hull?
(b) What is the angular velocity of a point just inside the outer hull?
(c) What is the angular momentum of the ship (For this part, ignore the mass of the ship outside the water tanks.)
(d) A year or so into the trip they realize that the rotation is making the pilots seasick. Not wanting to waste fuel using rockets to slow the rotation, they decide to use angular momentum to their advantage, and instead pump the water out of the central tanks into a thin shell around the outer hull. What is the new angular velocity after this operation? (Again, consider only the mass of the water.)
(e) What is the acceleration of a person just inside the outer hull after the operation in part (d)?

24. A 200 kg playground merry-go-round with a 2 m radius is subject to a frictional torque of 40 Nm.
(a) If the merry-go-round goes round with a linear velocity of 6 m/s on the outside edge, what is its angular velocity?
(b) What force must be applied by one of the child's parents pushing on the outside edge to keep the merry-go-round moving at a constant angular velocity? (Note $I = {}^1/_2\ MR^2$ for this system.)
(c) When the parent gets tired and lets go, how long does the merry-go-round take to stop?
(d) At a point when the merry-go-round has lost half of its initial angular velocity, how much energy has been lost to frictional heating of the system?

25. A student sits on a freely rotating stool holding two weights, each of which has a mass of 3.00 kg. When his arms are extended horizontally, the weights are 1.00 m from the axis of rotation and he rotates with an angular speed of 0.750 rad/s. The moment of inertia of the student plus stool is 3.00 kgm^2 and is assumed to be constant. The student pulls the weights inward horizontally to a position 0.300 m from the rotation axis.
(a) Find the new angular speed of the student.
(b) Find the kinetic energy of the rotating system before and after he pulls the weights inward.

26. A uniform cylinder of 0.5 kg mass and 5 cm radius lies on an inclined plane with a 30° angle of inclination.



(a) Draw a carefully labeled free-body diagram for the cylinder when at rest at its initial height of 1.5 m and calculate the external force F that must be applied to the cylinder as shown to keep it from rolling.
(b) If this external force is now removed, use conservation of energy principles to find the speed of the cylinder's center at the bottom of the incline.
(c) What is the cylinder's angular momentum at the instant that it reaches the bottom of the incline?

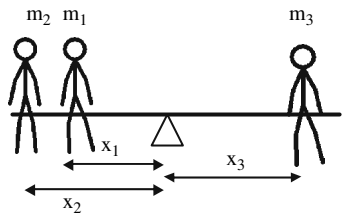
27. A physics professor lecturing about rotational motion uses as a prop a weighted bicycle wheel with a radius of 0.2 m and a mass of 5 kg, concentrated at the rim (i.e., ignore the hub and spokes when considering its motion).
(a) If the wheel is set to spinning at 150 revolutions per minute, what is the angular velocity of the wheel?
(b) If it takes 5.0 s to get the wheel up to speed, what torque was applied? What force does this require the professor to exert on the rim?
(c) Having stayed out late the night before, the professor drops the wheel on the floor. Assuming the wheel continues spinning at the same 150 rpm, how long does it take to roll into the wall, 10 m away?
(d) What is the kinetic energy of the rolling wheel?

28. A thin hoop with 2 kg mass and 1.5 m radius rolls down a 5 m long 30° inclined plane from rest.
(a) Find the center of mass velocity of the hoop at the bottom of the incline.
(b) Find the acceleration of the center of mass down the incline.
(c) How long does it take to get to the bottom?

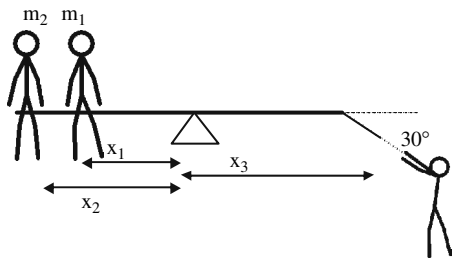
29. A centrifuge rotor, initially at rest, has a constant applied torque of 500 Nm causing it to speed up. Approximate the rotor as a uniform cylinder of 20 cm radius and 15 kg mass.
(a) If the friction force is negligible, find the angular acceleration of the rotor.
(b) How long does it take the rotor to reach 80,000 rpm?
(c) Suppose the applied torque is removed immediately upon getting up to speed and a small 30 Nm braking torque slows the rotor. How long does it take to stop?
(d) Find the total number of revolutions recorded on the centrifuge meter for this centrifuge run.

30. A custodian raising a bucket of coins from the bottom of a wishing well turns the handle attached to a spool of rope at a constant rate of 20 revolutions per minute.
(a) What is the angular velocity of the spool in rad/s?
(b) If the spool has a radius of 0.10 m, how much time is required to raise the bucket by 10 m?
(c) If the weight of the bucket exerts a torque of -2 Nm on the spool, what force must the custodian apply to the end of the 30 cm handle to keep the spool turning at a constant angular velocity?
(d) If the bucket is raised to the top, emptied, and allowed to drop back into the well, what is the angular velocity of the spool after the bucket has fallen 10 m, if the mass of the bucket is 10 kg and the cylindrical spool has a mass of 20 kg? (Assume frictionless bearings in the spool.)

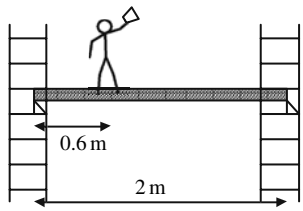
31. What must the mass of m_3 be in order to balance out the individuals of masses m_1 and m_2 situated on the seesaw as shown (Take $m_{seesaw} = 76$ kg, $m_1 = 18$ kg, $m_2 = 16$ kg, $x_1 = 1.2$ m, $x_2 = 1.4$ m, $x_3 = 1.4$ m)?



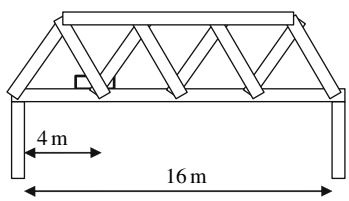
32. Suppose in the previous problem m_3 balances the seesaw by pulling now on the end 1.5 m from the fulcrum at an angle of 30° from the horizontal as shown. What force is necessary for balance?



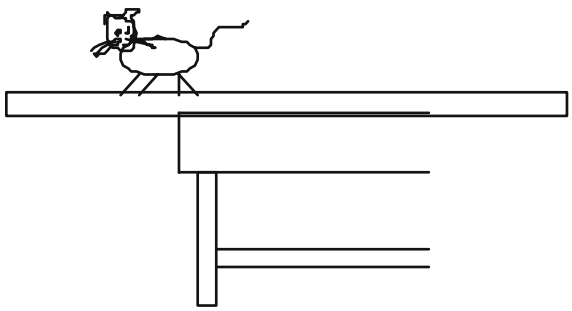
33. A housepainter who weighs 750 N stands 0.6 m from one end of a 2.0 m long plank that is supported at each end by ladder anchors. If the plank weighs 100 N, what force is exerted upon each anchor?



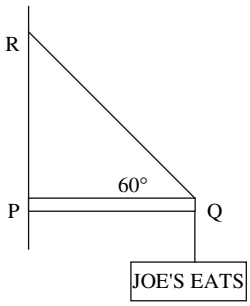
34. A car (14,500 N) travels across a simple truss bridge (230,000 N; 16 m long), that is supported by pylons at each end.
(a) What are the maximum and minimum forces exerted on each pylon due to the crossing?
(b) Suppose, all the while, a road crew (total weight 22,000 N) is situated 4 m from one end. What now are the maximum and minimum forces exerted on each pylon due to the crossing?



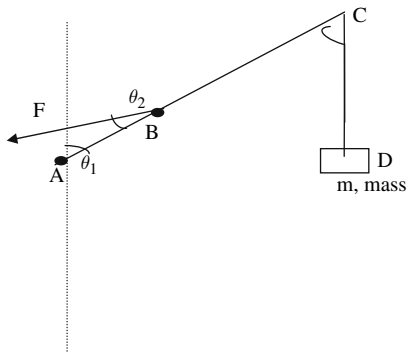
35. A carpenter places a 2 kg, 4 m long plank atop a workbench surface, with one end of the board overhanging the benchtop by ¼ of its length. A curious 8 kg cat then leaps up to the bench and begins to creep out along the overhang. How far can it go before the board tips, sending both cat and board to the floor?



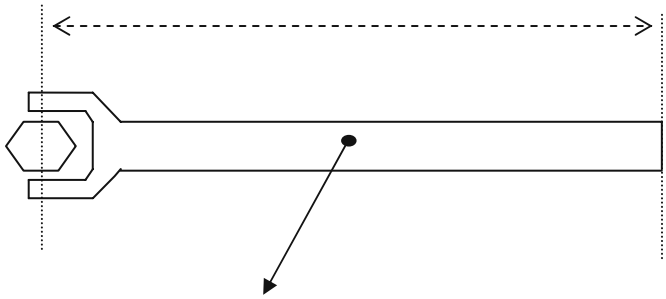
36. A 50 kg sign boom PQ is held horizontal by a wire cable that runs from Q to an anchor point on a building wall, R. The boom is free to pivot in the vertical plane about the hinge at P. If a sign suspended from the boom weighs 340 N, what is the tension in the cable necessary to hold the boom horizontal?



37. Lifting an object from a forward-leaning position. Consider the biomechanical stress that is sustained within the backbone and muscles of an individual who attempts to lift a weight from a stance in which the body leans forward. A represents the hinge point between back and hips. B represents a point in the lumbar region of the back. C is the hinge point between upper back and arms (with the shoulder bones intervening). Assume the individual pulls directly upward, from D to C. θ_1 indicates the amount of forward tilt of the spinal axis. Except for this tilt, an otherwise straight spine is assumed. θ_2 is the angle between the spinal axis and the line along which a set of lower back muscles exert tension. If the mass to be lifted is 2.5 kg, determine the magnitude of F, the amount of tension supplied by the lumbar muscles. Determine the compressive force on the lower spinal vertebrae, between A and B. Take $\overline{AB} = \frac{1}{4}\overline{AC}$, $\theta_1 = 45°$, and $\theta_2 = 8°$.



38. A specification chart notes that a certain machine bolt should be tightened to 85 N-m of torque. Because of interference with other adjacent machine parts, the mechanic can only grasp a wrench at the handle middle, and can pull along the direction shown, at an angle of 60° from the handle's long axis. If the wrench is 30 cm long, how much force must be applied in order to attain the specified torque?



39. A small 5 kg lead ball is on the end of a 0.5 m light rod that is hinged at one end so it is free to pivot and is held on the other end so that it is horizontal. When let go from rest find the
(a) Initial torque on the rod about the hinge
(b) Initial angular acceleration of the rod
(c) Angular velocity of the ball at its lowest point
(d) Angular momentum of the ball at its lowest point

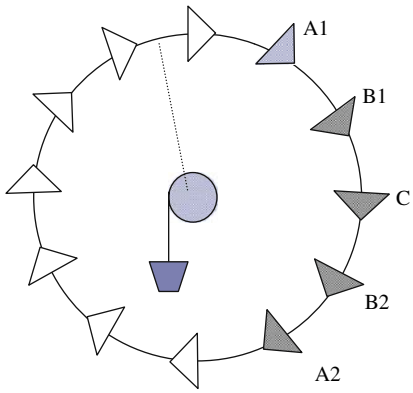
40. An old-fashioned child's toy top is set spinning by first winding string around it, and then tossing the top forward toward a smooth surface and then immediately pulling back sharply on the string. Suppose the string pull exists for 0.8 s during which time the top is set spinning at 15 turns per second. Treat the top as a solid sphere of radius 2 cm and of mass 100 gm. What is the pulling force that must be applied to the string to attain the motion described?

41. A uniform board, hinged at one end, is just barely supported in a horizontal position by an 8 N force applied at the other end and acting at a 30° angle with the horizontal.
(a) What is the weight of the board?
(b) What is the minimum force acting at the far end of the board that can keep it in a horizontal position?

42. A water wheel with an 8 foot radius collects water spilling from a millrace. The resulting weight imbalance produces a torque which turns the wheel. Suppose that only the five labeled buckets hold appreciable amounts of water, as follows: A1 = A2 = 0.5 cubic ft; B1 = B2 = 1 cubic ft; C = 1.5 cubic ft.

(a) What is the resulting torque on the wheel due to the water?

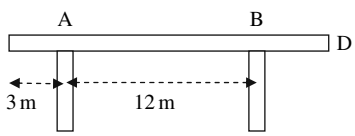
(b) If a capstan reel of radius 6″ is also mounted along the wheel axle, what is the maximum weight that can be raised by a rope wound round the reel?



43. A marching baton, twirled and tossed aloft, rotates end over end about its center of mass. Suppose its shaft is 45 cm long and has mass of 400 gm, and its rubber end caps are 40 gm each. Consider the shaft as a uniform rod and the endcaps as point masses situated at either end. If such baton rotates 3 times per second, what is its angular momentum?

44. A competitive diver executes a forward flip from a 3 m board. Suppose upon leaving the board the diver's body is assumed to be fully extended and is tilting forward at 0.5 turn per second. At the peak of the jump the diver tucks, bringing knees to chest and folding arms around knees. The maximum height reached by the diver is 1 m above the board. Representing the diver's body as a cylinder and assuming that tuck position reduces to one-half the overall length of the cylinder/body, show that it is possible for the diver to complete more than one complete somersault before falling to the water surface.

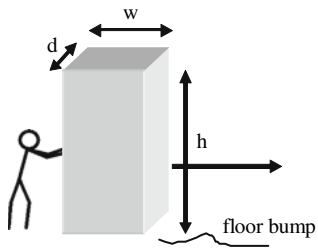
45. A 20 m long uniform beam weighing 600 N is supported on two 3 m long concrete columns A and B each having a cross-sectional diameter of 10 cm as shown.



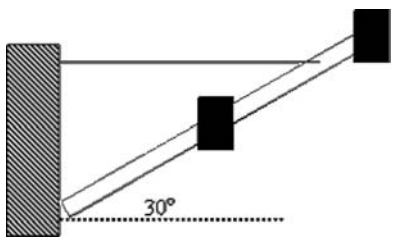
(a) Find the maximum weight a person can have and still walk to the extreme end D without tipping the beam.

(b) Find the forces that the columns A and B exert on the beam when the same person is standing at a point 2 m to the right of B.

46. A crate with rectangular faces (height h, depth d, width w) having roughly the outline of an upright refrigerator is slid sideways across the floor as indicated. Suppose that the leading edge of the crate strikes an irregularity in the floor and begins to tip. Determine an algebraic expression for the maximum angle through which the crate can tip about its bottom forward corner edge and still fall back to the upright position. Assume the crate and its contents uniformly occupy its volume.



47. A light 4 m long rod is hinged (frictionless) at one end, has two weights attached (one of 2 kg fastened at its center and one of 4 kg fastened at its other end), and is held in place at a 30° angle by a horizontal cable fixed at ¾ of the way along the rod from the hinge as shown.



(a) Find the tension in the cable

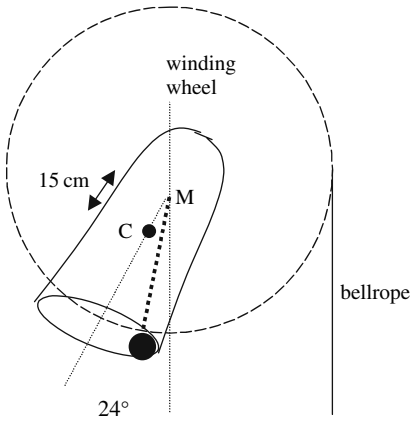
(b) If the cable is cut, find the initial angular acceleration of the rod.

48. Four children situate themselves on a playground whirly-go-round, a large metal platter that can rotate about a central vertical axis. The whirly-go-round platter has a mass of 90 kg and a radius of 1 m. Each child has a mass of 20 kg and sits 75 cm away from the center (25 cm in from the outer edge). After 15 s of sequential tugs at the attached metal bars, an adult sets the whirly-go-round spinning at 1 turn per second.

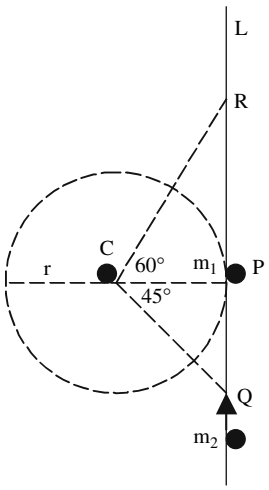
(a) What is the average force supplied by the adult?

(b) With the whirly-go-round now spinning at 1 turn per second, if the children allow themselves each to now move out to the edge of the whirly-go-round, what will be its new rotational speed?

49. A certain church bell weighs 800 lb (3600 N). In order for the clapper to contact the bell, the bell must be tilted 24°. Suppose the bell is mounted at M upon an axis 15 cm above the bell center of mass C. The winding wheel (diameter 1 m) for the bellpull has its center at the mounting axis of the bell. What force must be applied to the bellpull in order to impart tilt sufficient to ring the bell?



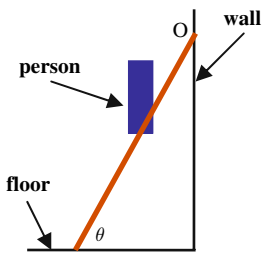
50. A mass m_1 of 400 gm travels a circular path around a center C at a radius r = 1.5 m at a constant speed of 2 m per second.



(a) What is the angular momentum of the mass about C?
(b) Suppose an identical mass m_2 travels, also at a constant 2 m/s, along the straight line l which is 1.5 m from C. Calculate the angular momentum of the mass when it is at point P, the point of intersection of the circular path and the line.
(c) Calculate the angular momentum of the mass m_2 about C when m_2 is at point Q.
(d) Calculate the angular momentum of the mass m_2 about C when m_2 is at point R.

Your findings should help you to see that the magnitude of angular momentum calculated for any mass depends, as does torque, on the point about which the value is calculated. Furthermore, different masses, with apparently different motions, can have the same angular momentum about a given point.

51. A 5.0 m long ladder with mass 100 kg is laid against a frictionless wall at an angle θ with respect to the floor as shown below. Suppose that the coefficient of static friction between the floor and ladder is 0.09 and that a painter of mass 60 kg has climbed up the ladder and has made it to a point 70% of the length of the ladder when the ladder begins to slip.
(a) In your own words, write a brief description of the problem stating the main physical principle(s) behind the problem.
(b) Draw a carefully labeled free-body diagram showing all of the forces that act on the ladder.
(c) From your free-body diagram, determine expressions for the normal forces due to the wall and the floor.
(d) Write an expression for the sum of the torques about the origin O (shown above) in terms of the angle θ and then evaluate your expression using the information given. (Hints: You will need the fact that $\sin(90 - \theta) = \cos\theta$ and for counter-clockwise rotations choose + for the direction of the torque.)



8 Ideal Fluids

The biological world could not exist apart from fluids. Water is the primary constituent of our bodies and of all animals and plants and it is difficult to imagine life without water. All life on Earth is also bathed in fluids, namely air or water, and the exchange of gases (oxygen and carbon dioxide) is required for all life as well. In this and the next chapter we study fluid mechanics, composed of the subjects of hydrostatics and hydrodynamics. These are generalizations of the statics and dynamics we have already studied and we use many of the fundamental principles and methods that have been developed. The major difference here is that we treat the fluid as a smooth continuous medium that continually exerts forces on immersed objects over their entire contact surface.

First pressure is introduced and we examine fluid flow of simple ideal fluids, those having no frictional losses of mechanical energy, showing how to apply the conservation laws we have learned to fluid motion. We show the power of these conservation laws in the context of a variety of different problems dealing with fluid dynamics. Then we study hydrostatics as a special case of hydrodynamics, considering the properties of a fluid in equilibrium and its effects on an immersed object. The chapter ends with a discussion of how pressure can be measured.

In the next chapter we study some more complex phenomena in fluids. These include a study of viscous fluids such as blood, in which there are frictional losses and complex behavior (we also discuss the human circulatory system from the perspective of fluid mechanics), as well as surface tension and capillarity of fluids.

1. INTRODUCTION

A fluid is a gas or liquid that, unlike a solid, flows to assume the shape of the container in which it is placed. This occurs because a fluid responds to a shear stress, or a force per unit area directed along the face of a cube of fluid, by flowing, rather than by an elastic displacement as in a solid. A drop of water on a kitchen counter flows when a towel is drawn over the surface whereas a pencil eraser bends when it is rubbed along the surface of a paper and then returns to its original shape. The molecules in a fluid are randomly located whereas those in a solid have some higher degree of order; intermolecular forces in a fluid are both somewhat smaller and are of a shorter range than in a solid, so that no elasticity exists in an (ideal) fluid. Gases, under so-called ideal gas conditions, have molecules that move completely independently of each other, without any intermolecular forces. Spreading out to fill any volume in which it is placed, a gas can have its average number of molecules occupying a unit volume change dramatically. Gases are therefore said to be compressible and their mass density (or just density) ρ, defined as

$$\rho = \frac{m}{V}, \tag{8.1}$$

J. Newman, *Physics of the Life Sciences*, DOI: 10.1007/978-0-387-77259-2_8,

is quite variable. On the other hand, liquids are characterized by their incompressibility and their density is a constant independent of the container size or shape or, to a good approximation, of the external forces acting on the liquid. What does determine the density of a liquid is the size of the molecular constituents and the intermolecular forces between them. Table 8.1 gives the densities of some materials.

Table 8.1 *Densities of Some Substances.*[1]

Substance	*Density(10^3 kg/m^3)*
Water	0.998
Water, 4°C	1.000
Mercury	13.6
Sea water	1.025
Ice	0.917
Ethyl alcohol	0.791
Whole blood	1.06
Blood plasma	1.03
Bone	1.9
Air	0.0013
Water vapor, 100°C	0.006

[1]At 20°C and atmospheric pressure unless noted.

Example 8.1 A cylindrical thin-walled plastic tube is filled with an unknown liquid. The tube has a 2.00 cm radius and is 20.0 cm long. When empty the tube weighs 0.200 N and when filled with the liquid it weighs 2.15 N. Calculate the density of the liquid. What might it be? The ratio of the liquid density to that of water at 4°C (1.0×10^3 kg/m^3) is known as the *specific gravity* of the liquid. Calculate the specific gravity of this liquid.

Solution: The volume of liquid the tube holds is given by

$$V = \pi r^2 L = \pi(0.02)^2(0.2) = 2.51 \times 10^{-4}\ \text{m}^3.$$

Subtracting the weight of the empty tube, the liquid has a weight of 1.95 N, or a mass of

$$m = \frac{F_w}{g} = \frac{1.95}{9.8} = 0.200\ \text{kg}.$$

We then find the liquid's density to be

$$\rho = \frac{m}{V} = \frac{0.200}{2.51 \times 10^{-4}} = 0.797 \times 10^3\ \text{kg/m}^3.$$

One liquid whose density is within 1% of this number is ethyl alcohol, therefore a likely candidate based simply on its density. The specific gravity measured by this procedure for this liquid is then 0.797, a dimensionless number. Note that although the cylinder volume and the mass of the liquid are specific to this particular problem, the density of this liquid and its specific gravity are properties of the substance and do not vary with the size of the cylinder. In other words, given a volume V, the mass of the contained liquid is determined by $m = \rho V$, where the density is a constant.

There are many materials that are not easily categorized into solid, liquid, or gas. *Gels* (cross-linked networks of polymer molecules) and *colloids* (suspensions of macromolecules or microscopic particles) are materials, many of which are important biomaterials, that can exhibit both liquidlike and solidlike properties depending on the conditions. Many gels, such as agarose, a polysaccharide, when dissolved in water at elevated temperatures behave as a liquid, but when cooled "gel" to form a material that has the elastic properties of a solid. A number of biologically interesting filamentous macromolecules can form *liquid crystals*, having some long-range order but behaving in other ways as does a liquid. Biological membranes are, in many respects, two-dimensional liquid crystals with aligned, but fluid, lipid molecules in a bilayer, as was discussed in the previous chapter.

Let's begin our discussion of fluids by considering an *ideal fluid*, one that is incompressible and has no viscous (frictional) resistance to flow. We treat our ideal fluid as a continuous medium without regard for its molecular composition and intermolecular forces. In the next chapter we consider the nonideal effects of viscosity.

2. PRESSURE

When a fluid is at rest it is said to be in *hydrostatic equilibrium* and there will be no net force on any portion of the fluid. Just as in particle mechanics, we know that this must be true because a net force on any portion of the fluid would result in motion, and we have assumed the fluid to be at rest. Although macroscopically the fluid is at rest, and the net force on every portion of the fluid is therefore zero, we know that the atoms or molecules of the fluid do move about. This thermal motion, or diffusion, is ever-present. We have discussed it briefly at the end of Chapter 2 and reconsider it when we study thermodynamics. To allow for this, in our continuous picture of a fluid we allow for local microscopic flows of fluid even under hydrostatic equilibrium.

If we imagine a square surface with unit area at some arbitrary location within the fluid (see Figure 8.1), then in the absence of any external forces such as gravity, there will be the same average momentum due to molecular motions crossing this surface per unit time in either direction, regardless of the position or orientation of the square. Any imbalance in the momentum flow, with its associated net force, would produce flow in our ideal fluid, which is assumed to be in hydrostatic equilibrium. The fluid on one side of this unit square thus exerts a force on the other side, which in turn exerts an equal and opposite force back. Now, imagine a cube of fluid with unit area sides (Figure 8.1). This fluid experiences forces from the external fluid on all six faces, canceling pairwise so that it experiences no net force.

We define the pressure within a fluid as the magnitude of the normal force per unit surface area due to the fluid on one side of the surface,

$$P = \frac{F}{A}. \tag{8.2}$$

Our discussion above indicates that in the absence of external forces, the pressure will be a constant throughout the fluid. The definition of pressure resembles a stress in a solid, namely a force per unit area, except that stressing a solid might well depend on the orientation of the solid if it is anisotropic. Note that pressure is defined to be a scalar quantity, and thus is independent of the orientation of the area. The metric unit for pressure is 1 N/m^2 which is given the name 1 pascal (Pa).

If a fluid is confined in a closed container and an external force is applied to a region of the surface bounding the fluid, there is an external pressure being applied. The external pressure applied does not remain localized near the surface where the pressure is applied, but the external pressure

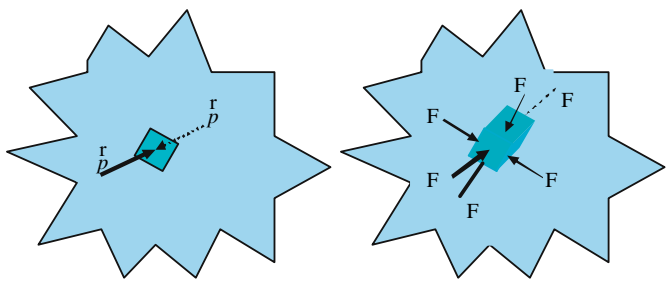


FIGURE 8.1 (left) *A region of a fluid* (darker colored) *with an imaginary unit area surface. The net momentum* $\vec{p}$ *crossing this surface must total zero when at hydrostatic equilibrium.* (right) *A cubical volume of a fluid* (darker colored) *with equal normal forces* F *acting on the six faces, resulting in a net zero force on the fluid in this cube from the surrounding fluid because the forces cancel pairwise.*

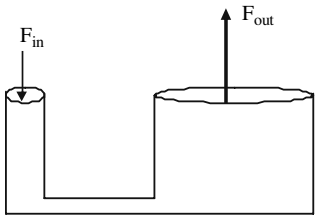


FIGURE 8.2 (left) *Schematic of a hydraulic lift. The output force is larger than the input force by the ratio (A_{out}/A_{in}), where the A's are the cross-sectional areas.* (right) *Mobile hydraulic lift for jacking up heavy equipment.*

on a confined fluid increases the pressure uniformly throughout the fluid by the same amount. This is known as *Pascal's principle*. Pressurized fluids occur in the human body in many forms: the vitreous humor of the eye (glaucoma being a disease with elevated eye pressure leading to damage of the optic nerve), the pressurized fluid in the cochlea of the ear, and blood under pressure in the chambers of the heart are all examples.

Hydraulic devices make use of Pascal's principle to amplify forces. As a first example of such a device consider the schematic diagram of a hydraulic lift shown in Figure 8.2. A smaller force F_{in} acting over a smaller area A_{in} determines the applied pressure $P = F_{in}/A_{in}$. The output end of the lift has a much larger area, A_{out}, and because the pressure within the fluid is essentially constant (exactly so if the heights are the same, as we show in Section 5), the output force F_{out} is determined from $P = F_{in}/A_{in} = F_{out}/A_{out}$, so that the output force is amplified to be

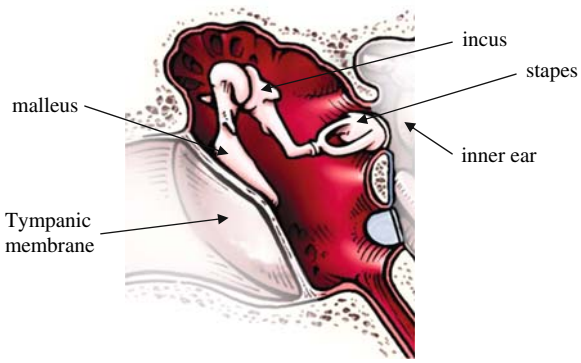
$$F_{out} = \frac{A_{out}}{A_{in}} F_{in}, \tag{8.3}$$

where the ratio A_{out}/A_{in} is the amplification, also known as the mechanical advantage. Hydraulic brakes in motor vehicles are based on this same idea.

A similar hydraulic effect is used to amplify sound in the middle ear. Consider the architecture of the middle ear shown in Figure 8.3. Bounded on the outer side by the tympanic membrane (ear drum) and on the inner side by the oval window, the middle ear is a small air-filled chamber, with a volume of about 2 cm^3, containing three small bones: the malleus (or hammer), the incus (or anvil), and the stapes (or stirrup). These are suspended by a set of ligaments and muscles so that the malleus is in close proximity to the tympanic membrane, and the "footplate" of the stapes is in the oval window. When the tympanic membrane vibrates in response to sound, the mechanical vibrations are transmitted to the inner ear through vibrations along the middle ear bones to the stapes. There is about a factor of 20 reduction in the effective area of the footplate of the stapes from that of the malleus. Because the mechanical force is constant through the bones (actually the force is also amplified roughly a factor of two due to some "lever action"), the pressure at the oval window is greatly amplified, due to its much smaller area. This hydraulic effect leads to amplification of sound waves entering the fluid-containing cochlea (inner ear) at the oval window. We discuss the overall functioning of the ear in more detail in Chapter 11.

Let's return to our discussion of a fluid at rest in a container and consider the situation when our imaginary unit area coincides with a portion of the container wall (Figure 8.4). Because this area now lies on the boundary, local random flow of fluid toward this area (and the resulting momentum flow) must be turned around via collisions with the wall to have a net zero macroscopic momentum flow in the fluid. Because in this case there is no fluid beyond the wall boundary to transport momentum back across our unit area, the wall must supply the reaction forces needed to "bounce" fluid back across the unit area so as to maintain the fluid macroscopically at rest. These forces are due to intermolecular interactions within the wall that supply, by collisions with the fluid molecules, the return flow of momentum needed to maintain hydrostatic equilibrium. The forces exerted by the wall must be perpendicular, or normal, to the wall and make up an external pressure supplied by the wall. If the wall were to exert forces parallel to its surface, these would generate fluid flow contrary to our assumption that the fluid is macroscopically at rest, and so the wall forces must be normal forces. We conclude (using Newton's third law) that fluids at rest exert only normal forces on boundary surfaces.

FIGURE 8.3 *The middle ear.*



Example 8.2 (a) A cylindrical tube filled with blood is held vertically. The tube has a radius and length of 1 cm and 10 cm, respectively. Calculate the pressure at the bottom of the tube. (b) Calculate the pressure exerted on the ground by a 100 kg man standing squarely on his feet, each sole having an area of 200 cm^2.

Solution: (a) The pressure at the bottom of the tube is equal to the weight of the blood divided by the cross-sectional area of the cylinder. (Here we are neglecting atmospheric pressure, the pressure due to the column of air above the tube; see Section 5.) Using the density of blood from Table 8.1, we find

$$P = \frac{mg}{A} = \frac{\rho(\pi r^2 h)g}{\pi r^2} = \rho g h,$$

so that the pressure is independent of the radius of the cylinder, and is equal to

$$P = 1.06 \times 10^3 \cdot 9.8 \cdot 0.1 = 1040 \text{ Pa}.$$

We later show that the result here that $P = \rho gh$ is generally true for a fluid in hydrostatic equilibrium. It actually represents the pressure increase at a distance h below the top surface. (b) Using the same principle, the pressure at the ground is given by

$$P = \frac{mg}{A_{\text{total}}} = \frac{100 \cdot 9.8}{2 \cdot 200 \cdot 10^{-4}} = 2.5 \times 10^4 \text{ Pa},$$

where the factor of 2 is due to the weight being equally supported by both feet and the factor 10^{-4} converts units from cm^2 to m^2.

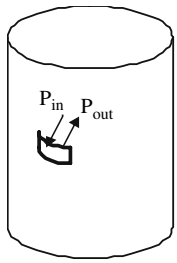


FIGURE 8.4 *The container wall supplies a normal force on the fluid at rest within it.*

3. DYNAMICS OF NONVISCOUS FLUIDS: TYPES OF FLOW

In this section we begin a description and analysis of fluid motion, the subject of fluid dynamics. Remember that we are treating a fluid as a continuous medium so that we characterize fluid motion not by the velocities of the individual fluid molecules, but by a mapping of fluid (vector) velocities as a function of both space and time throughout the volume of fluid. You can imagine a movie in which velocity vectors are drawn at a representative set of points in the fluid and you watch them change with time. We can distinguish two fundamentally different types of fluid motion, *steady flow*, or time-independent flow, and *unsteady flow*, or time-dependent flow. In steady flow the velocity mapping (Figure 8.5a) never changes and the flow pattern remains unchanged with time; in place of a movie, a picture will now do to map the flow pattern. Steady flow can also be visualized by drawing contour lines, known as *streamlines*, showing the trajectories of volume elements of the fluid (Figure 8.5b). Streamlines are drawn so that at any point the tangent to a streamline is the direction of the fluid velocity at that point (Figure 8.6). In addition, the magnitude of the velocity is indicated by the density of lines drawn; lines spaced more closely together indicate more rapid motion. Experimentally one can construct streamlines using dyes or "tracer particles" added to flowing fluids in order to analyze the fluid motion around an object. In this section, after briefly discussing unsteady flow, we limit our discussion to steady flow ideal fluids. The effects of viscosity are introduced in the next chapter.

In order to illustrate the various types of more complex fluid flow, let us imagine an experiment in which we confine a fluid

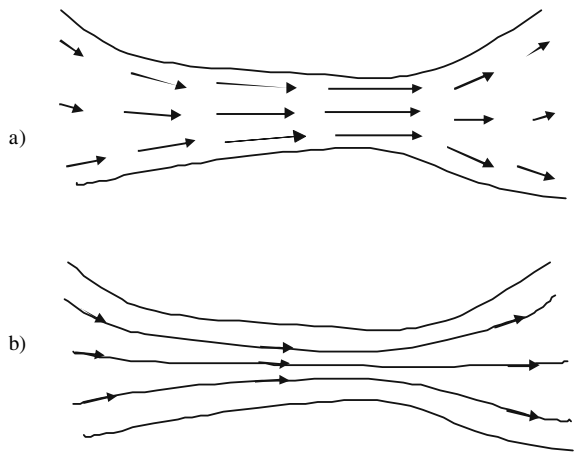


FIGURE 8.5 *Steady fluid flow mapped using either velocity vectors (**a**) or streamlines (**b**).*

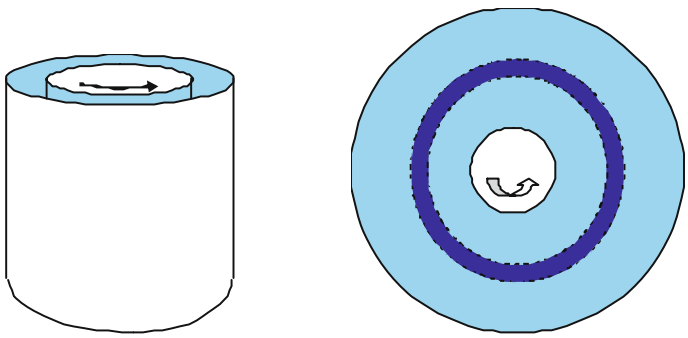
FIGURE 8.6 *Lava flow from a Hawaiian volcano showing streamline flow.*

between two concentric cylinders, with the outer one fixed and the inner one made to rotate at a variable rate that can be precisely controlled. Such a geometry is known as couette flow and is shown in Figure 8.7. At low rotational velocity of the inner cylinder, the flow will be steady with circular streamlines around the cylinder. This is one example of *laminar flow*, in which the fluid moves smoothly as if it were layered with the layers sliding smoothly over one another. Smoothly flowing streams or aerodynamic flow of air over and around a car are examples of laminar flow. Filtered air flow in hospital operating rooms is often designed to circulate under laminar flow to better filter all the air in the room. In couette flow because the outer cylinder does not rotate and the inner cylinder rotates at some constant angular velocity, the layers of fluid that slide over each other are cylindrical shells (Figure 8.7) with the outermost one stationary and the innermost one rotating most rapidly together with the inner cylinder.

If we imagine increasing the rotational velocity of the inner cylinder, eventually the fluid undergoes *turbulent flow* that is not only unsteady, but is also chaotic or unpredictable. Turbulence is familiar to you in the fast flow of water at rapids in a river or the flow of the air in a windstorm (Figure 8.8). In turbulent flow the velocity at any point changes chaotically in magnitude and direction.

Turbulence occurs in blood flow within our circulatory system and is quite important, for example, in the proper functioning of our heart valves. These are passive devices that open and close in response to the flow of blood and not from external muscle forces controlling them. As blood flows through a heart valve into one of the chambers of the heart, the turbulent back flow acts to shut the valve. Without the turbulent flow, heart valves would not close properly. As a second example of turbulence in blood flow, we note that a bleeding cut forms a clot much more rapidly if the flow of blood is turbulent, from a jagged wound, rather than laminar, from a thin smooth cut, such as a paper cut. This is true because the turbulent flow helps to shear blood platelets releasing the proteins necessary for clot formation.

FIGURE 8.7 (left) *The couette geometry.* (right) *Top view of laminar flow of fluid showing an annular layer flowing smoothly.*



As the fluid flow in the couette experiment changes from laminar to turbulent, there are a number of other types of unsteady flow that occur. Unlike turbulence, these flow patterns, although time-dependent, are periodic or quasi (nearly) periodic and are predictable despite being complex. Included in these flow patterns are vortices (rotational flow of the type seen as water drains in a bathtub) as well as amalgams of vortices and waves. The transition from steady to turbulent flow has been studied

FIGURE 8.8 *Steady to turbulent flow. Photo taken at the top of Horseshoe Falls near Niagara Falls.*

for a large number of fluid systems in recent years and has led to a deeper understanding of the nature of turbulence. There are also a number of other types of systems, including chemical, electrical, magnetic, and quantum, as well as simply mathematical, that have analogous behavior in a transition from deterministic to chaotic behavior.

4. CONSERVATION LAWS OF FLUID DYNAMICS

In our discussion of particle and rigid body mechanics we saw the importance of conservation laws. There are two conservation laws that we apply to the steady flow of ideal fluids in the absence of vortices, conservation of mass and conservation of energy, both of which give important results. Conservation of momentum can also be applied to fluid motion but it is beyond the scope of this book.

Figure 8.9 shows a fluid flowing in a cylindrical tube with a changing cross-sectional area. We are interested to learn how the speed of the fluid depends on the tube dimensions. For an ideal fluid, the velocity is constant over the cross-sectional area. We show later that a real viscous fluid has a velocity profile that varies over the cross-sectional area because of the drag forces slowing the fluid flow; in that case we can use the average velocity over the cross-sectional area in the following discussion and the result is still correct.

If the fluid of density ρ has a velocity v_1 in the portion of the tube with a constant cross-sectional area A_1, then in a time Δt, the mass of fluid that passes a given point in this section of the tube (see Figure 8.9) is given by

$$\Delta m = \rho A_1 v_1 \Delta t. \tag{8.4}$$

The product $A_1 v_1 \Delta t$ represents the cylindrical volume of fluid that will flow past the given point in a time Δt, and $Q = A_1 v_1$ is then the volume flow rate, or volume per second flowing past this point. Because mass is conserved, the fluid is incompressible, and no fluid escapes through the walls of the pipe, if we examine the fluid flow in the narrow region of the pipe, we must find the same mass of fluid flowing past a given point in this section of the tube in the same time Δt. Given that the fluid is incompressible, so that ρ is a constant, we find that

$$A_1 v_1 = A_2 v_2. \tag{8.5}$$

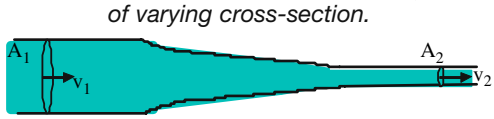


FIGURE 8.9 *Fluid flow in a cylinder of varying cross-section.*

Equation (8.5) can be written as $Q = Av$ = constant, and is known as the *continuity equation*. It tells us that when the cross-sectional area of a tube decreases, the velocity of flow must increase in an inversely proportional manner: v = (constant)/A. At first glance this may seem contrary to your intuition. Imagine a partially blocked water hose. You might think that the fluid would slow down where there is such a blockage. On the other hand most of us have had the experience of squirting water out of a hose, or sink or bathtub faucet, and realized that if the outlet area is decreased by blocking it with your hand, the flow of water can be speeded up. The variation of speed with cross-sectional area is a direct consequence of the principle of conservation of mass. The continuity equation compares fluid velocities in different regions of the flow, but what can we say about the actual fluid speed in a tube? To find out requires a bit more discussion and application of the conservation of energy principle, but first consider the following example.

Example 8.3 Water exits from a bathtub faucet with a speed v_0 (see Figure 8.10). Find an expression for the diameter of the flowing water stream as it falls, assuming laminar flow.

Solution: The stream of water narrows as its speed increases according to the continuity equation (Equation (8.5)), just as if it were confined to a tube. We can calculate the speed of the water after it falls a distance d because this is essentially a free-fall problem. Using standard kinematics, we have

$$v^2 = v_0^2 + 2gd,$$

so that the velocity after falling a distance d will be $v = \sqrt{v_0^2 + 2gd}$. From Equation (8.5), the continuity equation, the cross-sectional area, $A = \pi r^2$, will get narrower as the velocity increases according to

$$A = \pi r^2 = A_0 \frac{v_0}{v} = \pi r_0^2 \frac{v_0}{v}$$

where r_0 is the initial radius and r is the radius of the stream of water after it falls a distance d below the faucet. Solving for $2r$, we find

$$2r = 2r_0\sqrt{\frac{v_0}{v}} = 2r_0\sqrt{\frac{v_0}{\sqrt{v_0^2 + 2gd}}} = 2r_0\sqrt{\frac{1}{\sqrt{1 + \frac{2gd}{v_0^2}}}} = 2r_0\left(1 + \frac{2gd}{v_0^2}\right)^{-1/4}$$

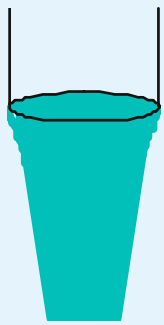


FIGURE 8.10 *A stream of water from a faucet. Why does it get narrower?*

For a given initial velocity of water, the stream narrows very slowly with distance d due to the ¼ power dependence. However, for small initial velocities the stream of water narrows appreciably over small distances d. For example, water slowly coming from a faucet at 10 cm/s shrinks to half its original diameter after falling only a few mm, but water pouring out at 5 m/s will travel almost 20 m before shrinking to half its diameter. Try this out at your sink at home!

Let's now examine the conservation of energy principle for an ideal fluid flowing in the absence of vortices. Figure 8.11 shows an idealized blood vessel with an ideal fluid flowing through it (the fluid properties of blood are detailed later in this chapter; for now we take blood to be an ideal fluid). Consider the colored fluid to be our system, originally bounded between the dotted blue lines in the top view. The two views

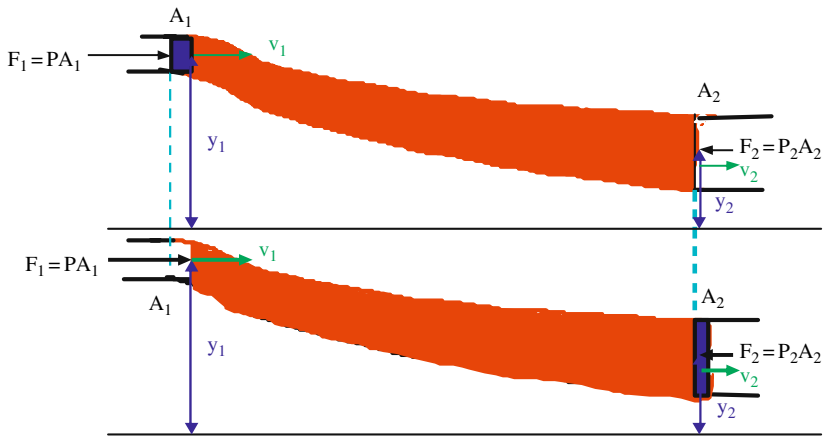


FIGURE 8.11 *A blood vessel with changing diameter shown at two different times. During the time interval between these two views our system has flowed, resulting in the boxed volume in the upper view* (blue) *to empty (or rather fill with nonsystem fluid) and the boxed region below (also* blue*) with the same volume to fill with system fluid.*

shown in the figure are taken a time Δt apart during the flow so that the colored fluid in the top view has moved, resulting in a boxed region in the bottom view moving beyond the original dotted line boundary, a boxed volume equal in volume to that in the top view. Of course there is fluid throughout the blood vessel, but we focus on the (red + blue) system fluid in what follows. In the transition between the two views the center of mass of the colored fluid of our system has moved, in general changing both its velocity as well as its height. We know that the total energy of the fluid in our system consists of kinetic and gravitational potential energy. The surrounding fluid, we show, does work on our system and we want to use the work–energy theorem to equate the total changes in kinetic and gravitational potential energy of the ideal fluid in our system from before to after this time difference Δt with the external work done on our system ($W_{\text{external}} = \Delta\text{KE} + \Delta\text{PE}_{\text{grav}}$).

Equal small volumes of fluid at two different locations along the blood vessel with different cross-sectional areas A_1 and A_2, will have different velocities because we have already shown that the volume flow rate $Q = Av$ is equal to a constant based on conservation of mass. Therefore as the fluid flows through the blood vessel, in a time Δt such that the volume of our system in the boxed region in the upper view (at a narrow location and thus a relatively fast velocity) moves to the right, and the equal boxed volume of the lower view (at a wider location and thus a slower velocity) fills with system fluid, the kinetic energy of the center of mass of the fluid in our system will decrease (in general, it may increase or decrease depending on the change in cross-sectional area). To find this change in system KE we need only consider the change in KE of the fluid in the two boxed regions. We can write this change in kinetic energy as

$$\Delta\text{KE} = \frac{1}{2}\rho\,(v_2^2 - v_1^2)Q\Delta t, \tag{8.6}$$

where $\rho Q\Delta t$ is the (same) mass of fluid in either boxed region in Figure 8.11.

According to the work–energy theorem, this change in kinetic energy is equal to the net work done on the fluid. There are two types of forces that contribute to this work: the gravitational force (if there is a change in height) and the pressure forces in the fluid. Work done by gravity is equal to the negative of the gravitational potential energy change, which is equal to

$$\Delta\text{PE}_{\text{grav}} = (\rho Q\Delta t)g(y_2 - y_1), \tag{8.7}$$

where again only the boxed volumes contribute to a change in the gravitational potential energy.

Work done by the pressure forces in the fluid can be found from the following argument. First, because the walls of the blood vessel only exert normal forces on our

ideal fluid and the fluid flows along the walls, there is no work done by the pressure supplied by the vessel walls. The fluid column to the left exerts a pressure on our system by supplying a force $F_1 = P_1A_1$ toward the right. Similarly the fluid to the right of our system exerts a pressure to the left resulting in a force toward the left $F_2 = P_2A_2$ that must be less than that acting toward the right in order for the fluid to flow toward the right. Each of these forces does work on the system fluid. Note that we chose the shape of the blood vessel to have uniform horizontal sections at either end to make this portion of the calculation easier, although the derivation is correct for all geometries. At the left end, positive work is done on the fluid in the amount

$$W_1 = F_1\Delta x_1 = P_1 A_1(v_1\Delta t), \tag{8.8}$$

where $v_1\Delta t$ is the distance Δx_1 over which the force acts in a time Δt. Using $A_1v_1 = Q$ and similarly calculating the (negative) work done on the fluid by the pressure force on the right, we find the net work done by the fluid pressure to be

$$W_{\text{net}} = (P_1 - P_2)Q\Delta t. \tag{8.9}$$

Combining Equations (8.6), (8.7), and (8.9) to write the change in mechanical energy (kinetic + gravitational potential) as equal to the net work done by the external pressure forces, we find

$$\left[\frac{1}{2}\rho(v_2^2 - v_1^2) + \rho\, g(y_2 - y_1)\right]Q\Delta t = (P_1 - P_2)Q\Delta t.$$

After dividing by $Q\Delta t$ and rearranging,

We have an expression for the conservation of energy for an ideal fluid

$$P_1 + \tfrac{1}{2}\,\rho v_1^2 + \rho g y_1 = P_2 + \tfrac{1}{2}\,\rho v_2^2 + \rho g y_2, \tag{8.10}$$

known as Bernoulli's equation.

FIGURE 8.12 *A 3-D magnetic resonance angiogram of a blood vessel with an aneurysm, the bulge near the center of the photo.*

Note that because the two positions were chosen arbitrarily, we can conclude that at any point in the fluid at any time the quantity

$$P + \frac{1}{2}\rho v^2 + \rho g y = \text{constant}. \tag{8.11}$$

Bernoulli's equation states that the sum of the fluid pressure and the mechanical energy density (the kinetic energy per unit volume plus the gravitational potential energy per unit volume, remembering that ρ is the mass per unit volume) in the fluid remains a constant of motion for the fluid, both as a function of time and position. We mention that for flow of blood in a real blood vessel, there are two major problems with this analysis, both of which are addressed in the next chapter. Blood is far from an ideal fluid, being viscous and quite complex in its properties. In addition, blood flow in the major arteries is not laminar but turbulent, and so requires a more complicated analysis for a complete understanding. Let's now examine a number of consequences of this restatement of the principle of conservation of energy for an ideal fluid.

First let's examine Bernoulli's equation for the case when y = constant, so that we can ignore gravity and have

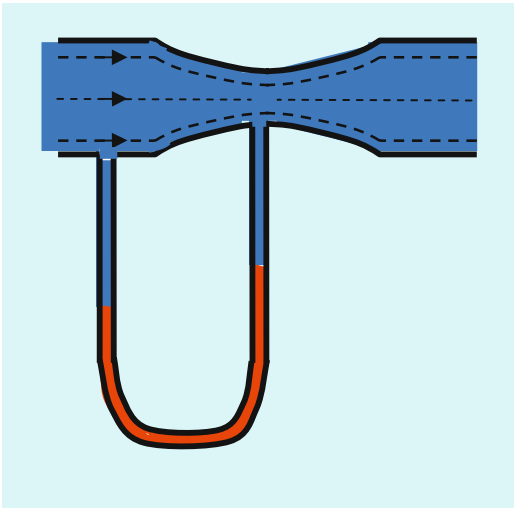
$$P + \frac{1}{2}\rho v^2 = \text{constant} \quad \text{(constant height)}. \tag{8.12}$$

FIGURE 8.13 *Atherosclerosis in a small artery.*

As an example application, consider the flow of an ideal fluid in a blood vessel that has an aneurysm, a weakened wall that has swollen as shown in Figure 8.12. If the cross-sectional area is greater at the aneurysm, the velocity of the fluid will be slower due to conservation of mass in the continuity equation. If the radius of the aneurysm is a factor of N greater than that of the blood vessel, the area will be greater by a factor of N^2 and from the continuity equation the velocity will be smaller by that same factor. Thus, in order for Bernoulli's equation (in the form of Equation (8.12)) to hold true, the pressure at the site of an aneurysm must be substantially increased. Such high pressure near a weakened portion of a blood vessel is extremely dangerous, especially because the vessel wall is already weakened.

Another important example is that of an artery with a partial blockage due, for example, to the buildup of plaque deposits, mainly composed of cholesterol (Figure 8.13). This disease is known as atherosclerosis. In a similar calculation to that just done for an aneurysm using the continuity equation, if the inside diameter of the artery is decreased by a factor N, the velocity of the blood in that region will increase by a factor of N^2. In this case the local pressure will drop substantially and, with a sizeable deposit of plaque, may drop to the point where the external pressure is enough to collapse the artery, cutting off the flow of blood. When this occurs in the coronary artery which supplies blood to the muscles of the heart, angina and eventual heart attack occur; if it occurs in the arteries leading to or in the brain, TIA, or transient ischemic attack, and eventual stroke occur.

FIGURE 8.14 *Venturi tube* (top horizontal portion) *and simple pressure meter (tubes with colored liquid) to measure fluid velocities in tubes of different cross-sectional areas through which fluid flows. Do you see why the fluid is higher in the center section?*



The principle governed by Equation (8.12) is also the basis for a number of flowmeter devices, used to measure fluid velocities and flow rates. An example is the Venturi meter shown in Figure 8.14. A fluid is being forced through the horizontal tube shown with varying cross-sectional area, which is known as a Venturi tube. According to the continuity equation the fluid velocity will vary inversely with the area. Using the assembly of vertical tubes with colored fluid, we can measure the local pressure differences (based on the colored fluid height differences as we show in the next section), and together with Equations (8.5) and (8.12), these can be used to determine the fluid velocity in the Venturi tube if the relative cross-sectional areas are known. Variations of this scheme can be used to measure flow of gases and liquids confined in tubes or even flowing around obstacles in the bulk. Blood velocities in arteries have also been measured using this method.

Example 8.4 Suppose that a catheter is inserted into the aorta, the largest artery of the body, to measure the local blood pressure and velocity (found to be 1.4×10^4 Pa and 0.4 m/s) as well as to view the interior of the artery. If the inside diameter of the aorta is found to be 2 cm and a region of the aorta is found with a deposit due to atherosclerosis where the effective diameter is reduced by 30%, find the blood velocity through the constricted region and the blood pressure change in that region. For this problem assume that blood is an ideal fluid and take its density from Table 8.1.

Solution: The unknown velocity can be determined from the continuity equation to be

$$v_2 = v_1 \frac{A_1}{A_2} = v_1 \frac{d_1^2}{d_2^2},$$

where the diameters d are given in the ratio of 1:0.7. The blood velocity in the constricted region is then found to be $v_2 = 0.4(1.0/0.7)^2 = 0.82$ m/s. Knowing the velocities and the density of blood, we can use Equation (8.12) to find the pressure difference at the constricted region. We find

$$P_1 - P_2 = \tfrac{1}{2}\rho\,(v_2^2 - v_1^2) = 0.5\,(1060)\,(0.82^2 - 0.4^2) = 270 \text{ Pa},$$

so that the local pressure in this region is reduced by only about 2%. We show in the next chapter that when viscosity effects are included the pressure reduction is much greater.

Another type of application of Bernoulli's equation occurs when the pressures at the two points of interest are equal. In this case Equation (8.11) reduces to

$$\frac{1}{2}\rho v^2 + \rho g h = \text{constant} \quad \text{(constant pressure)}. \tag{8.13}$$

One example application is the calculation of the efflux velocity from a water storage tank, as shown in Figure 8.15. We assume that the volume of the tank is very large so that the height difference between the water surface and the water tap $y_2 - y_1$ does not change appreciably (at least over short times) and the velocity of the water at its surface within the tank is negligible. Because the pressure at the top of the water tank and at the water tap are both equal to atmospheric pressure,

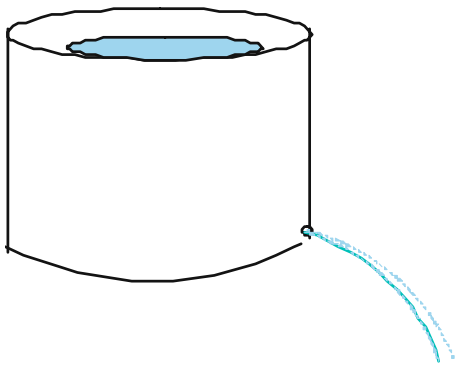
$$\frac{1}{2}\rho v^2 + \rho g y_1 = \rho g y_2, \tag{8.14}$$

where v is the efflux velocity of the water. Solving for v, we find that

$$v = \sqrt{2g(y_2 - y_1)}, \tag{8.15}$$

a result known as Torricelli's theorem. This expression should be reminiscent of our studies of the free-fall of an object through a height $(y_2 - y_1)$; the water has the same speed as it would have if it underwent free-fall vertically through the same height difference. Many residential homes receive water from storage tanks using a gravity delivery system. In very tall apartment buildings or skyscrapers, water must be pumped to holding tanks using positive pressure from below. (In Example 8.10 we show that sucking the water upward with negative pressure from above will not work!)

FIGURE 8.15 *Calculation of the efflux velocity from a large water tank.*



Example 8.5 An opened bottle of saline solution used as an intravenous drip has a fine capillary tube of 1 mm diameter attached. If the bottle is placed 1 m above the open end of the capillary tube, what will the flow rate from the tube be before it is attached to a hypodermic needle and inserted into a person's vein? Treat the solution as an ideal fluid.

Solution: As we have seen the efflux velocity of the saline solution is given by Torricelli's theorem to be

$$v = \sqrt{2g(\Delta y)} = \sqrt{2 \cdot 9.8 \cdot 1} = 4.4 \text{ m/s}.$$

Then the flow rate Q, given by $Q = Av$, is

$$Q = \pi r^2 v = \pi(0.0005)^2 4.4 = 3.5 \times 10^{-6} \text{ m}^3/\text{s} = 3.5 \text{ cm}^3/\text{s}.$$

Because of the narrow tube diameter, this result is about a factor of 10 larger than the correct answer when we have accounted for the viscosity of water. We learn about viscosity in the next chapter. You should realize that this is not the method currently used to deliver saline to a patient. A sealed bag of saline is connected via a peristaltic pump to a hypodermic needle. The pump ensures a constant delivery rate of fluid and the sealed bag ensures sterility.

5. HYDROSTATICS: EFFECTS OF GRAVITY

In Section 2 we argued that in the absence of external forces, the pressure within a fluid is uniform, having the same value throughout. In this section we consider the effects of gravity on the pressure within a fluid and on the effective weight of a submerged object within the fluid.

Consider a small element of volume within a fluid in hydrostatic equilibrium as shown in Figure 8.16. From our general result (Bernouilli's equation, Equation (8.10)), we can find a condition that must hold under hydrostatic equilibrium simply by setting the fluid velocities equal to zero,

$$P_1 + \rho g y_1 = P_2 + \rho g y_2. \tag{8.16}$$

It is instructive to derive this same result directly from first principles. According to Newton's laws for an object in equilibrium, the net force on this fluid element must be zero. If we consider the vertical forces acting on this volume, there are three such forces to include: the weight of the volume element and the downward and upward pressure forces from the surrounding fluid at the top and bottom of the volume element, respectively. If we call the pressure at the top surface P_1 and the pressure at the bottom surface P_2, and the volume element has a height h and a cross-sectional area A, then the balance of vertical forces implies that

$$P_2 A = P_1 A + \rho(Ah)g, \tag{8.17}$$

where (Ah) is the volume of the element and $\rho(Ah)g$ is its weight. Dividing through by the arbitrary cross-sectional area A, we find that the pressure in the fluid varies with the height h according to

$$P_2 = P_1 + \rho g h. \tag{8.18}$$

Note that this result agrees with Bernouilli's equation when the fluid velocities are set equal to zero (Equation (8.16)), where $h = y_1 - y_2$. The individual pressures P_1 and P_2 in this expression are called absolute (hydrostatic) pressures. Equation (8.18) relates the absolute pressure at two different points to each other and shows that in

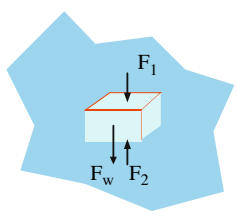


FIGURE 8.16 *Forces acting on a small volume of fluid, with* $P_1 = F_1/A$ *and* $P_2 = F_2/A$.

hydrostatics the pressure difference, $\Delta P = P_2 - P_1$, depends only on the fluid density and the height difference between the two points. In our derivation of Equation (8.18), we have assumed that the density does not vary with height. This is a good assumption for incompressibile (or nearly so) liquids, but a poorer one for gases. One conclusion from this result is that the pressure in a fluid is uniform within a horizontal plane, regardless of the shape of the container, varying only with height. The height difference is sometimes referred to as the *pressure head.*

Let's return to Equation (8.18) and choose one of the positions (#1) to be the surface of the fluid that is open to the atmosphere at sea level (Figure 8.17). In this case, the pressure P_1 is known as *atmospheric pressure*, P_{atm}, and represents the pressure due to the weight of the entire column of air in the atmosphere above a unit area on the surface. That this is so follows from considering the point at the surface of the fluid near sea level and a point vertically above it outside the Earth's atmosphere, where the absolute pressure is essentially 0, that of the vacuum. Applying Equation (8.18) to these two points results in a pressure difference just equal to the weight of a column of air in the Earth's atmosphere with a unit cross-sectional area. The average atmospheric pressure at sea level is equal to 1.01×10^5 Pa and is defined as 1 atmosphere (1 atm). This converts to about 14.7 pounds per square inch. This means that the average weight of air contained in a rectangular solid with a 1 m^2, or with a 1 in^2, cross-sectional area and a height equal to the height of the Earth's atmosphere, some 400 km, is just about 10^5 N, or 14.7 lb, respectively.

As we noted above, Equation (8.18) does not determine the absolute pressure at some height h, but the absolute pressure relative to that at another height. If we use atmospheric pressure as the reference point, the hydrostatic pressure in Equation (8.18), is given with respect to atmospheric pressure as

$$P = P_{atm} + \rho g h; \tag{8.19}$$

P is the absolute pressure and $\rho g h$ is called the gauge pressure, the difference in pressure from atmospheric pressure. In the next section we discuss the measurement of pressure, showing that most pressure measurements are referenced to atmospheric pressure and are therefore gauge pressures.

Example 8.6 A swimming pool has a sloped bottom starting at 1 m depth and dropping linearly to a 5 m depth at the middle where it levels off for the rest of the pool. Find the pressure on a small, 2 cm diameter, balloon when held at the bottom of either end of the pool. Also find the net compressive force on the balloon at each location due to the water.

Solution: The absolute pressure on the balloon at either location is given by Equation (8.19). We find that

$$P = P_{atm} + \rho g h = 10^5 + 10^3 \cdot 9.8 \cdot h,$$

so that the pressures are 1.1×10^5 and 1.5×10^5 Pa at 1 m or 5 m depth, respectively. The forces on the balloon are of three types, as we show below. Aside from the actual (negligible) weight of the balloon there is a buoyant force, discussed just below, and a compressive force due to the pressure of the water.

FIGURE 8.17 *Fluids rise to the same height regardless of the shape of the container.*

Because the balloon is small, the pressure on it can be taken as a constant. The compressive force is given by $F = PA$, where A is the total surface area of the balloon, so that

$$F = P4\pi r^2,$$

and the compressive forces are 138 N and 188 N at the two respective locations. These compressive forces tend to shrink the volume of the balloon (see Figure 3.17). Of course these compressive forces, when added as vectors give a net force of zero because of the symmetry. Therefore the balloon will not experience a translation or rotation due to pressure, but will have a compressive force acting on its volume. We also know that the buoyant force will provide a net upward force on the balloon, as we show after this next example.

Example 8.7 Your blood pressure varies not only periodically in time with your heartbeat, as we discuss in the next chapter, but also spatially at different heights in the body. This variation is due to differences in the weight of the effective column of blood in your blood vessels as a function of height in the body. Assuming that the average blood pressure at the heart is 13.2 kPa (corresponding to the average of a high and low pressure of 120/80, as it is commonly referred to, or 100 mm Hg; see the next section for a discussion of these units), find the blood pressure at foot level (1.3 m below the heart) and at head level (0.5 m above the heart). If a person experiences an upward acceleration as, for example, in an airplane during take-off or even in a rapid elevator in a tall building, the increased pressure can drain the blood from the person's head. What is the minimum acceleration needed for this to occur (take the head to be 25 cm in height)?

Solution: At the level of the feet, the blood pressure is increased over that at the heart by the amount

$$\Delta P = \rho gh = 1060 \cdot 9.8 \cdot 1.3 = 13.5 \text{ kPa},$$

so that the blood pressure there is 26.7 kPa (twice that at the heart, or roughly 200 mm Hg, by simple proportion). Similarly the blood pressure in the head is, using the same expression, reduced by

$$\Delta P = 1060 \cdot 9.8 \cdot 0.5 = 5.2 \text{ kPa}$$

to a value of 8.0 kPa (or 61 mm Hg). Do you see why blood pressure is measured with a cuff placed on your arm near your heart?

If the person is accelerating upward with acceleration a near the Earth's surface, then the effective value for g in the expression for ΔP would be replaced by $(a + g)$. This is so because the effective weight of the fluid becomes $\rho(g + a)V$, where V is the volume, in the derivation of Equation (8.17). Normally the blood pumped by the heart could rise to a maximum height determined by inserting the value for the blood pressure at the heart level into the equation

$$P = \rho gh,$$

(*Continued*)

to find a height of 1.27 m above the heart. Assuming the pressure supplied by the heart remains constant, this implies that on accelerating upward the blood would only rise to a height given by

$$h' = h\left(\frac{g}{a+g}\right)$$

Solving for a and setting $h' = 0.25$ m, we require an upward acceleration of

$$a = g\left(\frac{h-h'}{h'}\right) = g\left(\frac{1.27-0.25}{0.25}\right) = 4.1 \times g.$$

At accelerations near or above this value, without blood reaching the brain, the person will black out.

5.1. ARCHIMEDES' PRINCIPLE

In the development of Equation (8.18), we showed that at hydrostatic equilibrium the weight of the volume element of fluid is precisely balanced by the net upward force due to fluid pressure. Although in our derivation we used a volume element of uniform cross-section, the weight of any volume element, regardless of shape, will be supported by the net fluid pressure force, as long as the fluid is in hydrostatic equilibrium. Now, suppose that we want to find the force acting on an object of mass m submerged in our fluid. In addition to the downward pull of gravity that supplies a force equal to mg, there will also be an upward *buoyant force* due to the pressure of the fluid.

We can determine the strength of this buoyant force on our object by doing the following "thought experiment." Suppose that in the pure fluid we draw an imaginary closed surface that forms the exact boundary of the material object, having the same volume V as the object (Figure 8.18 left). The buoyant force on this imaginary object will be exactly equal to the weight of fluid contained within our imaginary closed surface. Why? Because the fluid is in hydrostatic equilibrium just sitting there. Remember that this net upward buoyant force arises from the pressure of the surrounding fluid. Now if we imagine draining the fluid from within this imaginary boundary and inserting the object in the identical location, the surrounding fluid will not "know" the difference and the buoyant force on the material object will be the same as it was on the original fluid. We conclude that the buoyant force on an object is equal to the weight of the fluid displaced by the object. This is a statement of *Archimedes' principle*.

The buoyant force may or may not result in the material object floating in equilibrium with no net force acting (neutral buoyancy). If the material object has a density just equal to that of the fluid, then the buoyant force will just support it, as it does in the case of the pure fluid in the same imaginary boundary. If the object has a greater density than the fluid it is immersed in, the weight mg will be greater than the buoyant force and the object will sink, whereas if the density is less than that of the fluid (e.g., a balloon under water or an ice cube in water), the buoyant force will be greater than mg and the object will rise to the surface. In the latter case, the object will float partially submerged in the fluid so that the displaced fluid volume is smaller than the volume of the object itself, and the buoyant force will be correspondingly smaller, just equal to the object's weight.

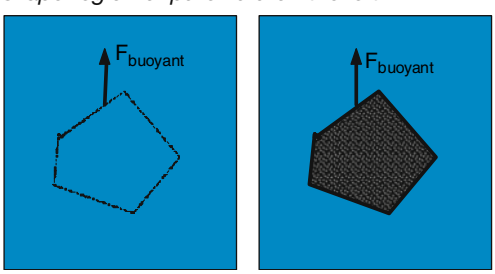


FIGURE 8.18 *"Thought experiment" to derive Archimedes' principle. Object on the right versus same shape region of pure fluid on the left.*

An object of weight mg immersed in a fluid will have an effective weight that is reduced by the buoyant force. The *effective weight* is precisely the weight that would be measured for the submerged object by, for example, a spring scale (Figure 8.19). Even objects in air can have

effective weights that are substantially less than their actual weight of mg, as is the case for a balloon, for example. The effective weights of animals are only slightly affected by air but are substantially reduced in water. For instance, a man weighing 800 N (about 180 lbs) in air will weigh about 55 N in water, or only 7% of his weight in air (see Example 8.8). Furthermore, because the density of animals is so close to that of water, small variations in the average density of an animal, obtained by varying the volume of air in the lungs or in an air-filled sac (the swim bladder in fish), can determine whether the animal will float or sink in water. Most land animals, quite independently of the animal's size, can float in water only when their lungs are inflated with air. This is true because, surprisingly, the lungs occupy about the same fraction (6%) of the total volume of most land animals. Fish can adjust the volume of air in their swim bladder to maintain neutral buoyancy by the exchange of dissolved blood gases.

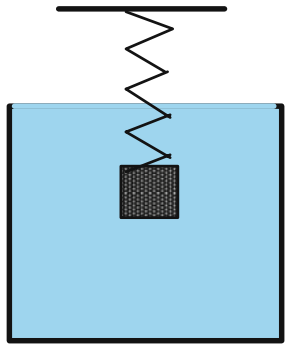


FIGURE 8.19 *The effective weight of an object immersed in a fluid is reduced by the buoyant force.*

Example 8.8 An 800 N man displaces a volume of 0.076 m^3 when submerged in a swimming pool. Calculate his effective weight when submerged in the pool. Repeat the calculation when he is in the ocean. Most land animals can float in water if they keep their lungs fully inflated, but sink if they exhale. Using the above data, find the % increase in body volume when the lungs are fully inflated, if you assume that with the lungs fully inflated the body is just neutrally buoyant in fresh water.

Solution: When in water, the man's weight and buoyant force add (as vectors) to yield his effective weight. The buoyant force on him is given by the weight of the displaced water or by the density of water times his volume times g. We can then write that

$$F_{W,\text{eff}} = mg - \rho_w gV,$$

so that in fresh water we have

$$F_{W,\text{eff}} = 800 - 1000 \cdot 9.8 \cdot 0.076 = 55\ N,$$

and in sea water we have

$$F_{W,\text{eff}} = 800 - 1025 \cdot 9.8 \cdot 0.076 = 37\ N.$$

Note that the slight (2.5%) increase in the density of sea water has reduced the man's effective weight by about 33%.

For the next part of the problem, we are told to assume that with inflated lungs the body is neutrally buoyant in fresh water. This implies that the increased volume must produce an additional 55 N of buoyant force. Setting $\rho_w gV = 55$ N, the extra volume of the inflated lungs corresponds to 0.0056 m^3. This is about 7% of the volume of the body, a typical value for the lung volume.

One further application of these ideas is the determination of average human body density in order to estimate body fat content. A simple method of estimating body fat is to simply take a person's height and weight and calculate a body mass index (BMI) given by mass (in kg)/height^2 (in m^2). There is a normal range of BMI, as well as a threshold for obesity (BMI = 30), where the numbers were originally meant to simply classify people.

A more accurate method of determining body fat content is to realize that fat has a lower density (900 kg/m^3) than water, whereas bone and muscle have a somewhat

higher density than water (so that fat-free mass has an average density of 1100 kg/m^3). If a person submerges herself in water and exhales to remove air from her lungs, then she will float if her average density is that of water and sink if it is greater. A simple formula can be used, based on the average bone and muscle content, to compute the amount of fat content. If a person were to float after exhaling all air, the fat content would be over 40%, indicating extreme obesity, so most people will sink in water with no air in their lungs. Measurements of displaced water (or in newer techniques, small changes in air pressure in a sealed chamber) can be used to determine body fat content to an accuracy of about 1%.

Example 8.9 The tallest iceberg ever measured was 168 m above sea level. Assuming it was in the shape of a large cylinder, find its depth below the surface. (Ignore the variation in the density of water or ice with depth or temperature.)

Solution: The iceberg must float with a volume below the surface such that the weight of the displaced water equals a buoyant force that just balances the weight of the entire iceberg. If h is the height of the cylindrical iceberg below the surface then we must have

$$F_{\text{buoyant}} = \rho_w Ahg = F_{\text{weight}} = \rho_{\text{ice}} A(h+168)g,$$

where A is the cross-sectional area of the cylinder (this cancels and does not affect the result) and the densities of sea water and ice can be obtained from Table 8.1. We find, canceling the common factor Ag that

$$1.025\, h = 0.917(h+168),$$

so that $h = 1430$ m, or 0.9 mi.

6. THE MEASUREMENT OF PRESSURE

Suppose that a long tube closed at one end is filled with a liquid and quickly inverted into an open bowl containing the same liquid, as shown in Figure 8.20. If the tube is sufficiently long, the fluid column will drain into the dish until the column of liquid has a characteristic height. The height of the column is determined by the balance of the pressure due to the atmosphere, P_{atm}, acting on the liquid in the open bowl and the pressure at the base of the column of liquid due to its weight ρgh. In the following example we calculate the height of a column of water and a column of mercury for this situation. It should become clear why such a device for measuring atmospheric pressure, known as a *barometer*, uses a column of mercury.

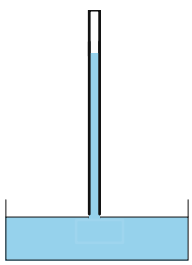


FIGURE 8.20 *A simple barometer.*

Example 8.10 Calculate the height of a column of water or mercury when a long tube closed at the bottom is filled and then inverted into an open container with the same liquid (see Figure 8.20). Based on this result, what is the maximum theoretical length of a functioning straw for sucking water up; that is, above what height would it be impossible to suck water up in such a straw.

Solution: The water in the tube will fall until the atmospheric pressure on the open container of water is sufficient to support the remaining column of water. Because the tube is closed there is no additional pressure on the top of the

column of water (the space above the column is evacuated) so that the downward atmospheric pressure on the open container of fluid, by Pascal's principle, is the same as the pressure at the base of the tube:

$$P_{\text{atm}} = \rho g h.$$

Solving for the height of the water column we find

$$h = \frac{1.01 \times 10^5}{0.998 \times 10^3 \cdot 9.8} = 10.2 \text{ m}.$$

An enormous column of water would be required!

If the calculation is repeated using a column of mercury with a density 13.6 times greater, we find a height of 0.76 m. Mercury, because of its high density and correspondingly small value of h, is often used in devices to measure pressure. In fact, a unit of pressure known as the torr is often used where 1 torr = 1 mm Hg. Atmospheric pressure is then often quoted as 1 atmosphere (atm) = 760 torr, as we have just calculated.

If we imagine a long empty straw immersed in water, then no matter how hard we suck on the straw—even connecting it up to a vacuum pump—the water will never rise more than 10.2 m. This can be seen by the fact that in our closed inverted tube, in the ideal case there is a perfect vacuum above the water. The only way to have water rise higher than this height (33.5 ft) is to push it higher by exerting a greater pressure on the column of water from below rather than by trying to lower the pressure from above.

The same principle used in a barometer to measure atmospheric pressure can be applied to measure the pressure at some location in any fluid. Figure 8.21 shows an open-tube *manometer*. The height difference in the fluid column is determined by the excess pressure (over atmospheric) at the closed end in contact with the fluid so that

$$P_{\text{gauge}} = P - P_{\text{atm}} = \rho g h. \tag{8.20}$$

This follows because pressure is only a function of height. The pressure at the lower level of the fluid column is equal to both the absolute unknown pressure P, and to the

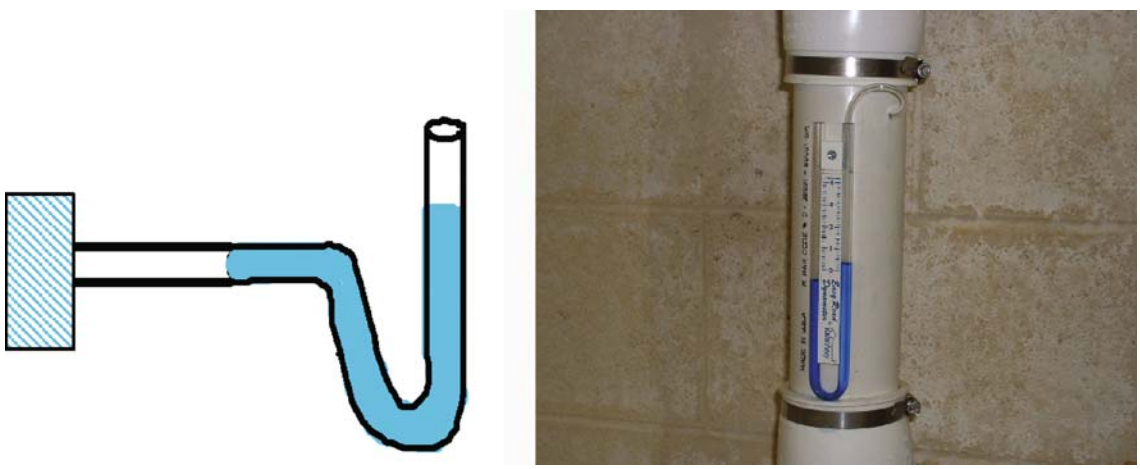


FIGURE 8.21 *Measuring the pressure within a fluid using a manometer. The photo on the right shows a manometer used to measure the pressure difference in an exhaust system for radon within a basement.*

FIGURE 8.22 *Sphygmomanometer being used to measure blood pressure*

sum of atmospheric pressure at the open end of the manometer plus the pressure due to the excess fluid column on the open side of the tube.

A device known as a *sphygmomanometer*, which uses a mercury manometer in combination with a cuff, is used to measure cardiac blood pressure (Figure 8.22). The cuff is wrapped around the upper arm, at the same height as the heart so that the pressure measured will be that at the aortic valve (output) of the heart. If the cuff were used on the leg of a standing person, for example, there would be a significant ρgh correction (where ρ is that of blood), as we have already seen in Example 8.7. The cuff is inflated with sufficient air to stop blood flow in the brachial artery and, while listening with a stethoscope just below the cuff, air is slowly let out of the cuff. When the pressure, monitored on the mercury manometer, is just below the systolic (or maximal) pressure, blood will periodically enter the brachial artery at the high-pressure portions of the cardiac cycle. The blood flow can be heard through the stethoscope due to the turbulent flow that produces an audible tapping noise, known as the Korotkoff sounds. As the air is further let out of the cuff, the blood flow increases but is still absent during the diastolic (minimal) pressure portion of the cycle, so that one hears blood flow still only during a portion of each cycle. Once the cuff pressure is reduced below the diastolic pressure, blood flow becomes continuous, although progressively less noisy as the turbulence decreases. Typical values for the systolic and diastolic gauge pressures are 120 mm and 80 mm of Hg, cited as a blood pressure of 120/80. Significantly higher blood pressures are indicative of a cardiovascular disease known as hypertension (high blood pressure), which can often be controlled using medication.

There are other pressure gauges that combine the principles of the pressure of fluid columns with mechanical devices in order to rotate needles or compress springs in proportion to the applied pressure (as in the tire gauges shown in Figure 8.23). Another type of pressure sensor uses a diaphragm, or membrane, such as that in a loudspeaker. The pressure applied distorts the diaphragm and an electrical signal is produced that is related to the applied pressure. There are also devices that use a *piezoelectric crystal*, a crystal which when compressed by small displacements due to pressure produces an electrical response that can be directly measured. We see these devices again in the next chapter when we discuss the generation and detection of ultrasound. Such devices represent one type of a class of devices known as *transducers*, devices that take energy in one form and transform it to another form. In the case of piezoelectric crystals the two forms of energy are mechanical and electrical.

FIGURE 8.23 *Tire gauge and high pressure gas gauge used to measure (gauge) pressure.*

CHAPTER SUMMARY

Pressure within a fluid is equal to the magnitude of the normal force per unit area

$$P = \frac{F}{A}. \qquad (8.2)$$

If a fluid is confined to some container, Pascal's principle states that the external pressure on the fluid increases the pressure uniformly throughout the fluid by the same amount. This leads to the hydraulic effect where the output force can be amplified according to

$$F_{\text{out}} = \frac{A_{\text{out}}}{A_{\text{in}}} F_{\text{in}}, \qquad (8.3)$$

where the ratio of cross-sectional areas results in the amplification.

A fluid flowing at a rate Q (volume per unit time) and with a velocity v through a tube or channel with a cross-sectional area A that changes will obey the continuity equation,

$$Q = A_1 v_1 = A_2 v_2. \qquad (8.5)$$

Conservation of energy in fluid dynamics (in the absence of appreciable viscosity, said to be an ideal fluid) takes the form of Bernoulli's equation

$$P + \frac{1}{2}\rho v^2 + \rho g y = \text{constant}, \qquad (8.11)$$

where ρ is the fluid density and y is its height. The text considers three special cases of this general equation:

$$P + \frac{1}{2}\rho v^2 = \text{constant} \quad \text{(constant height)} \qquad (8.12)$$

when there is no change in fluid height;

$$\frac{1}{2}\rho v^2 + \rho g h = \text{constant} \quad \text{(constant pressure)} \qquad (8.13)$$

when there is a constant pressure in the fluid. As a special example of this the efflux velocity from a large container of an ideal fluid is given by Torricelli's theorem:

$$v = \sqrt{2g(y_2 - y_1)}, \qquad (8.15)$$

and

$$P_2 = P_1 + \rho g h, \quad \text{(statics)}, \qquad (8.18)$$

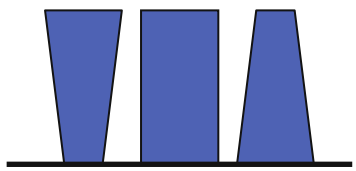
when the fluid is in static equilibrium ($v = 0$). In this latter case, usually the reference pressure is atmospheric and the static pressure reduces to atmospheric pressure plus the gauge pressure ($\rho g h$)

$$P = P_{\text{atm}} + \rho g h. \qquad (8.19)$$

Archimedes' principle states that the buoyant force on an object is equal to the weight of the fluid displaced by the object.

QUESTIONS

1. If two identical solid pieces of steel are glued together does the single piece formed have a different density than either piece? A different mass than the total of the two pieces? A different volume than the total of the two pieces?
2. Consider the previous question for the situation when you mix together one volume of water and 0.1 times that volume of salt. How does the final density compare with that of the starting materials? How does the final mass compare? The final volume? (You need to inject some independent thought here.)
3. The density of seashells is the same as that of aluminum. Does this mean that the molecules have the same mass?
4. Explain why when siphoning gas out of a car's fuel tank with a length of tubing, the tubing needs to be "primed," or completely filled, beforehand.
5. Why does water boil at a lower temperature than 100°C when at high altitudes? Give a molecular basis for this phenomenon.
6. Explain why the pressure at the bottom of the three water-filled vases shown is the same even though the weight of water in each is different.

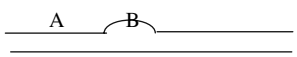


7. When a plane is rapidly descending to land, your ears will "pop" due to the rapid pressure change, and your hearing will then become clearer. What is this due to and explain why swallowing often helps to relieve this.

8. When a rapid high-rise elevator accelerates upward, some people feel a bit light-headed. What causes this?
9. Distinguish between steady and unsteady flow and between laminar and turbulent flow. In which categories are vortices, or swirling flow like that down a swiftly draining bathtub?
10. Why does the flow of water in a stream or river increase at a narrowing or obstacle due to rocks?
11. If you hold two pieces of paper vertically next to your mouth and you blow out between them, what will happen to the papers? Why? Try it!
12. If a styrofoam cup is filled with water and a pencil used to puncture a hole in the bottom, water leaks out, but stops while the cup is in the air if it is dropped. Why?
13. Why is it easier to float in fresh water with your lungs full of air than when empty? Why is it easier to float in the Great Salt Lake than in a freshwater lake?
14. Are you more or less buoyant in a hot mineral springs than in the cold ocean, assuming the same salt content?
15. A full cup of lemonade has ice cubes floating in it. Why doesn't the cup overflow when the ice melts?
16. Discuss the role of Bernouilli's equation (in the form of Equation (8.12)) in providing fresh air within animal burrows with several connecting holes to the surface. Consider how different wind velocities over the different holes provide a driving force for air exchange.
17. Why would a blood pressure reading measured on your thigh be in error? Would the pressure measured be too high or too low? What would happen if you were lying down while your blood pressure was measured on your thigh?

MULTIPLE CHOICE QUESTIONS

1. You are originally 1m beneath the surface of a pool. If you dive to 2 m beneath the surface, what happens to the absolute pressure on you?
   (a) It quadruples.
   (b) It less than doubles.
   (c) It doubles.
   (d) It more than doubles.
2. Which of the following is a false statement about an aneurysm (weakening of an artery wall)?
   (a) The flow rate through the artery at A is the same as that at B.
   (b) The velocity at B is less than that at A.
   (c) The pressure at B is less than that at A.
   (d) The density at B is the same as that at A.



3. A plastic bag full of empty, unsquashed aluminum soda cans has a volume of 1 m^3. The density of aluminum is 2700 kg/m^3 and the density of air is about 1 kg/m^3. The mass of the bag is 0.05 kg. The mass of the bag and its contents is (a) 2700 kg, (b) between 2700 kg and 2701.05 kg, (c) exactly 2701.05 kg, (d) a few kg.
4. A pail, held in the air, is initially completely filled with water. Ten cm down from its top, the pail has a hole of diameter less than 1 cm that is sealed by a piece of tape. When the tape is removed the speed of the jet of water that immediately spurts out (a) is about 1.4 m/s, (b) is about 4.5 m/s, (c) zero because air pushes into the pail at the hole, (d) depends on the value of the diameter of the hole and the diameter of the pail.

Questions 5 and 6 refer to: A boat floating in a lake contains a block of volume V_0. The density of the block is 5000 kg/m^3.

5. The volume of water displaced by the boat to keep it afloat includes what amount that is associated with the block? (a) $5V_0$, (b) V_0, (c) $V_0/5$, (d) none, because the block isn't in the water.
6. If the block is thrown overboard, the volume of the water displaced associated with the block is (a) $5V_0$, (b) $4V_0$, (c) V_0, (d) $V_0/5$.
7. Two hoses, one of 20 mm diameter, the other of 15 mm diameter, are connected one after the other to a faucet. At the open end of the hose, the flow rate of water is 10 L/min. Through which hose does the water flow faster? (a) the 15 mm hose, (b) the 20 mm hose, (c) the water velocity is the same in both cases, (d) the answer depends on which of the two hoses comes first in the flow (i.e., which is attached to the faucet).
8. The apparent weight of an immersed body has the same magnitude as (a) the weight of the body, (b) the difference between the weights of the body and the displaced fluid, (c) the weight of the fluid displaced by the body, (d) the average pressure of the fluid times the surface area of the body, (e) none of the above.
9. A piece of wood is floating in a bathtub. A second piece of wood sits on top of the first piece and never touches the water. If the top piece is taken off and placed in the water, what happens to the water level in the tub? (a) It goes up, (b) it goes down, (c) it does not change, (d) cannot be determined from the information given.
10. If you dangle two pieces of paper vertically a few inches apart and blow air between them (a) the papers will move apart because the air pressure exerts an outward force, (b) the papers will move together because the increased velocity reduces the pressure between the papers, (c) the papers will remain vertical, (d) the papers will move apart because the air friction causes an increased pressure.
11. It is conventional to give blood pressure as a gauge pressure measured in mm Hg. If a person's diastolic pressure is 76 mm Hg and atmospheric pressure is 760 mm Hg the absolute diastolic pressure is (a) 10% of atmospheric pressure, (b) 90% of atmospheric pressure, (c) the same as atmospheric pressure, (d) 110% of atmospheric pressure.
12. In a swimming pool the gauge pressure on a person's head is P_0 and the buoyant force on the person is F_0

when the person's head is submerged to a depth D. When the person's head is submerged to a depth $2D$ (and his orientation is the same) the pressure on his head and the buoyant force he experiences are about (a) P_0 and F_0, (b) P_0 and $2F_0$, (c) $2P_0$ and F_0, (d) $2P_0$ and $2F_0$, respectively.

13. A large truck passes close to your car traveling in the opposite direction at a fairly high speed. After the front of the truck passes by, you feel your car pulled toward the truck. This is most likely due to (a) the air moving over your car faster on the side closer to the truck than on the side farther from the truck, (b) the air moving over your car slower on the side closer to the truck than on the side farther from the truck, (c) Archimedes' principle, (d) Pascal's law.

14. The fundamental dimensions of pressure are (a) MLT^{-2}, (b) ML^{-3}, (c) MLT^{-1}, (d) $M\,L^{-1}\,T^{-2}$.

15. A sphere of volume 1 m^3 and a rectangular solid whose dimensions are $2\text{ m} \times 2\text{ m} \times {}^1/_4\text{ m}$ (i.e., also with volume 1 m^3) are immersed to a depth of about 2 m in water. The rectangular solid is oriented so that its 2×2 side is horizontal. Which of the following is true? (a) The buoyant force on the sphere is greater than that on the rectangular solid. (b) The buoyant force on the sphere is less than that on the rectangular solid. (c) The buoyant force on the sphere equals that on the rectangular solid. (d) The buoyant forces on each object depend on the materials from which they are made.

16. Which of the following is a direct result of the equation of continuity? (a) The pressure one meter below the surface of water is about 10% greater than atmospheric pressure. (b) The pressure on the inlet side of a horizontal pipe through which water is flowing at a constant speed equals the pressure on the outlet side. (c) When the open end of a long evacuated tube is inserted into a pool of water, water rises to about 10 m in the tube. (d) Placing your thumb partially over the opening of a hose causes the velocity of the water leaving the hose to increase.

17. When blood flows through into an aneurysm (a) the velocity slows and the pressure increases, (b) the velocity slows and the pressure decreases, (c) the velocity increases and the pressure increases, (d) the velocity increases and the pressure decreases.

18. Which of the following best describes how a plane's wing generates lift? (a) The wing is thick so the air pressure on the top is less than the air pressure on the bottom by an amount equal to $\rho_{air}gh$ (h is the thickness). (b) The wing is curved; to assure continuity of flow, air has to pass over the top faster than over the bottom, and a higher pressure on the bottom than the top results. (c) The curvature of the wing forces air to be more densely packed on its bottom than on its top and a higher pressure on the bottom than on the top results. (d) As the plane flies through the air the wing vibrates vertically causing momentum to be departed to the air, in the same manner that birds fly.

19. A passenger in the back seat of a moving car is smoking. The driver opens a front window slightly and the smoke is drawn out of the car through it. This is due primarily to (a) Bernoulli's equation, (b) Archimedes' principle, (c) Pascal's principle, (d) the equation of continuity.

20. A hot-air balloon rises to a maximum height and then stays there. A rock falls to the bottom of a lake, independent of how deep it is. The difference between these two effects is most directly related to (a) terminal velocity in air is smaller than in water, (b) pressure in air is greater than in water, (c) air is compressible but water isn't, (d) the balloon cools as it rises but the rock's temperature doesn't change.

21. A 1 m by 1 m square plate lies on the ground exposed to the air. An identical plate lies inside an evacuated chamber with sand piled on top of it. What mass of sand is required to make the downward force on both plates equal? About (a) 10,000 kg, (b) 100 kg, (c) 1 kg, (d) 0.01 kg.

22. Which one of the following is Bernoulli's equation not involved in explaining? (a) Why a roof can blow off a house in a hurricane, (b) the buoyant force on a floating iceberg, (c) dynamic lift on airplane wings, (d) how fast water sprays out from a hole in a water tank, (e) it can explain all of these.

PROBLEMS

1. What is the radius of a 0.1 m long 0.2 kg mass aluminum cylinder of density 2700 kg/m^3?

2. A 5 m diameter circular in-ground swimming pool has a 20 cm thick layer of ice over 150 cm deep water at 4°C. Find the total mass of the ice and water.

3. The least dense solid is an aerogel of silica, first produced in 1990. If a cubic slab 10 cm on a side weighs 0.05 N, find its density.

4. Given the mean radius of the Earth, 6.38×10^6 m, and the fact that the Earth's mean density is about 5.5 times that of water, find the mass of the Earth.

5. Find the pressure exerted on the ground by a 2600 lb (1180 kg) car with each tire having a surface area of 36 in^2 (232 cm^2) in contact with the ground. Assume the weight of the car is uniformly distributed to the four tires.

6. If a scuba diver 10 m below the water surface were to hold his breath and rise to the surface, what would be the pressure change in the lungs? Assuming the ideal gas law, $P \propto 1/V$, by what factor would the air in the lungs expand? Divers learn to exhale continuously on ascent because of this effect.

7. The lungs can exert a negative pressure, with respect to atmospheric pressure, of up to 1.3 kPa. To what height can you suck water through a straw?

8. What is the pressure difference, in Pa and in mm Hg, between sea level and Breckinridge, CO at an altitude of 9600 ft (2926 m)?

9. The deepest part of any ocean is thought to be in the Mariana Trench in the Pacific and is about 35,800 ft below the surface. What is the hydrostatic pressure at this depth?

10. Find the upward acceleration of your body at which the heart would no longer be able to pump blood to your head. Use an average blood pressure of 100 mm Hg and take the vertical distance from your heart to the bottom of your head to be 0.3 m. (Hint: First convert the blood pressure to mm blood.)

11. A pressure differential of about 120 mm Hg across the eardrum can cause it to rupture. To what depth can a diver go before this occurs? One solution is for a diver to equalize the pressure by raising the pressure in the mouth (and Eustachian tubes) by holding the nose and "blowing" out; this also works well to equalize the pressure when landing in an airplane.

12. A wooden wine barrel is filled with water and a very long narrow tube connected to its lid. If the tube is gradually filled with water, the barrel pressure increases. If the maximum force the lid on the barrel can sustain is 14,000 N, find the height of the column of water needed to burst the barrel. Take the barrel lid to have a 40 cm radius and the tube to have a 2.5 mm radius. Calculate the weight of the water in the tube and compare it to the force exerted on the barrel lid.

13. Water flows through a horizontal Venturi tube with a section with a large inner radius of 2.5 cm and a section with a smaller inner radius of 1 cm that is 4 cm long. If the flow rate into the larger diameter section of the tube is 30 cm^3/s, find the following (neglecting the viscosity of water).
(a) The water speed in both the larger and smaller cross-sections of the tube.
(b) The water pressure difference between the two sections of the tube.
(c) What is the buoyant force on a spherical bubble of 0.1 mm radius trapped within the tube?

14. Water is being pumped through a horizontal pipe of 1 cm inner diameter by a gauge pressure equal to one tenth of atmospheric pressure so that the flow rate is equal to 1000 cm^3/min.
(a) Find the velocity of the water (neglecting viscosity).
(b) In a region where dirt has accumulated and the inner diameter is reduced by half, find the internal gauge pressure in the water.

15. A gardener is watering his garden from a hose. With the water pressure full blast holding the hose horizontally he can just reach a distance of 12 m, but needs to water an area up to 18 m away. By what fraction must he reduce the cross-sectional area of the hose, still keeping the hose horizontal, to be able to water this area?

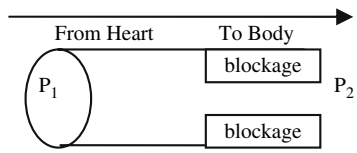
16. The Buckingham Fountain in Chicago is famous for its water displays. Suppose that you are watching the water from the fountain and notice that a strean of water is being shot upward. You also notice that the stream has a slight inclination to one side so that the descending water does not interfere with the ascending water. The upward velocity at the base of the column of water is 15 m/s.
(a) How high will the water rise?
(b) The diameter of the column of water is 7.0 cm at the base. What is the diameter at the height of 10 m?

17. What is the average speed of blood in the aorta? The volume flow rate of blood is known to be about 5 L/min; take the aorta diameter to be 1.8 cm.

18. The cross-sectional area of the aorta A_0 (the major blood vessel emerging from the heart) of a normal resting person is 3 cm^2, and the speed v_0 of the blood through it is 30 cm/s. A typical capillary (with diameter approximately 6 μm) has a flow speed v of 0.05 cm/s. How many capillaries does such a person have?

19. Suppose that the aorta has a radius of about 1.25 cm and that the typical blood velocity is around 30 cm/s and that it has an average density of 1050 kg/m^3.
(a) What is the average blood velocity in the major arteries if the total cross-sectional area of the major arteries is 20 cm^2?
(b) What is the total flow rate?
(c) If the blood in the circulatory system goes through the capillaries, what is the total cross-sectional area of the capillaries if the average velocity of the blood in the capillaries is 0.03 cm/s?
(d) If a typical capillary has a cross-sectional area of 3 $\times 10^{-11}$ m^2, about how many capillaries are there in the human body?
(e) What is the kinetic energy per unit volume for blood in the aorta, the major arteries, and the capillaries?
(f) If a capillary has an average length of 0.75 mm what is the average time that a red blood cell remains in a capillary?

20. The human heart is a mechanical pump. The aorta is a large artery that carries oxygenated blood away from the heart to various organs in the body. For an individual at rest, the blood in the aorta (of radius 1.25 cm) flows at a rate of 5 $\times 10^{-3}$ m^3/min.



(a) What is the velocity, in meters per second, of the blood in the aorta?
(b) Suppose that the blood flows continuously throughout the body (and not in spurts as it really does); what is the kinetic energy of the blood, per unit volume of blood, in the aorta? (Hint: The density of blood is 1050 kg/m^3.)
(c) Every time that the heart beats, it does work moving the blood into the aorta and then into the body. Suppose that the heart does work at a rate of 0.5 W.

What is the change in pressure across the aorta? (Hint: The power is the work done moving the blood per unit time.)

(d) Suppose that the difference in pressure in part (c) were due to an aortic blockage as shown above. What is the velocity of the blood through the blockage if the person were lying horizontally? This is a medical condition known as atherosclerosis.

21. In the Old West, a cowgirl fires a bullet into an open water tank creating a hole at a distance of 5 m below the water surface (which is open to the air). What is the speed of the water emerging from the hole?

22. A bottle of saline solution (with a specific gravity of 1.02) is attached to a 1.2 m long piece of tubing with a 1.0 cm inner diameter. If the tubing is held vertically, filled with saline, and clamped at the bottom, what is the gauge pressure at the bottom of the tube and what would be the initial efflux velocity of the saline if the clamp were released? (Take the height of the saline solution in the bottle to be 10 cm.)

23. The diameter of a capillary, as small as 10 μm, is very much smaller than that of the aorta. A naïve application of the continuity equation would lead to the conclusion that the blood velocity in a capillary is very much faster than in the aorta, but this is not true. Actually, the blood from the aorta branches out to a vast network of arteries and eventually capillaries with a total effective cross-sectional area of about 1000 times that of the aorta. Using this information and the numbers problem 17, find the velocity of blood in a capillary.

24. Similar to the last problem, calculate the velocity of air in the alveolar ducts of the lungs, assuming a tube-like diameter of 0.4 mm, knowing that the trachea has a diameter of 18 mm, the total effective cross-sectional area of the alveolar ducts is 5880 cm^2 and assuming an average flow rate of 500 cm^3 per 2.5 s. What is the air velocity in the trachea?

25. A spherical balloon filled with air to a diameter of 20 cm is submerged in water. Find the force needed to hold it under the water.

26. Spinal fluid pressure can be measured using a spinal tap in which a needle is inserted in the patient's lower back with the patient sitting upright on an examination table. The pressure due to the weight of the spinal fluid (given that its density is 1050 kg/m^3) in the spinal column increases the pressure.

(a) What is the pressure measured if the pressure around the brain is 10 mm Hg and the tap is at a point 75 cm lower than the brain?

(b) What is the pressure measured if the person is lying down?

27. The U.S. Navy has the largest warships in the world, aircraft carriers of the *Nimitz class* (an example of which would be the *USS Ronald Reagan*). Suppose that fifty 25 t airplanes (~22,500 kg) take off from the flight deck and the ship bobs up to float 22 cm higher in the water, in a region where $g = 9.78$ m/s^2. What is the horizontal area enclosed by the waterline of the ship? Compare this to the deck of an aircraft which has an area 20,000 m^2.

28. A 175 lb (779 N) man is submerged in water and after exhaling is found to have an apparent weight of 11.5 lb (51.2 N). Find his density and specific gravity.

29. Tom Sawyer and Huckleberry Finn want to build a raft to float down the Mississippi. Knowing their combined weight is 250 lbs (1110 N), what is the minimum number of logs required? Each log is 3 m long with a 0.1 m radius and a density of 750 kg/m^3.

30. A 4 m × 4 m × 0.3 m solid wood raft is floating in a fresh water lake.

(a) If the density of the wood is 600 kg/m^3 find the fraction of the raft above the water.

(b) How many 150 lb (670 N people) can the raft support just staying above the surface?

31. Using conservation of energy ideas contained in Bernouilli's equation, find the total electric power that could be generated if all the energy of the water going over Niagara Falls (750,000 gallons per second or 2835 m^3/s) were to be used to generate electricity. Take the water at the crest of the falls to have a velocity of 40 mph and fall an average height of 30 m.

Niagara Falls

32. Suppose that a 20 m/s wind blows over the roof of your house. Take the density of air to be 1.3 kg/m^3.

(a) Find the reduction in pressure, below atmospheric pressure in the absence of any wind, above the roof.

(b) If the roof has a surface area of 300 m^2, find the net force on the roof.

33. An airplane with a mass of 10,000 kg accelerates down a runway as it takes off. If the wings are designed to produce a faster air speed above than below the wing by 25% find the minimum speed the plane must travel on a windless day in order to take

off. Assume the area of each wing is 70 m^2 and the density of air is 1.3 kg/m^3.

34. A Boeing 777 has a mass of 2.43×10^5 kg and each wing has an area of 189 m^2. During level flight, the pressure on the lower wing surface is 700×10^4 Pa.

(a) What is the pressure on one of the upper wing surfaces?

(b) What is the upward acceleration of the aircraft if the pressure on the lower surface were to increase to 702×10^4 Pa? (This increase in pressure is due to the aircraft increasing its forward velocity and assumes that P_{upper} remains constant.)

9 Viscous Fluids

In the previous chapter on fluids, we introduced the basic ideas of pressure, fluid flow, the application of conservation of mass and of energy in the form of the continuity equation and of Bernoulli's equation, respectively, as well as hydrostatics. Throughout those discussions we restricted ourselves to ideal fluids, those that do not exhibit any frictional properties. Often these can be neglected and the results of the previous chapter applied without any modifications whatsoever. Clearly mass is conserved even in the presence of viscous frictional forces and so the continuity equation is a very general result.

Real fluids, however, do not conserve mechanical energy, but over time will lose some of this well-ordered energy to heat through frictional losses. In this chapter we consider such behavior, known as viscosity, first in the case of simple fluids such as water. We study the effects of viscosity on the motion of simple fluids and on the motion of suspended bodies, such as macromolecules, in these fluids, with special attention to flow in a cylinder, the most important geometry of flow in biology. The complex nature of blood as a fluid is studied next leading into a description and physics perspective of the human circulatory system. We conclude the chapter with a discussion of surface tension and capillarity, two important surface phenomena in fluids. In Chapter 13 we return to the general notion of the loss of well-ordered energy to heat in the context of thermodynamics.

1. VISCOSITY OF SIMPLE FLUIDS

Real fluids are viscous, having internal attractive forces between the molecules so that any relative motion of molecules results in frictional, or drag, forces. The work done by these drag forces, in turn, results in a loss of mechanical energy due to slight heating. We can think of viscosity as a measure of the resistance of a liquid to flowing, so that liquids such as paint or maple syrup have much higher viscosities than water. A quantitative definition of viscosity can be introduced using the example of laminar flow of a liquid between two parallel plates (Figure 9.1), the lower one fixed and the upper one pulled by an external force to move with a constant velocity v parallel to the surface of the plate. Clearly in the absence of drag forces the constant external force would lead to uniform acceleration of the top plate, but due to the drag forces the top plate quickly reaches a steady-state constant velocity. Because the liquid is viscous, it tends to stick to the surfaces of the plates, forming a boundary layer. Therefore the liquid layer at the fixed plate is at rest, whereas the liquid layer at the top plate moves with velocity v. For laminar flow, the velocity of the liquid varies linearly in the transverse direction (y-direction in Figure 9.1) from 0 to v over the separation distance between the plates of area A. Planar layers of fluid slide over one another.

Viscosity can be defined through the relation between the *shear stress*, or force per unit area F/A, needed to keep the upper plate moving with a constant velocity

J. Newman, *Physics of the Life Sciences*, DOI: 10.1007/978-0-387-77259-2_9,

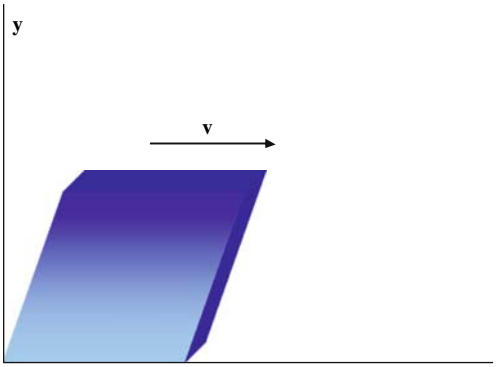


FIGURE 9.1 *A fluid sandwiched between two plates with the bottom plate fixed and the top plate moving at a constant velocity v.*

and the rate of variation of the velocity between the plates, $\Delta v/\Delta y$ (known as the *rate of strain*),

$$\frac{F}{A} = \eta \frac{\Delta v}{\Delta y}, \tag{9.1}$$

where η is the viscosity of the liquid. Contrast this with the stress–strain relation discussed in Chapter 3 for solids where the strain $\Delta x/\Delta y$ appeared on the right-hand side and not the rate of strain, appropriate here for fluids. Strain and rate of strain are connected in the usual way because the time rate of change of strain is given by $(\Delta x/\Delta y)/\Delta t = (\Delta x/\Delta t)/\Delta y = \Delta v/\Delta y$. The SI unit for viscosity is the Pa-s, but another commonly used unit is the poise (P; 1 P = 10 Pa-s). Table 9.1 lists viscosities of water and blood. Equation (9.1) can be taken as the definition of viscosity, originally due to Sir Isaac Newton. Fluids that obey this relation are said to be *Newtonian fluids*. The proportionality of the shear stress and rate of strain usually holds only at lower strain rates. Water and salt solutions are Newtonian, whereas blood, whose behavior does not follow Equation (9.1), is said to be a non-Newtonian fluid and is discussed in the next section.

Table 9.1 *Viscosities of Water and Blood*

Fluid	*Temperature*	*Viscosity (10^{-3} Pa-s)*
Water	0	1.8
	20	1.0
	37	0.7
Whole blood[a]	37	4.0
Blood plasma	37	1.5

[a] Varies greatly with hematocrit, or red blood cell content.

When a solid is put under shear stress, with an external force applied in a particular direction, it deforms and, for small stresses F/A, the strain, or response of the solid, is proportional to the stress. Once the stress is removed, the solid returns to its original shape (unless it has some plasticity, in which case it may flow). In a Newtonian liquid, however, a constant applied shear stress results in a constant *rate of strain* (Equation (9.1)) rather than constant strain. The larger the rate of strain, meaning the more abruptly the velocity changes with transverse distance, the greater the viscous force, and in turn, the greater the applied shear stress needed to keep the top plate moving at the same constant velocity. At higher shear stress there are deviations from this relation, and at still higher stress, turbulence will occur.

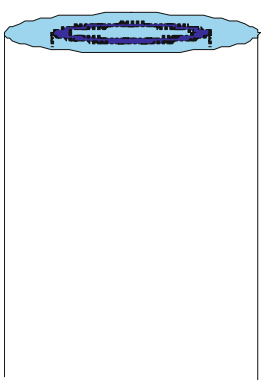


FIGURE 9.2 *Laminar capillary flow showing a concentric layer of fluid that flows at the same velocity along the length of the tube.*

Example 9.1 A sheet of plywood is covered with a 1 mm thick layer of tile adhesive and a square piece of ceramic tile measuring 30 cm on a side is placed on it. If a force of 10 N is applied parallel to the surface, find the velocity with which the tile slides. Assume laminar Newtonian flow and use a viscosity of 50 Pa-s for the adhesive.

Solution: We first calculate the stress as $F/A = 10/(.3)^2 = 110\ \text{N/m}^2$. Dividing this stress by the adhesive viscosity, the rate of strain is found to be $2.2\ \text{s}^{-1}$, so that the velocity of the tile is given as

$$v = \frac{\Delta v}{\Delta y} y = (2.2\,\text{s}^{-1})(1\ \text{mm}) = 2.2\ \text{mm/s}.$$

The capillary tube is a very common geometry for fluid flow in biology. It is relevant for blood flow, for example, as well as for *viscometry*, the methodology of

viscosity measurement. When a liquid flows through a tube without obstacles, the flow at low velocities is laminar with layers of liquid in concentric cylinders (Figure 9.2). The outermost layer is the boundary layer that remains at rest and the fastest flowing liquid lies at the center of the tube. The actual velocity profile across the tube is parabolic as indicated in Figure 9.3. The velocity varies across the capillary tube; thus in order to find the volume flow rate, Q ($= vA$ when the velocity was assumed uniform in the absence of viscosity), an average must be calculated across the cross-sectional area. This was first done in 1835 by Poiseuille, a French physician interested in blood flow (the viscosity unit poise is taken from his name), who found

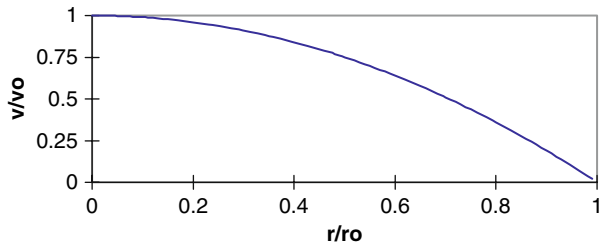


FIGURE 9.3 *The velocity profile across a capillary tube of radius* r_o.

$$Q = \frac{\pi P r^4}{8\eta L}, \tag{9.2}$$

where P/L is the applied pressure per unit length of the tube and r is the tube radius. Equation (9.2) is known as *Poiseuille's law*.

If we rewrite this equation in the form

$$\Delta P = \left(\frac{8\eta L}{\pi r^4}\right) Q, \tag{9.3}$$

where we write ΔP as the pressure difference across the tube of length L, then we can interpret the equation as follows. For a given ΔP across the tube, the resulting flow Q depends on the resistive term in parentheses. The larger this term, the slower the flow rate is for a given applied pressure. With a constant resistive term (fixed tube length, radius, and fluid viscosity), the greater the pressure difference acting on the liquid, the greater is the expected flow rate. A longer tube or larger viscosity provides a greater resistance to flow as might be intuitively expected. The very strong dependence of the resistive term on tube radius r^{-4} is surprising and extremely significant in controlling the flow rate of a liquid in a capillary tube. The resistance to fluid flow increases dramatically as the tube radius gets smaller. This can lead to important effects in the flow of blood in arteries because a partially clogged artery will require a much higher pressure differential to supply the same fluid flow rate.

Example 9.2 In giving a transfusion, blood drips from a sealed storage bag with a 1 m pressure head through capillary tubing of 2 mm inside diameter, passing through a hypodermic needle that is 4 cm long and has an inside diameter of 0.5 mm. If the blood pressure within the vein into which the blood is being transfused is at a gauge pressure of 18 torr, find how long it will take to give the patient 1 L of blood. How long will it take if the inside diameter of the needle is only 0.4 mm?

Solution: Since the flow rate depends so strongly on the radius of the capillary, the most resistance to flow will occur within the hypodermic needle and relatively little within the delivery tubing. We can therefore apply Equation (9.2) using the radius and length of the needle, ignoring the dimensions of the tubing. For the net driving pressure across the column of blood up to the vein we use a value of $P = (\rho gh - 18 \text{ torr}) = (\rho gh - 2400 \text{ N/m}^2)$, where the density of blood

(*Continued*)

is found in Table 9.1 and we have used the conversion from $P_{atm} = 10^5$ N/m^2 = 760 torr. We find a flow rate of

$$Q = \frac{\pi P r^4}{8\eta L} = \frac{\pi[(1.06 \times 10^3)(9.8)(1) - 2400](0.00025)^4}{8(4 \times 10^{-3})(0.04)}$$
$$= 7.7 \times 10^{-8} \text{ m}^3/\text{s} = 0.077 \text{ cm}^3/\text{s}.$$

With this flow rate, each cm^3 of blood will take 13 s to flow into the vein, so that it will take a total time of 3.6 h for a liter of blood to be transfused. If the r value is 0.2 mm then the flow rate will decrease by the factor $(2/2.5)^4 = 0.41$ and so it will take 3.6/0.41 = 8.8 h. We see that a decrease in the radius by a factor of only 0.8 increases the time required by almost 2.5 times, pointing out the very strong dependence on r.

Capillary viscometers make use of Poiseuille's law to measure the relative viscosity of liquids or solutions. They consist of a fine capillary tube in which a liquid is placed and measurements made of the time for a fixed volume of liquid to flow through the tube (Figure 9.4). Because the pressure P is equal to ρgh, where h is the height of liquid in the tube, we find from Equation (9.2) that for a given capillary tube $Q \propto \rho/\eta$, where the other parameters are independent of the liquid properties. Q is a flow rate and therefore $Q \propto 1/t$, where t is the time for a fixed fluid volume to flow through the capillary, so we have that

$$\frac{\rho}{\eta} \propto \frac{1}{t} \quad \text{or} \quad \eta \propto \rho t.$$

FIGURE 9.4 *Capillary viscometer used to measure solvent viscosity by a timing measurement.*

From measurements of the efflux times of the same volume of an unknown fluid and a standard fluid, we can take the ratio to write that

$$\eta_{unknown} = \eta_{stnd} \frac{\rho_{unknown}\, t_{unknown}}{\rho_{stnd}\, t_{stnd}}. \tag{9.4}$$

If the densities and standard viscosity are known, the viscosity of the unknown liquid can be determined from simple timing measurements. Results from such measurements can give accurate viscosity values for pure liquids or for solvents (typically solutions of small dissolved ions).

Thus far we have considered the flow of pure viscous fluids at low stress. At higher stress turbulence occurs and the flow profile in a capillary is much different than for laminar flow (Figure 9.5). In such turbulent flow there is a much greater effective internal friction due to vortices, and also the strain rate $\Delta v/\Delta r$ near the walls is much greater (note the more rapid velocity change near the boundary layer on the tube wall in Figure 9.5).

What happens when a flowing viscous fluid meets an obstacle, perhaps a biological macromolecule? We have already briefly considered this question in our discussion of motion in a fluid in Chapter 3, Section 2. There we introduced the dimensionless Reynolds number $\Re$, defined as

$$\Re = \frac{L\rho v}{\eta}, \tag{9.5}$$

where L is the characteristic size of the object. Imagine a sphere of radius r held fixed within a flowing fluid. Under laminar flow conditions, with $\Re$ on the order of 1 or less, the fluid will flow around the sphere in a symmetric pattern as shown in Figure 9.6 (top). There is a frictional force that acts on the sphere given by *Stokes' law*,

$$F_f = -6\pi\eta r v. \tag{9.6}$$

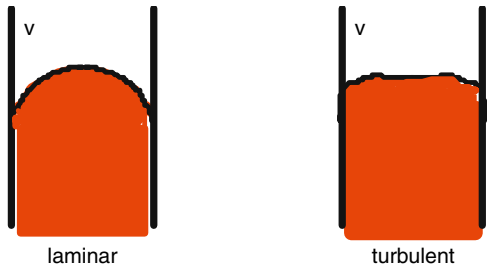


FIGURE 9.5 *Velocity profiles for laminar and turbulent capillary flow. Note that the profile of the fluid in the tube is not shown here, but rather how the velocity varies across the capillary.*

The frictional force in this case varies linearly with both the fluid velocity and viscosity as well as with the size of the sphere. As the fluid velocity is increased, the flow pattern will become more complex and asymmetric, and the frictional force will become dependent on the square of the fluid velocity, as already discussed (see Equation (3.5)). The fluid velocity downstream from the sphere is decreased as the Reynolds number is increased, and at a certain point the flow becomes unsteady with "eddies," or vortices, forming in the downstream region known as the *wake* of the object (Figure 9.6, middle); at even higher Reynolds numbers the flow becomes fully turbulent (Figure 9.6, bottom). By careful design of the shape of an object, the frictional forces can be reduced. Engineered streamlined designs have led to improved aerodynamic performance of cars and airplanes. In the world of animals, evolutionary design has also resulted in streamlined shapes particularly for many aquatic or flying animals.

The problem of determining the viscosity of a suspension of objects is a very complex one. When more than one object is present in a fluid, the wake produced by one object can interact with the other objects through what are termed *hydrodynamic interactions*. In 1906, Einstein solved the problem of determining the viscosity of a suspension of identical spherical particles η_s and found

$$\eta_s = \eta_o (1 + 2.5\Phi), \tag{9.7}$$

where η_o is the solvent viscosity and Φ is the (dimensionless) volume fraction occupied by the spheres. Note that this result does not depend explicitly on the particle radius. The larger the sphere is, the smaller the number of them required to occupy the same volume fraction and hence have the same solution viscosity. For particles of other shapes the factor 2.5 is replaced by a shape-dependent numerical factor.

Example 9.3 Find the viscosity of a 100 μM aqueous solution of a small spherical protein with radius 5 nm and molecular weight 40,000 at 20°C. This might be a solution of globular actin protein.

Solution: To proceed from Equation (9.7), we need to calculate the volume fraction occupied by the protein. Each protein molecule occupies a volume of $\frac{4}{3}\pi r^3 = 5.2 \times 10^{-25}$ m^3 and each protein molecule has a mass of $40{,}000/N_A$, where N_A is Avogadro's number, or a mass of 6.6×10^{-20} g $= 6.6 \times 10^{-23}$ kg. A 100 μM solution has a density of 40,000 g/mol $\times 10^{-4}$ mol/L $= 4$ g/1000 cm^3 $= 4 \times 10^{-3}$ kg/10^{-3} m^3 $= 4$ kg/m^3, using 1 cm^3 $= 10^{-6}$ m^3. We can then compute that in every unit volume (1 m^3) there are $4/(6.6 \times 10^{-23}) = 6.1 \times 10^{22}$ molecules occupying a volume of $(6.1 \times 20^{22})(5.2 \times 10^{-25}) = 0.03$ m^3. Thus the volume fraction is 0.03 and the viscosity is then found to be (using $\eta_o = 10^{-3}$ Pa-s for water) $\eta = [1 + (2.5)(0.03)] \times 10^{-3}$ Pa-s $= 1.075 \times 10^{-3}$ Pa-s, a 7.5% increase over pure water.

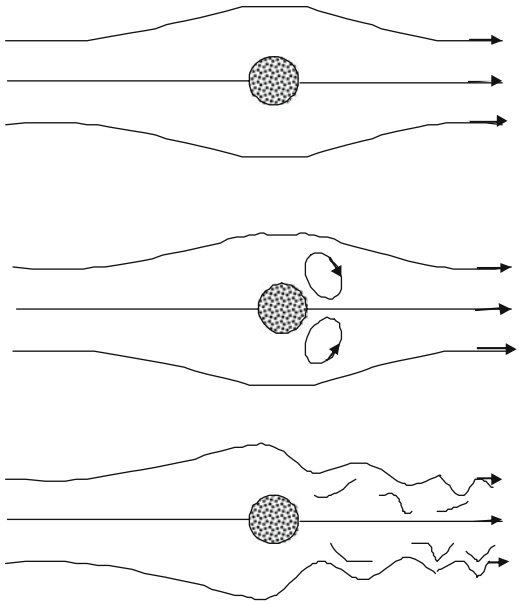


FIGURE 9.6 *Flow patterns around a sphere at increasing Reynolds numbers.*

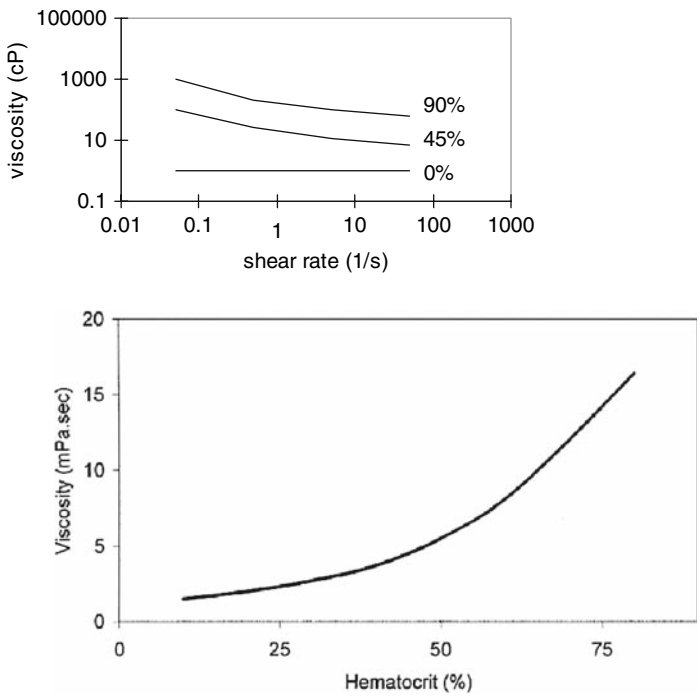
Viscosities of suspensions or solutions of macromolecules can be measured using capillary viscometers, just as for pure fluids, if the particles or macromolecules are small, so that they are not oriented by the flow in a capillary, and if a sufficient volume of material is available (typically 0.1 L). Other designs for viscometers have been developed to use smaller volumes and to extrapolate to zero shear rate in order to avoid orienting asymmetric particles.

When a DNA molecule is stretched by hydrodynamic forces during flow, it responds somewhat like a stretched rubber band, storing energy like a spring that can be recovered when the flow stops. Elastic properties of DNA and many other biomolecules seem to be very important in their functioning. Solutions of DNA and other fiberlike molecules (filamentous proteins and other elongated (bio)polymers) exhibit *viscoelasticity*, having both a measurable viscosity, or energy loss mechanism, as well as elastic storage of energy. One method by which such solutions can be studied involves more sophisticated viscometers, called rheometers (after *rheology*, the study of viscoelasticity), in which both the elasticity and viscosity are simultaneously measured to give information about the structure and functioning of these macromolecules.

2. BLOOD AND OTHER COMPLEX FLUIDS

The term "complex fluid" is usually used for a non-Newtonian fluid, meaning that the shear stress and rate of strain are not simply proportional as they are in Equation (9.1). Most biological fluids are complex, including blood. Even simple suspensions of asymmetric macromolecules are non-Newtonian due to orientation effects at higher strain rates: large transverse variations in velocity create torques on such molecules tending to align them in the flowing fluid, just as a stick aligns itself with the flow in a fast-moving stream. Other complex biological "fluids" include cellular cytoplasm, which has viscoelastic properties, and biological membranes, having two-dimensional fluidlike properties briefly discussed in Section 6 of Chapter 7. In this section we consider the composition and properties of blood as perhaps the most interesting example of a complex fluid.

FIGURE 9.7 (top) *The relation between viscosity and shear rate for whole blood (with the hematocrit shown) and blood plasma;* (bottom) *the viscosity of whole blood versus hematocrit.*



Human blood makes up about 1/13 of the total body mass and amounts to 5–6 L in the average adult male. When blood is centrifuged it separates into two portions. Plasma is the fluid component of blood and is composed by weight of about 92% water, 7% protein, and small amounts of organic and inorganic molecules as well as dissolved gases. It behaves as a Newtonian viscous fluid with a viscosity about 20% higher than that of water. The second phase that spins down in a centrifuge consists of cells, primarily red blood cells that make up over 50% of the volume of blood. Red cells, or erythrocytes, contain hemoglobin and carry oxygen throughout the body. There are also much smaller numbers of white blood cells and platelets in blood. The white cells, or leukocytes, come in five varieties and are capable of amoeboid motion and one variety, the neutrophils, can migrate out of small blood vessels and play a role in fighting infections by engulfing bacteria throughout the body in a process called phagocytosis. Platelets are small cells that are involved in blood clotting. All of these cells have finite life spans ranging from one or two days to several months and are replenished by the bone marrow.

Figure 9.7 (top) shows data for the viscosity of whole blood at three different hematocrits (the percent

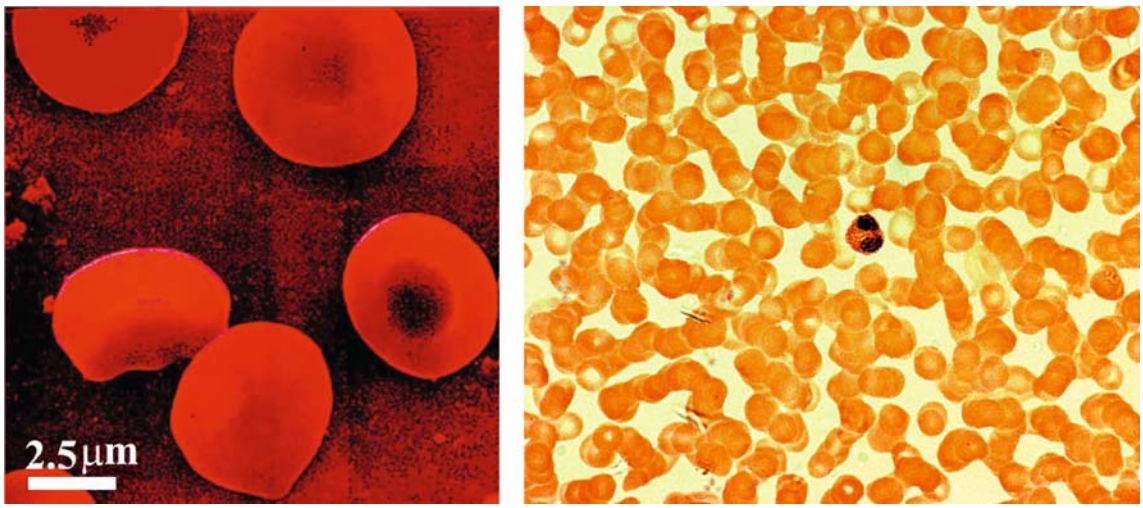
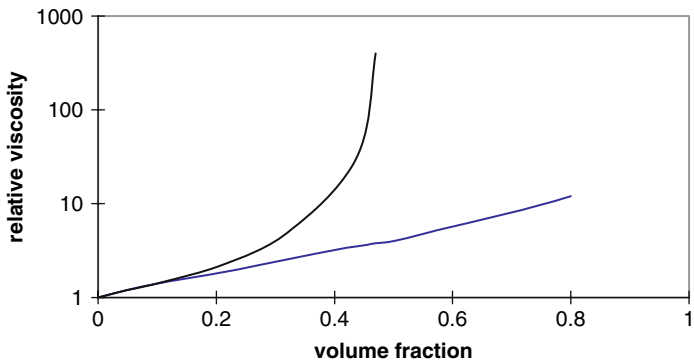


FIGURE 9.8 *Red blood cells* (left: *showing biconcave shape and* right: *red cells aggregating to form stacked cells, or rouleaux).*

of blood volume occupied by cells) as a function of the shear rate. The non-Newtonian feature is the variation of viscosity by large factors (note the log scales) with shear rate for blood with cells present. Blood plasma is a Newtonian fluid because its viscosity is independent of shear rate (lower curve). The red blood cells normally constitute about 50% of the blood volume, therefore it is clear that the non-Newtonian rheological properties of blood are primarily due to the red cells. In the bottom half of Figure 9.7 the low-shear viscosity of whole blood is shown as a function of the hematocrit. The strong dependence on the red cell content is also indicative of the large impact of the red cells on the rheological properties of blood. Red blood cells are disks that are biconcave (thinner in the middle than at the edges), are about 8 μm in diameter and have a tendency to stack together like coins, into aggregates called *rouleaux* (Figure 9.8). The extent of aggregation is strongly dependent on the shear rate; the aggregates will break up as the shear rate is increased, qualitatively explaining the decrease in viscosity at increasing shear rates shown in the top of Figure 9.7.

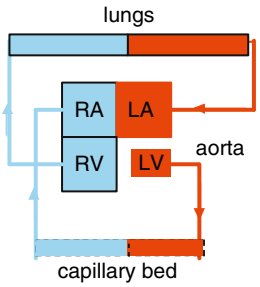
Blood is remarkably fluid. A 50% (by volume) suspension of small rigid spheres will be a solid, unable to flow at all, whereas blood is extremely fluid even at elevated hematocrits (Figure 9.9). This fluidity is due to the special properties of the red blood cells, particularly their membrane elastic properties and shape, which permit tremendous deformation of the red cells to allow flow. In many small blood vessels, the capillary diameters are on the order of the red cell diameter or even smaller and without great flexibility of the red cells, flow would be blocked. Diseased red cells, such as deformed cells in sickle cell anemia that lose their elastic properties, will clog small blood vessels. In the next section we take up the human circulatory system, including the heart, and expand on the flow properties of blood.

FIGURE 9.9 *Viscosity, relative to water, for human blood* (lower curve) *and a suspension of rigid plastic spheres* (upper curve) *as a function of the volume fraction occupied by particles. At volume fractions of about 50% or higher a suspension of plastic spheres behaves as a solid.*



3. THE HUMAN CIRCULATORY SYSTEM

In Western culture, the concept of blood circulation was established surprisingly late, in the 1600s, by William Harvey. The human circulatory system consists of a pump (the heart) and a complex branched distribution of "smart" delivery tubes that carry oxygen and nutrients



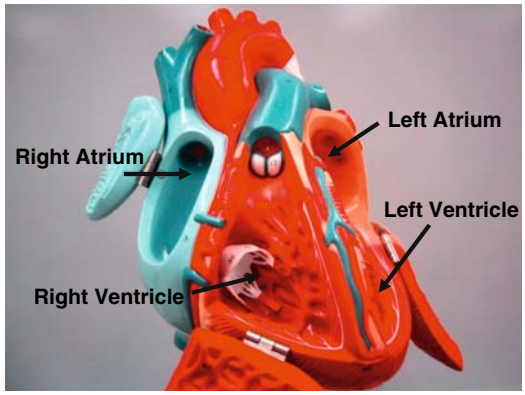


FIGURE 9.10 *(a) Schematic diagram of the heart and the flow of blood.* Red *indicates oxygenated blood and* blue *deoxygenated in this simplified scheme. (b) Schematic open model of the heart.*

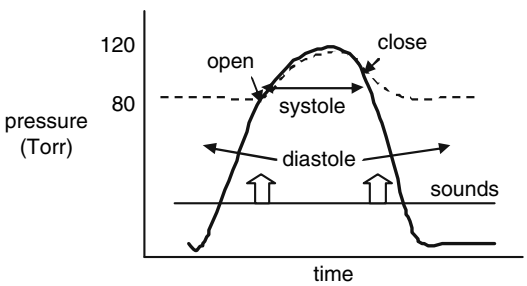
to, and remove waste products from, the body. Each side of the heart receives blood at low pressure and pumps this blood out at high pressure. In schematic form, shown in Figure 9.10a, oxygenated blood is pumped out of the left ventricle of the heart, through the aortic valve and the aorta to a branched network of arteries, smaller arterioles, and finally to the capillary beds throughout the body in which the exchange of gases and dissolved molecules with the body tissue occurs. Blood is collected from the capillary beds by the venules, which feed into the veins, all of which merge with either the superior (from above the heart) or inferior vena cava (from below the heart) or the coronary sinus (from the blood supply for the heart muscle itself) to return the blood to the right atrium. Thus the left ventricle and the right atrium of the heart together form the outlet and inlet of a pump that supplies nearly the entire body with blood.

A second parallel pump in the heart sends blood that has arrived from the right atrium through the tricuspid valve to the right ventricle, through the pulmonary arteries to the lungs where an exchange of gases occurs. The reoxygenated blood returns to the left atrium through the pulmonary veins where it enters the left ventricle through the bicuspid (mitral) valve to complete its cycle of flow. Thus, the right ventricle and left atrium are the outlet and inlet for a second pump of the heart. In the healthy mammalian heart the chambers of the left side of the heart are completely separated from those of the right after birth, and there is no mixing of oxygenated and deoxygenated blood in the heart. Despite this separation, the two sides are part of a single anatomical organ, and the heartbeat is coordinated by a single clump of cells, the pacemaker region. A schematic of the heart is shown in Figure 9.10b. In the rest of this section we consider several aspects of the circulatory system that relate to our previous discussions in this chapter. We return later (Chapter 15) to consider the electrical aspects of the heart, including the electrocardiogram (EKG).

The heart, about the size of an adult fist, pumps about 80 cm^3 of blood in each of the 70 beats/minute in a typical resting adult, so that about 5.5 L of blood are pumped throughout the body each minute. Because the total volume of blood in an adult is 5–6 L, we conclude that it takes just about a minute for blood to make a complete loop through the circulatory system. The total volume of blood is actually in dynamic equilibrium because fluids leave the blood vessels to exchange with tissue and to be filtered in the kidneys (discussed in Chapter 12). Figure 9.11 shows some events during a single cardiac cycle, divided into the systole, or contraction, phase and the diastole, or relaxation, phase. Note that the left and right ventricles contract together, as do the atriums. During systole, the ventricular pressure rises rapidly, after closure of the tricuspid or mitral valve, as the blood volume in the ventricle increases. When the aortic valve opens, the aortic pressure rises from its resting value of about 80 mm Hg to about 120 mm Hg. It is this pressure that is measured with a sphygmomanometer.

Figure 9.11 also shows the times at which valves open and close and those at which the heart sounds are most clear. The pulmonary artery

FIGURE 9.11 *A single cardiac cycle showing the left ventricular pressure* (bold) *and aortic pressure* (dashed) *as functions of time. Also indicated are the times at which valves open and close, when heart sounds are most clear, and the period of systole and diastole.*



pressure rises during the contraction of the right ventricle, but to a lesser extent than in the aorta; the peak pressure in the aorta is about six times that of the pulmonary artery. The greater pressure generated by the left ventricle is a result of thicker layers of muscle surrounding it compared to the right ventricle. Pressures in the atria are close to zero and only fluctuate a little during the cardiac cycle.

As an example of applying some of the fluid dynamics we have learned, we can make an estimate of the power developed by the heart in pumping blood, where power here is the time rate of transfer of energy to the blood. The heart supplies both the pressure and kinetic energy of the blood as it leaves the heart and enters the aorta. If we multiply Bernoulli's equation for constant height (see Equation (8.12)) by the volume flow rate Q, we obtain an expression for the power supplied by the heart as

$$Power = P_{\text{ave}}\, Q + \frac{1}{2}\rho v^2 Q,$$

where P_{ave} is the mean blood pressure in the aorta, ρ is the density of blood, and v is the average blood velocity in the aorta. If we take the mean aortic pressure to be 100 mm Hg when a person is at rest, then the PQ contribution of the left ventricle in one heartbeat is simply the product of the mean pressure and the volume change, 80 cm^3, resulting in a value of (100 mm Hg) $\times$ (1.01×10^5 Pa/760 mm Hg) $\times$ (80×10^{-6} m^3) = 1.06 J/heartbeat. Assuming 70 heartbeats per minute (or 1.2 beats/s) this translates into an average power of about 1.3 W. Because the pressure in the right ventricle is about 1/6 that of the left ventricle and the volume flow rate is the same, the PQ power contribution of the right ventricle is an additional 0.2 W. The kinetic energy term contributes a small additional amount of about 0.3 W when a person is at rest, so that the total power supplied by the heart is about 1.8 W. To find this kinetic energy contribution we use the fact that $Q = Av$, or $v = Q/A$, so that the kinetic energy term is $\frac{1}{2}\rho v^2 Q = \frac{1}{2}\rho(Q^3/A^2)$, proportional to Q^3.

When someone engages in very strenuous exercise, the flow rate of blood can reach 35 L/min (nearly 7 times the resting rate). In this case, assuming the mean pressure does not change significantly, the PQ power increases by a factor of 7 to about 10 W, and the kinetic energy power delivered to the blood rises dramatically to (0.3 W) (7^3) $\cong$ 100 W, because of its third-order dependence on Q. Where does this power go? Just ask someone doing exercise and they will tell you how hot they get and the amount of sweating they do in an attempt to cool off.

The key to the heart's success in maintaining pressure differentials in order to drive blood throughout the body is the four heart valves. Heart valves are crucial for the proper functioning of the heart and a number of heart diseases are traceable to defective valves. Perhaps surprisingly, the heart valves (and those of the veins mentioned below) are not controlled actively, but open and close passively in response to hydrodynamic forces. Consider the mitral valve, located between the left atrium and ventricle, shown schematically in Figure 9.12. In the diastole, when the pressure in the atrium exceeds that in the ventricle, the two thin membranes of the valve are pushed open and blood enters the ventricle. As the blood pours into the ventricle it strikes the ventricular walls and the flow breaks up into eddies, or vortices, that provide a back-pressure on the valve membranes, forcing them closed when the ventricular pressure exceeds the atrial pressure. A set of small muscles prevents the membranes from opening to allow backflow into the atrium; when properly functioning, the mitral valves prevent any blood from re-entering the atrium. Some types of heart murmurs are due to malfunctioning heart valves that allow backflow, producing characteristic sounds. Heart valves used in an artificial heart also make use of the same principles to provide passive control, rather than direct active control of the opening and closing of valves.

The cyclic variation in the aortic (and pulmonary arterial) pressure is the driving force producing blood flow throughout the body (and the lungs). We have seen from the continuity equation that the flow velocity in a tube is inversely

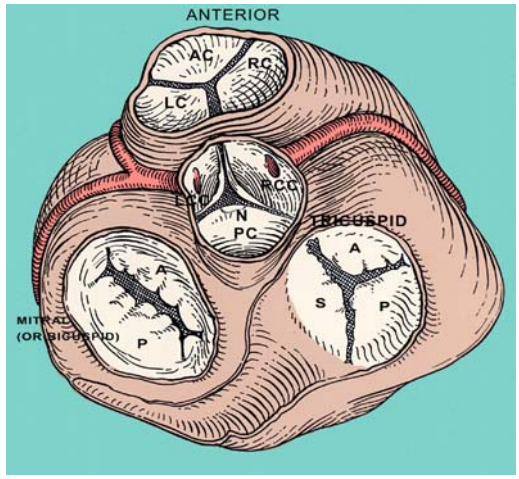


FIGURE 9.12 (left two panels) *The closure of the mitral valve by the backflow of blood in the left ventricle.* (right panel) *Schematic of the heart valves.*

proportional to the cross-sectional area, in order to conserve mass. In the circulatory system, one large artery (typical inside diameter 1 cm) divides into many arterioles (typical diameter 5 μm), each of which divides into many capillaries (typical diameters 0.6 μm). The capillaries have the smallest diameter therefore we might expect blood to flow fastest in these vessels. Although the capillaries do have the smallest cross-section, the total cross-sectional area of the estimated five billion capillaries is about five times that of the arterioles. Blood velocity in a capillary is therefore slower than in any other blood vessel, only about 0.07 cm/s. Capillary diameters are comparable to the dimensions of a red blood cell and so the flow of blood through capillaries, known as *bolus flow*, is quite special. As shown in Figure 9.13, to promote the flow of the red cells and the exchange of gases and chemicals across the vessel walls, the elastic red cells trap blood plasma between themselves that flows in eddies.

In some regions, blood flow from the arterioles can bypass the capillaries and flow directly into venules through an *arteriovenous (AV) shunt*. These shunts are able to regulate the flow of blood in order to control, for example, the extent of body cooling through blood flow in the skin. During exercise, as metabolism is increased, or when the external temperature is high, excess heat must be removed by evaporation and the capillaries near the skin surface are dilated by decreasing the AV shunt flow. Similarly in cold weather, the AV shunt is opened to decrease blood flow near the skin surface in order to reduce heat loss from the body. Another control mechanism outside the heart is *vasoconstriction*, a reflex process of reducing the diameters of blood vessels to increase flow rates in the case of blood loss or shock.

Blood flow in the larger arteries is known as *pulsatile flow*. As the ventricles pump blood into the major arteries, the blood cannot flow into the capillaries fast enough and so the arteries swell in diameter because the walls are elastic. As the pressure in the artery drops during diastole, the energy stored in the elastic vessel walls tends to smooth out pressure variations and this becomes more and more the case farther downstream from the aorta. This same elastic expansion of blood vessels can be felt as the pulse measured at one's wrist. By the time the blood leaves the capillary bed, the pressure in the veins is quite low. To help the return flow, larger veins, particularly in the limbs, have one-way valves along them. Excess fluid pressure in the feet, due to the extra pressure head, can sometimes result in fluid buildup and swelling (edema), especially without movement of the feet in order to promote blood flow in the venules and veins.

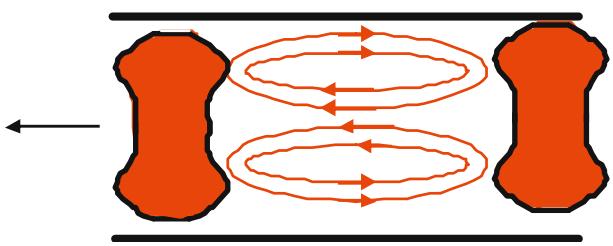


FIGURE 9.13 *Bolus flow of red blood cells moving to the left in a capillary, showing the eddy flow between cells.*

4. SURFACE TENSION AND CAPILLARITY

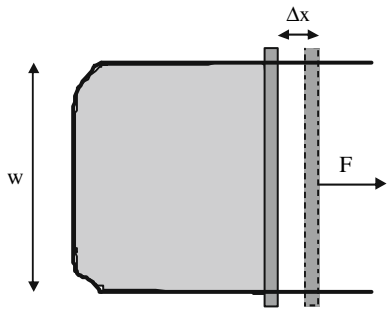


FIGURE 9.14 *A film of liquid is stretched by moving the crossbar a distance Δx. We use this to calculate the surface tension.*

The surface of a fluid represents a boundary that exhibits many special properties worthy of our attention. A thin layer of surface fluid in contact with air feels an excess attractive intermolecular force over the local interactions within the bulk fluid. The net force pulling the surface layer into the bulk fluid gives rise to a slightly greater density near the surface. Molecules that move into the surface layer have a higher energy than those in the bulk because there are fewer bonds to neighboring molecules and therefore work must have been done on them to move them to the surface. The measure of this extra energy is the surface energy per unit area γ, given in J/m^2, which depends on the particular fluids involved at the boundary (e.g., water and air). For pure water in air the surface energy density is unusually high, $\gamma = 0.073$ J/m^2.

Associated with the increase in energy in the surface layer of fluid is a surface tension. Consider the device shown in Figure 9.14 on which a liquid film, such as a soap film, is formed in air. In order to increase the surface area by sliding the crossbar a distance Δx, increasing the surface area by $(w\Delta x)$, a force F is needed. The work done by this force will equal the extra surface energy, therefore we find

$$F\Delta x = 2\gamma\,(w\Delta x), \tag{9.8}$$

where the factor 2 enters because there are two surfaces of the fluid exposed to air. From this we find another expression for γ,

$$\gamma = \frac{F}{2w}, \tag{9.9}$$

so that γ, already seen to be the surface energy density, is also a force per unit length, with units of N/m, and is also known as the *surface tension*. A force per unit length is appropriate for a fluid, rather than a force per unit area, or stress, used for a solid, because the fluid surface layer is imagined to be infinitesimally thin. The surface of the liquid is sometimes said to behave like a skin or rubber sheet. This is because the surface can support small insects such as water striders skimming the surface of a pond (see Figure 9.15 and Problem 19). However, unlike a rubber sheet, when a fluid

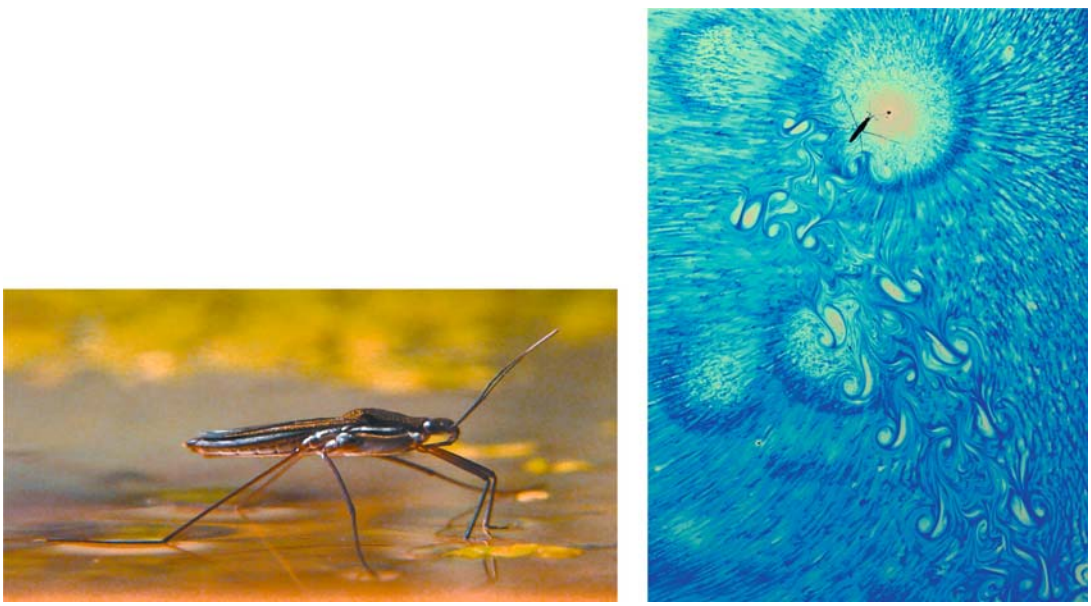


FIGURE 9.15 *A water strider glides over the water using surface tension to support itself. On the right, a dye was added to float on the water surface and when illuminated from below reveals the hydrodynamics of the strider's motion. Note that the strider is light-seeking.*

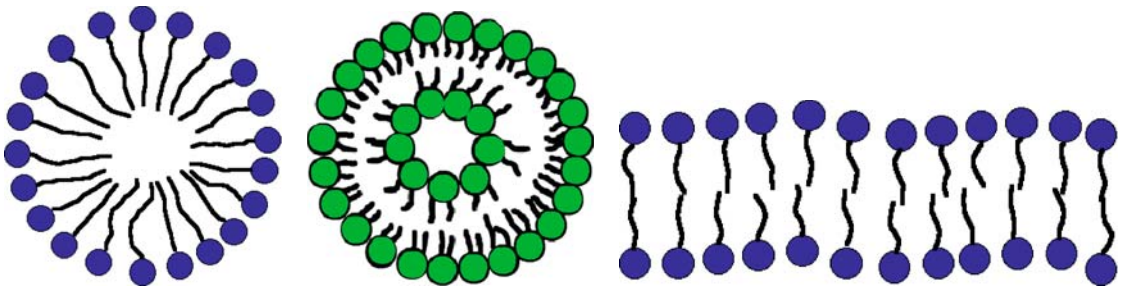


FIGURE 9.16 *Micelle* (left)*, vesicle* (center)*, and planar bilayer* (right) *all composed of lipids.*

surface is extended, additional molecules are added to the surface from the bulk fluid, and so the analogy is of limited use.

If a drop of liquid is formed in air, as from a dripping faucet, the intermolecular interactions will tend to minimize the surface energy by minimizing the surface area. Because a sphere has the minimum surface area for a given volume, in the absence of other forces liquid drops are spherical. For small drops the surface tension is much larger than gravitational forces and the drops are indeed spherical. Under "weightless" conditions, such as in a space shuttle flight, even large liquid drops are observed to be spherical. In the presence of gravity larger drops tend to get elongated vertically. We use this liquid drop idea to model the nuclei of atoms in Chapter 26 to understand the process of nuclear fission.

An important related example is the formation of micelles or vesicles of lipids in water. Recall that lipids have hydrocarbon tails that are hydrophobic and polar head groups. When mixed in water at low concentrations, lipids tend to form micelles, or spherical balls with the polar groups facing water on the outside and the hydrocarbon tails buried inside (Figure 9.16). At higher lipid concentrations, the lipids form vesicles or spherical lipid bilayers with water both inside and outside, as shown in the center figure. These are similar to cell membranes, although cell membranes also have many associated proteins bound to the lipids. Surface tension is an important factor in the overall structure of both vesicles and micelles.

In our bodies, the largest surface area in contact with air is the internal surface of the lungs. The total surface area in the lungs of an adult is tremendous, roughly 100 m^2, or the size of a large room. This large surface area is possible because of a branched network of small sacs or alveoli. Figure 9.17 shows an idealized section of an alveolus taken to be spherical. The air pressure inside, P_i, is normally greater than the pressure in the pleural cavity outside, P_o, and this net pressure difference is balanced by the surface tension in the wall of the alveolus which we treat as an idealized elastic membrane, like a small balloon. We can relate the surface tension in the alveolus to the pressure difference by imagining that we divide the alveolus into two hemispheres and balance the forces acting on each separate hemisphere (see Figure 9.17). The net tension force pulling to the left on the right hemisphere in the figure along the circular edge of the alveolus membrane is $2\pi r\gamma$. This force must be balanced by the net pressure force directed toward the right, which can be shown to equal the pressure difference $P_i - P_o$ multiplied by the projected area πr^2 (see boxed calculation). The balance of forces $2\pi r\gamma = (P_i - P_o)\pi r^2$ then implies

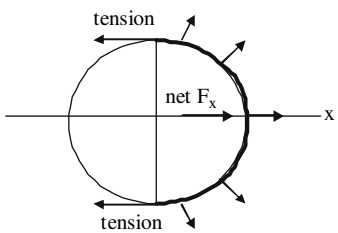


FIGURE 9.17 *Right hemisphere is in equilibrium under the tension forces from the left hemisphere and the pressure difference (radial forces) resulting in a net pressure force along the x-axis.*

$$P_i - P_o = \Delta P = 2\frac{\gamma}{r}, \tag{9.10}$$

which is known as *Laplace's law* for a spherical membrane. This relation also holds for a spherical drop of liquid.

In the lungs, both the radius of the alveoli and the pressure difference vary during breathing, in part due to motion of the diaphragm (with an area of 0.05 m^2). If an alveolus were to collapse to a diameter of about 0.2 mm, from Laplace's law using γ for a water/air interface, the pressure difference would predict that a force ($\Delta P A_{\text{diaphragm}} = (2 \cdot 0.073/0.1 \times 10^{-3}) \cdot 0.05$) of about 70 N would be required to breathe. This is more than the weight of a newborn and is an impossibly large force for the diaphragm to exert. To solve this problem we have a surfactant, a lipid–protein complex, present in the lungs. The addition of small quantities of impurities can dramatically reduce the surface tension at a surface. In this case surfactants reduce the surface tension by about a factor of 15, thus greatly reducing the needed force. Premature infants with hyaline membrane disease do not manufacture this surfactant and are prone to developing collapsed lungs. One treatment of this disease involves spraying a surfactant into the lungs to temporarily support breathing.

Suppose that a drop of liquid is placed on a plane substrate surface. Molecules on the surface of the drop have two competing forces, those of *cohesion* tending to keep the drop spherical, and those of *adhesion* to the substrate surface that will tend to spread the liquid on the substrate. The nature of the two materials involved will determine the *contact angle* θ, shown in Figure 9.18. Liquids with contact angles between 0 and 90° are said to wet the substrate surface. Pure water wets ultraclean glass at $\theta \approx 0$ so that the drop spreads freely on the glass, whereas on typical glass $\theta \approx 30°$. For angles larger than 90°, for example, mercury on glass where the mercury beads up, the liquid does not wet the substrate at all. Wetting characteristics are important in our lives; we use water repellents so that water beads up and will not wet surfaces, and we add wetting agents, generally molecules with hydrophobic and hydrophilic portions, to promote better contact of a liquid with a solid surface.

In biology, a most important consequence of wetting is capillary action, the rise of liquids that wet the surface of a capillary. Figure 9.19 shows a glass capillary immersed in a container of water, in which the water rises and has its characteristic meniscus and a similar tube immersed in a container of mercury showing the situation for a nonwetting liquid. We can calculate the height rise h of the water in the capillary with radius r, by considering the surface tension that supports the weight of the water column. Because the water wets the glass at a contact angle of θ, the vertical component of the surface tension is (see Figure 9.19 right)

$$F = 2\pi r \gamma \cos\theta, \tag{9.11}$$

where the factor $2\pi r$ is the contact perimeter and cos θ accounts for the vertical component.

With the weight of the water column given by $\rho \pi r^2 h g$, equating these two forces yields a column height of

$$h = \frac{2\gamma \cos\theta}{\rho g r}. \tag{9.12}$$

Equation (9.12) predicts that the smaller the radius of the tube, the higher the column of fluid can rise by capillary action. It also predicts the behavior of mercury in a glass capillary because cos θ is negative and not only will the meniscus be inverted, but surprisingly, the column of mercury will fall below its level in the large container, as shown in the figure.

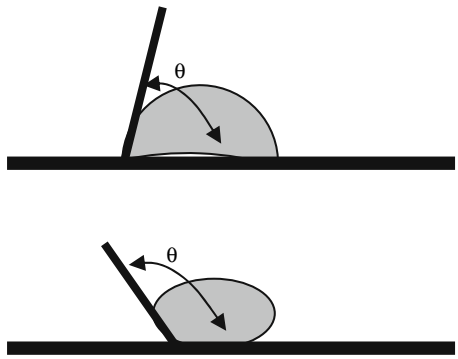
Clearly water transport in plants and trees (sap is mostly water) is an important application of capillary action, although in this case the upper end of

To find the net force on the right hemisphere due to the pressure difference ΔP in Figure 9.17, we need to add up the contributions from the normal force at each portion of the hemisphere. By symmetry, it should be clear that the direction of the net force will be to the right because for every area ΔA in the right hemisphere with normal force vertical component F_y or F_z, there will be a symmetrically located area with a component of $-F_y$ or $-F_z$. Using spherical coordinates, the x-component of force due to the pressure at ΔA is $F_x = P \cos\theta \, \Delta A$, where ΔA can be written as $(r \sin\theta \, d\phi)(r \, d\theta)$. Integrating to find the total force in the x-direction gives

$$F_x = \Delta P r^2 \int_0^{\pi} d\phi \int_0^{\pi} \sin\theta \cos\theta \, d\theta$$
$$= \Delta P \pi r^2,$$

which was used to find Equation (9.10). Note that πr^2 is the projected area along the x-axis.

FIGURE 9.18 *The definition of the contact angle for a drop on a surface; the top drop wets the surface and the bottom drop beads up on the surface, not wetting it.*



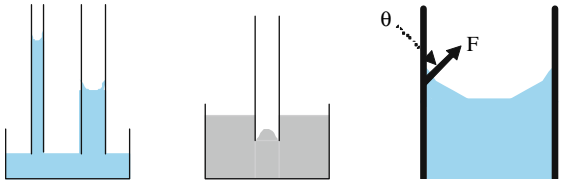


FIGURE 9.19 (left) *Capillaries immersed in water showing the meniscus and the fact that water rises higher in a narrower tube;* (center) *capillary immersed in mercury showing the inverted meniscus and the lower level in the capillary than in the surrounding container;* (right) *detail showing surface tension force calculated in Equation (9.11).*

the vascular system is not open to the atmosphere. Typical pore radii in the xylem of trees is 20 μm, so that the maximum height rise of water in such a capillary should be about 75 cm, using a contact angle of 0°, based on Equation (9.12). But how does water rise higher in trees, some of which are over 100 m tall? In the leaves of trees the interstitial pathways for water flow are believed to be on the order of 5 nm. As long as water is able to reach the leaves, it will be supported by the capillary action in the leaves, because with 5 nm pores Equation (9.12) yields a height of nearly 3 km, much taller than any tree. It is believed that as a tree grows, as long as the column of water is maintained, the capillary action in the leaves is sufficient to support the column of water. The flow of water is then regulated mainly by evaporation from the leaves, known as transpiration, effectively producing a "negative pressure" that pulls water up from the soil. We know that even a vacuum cannot pull water up to a height greater than 30 m; hence the term negative pressure, which is able to pull water to greater heights based on capillary action. If a tree has a portion of its xylem damaged so that the water column is interrupted, then beyond a height of 75 cm there is no mechanism to restore the flow of water.

CHAPTER SUMMARY

Viscosity η can be defined for a Newtonian fluid as the proportionality constant between the stress (F/A) and the rate of strain ($\Delta v/\Delta y$) (where the geometry is that of Figure 9.1):

$$\frac{F}{A} = \eta \frac{\Delta v}{\Delta y}, \tag{9.1}$$

For a Newtonian fluid flowing in a cylindrical tube, the flow rate Q is given by Poiseuille's law,

$$Q = \frac{\pi P r^4}{8\eta L}, \tag{9.2}$$

where P is the pressure difference across the tube, and r and L are the tube radius and length, respectively.

The viscosity of a suspension of spherical particles η_s, increases from the solvent viscosity, η_o, as the volume fraction Φ increases according to

$$\eta_s = \eta_o (1 + 2.5\Phi). \tag{9.7}$$

Blood is a complex fluid that exhibits non-Newtonian flow and has rheological properties that are very dependent on the hematocrit, the percent of blood volume occupied by cells (mostly red cells). The human circulatory system basically functions as two coupled pumps that send blood to the lungs for oxygenation and release of carbon dioxide, and to the capillary beds for distribution of oxygen and nutrients and collection of cellular waste products.

The surface tension γ at the boundary surface between two fluids (a liquid and air, e.g.) is given by the excess surface energy per unit surface area, or equivalently by the force per unit length (in the geometry of Figure 9.14),

$$\gamma = \frac{F}{2w}. \tag{9.9}$$

The pressure difference across a spherical membrane or drop of fluid of radius r is given by Laplace's law,

$$P_i - P_o = \Delta P = 2\frac{\gamma}{r}. \tag{9.10}$$

Capillary action causes a column of fluid of density ρ to rise a distance h

$$h = \frac{2\gamma \cos\theta}{\rho g r}, \tag{9.12}$$

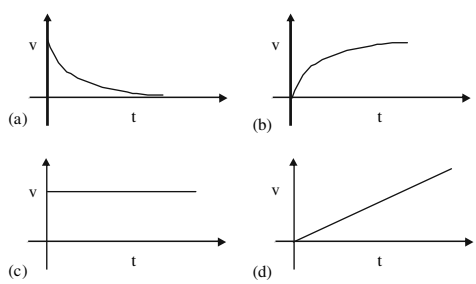
where γ and θ are the surface tension and contact angle that the fluid wets the capillary surface.

QUESTIONS

1. Give some examples of fluids with appreciable viscosity and try to put them in order of increasing viscosity.
2. Give an argument as to why the viscosity of normal fluids should generally decrease with increasing temperature.
3. Give some examples of laminar and turbulent flow of fluids.
4. Assuming Poiseuille's law applies, what would be the change in volume flow rate through a tube when the radius is halved? When the length is quadrupled? When the viscosity of the liquid is doubled? When the pressure head is doubled?
5. Explain how a simple timing measurement can determine the viscosity of a liquid in a capillary viscometer. What complications can you imagine would arise in the measurement of the viscosity of suspension of long polymers in a high-shear capillary tube?
6. Check that the Reynolds number is dimensionless.
7. Cigarette smokers generally have higher hematocrits than nonsmokers. This is probably due to the decreased oxygen efficiency of the red blood cells from the inhaled carbon monoxide in cigarette smoke (about 250 cm^3 per pack). What is the effect of the higher hematocrit on the velocity of blood flow?
8. Explain why an aneurysm in an artery leads to a locally elevated blood pressure.
9. How can a plaque deposit on an artery or arteriole wall lead to a decreased local blood pressure and the collapse of that vessel?
10. Hold your hands at your sides and observe a swollen vein in your arm or hand. Then raise your arm over your head. The vein will "disappear" as it shrinks in diameter. Why?
11. Describe, in words, the path of blood flow throughout the human circulatory system.
12. What is the difference between pulsatile and bolus flow of blood?
13. Why are artificial heart valves designed to be passively rather than actively controlled?
14. Insects that walk on water secrete antiwetting liquids that coat their legs. How does this help them?
15. Why don't the lungs consist of two large sacs rather than huge numbers of small alveoli? Examine Laplace's law for the answer.
16. Discuss how the competition between cohesion and adhesion determines the wetting of a material by a liquid. Adhesive tape (including Post-it type paper) uses this idea as well as large numbers of tiny bubbles that create vacuum suction attachments.
17. Explain the function of surfactants in our lungs.
18. What factors control how high a fluid will rise in a narrow capillary tube? Which ones depend on the fluid, the tube material, or the geometry alone?
19. Describe in words the source of the "negative pressure" that allows water to rise so high in plants and trees.

MULTIPLE CHOICE QUESTIONS

1. The SI units for viscosity are (a) kg/(m-s), (b) kg-m^2/s, (c) kg-m/s, (d) kg-s/m.
2. Which has the greatest effect on the flow of fluid through a pipe? That is, if you made a 10% change in each of the quantities below, which would cause the greatest change in the flow rate? (a) the fluid viscosity, (b) the pressure difference, (c) the radius of the pipe, (d) the length of the pipe.
3. In the flow of water through a capillary tube, if the diameter of the tube is tripled with no other changes, the flow rate will (a) increase by a factor of 9, (b) increase by a factor of 27, (c) increase by a factor of 16, (d) increase by a factor of 81.
4. Which of the following is a false statement about the flow of a liquid in a thin vertical tube?
   (a) The velocity is fastest at the center of the tube, (b) if the tube radius is doubled the flow rate will increase by a factor of 16, (c) the ratio of the flow times for two liquids depends only on the ratio of their viscosities, (d) the presence of suspended particles in the liquid decreases the flow rate.
5. For a given solution of particles in a solvent, the characteristic velocity at which there is a transition from laminar to turbulent flow is (a) proportional to the size of the particles, (b) is proportional to the density of the fluid, (c) is proportional to the viscosity of the fluid, (d) is independent of the size of the particles.
6. The flow of blood through a capillary requires a higher pressure where the blood enters the capillary than where it leaves. That is most directly related to (a) $F/A = \eta \Delta v/\Delta y$, (b) $\Delta P = \rho g \Delta y$, (c) $\Delta(P + \rho g y + \rho v^2/2) = 0$, (d) $\Delta(\rho A v) = 0$.
7. An object is dropped from rest at $t = 0$ into a viscous fluid. Which of the following best describes the object's speed as a function of time?

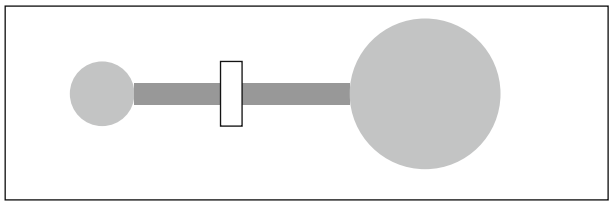


8. The incremental viscosity of a dilute solution of identical particles over that of the solvent depends on all but which of the following? (a) Particle size, (b) particle concentration, (c) particle shape, (d) the solvent viscosity.
9. Heart valves close in response to (a) sets of muscles, (b) hydrodynamic forces, (c) vasoconstriction, (d) surface tension.
10. Blood is called a complex fluid because (a) it has many different components, (b) it has a high viscosity, (c) its viscosity depends on shear rate, (d) blood plasma is a Newtonian fluid.
11. The fundamental reason that red blood cells can flow through small diameter capillaries at high concentrations whereas plastic spheres of the same size form a stiff "solid" is (a) the unique disk shape of the red blood cell, (b) the bolus flow of the red cells, (c) the flexibility of the red cell, (d) the vasoconstriction of the capillaries.
12. Heart sounds heard in a stethoscope are due to (a) turbulent flow between heart chambers, (b) pulsatile flow in the aorta, (c) laminar flow through the heart chambers, (d) the AV shunt.
13. The shape of a droplet of liquid on a surface is due to a combination of (a) pressure and cohesion, (b) adhesion and cohesion, (c) capillary action and adhesion, (d) capillary action and pressure.
14. In an open tube, water can only be suctioned to rise about 10 m. In a 20 mm radius tube water will only rise about 75 cm by capillary action. How can water rise to the top of trees, sometimes over 100 m tall? (a) By cohesive forces, (b) by adhesive forces, (c) by Laplace's law, (d) by transpiration generating negative pressure.
15. Teflon does not wet with water at all. The contact angle for water on teflon is (a) 0°, (b) 90°, (c) 180°, (d) 270°.

PROBLEMS

1. Assuming that the cream in a chocolate cream sandwich cookie behaves as a Newtonian fluid of 10 Pa-s viscosity (probably not a great assumption), find the force needed to slide one of the chocolate wafers off the cream at a speed of 2 mm/s if it is a 5 cm diameter disk and the cream filling is 2 mm thick.
2. Suppose that there is a partial blockage of the aorta, which normally pumps about 5 L of blood per minute. If the diameter of the aorta is reduced by 30%, find the average flow rate through the diseased aorta. What increase in blood pressure would be needed to obtain the normal flow rate? (Assume that Poiseuille's law applies.)
3. What pressure is needed to deliver saline solution through a hypodermic needle with 0.3 mm inner diameter and 2 cm length at a rate of 10 cm^3/min. Assume the saline has the same physical properties as pure water. Also express the pressure in units of cm of H_2O.
4. In giving an intravenous (IV) of saline solution to a patient, the storage bag is placed 1 m above the patient's arm and attached to a hypodermic needle. If the flow rate out of the needle just before it is inserted into the patient's arm is 50 cm^3/min, when it is inserted into a vein with a blood pressure of 20 torr (gauge pressure), how long will it take to give 1 L of saline? Assume the saline has the same physical properties as pure water.
5. Find the Reynolds number for blood pumped into the 1 inch diameter aorta from the heart, using this distance as the characteristic length involved. Take the volume of blood pumped each of 72 times per minute to be 70 cm^3. Is the blood flow turbulent? In fact, because of the pulsatile nature of the heart pump, blood flows into the aorta as a bolus or plug with relatively little turbulence.
6. An intravenous blood plasma drip enters a vein in the patient's arm from a bag raised a height h above the vein. If the diameter of the 5 cm long needle is 0.5 mm, find the height h that results in a 5 cm^3/min flow rate. (Assume the blood pressure in the arm is 18 torr.)
7. A salt solution of specific gravity 1.018 has its efflux time in a capillary viscometer measured to be 122.5 s compared to a time for distilled water of 116.4 s. What is the viscosity of the salt solution?
8. The viscosity of blood plasma is to be measured in a capillary viscometer at 37°C. Using water as a standard, the efflux time is found to be 95 s. Predict the efflux time measured with blood plasma. Use Tables 8.1 and 9.1 and assume that the ratio of the densities of the two fluids is temperature-independent. Suppose this viscometer were used to try to measure the viscosity of whole blood. Knowing that the shear forces are fairly high, would your result be higher or lower than the low-shear value?
9. Plastic microspheres with a 5 μm diameter are added to water to make up a suspension. If there are 10^9 such spheres in a 1 cm^3 volume of water, what is the expected viscosity of the suspension? If the same numbers of 1 μm diameter microspheres are used in the same volume of water, find the expected viscosity of this suspension.
10. Viscosity standard solutions are to be made up from distilled water and 10 μm diameter plastic microspheres. If solutions of 1.05 cP, 1.1 cP, and 1.4 cP are desired, starting from a large volume of stock solution of 10^9 spheres per cm^3, give a recipe to make up 100 ml of each of the three desired solutions.
11. A long fine glass capillary pipette with an inner diameter of 0.1 mm is immersed in distilled water. How high will the water rise if the glass is extremely clean? Repeat if the contact angle is 30°. Note that it is well known that the meniscus is taller for glass when it is extremely clean.
12. A spherical balloon is filled with air to a radius of 10 cm. Find its surface tension assuming the pressure inside is 5 kPa.

13. Imagine two bubbles of air of different sizes attached via a tube with a valve as shown, all immersed in water. With the valve closed, which bubble is at the higher pressure? Show that unless both bubbles start with the same size, when the valve is opened surface tension will cause the smaller bubble to shrink and the larger one therefore to grow. This illustrates a potential problem in our lungs in inflating the alveoli, or sacs connected via bronchioles. If the alveoli were not all expanded at the same rate, in theory only the largest would form. The result of this would be a minimizing of the total surface area. Fortunately the fluid that coats our alveoli contains a surfactant that both tends to reduce the surface tension, making it easier to expand alveoli, as mentioned in the text, and also reducing the dependence of pressure on the radius so that not all alveoli need be the same size to inflate.



14. Suppose an alveolus of the lung, with a radius of 0.15 mm, is coated with pure water and devoid of surfactant. What pressure would be needed to keep the alveolus inflated? Treat the alveolus as a spherical air bubble. Note that the maximum pressure that can be reached normally is about 18 mm when inhaling maximally.

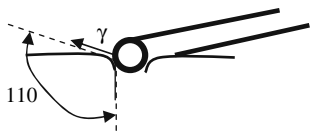
15. Suppose that in the previous problem the alveolus has surfactant present, reducing the surface tension to 0.03 N/m. How small can the alveolus collapse to, assuming it remains spherical, and still be inflatable by a strong inhalation with a maximal pressure of 18 mm Hg?

16. Laplace's law for cylindrical geometry is $\gamma = \Delta P r$, where r is the radius of the cylinder; note that this is a factor of 2 different from that for a sphere. Consider a cylindrical balloon that is partially inflated as shown. Because the balloon is in equilibrium, the pressure is uniform throughout and the surface tension, really the elastic tension of the balloon material, varies with the stretch of the surface. Rank order the surface tension, from high to low, for the labeled points on the balloon.

17. Using the previous problem, we can understand how thin-walled capillaries (thin to allow the exchange of gases) can withstand the blood pressure within them. Find the elastic tension (in N/m) in a 6 μm radius capillary with a blood pressure of 30 mm Hg.

18. Healthy young human arteries have a maximum elastic tension of about 500 N/m, a value that increases by more than a factor of two with age. Find the maximum pressure that such an artery can withstand before developing an aneurysm, or bulge often leading to rupture, and compute how many times greater this pressure is above the normal maximum systolic pressure of 120 mm Hg. An aneurysm, or bulging of an arterial wall, cannot occur in a heavy artery, but only in one with a weakened wall due to a connective tissue disorder.

19. Water striders are able to walk on the surface of water. This problem shows how they do it. Suppose that the insect's legs are nonwettable, so that the contact angle with water is 110° (assume this is from the vertical), and that the portion in contact with the water is cylindrical. If the insect has a mass of 0.01 g (and, remember, 6 legs) use Laplace's law for cylinders, $\gamma = \Delta P r$, to find the length of each leg that must be immersed in water to support the weight of the insect by the vertical component of the surface tension.



10 Waves and Resonance

Of all the types of waves we study, we are most familiar with water waves as seen in oceans, lakes, rivers, and bathtubs. We're also familiar with waves created by air currents through fields of grasses or wheat. In reality, we constantly experience waves of various types. Sound, light, radio, and other forms of electromagnetic radiation surround us every moment of our lives and although we do not directly "see" their waves, aside from visible light, these phenomena can all be understood in terms of waves. Furthermore, we show later that matter also behaves as a wave and that our current quantum physics picture of the world is intimately connected with a mathematical description known as the wave function. Waves are thus the key to our understanding of nature on a fundamental level.

In this chapter we first return to the type of motion known as simple harmonic motion that we used to describe a mass on a spring in Chapter 3. Here we extend our previous discussions to include the frictional loss of energy, known as damping, and the effects of a "driving force" used to sustain the motion. With the addition of energy by this external force comes the possibility of a resonance phenomenon in which the amplitude of oscillation can grow rapidly. This is an extremely important idea in physics that we will see often throughout the rest of our studies. We then introduce some fundamental concepts concerning waves and consider traveling waves along a string and along a coiled spring as mechanical examples of the two basic forms of waves, transverse and longitudinal. As waves travel along or through a medium, they meet and interact with boundaries or obstacles, and different interactions possible at a boundary are considered, including reflection and refraction. We also discuss one possible result from such boundary conditions, the creation of standing waves. These are important in such diverse areas as musical instruments, the human ear, and the basic functioning of a laser, all considered later in this book.

1. SIMPLE HARMONIC MOTION REVISITED: DAMPING AND RESONANCE

A linear restoring force is the basis of simple harmonic motion. Our example has been the spring force, $F = -kx$, first studied in Chapter 3. The characteristic of simple harmonic motion is the variation in oscillator position according to

$$x(t) = A\cos(\omega_0 t), \tag{10.1}$$

where ω_0 is the angular frequency that depends on the parameters of the particular type of simple harmonic oscillator. For example, in the case of a mass on a spring we have seen that the angular frequency is given by $\omega_0 = \sqrt{\frac{k}{m}}$. We have already introduced the definitions of the frequency, f, and period, T, which are related to the angular frequency in general by

$$f_0 = \frac{1}{T} = \frac{\omega_0}{2\pi}. \tag{10.2}$$

J. Newman, *Physics of the Life Sciences*, DOI: 10.1007/978-0-387-77259-2_10,

FIGURE 10.1 *Oscillating systems:* (from left) *pendulum at the Griffith Observatory, an automobile coil spring, and the Tacoma Narrows bridge, just before its collapse.*

The frequency f_0 is often called the *natural frequency* of oscillation of the isolated system because it is the frequency the system adopts if released and left unperturbed. Later in this section we consider cases when an external force, oscillating at a frequency different from the natural frequency, acts on the system (Figure 10.1). As we have also seen when we considered potential energy, the energy of a simple harmonic oscillator remains constant, exchanging periodically between kinetic and potential energy.

A second example of an oscillating system that can be modeled as undergoing simple harmonic motion is the so-called simple pendulum, consisting of a point mass suspended from a massless string or rod of length L. A true simple pendulum consists of a mass with dimensions small compared to L and a light string or rod. If the pendulum is made to oscillate in a plane, we can show that if the string makes a small ($<10°$) angle with the vertical that this angle will oscillate according to Equation (10.1) with x replaced by the angle A equal to the maximum angle, and ω_0 given by $\omega_0 = \sqrt{\frac{L}{g}}$. Thus, the motion of the simple pendulum is independent of its mass, depending only on its length.

Simple harmonic motion is an abstraction. All real oscillators lose energy over time due to frictional forces. This was first seen in the Chapter 3 section on viscoelasticity where we discussed models in which the elastic springs were combined with frictional dashpots to describe the viscous effects of the material. Let's now consider in more detail the effect of frictional forces on the simple harmonic motion of a mass on a spring. We model the frictional (damping) force as linearly dependent on the velocity of the mass. This is a good approximation when the damping forces are small. Then the net force on the mass is given by

$$F_{\text{net}} = -kx - bv, \tag{10.3}$$

where b is a frictional or damping constant.

What is the effect of this damping on the motion of the mass? If the damping is small we might guess correctly that the resulting motion would be an oscillation with slowly decreasing amplitude. The correct expression for the oscillator position with damping is

$$x(t) = (Ae^{-\frac{bt}{2m}})\cos(\omega_{damp}t), \tag{10.4}$$

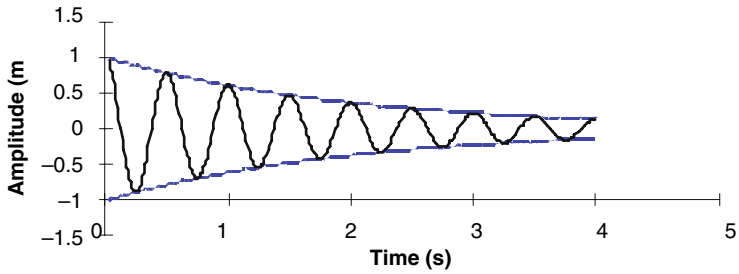


FIGURE 10.2 *Damped harmonic oscillations showing the exponentially decreasing envelope of the amplitude.*

where the angular frequency is a constant somewhat different than in the case of no damping and given by

$$\omega_{\text{damp}} = \sqrt{\frac{k}{m} - \frac{b^2}{4m^2}}. \tag{10.5}$$

Note that if $b = 0$ this expression reduces to the angular frequency in the absence of damping, as it must. The first term in parentheses in Equation (10.4) is an exponentially decreasing amplitude. Figure 10.2 shows a typical graph of Equation (10.4); the dashed lines are called the envelope of the equation and show the exponentially decreasing amplitude of oscillation.

The energy of a spring undergoing *undamped* simple harmonic motion is equal to the constant value $\frac{1}{2}kA^2$. The energy of the *damped* oscillator can be found by substituting the exponentially decreasing amplitude to find

$$E = \tfrac{1}{2}kA^2 e^{-\frac{bt}{m}}, \tag{10.6}$$

that itself decreases exponentially with time. Thus, once made to oscillate, a damped harmonic oscillator will maintain a fixed period of oscillation, given by $T = 2\pi/\omega_{\text{damp}}$, but will have an amplitude and energy that continuously decrease (Figure 10.3). The damped harmonic oscillator model can be used to describe many other systems in addition to springs. For example, molecules that interact with each other but lose

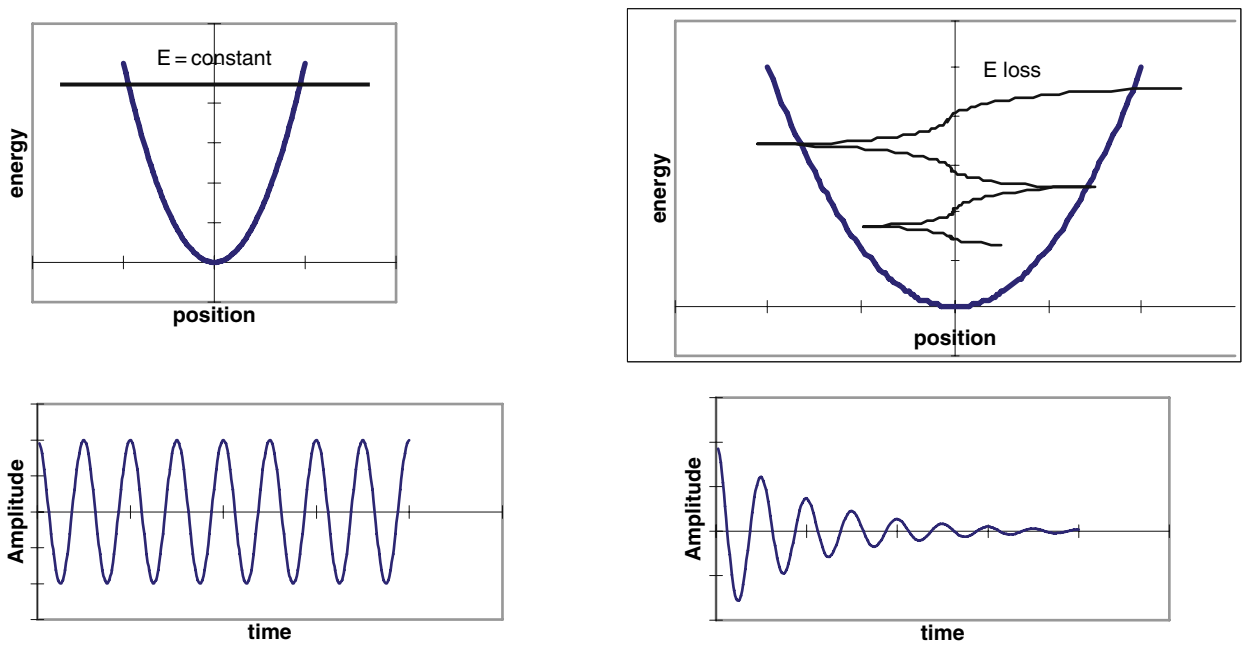


FIGURE 10.3 Left: *Undamped simple harmonic motion showing constant energy and amplitude;* Right: *Damped harmonic motion with decreasing energy and amplitude. The peculiar shape of the energy loss curve is due to the nonlinear dependence of position on time.*

energy via collisions or other mechanisms can also be modeled using spring and damping constants that can be related to the interaction parameters. Also a real pendulum with damping forces can be modeled in a parallel way.

Example 10.1 A 0.2 kg mass is attached to a spring with a spring constant of $k = 40$ N/m and a damping constant of $b = 0.02$ kg/s and allowed to come to equilibrium. If the spring is then stretched a distance of 10 cm and released from rest, find the following: (a) the initial energy; (b) the natural frequency; (c) the actual period of the motion; (d) the time for the amplitude to decrease to 5 cm, half of its initial value; and (e) the time for half the energy to be dissipated.

Solution: (a) The initial energy is equal to $\frac{1}{2}kA^2$ (this is also the $t = 0$ value of energy obtained from Equation (10.6)) and is therefore $E_i = 0.5(40)(.1)^2 = 0.2$ J. (b) The natural frequency is defined by Equation (10.2). Recalling that for a spring $\omega_0 = \sqrt{\frac{k}{m}}$, we have that

$$f_0 = \sqrt{\frac{40}{.2}}\ /2\pi = 2.25 \text{ Hz}.$$

(c) The actual period of the motion is $T = 2\pi/\omega$, where ω is the actual angular frequency of the oscillation, affected by the damping, and given by Equation (10.5). We have that $T = 2\pi/\sqrt{\frac{40}{.2} - \frac{0.02^2}{4 \cdot 0.2^2}} = 0.44$ s. This value is extremely close to the period in the absence of damping; the second term in the square root is negligible; in fact, in order for that term, $b^2/4m^2$, to make a 5% change in the period, b must as large as 0.2 kg/s.

(d) Because the amplitude decays exponentially, we can write from Equation (10.4) that $A(t) = A(0)e^{-bt/2m}$. Substituting we have $0.05 = 0.1\ e^{-0.02t/(2)(0.2)} = 0.1\ e^{-0.05t}$, or $0.5 = e^{-0.05t}$. We solve this equation by taking the natural logarithm of both sides of the equation: $\log 0.5 = \log(e^{-0.05t}) = -0.05\ t$, so that $t = -(\log 0.5)/0.05 = 13.9$ s.

(e) The time for half the energy to be dissipated is found in a similar way using Equation (10.6) in the form $E(t) = E(0)\ e^{-bt/m}$. Because we want the time for $E(t)/E(0) = 0.5$, we write $0.5 = e^{-bt/m}$ and again take the natural logarithm of both sides, to find $t = -\log(0.5)m/b = 6.9$ s, or half the time for the amplitude to drop to half its starting value, as expected from the factor of two difference in the exponents.

In practice, oscillators have their amplitude maintained by adding energy from the outside; for example, the pendulum on a grandfather clock maintains its amplitude of oscillation from the energy of a spring or a mechanical gear mechanism that requires winding. We can account for an external force F_{ext} by adding a term to Equation (10.3) so that the net force on the oscillator mass is now

$$F_{\text{net}} = -kx - bv + F_{\text{ext}}. \tag{10.7}$$

If the external force is sinusoidal, with a frequency f_{ext} known as the external *driving frequency* then, after sufficient time to reach a *steady state* in which the motion remains periodic, the oscillator position is given as

$$x(t) = A(\omega_0, \omega_{\text{ext}})\cos(\omega_{\text{ext}}\, t + \varphi), \tag{10.8}$$

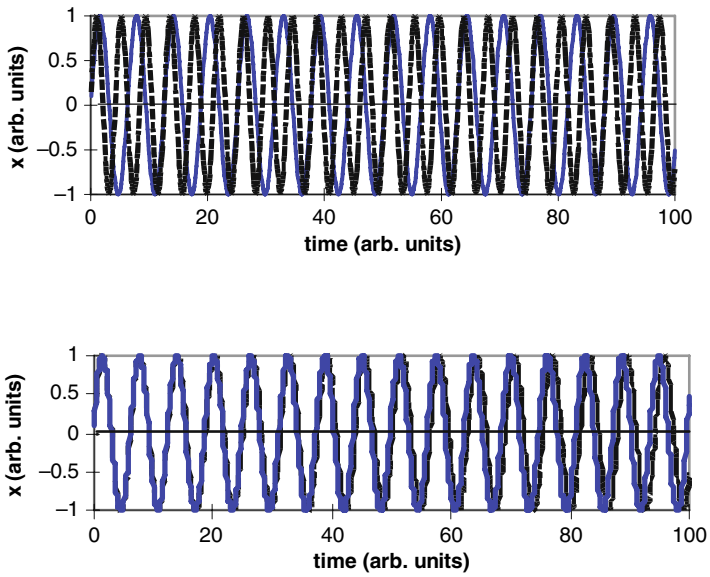


FIGURE 10.4 Top: *Two sine curves with frequencies differing by 50%.* Bottom: *Same, with a frequency difference of only 1%. If the two sine curves represent F and v, then when they are nearly in phase* (bottom) *resonance will occur.*

where the amplitude A depends on both the natural angular frequency of the oscillator ω_0 and that of the external driving force, the oscillation frequency is that of the external force, but a phase shift φ appears so that the driving force and oscillator response are not necessarily in synchrony in time.

After reaching this steady-state condition, the energy added to the oscillator by the driving force in one cycle of oscillation must equal the energy loss through dissipation by the frictional damping force $(-b\vec{v})$ in that same period of time T. The input energy in one cycle is given by the product of the power and the period $E = (\vec{F}_{\text{ext}} \cdot \vec{v})\text{T}$, where the input power is averaged over one cycle of time. If this input energy is small, then the dissipation (velocity term) must be equally small, and so the velocity and hence the oscillator amplitude will be correspondingly small. On the other hand, if the energy input is large then the dissipation must be large, so that the oscillator velocity and therefore amplitude will also be large. We call this phenomenon *resonance*.

What controls the average energy input? Well, clearly the strength of the driving force will be a factor here. However, for a given driving force amplitude, what determines the energy input is how close its driving frequency is to the natural frequency of the oscillator. This is true because the energy input depends on $\vec{F}$ and $\vec{v}$ pointing in the same direction and since both are sinusoidal functions oscillating in the $+x$ and $-x$ directions, as is shown in Figure 10.4, if their two frequencies are very different, the average time they are pointed in the same direction will be much smaller than if their two frequencies are close. Quantitatively, the amplitude of the driven damped harmonic oscillator is given by

$$A = \frac{F/m}{\sqrt{(\omega_0^2 - \omega^2_{\text{ext}})^2 + \left(\frac{b\omega_{\text{ext}}}{m}\right)^2}}. \tag{10.9}$$

Figure 10.5 shows how the amplitude depends on the external driving force. A pronounced maximum, or resonance, occurs as the driving frequency approaches the natural frequency of the oscillator.

Every day examples of resonance abound. When a child on a swing is pushed by a friend, maximum amplitude is reached when the pushes come in sync

FIGURE 10.5 *The amplitude of a driven harmonic oscillator with small damping. When ω approaches ω_0 resonance occurs.*

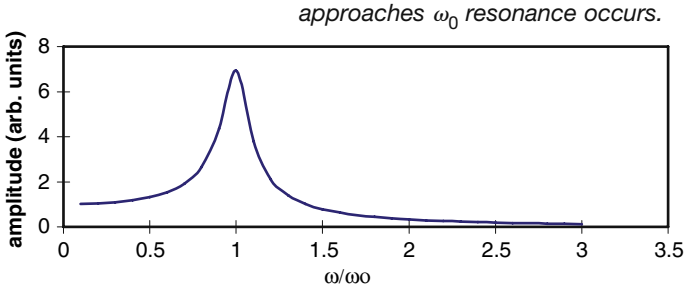


FIGURE 10.6 *One result of the 1989 earthquake near San Francisco, CA. The earthquake vibrations overlapped with the suspended highway resonant frequencies causing large amplitude vibrations leading to its collapse.*

with the natural frequency of oscillation. Hikers marching in step over a suspended bridge can cause large amplitude vibrations of the bridge. Occasionally a similar phenomenon will destroy a poorly designed bridge or highway when energy from wind or earthquakes causes large amplitude oscillations that can weaken the structure. This was the cause of a major highway collapse during the 1989 earthquake in the San Francisco Bay area, for example (Figure 10.6). We also show in the next chapter that resonance plays a major role in the design of musical instruments as well as in the sensitivity of our ears to different frequencies of sound. Many electronic circuits have resonances; when you tune a radio or change the channel on a TV you are choosing a particular resonant frequency. A variety of biophysical techniques also involve resonances, including nuclear magnetic resonance (NMR, and its imaging version, magnetic resonance imaging or MRI), and electron spin resonance (ESR).

2. WAVE CONCEPTS

Mechanical waves are vibrational disturbances that travel through a material medium (in this section we assume no energy dissipation). Examples include water waves, sound waves traveling in a medium such as air or water, waves along a string (as in a musical instrument) or along a steel beam, or seismic waves traveling through the Earth. A general characteristic of all waves is that they travel through a material medium (except for electromagnetic waves which can travel through a vacuum) at characteristic speeds over extended distances; in contrast, the actual molecules of the material medium vibrate about equilibrium positions at different characteristic speeds, and do not translate along the wave direction.

Mechanical waves on a stretched string can be directly visualized. Imagine that we tie one end of a string to a fixed point and stretch it tightly. We can send a wave pulse down the string by giving the held end a single rapid up and down oscillation (Figure 10.7). The motion of the string is vertical whereas the pulse travels horizontally along the string. The vertical forces acting from one region of the string to the next near the leading edge of the pulse are what sustain the pulse and cause it to move along the string. If we continue to oscillate the held end at a fixed frequency f, then we set up a series of identical oscillations, or a *periodic wave*, that travels down the string (Figure 10.8). Such waves are called *transverse*, because the medium oscillates in a plane perpendicular to the direction in which the wave travels.

Suppose we replace the string by a stretched spring tied at one end. If we oscillate the free end of the spring either once, or continuously, along the horizontal direction (along its axis), we set up a *longitudinal pulse*, or periodic wave, in which the motion of the material medium is an oscillation along the direction of propagation of the wave (Figure 10.8).

From a flash photo at some instant of time of the string undergoing continuous oscillations, we can see that the wave consists of a repeating series of positive (above axis, where the axis is the unperturbed string) and negative (below axis) pulses. The distance between corresponding points of one pulse and the next is called the *wavelength*, λ. Because the *waveform*, or shape, is repetitive, or periodic, corresponding points can be neighboring maxima, crests, of the wave, or minima, troughs, of the wave, or any set of neighboring corresponding points (Figure 10.9).

FIGURE 10.7 *Transverse wave pulse on a string.*

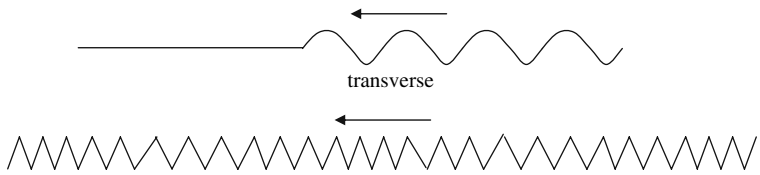


FIGURE 10.8 *Continuous transverse and longitudinal waves traveling to the left along a string or spring, respectively.*

A similar analysis applies to the longitudinal waves of the spring, where now positive and negative refer to the compression or extension of the spring compared to its unperturbed configuration. In this case it is easier to see the wave variation with time clearly by performing the intermediate step of graphing the longitudinal displacement as a function of time to obtain a curve similar to Figure 10.9.

As a wave moves along the string, we can ask with what speed it is traveling. If we look at an arbitrary point along the string, we will see exactly one wave move by in a period, the time $T = 1/f$ required for one oscillation. The distance the wave travels in this time is exactly one wavelength. Therefore, the velocity of the wave is given, quite generally, by

$$v = \frac{\lambda}{T} = \lambda f. \tag{10.10}$$

This same expression holds for longitudinal waves as well and is applicable to all types of waves, from mechanical to electromagnetic.

In addition to mechanical waves on a string or spring, there are several important examples of other waves that we study in this book. Sound waves are mechanical pressure waves traveling in an elastic medium, fluid or solid, causing density variations with regions of lower and higher density. In a solid these waves can be both transverse and longitudinal (as in an earthquake when seismic waves travel through the Earth), but in a fluid, such as air or water, sound waves are only longitudinal. Water waves are also a combination of transverse and longitudinal waves that produce a rolling motion so that as a wave passes by, the water actually travels in an elliptical path. (If you've ever floated in the ocean surf, you will remember that your motion is both up and down as well as horizontal so that you periodically oscillate in a looplike rolling motion.) Electromagnetic waves are transverse waves that are studied in some detail later where we show that these waves do not require a medium in which to propagate but can travel through a vacuum at the speed of light.

Every type of periodic wave has its source in some periodic vibration. For example, sound may be produced by the vibrations of a string, a membrane (drumhead), an air column, or a tuning fork; vibrations of electrons can produce electromagnetic waves of a variety of types including visible light and radio waves. Furthermore, different types of waves will interact with matter in different ways that we study in the course of the remainder of this book.

Waves that can be described by a sinusoidal variation are called *harmonic* waves. At any fixed position such waves vary with time according to Equation (10.1). The wave will also vary with position at a fixed time. For waves on a string, the spatial variation at a fixed time can be captured by a snapshot of a harmonic wave frozen in time that would appear as a sinusoidal curve. We could then describe the vertical position of the string measured from its equilibrium horizontal position in the snapshot photo as

$$y(x) = A\sin(kx), \tag{10.11}$$

where k, known as the *wave number*, is related to the wavelength through the relation

$$k = \frac{2\pi}{\lambda}. \tag{10.12}$$

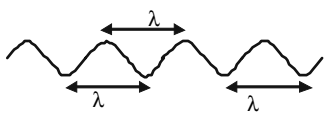


FIGURE 10.9 *Wavelength of a repetitive, or periodic, wave is independent of from where it is measured.*

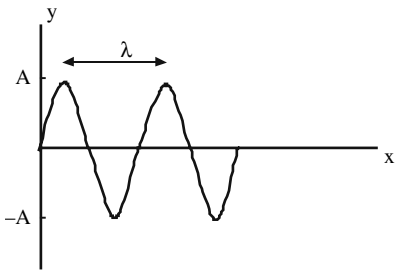


FIGURE 10.10 *Spatial parameters of a harmonic wave.*

Thus as we move horizontally along the snapshot of the string, in the x-direction, the vertical variation in the height of the string is sinusoidal with an amplitude A and a spatial repeat distance of λ (Figure 10.10). Because the sine function has a period of 2π radians, writing the argument as $2\pi(x/\lambda)$ ensures that each time x increases its value by λ, the argument of the sine function will have increased by 2π, maintaining the same value for the function $y(x)$.

Each point on the string actually oscillates in the vertical direction as time goes by so that y, the vertical coordinate, varies not only with x, the distance along the string, but also with time. This is a generalization of Equation (10.1) in which a one-dimensional harmonic oscillator was described using $x(t)$. For a wave on a string the y-coordinate of each point along the string (with a different x-coordinate) varies in time according to an equation similar to Equation (10.1) but with x replaced by y. In the next section we show how we can connect the motion of each point along the string in a simple mathematical way.

3. TRAVELING WAVES

The frozen-in-time snapshot of a sinusoidal wave on a string in the last section actually is traveling along the string in a way that maintains the shape of the wave as it moves along the string. We can describe such a traveling harmonic wave mathematically by writing an expression for the vertical displacement of the string as a function of both x, the horizontal position along the string, and t, the time, as

$$y(x,t) = A\sin(kx - \omega t), \tag{10.13}$$

where ω is the angular frequency of oscillation (remember that $\omega = 2\pi f$). In this section we ignore what happens to the wave at the end of the string by imagining the string to be very long. We consider the effects of a boundary, for example, the tied end of the string, in the next section.

Let's consider the meaning of Equation (10.13) more carefully. If we fix the value of t, then we are looking at the spatial variation of the wave frozen in time as we just did in the last section. Different constant nonzero values of t in Equation (10.13) simply shift the argument of the sine function in Equation (10.11) without any other changes. Note that for a wave to travel along a string, the string must be elastic, or able to stretch. That this is so is obvious on considering that the contour length along the sine curve is clearly greater than the straight line distance along the string axis. The stretch of the string varies along its length and is proportional to the slope of the string. Where the slope is greatest, at the $y = 0$ crossings or nodes, the string is stretched the most, however, where the slope is zero, at the amplitude where y is a maximum or minimum, the string is unstretched.

If we fix, instead of time t, the value of x so that we are looking at the time dependence of the wave at a fixed point on the string, Equation (10.13) reveals a sinusoidal oscillation of the string up and down with an amplitude A and an angular frequency ω or period T (Figure 10.11). Each element of the string moves only vertically. This is precisely the motion of the string to be expected as the waveform given by Equation (10.11) moves by with a velocity v. In this case the waveform remains constant but moves along the positive x-direction at a velocity such as to keep the argument $(kx - \omega t)$, and hence y, equal to a constant. This will occur if $v = x/t = \omega/k = (2\pi f)/(2\pi/\lambda) = \lambda f$, in agreement with Equation (10.10). Thus as the clock ticks on and t increases, the entire waveform, representing $y(x, t)$ moves along the positive x-axis at velocity v. In the case of a wave traveling toward the negative x-axis, the argument in Equation (10.13) simply gets replaced by $(kx + \omega t)$, so that there is a negative velocity with the same magnitude as that in Equation (10.10).

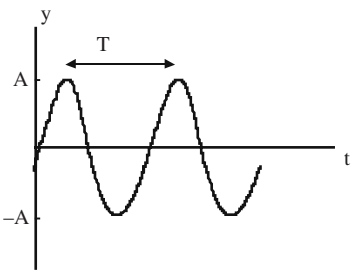


FIGURE 10.11 *Time-dependence of a wave on a string at a particular x-position along the string.*

What determines the frequency and wavelength of the waves traveling along the string? In the case we have been discussing in which one end of the string is made to oscillate, the frequency is determined by the external driving frequency. The wave

velocity for small amplitude waves is determined by two quantities: the tension in the string F_T and an intrinsic property of the string, its mass density or mass per unit length, according to

$$v_{\text{wave}} = \sqrt{\frac{F_T}{(m/L)}}. \tag{10.14}$$

The wavelength of the traveling waves is then determined by the frequency and the wave speed, according to Equation (10.10). From this discussion, we expect that the greater the tension is in the string, the faster the waves travel, and, for a given frequency of oscillation, the longer the wavelength. Similarly, for the same driving frequency and length of string, a more massive string will result in a slower wave speed and therefore a shorter wavelength.

Example 10.2 A traveling wave on a string is described by the equation $y = 0.025 \sin(1.5x - 200t)$ where x and y are measured in m and t in s. If the string has a mass per unit length of 0.003 kg/m, find the following quantities: the amplitude, wavelength, frequency, period, the velocity of the wave, and the tension in the string.

Solution: From the general form of a traveling wave on a string, given by Equation (10.13), we can identify directly from the given equation that the amplitude $A = 0.025$ m, the wave number $k = 2\pi/\lambda = 1.5\ \text{m}^{-1}$, and the angular frequency $\omega = 2\pi/T = 200$ rad/s. We can therefore straightforwardly compute the wavelength to be $\lambda = 2\pi/k = 4.2$ m and the period to be $T = 2\pi/\omega = 0.031$ s. The frequency is the inverse of the period and is therefore equal to $f = 1/T = 31.8$ Hz. Because the wave travels a distance of one wavelength in a time equal to one period, the wave velocity is given as $v = \lambda/T = 130$ m/s. From this value and the equation connecting the speed of a wave on a string to the tension in the string (Equation (10.14)), we can solve for the tension, $F_T = v^2(m/L) = 53$ N.

Having described the waveform, the relationships between the variables describing the waveform and the velocity of a wave on a string, we can ask the obvious question: if a wave is not the translational motion of the material medium itself, what is it that is transported with the wave velocity? The answer is energy. Continuing with our example of the string, the energy that is input to the system from the external driving force at one end is transmitted along the string at velocity v_{wave}. With a single pulse sent down the string it is clear that the kinetic energy of the transverse motion of the string is translated along the string with the pulse. If we imagine the string to be divided up into short segments along the x-direction, we can ask where the segments have their maximum and minimum kinetic and potential energy when a harmonic wave travels along the string. Because each element moves vertically, oscillating harmonically about $y = 0$ as a function of time, the kinetic energy of an element is a maximum as it moves through the $y = 0$ position (Figure 10.12).

At the amplitude, $y = \pm A$, the segment is instantaneously at rest and therefore has no kinetic energy. The stretch of the string is proportional to its slope, and the elastic potential energy is proportional to the product of the tension force and the stretch, therefore we see that the elastic potential energy is also maximum at $y = 0$ where the slope of the string is a maximum. Again at the amplitude, $y = \pm A$, the slope of the string is zero and therefore so is the elastic potential energy. (Note that this is in contrast to a mass on a spring, where the elastic potential energy is a maximum at the amplitude.) In fact, it can be shown that the kinetic and potential energies are exactly equal for harmonic waves traveling along an elastic string, with the peaks in energy located at the $y = 0$ crossings and moving with the wave velocity

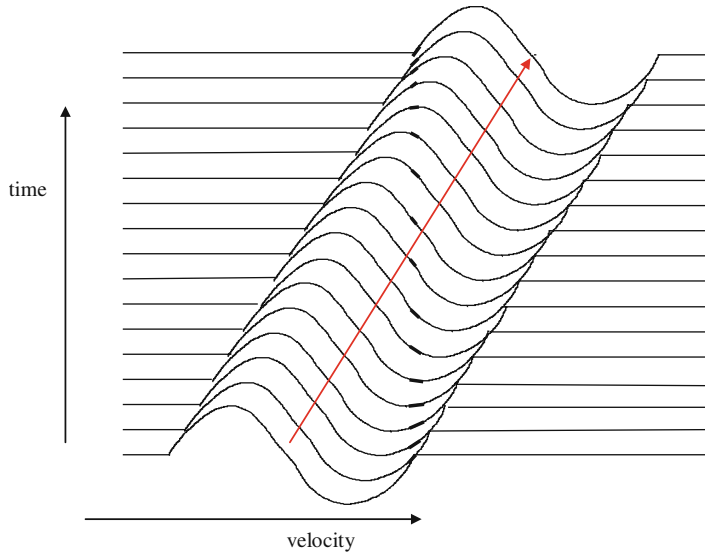


FIGURE 10.12 *A time series of the motion of a wave on a string. The thick line segment represents the same piece of string oscillating as the wave passes by. The red arrow indicates the location of the maximum energy of the pulse as it moves along.*

along the string. So it is energy that travels along the string and constitutes the wave. The elements of the string behave as harmonic oscillators each carrying a total energy proportional to the square of the wave amplitude and transmitting that energy along the string through the elastic interactions with neighboring string elements. This is a general result of harmonic waves: the total energy carried by the waves is proportional to the square of the wave amplitude.

We have only discussed traveling waves along an elastic string. Traveling longitudinal harmonic waves can also be produced on a coiled spring by oscillating one end longitudinally at a fixed frequency (see Figure 10.8). The variations in the compression and expansion of the spring result in a wave traveling down the spring. If $y(x, t)$ represents the local displacement (assumed small) of the spring from its equilibrium position as a function of both the position along the spring, x, and the time, t, then Equation (10.13) fully describes such longitudinal waves as well.

Both of the examples of traveling waves cited are one-dimensional cases with waves traveling along the x-direction. When a rock is dropped in a pond of water, waves spread out radially along the two-dimensional surface of the pond with the *wavefronts* (or shape of the crests) forming circles. Light waves from a light bulb travel radially outward in space in three dimensions with spherical wavefronts, as do sound waves from a person who is speaking. We study some of these examples later in the text, but we note that the fundamental definitions introduced in this chapter are still appropriate but that our one-dimensional pictures need to be generalized for these other situations.

4. WAVES AT A BOUNDARY: INTERFERENCE

When traveling waves reach boundaries between two different media several different phenomena can occur. In the case of one-dimensional waves, at a boundary part of the *incident wave* will continue into the new medium as the *transmitted wave*, traveling at a different velocity due to the medium's different properties, and the balance of the wave's energy will be reflected back within the incident medium as the *reflected wave*. In the case of waves traveling in the positive x-direction along a string with a particular linear mass density m/L tied to another string with a different mass density at a knot between the two strings, the knot serves as the boundary. As the incident wave (or pulse) arrives at the boundary, there will be both a transmitted and a reflected wave (pulse). A portion of the energy will enter the new medium and the transmitted wave (pulse) will continue to travel in the positive x-direction but at a different velocity according to Equation (10.14). (F_T will be the same but m/L is different.) The reflected wave (pulse) will contain the balance of the incident energy and will return along the string traveling along the negative x-direction.

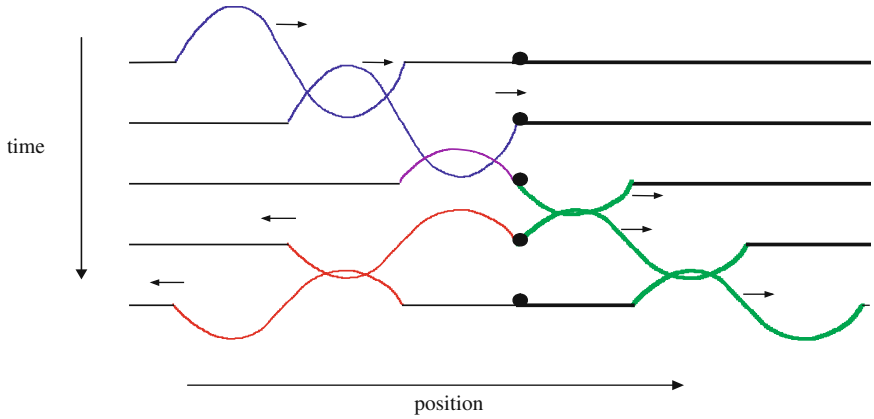


FIGURE 10.13 *A time sequence of events, from top to bottom, when a single pulse waveform traveling to the right* (blue) *meets a boundary. The string on the right is heavier than that on the left, so that the reflected* (red) *and transmitted* (green) *pulses are as shown with the reflected wave inverting as it is reflected. Note that in the center picture the incoming and reflected waves in the lighter string overlap* (red + blue = purple) *and add together at that instant. If the strings were reversed so that the wave entered on the heavier string, there would be no inversion of the wave on reflection.*

Consider the example of an individual pulse traveling to the right along our string as shown in Figure 10.13. The figure shows the time sequence of events that occur when this pulse reaches the knot between two different strings. A portion of the amplitude of the pulse continues into the second string traveling to the right. The reflected pulse passes through the incident pulse emerging in reverse order traveling to the left. During the time that the two pulses overlap along the string their amplitudes are seen to add together. This is an example of the *superposition principle*, an extremely important concept in wave physics. We have already seen the superposition principle in action when, in Chapter 5, we noted that the net vector force was the sum of the individual vector forces acting on an object. For waves, this principle states that the wave displacement at any point is the algebraic sum of the individual displacements of the overlapping waves at that point. Said differently, the net waveform is the algebraic sum of the individual waveforms.

A consequence of the superposition principle is the phenomenon known as *interference*. Two transverse waves traveling in the same direction along the same string will add together to produce a resultant wave that is the observed waveform. Mathematically the expressions for the two waves add algebraically. If they have the same wavelength (and, because the velocities are the same, also the same frequency) and are *in phase*, so that their crests and troughs march together along the string, then the resultant amplitude will be their sum. In this case if the two waves are identical, each of amplitude A, the resultant wave will have an amplitude of $2A$ (Figure 10.14 left). These two waves are said to combine by *constructive interference*.

If the waves have equal amplitude A, and the resulting waves are completely *out of phase*, so that the crest of one travels together with the trough of the other, then the two waves combine by *destructive interference* and, in this case of equal amplitudes, will completely eradicate each other resulting in no disturbance of the string at all. If the two out of phase waves have different amplitudes A_1 and A_2, as in the center panel of Figure 10.14, then the destructive interference leads to partial cancellation of the waves and an amplitude equal to $|A_1 - A_2|$. When the two waves are partially out of phase, as in the right panel of Figure 10.14, they will add together to produce a wave with the same wavelength but an amplitude that is between 0 and $A_1 + A_2$ ($=2A$ if the amplitudes are equal) depending upon their phase difference (or relative position of their crests).

We can explore the interference of two equal amplitude waves a bit further by writing each of the two waves that overlap in the form of Equation (10.13), but with one wave shifted by an arbitrary phase φ with respect to the other so that

$$y_1 = A \sin(kx - \omega t) \quad \textit{and} \quad y_2 = A \sin(kx - \omega t + \varphi). \tag{10.15}$$

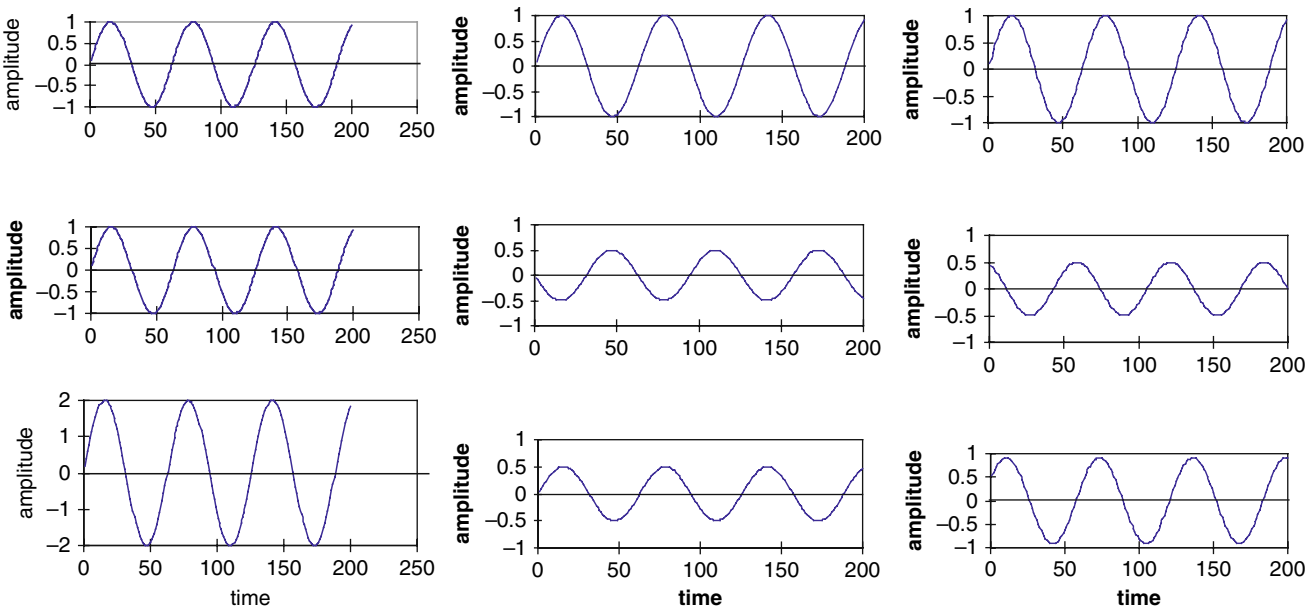


FIGURE 10.14 *The superposition of two harmonics with the same frequency.* (left) *Equal amplitude waves in phase;* (center) *unequal amplitude waves 180° out of phase;* (right) *unequal amplitude waves with arbitrary phase.*

When these two waves overlap, the principle of superposition tells us that the total wave amplitude will be

$$y = y_1 + y_2 = A(\sin(kx - \omega t) + \sin(kx - \omega t + \varphi)).$$

Using a trigonometric identity (namely, $\sin\alpha + \sin\beta = 2\sin\frac{1}{2}(\alpha + \beta)\cos\frac{1}{2}(\alpha - \beta)$), we can simplify this expression to find

$$y = [2A\cos\frac{1}{2}\varphi]\sin(kx - \omega t + \frac{1}{2}\varphi). \qquad (10.16)$$

This result shows that the superposition is also a traveling wave with the same wavelength and frequency, but shifted in phase by $\varphi/2$ and with an amplitude, given by the terms in the square bracket, that depends on the phase angle and lies between 0 and $2A$.

If the two traveling waves are in phase, or interfere constructively with $\varphi = 0$, then Equation (10.16) yields a net amplitude equal to the sum of the separate amplitudes ($2A$), as we saw earlier. On the other extreme, if the two traveling waves interfere completely destructively with $\varphi = \pi$, or 180°, then the two waves will exactly cancel, giving an amplitude identically equal to 0. Equation (10.16) gives the result for the general case of arbitrary phase angle.

As a further example of interference, consider the case of two waves of slightly different wavelength (or frequency) traveling in the same direction along the same string. Figure 10.15 shows two waves that differ in frequency by 10% (red and

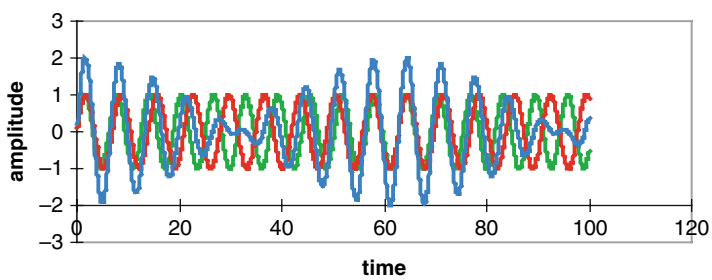


FIGURE 10.15 *The superposition* (in blue) *of two equal amplitude sinusoidal waves of slightly (10%) different frequency, illustrating the phenomenon of beats.*

green sine curves) and their superposition (in blue). Notice that in addition to a periodic variation at the average frequency, the resultant wave has a slower periodic variation that occurs at the difference frequency. In the figure this lower frequency component has a period equal to ten times that of the higher frequency component; you can count ten peaks between a longer period repeat. The slower variation is due to the interference of the two waves that leads to more or less cancellation in a period fashion. This phenomenon is known as *beats* and in the case of sound waves results in an audible low frequency variation in loudness. When two tones are played that are very close in frequency, one hears the average frequency tone modulated in loudness at the difference or *beat frequency*. This phenomenon is discussed in more detail in Section 3 of the next chapter. Beats can be used to tune an instrument when a standard frequency is used to generate one of the tones; the instrument is tuned so as to lower the beat frequency, lengthening the period of the loudness variations. In the limit of an infinite beat period the two frequencies are identical.

5. STANDING WAVES AND RESONANCE

We now consider the situation on a string when we force one end to oscillate in simple harmonic motion at some frequency f and fix the other end of the string so that it cannot move. In this case as the wave reaches the fixed end, all of its energy is reflected, and the reflected wave reverses its sign. This reversal of sign is a byproduct of the requirement of a fixed point; if the wave did not reverse itself on reflection then the amplitude would not always add to zero at the fixed point. Reversal of sign of the reflected wave also occurs for the case of two strings tied together when the wave travels from the lighter to the heavier string, a situation shown in Figure 10.13.

If the string has a length L, then the round-trip of the wave down and back along the string requires a time equal to $2L/v_{\text{wave}}$. If the reflected wave arrives back at the oscillating end of the string at a time just equal to a period of oscillation $1/f$ of the driven end of the string, then the waves traveling to the right and the left will be exactly in phase and constructively interfere, producing a standing wave as shown in the sequence of events in Figure 10.16.

We can understand this result by adding together two waves of equal amplitude that are traveling along the string in opposite directions. Given

$$y_1 = A\sin(kx + \omega t) \quad \textit{and} \quad y_2 = A\sin(kx - \omega t),$$

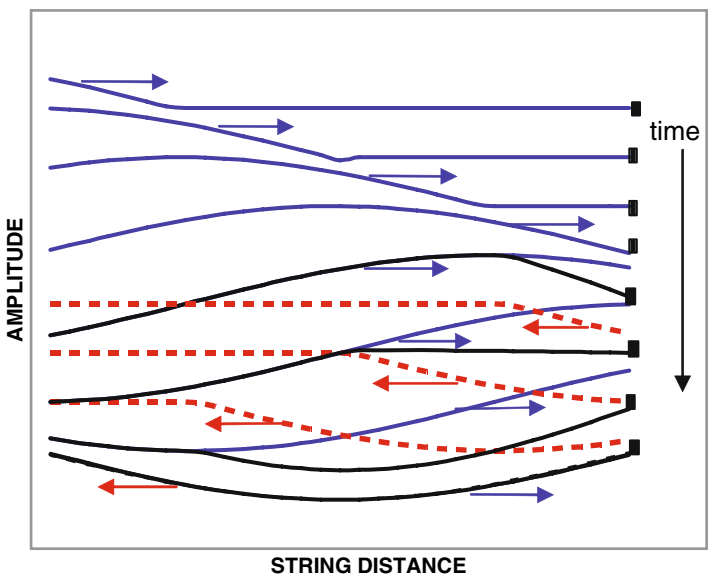


FIGURE 10.16 *A sequence of eight equal time views spanning one period and showing a string tied down at the right and driven at the left. A wave pulse travels to the right* (blue) *in the first four views, reaching the knot. In subsequent views, a reflected wave* (red dashed curve or red arrow) *returns to the left, and the incident wave continues to the right; the black curves are the superposition of the incident* (blue) *and reflected* (red) *waves and may overlap the red/blue curves. Note that the reflected* (red) *wave returns to the left end just in phase with the driver* (or incident blue wave)*, setting up a standing wave with one-half the wavelength just fitting along the string.*

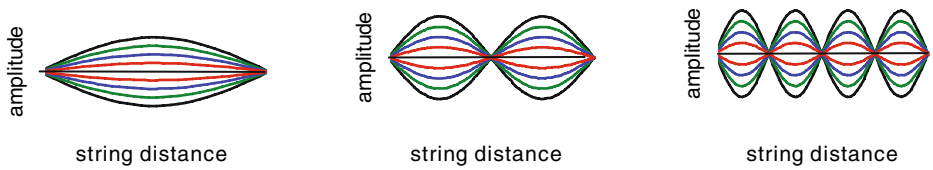


FIGURE 10.17 *Time sequences showing the fundamental* (left)*, second* (center)*, and fourth* (right) *harmonic standing waves on a string. (Note that the time intervals between snapshots are not equal; the string spends more time out near its amplitude where its transverse velocity is slower and less time near the horizontal equilibrium position where its velocity is most rapid.)*

employing the same trigonometric identity that we used to get Equation (10.16), we have for the sum

$$y = y_1 + y_2 = 2A \sin kx \cos \omega t. \tag{10.17}$$

What is striking about this result is that the sum of these two traveling waves is no longer a traveling wave. At any value of x the amplitude oscillates at angular frequency ω, but there is no waveform that travels along the string. In fact there are periodic positions along the string (corresponding to kx equal to either 0 or multiples of π) where the amplitude is always equal to zero. This type of wave is known as a *standing wave*.

Because the string length and wave velocity are fixed, for most continuous oscillation frequencies the waves traveling to the right and left will have no particular phase relation, with the wave returning to the left end at different values of transverse displacement at the start of each of the forced oscillations. The result of such a situation will be a net destructive interference and no sustained displacement of the string. Only for a particular set of frequencies, called the *resonant frequencies*, will standing waves be produced. The lowest possible resonant frequency is called the fundamental frequency, or first harmonic, and is the situation shown in Figure 10.17 (left) in which half of a wavelength fits on the string. The wavelength is then equal to $\lambda = 2L$, so that the fundamental frequency is equal to $v_{\text{wave}}/(2L)$, or the inverse of the round-trip time.

As the frequency is increased beyond the fundamental, there will be a sequence of discrete frequencies, called harmonics, at which resonance will occur. At the second harmonic frequency, for example, the wave will reach the right end and reflect back in the same round-trip time but now corresponding to two complete oscillations, so that the resonant frequency is twice that of the fundamental. The wavelength is then equal to $\lambda = L$, with the second harmonic frequency given by v_{wave}/L, precisely twice the fundamental frequency. In this case, the waves traveling to the right and left will always produce a point at the center of the string at which there is no displacement. Such a point is called a node and, as can be seen in Figure 10.17, each higher harmonic adds one additional node along the string. The wavelengths of these resonances are given by

$$\lambda_n = \frac{2L}{n}, \qquad n = 1, 2, 3, \ldots, \tag{10.18}$$

where n is the harmonic number; $n = 1$ refers to the fundamental or first harmonic, $n = 2$ to the second harmonic, and so on. The corresponding resonant frequencies are given by

$$f_n = \frac{v_{\text{wave}}}{\lambda_n} = nf_1. \tag{10.19}$$

The fourth harmonic is shown in Figure 10.17. The second and higher harmonics are also known as the overtones, with the second harmonic also called the first overtone, the third harmonic also called the second overtone, and so on.

Example 10.3 A steel guitar string with a 10 g mass and a total length of 1 m has a length of 70 cm between the two fixed points. If the string is tuned to play an *E* at 330 Hz, find the tension in the string.

Solution: From the frequency and the fact that the fundamental has a wavelength equal to twice the distance of 0.7 m, we find that the wave velocity must be equal to $v = f\lambda = (330)(1.4) = 462$ m/s. Then given the mass per unit length of 0.01 kg/1 m = 0.01 kg/m, we can use Equation (10.14) to find the tension. From $v = \sqrt{\frac{F_T}{m/L}}$, we can solve for F_T to find

$$F_T = v^2\left(\frac{m}{L}\right) = (462)^2\,(0.01) = 2130 \text{ N}.$$

Enormous tensions are needed in stringed instruments. Steel, nylon, or natural fibrous materials such as catgut are used to support these tensions.

Standing waves on a string are one example of the more general phenomenon of resonance, introduced in Section 1 for the case of simple harmonic motion. In general, resonance is the addition of energy to a system at one of the natural frequencies of the system. In the case of the string, standing waves occur if the driving force frequency is equal to the fundamental or any harmonic frequency of the system, as determined by the length and mass per unit length of the string as well as its tension. As the driving frequency is tuned, a series of resonances with different amplitudes is produced (Figure 10.18).

Standing waves can be set up in any object that is made to vibrate, including all musical instruments at sound frequencies, and bridges, buildings, and other manmade constructions, as well as ocean water at subsonic frequencies (Figure 10.19). We study some of these in connection with sound a bit further in the next chapter. Resonance can occur in many other types of systems including atomic or molecular systems. In these cases involving the microscopic world, electromagnetic oscillations, comparable to mechanical or sound vibrations, produce the resonance. Nuclear magnetic resonance (NMR—the basis for MRI—magnetic resonance imaging) occurs when electromagnetic radio waves are tuned to have the energy needed to produce spin flips in the nuclei and are studied later in this book. A variety of other spectroscopic techniques that involve the interactions of various types of electromagnetic radiation with matter can be analyzed using the concept of resonance.

Even the simpler case of resonance in damped forced harmonic motion, as discussed in Section 1, can serve as the basis for analyzing a variety of physical systems ranging from the mechanical pendulum in a grandfather clock, or a child being rhythmically pushed on a playground swing, to electromagnetic and quantum systems in which radiation acts as

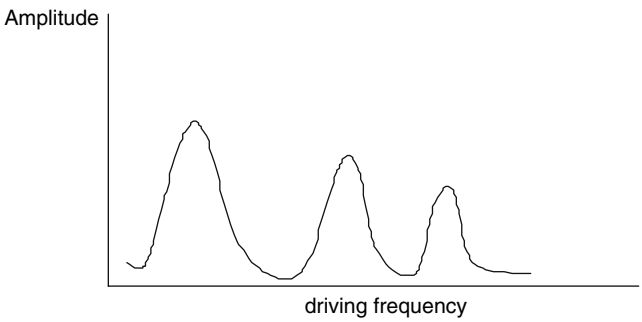


FIGURE 10.18 *Multiple resonances in a real system (such as a string tied at one end) will occur as the driving frequency is varied.*

FIGURE 10.19 *Standing wave sand markers where the ocean and a stream meet.*

the driving force and the damped oscillations are those of electrons or nuclei in molecules. Just as we saw in Chapter 4 (Section 4) that springs are the natural "picture" that we can use to approximate the forces acting near equilibrium, the addition of a damping and a driving force allow for interactions of the spring with both internal forces (the frictional loss of energy) and external forces (the addition of energy to the system). In biological systems, receptors (of sound, light, or specific molecules) usually involve a resonance. For example, in the next chapter on sound we learn about Helmholtz resonance in the ear and the Békésy resonant waves in the cochlea.

CHAPTER SUMMARY

In the presence of a damping force proportional (through the damping constant b) to velocity, the position as a function of time of a mass m attached to a spring with spring constant k is given by

$$x(t) = (Ae^{-\frac{bt}{2m}})\cos(\omega_{damp}t), \qquad (10.4)$$

where the angular frequency is

$$\omega_{\text{damp}} = \sqrt{\frac{k}{m} - \frac{b^2}{4m^2}}. \qquad (10.5)$$

When the b is equal to zero, these equations reduce to the simple harmonic motion case:

$$x(t) = A\cos(\omega_0 t), \qquad (10.1)$$

with the natural frequency $\omega_0 = \sqrt{\frac{k}{m}}$.

When the oscillator is driven by an external force F oscillating at ω_{ext} the position of the oscillator is given by

$$x(t) = A(\omega_0, \omega_{\text{ext}})\cos(\omega_{\text{ext}}\, t + \varphi), \qquad (10.8)$$

$$A = \frac{F/m}{\sqrt{(\omega_0^2 - \omega_{\text{ext}}^2)^2 + \left(\frac{b\omega_{\text{ext}}}{m}\right)^2}}. \qquad (10.9)$$

Such an oscillator exhibits the phenomenon of resonance: the amplitude rises rapidly as the external frequency approaches the natural frequency of the spring and more external energy is able to be absorbed by the system. This model of the driven damped harmonic oscillator using a spring is broadly applicable to a variety of other types of systems.

All periodic waves (with wavelength λ and frequency f) travel at a speed given by

$$v = \lambda f. \qquad (10.10)$$

Periodic traveling waves (e.g., waves on a string) can be written showing their displacement as a function of both position x and time t:

$$y(x, t) = A\sin(kx - \omega t), \qquad (10.13)$$

where k is the wave number,

$$k = \frac{2\pi}{\lambda}. \qquad (10.12)$$

Waves obey the principle of superposition: they pass through each other undisturbed and where they overlap in space, the net amplitude of the wave is equal to the algebraic sum of those of the overlapping individual waves. Interference is a consequence of superposition. Two waves traveling along a string with equal amplitude, wavelength, and frequency but with a phase difference φ between them superimpose to yield a net traveling wave that has an amplitude given by the term below in square brackets and is shifted in phase by $\varphi/2$ from either original wave:

$$y = [2A\cos\frac{1}{2}\varphi]\sin(kx - \omega t + \frac{1}{2}\varphi). \qquad (10.16)$$

In the special case of two overlapping waves of equal amplitude traveling in opposite directions along a string of length L (perhaps from reflections at the ends), standing waves can be produced:

$$y = y_1 + y_2 = 2A\sin kx\cos\omega t. \qquad (10.17)$$

These only occur when the frequency (or wavelength) satisfies the resonance conditions:

$$f_n = \frac{v_{\text{wave}}}{\lambda_n} = nf_1, \qquad (10.19)$$

$$\lambda_n = \frac{2L}{n}. \qquad n = 1, 2, 3, \ldots. \qquad (10.18)$$

QUESTIONS

1. Give several examples of everyday phenomena that approximate harmonic motion. In each case name the source of damping.
2. What are a few examples of forced harmonic motion?
3. Carefully define the amplitude, phase angle, driving frequency, and natural frequency for driven harmonic motion.
4. Name some examples of resonance phenomena, giving the approximate resonant frequency involved.
5. What is the difference between equilibrium and steady state? Which one requires an input of energy?
6. Define wavefront. What is the wavefront shape of each of the following?
   (a) A three-dimensional wave emanating from a point in all directions
   (b) An in-phase wave traveling along the x-direction
   (c) A two-dimensional wave (such as on a drum membrane or the surface of a lake) emanating from a point
7. What is the difference between transverse and longitudinal vibrations of a spring? Distinguish between the wave velocity and spring velocity in each case.
8. What distinguishes a harmonic wave from any other type of wave?
9. Equation (10.13) defines a wave traveling along the positive x-axis, because as time increases x must increase for points of constant phase. How would you write an expression for a wave traveling along the negative x-direction?
10. Why is the potential energy of a stretched string zero at the amplitudes of a traveling wave and maximum at the zero-crossings?
11. Two waves, each of amplitude A with intensities proportional to A^2, overlap in space producing an interference effect. Although the total intensity of the two separate waves is proportional to $2A^2$, the net amplitude where they overlap can range from 0 to $2A$, so that the net intensity can range from 0 to $4A^2$. Discuss this in terms of conservation of energy.
12. Discuss in words how a node is produced for a standing wave on a string. How does the string move at an antinode?
13. A string of length L, mass per unit length μ, and tension F_T is vibrating at its fundamental frequency. Describe the effect that each of the following conditions has on the fundamental frequency.
   (a) The length of the string is doubled with all other factors constant.
   (b) The mass per unit length is doubled with all other factors constant.
   (c) The tension is halved with all other factors constant.
14. When two different stringed instruments play the same fundamental note, what is it that allows you to distinguish the tone from the two instruments, for example, a violin and a viola?
15. Two strings are tied together in a knot. One string has a length L and a mass m, and the other one has half the length and twice the mass. If the strings are stretched taut and put under tension and a transverse wave travels down the longer string, through the knot into the shorter string, what is the ratio of the wave speeds in the shorter to longer string? What is the ratio of the frequencies in the two strings? The wavelengths?
16. Consider the same two strings tied together as in the previous question. If a positive wave pulse (above the axis of the strings) is sent down the longer string, what will be the polarity of the pulse reflected at the knot? If a positive pulse is sent the other way along the string, what will be the polarity of the reflected pulse in this case?
17. When a snowstorm occurs, often there is a variation in the amount of snow on an electric high-voltage wire strung between two support poles as a result of standing waves. Show what you might expect the pattern to look like for the lowest resonant modes.

MULTIPLE CHOICE QUESTIONS

1. Which of the following is not true of simple harmonic motion of a mass on a spring? (a) The maximum acceleration occurs at the amplitude of motion, (b) the resonant frequency is proportional to the square root of the mass, (c) the period is independent of the amplitude of the motion, or (d) the kinetic and potential energies of the mass exchange with each other at twice the resonant frequency.
2. A disturbance in a string has a node at $x = 0$ m, at $t = 0$ s. At $t = 1$ s, the same node is observed to be at $x = 5$ m. This disturbance must be (a) a wave traveling in the negative x-direction with speed 5 m/s, (b) a wave traveling in the positive x-direction with speed 5 m/s, (c) a standing wave with nodes separated by 5 m, (d) either a standing or traveling wave with frequency equal to 1 Hz.
3. A transverse sinusoidal wave travels along a string with a constant speed 10 m/s. The acceleration of a small lump of mass on the string (a) varies sinusoidally in time in a direction perpendicular to the string, (b) varies sinusoidally in time in a direction parallel to the string, (c) is 10 m/s^2, (d) is zero.
4. In a periodic transverse wave on a string the value of the wave speed depends on (a) amplitude, (b) wavelength, (c) frequency, (d) none of the choices (a)–(c).
5. Two strings are held under the same tension. String A has a mass per unit length that is two times that of string B. The wave speed in A is (a) the same as in B, (b) one half that in B, (c) two times that in B, (d) none of the above.
6. Suppose the tension in a string is given by T and the mass per unit length by μ. What are the fundamental dimensions (i.e., M, L, and T) of the quantity $\sqrt{T/\mu}$? (a) LT^{-1}, (b) MLT^{-2}, (b) L^2T^{-2}, (d) $L^{1/2}M^{-1/2}T^{-1/2}$.

Questions 7–9 refer to: A transverse traveling wave on a string is described by the mathematical expression $y = (0.10)\sin(2\pi x + 10\pi t)$, where x and y are measured in meters and t is measured in seconds.

7. The frequency of this wave is (a) 10 Hz, (b) 5 Hz, (c) 2 Hz, (d) 1 Hz.

8. This wave is traveling in which direction? (a) $+y$, (b) $-y$, (c) $+x$, (d) $-x$.

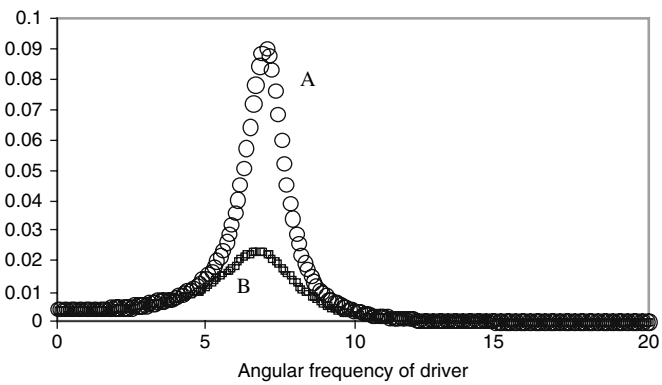
9. The speed with which this wave travels is (a) 1 m/s, (b) 2 m/s, (c) 5 m/s, (d) 10 m/s.

10. Given the traveling wave $y(x, t) = 0.1 \sin(\pi x - \pi t/2 + \pi/2)$, with x and y in meters and t in seconds, its frequency is (in Hz)
(a) 0.25, (b) 2.0, (c) π, (d) $\pi/2$, (e) none of the above.

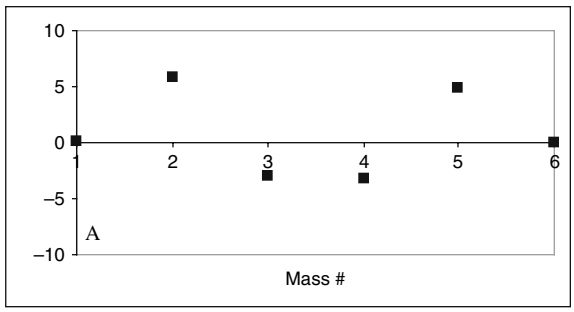
11. Antinodes and nodes occur(a) in standing waves, (b) in traveling waves, (c) during beats, (d) in longitudinal waves, (e) none of the above.

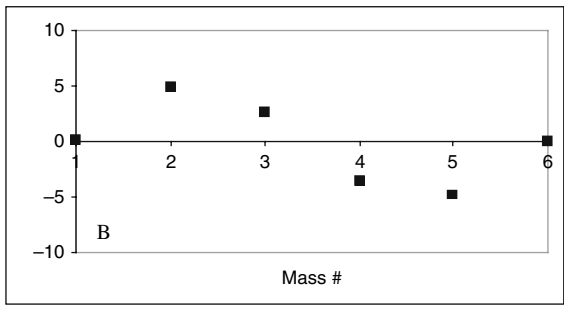
12. When a string tied down at both ends is plucked, the resonant frequencies are characterized by all of the following except (a) there must be nodes at both ends, (b) they must satisfy the equation $f = v_{wave}/\lambda$, (c) the fundamental frequency is the lowest allowed resonant frequency, (d) the fundamental wavelength is L.

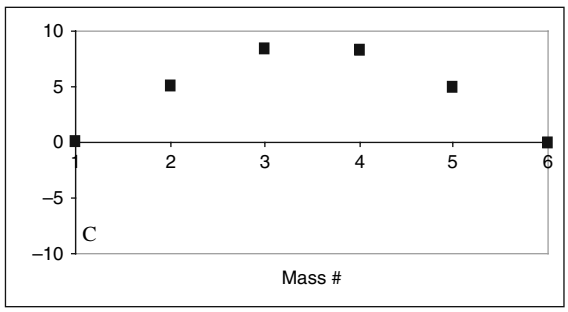
13. Two identical masses are each attached to a spring. The springs are also identical. The masses are driven by the same periodic external force and the response curves (amplitude versus driving frequency) are shown to the right. Which of the following best describes what is seen in the graphs? (a) Mass B is not as well attached to its spring as is mass A. (b) Mass A is at resonance but mass B is not. (c) Mass A experiences more friction than mass B. (d) Mass A experiences less friction than does mass B.

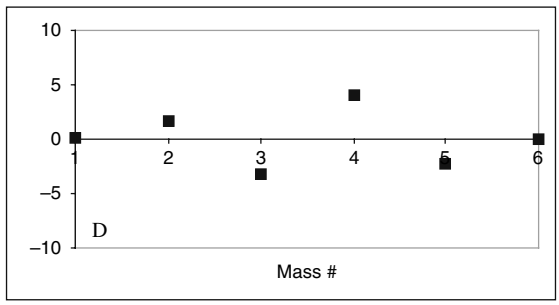


14. Four masses capable of moving along a line are interconnected by springs. These masses are driven into resonance by an external force. The following graphs show the masses' displacement from equilibrium, at a given instant, in each of the allowed resonant modes. (The end masses in these graphs don't move; they're not part of the system.) Rank order the frequencies associated with each graph. (Hint: connect the dots.) (a) A > B > C > D, (b) D > C > B > A, (c) A > D > C > B, (d) D > A > B > C.









15. When two identical harmonic waves of amplitude A interfere, the net result can be all but which of the following: (a) no wave, (b) a harmonic wave with an amplitude of $2A$, (c) a harmonic wave with an amplitude of $1.5A$, (d) a harmonic wave with an amplitude of $4A$.

16. In forced harmonic motion, as the frequency of the external oscillation driving force approaches the natural frequency of oscillation in a phenomenon called resonance, which of the following occurs? (a) The

period becomes increasingly long. (b) The amplitude becomes increasingly large. (c) The frictional damping becomes increasingly large. (d) The energy steadily decreases.

17. You observe a string under tension (fixed at one end and supporting a hanging weight at the other) to form a standing wave when the driving frequency is 40 Hz. If you replace the 200 g hanging weight with a 100 g weight (but don't change the wire length) the standing wave with the same shape will occur at about what frequency? (a) 40 Hz, (b) less than 40 Hz, (c) greater than 40 Hz, (d) you can't form a standing wave with the same shape under these conditions.

18. You are told that the mass per unit length of a wire is 1×10^{-3} kg/m and that a 0.1 kg mass is to be used to stretch the wire, by hanging from one end with the other end held fixed. Which of the following is true about the wave speed in the wire? The wave speed (a) depends on the length of the wire, (b) depends on the frequency with which the wire is vibrated, (c) is approximately 1000 m/s, (d) is approximately 30 m/s.

19. The fundamental standing wave on a string of length 1 m that is fixed at both ends vibrates at a frequency of 300 Hz. The speed of waves on this string must be (a) 100 m/s, (b) 150 m/s, (c) 300 m/s, (d) 600 m/s.

20. Suppose a vibrating wire is exactly 1 m long. The standing wave corresponding to the third harmonic on this wire has a frequency of 30 Hz. The wave speed of a transverse wave on this wire (a) is 10 m/s, (b) is 20 m/s, (c) is 30 m/s, (d) cannot be determined from the information given.

21. In restringing a violin A string (fundamental $f =$ 440 Hz), if a string with twice the mass/length is incorrectly used and the tension is adjusted to play the correct fundamental, by what factor is the tension different from what it should be using the correct string: (a) $^1/_2$, (b) $\sqrt{2}$, (c) 2, (d) 4.

22. A car travels over a dirt road that contains a series of equally spaced bumps (a so-called "washboard" road). While traveling at a given speed the driver experiences a very jarring ride. When the driver drives at a higher speed, however, the ride gets smoother. That is because (a) the car actually leaves the ground at higher speeds, (b) the faster moving car actually crushes the bumps and makes the road smoother, (c) the car's shock absorbers have more friction at higher speeds, (d) going faster in the car forces the suspension to oscillate at a frequency higher than its natural frequency.

23. A damped driven oscillator has an equation of motion given by $ma = -kx - bv + F_0 \cos(\omega_d t)$, where ω_d is the angular frequency of the driving force. At resonance ma must equal (a) $-kx$, (b) $-bv$, (c) $+F_0\cos(\omega_d t)$, (d) zero.

PROBLEMS

1. You are watching a mass oscillate on a spring. You measure the period to be a constant 1.1 s but you see that the 10 cm initial amplitude of the oscillation halves after 10 s. Write an expression for the time-dependence of the position of the mass $x(t)$ in terms of t with all other factors given as numbers.

2. In the previous problem, how long will it take the mass to lose half its initial energy?

3. A 0.5 kg mass attached to a linear spring, with spring constant 5 N/m and damping constant 0.2 kg/s, is initially displaced 10 cm from equilibrium.
(a) What is the natural frequency of oscillation?
(b) What is its period of oscillation?
(c) How long does it take for the amplitude to decrease to 10% of its starting value?
(d) How many oscillations have occurred in this time?
(e) What fraction of the initial energy remains after this time?

4. A 0.2 kg mass is attached to a vertical hanging spring, stretching it by 10 cm. The mass is then pulled down an additional 10 cm and released. It is found that the amplitude decreases to 5 cm in 30 s.
(a) What is the spring constant?
(b) Find the natural frequency of oscillation.
(c) What is the damping constant of the spring?
(d) Write the equation of motion for the mass as a function of time.
(e) Write an equation for the energy of the mass as a function of time.

5. A 1 kg mass is attached to a vertically hanging spring with spring constant 10 N/m and damping constant 0.1 kg/s. Suppose a harmonic driving force with fixed amplitude of 1 N and variable frequency is applied to the mass. Construct the resonance curve showing the amplitude of oscillation as a function of the driving frequency near the natural frequency of oscillation of the mass. Use a set of about 10 points to show the main features of the curve.

6. A vertical spring with a spring constant of 8 N/m and damping constant of 0.05 kg/s has a 2 kg mass suspended from it. A harmonic driving force given by $F =$ $2\cos(1.5t)$ is applied to the mass.
(a) What is the natural angular frequency of oscillation of the mass?
(b) What is the amplitude of the oscillations at steady state?
(c) Does this amplitude decrease with time due to the damping? Why or why not?

7. A 4 m long rope weighing 1.4 N is stretched so that the tension is 10 N. The left end is then made to oscillate vertically at 4 Hz by shaking the rope up and down a total distance of 10 cm.
(a) What is the speed of the traveling waves on the rope?

(b) What is the wavelength of the waves?
(c) Write the equation of the traveling waves along the rope (ignoring the reflected waves from the far end).

8. A traveling wave on a string is described by the equation $y(x,t) = 0.1 \sin(25x - 500t)$.
(a) What are the wavelength, frequency, and amplitude of the wave?
(b) What is the wave velocity?
(c) If the mass density of the string is 0.001 kg/m, find the tension in the string.

9. A wave traveling on an elastic string has a 5 cm amplitude, a 25 cm wavelength, and a period of 0.01 s.
(a) Write an equation for the traveling wave $y(x,t)$ traveling in the positive x-direction.
(b) Find the wave speed.
(c) If the string is under a tension of 10 N, find the mass density of the string.

10. A 10 m elastic cord with a mass of 0.42 kg has its left end tied to a wall and is pulled with a force of 50 N at the right end. When the right end is vibrated vertically according to the equation $y = 0.04 \sin(2.5t)$, where y is in meters and t in seconds, write the equation for the wave traveling to the left.

11. A string is tied at one end to a fixed point and the other is attached to a 1 kg weight after passing over a frictionless pulley. The 4 m long string weighs 0.1 kg and the distance between the fixed point and the pulley is 3.5 m.
(a) Find the speed of transverse waves on the string.
(b) What is the fundamental frequency?
(c) What is the wavelength of the fourth harmonic?

12. Derive Equation (10.16) for the superposition of two equal amplitude traveling waves with a phase difference ϕ between them.

13. Two traveling waves with the same amplitude A, frequency f, and wavelength λ, but out of phase with each other by one quarter of a wavelength, are both traveling to the right and superpose in space. Find the amplitude, wavelength, and frequency of the resulting wave in terms of the given symbols. Write the equation of the resulting traveling wave $y(x, t)$.

14. A sinusoidal wave is traveling at 300 m/s along a string with a mass-per-unit-length of 0.002 kg/m. If the wave has an amplitude of 0.01 m and a wavelength of 0.05 m find the following.
(a) The equation for the traveling wave, $y(x,t)$
(b) The tension in the string
If a second identical traveling wave is on the same string but is shifted by 45° with respect to the first, find
(c) The net amplitude where the two waves overlap on the string
(d) The equation for the net traveling wave, $y(x,t)$

15. Standing waves are set up on a 1.5 m long string under tension and fixed at both ends. If the distance between nodes along the string is 0.25 m what is the wavelength of this mode and what harmonic is it?

16. A 3 m long string with a mass-per-unit-length of 0.005 kg/m is tied down at one end and has a 5 kg mass hanging over a pulley from the other end of the string putting it under tension. If standing waves are set up, find the frequency of the fundamental mode and of the fourth harmonic.

17. According to the Guinness book, the world's largest double bass instrument was 14 feet tall and had 4 strings (of equal length) totaling 104 feet in length. If the heaviest of these strings had a mass of 2 kg, find its fundamental frequency when under a tension of 5000 N. This sound would be felt but not heard.

11 Sound

Sound is one of our most important forms of communication. The science of sound is known as *acoustics.* In this chapter we learn about the physical properties of sound and how to describe sound in the language of waves. We study how sound can be produced in speech as well as musical instruments, and how our ear works to detect sound and transform its energy into electrical signals to be interpreted by our brain. Depending on the relative motion of the sound source and detector, the frequency of sound is changed according to the Doppler effect, studied next in this chapter. Ultrasound is simply sound at frequencies beyond the detection capabilities of our ears. It has a number of medical and scientific applications that we study, including ultrasonic imaging, routinely used for fetal monitoring and for imaging internal organs of the body.

1. BASICS

What happens when someone is speaking to you that enables you to hear them? The sound you hear is first generated by the person forcing a set of vocal chords in their larynx to vibrate while expelling air. The intonation and pitch are controlled by various muscles, the tongue, lips, and mouth. Sound emitted by the person then travels through the air to your ears where in a series of remarkable steps it is converted into an electrical signal that travels to the auditory center of your brain. We interpret sound to have several properties, including loudness, pitch, and tonal qualities or timbre, but what is sound, how does it travel through the air, and what physical qualities does it have that correspond to the properties just mentioned?

When vocal chords vibrate, they force molecules of air in the larynx to vibrate through collisions that periodically transfer momentum to the surrounding air (Figure 11.1). Consider a zone or band of air molecules in the vicinity of a vocal chord and let's follow those particular molecules through one oscillation in Figure 11.2. The vocal chord's motion to the right increases the local momentum of our neighboring band of molecules thus increasing the local pressure (in the figure we code the increased local momentum or pressure with a darker band). There is also a corresponding increase in the local density above the mean density as our molecules collide with those just to the right and a subsequent corresponding decrease in the local pressure and density below the mean of the band of molecules just to the left of the vocal chord. As momentum of our band on the right is transferred through collisions with neighboring molecules farther to the right and the vocal chord oscillates to the left, our band of molecules slows down, reducing its pressure and density, and a net restoring force to the left is applied from the pressure (and density) imbalance. Then, as the vocal chord moves again to the right, our molecules collide with others from the left that have been pushed to the right and this process repeats itself. Thus

J. Newman, *Physics of the Life Sciences*, DOI: 10.1007/978-0-387-77259-2_11,

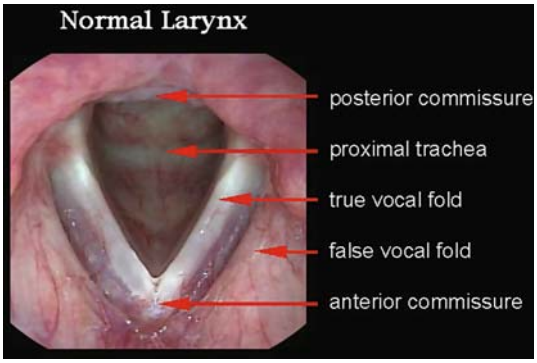


FIGURE 11.1 *The larynx, showing the vocal chords that vibrate to produce sounds.*

any particular molecule will oscillate longitudinally about some position and as a result, there is a local pressure and density variation in time at any point.

The local pressure and density adjacent to the vocal chord vary periodically, however, the collisions with neighboring molecules cause the pressure variation to propagate outward in space. Sound is this spatially periodic pressure (and density) longitudinal wave that travels outward from the source. In a band where the pressure is high, so is the density of the molecules and this pressure tends to push the molecules apart. Similarly in a band of lower density, the neighboring higher-pressure bands tend to restore the density and pressure toward their mean values. The air is said to be *compressed* and *rarefied* in a periodic manner. The centers of the bands of higher and lower density (and pressure) instantaneously have zero displacement because molecules from either side have moved either toward or away from them, respectively (Figure 11.3). These positions are called displacement *nodes*. Furthermore, the maximum displacements of the molecules, or *antinodes*, occur precisely at the bands of zero density variation located between those of high and low density extremes. This agrees with our discussion of the energy propagated along a traveling wave on a string in Chapter 10, where we showed that the maximum energy occurs at the displacement nodes where the slope of the string is greatest. For sound, the pressure nodes are the positions where the pressure equals atmospheric and there is no pressure (or density) variation. We can summarize the situation by stating that the displacement antinodes occur at the pressure nodes and the displacement nodes occur at the pressure antinodes. We return to this idea in our discussion of musical instruments in Section 4.

We can write the pressure variation from atmospheric pressure in the form

$$\Delta P = \Delta P_{\max} \sin(kx - \omega t), \tag{11.1}$$

where a positive value of ΔP corresponds to compression and a negative value to expansion and the other variables are just as defined in Chapter 10 in our discussion of traveling waves. There is a similar expression for the displacement of air molecules

$$\Delta s = \Delta s_{\max} \cos(kx - \omega t). \tag{11.2}$$

According to our previous discussion, points of maximum displacement correspond to points of zero pressure variation; the change from a sine to a cosine function accounts for this difference because when the sine is zero, the cosine function has an extreme value of ± 1 (see Figure 11.3). Values for $\Delta P_{\max}$ are usually very small fractions of the ambient pressure (the maximum value that does not cause pain to the ear is only 0.03% of atmospheric pressure) whereas values for $\Delta s_{\max}$ are extremely small (with a value of about 10 μm corresponding to the pain threshold just cited).

From our discussion, we might guess that the velocity of sound is related to the mean velocity of the molecules themselves and this is true in an ideal gas. The speed of sound, in general, depends on two parameters of the medium: its density ρ and a parameter of its elastic properties. For a fluid medium, the velocity of sound is given by

$$v = \sqrt{\frac{B}{\rho}}, \tag{11.3}$$

where B is the bulk modulus, the elastic constant of proportionality between the pressure variation and the resulting volume strain (see Figure 3.17 and its discussion). This equation has the same form as Equation (10.14) for the velocity of a mechanical wave on a string. There the tension serves as the elastic parameter and the linear mass density (mass/length) is the volume mass density analog. For a long solid rod, such as a railway track, the velocity of sound is given by a similar expression but with the elastic modulus E replacing the bulk modulus in Equation (11.3).

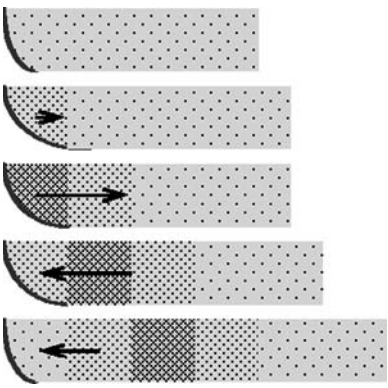


FIGURE 11.2 *Schematic of density variations in air emanating from a vibrating vocal chord over one oscillation. The arrows indicate the oscillatory velocity of the local molecules. These density oscillations comprise the sound wave and travel outward at the speed of sound.*

The speed of sound in air at 20°C and 1 atm pressure is 343 m/s (about 770 miles/hour). Aircraft that break the "sound barrier" fly faster than this speed, known as Mach 1. The Mach number is the ratio of the airspeed to the speed of sound. Beyond Mach 1, also known as supersonic speeds, a shock wave is created. This is a directed wave in which the gas density and pressure change dramatically as the wave passes.

Because the density of gases is dependent on temperature, the speed of sound in air actually increases approximately 0.6 m/s for each 1°C increase in temperature, as the density decreases. In liquids and solids, which are much less compressible or much "stiffer" than gases with correspondingly higher bulk or elastic moduli, the speed of sound is much faster. Table 11.1 lists the velocity of sound in various materials.

Table 11.1 *Densities and Velocities of Sound*

Material (20°C Unless Noted)	*Density (kg/m^3)*	*Speed (m/s)*
Air	1.20	343
Water	998	1,482
Seawater	1,025	1,522
Body tissue (37°C)	1,047	1,570
Glass (pyrex)	2,320	5,170

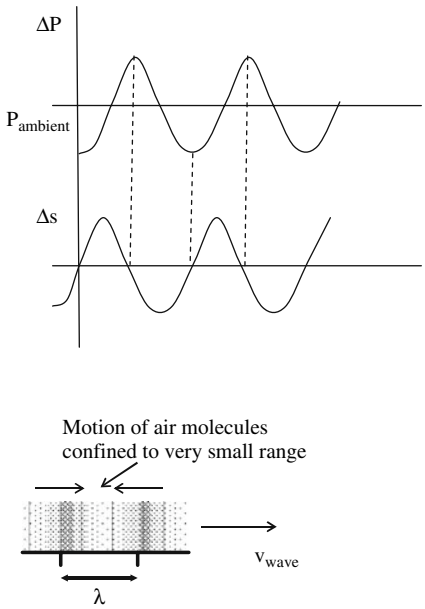


FIGURE 11.3 *Pressure or density variation along a sound wave in air. Zero displacements of air occur at the centers of the densest and least dense bands whereas maximum displacements occur where the density equals the mean density located midway between these bands.*

Frequency, wavelength, and intensity are other parameters characterizing sound. Audible sound corresponds to frequencies in the range of about 20–20,000 Hz. Lower frequencies than this are called infrasonic, whereas higher frequencies are called ultrasonic and are discussed later in this chapter. From the general relation $\lambda = v/f$, wavelengths of sound waves can range from cm to many meters. The *pitch* of sound is the audible sensation corresponding most closely to frequency; increasing frequency corresponds to increasing pitch.

Intensity represents the energy per unit time (or the power) crossing a unit surface area. Units for intensity are therefore given by J/s/m^2 or W/m^2. The intensity of sound is discussed in some detail in the next section. *Loudness* is the audible sensation corresponding most closely to intensity, although there is no direct relation. For example, at frequencies that are barely audible, a sound will not seem loud even if the intensity is quite large. We discuss loudness later in the chapter after discussing the ear and hearing.

2. INTENSITY OF SOUND

Sound is a longitudinal traveling wave that carries energy in the form of mechanical oscillations of the medium. For a one-dimensional longitudinal traveling wave, such as travels along an ideal spring as seen in Chapter 10 (where we neglect damping), the amplitude of the wave remains constant along its direction of travel. In this case, the energy per unit time, or power, traveling with the wave velocity is constant. The wave can be pictured as traveling along a fixed direction of propagation and represented as a *plane wave*, one having parallel wavefronts. These are the surfaces constructed by connecting all in phase points along the direction of propagation. For a one-dimensional wave, points of common phase are planes with normals along the wave velocity direction. Sound traveling along a railroad track is an example of such a one-dimensional sound wave, although there is some damping or attenuation of sound over large distances.

In three-dimensional examples, however, as the wave spreads out spatially, the energy crossing a unit cross-sectional area decreases with increasing distance from the sound source (see Figure 11.4). It is therefore more common to speak about the

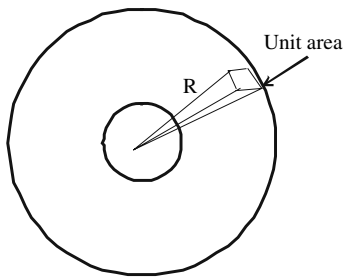


FIGURE 11.4 *The power radiated from a point source into the pyramid shown with vertex at the source is a constant, thus the power density, or intensity, must decrease according to Equation (11.4).*

intensity of a three-dimensional wave than about its power. In this case, if the sound originates at a localized source and flows outward in all directions, the wavefronts are spherical and their surface area increases with radius from the source as $A = 4\pi r^2$. If the power emitted by the source of sound is constant, then as the spherical wavefront travels outward, the total amount of energy crossing any spherical shell centered at the source is the same. Therefore the energy per unit time crossing a unit area must decrease at increasing distances from the source. Mathematically, the intensity of sound is related to the power P, generated by the source and the distance r from the source by

$$I = \frac{P}{A} = \frac{P}{4\pi r^2}. \tag{11.4}$$

If the power is constant then we see that the intensity is inversely proportional to the square of the distance from the source

$$I \propto \frac{1}{r^2}. \tag{11.5}$$

This is a general characteristic of spherical waves of any type and has only to do with the geometry of space.

For all waves, whether mechanical, sound, light, or any other type, the intensity I of the wave is proportional to the square of the wave amplitude. We know that this is true in the case of a spring because the total spring energy is

$$\frac{1}{2}kx_{\max}^2$$

and the intensity will therefore be proportional to $x_{\max}^2$. In the case of sound, the intensity is given by

$$I = \frac{\Delta P_{\max}^2}{2\rho v}, \tag{11.6}$$

where the intensity, pressure wave amplitude, and density values all refer to the same spatial location. This expression can also be shown to be proportional to the square of the amplitude of vibration of the medium, $\Delta s_{\max}$. Recall from the last section that these amplitudes are very small with typical $\Delta P_{\max}/P_{\text{atm}}$ and $\Delta s_{\max}$ values of under a few percent and submicrometer distances, respectively.

Sound intensities vary over an enormous range. The least intense sound that can be heard by the human ear is called the threshold of hearing and is taken as 10^{-12} W/m^2. Of course, this value actually varies from person to person as well as with a person's age. As the intensity increases so does the perceived loudness. The most intense sound that the human ear can respond to without harm is called the threshold of pain and is taken as 1 W/m^2. Because of the enormous range of intensities to which the ear responds, 12 orders of magnitude, sounds that are 10 times more intense do not seem 10 times as loud to the ear. In fact, the ear responds nearly logarithmically to sound intensity, the sound loudness doubling for each decade increase in intensity. A useful scale for intensity level is the decibel scale for which the sound intensity level β is given by

$$\beta = (10\ \text{d}B)\log\frac{I}{I_o}, \tag{11.7}$$

where the logarithm is the common logarithm, with base 10, I_0 is a reference intensity, taken as the threshold of hearing (10^{-12} W/m^2), and the unit of sound intensity is the decibel or dB (where 1 dB = 1/10 bel, named in honor of Alexander Graham Bell). The scale is chosen so that at $I = I_0$ the intensity level is 0 dB, whereas at the threshold of pain, $I = 10^{12}\, I_0$, the intensity level is 120 dB (check this by substitution

in Equation (11.7)). Table 11.2 gives examples of various sounds and their corresponding intensity levels. We return to a discussion of the response of the ear to sound intensity in Section 5 below.

Table 11.2 *Intensities of Sounds*

Sound	*Intensity (W/m^2)*	*Intensity Level (dB)*
Threshold of hearing	10^{-12}	0
Whisper	10^{-10}	20
Normal conversation (at 1 m)	10^{-6}	60
Street traffic in major city	10^{-5}	70
Live rock concert	10^{-1}	110
Threshold of pain	1	120
Jet engine (at 30 m)	10	130
Rupture of eardrum	10^{4}	160

Example 11.1 Find the ratio of the intensity of two sounds that differ by 3 dB.

Solution: Let the two intensities be I_1 and I_2. According to Equation (11.7), the two sounds have dB given by $\beta_1 = 10 \log I_1/I_0$ and $\beta_2 = 10 \log I_2/I_0$, so that if the two sounds differ by 3 dB, we have that

$$\beta_2 - \beta_1 = 3\ dB = \left(10 \log \frac{I_2}{I_0} - 10 \log \frac{I_1}{I_0}\right) =$$

$$(10 \log I_2 - 10 \log I_0 - 10 \log I_1 + 10 \log I_0) = 10 \log \frac{I_2}{I_1}.$$

Solving for the ratio of the intensities, we find $I_2/I_1 = 10^{0.3} = 2.0$. Any two sounds differing by 3 dB have intensities that differ by a factor of two. The best human ears can hear a difference in loudness corresponding to about 1 dB. To what ratio of intensities does this correspond?

3. SUPERPOSITION OF SOUND WAVES

REFLECTION, REFRACTION, AND DIFFRACTION

When sound waves traveling in more than one dimension come to a boundary between two different media, additional considerations beyond what we have seen in the last chapter are required. Consider the case of a plane boundary between two different media and let's imagine a sound wave traveling through one medium and impinging on the boundary. Let's take the wave to be a plane wave, with all points along a plane wavefront in phase, an often-used idealized wave that is traveling in synchrony in a particular direction. The wavefronts are drawn perpendicular to the propagation direction as shown in Figure 11.5. When this wave meets the boundary, as in the case of waves on a string, there will be a reflected wave as well as a transmitted wave. If the wave approaches the boundary along the perpendicular, or normal, to the planar boundary, then the reflected and transmitted waves will remain along that direction and the problem is quite similar to the one-dimensional case of waves on a string.

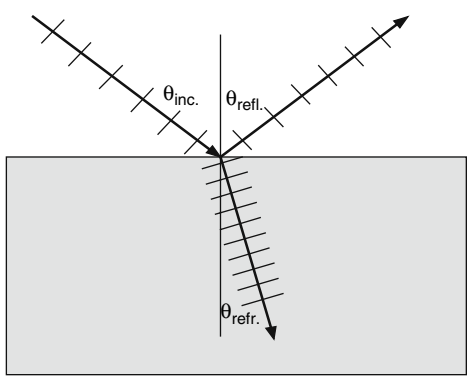


FIGURE 11.5 *Reflection and refraction of an incident plane wave at a planar boundary between two different media.*

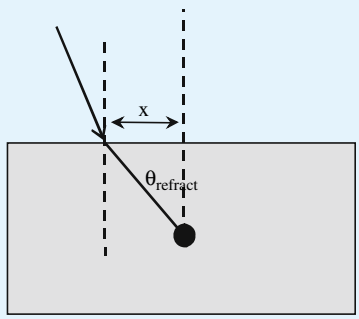
If the wave approaches the boundary along a line making an angle θ_{incident} with the normal to the planar boundary then the reflected and transmitted waves do not travel along the same line. In such a case the reflected wave remains in the incident medium, remains a plane wave, and propagates in a direction making an angle $\theta_{\text{reflection}}$ with the boundary normal that is equal to the incident angle as shown in Figure 11.5. The incident wave, reflected wave, and normal to the surface all lie in a common plane, known as the plane of incidence. These two sentences comprise a statement of the *law of reflection*: the reflected wave lies in the incidence plane at an angle of reflection equal to the incident angle. When we study sound further and optics later on we show some consequences of this law for acoustic and light waves. Although seemingly simple, this law is fundamental to ultrasonic imaging, the functioning of mirrors, the imaging of x-rays, and a wide variety of applications in optics.

The transmitted wave enters the second medium but is deviated from the original propagation direction. Due to the different speed of the wave in the second medium, the wavelength (but not the frequency) is changed and the wave direction is bent or *refracted* (Figure 11.5). The angle of refraction, or the angle between the direction the transmitted wave travels and the normal to the surface, can be related to the incident angle and the ratio of wave velocities in the two media by

$$\frac{\sin\theta_{\text{incident}}}{\sin\theta_{\text{refracted}}} = \frac{v_{\text{incident}}}{v_{\text{refracted}}} \tag{11.8}$$

which is known as the *law of refraction.*

Example 11.2 An ultrasonic wave is incident on a person's abdomen at a 20° angle of incidence. Where should it be directed so as to hit a kidney stone located 7 cm beneath the surface as shown in the figure? The ultrasonic waves are emitted directly into an aqueous gel coating the abdomen. Take the speed of sound in the gel to be $v_{\text{gel}} = 1400$ m/s and in body tissue $v_{\text{tissue}} = 1570$ m/s, and specify the location in terms of the transverse distance x from the normal to the surface going through the kidney stone.



Solution: The wave entering the abdomen tissue will refract at the surface entering at an angle of refraction given by

$$\sin\theta_{\text{refract}} = \sin\theta_{\text{inc}}\left(\frac{v_{\text{tissue}}}{v_{\text{get}}}\right) = \sin 20\left(\frac{1570}{1400}\right) = 0.38,$$

so that $\theta_{\text{refract}} = 22.6°$. To then hit the kidney stone 7 cm beneath the surface, we must have that $\tan\theta_{\text{refract}} = x/(7\text{ cm})$, so that $x = 2.9$ cm along the surface from the normal. Note that without making the correction for refraction the distance x would be $7(\tan 20°) = 2.5$ cm, and the wave would probably miss the kidney stone.

FIGURE 11.6 *Diffraction of water waves around obstacles. Ripples spreading out from bottom center diffract around rocks and are seen in their "shadow" region.*

One other general property of waves should be briefly mentioned here. When a wave meets either an obstacle or a hole in a reflecting boundary, it spreads out behind the obstacle or hole into the "shadow" region (Figure 11.6). The extent of this *diffraction*, or bending, of the wave depends on the wavelength of the wave relative to the size of the obstacle or hole. If the physical dimensions of the object are much larger than the wavelength then there will be little diffraction of the wave but if the object is comparable or smaller than the wavelength there can be dramatic spreading of a wave around an obstacle or behind the edges of a hole. When we study optics we show that diffraction sets fundamental limits on our ability to "see" microscopic objects.

TEMPORAL SUPERPOSITION

Up until now we have been discussing sound as if it were of a single frequency, as in Equation (11.1). Almost all of the sounds we hear cannot be described in such simple terms, but can be thought of as the superposition of a variety of pure sine waves each of a different frequency and amplitude. Figure 11.7a shows a time record of the amplitude of vibration of air for a relatively simple sound. An analysis of this sound record (waveform) is usually presented in the form of a *spectrum*, in which the amplitudes of the different frequency components are plotted as a function of the frequency (Figure 11.7b). In simple cases there will be a small number of discrete frequency components present, as in our example in which there are four components. These are the resonant frequencies of the sound source. As we discussed in Chapter 10, the lowest frequency is called the *fundamental* whereas often the other frequency components in the spectrum will be integral multiples of the fundamental and are known as *harmonics.*

The mathematics involved in the superposition of harmonics of varying amplitude is known as *Fourier series* and is illustrated in Figure 11.8 for the example of the previous figure. The four different sine curves, with relative amplitudes and frequencies given by the spectrum in Figure 11.7b, add together to reproduce the sound waveform of Figure 11.7a. In fact, any periodic waveform, no matter how complex, can be represented as the superposition of harmonics according to Fourier's theorem.

Musical sounds are characterized by spectra that are constant over periods of time of at least fractions of a second, the duration of the musical notes being played. The waveform of a musical sound is therefore repetitive over at least that time interval. Noise, on the other hand, is characterized by a chaotic frequency spectrum that

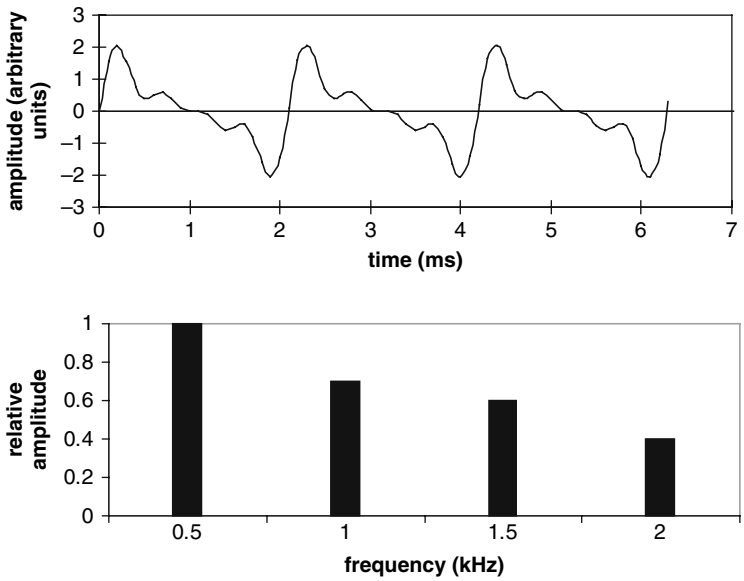


FIGURE 11.7 *(a) Amplitude versus time for a simple sound. (b) Spectrum of frequency components for the sound in (a).*

changes rapidly with time and is nonrepetitive. Example spectra from complex music and from noise are shown in Figure 11.9. Each musical instrument has its own unique spectral tone that accompanies the playing of any particular note. Detailed analysis of the Fourier composition of these tones from different instruments has led to digital synthesizers that can mimic the sounds from a large variety of musical instruments with high quality. For each note played by these "computers" to mimic an instrument, the appropriate set of overtones is added to give the proper tone quality for that particular instrument. The analysis and synthesis of musical tones has progressed to the point where some digital synthesizers can actually give better tone quality than even moderately priced individual instruments.

Let's examine the particularly simple case of the temporal (time) superposition of two pure tones of the same amplitude that are relatively close together in frequency. What will we hear if this occurs? We show just below that we'll hear a sound at the average of the two frequencies that has an intensity that varies slowly in time in a whining fashion. The tone of the sound does not change but the intensity oscillates at the difference, or *beat*, frequency resulting in a slow repetitive whine as briefly discussed in Section 4 of Chapter 10 (see Figure 10.15).

If we listen to these two sounds at the same spatial location, we can write expressions for the time variation of their amplitudes as

$$y_1 = A \sin \omega_1 t \quad \text{and} \quad y_2 = A \sin \omega_2 t. \tag{11.9}$$

Superposition of these two sounds results in a time-varying signal given by

$$y = y_1 + y_2 = A(\sin \omega_1 t + \sin \omega_2 t). \tag{11.10}$$

By using the same trigonometric identity previously used to get Equation (10.16),

$$\sin \theta + \sin \varphi = 2 \cos \frac{1}{2}(\theta - \varphi) \sin \frac{1}{2}(\theta + \varphi),$$

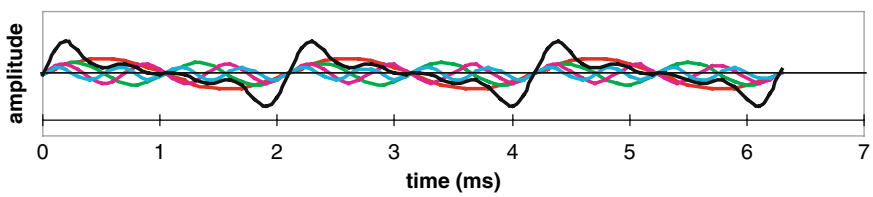


FIGURE 11.8 *The waveform from Figure 11.7, in black, is the sum of the four colored sine curves with frequencies and amplitudes from Figure 11.7b shown in this Fourier series addition.*

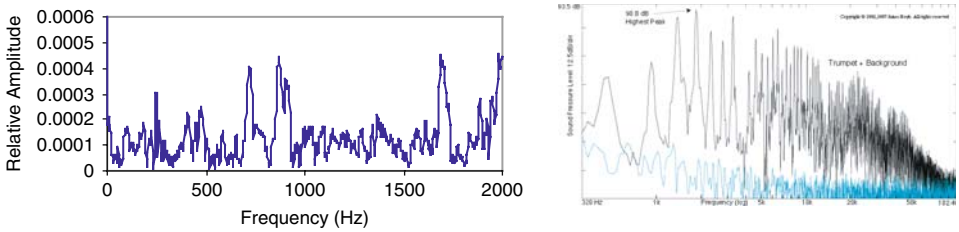


FIGURE 11.9 (left) *Noise frequency spectrum from hitting a table with a plastic ruler;* (right) *black curve is spectrum from a trumpet.*

we can rewrite Equation (11.10) as

$$y = \left[2A\cos\left(\frac{\omega_1 - \omega_2}{2}\right)t\right]\sin\left(\frac{\omega_1 + \omega_2}{2}\right)t. \tag{11.11}$$

If the two angular frequencies are nearly equal, then the average value (in the second term) is approximately equal to each original frequency, whereas the difference term has a much lower frequency, close to zero. We can think of this as resulting in a time-varying amplitude prefactor multiplying a sine term with angular frequency equal to the average

$$y = [2A\cos\Delta\omega t]\sin\varpi t, \tag{11.12}$$

where $\Delta\omega = (\omega_1 - \omega_2)/2$ and $\varpi = (\omega_1 + \omega_2)/2$ and the square bracket emphasizes that this term is a more slowly varying amplitude. Because the intensity is proportional to the square of this amplitude, a beat, or maximum sound, will occur when $\cos\Delta\omega t$ is equal to either 1 or -1. This occurs at an angular frequency of twice $\Delta\omega$ or at $\omega_1 - \omega_2$. The corresponding beat frequency is

$$f_{\text{beat}} = f_1 - f_2, \tag{11.13}$$

and it is at this frequency that one hears the loudness pulsate. Listening to beats is a commonly used method of tuning musical instruments. Using calibrated standard tones, the instrument is adjusted to make the beat frequency as long as possible, eventually disappearing when the two tones have matched frequencies.

Example 11.3 Suppose that two small speakers each play a pure tone. If one speaker emits a frequency of 1000 Hz and you hear a beat frequency of 5 Hz, what is the wavelength difference between the two tones?

Solution: The frequency of the second tone is either 1005 or 995 Hz, both of which would produce 5 beats/s. Using the speed of sound in air from Table 11.1, the wavelength of the first tone is (343/1000) = 0.343 m. The second tone has a wavelength of either (343/1005) = 0.341 m, or (343/995) = 0.345 m, both giving a wavelength difference of about 2 mm.

SPATIAL SUPERPOSITION

After having examined the superposition of two different frequency sound waves, we now turn to the situation when two sounds, produced at different locations, combine at some point in space. In this case we can write the two sound waves in one dimension as

$$y_1 = A_1\sin(kx - \omega t + \varphi_1) \text{ and } y_2 = A_2\sin(kx - \omega t + \varphi_2),$$

where A_1 and A_2 are the amplitudes, φ_1 and φ_2 the phase angles, and k and ω, as usual, are given by $k = 2\pi/\lambda$ and $\omega = 2\pi/T = 2\pi f$, and we have assumed the two sounds have the same frequency and wavelength. The phase angles account for the relative shift of the sine curves with respect to the origin of coordinates because in general the two waves originated at different locations with different phases. Setting $x = 0$ in the expressions for y_1 and y_2, the phase angles are seen to determine the amplitudes at a given time at the origin and thereby at any other point x. At a point where these two sound waves overlap the net amplitude is simply the sum of the individual amplitudes and the intensity is proportional to the square of those amplitudes. To simplify the problem, suppose that the two amplitudes are also equal to each other (we have considered a similar problem in Section 4 of Chapter 10 for waves on a string). Then using a similar argument that lead to Equation (11.11) above, we can write that

$$y_{\text{net}} = y_1 + y_2 = 2A\left[\cos\left(\frac{\varphi_1 - \varphi_2}{2}\right)\right]\sin(kx - \omega t). \tag{11.14}$$

We see that when these two sounds combine at a point in space, the net amplitude depends on the relative phases of the two waves. If the two waves have some definite phase relationship that remains constant in time (i.e., the phase angles φ_1 and φ_2 are constants), the two waves are said to be spatially *coherent* and exhibit *interference.* At each point in space, if the two sine waves are "in phase", meaning they have zero phase difference, then because $\cos(0) = 1$, the net amplitude is $2A$, just as you would expect when two identical sine curves exactly overlap in space (Figure 11.10). This is known as *constructive interference.* If the two sine waves are out of phase by 180°, or π radians, then because $\cos(90°) = 0$, the two waves exactly cancel, again just as expected if the waves are shifted with respect to each other by half a wavelength. This is known as *total destructive interference.* At any intermediate situation Equation (11.14) gives the net amplitude and there will be some intermediate situation with the amplitude in general lying between 0 and $2A$.

Because the intensity is proportional to the square of the amplitude, the intensity of the combined sound wave will be between 0 and $4I$, where I is the intensity of each of the two sounds. This should seem strange at first glance because the intensity is a measure of the energy carried by the sound wave, and energy must be conserved. So if each wave carries an intensity I, how can the sum ever be larger than $2I$? What's going on here? It is clear that if the intensity of the combined sound wave is averaged over a large region of space that the average intensity must be $2I$, since each sound wave carries intensity I. The phenomenon of interference leads to a redistribution of the energy, concentrating it in some regions and depleting it in others, depending on

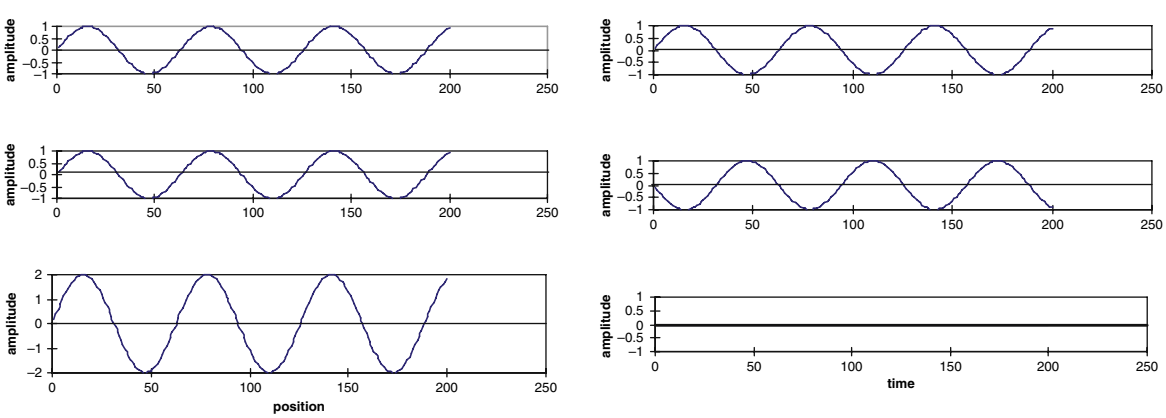


FIGURE 11.10 *Interference between two waves.* (left) *Two in phase waves, with their constructive interference superposition at* bottom; (right) *two equal amplitude out-of-phase waves showing complete destructive interference when added together at* bottom.

the phase relationship of the waves; maxima have intensity $4I$, but minima have zero intensity.

Although we have limited our discussion to one-dimensional waves, real sound waves travel in real space. In Figure 11.11 we show two experimental measurements of the superposition of two waves emanating from "point sources" and traveling radially outward. On the top is a photo of the surface ripples in a water tank and the measurement on the bottom using NMR techniques is sensitive to the local density and shows an image of sound waves traveling through a material simulating human tissue. We show later how this methodology can be used to image inside the human body. In the last section of this chapter we return to take a further look at imaging inside human tissue with ultrasound.

Another example of interference effects in three dimensions involves designing a musical auditorium or concert hall where the phenomenon of interference can lead to disasters. Because sounds reverberate off walls as well as travel directly out to someone in the audience, the listener hears the superposition of a complex collection of sound waves. Depending on the phase relationships of the different sound waves, there can be "dead spots" in an auditorium where there is significant destructive interference. Special baffles as well as ceiling and wall designs and materials are used to reduce direct reflections in order to avoid this problem.

We return to the very important and general phenomenon of interference when we discuss other types of waves, including light and also matter waves in our discussion of quantum mechanics.

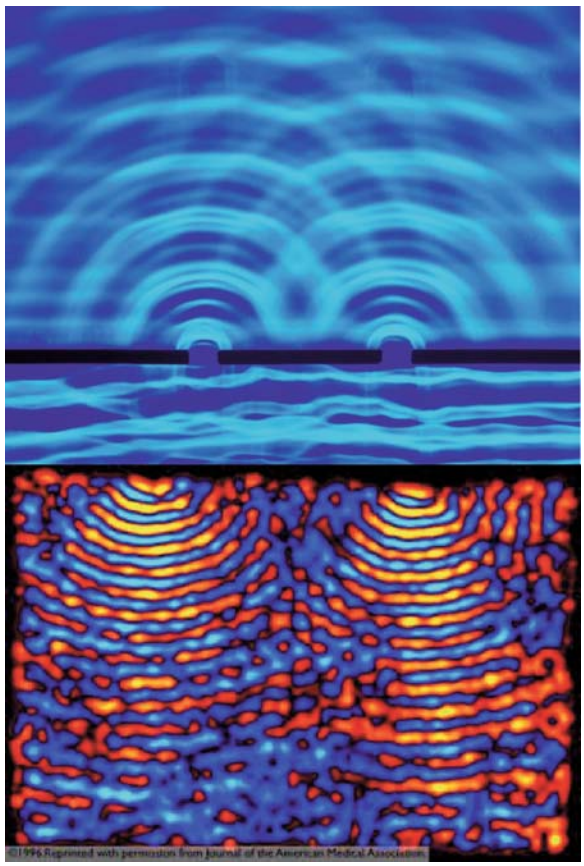


FIGURE 11.11 (top) *Interference of ripples of water waves in a tank;* (bottom) *magnetic resonance techniques used to image the interference between two sound waves inside a material medium from "point" sources at the top. Note the similarities.*

4. PRODUCING SOUND

Aside from incidental sounds generated from chemical or other forms of energy, such as the crackling of a campfire or the noise when a branch of a tree falls (even in a forest with no one around), the production of sound usually involves two requirements: a way to generate mechanical vibrations and a resonant cavity structure to amplify and "shape" the sound. Here we discuss the generation of music from a variety of instrument types. Each of these generates mechanical vibrations of a string, wire, or drumhead (as in stringed instruments, pianos, or drums, respectively), or of the air directly by vibrations of a reed (woodwinds) or the lips (brasses). The music generated then acquires its tone and quality from a resonant cavity such as the hollow wooden body of a stringed instrument or the tube of a woodwind or brass instrument. A loudspeaker produces sound by converting an electrical signal into mechanical vibrations of a diaphragm. The mechanism for this conversion is the electromagnetic force, discussed later, used to vibrate the diaphragm. In this case the shape and design of the diaphragm help to amplify and direct the sound.

Let's first review the generation of sound by a string held under tension, discussed in Section 5 of Chapter 10, as a model for a stringed instrument such as a violin. Excitation by plucking or bowing the string results in standing waves. The fundamental frequency is determined by the requirement of nodes at only both fixed ends of the string so that the fundamental wavelength is twice the string length yielding

$$f_1 = \frac{v}{2L}, \qquad \text{string}, \tag{11.15}$$

where v is the wave speed and L is the string length between fixed points. Recall that the wave speed on a string is given by

$$v = \sqrt{\frac{T}{m/L}},$$

FIGURE 11.12 *Examples of simple standing wave patterns on the back-plate of a violin. The dark lines, formed by black sand, represent nodal lines where the wood does not vibrate.*

where T is the tension in the string and m/L is its mass per unit length. In a violin, the four strings each have a different mass per unit length and the tensions are adjusted to tune the fundamental frequency appropriately. Recall also that the harmonics are given as integral multiples of the fundamental frequency. When a string on a violin is played, not only does the string vibrate, so does the entire volume of air within the wooden cavity as well as the wood itself. These vibrations not only help to amplify the sound by more effectively causing the air to vibrate, but also add depth and quality to the sound. Figure 11.12 shows two examples of simple vibration patterns of a violin. In general the standing wave patterns of the wood of the violin can be quite complicated.

Wind and brass instruments have a resonant tube that serves to amplify only those frequencies that produce a standing wave pattern. There are two main configurations that occur in different musical instruments: tubes with two open ends, such as in a flute (Figure 11.13a) or organ pipe, where the blowhole serves as an open end, and tubes with one open and one closed end, such as a trumpet or trombone, where the lips act as a closed end. Figure 11.13b shows a simple schematic of both cases.

The conditions at the tube ends, known as the boundary conditions, are what determine the nature of the standing waves produced. At a closed end, because air is not able to oscillate longitudinally due to the wall, there must be a node of displacement and the sound is completely reflected, neglecting losses. At the open end, the sound wave is partially reflected and partially transmitted out of the resonant tube. Although it is less obvious, there must be a displacement antinode at the open end. We can see this by first observing that because atmospheric pressure outside the tube serves to maintain a constant pressure at the open end, there must be a node of pressure variation there. Any increase or decrease from atmospheric pressure at the open end is immediately compensated for by bulk flow of outside air to maintain a constant pressure node. As discussed in Section 1, positions of pressure nodes correspond to displacement antinodes, and so we see that the proper boundary condition at a tube open end is a displacement antinode.

From these boundary conditions it is straightforward to detail the fundamental and harmonic frequencies allowed for each configuration of a resonant tube. For tubes that are open at both ends, the fundamental resonant mode has a displacement antinode at each end so that half of one wavelength corresponds to the tube length L (see Figure 11.14a). Therefore the fundamental wavelength is $2L$ and the fundamental frequency is $v/2L$. Each higher harmonic adds an additional node giving a set of resonant mode wavelengths

$$\lambda_n = \frac{2L}{n}, \quad n = 1, 2, 3, \ldots, \tag{11.16}$$

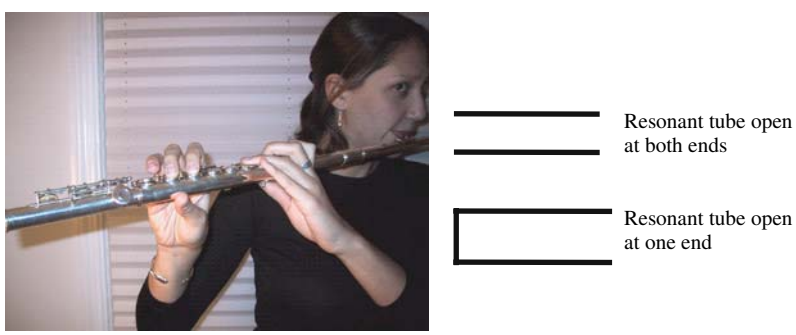


FIGURE 11.13 (left) *Emily playing a flute as a resonant tube;* (right) *simple models for wind and brass instruments.*

where the integer n is the harmonic number. Corresponding to these wavelengths are the resonant frequencies of the open tube

$$f_n = \frac{v}{\lambda_n} = \frac{nv}{2L} = nf_1, \quad n = 1, 2, 3, \ldots \quad \text{(open tube)}, \tag{11.17}$$

where v is the speed of sound.

For tubes that are open at one end and closed at the other, the fundamental has an antinode at the open end and a node at the closed end so that only 1/4 wave fits in the tube length L (see Figure 11.14b). Therefore the fundamental wavelength is equal to $4L$. Each higher harmonic adds one additional node within the tube giving a set of resonant wavelengths

$$\lambda_n = \frac{4L}{n}, \quad n = 1, 3, 5 \ldots, \tag{11.18}$$

where in this case only odd harmonics are present. The corresponding resonant frequencies in this case are

one side closed tube $$f_n = \frac{v}{\lambda_n} = \frac{nv}{4L} = nf_1. \qquad n = 1, 3, 5. \ldots \tag{11.19}$$

We see that for a tube closed at one end, only the odd harmonics are present. The differences in each of these cases (as well as those of resonant modes on a string) are due to the different boundary conditions.

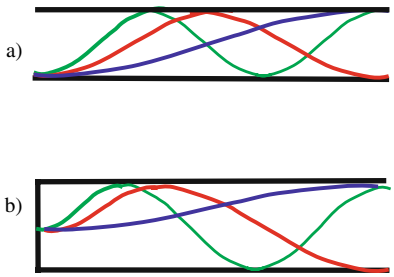


FIGURE 11.14 *The first three resonant modes of (a) a tube open at both ends: n = 1 blue; n = 2 red; n = 3 green; and (b) a tube open at one end: n = 1 blue; n = 3 red; n = 5 green.*

Example 11.4 Compare the resonant frequencies from two tubes, one open at both ends with twice the length of the second one which is closed at one end. Will they have the same fundamental and harmonics?

Solution: For the open tube the resonant frequencies are given by

$$f_n = n\frac{v}{2L_o}, \quad n = 1, 2, 3, \ldots,$$

whereas the tube closed at one end will have resonant frequencies given by

$$f_n = n\frac{v}{4L_c}, \quad n = 1, 3, 5, \ldots.$$

Because $L_o = 2\ L_c$, the fundamental frequencies (for $n = 1$) will be the same for the two tubes. However, notice that the closed tube will be missing every other harmonic that the open tube will have, although the common frequencies will match.

For a circular drumhead, the standing wave patterns observed when the drumhead is made to vibrate are two-dimensional and arise from the condition that there must be a node at the fixed circular boundary. The fundamental has a single antinode at the center of the drumhead so that the entire membrane oscillates together. Higher-order modes of vibration include a variety of interesting patterns, some of which are shown in Figure 11.15.

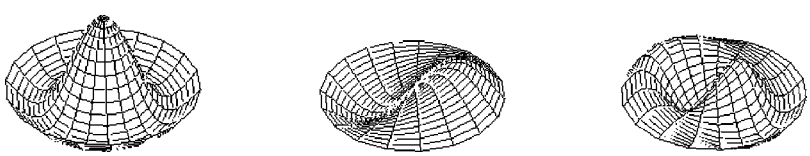


FIGURE 11.15 *Examples of modes of vibration of a circular drumhead.*

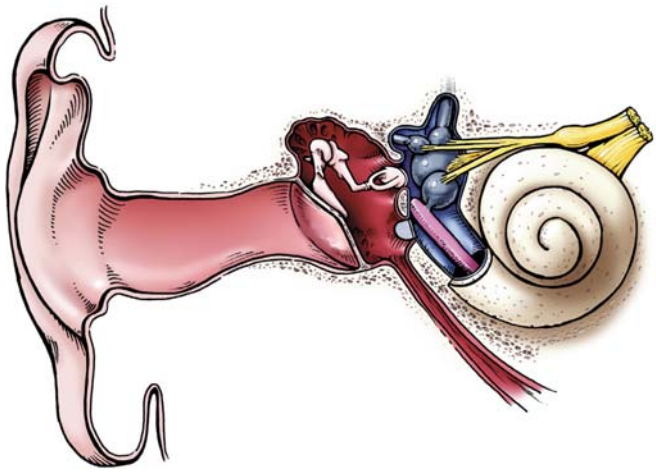


FIGURE 11.16 *Overall structure of the ear.*

5. THE HUMAN EAR: PHYSIOLOGY AND FUNCTION

Hearing is one of the primary sensory systems in man as well as in many animals. It gives us information about our surroundings, allows for oral communication, and gives us pleasure in listening to music. Although hearing is one of the earliest biophysical systems studied, until quite recently there was surprisingly little known about the fundamental physical processes involved. This is due, in part, to the extremely complex and nonlinear nature of these processes and also to the location of the ear within the skull in close proximity to the brain, making it difficult to study in detail while intact and functioning normally. Here we summarize the important features and functions of the various portions of the ear.

The ear is composed of three sections, the outer (or external), middle, and inner ear, each of which has a specific purpose in the transduction of sound from a pressure wave in the air to an electrical signal that is interpreted as sound by the brain (Figure 11.16). The outer ear consists of the external *pinna* and the outer *auditory canal* that ends at the *tympanic membrane* (or ear drum). In the air-filled middle ear lie the three tiny bones, the *ossicles*, known as the *malleus* (hammer), *incus* (anvil), and *stapes* (stirrup) already introduced in Section 2 of Chapter 8 in connection with the hydraulic effect. The middle ear is bounded by the tympanic membrane on the outer side and the oval window on the inner side. There is also a connection, through the round window to the Eustachian tube that connects with the pharynx. This is important in equalizing pressure between the middle and outer ear and can lead to painful infections when clogged. Beyond the oval window lies the inner ear, a complex multichambered cavity that contains both the *semicircular canals* involved in balance (but not in hearing) and the *cochlea*, the transduction center of hearing.

OUTER EAR

Serving two functions, the outer ear amplifies sound and protects the delicate tympanic membrane. Protection is accomplished by providing a narrow (~0.75 cm diameter) long (~2.5 cm) tube or ear canal, lined with hairs and wax-secreting cells. In many animals the pinnae can be directed at the source of sound and can help not only to increase sensitivity to sounds but also to locate their source. In humans the pinnae serve no known purpose other than wiggling to make people laugh.

Amplification occurs because the ear canal serves as a resonator. Recall that a tube with one closed and one open end has a fundamental resonant wavelength equal to four times the tube length. If we approximate the ear canal as such a tube, we find that the resonant wavelength is about 10 cm, corresponding to a frequency of 3430 Hz (using the velocity of sound in air as 343 m/s). In fact our ears are most sensitive near this frequency as discussed later. Although the closed end of the ear canal, the tympanic membrane, is fairly thick (~0.1 mm) and stiff, both it and the walls of the ear canal are elastic and there is not a sharp resonance, but a broad resonance spanning about three octaves (frequency doublings) with a peak at about 3300 Hz. Typically sound in the range from 1.5 kHz to 7 kHz is amplified by about 10–15 dB (a factor of 10–30) by the outer ear.

As we show in the next section, sound in air does not penetrate water very well. Just think of how quiet it gets when you submerge your head under water in a bath or when swimming. Over 99.9% of the sound energy traveling in air is reflected from water. How then does sound, traveling in air, enter the cochlea, a fluid-filled tiny coiled structure, in order for us to hear?

MIDDLE EAR

The middle ear functions to efficiently transmit and amplify sound from the vibrating tympanic membrane (ear drum) to the oval window at the entrance to the cochlea. The ossicles are suspended by a set of ligaments and muscles so that the malleus is in close proximity to the tympanic membrane, and the "footplate" of the stapes is in the oval window, basically a hole in the bone surrounding the inner ear (see Figure 11.17). Fluctuating pressure differences between the outer and middle ear will cause the tympanic membrane to vibrate. (Excess pressure within the middle ear is relieved via the Eustachian tube. When in a rapidly descending airplane, the pressure buildup in the middle ear can be painful and can even cause a temporary hearing loss. A similar pressure increase can occur in an infected ear.) The ossicles provide a transmission and amplification mechanism in two basic ways.

First, there is some "lever action" of the mechanical force transmission from the malleus to the stapes, providing roughly a 30% increase in the force. In addition, there is a large (~17-fold) reduction in area from that of the tympanic membrane to that of the portion of the stapes in contact with the oval window. This reduction in area results in a similar phenomenon to "hydraulic pressure" with an increase in pressure. The ratio of the pressure at the oval window to that at the tympanic membrane is given by

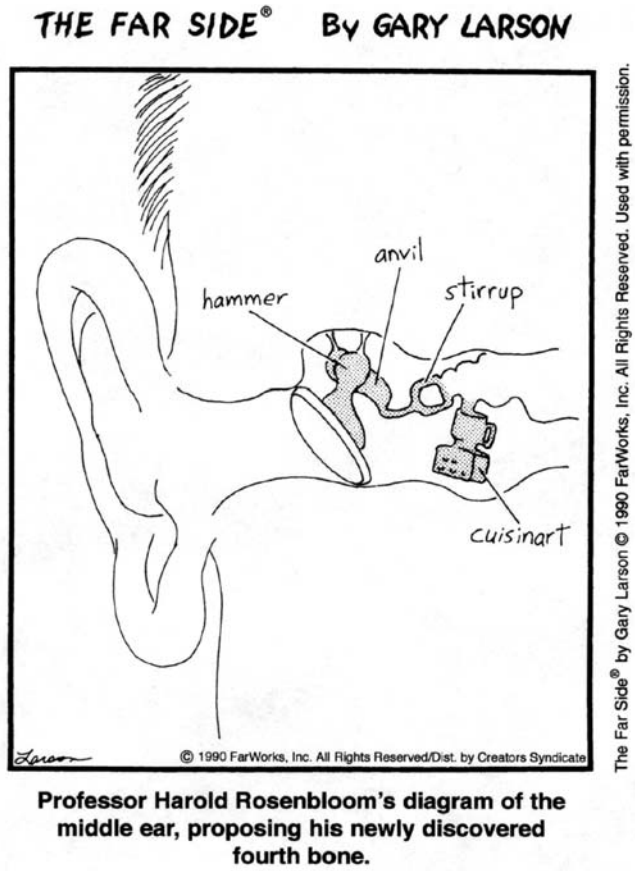


FIGURE 11.17 *The middle ear (see also Figure 8.3).*

$$\frac{P_{\text{oval}}}{P_{\text{tymp}}} = \frac{\left(\frac{F_{\text{oval}}}{A_{\text{oval}}}\right)}{\left(\frac{F_{\text{tymp}}}{A_{\text{tymp}}}\right)} = \frac{F_{\text{oval}}}{F_{\text{tymp}}}\frac{A_{\text{tymp}}}{A_{\text{oval}}} = (1.3)(17) = 22. \qquad (11.20)$$

Thus, the overall theoretical pressure amplification (ignoring damping losses) of this simple model is about a factor of 22, comparing quite well with the actual experimental value of about 17. The middle ear effectively changes the larger amplitude, smaller pressure vibrations of the tympanic membrane to smaller amplitude, larger pressure vibrations at the oval window. This is precisely what is needed in order to effectively couple the sound waves into the fluid of the cochlea. The middle ear is said to act as an impedance matching system (see the next section), allowing the maximum transmission of energy.

FIGURE 11.18 *The cochlea of the inner ear.*

INNER EAR

It is the cochlea of the inner ear that converts sound energy into an electrical signal sent via the auditory nerve to the auditory centers of the brain for interpretation. Humans can hear without a tympanic membrane and without ossicles, although there is significant loss of hearing under these conditions, but the cochlea has been thought to be essential for hearing. Recent cochlear implants have had some success in direct coupling to auditory nerves. Each inner ear is actually a cavity in the temporal bone (the hardest bone in the body) with six independent sensory organs (Figure 11.18): there are two detectors of linear acceleration, the saccule (mainly detecting vertical accelerations) and utricle (mainly detecting horizontal accelerations); three

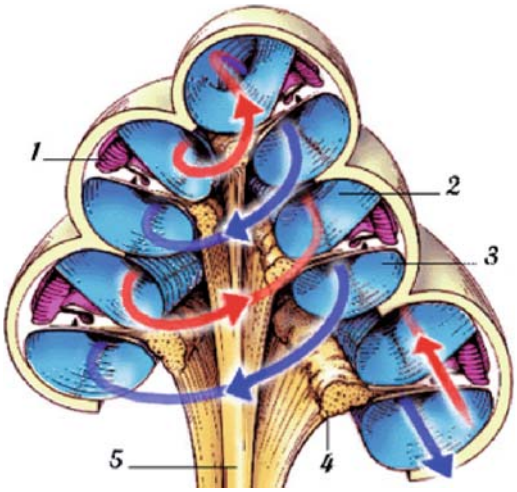


FIGURE 11.19 *A cross-section of the cochlea showing the three parallel ducts that spiral around the organ.*

semicircular canals, each monitoring angular acceleration about a different orthogonal axis and aiding in maintaining balance; and the cochlea, a fluid-filled, snail-shaped cavity with three turns having a total length of about 35 mm and ending in a closed apex. All of these detectors function in essentially the same way. Each contains hair cells that are mechanically sensitive and serve as the basic transducers, converting mechanical forces, due to accelerations or sound waves, into electrical signals.

Along the cochlea there are three parallel ducts filled with fluid (Figure 11.19). The total fluid volume is about 15 μl, roughly a drop of water. The *basilar membrane* separates two of these, the scala tympani and the scala media, or cochlear duct, and is the site of the organ of Corti where the hair cells are located and the transduction occurs. The third, the scala vestibuli, is separated from the cochlear duct by Reisner's membrane and connects with the scala tympani at the apex through a small opening.

If we imagine the cochlea to be unwound and examine a detail of the organ of Corti (Figure 11.20), all of the "action" occurs between the basilar and tectorial membranes along the length of the cochlea. There are about 16,000 hair cells in this region, each of which has a hair bundle, composed of about 50–100 stereocilia projecting from their apex into the surrounding fluid in precise geometric patterns. Each stereocilia is a thin (0.2 μm) rigid cylinder composed of cross-linked actin filaments that are arranged to increase uniformly in length from about 4 μm at the stapes end to about 8 μm at the apex end of the cochlea (Figure 11.21). The stereocilia are so rigid that applied forces do not bend them; instead they pivot at their base. Within a hair bundle, all the stereocilia are interconnected by filamentous cross-links so that the entire hair bundle moves together. For this to occur, stereocilia must slide along their neighbors by breaking and reattaching filamentous cross-links in a complex and incompletely understood process. It is thought that this relative sliding mechanism results in ion channels opening and closing along the stereocilia membrane that, in turn, lead to the propagation of electrical signals down to the hair cell base. These electrical signals then trigger the release of a chemical neurotransmitter near synaptic junctions leading to nerve cells comprising the auditory nerve. We study nerve conduction in much more detail later in this book.

So, in principle, we see the path by which sound waves in air are eventually converted into an electrical signal along a nerve fiber. Sound waves collected by the outer ear vibrate the tympanic membrane. In turn, through mechanical vibrations, the stapes sets up traveling waves along the basilar membrane and other structures of the cochlea. For the stapes oscillations to effectively produce vibrations within the fluid of the inner ear, there must be another site for pressure relief because the fluid is incompressible; this is the round window. There are actually two types of hair cells, known as inner and outer. The outer hair cells are attached to the tectorial membrane and have efferent (motor)

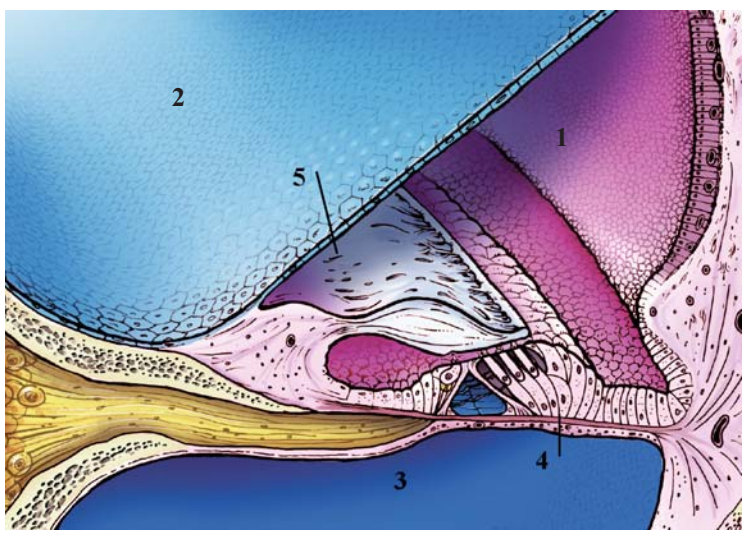


FIGURE 11.20 *The organ of Corti, showing the three chambers (tympani (3), vestibuli (2), and media (1)), basilar membrane (4), and tectorial membrane (5).*

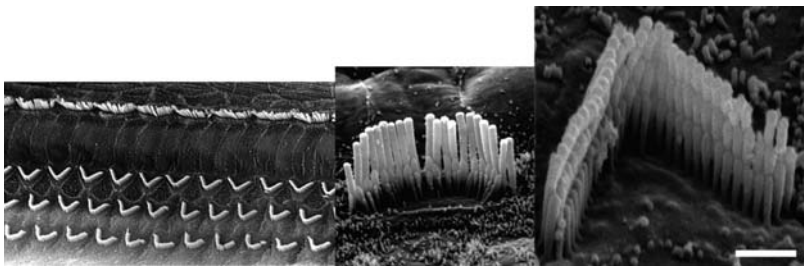


FIGURE 11.21 (left) *Electron microscope detail of hair cells of the cochlea, inner hair cells in a nearly linear array in the background and outer hair cells in a characteristic pattern;* (middle) *inner hair cells;* (right) *outer hair cells. (bar = 3 μm).*

neuron connections so that they do not provide information to the brain, but instead play an active feedback role, taking signals from the brain and modifying the elastic interaction between the basilar and tectorial membranes. Such processes are inherently both extremely complex as well as nonlinear. The inner hair cells on the organ of Corti are sheared by relative motions of the basilar membrane in the surrounding fluid to produce an electrical change in the stereocilia membrane leading to a series of electrochemical events that culminate in the recognition of sound in the auditory cortex of the brain.

Although we have given a reasonably complete outline of the primary mechanism for the transduction of sound to nerve impulse, a number of general unanswered questions remain, among them: how do we distinguish sounds of different frequency and intensity?

FREQUENCY RESPONSE

Our early understanding of how we hear different frequencies of sound is due to von Békésy during the 1940s to 1960s, although a more complete picture came only in the 1980s. The key point is that the basilar membrane acts as a frequency filter in an as yet incompletely understood, but remarkable way. Vibrations of the stapes result in traveling waves of varying amplitude along the basilar membrane. These waves have a maximum amplitude that occurs at different distances along the cochlear spiral from the stapes, with higher frequencies having a maximum closer to the stapes and lower frequencies having their maximum further toward the apex (Figure 11.22). At high enough frequencies there is no displacement at all near the apex. The variation in the position of the wave amplitude maximum reflects variations in the basilar membrane thickness, elastic properties and structure along the spiral. The cochlea ducts all become narrower toward the apex, however, the basilar membrane thickens and widens so as to act as a frequency filter. Only in the 1980s was it shown that the membrane stiffness turns out to decrease exponentially along the spiral by almost a factor of 1000 (Figure 11.23), large enough to account for the frequency range of hearing, so that the location of the maximum wave amplitude varies with the logarithm of the frequency. These experiments

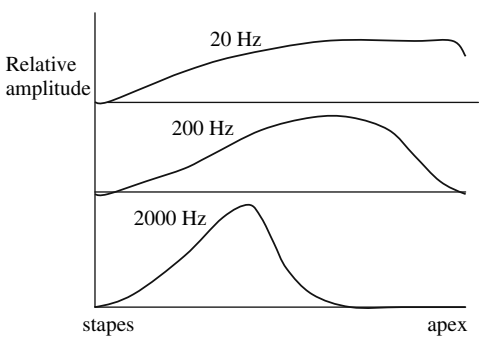


FIGURE 11.22 *Frequency response of the basilar membrane as a function of distance from the stapes.*

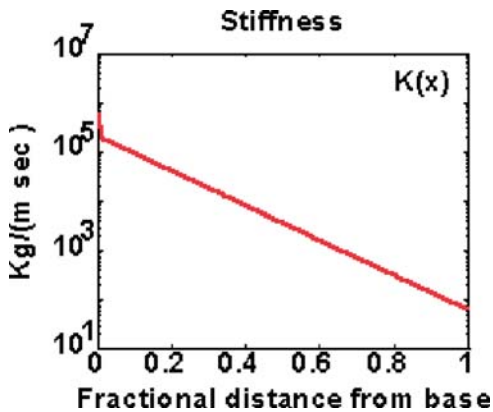


FIGURE 11.23 *Stiffness of the basilar membrane versus distance into the cochlea (Note log scale on y-axis).*

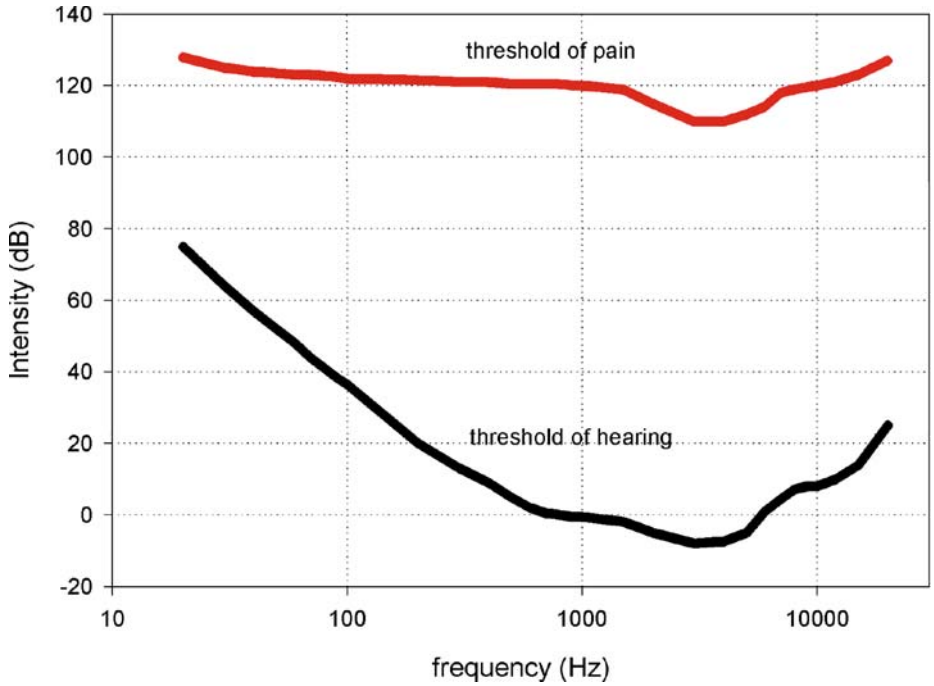


FIGURE 11.24 *The sensitivity of the human ear.*

were done using laser holographic techniques (see Chapter 25) to visualize the variation in membrane modes of vibration with the frequency of stimulation.

The human ear can typically detect sound within the frequency range of from 20 to 20,000 Hz, although the upper limit decreases dramatically with age. The ear is not equally sensitive to all frequencies in this range, however, being most sensitive between about 200 and 4000 Hz (see Figure 11.24). This range is sufficient to hear speech, although a wider range is clearly beneficial for a fuller appreciation of music.

INTENSITY EFFECTS

The human ear has a tremendous range of response to sound intensity. At our most sensitive frequency of 3 kHz, the ear responds to intensity levels as low as 10^{-12} W/m^2, the threshold of hearing, taken as 0 dB, as discussed above in Section 2. Taking the area of the tympanic membrane as 0.5 cm^2, the total threshold power incident on the ear is equivalent to only 0.5×10^{-16} W. This corresponds to, for example, the average power generated by dropping a tiny pin made from 100 million aluminum atoms from a height of 1 m every second (remember the telephone commercial). Using Equation (11.6), this intensity corresponds to a maximum pressure variation of about 2.8×10^{-5} Pa (recall that atmospheric pressure is 1×10^5 Pa). Amazingly, this minimally detected pressure variation corresponds to an amplitude of vibration of air molecules about 10 times smaller than the radius of a single atom! The ear is an exquisitely sensitive detector. At this same frequency, our ears can also tolerate sounds a million million times louder, or 1 W/m^2, known as the threshold of pain. Using the decibel scale this corresponds to 120 dB. At this intensity level, air molecules have a displacement amplitude of about 11 μm and beyond this level, sound becomes painful.

6. THE DOPPLER EFFECT IN SOUND

The Doppler effect in sound occurs when either the source of sound or the listener (detector) are moving. It is commonly experienced from the characteristic frequency changes heard from the siren on a fire truck as it rushes by. The sudden drop in pitch heard as the truck goes by is due to the Doppler effect. Although not as obvious, the

frequency of the siren is also actually higher as the fire truck approaches the listener than it would be if the truck stopped. This phenomenon occurs for all types of waves including light, a form of electromagnetic wave that we discuss in detail later in this text.

In the case of light, when the frequency shifts, the color of the light changes. The well-known red shift of starlight in astronomy is due to the fact that stars are rapidly receding from us. Characteristic frequencies of light are emitted by various atomic elements as we show in Chapter 25. By comparing the frequencies of emitted light from atoms in the laboratory with that emitted from stars, the frequency shifts can be used to determine the recessional velocities of stars using similar equations to those derived below. This is the ultimate source of our knowledge of the extent and age of the universe.

We can understand the Doppler effect by imagining that a point source of a pure frequency sound emits a continuous set of spherical wavefronts, each one wavelength λ apart and that travel at velocity v, as shown in Figure 11.25. If the source and observer are stationary then the frequency of the sound is determined simply by counting the number of wave crests received per second. Because in a time t the number of wavefronts reaching the detector is vt/λ, the frequency is given by dividing this by time to find the usual expression $f = v/\lambda$.

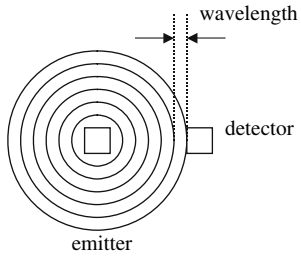


FIGURE 11.25 *Spherical waves from a stationary source detected by a stationary observer.*

Imagine that the detector now moves with a constant velocity v_D along the line towards (or away from) the source. In this case, the number of wavefronts reaching the detector will increase (or decrease) because of the increased (decreased) relative speed of the waves as seen by the detector, so that the detected frequency will be

$$f' = \frac{(v \pm v_D)t}{\lambda t} = \frac{v \pm v_D}{\lambda}. \tag{11.21}$$

This can be rewritten in terms of the frequency detected when the source and detector are both stationary by substituting $\lambda = v/f$ to find

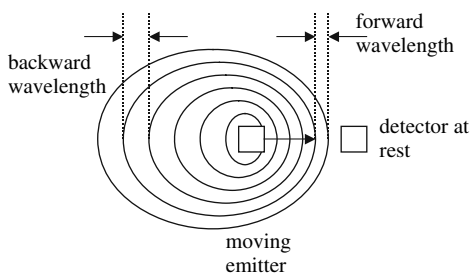
$$f' = f\left(1 + \frac{v_D}{v}\right). \quad (+\text{sign for D approaching; } -\text{sign for D receding}). \tag{11.22}$$

When the detector velocity is zero, Equation (11.21) predicts correctly that there is no frequency shift. If the detector approaches the source the frequency rises above f, whereas if it recedes from the source the frequency drops below f.

A similar phenomenon occurs if the detector is stationary but the source moves toward or away from the detector at a constant velocity of v_S. In this case the motion of the source changes the distance between wavefronts emitted depending on direction. As shown in Figure 11.26, the wavelength is decreased in the forward direction and increased in the backward direction due to the motion of the source. A stationary observer along the line of motion will hear a higher frequency as the source approaches and a lower frequency as the source recedes. This is the explanation of the fire truck siren effect for a stationary observer. In mathematical form the detected frequency is changed due to the wavelength compression or expansion ($\lambda' = \lambda \mp v_S T$, where T is the period, $T = 1/f$) so that the detected frequency is

$$f' = \frac{v}{\lambda'} = \frac{v}{\lambda \mp v_S T} = \frac{v}{\frac{v}{f} \mp \frac{v_S}{f}}. \tag{11.23}$$

FIGURE 11. 26 *Doppler effect for moving emitter and stationary detector. The wavefront spacing in the forward direction is decreased whereas that in the backward direction is increased.*



Rewriting this we have a result for the frequency detected from a moving source

$$f' = f\left(\frac{1}{1 \mp v_S/v}\right). \quad (-\text{for motion toward D; } +\text{for motion away from D}). \tag{11.24}$$

In the more general case in which both source and detector are moving, but still along the line joining them, the detected frequency, from Equation (11.21) and (11.23), is

$$f' = f\left(\frac{1 \pm v_D/v}{1 \mp v_S/v}\right), \tag{11.25}$$

where the upper signs are used when the relative motion brings the source and detector closer and the lower signs apply when that distance is increasing.

The Doppler effect can be used to measure the velocity of moving objects by aiming a wave at the object and measuring the frequency of the reflected wave. This technique is probably most familiar to you in the form of radar. Police radar uses high-frequency radio waves (a form of electromagnetic radiation) to detect the velocity of cars on a highway; weathermen use Doppler radar to measure the velocities of clouds to make forecasts. A medical application of the Doppler effect is the use of ultrasound to determine blood velocities as discussed in the next section.

7. ULTRASOUND

Sound at frequencies above 20,000 Hz is called ultrasound. Although our ears do not respond to sounds of those frequencies, many animals can hear at frequencies ranging up to 100 MHz. Ultrasound may be familiar to you from its use in ultrasonic cleaning baths (for jewelry or glassware), cool mist humidifiers, and fetal monitoring, a very common method of imaging a fetus within the womb. In this section we study some of the physical properties of ultrasound and its interaction with matter. We also learn the fundamental ideas behind medical imaging using ultrasound.

Ultrasound differs from audible sound only in its higher frequency and correspondingly shorter wavelength. In most of the applications we discuss, ultrasound is traveling through water or biological tissue in which the speed of sound is quite a bit faster than in air. Referring back to Table 11.1 we see that the velocity of sound in water and various biological tissues is quite fast (nearly a mile per second). For 1.5 MHz ultrasound, the wavelength in water (using the speed of sound as 1480 m/s) is just about 1 mm. The fact that the wavelength is so short is important because the wavelength ultimately limits the possible obtainable resolution when imaging with ultrasound.

Ultrasonic waves traveling in a material undergo several interactions. Some portion of the wave is absorbed as it travels through the material. This is usually described by an *absorption coefficient* α that describes the loss in intensity of the wave as it travels along

$$I(x) = I_o\, e^{-\alpha x}, \tag{11.26}$$

where I_0 is the intensity at some arbitrary point labeled $x = 0$ and $I(x)$ is the intensity transmitted through the material after the wave has traveled a further distance x. The smaller the absorption coefficient, the longer the wave can travel through the medium without appreciable loss. In pure water absorption over the distances of 0.1–0.2 m used in imaging systems is negligible. The absorption coefficient in human soft tissue depends on the frequency of the ultrasound, increasing with frequency in the MHz range with a typical value of about 12% per cm of distance per MHz. Thus, 1 MHz ultrasound loses 12% in the first 1 cm, an additional 12% in the second cm, and so on, so that after 10 cm, there is only 28% of the original signal intensity left, the rest being absorbed. At 5 MHz, in the first 1 cm 60% of the intensity is lost, so that after 10 cm there is less than 0.01% of the original intensity left, all the rest being absorbed.

This particular interaction of ultrasound with tissue is used in two different ways. At low-intensity levels, the absorbed energy heats the tissue. This interaction is clinically used in *diathermy* to locally heat tissue. At higher powers a new phenomenon

occurs, known as *cavitation.* At these higher-intensity levels the local pressure variation is sufficient to tear apart the medium, forming spherical holes or cavities. Medical applications of cavitation include the disruption of kidney stones or tumors using focused ultrasound. Other applications include cleaning solid surfaces (such as glassware or jewelry) and disrupting cells and cell constituents for scientific applications.

When an ultrasonic wave reaches a boundary between two different media, some of the wave is reflected back and the rest of the wave is transmitted (Figure 11.27). The *acoustic impedance* z, a parameter defined as the product of the mass density and the velocity of sound in the medium, $z = \rho v$, determines the fraction of the wave that is reflected. If z_1 and z_2 are the acoustic impedances of the two media at a planar boundary then the fraction of the incident intensity that is reflected back is

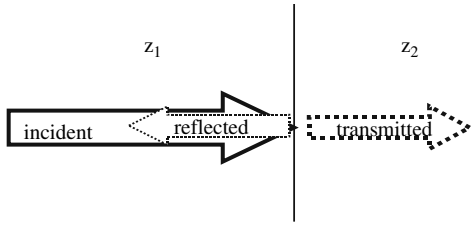


FIGURE 11.27 *The acoustic impedance of the two media determines the division of the incident acoustic energy into reflected and transmitted waves.*

$$\frac{I_{\text{reflected}}}{I_{\text{incident}}} = \frac{(z_1 - z_2)^2}{(z_1 + z_2)^2}. \tag{11.27}$$

If the two impedances are equal, then Equation (11.27) confirms that there will be no reflection and all the intensity will be transmitted (because $I_{\text{transmitted}} + I_{\text{reflected}} = I_{\text{incident}}$, we have that

$$\frac{I_{\text{transmitted}}}{I_{\text{incident}}} + \frac{I_{\text{reflected}}}{I_{\text{incident}}} = 1).$$

If one impedance differs from the other by a factor of 10 then Equation (11.27) predicts 67% of the intensity will be reflected. Table 11.3 lists the acoustic impedance of some materials relevant for biological imaging. Different tissues in the body all have impedance values similar to those of water except for bone, whereas air has a much lower value, implying that the lungs should have a distinctly lower impedance. These values are important in describing the "contrast" of different tissues to ultrasound. That is, if neighboring tissues have similar impedances, there will only be a small reflection of intensity at their boundary, but at bone or lung interfaces there will be a much larger reflected signal. In addition, at an air–tissue interface, only a small fraction of the intensity will be transmitted, so that it is difficult to "couple" ultrasound into the body. We return to these ideas shortly when we consider imaging methods.

Table 11.3 *Acoustic Impedances*

Material	*Acoustic Impedance (kg/m^{2}s)*
Air	430
Water	1.48×10^6
Fat	1.33×10^6
Muscle	1.64×10^6
Bone	6.27×10^6

In order to generate ultrasound, a mechanism for producing vibrations at MHz frequencies is required. The diaphragm of a loudspeaker cannot be made to vibrate at these high frequencies, however, there are special materials, known as *piezoelectric* ceramics, which oscillate at such frequencies in response to a MHz time-varying electrical signal. Other materials, known as *magnetostrictive* ceramics, respond similarly to time-varying magnetic signals. Furthermore, these materials work reversibly, just as a loudspeaker does. Loudspeakers normally interchange electrical energy for sound energy, taking an oscillating electrical signal and producing vibrations of the speaker, leading to sound. A microphone that converts sound into an electrical signal is basically just a small speaker working in reverse. Sound impinging on the speaker

FIGURE 11.28 *An ultrasonic fetal monitor at work.*

produces vibrations that cause a small electric signal to oscillate at the same frequency. We show how this works later when we learn about electromagnetism.

Devices that change one form of energy into another form are known as *transducers.* Ultrasonic transducers are very efficient devices that can be used as a source or detector of ultrasound because the conversion of acoustic energy to electrical or magnetic energy is reversible in these devices. In other words, an applied high-frequency electric or magnetic signal can produce the mechanical oscillations that yield ultrasound, or an ultrasonic wave impinging on the transducer will induce mechanical oscillations that, in turn, produce a time-varying electric or magnetic signal that "detects" the presence of ultrasound.

Ultrasonic transducers must be very sensitive in order to "see" the reflections from soft tissue boundaries because the acoustic impedances are very similar and the reflections are correspondingly weak. For example, at a boundary between fat and water only 0.5% of the incident wave is reflected, as a short calculation using the data in Table 11.3 and Equation (11.27) indicates. In ultrasonic imaging, the transducer is mounted in a microphone-type housing with a fluid-filled tip that is pressed against the skin, coated with a layer of gel to eliminate an air gap through which ultrasound would not penetrate (Figure 11.28). The single transducer is used as both source and detector of pulses of ultrasound as we now describe.

Ultrasonic imaging is based on the *pulse–echo* method. A short pulse of ultrasound, typically of several MHz in frequency, is directed into the soft tissue of the body. Reflections from boundaries with different acoustic impedance arrive back at the transducer in times that depend on the round-trip distance and on the average speed of sound (which we take as 1570 m/s for soft tissue: see Table 11.1). From the delay time between the emission of the pulse and the detection of the echo, we can reconstruct the distance to the boundary as

$$d = \frac{1570t}{2}, \tag{11.28}$$

where d is measured in meters, t is the delay time, and the factor of 2 accounts for the round-trip of the pulse. This pulse–echo method is the same as is used in sonar to map the ocean's floor or by flying bats to navigate. In ultrasonic imaging, this simplest of methods is called an *A-scan* and gives information on not only the depths of boundaries corresponding to each reflection, but also information as to the acoustic impedance (and therefore the tissue type) of each region based on the intensity of the pulse echo. Note that the transducer must be both very sensitive to detect the low intensities of the echoes and have a fast response time. A-scans, however, give only information on the depth of tissue boundaries; they do not give any spatial information in the directions transverse to the direction of travel of the pulse.

By recording the information from an A-scan differently and by scanning the incident pulse along a transverse line, an image of the major acoustic boundaries can be displayed on a computer screen in a *B-scan.* The pulse–echo information is recorded so that one axis of the image corresponds to the echo depth in the tissue and the image brightness corresponds to the intensity of the echo. Without any scanning a strong single echo would appear as a bright dot, a weaker echo as a fainter dot, and multiple reflections as a series of such dots along the axis. If the incident pulses are scanned along a transverse line, then because the pulse duration is short and the reflection times are short, an entire sequence of such scans can be independently accumulated to yield the outline of tissue boundaries. This is done by displaying the scanning distance along an orthogonal axis. The time for a complete scan is short enough to persist on the computer screen, much the same way as television works.

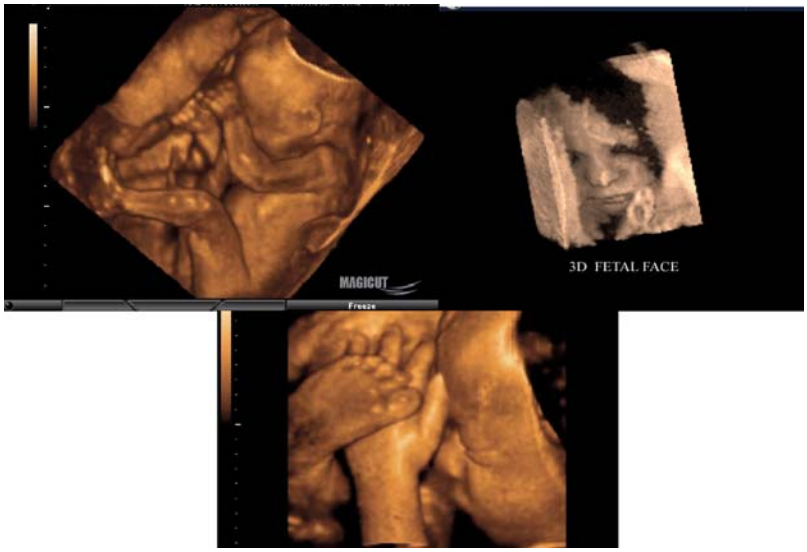


FIGURE 11.29 *High resolution 3-D ultrasound images of a fetus.*

Techniques have been developed to produce narrow beams of ultrasound that are scanned rapidly and continually to produce a continuous real-time image. Figure 11.29 shows examples of a B-scan. Note that false color is added to the pictures to enhance the contrast for our eyes. Each color corresponds to a different level of intensity according to some grayscale level in which intensity is scaled between black and white with shades of gray. The intensity levels of the pulses used in imaging are sufficiently low ($< 3 \times 10^4$ W/m^2) so that this method is considered a safe and completely noninvasive technique. It is widely used in fetal monitoring and in imaging internal organs of the body. The spatial resolution is limited to about 1 mm due to the frequency of ultrasound; higher frequencies would give better resolution in principle, but the absorption increase with frequency is prohibitive.

A third type of imaging, known as the *M-scan* or motion-scan, is similar to the A-scan but measures the position of a moving target, such as a heart valve, in a time sequence of pulse echoes. A more sophisticated version, known as *Doppler scans*, makes use of the Doppler shift of sound (see the previous section) to produce velocity profile images. This technique is useful in mapping motions within the heart and gives a two-dimensional image similar to a B-scan, except that the false color does not indicate the intensity of the reflection but rather its frequency shift (related to the velocity of the target). Figure 11.30 gives an example of this type of image. Ultrasonic imaging is the first of a number of imaging methods that we study, including CT scans (using x-rays), MRI (using radio waves), and PET (using the emission products of radioactive particle decays). These techniques have revolutionized medical care as well as our knowledge of the human body.

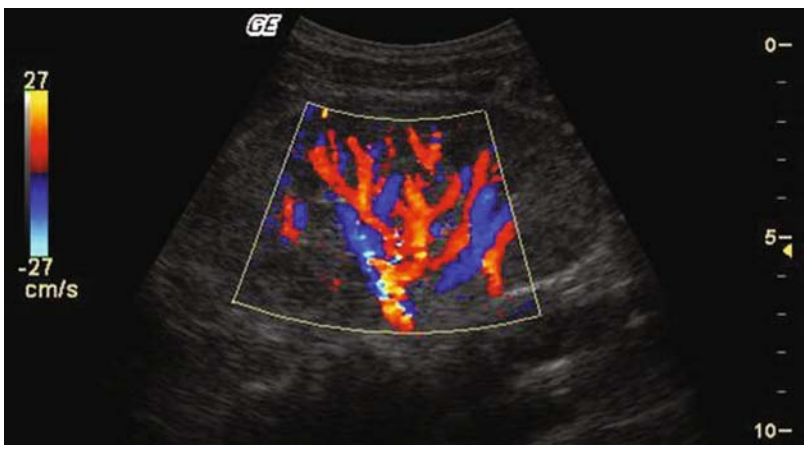


FIGURE 11.30 *Doppler scan of the adult kidney with color code indicating flow rates.*

CHAPTER SUMMARY

Sound is a longitudinal pressure wave that can be described by either a traveling pressure wave or a displacement (of air, or whatever medium it travels in) wave:

$$\Delta P = \Delta P_{\max} \sin(kx - \omega t), \qquad (11.1)$$

$$\Delta s = \Delta s_{\max} \cos(kx - \omega t). \qquad (11.2)$$

Sound intensities are proportional to the square of ΔP and are measured using the decibel scale

$$\beta = (10\ \mathrm{d}B)\log\frac{I}{I_o}, \qquad (11.7)$$

where I_o is a reference intensity (here taken as 10^{-12} W/m^2).

When sound waves strike a boundary between two different materials, in which the speed of sound differs, some fraction of the intensity is reflected and the rest is transmitted but is refracted, or bent, according to the law of refraction,

$$\frac{\sin\theta_{\text{incident}}}{\sin\theta_{\text{refracted}}} = \frac{v_{\text{incident}}}{v_{\text{refracted}}}. \qquad (11.8)$$

Two overlapping sound waves of different frequencies will exhibit a phenomenon known as beats, in which the net sound produced by interference will have the average frequency, but will have an amplitude that oscillates at the difference, or beat, frequency,

$$y = \left[2A\cos\left(\frac{\omega_1 - \omega_2}{2}\right)t\right]\sin\left(\frac{\omega_1 + \omega_2}{2}\right)t. \qquad (11.11)$$

Sounds produced by wind or brass instruments can be modeled by closed or open tubes, or columns of air, leading to a set of resonant frequencies able to be excited in each type of tube according to

$$\text{open tube} \quad f_n = \frac{v}{\lambda_n} = \frac{nv}{2L} = nf_1, \quad n = 1, 2, 3, \ldots, \qquad (11.17)$$

$$\text{open side closed tube} \quad f_n = \frac{v}{\lambda_n} = \frac{nv}{4L} = nf_1. \quad n = 1, 3, 5 \ldots. \qquad (11.19)$$

The relationship between the structure and function of the three parts of the ear is discussed, showing how a pressure wave incident on the outer ear ends up as an electrical signal produced by the hair cells of the inner ear.

Sound waves that are either produced by a moving source, detected by a moving sensor, or both, will have their frequency f shifted, to f', according to the Doppler effect,

$$f' = f\left(\frac{1 \pm v_{\mathrm{D}}/v}{1 \mp v_{\mathrm{S}}/v}\right). \qquad (11.25)$$

Ultrasound, sound waves at frequencies above those capable of human detection (>20,000 Hz), can be used to probe inside the human body by detecting reflections from "objects" (organs, a fetus, blood, etc., with different acoustic impedance) and measuring pulse echos to determine depth information.

QUESTIONS

1. Give a conceptual argument based on the nature of a pressure wave as to why the speed of sound should be greater in a liquid than a gas and still greater in a solid.
2. If we lived in "Flatland," the two-dimensional world of Edwin Abbott, and sound were confined to our two-dimensional world, repeat the argument in Section 2 to find how intensity would vary with distance from the source.
3. What is the ratio of intensities of two sounds that differ by 1 dB? What is the intensity level difference (in dB) between two sounds that differ by a factor of 2 in intensity?
4. Discuss the differences and similarities between temporal and spatial superposition of sounds.
5. Why do two sound waves need to be coherent in order to exhibit interference phenomena?
6. Suppose that you are given a set of three consecutive resonant frequencies from a resonant tube. You do not know if the tube is open at one end or at both. Comparing Equations (11.17) and (11.19) how could you tell?

7. Musicians commonly tune their instruments to "A" = 440 Hz. Two violinists prepare to play a duet together. One of them claims his instrument is tuned perfectly to A. The partner is also sure that his instrument is tuned to A. They draw their bows across their respective instruments and hear a beat of 2 Hz. Is there any way they can tell whose instrument is in perfect tune?
8. Review the basic sequence of events that lead from an incident sound wave to a signal along the auditory nerve.
9. There is also a Doppler effect for light. If a source of visible light is receding from an observer, based on the discussion in Section 6 for sound, do you expect a shift of detected frequency toward the red or toward the blue? What if the source is directed towards the observer? This effect is used, with other measurements, to determine the recessional velocities of stars.
10. From a consideration of acoustic impedance, why would ultrasound be better for detecting a bone fracture than for detecting fat blockages in arteries?
11. The resolution of ultrasound is dependent on the wavelength, increasing with decreasing wavelength. Why doesn't ultrasonic imaging use much higher frequencies (shorter wavelengths) in order to increase the resolution to be much better than about 1 mm (Hint: consider absorption and its effects)?

MULTIPLE CHOICE QUESTIONS

1. Ultrasonic imaging is not based on (a) pulse echo techniques, (b) differences in acoustic impedance, (c) cavitation, (d) scanning.

Questions 2–5 refer to an acoustic resonator tube with a speaker mounted at one end and a solid piston able to slide in the tube mounted at the other end.

2. The ends of an acoustic resonator tube correspond to which of the following pressure conditions: (a) antinode at the speaker, antinode at the piston; (b) antinode at the speaker, node at the piston; (c) node at the speaker, antinode at the piston; (d) node at the speaker, node at the piston.
3. You set the frequency of the speaker to 1000 Hz. As you draw the piston head back from the speaker the first resonance you hear occurs when the head is at 2.5 cm. The next resonance you hear is most likely to occur at (a) 25 cm, (b) 20 cm, (c) 12.5 cm, (d) 7.5 cm.
4. Suppose you have a tube 0.25 m long with a speaker at one end and with the other end open. If you gradually increase the frequency of the speaker from zero at about what frequency will you hear the first resonance? (a) 350 Hz, (b) 700 Hz, (c) 1050 Hz, (d) 1400 Hz.
5. Suppose the tube is replaced with a tube that is open instead of blocked by a piston head. Suppose further that a fundamental resonance is produced for an input frequency of 350 Hz. At about what frequency will a first overtone be produced in the same tube? (a) 117 Hz, (b) 175 Hz, (c) 700 Hz, (d) 1050 Hz.
6. An organ pipe of length 0.5 m has two open ends. The fundamental and first overtones in this pipe have frequencies of about (a) 350 Hz and 700 Hz, (b) 350 Hz and 1050 Hz, (c) 700 Hz and 1400 Hz, (d) 175 Hz and 525 Hz, respectively.
7. A fundamental standing wave is produced in the vibrating wire at an input frequency of 22 Hz. The first overtone will be produced when the input frequency is set at (a) 7 Hz, (b) 11 Hz, (c) 44 Hz, (d) 66 Hz.
8. Two people talk simultaneously, each creating a sound intensity of 50 dB at a given point. The total sound intensity at that point is (a) 0 dB, (b) 50 dB, (c) 100 dB, (d) between 0 dB and 100 dB.
9. A car heads toward a wall at high speed while its horn is blowing. The frequency of the horn when the car is at rest in still air is f. An observer sitting on the wall hears the horn having a frequency f'. The driver hears an echo from the wall that has a frequency (a) equal to f, (b) equal to f', (c) greater than f', (d) less than f'.

Questions 10–12 refer to: A room is filled with air with a pressure P_0. A speaker creates a sound wave in the room described by $\Delta P = \Delta P_{max} \sin(2\pi x - 700\pi t)$. The average intensity of this wave is I.

10. Under typical conditions ΔP_{max} is (a) about the same as P_0, (b) much greater than P_0, (c) much less than P_0, (d) about 350 m/s.
11. At one point in the room a wave directly from the speaker combines with a wave that reflects off a wall to produce a stationary node. This will occur if the difference in distances traveled by the two waves is (a) 0 m, (b) 0.5 m, (c) 1.0 m, (d) 3.14 m.
12. Suppose you wanted to increase the intensity of the wave from I to $4I$. You would have to change (a) ΔP_{max} to $2\Delta P_{max}$, (b) ΔP_{max} to $4\Delta P_{max}$, (c) the $2\pi x$ to πx and the $700\pi t$ to $1400\pi t$, (d) the $2\pi x$ to $4\pi x$ and the $700\pi t$ to $350\pi t$.
13. You have an empty 20 oz. soda bottle and an empty 32 oz. soda bottle, both roughly the same diameter. You blow air over the opening of one and produce a fundamental standing wave. Then you blow air over the opening of the other and produce another fundamental standing wave. Which is true: (a) The fundamental tone in the 20 oz. bottle is lower in frequency than in the 32 oz. bottle. (b) The fundamental tone in the 20 oz. bottle is higher in frequency than in the 32 oz. bottle. (c) The tones are both fundamentals and therefore are the same frequency. (d) The speed of the airflow must be the same for both bottles.
14. You have an empty 20 oz. soda bottle and you blow air over the opening to excite a fundamental standing wave. Now, you slice off the bottom of the bottle (it's plastic) without changing its length very much. You blow over the opening and excite a fundamental

standing wave in the bottle with its bottom end open. The frequency of the standing wave in the second case (a) is higher than that in the first case, (b) is lower than that in the first case, (c) is the same as that in the first case, (d) no sound is produced in the second case.

15. Which one of the following is true? (a) The air pressure in a room is 1 atm; therefore the amplitude of a sound wave in the air must be about 1 atm. (b) A horizontal string is 1 m off the floor; therefore the amplitude of a transverse wave on the string must be about 1 m. (c) A traveling water wave carries mass along with it. (d) A traveling wave of people alternately standing and sitting in a baseball stadium carries energy along with it.

16. How much louder (in dB) is a sound heard 2 m from a point source than when it is heard by the same ear 4 m from the source? (a) 4, (b) 2, (c) 10 log 4, (d) 10 log 2, (e) none of the above.

17. In a resonant tube open at one end and closed at the other, the resonant frequencies are determined by all of the following except (a) the speed of sound, (b) the length of the tube, (c) the boundary conditions at the ends of the tube, (d) the temperature of the air, (e) the tube diameter.

18. The intensity of sound wave A is 10 dB greater than that of sound wave B. Measured in W/m^2 the intensity of A must be greater than the intensity of B by (a) a factor of 2 times, (b) a factor of 10 times, (c) 10 N/m^2, (c) 10^5 N/m^2.

19. Suppose that the speed of sound in still air is 350 m/s. A source of a pure tone of 1000 Hz moves through the air at a speed of 30 m/s. An observer at rest with respect to the air hears the tone at a frequency of 1094 Hz. This is primarily because the (a) speed of sound to the observer is 380 m/s, (b) speed of sound to the observer is 320 m/s, (c) wavelength of the tone as measured by the observer is 0.32 m, (d) wavelength of the tone as measured by the observer is 0.38 m.

20. Three speakers, all connected to the same amplifier, all put out the same single frequency tone. At one point in the vicinity of the speakers the three tones add coherently, producing an intensity maximum. If the intensity of each individual speaker at that point is I (in W/m^2) the intensity of sum of tones is (a) $9I$, (b) $3I$, (c) I, (d) zero.

21. The auditory canal of a human ear is about 2.5 cm long. From this we can infer that humans are especially sensitive to sound with a wavelength of about (a) 2.5 cm, (b) 5 cm, (c) 7.5 cm, (d) 10 cm.

PROBLEMS

1. A beaver swims near its den on the shore of a lake 800 feet wide. Startled, it slaps its tail on the water surface before diving underwater. How long does it take the sound of the slap to cross the lake to a beaver near the opposite shore if the second animal is
(a) Above the water surface?
(b) Underwater?

2. A hunter stands 200 m away from one side of a steep-walled canyon that is itself 600 m wide. If he fires a gun, describe the sequence of echoes that is heard.

3. Write an equation for the speed of sound at any temperature given the information in Section 1 of the chapter.

4. Determine how big a change there is in the speed of sound due to seasonal extremes in outdoor air temperature, taking the warm summer upper value to be 30°C and the cold winter lower value to be −10°C.

5. Compute the two wavelengths of sound, λ_{low} and λ_{high}, corresponding to the 20 Hz low- and the 20 kHz high-frequency limits of human hearing. Assume 343 m/s for the speed of sound.

6. An ironworker at a large construction site guides a steel girder into place with a mallet, slamming the mallet down onto the steel every 1.5 s. A foreman watching the ironworker from some distance away discerns no time lag between sight of the mallet impact and the sound of the clang of the steel. How far away is the foreman?

7. Fill in the table with the lengths of resonant tubes that will produce fundamental frequencies at the low and high limits of human hearing, 20 Hz and 20 kHz, respectively,

	Tube, Open Both Ends	*Tube with One End Closed*
Low freq.		
High freq.		

8. If the intensity of sound from a jet engine is 10 W/m^2 at a distance of 30 m, how far away from the jet do you have to be for the intensity to be 0.1 W/m^2?

9. How much acoustic energy is emitted by a source every second if the sound intensity is 80 dB at a distance away of 20 m?

10. At a distance of 10 m away, the equipment of a road repair crew emits sound of 90 dB intensity.
(a) How much farther away would a passerby have to remove himself so that the sound intensity would be a somewhat more tolerable 80 dB?
(b) If a member of the repair crew must work at a distance of 1 m from the noisy equipment, to what sound intensity, in dB, is he exposed?

11. Using values for the variation in air pressure due to sound waves and the dimensions of the eardrum (tympanic membrane), both given in the chapter, calculate the force on the eardrum for sound at maximum safe intensity.

12. A crying child emits sound with an intensity of 8.0×10^{-6} W/m^2.
(a) What is the intensity level in decibels for the child's sounds?

(b) Suppose that two children are crying with the same intensity. What is the intensity level in decibels for the two children crying together?
(c) Derive a general rule for the intensity level in decibels (based on parts (a) and (b)) if there were four children, eight children, or any even number of children.
(d) How long does it take you to hear the children crying if you are 100 m from them when they start crying?

13. Suppose that you hear a clap of thunder 5 s after seeing the lightning stroke. If the speed of sound in the air is 343 m/s and the speed of light in air is 3×10^8 m/s, how far are you from the lightning strike?

14. A listener moves with respect to a musician who plays a steady middle C note of 262 Hz.
(a) Determine the speed with which a listener must approach a musician such that the perceived pitch is shifted upward a half step to C# (C-sharp) = 277 Hz.
(b) If the musician were instead playing C#, would the note be perceived by the listener as C if the listener recedes from the musician at a speed equal to that of the previous case?
(c) Suppose it was the source (i.e., the musician) that was in motion. What is the magnitude and direction of such motion that would result in the middle C in fact being played by the musician to be perceived by the listener as C#?

15. The musical scale of "equal temperament" has its notes tuned as shown in the table below. Suppose a string is stretched at such tension that the fundamental of the string oscillation is the lowest C of the scale. Determine the lengths for the same string that will produce fundamentals for all of the notes, assuming a sound velocity of 350 m/s.

Note	*Freq. (Hz)*	*String Length (m)*
C	262	
D	294	
E	330	
F	349	
G	392	
A	440	
B	494	
C	523	

16. Suppose a string similar to that of the previous problem is one meter long and carries tension for C = 262 Hz. Determine the set of tensions necessary, in terms of the initial tension T, for the rest of the notes of the scale using strings of the same length.

17. A piano has about 240 strings (one key controls several strings). Increasing the string tension increases the pitch (i.e., the frequency of the fundamental). Higher tension also increases sound volume. Therefore, it is musically advantageous to have the strings for the lowest notes have as high a tension as possible. Piano wires have diameters ranging from 31 to 55 mils (0.79–1.4 mm) made of steel only, or of steel cores wound with copper. Determine the string type and size that will result in the largest volume of sound for the lowest notes. Assume the length is fixed, determined by the dimensions of the piano. Note density of steel = 7.8×10^3 kg/m^3; density of copper = 8.9×10^3 kg/m^3.

18. What will be the fraction of ultrasound intensity reflected from the surface of the heart? Consider the heart to be a muscle, surrounded by water.

19. How long is the time gap between ultrasound reflections from the front and back of the heart, assuming the heart to be modeled as a cube of edge length 15 cm?

20. If we use the value given in the text for an absorption coefficient of 0.12/cm/MHz, what distance in water will result in an absorption of a 5 MHz ultrasound beam
(a) of 10%?
(b) of 90%?
(c) Suppose instead the frequency is reduced to the nominal minimum of 1 MHz. Calculate the distances traveled for the same fractional absorption.

21. A basic property of measurement with waves of any type is diffraction, wherein the interaction of the object under study with the wave gives rise to a distortion of the direction of wave travel. Diffraction effects impose an effective lower limit on the determination of size of the target object and this limit can be taken to be roughly equal to the wavelength of the wave. By calculating the wavelength of an ultrasound beam of frequency of 10 MHz in water, what is the size limit for objects under observation with ultrasound?

22. A drummer begins to drum on iron railway tracks with a regular beat. You are nearby with your ear near the tracks and hear two sets of drumming, one starting 0.8 s after the other. (The speed of sound in air is 345 m/s and in iron is 5,000 m/s.)
(a) How far away are you from the drummer?
(b) If the delayed sounds are 5 dB less intense than the first set of drumming heard, find the ratio of the intensities of the two sounds.
(c) If the drummer drums at a frequency of 4 Hz, what frequency will a person hear on a train approaching at 60 mph (conversion factor: 1 mph = 0.45 m/s)?

23. A scientist playing with musical instruments has a 1 m long guitar string with total mass 0.010 kg hooked up to a mechanical oscillator.
(a) If the string oscillates in the second harmonic with $f_2 = 330$ Hz, what is the tension in the string?

(b) If the scientist doubled the oscillation frequency, how many oscillating lobes would there be?
(c) Also in the laboratory is a pipe, open at both ends, which the scientist wants to have resonate in the fundamental mode at the same 330 Hz from part (a). How long should this pipe be?
(d) The pipe in part (c) is slightly too long, such that the beat note between the fundamental mode of the pipe and the 330 Hz from part (a) is 5 Hz. How much should it be shortened to reach the resonance sought in part (c)?
(e) A second pipe in the laboratory has resonances at 330 Hz, 550 Hz, and 770 Hz. Is this pipe open or closed?

24. A nerdy scientist proposes to measure how fast he is traveling toward vertical cliffs by blasting a pure 1000 Hz tone and listening for beats produced by the echo. If he hears a beat frequency of 2 Hz, what is his speed? (Use $v_{\text{sound}} = 343$ m/s and remember that he is both a moving source and a moving detector.)

25. A stationary bat sends out an ultrasonic tone at 60,000 Hz searching for food. At what frequency does the bat hear the echo from a dragonfly moving away from the bat at 5 m/s?

26. A Doppler beat device is used to measure the velocity of blood flowing in an artery. Taking the velocity of sound in tissue as 1500 m/s, what is the velocity of blood flowing away from the detector emitting ultrasound at 1 MHz that results in a beat frequency of 15 Hz?

12 Thermal Energy

In the broadest context, thermodynamics is the branch of physics concerned with the study of macroscopic systems with extremely large numbers of constituent molecules. Most prominent in this study is energy and its transformation and exchange with the surroundings. Not only thermal energy, but all forms of energy are included in the domain of thermodynamics. Even in cases in which the basic interactions between the individual molecules are very simple, because of the sheer number of molecules in a macroscopic volume of matter (1 cm^3 of an ideal noninteracting gas has about 3×10^{19} individual molecules), it is impossible to analyze such a system directly using Newton's laws of motion. Even more to the point, the information gained from the enormous calculational exercise of following the trajectory of each molecule would be unintelligible and useless without reducing that knowledge to some small set of macroscopically averaged quantities that could be directly measured. Thermodynamics deals with such systems by calculating these average quantities using statistical arguments, as we show.

In this and the next chapter we learn basic terminology and ideas, study the fundamental laws of thermodynamics and some of their implications, as well as study a number of biological applications of these laws. In this chapter we start by defining temperature, its measurement, and the thermal expansion of materials. Then we turn to the main topic of the chapter, thermal energy and the conservation of energy principle known in the context of thermodynamics as the first law of thermodynamics. Some general applications including thermal properties of matter, colligative properties of solutions, and the transfer of heat are discussed in the last three sections of the chapter. In the following chapter we look beyond the first law and discuss a broad array of topics of biological interest. Thermodynamics is a very rich subject with connections to all areas of biology and chemistry. In these next two chapters, we illustrate the importance of a basic knowledge of thermodynamics to the study of a wide range of subject matter.

1. TEMPERATURE AND THERMAL EQUILIBRIUM

The notion of whether an object is hot or cold is a relative one. Something that is hot to one observer may be cold to another. To someone who has been outdoors in the cold of a northern winter for several hours, a house kept at a temperature of 60°F may be quite warm, whereas to someone visiting from southern Florida, the same house might be very cold. This may seem like a mundane point, but it is related to an important concept of thermodynamics: heat flows from a hotter object to a colder object. The *temperature* of an object is a quantitative measure of its "hotness," a term that we replace below with "internal," or "thermal energy."

When two objects at different temperatures are placed in *thermal contact* with each other, meaning that energy is allowed to exchange between them, heat will flow

J. Newman, *Physics of the Life Sciences*, DOI: 10.1007/978-0-387-77259-2_12,

from the hotter object to the colder object until eventually the two objects reach the same temperature. When this common final temperature is reached, the two objects are said to be in *thermal equilibrium.* As long as they are isolated from other objects and cannot exchange any heat with their surroundings, they will remain at that temperature. For example, a thermos bottle filled with warm juice and ice cubes that melt, will arrive at some intermediate temperature that remains constant for a long period of time (however, the thermos bottle, being imperfect, will eventually allow its contents to reach the ambient temperature of the surroundings, coming into thermal equilibrium with its environment).

Although two different observers may disagree on whether an object is "hot" or not, they will agree on the temperature of that object. This is indirectly a statement of the *zeroth law of thermodynamics*, a law that deals with the conditions under which two objects may be said to be in thermal equilibrium without ever bringing them into contact with each other. A third "measuring object" is used to test this. Formally, if each of two objects, when put separately in thermal contact with a third measuring object, is found to be in thermal equilibrium with the measuring object, then the two objects are known to be in thermal equilibrium with each other, even without coming into thermal contact with each other. This may seem obvious but, because it really required experimental confirmation and is of fundamental importance, it is stated as a law (albeit the zeroth). By using a measuring object (or thermometer), one can separately determine a property of each object (its temperature) in order to know whether heat will flow if the two objects are brought into thermal contact. Whether heat flows when the two objects are brought into contact does not depend on any other variables, including their mass, color, shape, electric charge, and so on, but only on their temperature.

In order to measure the temperature of an object, we first need to define some scale of temperature. Since temperature is a scalar quantity, we need to define the unit of temperature and also some origin or set point; together these define the temperature scale. Two commonly used temperature scales are the *Celsius* (aka centigrade) and *Fahrenheit* scales. The Celsius scale is determined by fixing the temperature span between the freezing and boiling points of water to be 100°C, and by defining the freezing point of water to be 0°C. Alternatively, the Fahrenheit scale uses 180°F to span between the same two physical points, and uses 32°F as the freezing point of water. These two temperature scales are simply related to each other (as you should verify) by

$$T_F = \frac{9}{5}T_C + 32°F. \tag{12.1}$$

Of the two, the Celsius scale is favored in scientific work and is used here.

A question arises as to whether there are upper or lower limits to temperature. As far as we know there is no upper limit to temperature. For example, temperatures of 10^9°C are present within the hottest stars. On the other hand, there is a lower limit of temperature in nature, one that can only be approached, but never attained, as we show. Using this lower limit of temperature as the set point, known as a temperature of *absolute zero*, we define the fundamental or absolute *Kelvin* temperature scale by choosing the temperature of the so-called triple point of water as 273.16 K. (The triple point of water is that temperature at which ice, water, and water vapor coexist within a sealed container and corresponds to 0.01°C.) Note that temperatures measured on the Kelvin scale are not cited as degrees Kelvin, but simply as Kelvin, because of their more fundamental significance. Table 12.1 lists a variety of corresponding temperatures in the three different temperature scales we have introduced. Note that the unit size of 1 K and 1°C are the same, so that

$$T_C = T_K - 273.15°C, \tag{12.2}$$

where the 0.01°C difference in the triple point and freezing point of water in the definitions of K and °C is noted.

Table 12.1 *Comparison of Various Temperatures in Different Units*

Temperature	*Celsius (°C)*	*Kelvin (K)*	*Fahrenheit (°F)*
Helium liquefies	−269	4.2	−452
Nitrogen liquefies	−196	77	−321
Dry ice (CO_2 freezes)	−78	195	−108
Freezing point of water	0	273	32
Human body (core)	37	310	98.6
Boiling point of water	100	373	212
Gas flame (stovetop)	1630	1900	2970
Surface of sun	5730	6000	10,350
Center of Earth	15,700	16,000	28,300
Center of sun	10^7	10^7	1.8×10^7

Example 12.1 Find the general relation between the Fahrenheit and Kelvin temperature scales and determine absolute zero in °F.

Solution: We can find the general relation by substituting Equation (12.2) for T_C into Equation (1) for T_F. After substitution we find that

$$T_F = \frac{9}{5}(T_K - 273.15) + 32 = \frac{9}{5}T_K - 459.67°F,$$

so that when $T_K = 0$ then $T_F = -459.67$°F.

We are familiar with several types of thermometers used to measure temperatures close to the ambient atmospheric temperature (Figure 12.1). Perhaps the most familiar is the mercury-in-glass thermometer that uses the thermal expansion (see Section 2 below) of a column of liquid mercury with increasing temperature as an indicator of temperature. Another thermometer uses the variation in thermal expansion of two dissimilar metals (a bimetallic strip) wound into a coil that controls a pointer. Other thermometers use changes in electrical properties to measure temperature (thermocouples, platinum-resistance thermometers). A variety of other specialized thermometers are used for different ranges of extreme temperatures, both low and high, but are not discussed here.

FIGURE 12.1 *Common thermometers. From left, mercury-in-glass, outdoor, cooking thermometers.*

Having introduced the notion of temperature and thermal equilibrium, we need to make a few general comments before continuing our discussion of thermodynamics. Earlier, we mentioned that two objects in thermal contact but at different temperatures will eventually reach thermal equilibrium. If this is generally true, how do we explain that while we are in thermal contact with the atmosphere, we manage to maintain our body temperature? How do warm-blooded organisms, clearly not at the temperature of their environment, manage to survive? The answer lies in distinguishing between two fundamentally different types of thermal systems: open and closed.

A *closed system* is one that does not exchange mass with its surroundings; such a system typically is physically isolated but still can exchange energy with its surroundings through the bounding walls. Other systems, including living organisms, are *open systems*, exchanging mass, as well as energy, with the surroundings. Animals, for example, require the exchange of water, nutrients, oxygen, and waste products in order to survive. Our previous statements about reaching thermal equilibrium were restricted to closed systems. Open systems are not in thermal equilibrium and are known as *nonequilibrium systems.* Living organisms, for example, constantly replace most of their constituent molecules: skin, muscle, and blood cells; nearly all of our constituents are recycled over various time periods.

Although not in thermal equilibrium, many open systems reach what is known as *steady state.* In this case there is a balance between the input and output of total energy. Such a distinction can also be made in a chemical reaction. When the total amounts of reactants and products are fixed and no mass is exchanged with the surroundings, the reaction will reach a chemical equilibrium. On the other hand, when new reactants are constantly supplied to the system at a sufficient rate to come together and maintain a steady production of products that then leave the system, we call this nonequilibrium situation one of steady-state behavior. We show examples of open systems in our later discussions of thermodynamics. For now, as we introduce the basic concepts of thermodynamics, we limit our discussion to closed systems.

2. THERMAL EXPANSION AND STRESS

Almost all substances expand when heated and contract when cooled. This is true of most liquids and solids as well as gases, discussed in the next section. You may have used this idea to open a tightly sealed glass jar with a screw-top metal cover by warming the cover under tap water. The cover expands somewhat more than the jar and can then be more easily opened. A second, now famous, example is the effect of freezing temperatures on rubber o-rings, dramatically demonstrated by Richard Feynman, Nobel laureate in physics, in connection with the failed Challenger shuttle mission. Unusual freezing weather in Florida led to the contraction of an o-ring seal that became brittle and leaked fuel, causing an eventual explosion of the rocket and loss of life (Figure 12.2).

The origin of thermal expansion or contraction ultimately lies in the molecular motions and interactions of the material. As we study further in the next section, when heated, molecules move about more rapidly, and therefore make harder collisions with neighboring molecules, pushing the material apart. In most materials thermal expansion is uniform in all directions, although in certain crystalline materials with different crystal structure along different spatial directions, expansion may occur to different extents along the different "crystal axes," although these are not discussed further here.

A solid rod of length L is found to expand by an amount that is directly proportional to the temperature increase and to its length according to

$$\Delta L = \alpha L \Delta T, \tag{12.3}$$

FIGURE 12.2 (left) *Challenger just before explosion; flames from leak are visible at top center.* (right) *Richard Feynman doing a tabletop experiment to test the brittleness of a cold o-ring as a demonstration of the source of leaking fuel in the rocket.*

where α is the *coefficient of linear expansion.* For most solids α is quite small as Table 12.2 shows.

Table 12.2 *Coefficients of Expansion for Various Materials*[*]

Material	*Coefficient of Linear Expansion (10^{-6}/°C)*	*Coefficient of Volume Expansion (10^{-3}/°C)*
Solids	α	$\beta = 3\alpha$
Quartz	0.4	
Glass	9	
Steel	12	
Aluminum	24	
Lead	29	
Ice	51	
Liquids		
Mercury		0.18
Ethyl alcohol		1.1
Water		2.1

[*] Room temperature values listed except for ice which is at 0°C.

Example 12.2 A bridge over the New River Gorge in West Virginia has a steel arch with a span of 1700 ft. Find the change in length when the temperature drops by 70°F.

Solution: From Table 8.2, the coefficient of linear expansion for steel is 12×10^{-6}/°C. A 70°F temperature change corresponds to a 70(5/9) = 38.9°C change. Then the length change will be given by $\Delta L = \alpha L \Delta T = (12 \times 10^{-6})(1700)(38.9) = 0.79$ ft. Notice that whatever length units are used for L appear in ΔL and so no units conversion is needed.

More often than not thermal expansion is a problem, not a solution. Roadways and sidewalks buckling from the heat, walls developing cracks from extreme heat or

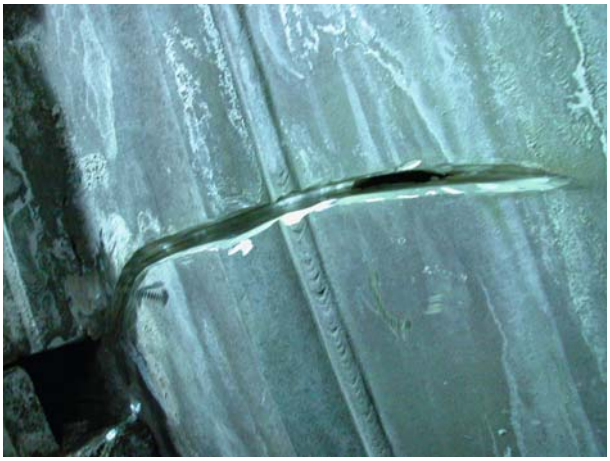


FIGURE 12.3 *Thermal stress rupture of an 8.5 foot diameter tube used in a calciner, a device for treating liquid waste and turning it into high-level solid waste.*

cold, and severe thermal stresses placed on large structures such as bridges or tall buildings are all common problems (Figure 12.3). These arise from the mismatch in thermal expansion of different materials in contact with one another. In our discussion of stress and strain in Chapter 4, we saw that an applied stress, or force per unit cross-sectional area, resulted in a proportional strain, or fractional change in length, for small stresses

$$\frac{F}{A} = Y\frac{\Delta L}{L},$$

where Y is the elastic, or Young's, modulus. In the context of our current discussion, if the material is heated so that it expands and increases the strain, a stress is produced that is known as *thermal stress.* Substituting for the strain from Equation (12.3), we find that the thermal stress can be written as

$$\frac{F}{A} = Y\alpha\Delta T. \tag{12.4}$$

Despite the relatively small values for coefficients of linear expansion of metals, thermal stresses can be enormous because of very large Young's moduli. Thermal stress is also the basis of the bimetallic strip used as a thermometer (Figure 12.4).

Example 12.3 A poorly designed bridge roadway has a 3 m steel beam butted against a concrete wall without an expansion gap at 20°C. If the beam has a 4×6 cm rectangular cross-section, find the force exerted on the concrete wall when the temperature rises to 30°C. Will the concrete buckle? Take the Young's modulus of steel as $Y = 200 \times 10^9$ N/m^2 and the ultimate strength of concrete as 20×10^6 N/m^2.

Solution: From Equation (12.4), substituting $\Delta T = 10$°C, $\alpha = 12 \times 10^{-6}$ /°C from Table 12.2, and $Y = 200 \times 10^9$ N/m^2, we find that $F/A = 2.4 \times 10^7$ N/m^2. From the cross-sectional dimensions we then find the force to be $F = (2.4 \times 10^7)(0.04 \times 0.06) = 5.8 \times 10^4$ N. To determine if the concrete buckles, we must know if the applied thermal stress exceeds the ultimate strength of concrete, 20×10^6 N/m^2; because it does, we know that indeed the concrete will buckle under the thermal stress. This points out the need for expansion joints even when the temperature variations are relatively mild.

Although we have singled out a linear dimension, expansion occurs in all directions. Imagine a cube of metal of volume V (length L on each edge) that is heated so that it expands and increases its volume by ΔV. We can calculate the expanded volume of the metal to be

$$V + \Delta V = (L + \Delta L)^3 = L^3\,(1 + \frac{\Delta L}{L})^3 = L^3\,(1 + \alpha\Delta T)^3 = V(1 + \alpha\Delta T)^3.$$

Because α is so small, when we cube the expression in parentheses, keeping only linear terms in α, we find it is equal to $(1 + 3\alpha\Delta T)$ so that

$$\Delta V = \beta\, V\Delta T, \tag{12.5}$$

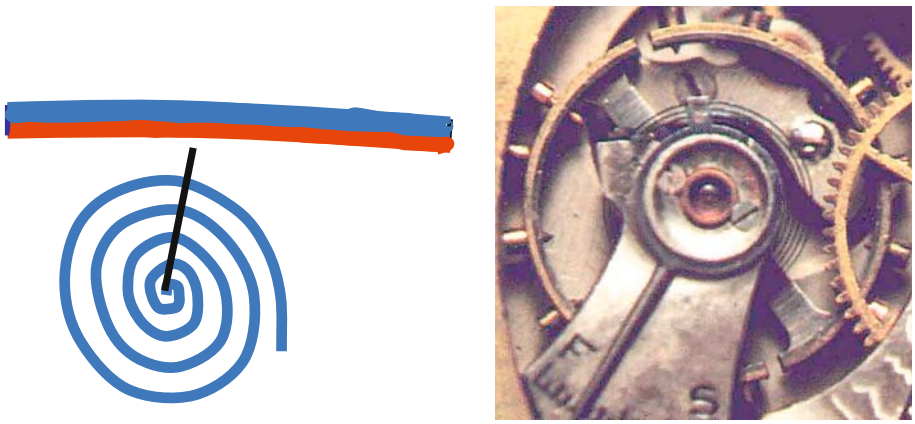


FIGURE 12.4 (left) *The bimetallic strip bends as the temperature changes due to differences in thermal expansion of the two metals; the coil is used in a thermostat to make or break electrical contact.* (right) *A older watch balance wheel with a bimetallic circular strip (the yellowish brass circle with the grey inner steel circular band), designed to compensate for temperature changes in the frequency of oscillation of the balance wheel.*

where β, the *coefficient of volume expansion*, is equal to 3α. In general, arbitrary shaped solids maintain their shape when heated.

Suppose that a solid object has a hole within it. When heated, the solid expands, but what happens to the hole? On first glance one might imagine that the solid expands equally in all directions, including into the hole and thus the hole contracts, but this would be false. To see this consider a steel machine nut threaded onto a steel bolt (Figure 12.5). When heated, the hole in the nut expands to the same extent as the bolt. Thus, we can treat holes in solid objects as expanding in the same way that the solid does.

Example 12.4 A steel bolt with a diameter of 0.2500 inches is to be inserted into a hole in an aluminum plate that is only 0.2495 inches in diameter. Is this possible, and if so find the minimum temperature to which the materials must be heated in order to accomplish this, and the diameter of the hole at this temperature.

Solution: The coefficient of thermal expansion of aluminum is greater than that of steel, therefore the hole will expand faster than the bolt when they are heated. In order to just fit the bolt in the hole we require that ΔL for the hole equal ΔL for the bolt + (0.2500 − 0.2495) inches. Writing this out, we want

$$\Delta L_{\text{hole}} = \alpha_{\text{Al}}\,(0.2495)\Delta T = \Delta L_{\text{bolt}} + 0.0005 = \alpha_{\text{steel}}\,(0.25)\Delta T + 0.0005$$

Solving for ΔT, the needed temperature increase, in terms of the coefficients of expansion, we find that

$$\Delta T = \frac{0.0005}{(0.248\alpha_{\text{Al}} - 0.25\alpha_{\text{steel}})} = 170°\text{C},$$

using the data in Table 12.2. The hole diameter at this temperature can be found from the first part of the above equation, $\Delta L_{\text{hole}} = \alpha_{\text{Al}}(0.2495)\,\Delta T = 0.001$ inches. Then the hole and bolt diameters are both 0.2505 inches at this temperature.

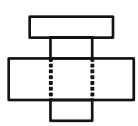


FIGURE 12.5 *A steel machine nut threaded onto a steel bolt.*

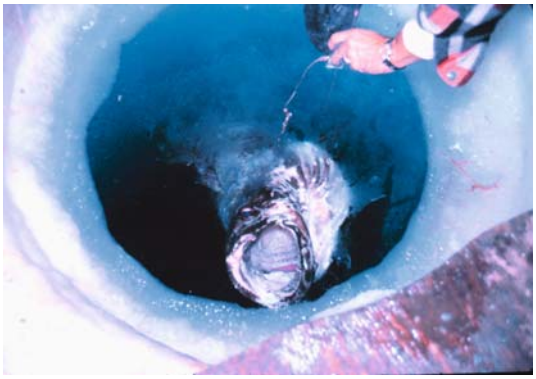


FIGURE 12.6 *Ice fishing: the water below remains at 4°C, allowing fish to survive the winter.*

Equation (12.5) also applies to the expansion of liquids and values for the coefficient of volume expansion of some liquids are included in Table 12.2. Water is an extremely important exception to the general rule of expansion of a liquid with increasing temperature. Above 4°C, water behaves as a normal liquid, expanding as it is heated so that its density decreases. As water is heated from 0°C it behaves anomalously by increasing its density (decreasing its volume) until it reaches 4°C. Thus, the density of water is a maximum at 4°C rather than at 0°C. This unusual property has to do with the strong hydrogen bonding properties of water that lead to microcrystalline ice structures forming at temperatures above but close to the freezing point. We return to the ordered structure of water in the next chapter, in connection with entropy.

This unusual behavior of water has profound consequences for aquatic life. In the winter, as the water in a lake or river cools but is above 4°C, the colder water, being more dense, sinks producing convective flow that keeps the water at a fairly uniform temperature. When the water temperature drops below 4°C, colder water floats because it is less dense than warmer water. Ice eventually forms on the surface when the temperature falls below 0°C and floats because it too is less dense than water (Figure 12.6). The layer of ice actually helps to prevent the water beneath from freezing by forming a layer of insulation and reducing convective flow. This wonderful process allows aquatic life to survive beneath a frozen lake or river surface in water at a temperature of 4°C. If water did not have this unusual property, the coldest water would be densest and would sink so that lakes and rivers would completely freeze in cold winters. By the way, ocean water does not freeze because of the presence of salts, lowering the freezing point of water; this is discussed in Section 6 below.

3. INTERNAL ENERGY AND THE IDEAL GAS

When you drop a 1 kg mass onto your hand all of the atoms—that is, all of its electrons and nucleons—are simultaneously moving downward as a coherent swarm. This coherent motion can easily be measured with a meter stick, for example. As they all fall, the atoms (as well as the stuff from which they are made) are also moving incoherently. Because the latter motions are microscopic, we can't see or measure them directly. The latter, incoherent, unseen, microscopic motions are said to be *internal* and the kinetic energy associated with them is called *internal kinetic energy* (see Figure 12.7). The internal kinetic energy of a macroscopic body is very much greater than the *external* kinetic energy associated with macroscopic motion of the center-of-mass and the macroscopic motion around the center-of-mass. The falling 1 kg mass stings your hand when you catch it because, in effect, you are stopping the coherent motion of all 10^{24} atoms. Those atoms transfer their coherent kinetic energy to the atoms in your hand, and they, in turn, obtain a bit more incoherent motion as a result. The nerves in your hand sense this increase in kinetic energy and send the signal, "sting," to your brain.

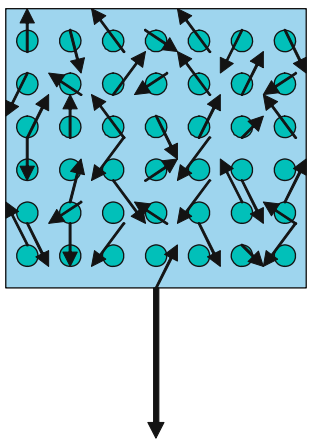


FIGURE 12.7 *The entire object moves with downward coherent motion, while the individual atoms are moving randomly with incoherent motions.*

The transfer of incoherent, internal kinetic energy from one body to another is related to the sensation of temperature. If more internal kinetic energy is transferred from a body to your hand when you touch it (i.e., when the atoms in the surface of your hand come close to the atoms in the surface of the body) than is transferred from your hand to the body, the body feels "hot." Similarly, if more internal kinetic energy is transferred from your hand to the body, the body feels "cold." Although it is not possible to measure internal motions with something as crude as a meter stick, we can infer them with an ordinary device: a thermometer. If all the nucleons, electrons, and jiggling atoms in a body were in their ground state, it would be impossible to remove any kinetic energy from that body. That's because the ground state, by definition, is the lowest allowed energy state. Such a state defines the *absolute zero*

point of the Kelvin temperature scale. Bodies with any degree of internal excitation have temperatures above absolute zero. In the Kelvin scale, temperature is determined by the average internal kinetic energy per atom of a body, above the ground state. When we say a body is "hot," what we are really saying is that the body has a high degree of internal excitation per atom.

All objects consist of molecules that interact through a variety of different electrical mechanisms. Those interactions, no matter how complex, can be pictured as a potential energy curve for a typical molecule that has a minimum at the equilibrium position. As we saw in Section 4 of Chapter 4, near enough to this energy minimum, the curve can always be approximated as parabolic, so that the interactions can be represented by a linear spring with potential energy $\frac{1}{2}kx^2$. Each molecule of the object has kinetic and potential energy while vibrating about its equilibrium position as if attached to such a spring. The *internal energy* of an object is the sum of this kinetic and potential energy and is the quantitative measure of the object's "hotness". Energy due to overall interactions and motions of the entire macroscopic object, such as overall translation or rotation, are not included in the internal energy. These constitute the mechanical energy in our discussions of the mechanics of such objects. After introducing a few general terms, we begin our study of thermodynamics with the specific example of an ideal gas. This leads us to a quantitative connection between internal energy and temperature.

Every thermodynamic study divides the universe into two parts: a *system*, the collection of objects under study, and the *surroundings*, everything else. As noted above, we consider only closed systems at present, those that do not exchange any mass with their surroundings. Thermodynamic systems can be described using a common language. The system consists of an extremely large number of molecules so that there are a correspondingly large number of possible *states*, or configurations with different possible total energy values. Different systems (thermal, electrical, magnetic, etc.) will need different sets of *state variables* to describe their possible states. For example, state variables for a gas are the pressure, volume, and temperature.

We next want to consider the relationship between the pressure and temperature when a specific amount of gas is confined to a volume *V.* Let's consider a collection of *N* identical gas particles contained in a cubical box with edges of length *L.* We aim to calculate the pressure that the particles exert on the walls of the container. We assume that the gas is at a low density so that the volume occupied by the gas molecules themselves is a negligible fraction of *V*, although the gas completely fills the volume. We also assume that the gas is in thermal equilibrium; particles with different velocities can interact with each other through elastic collisions that serve to scramble their velocities and produce the thermal equilibrium. The gas is ideal in that the only interactions between the gas particles are via direct elastic collisions; there are no long-range interactions. As the gas particles move about, they also collide elastically with the container walls. These collisions produce the measurable pressure on the container walls that we wish to calculate.

Let's focus on one particular gas particle that moves with constant momentum until hitting a wall as shown in Figure 12.8. The collision at the wall, being elastic, returns the particle with the same kinetic energy, but has reversed the *x*-component of its momentum, the component perpendicular to the wall, while keeping the other components unchanged. This particle will bounce back and forth periodically making a collision with the wall at the right with a repeat period $\Delta t = 2L/v_x$. From Newton's second law, we can find the average force exerted on the wall by this one particle to be in the *x*-direction and given by

$$F_x = \frac{\Delta p_x}{\Delta t} = \frac{mv_x - (-mv_x)}{(2L/v_x)} = \frac{2mv_x}{(2L/v_x)} = \frac{mv_x^2}{L}. \qquad (12.6)$$

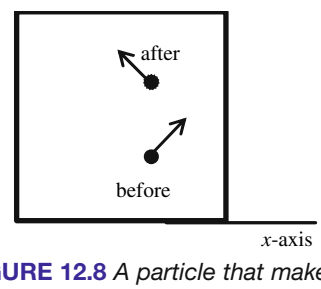


FIGURE 12.8 *A particle that makes an elastic collision with the container wall and rebounds with its momentum in the x-direction reversed.*

Now, with *N* particles in the box, we allow for a variation in the velocity of different particles (discussed below) and calculate the total force on the wall by

multiplying by N and using the average value for the square of the x-component of velocity. Noting that the pressure P exerted on the wall is given by dividing this force by the wall area L^2 (so that the denominator becomes L^3, equal to the volume V), and so we find

$$P = \frac{Nm\overline{v_x^2}}{V}, \tag{12.7}$$

where the bar indicates the average value. Because there is no preferred direction in the box, the averages of each term in the expression for the square of the velocity are equal so we can write

$$\overline{v^2} = (\overline{v_x^2} + \overline{v_y^2} + \overline{v_z^2}) = 3\overline{v_x^2}. \tag{12.8}$$

Substituting $\frac{1}{3}\overline{v^2}$ for $\overline{v_x^2}$ in Equation (12.7), we find

$$P = \frac{Nm\overline{v^2}}{3V}. \tag{12.9}$$

The term $\overline{v^2}$ is called the mean square velocity, and its square root is called the root mean square, or rms, velocity.

It is important to realize that the average of the square of the velocity is not equal to the square of the average velocity; the order of those two operations of squaring and averaging is important. This is easily seen by calculating those two quantities for a small set of numbers, for example, {1, 3, 5}. The average value of these three numbers is 3, whose square is 9; thus $\bar{v}^2 = 9$. On the other hand $\overline{v^2} = (1 + 9 + 25)/3 = 11.7$, whose square root, the rms value, 3.4, is quite different from the average.

Recognizing that the term $m\overline{v^2}$ is equal to twice the mean kinetic energy of a particle, we see that

$$PV = \frac{2N(\overline{KE})}{3} = \frac{2(KE_{\text{total}})}{3}, \tag{12.10}$$

where we have used the fact that N times the mean kinetic energy is equal to the total kinetic energy of the system.

Experimentally it is found that if N molecules of gas are confined in a container of volume V at an absolute temperature T that the pressure is given by the ideal gas law

$$PV = Nk_B T, \tag{12.11}$$

where k_B, Boltzmann's constant, is given by $k_B = 1.38 \times 10^{-23}\, J/K$. Comparing Equations (12.10) and (12.11) we find an expression that relates the mean kinetic energy of a molecule of an ideal gas to the absolute temperature

$$\overline{KE} = \frac{3}{2}k_B T. \tag{12.12}$$

This fundamental relation shows that the microscopic motion of the individual molecules of the system is directly related to the temperature of the gas.

We see that the product of Boltzmann's constant and the absolute temperature is a measure of the mean kinetic energy of the constituent gas molecules. The total kinetic energy of the system will then be equal to $U = 3/2Nk_B$T where this energy is called the thermal or internal energy because it arises from motion within the system rather than overall motion of the container itself. It is this internal energy U that changes when heat flows into or out of the gas through the fixed container walls, raising or lowering the temperature.

Because the mean kinetic energy of a molecule can be written as

$$\overline{KE} = \tfrac{1}{2}\, m v_{\text{rms}}^2,$$

Equation (12.12) gives us an expression for the rms velocity of an ideal gas particle at temperature T,

$$v_{\text{rms}} = \sqrt{\frac{3k_B T}{m}}. \tag{12.13}$$

This expression does not imply that all the gas particles have this same velocity, but only that this particular average velocity (calculated as the square root of the mean of the squares of individual velocities of the gas particles) is related to the temperature of the gas. In fact, there is a wide range of different velocities of gas particles.

The Maxwell speed distribution, shown in Figure 12.9 for hydrogen molecules, gives the relative numbers of molecules with different velocities in an ideal gas. Note that the possible range of speeds is quite large, from near zero up to quite fast velocities, and that the curve is not symmetrical around the peak value, having a larger "tail" extending to faster velocities. As the temperature of the gas increases (see the figure), the velocity distribution shifts toward higher velocities and there will be more gas particles moving faster. These curves are normalized so that the area under the curves remains constant (equal to one). This explains the diminished peak amplitude as the temperature increases.

Each gas molecule in an ideal gas is considered to behave as a point mass having only translational kinetic energy so that there are three quadratic terms in the expression for the total energy of each particle: $E = \tfrac{1}{2}\, m v_x^2 + \tfrac{1}{2}\, m v_y^2 + \tfrac{1}{2}\, m v_z^2$. It is clear that, because of the isotropic nature of the gas, on average the energy associated with each of these terms is the same, therefore we can identify $\tfrac{1}{2}\, k_B\, T$ worth of internal energy with each quadratic term, each so-called *degree of freedom*, in the energy expression, for a total of $\tfrac{3}{2}\, k_B\, T$ as in Equation (12.12). Although we do not prove that it is true, in more complex cases in which a molecule has additional energy associated with rotational or vibrational motion, for each additional degree of freedom in the energy expression (any quadratic term in a variable) classical thermodynamics dictates that there is an additional $\tfrac{1}{2}\, k_B\, T$ of energy per molecule. This is the *equipartition theorem*, stating that each degree of freedom of a molecule has on average an associated $\tfrac{1}{2} k_B T$ worth of internal energy. The failure of this classical physics theorem at very low temperatures was one of the motivations for the development of quantum mechanics in the early years of the 20th century.

An alternative expression to Equation (12.11) that is perhaps more familiar to the reader from chemistry uses a different unit for the amount of gas rather than the number of molecules. One *mole* (mol) of a substance is defined as that same number of atoms or molecules of the material as there are atoms contained in 12 g of the isotope carbon-12. The name Avogadro's number, N_{A}, is given to this number of atoms or molecules and it is experimentally found that

$$N_A = 6.02 \times 10^{23} \text{ molecules/mol}. \tag{12.14}$$

A *mole* of any substance corresponds to Avogadro's number of molecules. However, whereas a mole of carbon-12 has a mass of 12 g, a mole of another element or molecule will have a different mass, known as its *atomic* or *molecular mass.* Keep in mind that the term mole refers simply to a fixed number of molecules. If we use the symbol n for the number of moles of a gas, then n is simply equal to N/N_A.

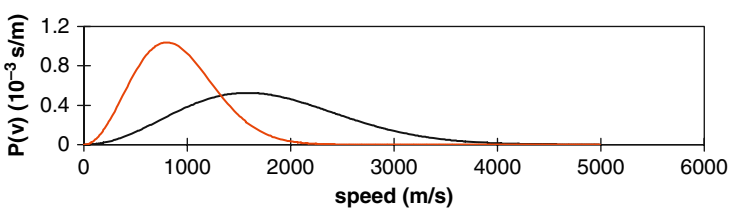


FIGURE 12.9 *Maxwell speed distribution for hydrogen molecules. Black curve is for T = 300 K; red curve is for T = 77 K.*

The ideal gas law can be rewritten by replacing the number of molecules by the number of moles so that we can write

$$PV = nRT, \tag{12.15}$$

where R is the molar gas constant. By comparing Equations (12.11) and (12.15) we see that $nR = Nk_B$, so that $R = N_A\, k_B = 8.31$ J/(mol-K). The molecules of an ideal gas move about independently, only interacting when they come into physical contact (so-called hard sphere repulsion). Although we do not consider more complex non-ideal behavior, we note that at higher densities longer-range interactions between gas molecules become significant and deviations from Equation (12.15) do occur.

Example 12.5 A compressed air tank holds a volume of 0.01 m^3 and is at a pressure of 50 atm (5×10^6 Pa). Taking air to be 80% nitrogen and 20% oxygen, compute the number of moles of air and the density of the air in the tank at 20°C. How many molecules of oxygen and of nitrogen are there in each cm^3 of volume within the cylinder and what is each of their rms velocities?

Solution: Using the ideal gas law, the number of moles of air is given by

$$n = PV/RT = (5 \times 10^6)(0.01)/(8.31)(293) = 20.5 \text{ mol}.$$

Because air contains 80% nitrogen molecules with molecular weight 28 g/mol and 20% oxygen molecules with molecular weight 32 g/mol, the mean molecular weight of air is

$$M = (28)(0.8) + (32)(0.2) = 28.8 \text{ g/mol}$$

and 20.5 mol of air then has a weight of 590 g. The density of air in the tank is then

$$0.59 \text{ kg}/0.01 \text{ m}^3 = 59 \text{ kg/m}^3.$$

Each mole of the air contains N_A molecules, so that there are a total of $(20.5)(6.02 \times 10^{23}) = 1.23 \times 10^{25}$ air molecules in the tank. In 1 cm^3, or 10^{-6} m^3, there are then about 1.23×10^{21} air molecules, 80% (or 9.8×10^{20} molecules) nitrogen and 20% (2.5×10^{20} molecules) oxygen.

According to Equation (12.13), the rms velocities of the molecules depend only on the temperature and the molecular mass. We find the nitrogen molecules move with an rms velocity given by

$$v_{\text{rms}} = [3(1.38 \times 10^{-23})(293)/(28)(1.67 \times 10^{-27})]^{1/2} = 2510 \text{ m/s},$$

whereas the oxygen molecules move with a slower velocity of 2350 m/s because of their larger mass.

4. THE FIRST LAW OF THERMODYNAMICS

The principle of conservation of energy is one of the cornerstones of modern science. For thermodynamic systems, the first law of thermodynamics is a statement of conservation of energy. Recall that the temperature (in K) of an object is proportional to its internal energy per mole or per particle. Temperature is therefore a measure of the concentration of internal energy within the object. When an object is allowed to come into thermal contact with its surroundings, the larger system of (object + surroundings) will eventually reach thermal equilibrium with a uniform temperature, indicating a constant concentration of thermal energy throughout the system. Thus thermal equilibrium can be viewed as that final state at which the internal energy has been

redistributed uniformly. In a macroscopic sense the density of internal energy is uniform at thermal equilibrium. We mention that because there is a distribution of gas particle speeds, as discussed above, there are also microscopic variations of temperature over smaller volumes; these fluctuations of the local temperature have important consequences, for example, giving rise to scattering of light from the gas (a perfectly uniformly ordered system—such as a high-quality gem diamond—will not scatter light so that a beam of light in the system will not be visible).

If a closed system interacts with its surroundings it can increase (or decrease) its internal energy U, and correspondingly its temperature, in two ways: by the inward (outward) flow of heat or by work being done on (or by) the system (Figure 12.10). We use the standard sign convention in which Q (>0) is the heat added to the system from the surroundings and W (>0) is the work done by the system on its surroundings. Then,

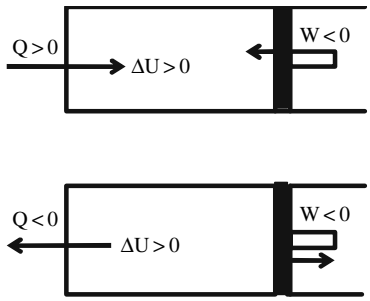


FIGURE 12.10 *A closed thermodynamic system can change its internal energy through heat flow or work.*

Conservation of energy leads to a statement of the first law of thermodynamics,

$$\Delta U = Q - W, \tag{12.16}$$

where attention must be paid to the signs. Negative values of Q or W indicate heat leaving the system or work done on the system, respectively.

Heat is a term that is used for the flow or transfer of internal energy between objects. Thus, an object does not contain heat, but does contain internal, or thermal, energy in proportion to its temperature. Internal energy is a physical property of an object. As we have discussed, when two objects at different temperatures are in thermal contact, heat flows from the hotter to the colder object until thermal equilibrium is established and both objects reach the same temperature or concentration of internal energy. In effect, when two objects are in thermal contact internal energy is redistributed by heat flow until the internal energy concentration is uniform throughout. The amount of heat that flows out of the hotter object is not determined by the object itself, but depends on the thermal properties and temperature of the other object as well, as we show in the next section. Thus, heat is not a state variable.

The amount of work done by or on a system is also not a physical property of the system itself, depending on external macroscopic forces and displacements. Work is also not a state variable. Equation (12.16) is therefore a somewhat strange relationship, stating that a change in a physical property of a system, ΔU, a state variable, can be written as the sum of two physical processes, neither of which is itself a physical property of the system. This is why we do not write the heat flow or work expressions as ΔQ or ΔW, since that notation would imply changes in some state variable, but rather as just Q or W.

If a closed system has no heat flow in or out and does no work, so that $Q = W = 0$, then $\Delta U = 0$ and the internal energy must remain constant; the system is said to be *isolated.* An example of an isolated system is a (perfect) thermos bottle that does not allow any exchange of heat with its surroundings. In general the internal energy of a thermodynamic system can change from both heat flow and work. In the rest of this section, we distinguish several special cases that are of interest. However, to put things in a more concrete fashion, let's first consider the work done by a system composed of an ideal gas.

Imagine that we put n moles of a gas into a cylinder of volume V with a movable piston of cross-sectional area A as shown in Figure 12.11. The gas pressure P exerts a net force on the piston that can cause it to move a distance Δx. In moving the piston a small distance (during which time the pressure can be assumed to be constant) the gas does a small amount of work on the piston given by

$$\delta W = F\Delta x = PA\Delta x = P\Delta V, \tag{12.17}$$

where ΔV is the change in volume of the gas as the piston moves.

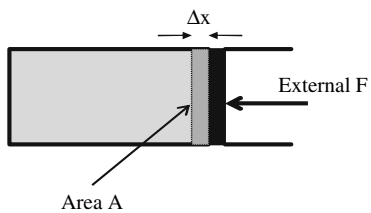


FIGURE 12.11 *A gas of volume V exerting a pressure P on a movable piston of area A with an external force F ensuring a quasistatic expansion of the volume.*

In general the force on the piston would make it accelerate. In order to ensure that all portions of the system remain at thermal equilibrium (so that the pressure and temperature of the gas are the same throughout the volume), we can imagine that an external force is applied on the piston to maintain a *quasistatic process.* That is, the external force is adjusted to keep a zero net force on the piston so it moves at a slow constant velocity. Equation (12.17) is a very general expression for the work done by any fluid but it applies only to a very small change in volume during which the pressure does not change. To proceed further we limit ourselves to the case of an ideal gas.

For an ideal gas we know that the three variables *P, V*, and *T* are connected by Equation (12.15), so that if one is held constant, the other two variables must change in a corresponding manner. We can distinguish four limiting cases of behavior for discussion of the first law.

(1) If the pressure of the gas is held constant, known as an *isobaric* process, then we can simply add up the contributions in Equation (12.17) to find that in general the total work is

$$W = P(V_{\text{final}} - V_{\text{initial}}). \quad \text{(isobaric)}. \tag{12.18}$$

The work will be positive if done by the system on the surroundings so that the final volume is greater than the initial. If the surroundings do work on the gas system decreasing its volume, then the work is negative leading to an increase in internal energy according to Equation (12.16). In such an isobaric process, as the volume varies the temperature will change according to Equation (12.15) and lead to a corresponding change in internal energy. For example, if work is done on the system decreasing its volume then the temperature must drop, according to the ideal gas law. But if only work were done on the system then the first law says that the internal energy would increase. Clearly then there must also be a flow of heat out of the system giving a net decrease in internal energy. In general the internal energy of the gas will change according to the first law because of both work done and heat exchange with the surroundings.

(2) We show in the box below that if instead the temperature is held fixed, known as an *isothermal* process, then the work done for the case of an ideal gas is

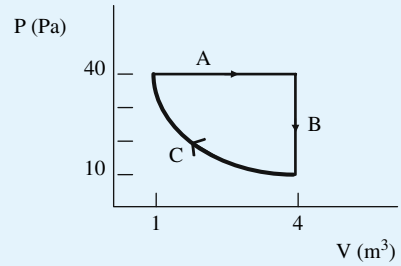
$$W = nRT \ln\left(\frac{V_{\text{final}}}{V_{\text{initial}}}\right), \quad \text{(isothermal)}, \tag{12.19}$$

where ln is the natural logarithm. In this case, because the process is isothermal then in general there can be no change in internal energy and the work and heat flow must be balanced to maintain $\Delta U = 0$. As the gas volume changes slowly in the quasistatic process, heat must flow to maintain the system at a constant temperature. If the gas expands, doing positive work, then heat must flow into the system to maintain the temperature; if work is done on the system causing the gas to compress, then heat must flow out of the system. Equation (12.19) represents the specific result for the work in the case of an ideal gas, but the result that $\Delta U = 0$ holds for all isothermal processes.

(3) If the volume is held constant, an *isochoric* process, then according to Equation (12.17) in general no work is done and the first law reduces to $\Delta U = Q$. In this case the heat flow directly determines the change in internal energy and hence the temperature of the gas.

(4) A fourth process of interest in which none of the variables *P*, *V*, or *T* is held fixed, but in which there is no exchange of heat, is known as an *adiabatic* process. In this case because $Q = 0$, the first law becomes $\Delta U = -W$, and, in general, the work done directly determines the internal energy change or temperature.

Example 12.6 Classify each of the labeled processes A, B, and C shown in the *PV* diagram for an ideal gas and find the work done by the gas for each and the total work for all three. In curve C, $P \sim 1/V$.



Solution: For curve A the pressure remains constant and so the process is an isobaric one. The work done by the gas is then simply, reading values from the graph, $W_A = P\Delta V = (40\text{ Pa})(4 - 1)(\text{m}^3) = 120$ J. The gas does positive work in expanding its volume. Note that this result represents the area under curve A. For curve B, the volume does not change, the process is isochoric, and no work is done, $W_B = 0$. For curve C, because $P \sim 1/V$, PV remains a constant and the process is isothermal. In this case the work done by the gas is given by Equation (12.19), representing again the area under the curve, and is $W_C = nRT \ln(1/4)$. Using the ideal gas law we can rewrite this as $W_C = PV \ln(1/4) = 40 \ln(1/4) = -55.5$ J, after we read the product of PV from the graph. The net work for this complete cycle, returning to the same values of P and V is the sum of our three values, $W_{net} = W_A + W_B + W_C = 64.5$ J, and represents the area enclosed by the three curves.

Many interesting situations are described by one of the above four limiting cases, but it is also commonly the case that an overall process is either a sequential combination of those cases or is still more complex. We introduce some other combinations of thermodynamic variables below that are useful for the study of chemical reactions and biological systems.

If we attempt to apply the first law to warm-blooded living organisms, we see that Q will be less than zero under ordinary situations, because body temperature is normally above ambient temperature. Also W is usually also positive because living organisms generally do work on their surroundings rather than have work done on them. Thus, under normal circumstances $\Delta U < 0$ for living organisms and their temperatures would seem to necessarily approach that of the surroundings. The flaw in this argument is that living creatures are not closed systems, but constantly exchange mass with their surroundings, whether it is in the form of gases, nutrients, or waste products. This exchange of mass and its metabolism supplies the necessary chemical energy to maintain life. We return to this discussion in the next chapter when we introduce the second law of thermodynamics.

From the very statement of the first law of thermodynamics, it is clear that internal energy, work, and heat all can be measured in the same energy units (e.g., joules). Historically there are other commonly used units for heat that should be mentioned. Before it was realized that heat is the flow of thermal energy, it was believed to be the flow of a substance that was called caloric and measured in units

Our expression, Equation (12.17), for the work done by a fluid system for an infinitesimal volume expansion can be written replacing ΔV by dV and then integrating to write a general expression for the total work done as

$$W = \int_{V_{initial}}^{V_{final}} P\,dV.$$

According to this very general result, the work done is the area under a curve representing the pressure variation of a system as a function of its volume. For an ideal gas, we know that the pressure varies as the volume changes according to Equation (12.15), so that we can write

$$W = \int_{V_{initial}}^{V_{final}} \frac{nRT}{V}\,dV.$$

In an isothermal process for which the temperature T is a constant and remembering that

$$\int \frac{dx}{x} = \ln x,$$

we can write that the work is given by

$$W = nRT \ln V \Big|_{V_{\text{initial}}}^{V_{\text{final}}}$$

which results in Equation (12.19).

of calories (cal). James Prescott Joule in the 1840s first showed that heat could do mechanical work and established the *mechanical equivalent of heat*, the heat required to raise the temperature of 1 g of water by 1°C (specifically from 14.5 to 15.5°C), known today to be

$$1 \text{ cal} = 4.186 \text{ J}. \tag{12.20}$$

This value varies slightly (by less than 1%) as the water temperature is changed within 0 to 100°C. Other units used specifically for measuring heat are the kilocalorie (1 kcal = 4186 J = 1 Cal (with a capital C) note that the Cal is the unit used in reporting energy content on packaged food) and the British Thermal Unit (1 BTU = 1055 J, still used predominantly in engineering).

5. THERMAL PROPERTIES OF MATTER

When heat flows into or out of a material its internal energy and temperature will change. For a given material, it is found experimentally that the amount of heat needed to produce a temperature change of ΔT is proportional to both the mass of material and to the temperature change and is given by

$$Q = cm\Delta T, \tag{12.21}$$

where c is called the specific heat of the material given in units of J/(kg-K) or kcal/(kg-°C). A block of material with twice the mass of another made from the same substance will require twice the heat transferred to it in order to warm both blocks by the same temperature. The specific heat of a material is actually dependent on its detailed electronic structure and can be calculated using quantum mechanics. It is a measure of the heat release or absorption capability of the material as the temperature changes. Temperature changes correspond to internal energy changes, and in a simple way we can understand these for a solid to be due to changes in the potential energy of molecules bound by effective springs. Thus the specific heat is related to the potential energy of interaction represented by these springs.

Specific heats for most materials are dependent on the temperature, but vary slowly near room temperature and can often be assumed constant. Table 12.3 lists the specific heats of several materials. Those with higher specific heats require

Table 12.3 *Specific Heats of Various Materials*

Material	*Specific Heat* *kcal/kg-°C*	*J/kg-°C*
Aluminum	0.22	900
Copper	0.093	390
Glass	0.20	840
Human body (mean at 37°C)	0.83	3500
Ice (−5°C)	0.50	2100
Iron or steel	0.11	450
Mercury	0.033	140
Silver	0.056	240
Steam (110°C)	0.48	2010
Water	1.00	4186
Wood	0.4	1700

more heat per unit mass in order to increase their temperature than other materials or, in turn, give off more heat per unit mass when their temperature drops. Water has one of the highest specific heats of all substances making it a valuable source of heat, for example, in hot water heaters and in our bodies.

Example 12.7 A liter of tea at 100°C is poured into a glass-lined thermos bottle at room temperature (20°C). If the glass bottle has a mass of 0.2 kg, find the final temperature of the tea in the sealed thermos.

Solution: Heat will flow from the tea (water) to the glass until the two are in thermal equilibrium at the same final temperature T. We can write that $Q_{\text{loss from tea}} = Q_{\text{gain to glass}}$, so that $c_{\text{water}}\, m_{\text{water}}\, (100 - T) = c_{\text{glass}}\, m_{\text{glass}}\, (T - 20)$. Using values in Table 12.3 for the specific heats and the density of water, we have that

$$(1)(1\ \text{g/cm}^3)(1000\ \text{cm}^3)(100 - T) = (0.2)(200\ \text{g})(T - 20)$$

Solving for T we find that $T = 96.9$°C. The relatively large specific heat of water results in a final temperature much closer to the water starting temperature.

Our discussion so far has been limited to materials not changing their phase, remaining either solid, liquid, or gas. Because electronic interactions are dramatically different for a material depending on its phase, we expect that the thermal properties of a material will strongly depend on what phase it is in. In fact as a solid is melting, heat must be input to break the orderly bonding in the solid to form the liquid and during this melting transition the temperature does not change. For example as a block of ice at 0°C melts, the water–ice mixture remains at 0°C until the ice is totally melted. Additional heat added will then increase the temperature of the water. The heat needed to change the phase of a unit mass of material is known as the *heat of transformation*. Possible transformations are shown in Figures 12.12 and 12.13.

The corresponding amount of heat per unit mass required for the transformation is known as a *latent heat L*, where

$$Q_{\text{transformation}} = Lm. \tag{12.22}$$

We distinguish between latent heat of fusion (for melting or freezing), latent heat of vaporization (for evaporation or condensation), and latent heat of sublimation (for phase changes directly from solid to gas, as in solid CO_2 known as

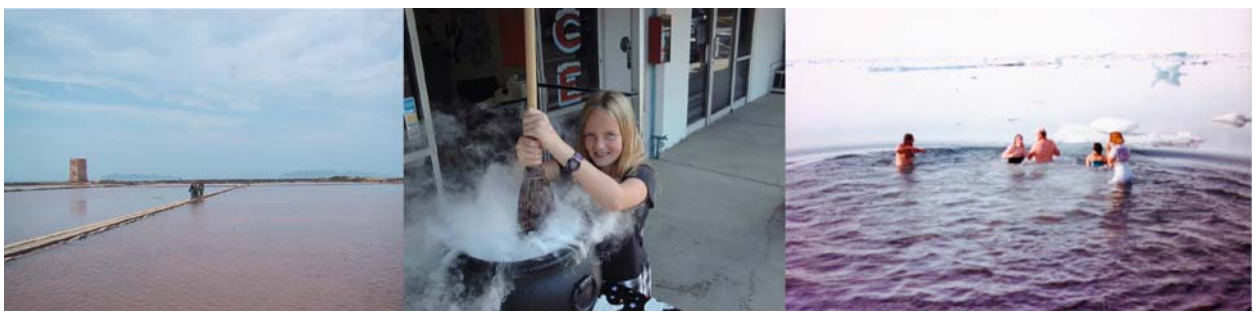


FIGURE 12.12 *Phase changes of water (from* left*) Evaporation in salt flats, sublimation of dry ice, and freezing water!*

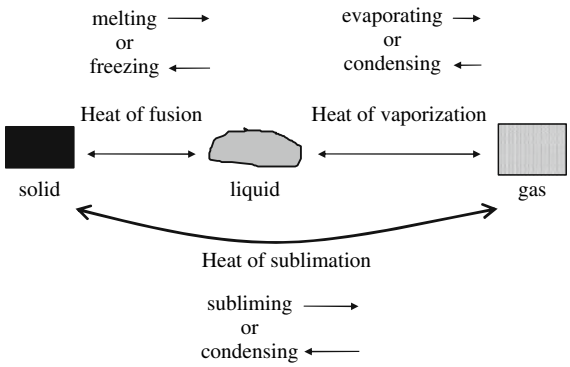


FIGURE 12.13 *Possible phase changes and their associated heats of transformation.*

"dry ice"). Each of these processes is reversible in terms of energy requirements; that is, the amount of heat required to melt a block of ice to water is the same as the heat given off when that same mass of water freezes to ice. Table 12.4 lists some materials with their melting and boiling point temperatures, together with their corresponding latent heats.

In more complex systems, other phase transitions are possible. For example, in biological membranes there are specific transitions that occur at particular temperatures in which the lipids and proteins in the membrane arrange themselves in more or less well-ordered states. These transitions also involve latent heats that can be determined from thermodynamic measurements.

Table 12.4 *Latent Heats of Various Substances*

Material	*Melting Point (°C)*	*Heat of Fusion (kJ/kg)*	*(kcal/kg)*	*Boiling Point (°C)*	*Heat of Vaporization (kJ/kg)*	*(kcal/kg)*
Helium	—	—		−269	25	6.0
Nitrogen	−210	25.7	6.1	−195.8	200	48
Ethyl alcohol	−114	104	24.8	78	854	204
Mercury	−39	11.3	2.7	357	296	71
Water	0	333	79.7	100	2260	539
Carbon dioxide	−79	Sublimates		—	578	138
Aluminum	660	399	95.3	2467	10550	2520
Tungsten	3410	184	44	5660	4940	1180

Example 12.8 Construct a quantitative graph showing the heat input to a 100 g block of ice at −20°C as a function of its temperature as the ice first warms, melts to water, heats to its boiling point, vaporizes, and heats to 150°C.

Solution: Starting with the ice at −20°C an amount of heat equal to

$$c_{\text{ice}} m\Delta T = (0.5 \text{ kcal/kg-°C})(0.1 \text{ kg})(20°\text{C}) = 1 \text{ kcal}$$

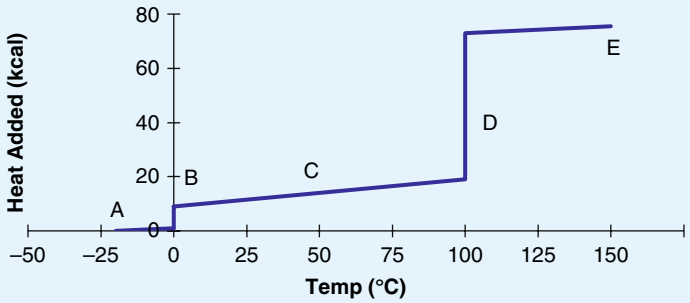
is needed to bring the ice to T = 0°C. This is plotted as the straight line labeled A below; the temperature rise is proportional to the heat added over this temperature range. As additional heat is added, the ice melts and the temperature remains at 0°C until the ice is completely melted. Because the latent heat of fusion of water is 79.7 kcal/kg, a total additional amount of heat equal to 79.7 (0.1 kg) = 8.0 kcal is needed to melt the ice. This portion of the graph is a vertical line labeled B because there is no temperature change. Once the ice has melted, any additional heat added will warm it according to Equation (12.21). To raise the water temperature to 100°C requires additional heat equal to

$$Q = c_{\text{water}} m\Delta T = (1 \text{ kcal/kg-°C})(0.1 \text{ kg})(100°\text{C}) = 10 \text{ kcal}.$$

This portion of the graph is plotted as the line labeled C.

Once the water is all at 100°C, additional added heat will cause it to boil and change to steam. To vaporize all the water requires (539 kcal/kg)(0.1 kg) = 54 kcal of heat and this part of the graph is drawn as the vertical line labeled D.

To now heat the trapped steam even further to 150°C, additional heat needs to be added according to the specific heat of steam of 0.48 kcal/kg-°C for a total amount of (0.48)(0.1 kg)(50°C) = 2.4 kcal. This final line of the graph is labeled E. In the graph note that the largest heat inputs occur during the phase changes, particularly the evaporation.



Next, we introduce a new state variable, the *enthalpy*, H, that can be used to characterize chemical bond energies and heats of chemical reactions and is defined as

$$H = U + PV. \tag{12.23}$$

We show in the box that for a process that occurs at constant pressure, the change in enthalpy is equal to the (reversible) heat flow between the system and its surroundings, or

$$\Delta H = Q. \quad \text{(isobaric process).} \tag{12.24}$$

Why do we bother to introduce H, if under isobaric conditions its change is just equal to Q? Recall that Q is not a property of a system, not a state variable, but rather depends on the system and its surroundings. On the other hand, H is a well-defined property of a system (because U, P, and V are well-defined state variables). Therefore under isobaric conditions, which are fairly common, enthalpy changes tell us about heat flow during the process.

From the definition of H (Equation (12.23)), we find that in general $dH = dU + PdV + VdP$.

Using the first law of thermodynamics for the case when only pressure–volume work is involved ($dU = Q - PdV$) and noting that if the pressure is constant ($dP = 0$), we have $dH = Q - PdV + PdV = Q$, which is rewritten as Equation (12.24).

Enthalpy may be used to characterize a chemical reaction, such as

$$A + B \rightarrow C + D,$$

where each of the reactants (A and B) and products (C and D) are characterized by an enthalpy and the net change in enthalpy.

$$\Delta H = H_C + H_D - H_A - H_B$$

is an important piece of information about the energetics of the reaction. If ΔH is positive, the reaction is called *endothermic*, with a net absorption of heat to the system, whereas if ΔH is negative, the reaction is called *exothermic*, with a net liberation of heat. Endothermic reactions require input of energy to occur although exothermic reactions may occur spontaneously.

The strength of chemical bonds may be measured by their enthalpy, in this case a measure of the energy required to break the bond. Using tabulated values (see Table 12.5) one can estimate the total bond energy of any particular molecule by adding up the individual bond energies. This procedure works quite well in many situations although there are some notable exceptions. One such exception is the benzene ring that has a lower overall energy than one would calculate from the individual bonds (three single C—C, three double C═C, and six C—H bonds) due to "resonant energy," a quantum mechanical phenomenon that stabilizes the ring structure compared to a linear molecule.

Table 12.5 *Average Bond Dissociation Energies*

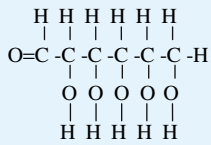
Bond	*ΔH (kcal/mol)*
C—C	83
C═C	146
C≡C	200
C—H	99
C—N	70
C—O	86
C═O	178
N—H	93
O—H	111
O—O	119

Example 12.9 Estimate the enthalpy change for the synthesis of glucose ($C_6H_{12}O_6$) from carbon dioxide and water. This is the most important result of photosynthesis in green plants. The overall chemical reaction is

$$6\ CO_2 + 6\ H_2O \rightarrow C_6H_{12}O_6 + 6\ O_2.$$

Solution: We solve this problem by estimating, using Table 12.5, the energy needed to break all the bonds of the starting reagents and form all the bonds of the products. To break all of the 12 C═O bonds of CO_2 requires $12 \times 178 = 2136$ kcal/mol. Similarly, to break all 12 of the O—H bonds of water requires $12 \times 111 = 1332$ kcal/mol, for a total of 3468 kcal/mol required to break the bonds of the starting reagents.

In forming glucose, with all its bonds as shown, the following energies are liberated: $5 \times 83 = 415$ kcal/mol, for the C—C bonds along the linear backbone of the molecule; $7 \times 99 = 693$ kcal/mol, for the C—H bonds; $5 \times 111 = 555$ kcal/mol, for the O—H bonds; $5 \times 86 = 430$ kcal/mol, for the C—O bonds; and $1 \times 178 = 178$ kcal/mol, for the C═O bond. To this must be added the $6 \times 119 = 714$ kcal/mol, for the O—O bonds in the O_2 molecules, so that the total energy released in the product formation is 2985 kcal/mol. The net heat of formation is then given by the difference between the 2985 kcal/mol liberated and the 3468 kcal/mol needed for bond dissociation yielding a value of 483 kcal/mol. According to the calculation the reaction is endothermic, or heat consuming, and requires energy input whereas the reverse reaction, the "burning" of glucose to form carbon dioxide and water is exothermic, or heat releasing, and can occur spontaneously. The actual energy required in the formation of glucose is 673 kcal/mol; this crude calculation underestimates the correct answer by about 30%.



Net values for enthalpies can be measured using a technique called *calorimetry* in which the heat input or output is determined as a chemical reaction proceeds. The particular values of enthalpy actually depend on temperature, pressure, and other experimental conditions. A modern version of such measurements is the *differential scanning calorimeter* in which the heat input or output is measured as the temperature is scanned. This technique is a sensitive way to detect phase transitions in biopolymers (Figure 12.14).

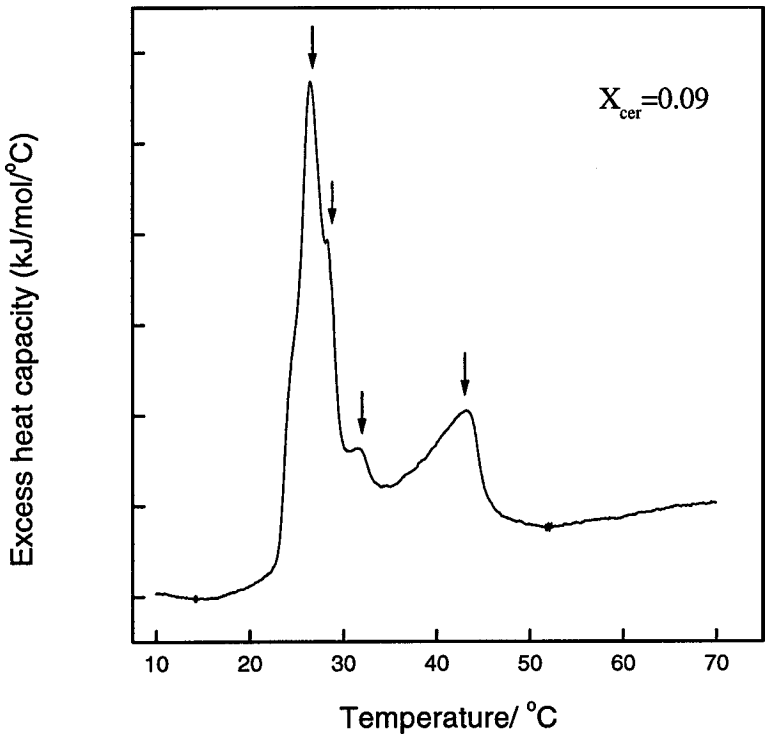


FIGURE 12.14 *Differential scanning calorimetry of a particular lipid bilayer showing four distinct peaks of endothermic activity.*

6. VAPOR AND OSMOTIC PRESSURE; MEMBRANE TRANSPORT AND THE KIDNEY

In this section we take up several related properties of solutions, collectively known as *colligative properties*, that deal with the thermodynamic effects of the addition of small amounts of solutes to a solvent. These dilute solution effects can be described using a formalism quite similar to that of an ideal gas because the individual solute molecules do not interact with each other. Because of this, the form of the equation of state is seen to be similar to that for the ideal gas.

Before we consider the effects of solute molecules on the properties of a liquid, we need to first briefly discuss thermal effects at a boundary surface of a pure fluid, such as water. Imagine a cup of warm water exposed to air at room temperature. Because the water molecules move about with a distribution of velocities, those with rapid velocities that reach the surface may have sufficient energy to escape from the liquid surface, entering the gas phase in a process known as *evaporation.* As only the most energetic molecules escape from the surface, the remaining water molecules have a lower average energy and the water thus cools by evaporation. We note here that this process is responsible for cooling our bodies by evaporation when we sweat. Furthermore, this same process also occurs when a liquid in a container is cooled by "pumping," or by attaching the container to a vacuum pump that pulls the faster molecules from the surface of the liquid. This method can be used to cool liquids well below the ambient temperature, often even freezing them.

If a thermos bottle is partially filled with warm water and sealed so that no heat is lost to the outside, the water molecules will evaporate until eventually an equilibrium is reached in which the number of molecules evaporating from the surface equals the number of gas molecules that collide with the water surface, giving up most of their energy and condensing to the liquid phase. At this point the pressure of the gas phase is known as the *equilibrium* (or *saturated*) *vapor pressure.* The value of this pressure depends only on the temperature of the liquid, and not on the volume above its surface. A larger volume would cause more evaporation to occur but would arrive at the same final equilibrium vapor pressure.

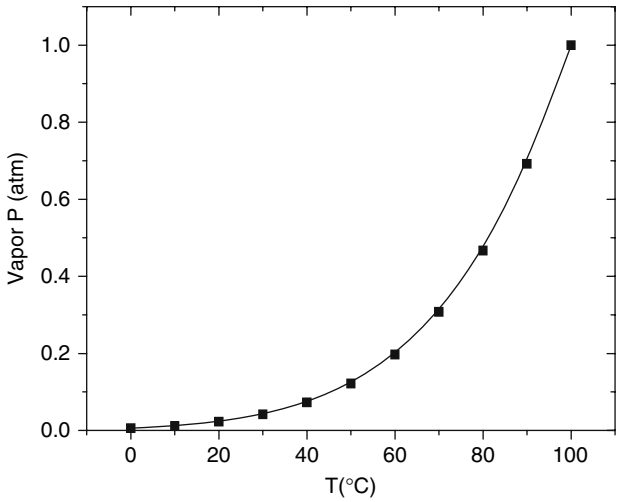


FIGURE 12.15 *The temperature dependence of the vapor pressure of water.*

Suppose we return to our cup of warm water and now heat the water. As the temperature rises, more evaporation occurs and the vapor pressure near the surface rises. When the vapor pressure exceeds the ambient pressure on the liquid surface (from atmospheric pressure, unless the container is sealed), boiling occurs. Bubbles filled with vapor form in the liquid and expand and rise to the surface. As long as the vapor pressure is at least equal to the ambient atmospheric pressure at the surface bubbles can support themselves against the external pressure. The vapor pressure of water is equal to atmospheric pressure at 100°C (at sea level), so that water will boil at this temperature. If we were hiking at an elevation of 3000 m rather than at sea level, atmospheric pressure is only about 70% that at sea level and water will boil at a lower temperature. The temperature dependence of the vapor pressure for water is shown in Figure 12.15, where it can be seen that at a pressure of 0.7 atm water will boil at about 90°C. The lower boiling temperature at high elevations requires foods to be cooked for a longer time (Figure 12.16). Remember that no matter how much heat flows into the water, its temperature will not rise above the boiling point. A pressure cooker is designed to increase the boiling temperature of water, in order to speed cooking. The higher pressure inside the cooker raises the vapor pressure and allows boiling to occur at a higher temperature.

Now that we have an appreciation of vapor pressure and boiling, let's consider what happens when salt is added to the heated water. Experimentally it is found that the vapor pressure of the solvent (water in our case) decreases when a solute (salt) is added. You cooks out there will recognize this, because salt is often added to rapidly boiling water to quench the boiling. Quantitatively, for dilute solutions it is found that the vapor pressure decreases according to Raoult's law,

$$P = XP_0, \tag{12.25}$$

where P_0 is the vapor pressure of pure solvent and X is the mole fraction (fraction of total number of moles) of the solvent. The decrease in vapor pressure can be understood in this ideal case as simply due to the decreasing mole fraction

FIGURE 12.16 *Why does it take longer to cook foods at higher elevations (here in the Andes of Equador)?*

represented by the volatile solvent. The nonvolatile solute does not contribute to the vapor pressure. As a consequence of the reduced vapor pressure with solute present, a higher temperature is needed before the vapor pressure equals atmospheric pressure and boiling occurs. The boiling point rises in the presence of a solute by an amount proportional to the concentration of solute. Salt added to boiling water will stop the boiling; continued heating of the water will lead to boiling again but at a higher temperature. An exactly analogous situation occurs in the process of melting from a solid to a liquid or freezing of a liquid to a solid. In the presence of a solute the freezing point of a liquid is lowered. This important phenomenon helps keep sea water from freezing. It is also the basis for salting roadways that are covered with freezing water to prevent icing.

Another colligative property, particularly important in biology, is *osmotic pressure.* Osmotic pressure is a solution phenomenon quite akin to vapor pressure. We have seen that a decreased vapor pressure occurs at an air–solution boundary because the solute is not volatile. An analogous situation can occur in solution if there is a semipermeable membrane present that allows water to freely pass through but has pores too small to allow a solute to penetrate. The membrane acts as if it were the air–solution interface.

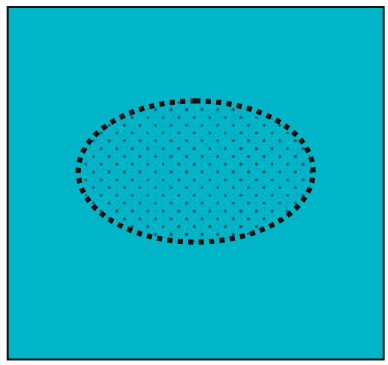


FIGURE 12.17 *A semipermeable membrane containing a protein solution immersed in pure solvent and swollen with water due to an elevated osmotic pressure.*

To give a more concrete example, consider the situation when a sealed semipermeable membrane (not unlike that used to encase sausage or hot dogs) containing a protein solution is immersed in pure water (Figure 12.17). The membrane allows the exchange of water but keeps larger proteins from leaving. Because the water inside the tube is at a lower concentration than the pure water outside, its pressure is decreased somewhat, just as the vapor pressure would be. As a result water enters the membrane and it swells until reaching an equilibrium at which the water pressure is the same across the membrane (analogous to equilibrium vapor pressure). Although the water pressure is equal across the membrane, the protein (solute, in general) exerts a pressure as well and so the internal pressure within the membrane is greater.

The osmotic pressure π is defined as the pressure difference across the membrane at equilibrium and can be shown to satisfy

$$\pi = \frac{nRT}{V}, \tag{12.26}$$

where n is the number of moles of protein, V is the volume of the solution within the membrane, and the protein is assumed to be dilute. This equation, known as the van't Hoff law for osmotic pressure, is just the ideal gas law that surprisingly works in the dilute solution case because the ideal noninteracting proteins behave as an ideal gas within the water.

The flow of water due to osmotic pressure differences is known as *osmosis.* The swelling that occurred in our example is a process that occurs in biology and is known as osmotic shock. Cells immersed in a solution with lower ion content (hypotonic) will swell, leading eventually to a rupture of the cell membrane. Such a process is commonly used in biochemistry to disrupt cells. (It is precisely the same phenomenon that swells hot dogs in boiling water causing them to split open.) There are a number of other important applications of osmosis in biology and medicine; we discuss two of them here, namely dialysis in the laboratory and the functioning of the kidneys, including kidney dialysis.

Dialysis is a technique used in biochemistry in a way similar to our example of the swelling membrane in order to change the solvent in which macromolecules are immersed. The starting solution of macromolecules is sealed in a dialysis tube (semipermeable membrane) and bathed in a large volume of the desired final solvent for a long period of time. Typically both solvents are water-based but contain different small ions that are also free to pass through the membrane whereas the macromolecules cannot. After some time, depending on the sample and external volumes, the outer solvent is replaced by a fresh large volume, and after several changes of outer solvent the inner solvent has been essentially completely replaced while the macromolecules remain inside.

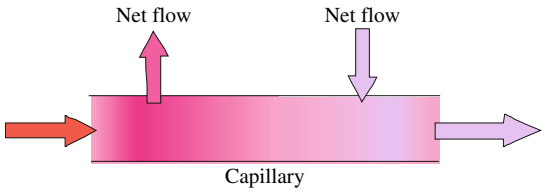


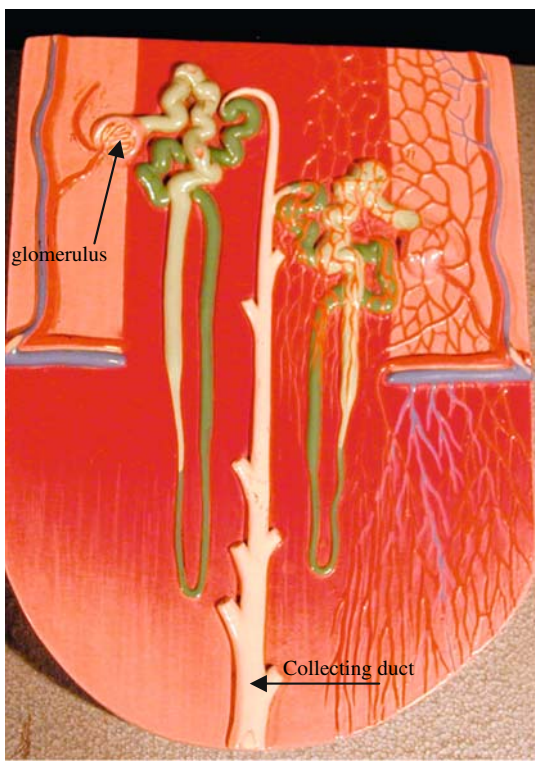
FIGURE 12.18 *Blood flow and osmosis through capillary walls.*

Osmosis is an important factor in the exchange of blood gases and small molecules such as sugars through the capillary walls. As we have seen in Chapter 9, the total hydrostatic pressure in the capillary drops from the arterial end to the venous end, causing blood to flow through the capillary. The osmotic pressure inside the capillary is about 20 torr higher than outside the capillary. This osmotic pressure difference results in a higher internal than external capillary pressure in the first (arterial) half of the capillary, and a higher external than internal pressure in the second (venous) half of the capillary. Accordingly there is a net outward flow of fluid during the arterial half and a net inward flow during the venous half of the capillary (Figure 12.18). Clearly these osmotic flows aid in distributing nutrients and oxygen and in collecting wastes and carbon dioxide.

This passive control of the exchange of small molecules through the pores of the capillaries can be disrupted in many ways leading to edema, a swelling due to excess fluid buildup in the tissues. An abnormally low blood protein level can reduce the osmotic pressure sufficiently to increase the net outward flow of fluids from the capillaries. This can occur in diseases of the kidney in which the nephrons, the basic building blocks of the kidney (see below), become permeable to larger macromolecules so that protein is lost in the urine, or in diseases of the liver, leading fluids to collect in the abdomen. Other possible sources of edema include right heart failure in which the returning blood is not processed fast enough to avoid backing up fluids, and injury or infection of tissue, in which the capillaries dilate increasing blood flow and leakage of fluids.

The kidney, a vital organ of the body, maintains and regulates the solute composition in the blood plasma. Consisting of a collection of about one million independently functioning units called nephrons, each kidney filters an incredible 850 L of blood every day in order to remove waste products. All of the blood in the body is thus processed every five minutes throughout one's entire life! Simply put, each nephron consists of two functional parts: the Bowman's capsule, containing the glomerulus, and the tubule, a relatively long (2–4 cm) duct with walls that are only a single membrane thick (Figure 12.19). Arterial blood passes by the glomerulus, consisting of an extensive membrane (with a total area of several m^2) serving as a semipermeable filter. This membrane has pores (with diameter <5 nm) making up roughly 5–10% of the surface and allowing a huge volume of blood to be processed rapidly. A set of arterioles regulates the blood pressure within the glomerulus so that the hydrostatic pressure is high (~70 torr).

FIGURE 12.19 *Schematic model of a nephron.*



From the large volume of blood processed, roughly 180 L of filtrate (consisting of blood plasma and low molecular weight solutes, but no blood cells) passes through the membrane and is collected in the tubules of the nephrons each day. This volume vastly exceeds that excreted each day in the urine (about 1.5 L). For example, the glomerulus membrane filters out roughly 2.5 pounds of sodium chloride daily with all but 5–10 g being reabsorbed by the capillary bed. The kidneys thus function by massive filtration and reabsorption as part of an extremely sensitive control mechanism, believed to be controlled by active transport, which rapidly regulates the solute balance in the blood plasma.

Kidney failure can be due to a variety of causes, including nephron destruction over a period of time, too high a permeability of the glomerulus membrane, or failure of the active transport mechanism in the tubules preventing reabsorption of specific solutes. The artificial kidney, or *renal dialysis*, may then be used to control solute levels in the blood plasma. In principle, the artificial kidney is quite simple (Figure 12.20). It relies on passive diffusion through a semipermeable membrane and to be effective, it must work in a period of a few hours to balance solute concentrations, removing wastes. Continuous flow

FIGURE 12.20 *A hemodialysis unit with a detail of the actual dialysis filters* (center) *and its schematic use.*

filtering can be done, but at much slower rates (~0.2 L/min) than in the kidney (~5 L/min). To keep the blood from clotting, heparin, an anticoagulant, is added as the blood enters the hemodialysis unit but is neutralized by the addition of protamine as the blood returns to the body. Although renal dialysis is a successful therapy, about 50% of those on dialysis will develop a critical cardiovascular problem.

7. HEAT TRANSFER MECHANISMS

This section discusses the three basic ways in which heat can be transferred from one object to another: conduction, convection, and radiation. Before discussing each in some detail let's introduce the major concepts. Our bodies are heat engines, fueled by the food we eat. Of all the energy gained in the metabolism of our food, we use less than 20% of the energy generated to do work while dissipating roughly 80% as heat. For a typical adult, just lying at rest generates about 90 kcal/h, the *basal metabolic rate.* Any activity will increase this rate (see Table 12.6).

Table 12.6 *Metabolic Activity Rates*[*]

Activity	*Heat Production Rate* *kcal/h*	*W*
Sitting at rest	100	115
Slow walking	225	260
Cycling (15 km/h)	360	420
Climbing stairs (2/s)	600	700
Running (15 km/h)	1000	1150

[*] Based on typical 65 kg person.

In a 24 h period, the basal metabolic rate generates about 2100 kcal of heat that needs to be removed from the body. If this heat were not removed, we can estimate that the average body temperature rise, given by $\Delta T = Q/(m\ c_{\text{body}})$, would be dramatic. Using the specific heat of water and a mass of 80 kg, the temperature rise would be about 1°C per hour. How does the body get rid of this heat so that its temperature remains relatively stable?

Passive conduction, in which heat travels through the body tissue to its surface just as it does along a metal frying pan handle when the pan is heated, is not efficient enough because the body tissue is not a very good thermal conductor. Instead heat is carried near the surface of our bodies by the blood, acting as a convective medium to transport heat just as air from a hot oven does by bulk air currents when the oven door is opened. We then lose heat from our capillary beds near the skin surface by conduction through the relatively thin skin layer. Finally, the heated surface of skin loses energy through a variety of possible

processes, including convective losses from circulating air, sweating, and from the emission of thermal radiation, discussed below.

The body has a variety of involuntary safety mechanisms that attempt to regulate heat losses. When we are overheated either from exercise that generates excess heat or from a high air temperature, there is an involuntary shift in blood flow so that more blood flows near the body surface. This is the source of the typical skin reddening that occurs from heavy exercise. Sweating is also a mechanism to cool the body surface by evaporation. When cold, the body involuntarily attempts to maintain the temperature of the body core by reducing the blood flow to the body surface and to the limbs in general. This is the origin of the danger of frostbite especially to toes and fingers. Another mechanism for warming the body is involuntary shivering, designed to burn more fuel in the muscles in an attempt to maintain a constant internal temperature. We obviously also can consciously control heat loss or gain by the clothing we wear.

In the rest of this section we consider each of the three heat transfer mechanisms in some more detail. We begin with *heat conduction.* Imagine a rod, of length L and cross-sectional area A, with both ends held at different fixed temperatures (hot and cold), T_H and T_C. Individual molecules in the rod do not travel long distances, but rather oscillate about fixed equilibrium locations. Hotter molecules oscillate more rapidly and transfer some of their energy to cooler neighboring molecules via collisions. Heat will flow from the hot to the cold end through the huge number of molecular collisions that transfer energy along the rod. After reaching a steady state, the temperature along the rod will not change with time but will vary linearly with distance along the rod between temperatures T_H and T_C (Figure 12.21).

What factors determine the rate at which heat is transferred along the rod? The larger the cross-sectional area A of the rod, the more rapidly heat can be transferred because of the increased area for collisional transfer of internal energy. We might also expect that the rate of heat conduction along the rod depends linearly on the thermal gradient (or variation of temperature along the rod $\Delta T/\Delta L$) so that either a larger temperature difference between the ends of the rod or a shorter rod will lead to an increased thermal conduction rate. Finally, we expect that the conduction rate will depend on an intrinsic thermal property of the material, known as the *thermal conductivity k.* Putting these factors together, the thermal conduction rate is given by

$$\frac{Q}{t} = kA\left(\frac{T_H - T_C}{L}\right). \tag{12.27}$$

The thermal conductivity varies with material due to different efficiencies of the collision mechanism in transferring energy. Metals are good *thermal conductors* for the same reason they are good conductors of electricity: they have large numbers of relatively unbound (free) electrons that can diffuse about making collisions effectively to transfer energy. Materials such as styrofoam or down are poor thermal conductors, also known as *insulators.* Table 12.7 lists thermal conductivities of a variety of materials. Note that air is a very good thermal insulator. Animals and humans make use of this to

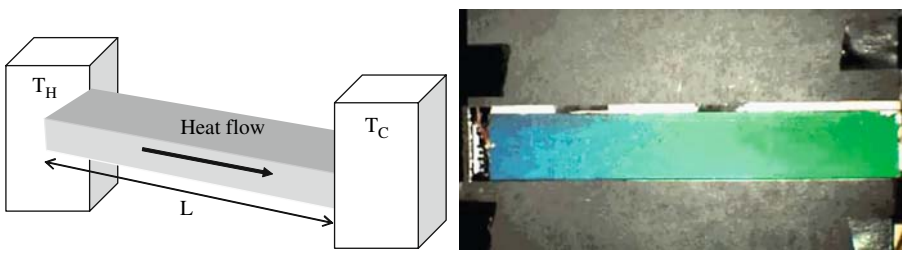


FIGURE 12.21 *Thermal conduction in a bar:* (left) *schematic and* (right) *photo of an aluminum bar kept at fixed temperatures at each end with liquid crystal color coding of temperature.*

keep warm in cold weather by trapping air in fur or feathers or in clothing or blankets (using down, e.g.) or in double-paned glass windows that have an air gap between the panes. In acting as a good insulator, air must be trapped so as to avoid *convection*, or bulk flow of matter that carries thermal energy, discussed next after an example.

Table 12.7 *Thermal Conductivities of Various Materials*

Material	*Thermal Conductivity (kcal/s-m-°C)*
Water	1.4×10^{-4}
Air (dry)	0.06×10^{-4}
Body tissue	0.5×10^{-4}
Fiberglass	0.1×10^{-4}
Down	0.06×10^{-4}
Glass	$\sim 2 \times 10^{-4}$
Metals:	
Steel (stainless)	3.3×10^{-2}
Aluminum	5.6×10^{-2}
Copper	9.6×10^{-2}
Silver	10×10^{-2}

Example 12.10 How much energy is lost in 1 h by conduction through a single-pane glass window that is 1.4 × 1.1 m and is 0.5 mm thick if the outside temperature is 0°C and the inner surface temperature is 18°C?

Solution: The temperature gradient across the window is $(18°C/0.5 \times 10^{-3}\ m) = 3.6 \times 10^4$ °C/m and the cross-sectional area is 1.54 m^2. The rate of heat conduction is then given by $Q/t = (2 \times 10^{-4})(1.54)(3.6 \times 10^4) = 11.1$ W, so that in 1 h (or 3600 s) 4.0×10^4 J or 40 kJ of energy would be conducted through the glass.

As mentioned earlier we regulate our body temperature predominantly by control over convective blood flow. The temperature of the human body is both nonuniform, being warmer in the core than the limbs and surface, and fluctuates in time over a day by about 1°C. When the body is cold, muscles around the elastic veins constrict (vasoconstriction) limiting blood flow near the body surface to reduce heat losses. In fact, the body has two venous return paths to the heart, one deep in the body and one near the body surface with a "valve" controlling which path is used. When you are overheated, the valve to the superficial veins is opened and these veins also dilate (vasodilation) causing the skin to become "flushed" allowing an efficient heat exchange with the surroundings to cool the blood.

To conserve heat more efficiently, the body uses a system of "countercurrent heat exchange" in which the major core arterial and venous blood vessels in the limbs (with flows in opposite directions, hence the term countercurrent) exchange heat with each other. The returning cooled venous blood is thus warmed before reaching the body core, and the warmer arterial blood is cooled so that less heat is lost in the limbs. Without such a mechanism, cooler blood from the limbs would need to be warmed from within the body core and the limbs would be warmed beyond their need, draining heat from the body core.

Fluids that are not at a uniform temperature, due to external heating or cooling, for example, will flow because of differences in the density of the fluid as a function of its temperature. As the fluid is heated it expands and the decrease in density makes that region of fluid more buoyant. *Thermal convection* is the flow of heat via the bulk flow of a fluid. Convection currents are very common on the Earth. Winds and ocean currents

are major factors determining the weather. Forced convection, by use of a pump or fan, is a common way to heat or cool a system: convection ovens speed cooking, hot air heating system fans circulate heated air in a house, cooling fans in computers and other electronic equipment keep devices from overheating, and the water pump and radiator fans in a car cool the engine by convection of water and air, respectively.

Heat transfer by conduction through an object or by convective flow of a fluid clearly requires the presence of molecules. Heat transfer by *thermal radiation* (no relation to nuclear radiation) can occur through a vacuum, in the complete absence of matter. Radiation refers to the transfer of photons, the elementary quanta of electromagnetic radiation, between objects at different temperatures. All objects emit radiation. Hotter objects emit radiation that is visible to our eyes, such as a hot toaster coil, embers in a campfire, or the sun. Objects need to reach about 1000 K before they emit a visible red glow due to the emission of photons with an energy that we interpret as red light. At progressively higher temperatures more energetic photons are emitted, until at around 1700 K objects glow white hot from the mixture of photons with energies corresponding to all visible colors. Beyond that ultraviolet radiation is also emitted, as from the sun, and it is this radiation that can produce sunburn. Later in the book we discuss the properties of radiation and their interaction with matter. Below 1000 K, and even at ambient temperatures, objects emit infrared radiation that we cannot see. Night vision detectors and infrared thermography can be used to image thermal sources such as heated buildings or machines and people (Figure 12.22).

Experimentally it is found that the time rate of emission of radiation is very strongly dependent on the surface temperature, varying as T^4. The *Stefan–Boltzmann law* gives this rate, or the radiated power, as

$$P = e\sigma A T^4, \tag{12.28}$$

where A is the surface area of the object, σ is the Stefan–Boltzmann constant, a universal constant equal to 5.67×10^{-8} W/m^2-K^4, and e is the *emissivity* of the object. The emissivity, varying between 0 and 1, is the property of an object characterizing its quality as an emitter of radiation. Light-colored materials with shiny surfaces have e values close to zero, whereas objects with a black dull finish have e values near 1.

All objects not only radiate but absorb radiation as well. If an object at temperature T_1 is in a "temperature bath" large enough to be at a fixed temperature T_2, then the net rate of radiant emission will be given by

$$P = e\sigma A(T_1^4 - T_2^4), \tag{12.29}$$

where the second term is the power absorbed from the bath. One might question why the constant e should be the same for emission and absorption. The answer lies in considering what happens when $T_1 = T_2$. In this case no net power can be radiated, so the coefficient e in the absorption term in Equation (12.29) must be the same as in the emission term in order that $P = 0$, and this must then be generally true.

An object can maintain a temperature different from its surroundings if it has either a source of internal energy or a sink for removal of internal energy and if the object balances

FIGURE 12.22 *Examples of infrared imaging color-coded thermograms:* (left) *the imprint of a hand 5 min after touching a wall;* (center) *a shoe;* (right) *color-coded house showing heat leaks.*

the rate of uptake and loss of energy. This situation is one of steady state rather than thermal equilibrium and warm-blooded animals are a primary example of this phenomenon. We produce energy from metabolism at the same net rate that we lose energy to our environment in order to maintain an approximately constant temperature.

Another example of steady-state thermal behavior is the atmosphere of the Earth. The thermal balance between the net absorption of energy from the sun and the net emission of radiation to space determines the Earth's mean temperature. The gases in the Earth's atmosphere transmit the sun's radiation, but reflect some portion of the infrared radiation from the warmed surface of the Earth, thus trapping some of the heat that would otherwise escape from the Earth. This is known as the *greenhouse effect.* The name comes from how a garden greenhouse functions to transmit the sun's light, but prevent the loss of heat. Necessary for most life on Earth, the greenhouse effect causes the average temperature of the Earth to be about 32°C warmer than it would be otherwise.

Global warming is a consequence of an imbalance in this steady state due to excessive absorption, caused by increasing amounts of "pollutants" in the atmosphere. These molecules, including carbon dioxide, nitrous oxide, ozone, methane, and other molecules together known as greenhouse gases, strongly absorb in the infrared and have been increasing in concentration. Dramatically increasing amounts of manmade greenhouse gases, most notably carbon dioxide from the burning of fossil fuels such as oil, coal, and natural gas, have led to an enhanced greenhouse effect in which there has been a relatively rapid rise in average temperature of the Earth. CO_2 levels were about 280 ppm (parts per million) at the start of the industrial revolution in the late 18th century and have increased to about 380 ppm today; we know that CO_2 levels have not been this high in the past 420,000 years and probably have not been this high in 20 million years.

Part of the grave nature of this enhanced greenhouse effect is that many aspects of global climate are coupled together. An increasing number of scientists believe that the greenhouse effect can lead to a positive (but not in the good sense) feedback process, whereby increasing global temperatures may lead to the release of trapped greenhouse gases (especially in marine sediments in the oceans and in the polar icecaps), spiraling the world's temperature to even higher values. Furthermore, increasing temperatures are leading to the polar ice caps shrinking, which may produce a major increase in the height of the oceans. Perhaps even more threatening, although uncertain, is the impact of higher temperatures on potable water, on agriculture, and on the development of strains of bacteria and virus that will cause new diseases. It is imperative that the world community address these issues now and begin steps designed to cut greenhouse gas emissions.

CHAPTER SUMMARY

Temperature is a measure of the thermal energy of an object. Three common temperature scales are the absolute (Kelvin), the Celsius (centigrade), and the Fahrenheit scales, with only the Kelvin scale having a nonarbitrary zero level. These are related to each other by

$$T_C = T_K - 273.15°C, \qquad (12.2)$$

and

$$T_F = \frac{9}{5} T_C + 32°F. \qquad (12.1)$$

When an object of length L is heated so that its temperature changes by ΔT, it will expand by ΔL according to

$$\Delta L = \alpha L \Delta T, \qquad (12.3)$$

where α is the coefficient of linear expansion.

For an ideal gas (one with no long-range interactions), the pressure P and volume V are related through

$$PV = \frac{2(KE_{\text{total}})}{3}, \qquad (12.10)$$

(*Continued*)

so that, using the ideal gas law

$$PV = Nk_BT, \qquad (12.11)$$

we have that the average kinetic energy of a molecule is given by

$$\overline{KE} = \frac{3}{2}k_BT. \qquad (12.12)$$

The first law of thermodynamics is a statement of conservation of energy, in which the change in internal energy of a system ΔU is equal to the heat flow in (Q) minus the work done by (W) on the system:

$$\Delta U = Q - W. \qquad (12.16)$$

Heat flow into or out of an object of mass m and specific heat c will lead to its temperature changing according to

$$Q = cm\Delta T, \qquad (12.21)$$

as long as there is no phase change. For the object to change phase at a fixed temperature, a specific amount of heat per unit mass, the latent heat, is required:

$$Q_{\text{transformation}} = Lm. \qquad (12.22)$$

Enthalpy H, defined as

$$H = U + PV, \qquad (12.23)$$

can be used to describe bond energies and chemical reactions.

Colligative properties of solutions are discussed, including vapor and osmotic pressures, and the basic filter functioning of the kidneys (and artificial dialysis) is described in terms of osmotic pressure.

Heat can be transported by convection, conduction, or radiation. Convection is the flow of heat via bulk motions of the surrounding fluid. The rate of thermal conduction is proportional to the thermal gradient ($\Delta T/L$) according to

$$\frac{Q}{t} = kA\left(\frac{T_H - T_C}{L}\right), \qquad (12.27)$$

where A is the cross-sectional area perpendicular to the heat flow and k is the thermal conductivity. Heat is radiated by all objects at some temperature T_1, surrounded by a medium at temperature T_2, at a rate given by

$$P = e\sigma A(T_1^4 - T_2^4), \qquad (12.29)$$

where A in the surface area of the object, e is the emissivity (a pure number between 0 and 1), and σ is the Stefan–Boltzmann constant, $\sigma = 5.67 \times 10^{-8}$ W/m^2-K^4.

QUESTIONS

1. Two bars made of different metals are placed in contact with each other and come to thermal equilibrium. Are the temperatures of the two bars the same? Are their internal energies the same? Are the rms velocities of their atoms the same?
2. What do you think will happen if you add an equal volume of liquid helium to liquid nitrogen?
3. Give some examples of engineering safety measures that allow for the differential thermal expansion of neighboring materials.
4. A tightly sealed jar with a metal cover is often put under hot water to make it easier to open. Why?
5. Distinguish carefully between average velocity and rms velocity in three dimensions. Which is important in determining temperature? For a solid what do you think the average velocity of the constituent molecules would be?
6. Discuss why the first law of thermodynamics is referred to as a conservation of energy equation for an isolated system.
7. In an isothermal process with an ideal gas, what happens to the gas pressure as its volume increases?
8. In an isobaric process in which an ideal gas expands its volume, what happens to the temperature? Be careful: the ideal gas law might seem to imply that as the volume increases at constant pressure the temperature should go up, whereas the fact that the gas has done work implies that its temperature should go down. Which is correct and why?
9. How much heat is required to melt 1 kg of ice at 0°C to water at 0°C compared to the heat needed to turn that same mass of water at 100°C to steam at 100°C? Discuss why so much more heat should be required to produce steam.
10. When water boils, salt can be added to stop the boiling. Discuss why this works.
11. Why do hot dogs swell when boiled in water, often to the point of splitting?
12. Why does a piece of metal feel cooler to your hand than a piece of wood at the same temperature?

13. Discuss the various mechanisms by which you maintain your body temperature when doing heavy exercise. Frame your answer in terms of the physics of heat transfer.
14. Why do you think it is true that the more strenuous exercise a person does, the cooler is the average skin temperature?
15. Why is the glass liner of a thermos bottle coated with silvered paint?

MULTIPLE CHOICE QUESTIONS

1. Liquid nitrogen is used by dermatologists to remove precancerous growths on the skin by flash-freezing the unwanted cells. The temperature of liquid nitrogen is approximately (a) 77 K, (b) 273 K, (c) 373 K, (d) −273 K.
2. The difference in Fahrenheit temperature between the steam point and ice point of water is (a) 100, (b) 180, (c) 212, (d) 273 degrees.
3. The internal energy of a beaker of gas in thermal equilibrium at room temperature is more than 10,000 J, whereas the internal momentum of the gas is zero. That is most closely related to the fact that (a) the gas molecules aren't moving at room temperature, (b) kinetic energy is a positive number, independent of the direction of motion, whereas momentum is a vector, (c) the kinetic energy of an atom is totally independent of its momentum, (d) electrons cannot be excited by room temperature collisions.
4. Two identical containers of gas (same volumes, same number of atoms) are at different temperatures. Which of the following is higher in the gas that has the higher temperature: its (a) volume, (b) density, (c) internal energy, (d) average atomic spacing?
5. You want to raise the temperature of an ideal gas to a maximum value with a fixed Q joules of heat. Which of the following is the best process for doing so? (a) Hold the volume constant. (b) Hold the pressure constant. (c) Hold the internal energy constant. (d) It doesn't matter because all processes will yield the same final temperature.
6. Two closed containers both contain 1 mol of the same ideal gas. The gas in container A has a volume of 1 L and a pressure of 1 atm. The gas in container B has a volume of 1/2 L and a pressure of 2 atm. When the containers are placed in good thermal contact with each other (with no exchange of gas) which of the following changes occur? (a) The pressure in A increases. (b) The pressure in B increases. (c) There are no changes in either container. (d) There isn't enough information to determine what happens.
7. Body A and body B are in thermal contact and are in thermal equilibrium. Which of the following is true? In thermal equilibrium, (a) the total amount of energy due to atomic motion is the same in A as it is in B, (b) each of the atoms in A and in B have exactly the same amount of energy, at any instant, (c) the atoms in both A and B stop moving, (d) the average amount of energy transferred by atomic collisions from A to B is the same as the average amount transferred from B to A from instant to instant.
8. The average translational speed of each molecule in an ideal gas is doubled. The Kelvin temperature of the gas (a) decreases by a factor of four, (b) decreases by a factor of two, (c) increases by a factor of two, (d) increases by a factor of four.
9. An aluminum block slides across a horizontal wood surface. Initially, the block's center-of-mass is traveling with a kinetic energy of 50 J. Later the block is at rest. That is because (a) the internal energy of the block and the surface has increased by about 50 J, (b) the air is flowing with an increase of 50 J of kinetic energy, (c) 50 J of kinetic energy is converted into 50 J of potential energy, (d) 50 J of energy is destroyed in this process.
10. Volume can be used to measure the temperature of a solid because (a) atoms swell as their temperature increases, (b) atoms collide more often at higher temperatures, (c) the potential energy of atomic interaction gets weaker at higher temperatures, (d) the average separation between atoms increases as atomic kinetic energy increases.
11. Zero degrees Kelvin is defined as the temperature at which (a) ice coexists with sea water at 1 atm, (b) ice coexists with pure water at 1 atm, (c) steam coexists with pure water at 1 atm, (d) one mole of argon gas would exert zero pressure.
12. A system of fixed mass does 300 J of work on its surroundings and takes in 500 J of heat from its surroundings. As a result, the internal energy change of the system is (a) +800 J, (b) +200 J, (c) −200 J, (d) −800 J.
13. A closed system interacts with its surroundings. For which of the following is $\Delta U > 0$? (a) $W = -500$ J, $Q = 0$, (b) $W = +500$ J, $Q = +300$ J, (c) $W = +100$ J, $Q = +100$ J, (d) $W = -100$ J, $Q = -100$ J.
14. Two different reversible processes connect the same two equilibrium states. Which of the following must be the same for the two processes? (a) ΔU and ΔT, (b) Q and W, (c) Q and ΔT, (d) ΔU and W.

Questions 15 and 16 refer to: A column of liquid mercury inside an evacuated glass tube is used as a thermometer. As the equilibrium temperature of the thermometer increases, the length of the mercury column increases.

15. In order for this system to measure temperature as stated (a) the density of mercury must be greater than that of glass, (b) the density of mercury must equal that of glass, (c) the coefficient of volume expansion of mercury must be greater than that of glass, (d) the coefficient of volume expansion of mercury must equal that of glass.

16. When the bulb (bottom) of a mercury–glass thermometer is placed in contact with a hot body the length of the mercury column initially falls. That is because (a) the hot body initially draws energy out of thermometer, (b) the glass temperature is initially higher than the mercury temperature when placed in contact with the hot body, (c) at the temperature being measured glass expands more per unit volume than mercury, (d) the thermometer is upright; this wouldn't happen if the thermometer were lying on its side.

17. A patient with a high fever is given an isopropyl alcohol rubdown. That is because alcohol (a) evaporates rapidly at temperatures around 40°C, (b) has a large heat capacity, (c) has a large coefficient of thermal conductivity, (d) has a large coefficient of volume expansion.

18. 310 K is closest to the temperature of (a) the interior of a living human, (b) a comfortable room, (c) ice water, (d) liquid nitrogen.

19. A copper block of mass 2 kg is placed in good thermal contact with a copper block of mass 1 kg. When the blocks are in thermal equilibrium (a) they have the same amount of internal energy, (b) the more massive one has half the internal energy of the less massive one, (c) the more massive one has twice the internal energy of the less massive one, (d) they contain the same amount of heat.

Questions 20 and 21 refer to: The specific heat of aluminum is twice as big as the specific heat of iron. One kg of aluminum is placed in good thermal contact with 1 kg of iron. After the two bodies have come into thermal equilibrium 1 J of heat has flowed out of the aluminum.

20. Assuming no other body is involved, which of the following is true? (a) 0.5 J of heat has flowed into the iron. (b) 1 J of heat has flowed out of the iron. (c) 1 J of heat has flowed into the iron. (d) 2 J of heat has flowed into the iron.

21. The temperature change of the aluminum is (a) half as large as the temperature change of the iron, (b) the same as the temperature change of the iron, (c) twice as large as the temperature change of the iron, (d) unrelated to the temperature change of the iron.

Questions 22 and 23 refer to an experiment in which two equal mass cylinders of different materials (*A* and *B*) at different temperatures are put into thermal contact within an insulated container.

22. In this experiment $c_A \Delta T_A = -c_B \Delta T_B$. This is because (a) the temperature loss of *A* equals the temperature gain of *B*, (b) the internal energy loss of *A* equals the internal energy gain of *B*, (c) the process involved is adiabatic, (d) the process involved is isothermal.

23. One room temperature cylinder and one cold cylinder are placed in thermal contact, but not inside the Styrofoam container. The initial temperatures of the cylinders are measured to be 20 and 0°C, respectively. The temperature at which the two cylinders come into equilibrium will be (a) about 10°C, (b) a few degrees above 10°C, (c) a few degrees below 10°C, (d) none, because the two cylinders will never come into equilibrium if they are not in the Styrofoam container.

24. The source of the tremendous energy associated with a hurricane is water vapor condensing into liquid droplets. Which of the following is most closely related to this effect? (a) Volume expansion, (b) specific heat, (c) vapor pressure, (d) latent heat of vaporization.

25. Work done by a fluid requires (a) the fluid's volume to change, (b) the fluid's pressure to change, (c) the fluid's internal energy to change, (d) a heat flow.

26. A desk has a wooden top and metal drawer. When you place your hand on the wood top it feels warmer than when you place your hand on a drawer. That is because (a) wood has more internal energy per atom than metal at room temperature, (b) wood has a lower heat capacity than metal at room temperature, (c) wood has a lower coefficient of thermal conductivity than metal at room temperature, (d) your hand makes better thermal contact with wood than it does with metal.

27. A snowbank covered with fine dark soot melts much faster on a sunny day than a snowbank with a bright white surface. That is because (a) the emissivity of a dark snowbank is greater than that of a white snowbank, (b) the melting point of snow lying below a layer of soot is less than for pure snow, (c) the specific heat of snow lying under a layer of soot is less than for pure snow, (d) the thermal conductivity of snow lying under a layer of soot is greater than for pure snow.

28. A liter container of O_2 gas obeys the relation $PV = Nk_BT$. The number of atoms in the gas equals (a) $2N$, (b) $3N/2$, (c) N, (d) $N/2$.

29. Suppose a liter of N_2 gas has its absolute (Kelvin) temperature doubled. The average translational speed of an N_2 molecule will (a) increase by a factor of 2, (b) increase by a factor of 3/2, (c) increase by a factor of $(2)^{1/2}$, (d) remain the same.

PROBLEMS

1. Find the temperature at which the numerical values of the Celsius and Fahrenheit temperatures agree.

2. When it is 0°F in the northeastern United States, someone in southern Europe mistakenly thinks that it is 0°C. By how many °C and °F are they wrong?

3. If a cube of aluminum is heated from 20 to 100°C and one edge expands by 0.19 mm, by how much has its volume increased? By what percent has its volume increased?

4. Water at 10°C is poured into molds to make cubic blocks of ice at −20°C. If the molds are 20 cm on a side what fraction of the mold should be filled for the ice to just make a cube?

5. Some lasers use invar rods to define the optical length of the laser because invar has a very low thermal expansion coefficient (0.9×10^{-6}/°C). Calculate the

length change of a 1.2 m invar rod when heated from room temperature (20°C) to 45°C.

6. How many ideal gas molecules are there in a 1 cm^3 volume at STP (standard temperature and pressure of 0°C and atmospheric pressure)? If the gas is argon, what is its mass density?

7. What is the rms velocity of the oxygen molecules in the air at room temperature (20°C)?

8. What is the ratio of the rms velocity of oxygen to that of nitrogen molecules in air?

9. Two pressurized gas tanks at room temperature have a sealed gas valve between them. One compartment with a 20 cm^3 volume has nitrogen gas at a 100 atm pressure and the other one with a 50 cm^3 volume has nitrogen at a 25 atm pressure. If the valve is opened, what will be the final common pressure?

10. Find the work done when 1 mol of an ideal gas at room temperature is isothermally heated to twice its volume.

11. How many moles of water are there in 1 L? How many water molecules is this?

12. Given the following ten numbers, calculate their rms value and compare it to their average value: 3, 5, 2, 8, 4, 3, 6, 5, 6, 4.

13. Suppose an 80 kg person takes in 3000 calories (1 calorie = 1 kcal = 4180 J) of food energy per day. Out of this, the body requires about 1800 just to maintain itself at rest (the basal metabolic rate). If our muscles are 20% efficient and the balance of the food energy is used to do work, calculate how many flights of 3 m tall stairs the person can go up in a day without using any stored energy.

14. The caloric value of fats is more than twice that of protein or carbohydrates. Each gram of fat is equivalent to about 9.3 Cal, whereas a gram of protein or carbohydrate is equivalent to about 4.1 Cal. If a 200 lb (90.8 kg) man wanted to "burn" off 5 lb (2.3 kg) of fat, how long would it take him bicycling where heat is generated at 6 Cal/min? Running, where heat is generated at 15 Cal/min? This points out why it is easier to lose weight by reducing calorie intake.

15. Ice is added to 500 cm^3 of water at 30°C. How many grams should be added so that the final temperature will be 5°C when it has all melted?

16. Although different types of foods, with different compositions of protein, carbohydrates, and fats, give different caloric values, the different pathways of oxidation of these nutrients all end up generating about 5 kcal of energy per liter of oxygen. Based on this, how many liters of air (20% oxygen) must be breathed in a day in order to burn 3000 Cal if all the O_2 were burned. Then calculate the actual fraction of the oxygen consumed by that person using an average breathing rate of 12/min with a lung volume intake of 0.5 L.

17. Physical fitness of an individual can be measured by the rate of maximum oxygen uptake during exercise. This reflects the person's ability to sustain aerobic energy pathways during exercise, producing heat at a rate of 5 kcal per liter of oxygen. Anaerobic pathways are only about 50% as efficient and lead to the buildup of lactic acid in the muscles, causing muscle cramping. A typical maximum oxygen uptake rate for a normal young male is about 2.5 L/min, although a well-trained athlete might double this rate. Assuming the muscles are 20% efficient, find the maximum mechanical power output in W possible for an athlete (remember to subtract the basal metabolic rate of about 80 kcal/h).

18. Estimate the rate of heat loss from the skin due to thermal conduction and thermal radiation when doing moderate exercise. Under these conditions the skin surface temperature is 31°C. Take the ambient temperature to be 23°C, the emissivity to be 0.97, the body surface area to be 2 m^2, and assume a 5 cm layer of air through which heat is conducted in the absence of any convection, so that at 5 cm from the skin the temperature is reduced to the ambient temperature.

19. In addition to conduction and radiation cooling of the body when exercising, evaporation of perspiration also is very effective in giving off heat. During moderate exercise the two to three million sweat glands produce about 8 g of sweat per second. From the heat of vaporization of water at 37°C of 580 kcal/kg, find the rate of energy loss from the body due to sweating, assuming that 5% of the sweat evaporates.

20. The SR-71 Blackbird is the world's fastest airplane, flying at altitudes over 80,000 feet and at over three times the speed of sound (Mach 3). The aircraft is 107 feet 5 inches long and when it lands after a long flight it is too hot to be touched for about 30 min and is 6 inches (15.24 cm) longer than at takeoff.

(a) How hot is the Blackbird when it lands, assuming that the coefficient of linear expansion is $24 \times 10^{-6}\ K^{-1}$ and its temperature at takeoff is 23°C.

(b) Suppose that the plane looks like an isosceles triangle. If the wingspan (base of the triangle) is 55 feet 5 inches, what is the new cross-sectional area of the plane when it lands?

(c) Suppose that the plane loses heat by radiation. What is the net rate of heat loss through the upper surface of the plane if the emissivity of the plane is 0.80?

13

Thermodynamics: Beyond the First Law

Our discussion of thermodynamics in the last chapter was limited to energy considerations. Although energy conservation is a necessary requirement for any process to occur, it is not a sufficient condition. There are many energy-conserving processes that occur spontaneously, but that are not reversible even though that reversed process would also conserve energy. In this chapter we continue our introduction to thermodynamics with a discussion of entropy and the second law of thermodynamics. We relate entropy to the degree of disorder in an isolated system through a microscopic picture and we show that this disorder always increases with time. Life is a constant struggle to maintain a high degree of order. The corresponding reduction in entropy is accomplished at the expense of even more disorder in our environment in order to satisfy the second law of thermodynamics. We next discuss Gibbs free energy, related to chemical potential, the most important energy concept in biology. This thermodynamic state variable is a measure of the energy available for useful work at constant temperature and pressure, the usual conditions of life. The chapter concludes with several biological applications of these concepts, including ATP hydrolysis, photosynthesis, and conformational changes in biomolecules.

1. ENTROPY AND THE SECOND LAW OF THERMODYNAMICS

Many processes in nature that conserve energy and do not violate any of the other fundamental principles we have introduced so far in our study of physics simply do not occur. Now when a basic physical process never happens even though it seems to satisfy all of the fundamentals in our theories of knowledge, there is something amiss. From many historical examples, it is usually the case that there is some new principle that would be violated by the occurrence of such a process. We begin this section with a brief discussion of some examples of processes in different areas of physics that never occur, leading to a qualitative presentation of the common principle that prohibits them.

In mechanics, all sliding objects eventually come to rest because their kinetic energy has been lost due to what we call friction, the process by which mechanical energy is transferred to heat. Energy has not been lost, but the "useful" form of energy, which in mechanics is the sum of kinetic and potential energy, called mechanical energy, has been lost through its transfer to internal energy. Once a sliding object comes to rest, it is never the case that the internal energy of the object and surroundings spontaneously transfers back to the object in the form of mechanical energy making it move again. We conclude that although energy would be conserved in the reverse process, once "organized" energy, such as kinetic energy in which all molecules of the moving object translate together, is converted to random thermal motions of molecules, the process is irreversible. It is too improbable that all the molecules will spontaneously coordinate their motions in order to propel the object again.

J. Newman, *Physics of the Life Sciences*, DOI: 10.1007/978-0-387-77259-2_13,

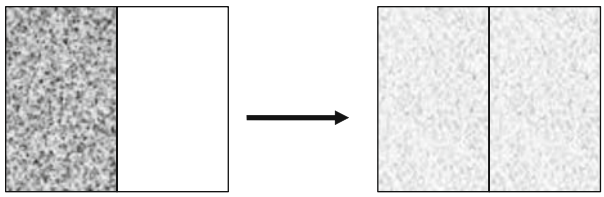


FIGURE 13.1 *When a small hole is made in the partition between the two chambers shown on the left, the gas distributes itself uniformly as shown on the right. The reverse process never occurs.*

In fluid mechanics there are many similar examples. Suppose that a gas is confined to half of a closed container by means of a partition. If a hole is punctured in the partition the gas will leak into the other half of the container, eventually reaching a uniform distribution throughout the container (Figure 13.1). Energy conservation would not be violated if the molecules on one side of the partition spontaneously re-entered the other side and all the gas returned to only one side of the container. We know, however, that this process is fundamentally irreversible because it is again too improbable that such a sequence of events would occur.

An example from thermodynamics further illustrating this point is the inevitable cooling of hot coffee. A thermos bottle can be used to reduce the rate of heat loss from the coffee compared to when it is just in a cup. "Vacuum" bottles reduce the conduction of heat, and the silver coating on the glass inner bottle reduces radiation losses. Despite this, the coffee eventually will lose heat to its surroundings and the cooling process is irreversible. Irreversibility here means that the original situation cannot be restored without additional energy input. Of course the coffee can be heated again, but the heat lost to the surrounding air cannot be collected and used alone to reheat the coffee to its initial temperature without additional energy input, even though such a process would conserve energy.

What is common to all of these examples is the notion of the probability of the occurrence of an event and of its time-reversal event. The bookkeeping of energy conservation is satisfied for both events, however, the likelihood of the reversal is essentially zero. Here we see how a methodology, known as *statistical mechanics*, can be developed for calculating the likelihood of events. We start with some simple notions from coin-tossing problems.

If we flip a legitimate coin in the air, there is an equal probability of getting a head or tail when it lands. Flipping three coins in the air at the same time results in a variety of possible "*outcomes*" including 0, 1, 2, or 3 heads, but these do not occur with equal probability. There is only one combination that gives either 0 or 3 heads, whereas there are 3 possible combinations that will result in either 1 or 2 heads, giving a total of 8 possible distinct "*states*" for the coin flip (Figure 13.2). As more and more coins are flipped together the total number of different possible states grows rapidly (with N coins, the number is 2^N; with $N = 100$, the number is about 10^{30} or more than the number of protons in your body!), and the number of possible outcomes is much smaller (in the case of N coins, there are simply $N + 1$ possible outcomes; what are they?). No matter how many coins are flipped, the number of states resulting in the most "ordered" outcomes of all heads or all tails remains just 1 so that those events become essentially impossible as the number of coins increases to a number of 100. Flipping 100 coins and finding 100 heads would be the equivalent to a cold cup of coffee spontaneously heating up by absorbing heat from the room temperature air.

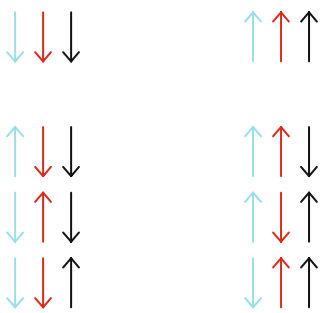


FIGURE 13.2 *The three coin flip experiment with up and down arrows indicating heads or tails.*

Example 13.1 Find the number of states and outcomes for the case when 4 fair coins are tossed and then find the probabilities of each of the outcomes.

Solution: With 4 coins there are 5 possible outcomes (ranging from 4 heads to 0 heads) and $2^4 = 16$ possible states. There is only one way to have 4 heads and only one way to have 0 heads, so that the probabilities for each of these is 1/16 = 6.25%. There are 4 different ways to have 1 head—each of the 4 coins could be a head—and similarly there are 4 different ways to have 3 heads—each of the 4 coins could be a tail—so that each of these has a probability of 4/16 = 0.25 = 25%. For 2 heads, the two coins that are heads could be any of the four coins and we can find 6 ways for this to occur, so the probability for 2 heads is 6/16 = 37.5%. Of course, this last value could have been obtained from noting that the probabilities must add up to 1 (or 100%). Check this.

Of the 101 different possible outcomes of the 100 coin flip experiment, which are most likely? As you should already have guessed the most likely case is 50 heads and 50 tails. Probability theory can tell us that if this experiment is repeated over and over again that about 90% of the time we will find between 45 and 55 heads out of the 100 coins. The distribution of possible outcomes is fairly sharply and symmetrically peaked around 50.

In real physical systems, what is analogous to the notion of "states" and "outcomes" in the coin-toss experiment? To answer this we need to jump ahead a bit. We show later in this book that the world is governed by quantum mechanics and that, for atoms and molecules, possible energy values are quantized, or discrete, so that there are a countable number of different values for the energy of an atom or molecule. Atoms or molecules cannot have any random value of energy, but must exist in a quantum state with one of a discrete list of possible energies that can be labeled by a number, a so-called quantum number. These energy states can be pictured as energy levels, which may be familiar to you from a previous physics or chemistry course, with the atom or molecule "residing" in a particular level. In Figure 13.3 the energy levels have a quantum energy separation of ε with this particular atom in the third "excited state" with an "excitation level" of 3 and an allowed state of motion corresponding to a total energy of 3ε. We can think of the atom in this state as having 3 quanta of energy, each worth ε joules.

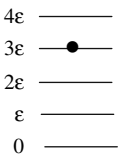


FIGURE 13.3 *Typical energy level diagram for an atom, with the lowest (ground) state and several excited states shown (there are many more levels above the fourth excited state not shown; also, in many cases the energy levels are not equally spaced). A typical energy spacing is 10^{-21} J for atoms in a solid and 10^{-23} J for atoms in a gas.*

Now, suppose we have a large number N_A of such atoms. Each of the atoms has its own excitation level and its own corresponding energy. To find the internal energy of the large system due to atomic motions we just add up all of the individual atomic motional energies. Thus, we can write the internal energy of the system as $N_E\varepsilon$, where N_E, the total excitation level of the whole system, is just the sum of all of the atomic excitation levels. (For example, if there were three atoms in the system with excitation levels 4, 5, and 6, the excitation level of the system would be 15.) N_E is the total number of energy quanta the system contains. Typically, for a macroscopic system both N_E and N_A will be huge, perhaps 10^{25} or so.

A *microstate* of this system is one of the very large number of states described by a particular set of excitation levels, one for each atom in the system. It is one of the premises of statistical mechanics that at equilibrium all allowed microstates of a system (those satisfying conservation of energy) are equally probable. Microstates are analogous to the 2^N different possible "states" of the coin-flip experiment. Unlike the heads or tails options for a coin, we are dealing with rolling a huge number of special dice, one for each atom, each of the dice with an enormous number of faces representing the different excitation levels of an individual atom rather than the usual six faces.

However, just as with the coin-flip experiment, when all is said and done, what is most important are the "outcomes": how many heads we will get with what probability for N coin flips. The details of which particular coin landed as a head or tail are not important. In our atomic system, the analog to an outcome is a *macrostate.* This is specified by the total numbers of atoms with each of the possible excitation levels, known as the *occupation numbers.* Occupation numbers together with the associated excitation levels represent the information needed to determine the total energy of the system. There will be many microstates corresponding to each particular macrostate, just as there are many different possible coin-flip sequences that result in the same outcome (except for all heads or all tails). Because, as we have noted, each microstate is equally likely to occur, the probability of a particular macrostate will depend solely on the number of microstates corresponding to a given macrostate. Thus, as we saw in the coin-flip experiment, the possible outcomes (or macrostates) may be limited by probability to those that are most likely to occur based on those with the largest number of states (microstates) leading to that outcome.

Example 13.2 Suppose that there are four identical atoms each with equally spaced energy levels given in Figure 13.3 and with a total energy of 6ε. Find all the possible macrostates of the system by defining their occupation numbers.

Solution: Because the total energy is 6ε, we need to include energy levels up to that value, because one possible macrostate has 3 atoms in the zero energy ground state and 1 atom with excitation level 6. If we write out the occupation numbers of this state as (3,0,0,0,0,0,1) where from left to right we show the number of atoms at increasing excitation levels from the ground state to 6, we can use this notation to find the other possible macrostates. These can be written as: (2,1,0,0,0,1,0), (2,0,1,0,1,0,0), (2,0,0,2,0,0,0), (1,2,0,1,0,0), (1,1,1,1,0,0,0), (0,3,0,1,0,0,0), and (0,2,2,0,0,0,0). We note that not all of these macrostates are equally likely. For example, there are 4 microstates that could correspond to the macrostate given by (3,0,0,0,0,0,1), corresponding to a different one of the 4 atoms having excitation level 6. For the macrostate given by (1,1,1,1,0,0,0) there are 4 choices for filling the first state, 3 for the second, 2 for the third, and the remaining atom fills the fourth so that there are 4! "4-factorial" = (4)(3)(2)(1) = 24 different possible microstates in this case. Thus this macrostate is six (24/4) times as likely as the one in which only one atom has all the energy.

The information on the numbers of microstates in a given macrostate (the occupation numbers) is contained in a function Ω, known as the statistical weight of the system, that is directly related to the *entropy S* of the system

$$S = k_B \ln \Omega, \tag{13.1}$$

where k_B is the Boltzmann constant. Entropy is thus a statistical function depending ultimately on the occupation and quantum numbers but indirectly on the state variables, such as pressure, temperature, and volume, and is a measure of the likelihood of that particular macrostate, given total values for energy and other conserved quantities. One immediate question is how much choice there is in the macrostate that the system occupies. In our coin-flip experiment with only 100 coins we saw that the probabilities are fairly sharply peaked with a probability of about 90% that the outcome is between 45 and 55 heads. With the typically much larger numbers of microstates in thermodynamic systems, the range of parameters of the final macrostate is extremely sharply peaked.

Having introduced some concepts that can be used to describe a thermodynamic system (with large numbers of atoms), we're now in a position to state a new law of physics.

The second law of thermodynamics states that the total entropy of a closed system always increases,

$$\Delta S \geq 0, \tag{13.2}$$

with $\Delta S = 0$ only in the special case of a reversible process.

A reversible process is an idealization of a process that is performed slowly enough so that the system remains in equilibrium throughout, a so-called *quasistatic* process. In general, the total entropy of a closed system must increase; this is fundamentally a statistical statement about probabilities of occupation numbers. As we saw in the last chapter the internal energy of a system can change over time by either work being done or by a flow of heat. Given a variety of different events that can occur (satisfying energy conservation and other conserved quantities), the one having the most

possible microstates will be the one that occurs. The number of different microstates of a particular macrostate is intrinsically related to its increased "randomness."

Mechanical energies are more "organized" and much less "random" in nature than thermal energies. The second law implies that, although both forms of energy may be equal in magnitude, statistics drives reactions or events toward producing thermal energy from mechanical energy in order to maximize entropy. Frictional forces are nonconservative precisely because the thermal energy they produce cannot be reversibly transformed back to mechanical energy. A general conclusion is that whenever the entropy of a closed system increases, the amount of energy available to do work is decreased. Increasing entropy degrades the usefulness of energy. To see this from a microscopic picture, let's now return to our atomic model system of N_A atoms with energy levels shown in Figure 13.3 and ask how we can change the energy of the system.

We can change the internal energy of a system of atoms in three ways. We might add or subtract atoms (change N_A). We might increase or decrease the atomic energy level spacing (ε). And we might increase or decrease the total number of energy quanta of the system (N_E), leaving the number of atoms fixed. As in previous discussions, we only consider closed systems, ones with fixed numbers of atoms, so that only the latter two options are available.

So, how can ε be changed? The exact value of ε depends on the details of how the atoms interact with each other and their container, but if the average region in which an atom is confined has length L, the value of ε is roughly proportional to $1/L^2$. (We study this in some detail in Chapter 25, but the proportionality arises from quantum mechanics.) That is, by changing the volume that the system is confined in we change ε. In fact, if we change the volume very slowly, each atom will stay in its allowed state of motion and the total excitation number will not change. (This is a formal result derived from quantum mechanics.) Very slow change in volume only changes ε and not N_E. That sounds a lot like what we have previously called work.

Similarly, we can change N_E without changing ε. We place our system in close contact with a second system so that the atoms at the interface can swap energy. If one system has more energy per atom than the other (a larger value of $\varepsilon N_E/\mathrm{N_A}$) and if atomic interactions can effectively be taken to be random processes, then random scrambling will cause a preferential flow of quanta of energy from the system with the higher energy per atom to the system with lower. This is demonstrated in Figure 13.4. Here white means "hot" (high number of energy quanta) and dark means "cold" (low number of quanta). Random swapping of energy quanta between atoms preferentially moves energy from the hot side to the cold because there are more quanta to select from on the hot side. (Quanta from the cold side move to the hot side also, but there are just fewer of them at first from which the random swapping process can choose.) As time goes on the quanta become more-or-less evenly distributed

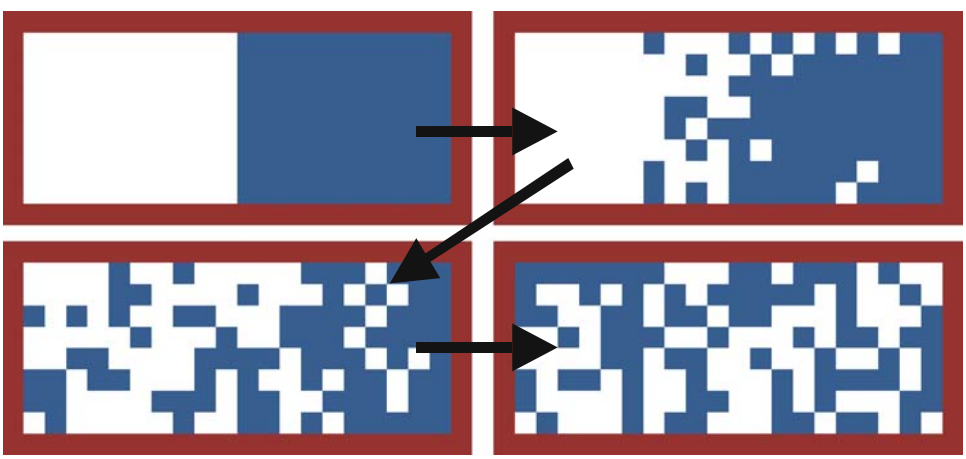


FIGURE 13.4 *Sequence of snapshots of the flow of energy quanta from the initially hot* (left) *side to the colder* (right) *side of a system.*

throughout the container. All of this sounds a lot like what we have called heat flow. So here's the atomic level interpretation of work and heat flow. *Changing the energy level spacing ε of the atoms in a system corresponds to work; changing the number of energy quanta a system has corresponds to heat flow.*

Entropy has something to do with assessing how much internal energy is available to do work. To be useful, internal energy has to be concentrated. The more dilute or disorganized the internal energy, the less useful it is and the larger the entropy. Microscopically, entropy is a measure of the number of different ways you can distribute N_E quanta over N_A atoms. This is precisely the statistical weight of Equation (13.1) and the occupation numbers represent the bookkeeping needed to keep track of this. The more ways you can divvy up the fixed total energy in packages of quanta of energy over the atoms of the system, the less concentrated the energy will be and the less useful it will be. The more ways you can divvy up quanta over the atoms of the system, the more "mixed up" the energy is, the more disordered it is, and the greater is the system's entropy. Microscopically, *entropy is a measure of disorder*.

The formal expression for counting all of the different arrangements of energy is

$$A = \frac{(N_E + N_A - 1)!}{N_E\,!(N_A - 1)!},$$

where "!" means "factorial:" $N! = N(N-1)(N-2)(N-3)\ldots 1$, as in Example 13.2. The number of arrangements of energy quanta over atoms increases extremely rapidly as either N_E or N_A increase. For example, suppose $N_A = 10$ and $N_E = 10$. Then $A = 92{,}378$. If the number of atoms just doubles to $N_A = 20$, still with $N_E = 10$, then $A = 3{,}628{,}800$, an increase of a factor of 40. Similarly, if $N_A = 10$ but the number of quanta doubles to $N_E = 20$, then $A = 10{,}015{,}005$, an increase of a factor of 110.

One immediate consequence of all of this is that unusually concentrated arrangements of energy in a large system are extremely unlikely. Suppose we have 20 atoms and 40 energy quanta. The number of ways to arrange 40 quanta over the 20 atoms is $A = 1.22 \times 10^{17}$. The number of ways of arranging the 40 quanta on just 18 of the atoms is 3.56×10^{14}. Thus, if all of the different arrangements of energy over atoms are equally likely—that's the random swapping, microscopic form of "thermodynamic equilibrium"—the chance of finding this system with all of its energy located on just 18 of the 20 atoms is $3.56 \times 10^{14}/1.22 \times 10^{17} = 0.0011$; that is, there's about a tenth of a percent chance of this happening spontaneously. This result is for just 20 atoms and 40 quanta. In a real macroscopic system where the numbers of atoms and quanta are about 10^{25} the chance that any even slight spontaneous concentration of energy would occur is unimaginably small (although, of course, it could happen). The point is, if we start a system off with its energy concentrated and let random atomic swapping processes mix energy units around for a while, the chance that the energy will spontaneously (just via the swapping processes) reassemble itself into a concentrated state is essentially zero. Increasing the number of ways to distribute the available energy among the atoms of a system degrades the usefulness of the energy. Thus, from this perspective thermodynamic equilibrium is just a matter of counting: there are vastly more states a large system can be in with its energy scattered about (and less useful) than states with energy clumped (and more useful).

It can be shown that entropy can be defined in an equivalent, strictly thermodynamic way based on the heat flow into or out of a system and its temperature. In these terms the second law of thermodynamics is written as

$$\Delta S \geq \frac{Q}{T}, \tag{13.3}$$

where Q is the heat input to the system at absolute temperature T and the equality holds again only for processes that are quasistatic. Loosely speaking, $T\Delta S$ is a measure of the energy content of the "order" in a system. From Equation (13.3) it is seen

that entropy has units of J/K or kcal/K, but often is expressed in molar units of kcal/(mole-K).

So, then, how do we understand the macroscopic relation $\Delta S \geq Q/T$ in terms of atoms? This macroscopic relation says that when heat flows into a system at constant temperature, the system's entropy increases. But, microscopically entropy is disorder. So how does $Q/T > 0$ imply greater disorder? Well, if heat flows into a system at constant temperature, the system's volume has to increase, otherwise, the internal energy would increase and the temperature would increase also. Because the volume increases, the region of confinement of each atom increases and the energy level spacing decreases (recall that $\varepsilon \approx 1/L^2$). The internal energy of the system starts out at εN_E, but ε changes to a smaller value when volume increases. To keep the internal energy constant (temperature is constant) N_E must therefore increase. This means that the number of quanta in the system increases when heat flows in at constant temperature. Increase in quanta, as we have argued, produces an increase in the number of ways of dividing up quanta among atoms, or more disorder. This is the microscopic reason why entropy change is Q/T.

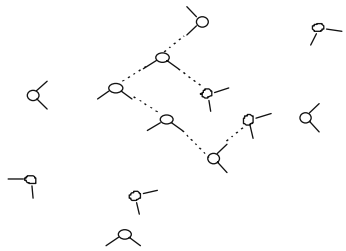


FIGURE 13.5 *Cluster of water formed by hydrogen bonding.*

Remember that the second law speaks of the total entropy of a *closed system.* We have seen that there are two classes of systems: closed, exchanging only heat but not mass with the surroundings, and open, exchanging mass as well as heat. The second law applies directly only to closed systems. Open systems can appear to violate the second law and have a decreasing entropy. *Life itself is fundamentally a process that reduces entropy in a series of self-organizing processes*. There is no violation of the second law because life cannot occur as a closed system. When the surroundings are included, the total entropy of the larger closed system always tends to increase. We are able to create ordered structures within our cells and organs at the expense of excess energy that we acquire from food. Said differently, we are able to live (and reduce our entropy) by increasing the entropy of our surroundings even more.

An interesting and important example of a molecular application of entropy is the structure of water. Water molecules are polar structures that form long-range hydrogen bonds that we study in Chapter 15. Those bonds are relatively weak and constantly break and reform on a picosecond (10^{-12} s) timescale. Because each water molecule has two hydrogen atoms and therefore can have two possible hydrogen bonds, water can form a network of bonds (illustrated in Figure 13.5), known as a cluster, that may persist for ~30 ps before "dissolving." Pure water can be pictured as a dynamic assembly of clusters that constantly break and re-form so that there is a fairly high degree of ordering in the water. In fact, the highly unusual thermal expansion property of water below 4°C, discussed in Section 2 of the previous chapter, is due precisely to the nature of the growing cluster formation as the temperature approaches the freezing point.

When a macromolecule is immersed in water it disrupts the organized clustering of water molecules in its neighborhood. Due to this effect polar regions on the macromolecule will tend to lie near water whereas hydrophobic portions tend to pack together internally to minimize contacts with water. An unfolded macromolecule will spontaneously fold into a characteristic native conformation (see Section 3 below). This phenomenon appears to be driven by strong interactions between the hydrophobic portions of the macromolecules, and is therefore called the *hydrophobic interaction*, but in fact the dominant interactions are entropic and are driven by the water hydrogen bonding. Minimum energy with the macromolecule impurity present is achieved in the more ordered state with water structure maintained as well as possible. The same effect occurs in membranes where the hydrophobic lipids aggregate within the membrane bilayer so that the polar heads can be exposed to water, minimizing the decrease in ordering of the water. This explains the very common bilayer structure of biological membranes.

2. GIBBS FREE ENERGY

So far in our discussion of thermodynamics we have studied two energy functions, the internal energy U and the enthalpy H, $H = U + PV$, both introduced in the last chapter. We have also seen that in a closed system the entropy will be maximized,

We can show, in a straightforward way, that the Gibbs free energy must decrease with time t in an isobaric isothermal process. Using the first law, $dU = Q - W$, for our system, and writing $W = PdV$, we have

$$Q = dU + PdV = dU + d(PV) = dH,$$

where we have assumed an isobaric process ($dP = 0$, so that $d(PV) = PdV$), and used the definition $H = U + PV$. Inserting this expression for Q into the thermodynamic form of the second law (Equation (13.3)), we have

$$Q = dH \leq TdS,$$

and by differentiating with respect to time we can write

$$\frac{dH}{dt} \leq T\frac{dS}{dt}.$$

Putting both terms on the left side of the inequality

$$\frac{d(H - TS)}{dt} = \frac{dG}{dt} \leq 0,$$

where we have also assumed an isothermal process ($dT = 0$). We conclude that the free energy decreases with time for all such systems until a minimum is reached at which thermal equilibrium has been established. In the special case when no heat flows ($dH/dt = 0$) then the decrease in free energy is matched by the increase in entropy alone.

To investigate the significance of the free energy, we start with its definition (Equation (13.4)) and write (using the product rule) that

$$dG = dU + PdV + VdP - TdS - SdT,$$

so that in an isobaric isothermal process ($P = T =$ constant), we have

$$dG = dU + PdV - TdS.$$

Writing the first law as $dU = Q - W$, and noting that for a reversible process $Q = TdS$, we have on substituting for dU, and in an open system the entropy of the {system + surroundings} will be maximized. It is useful to introduce another energy function, the Gibbs free energy G, that is particularly useful in open systems at constant temperature and pressure, the usual conditions in biology. We show that the free energy of an open system tends to decrease and that events (such as chemical reactions) will proceed spontaneously so long as the free energy decreases.

The *Gibbs free energy* is defined by

$$G = H - TS = U + PV - TS. \tag{13.4}$$

Under conditions of constant temperature and pressure, the only energy changes that can occur within an open system are $P\Delta V$ work, heat flow to or from the surroundings and other forms of useful work such as chemical or electrical work. Under those conditions, changes in free energy represents just those changes in "useful" work, hence the term "free," meaning available to do such useful work. The discussion in the box shows this and that the Gibbs free energy must always decrease as a system approaches equilibrium and must remain at that minimum value at equilibrium.

For an isothermal process $\Delta G = \Delta H - T\Delta S$, thus depending on the signs of ΔH and ΔS for a particular system we can distinguish four different possibilities (see Table 13.1). If $\Delta H < 0$ and $\Delta S > 0$ then ΔG is certainly negative and the process will occur spontaneously, decreasing the free energy until equilibrium occurs. Similarly if $\Delta H > 0$ and $\Delta S < 0$, then ΔG is positive and the process cannot proceed spontaneously, but could only proceed with some outside energy source. The two other cases are not as clear. If $\Delta H > 0$ and $\Delta S > 0$, then ΔG will be positive at low temperature, but may become negative at high temperature. Similarly if $\Delta H < 0$ and $\Delta S < 0$, then ΔG will be negative at low temperature, but will become positive at high temperature. In these cases the process will only be spontaneous below or above a threshold temperature.

Table 13.1 *Spontaneity of Thermodynamic Processes*

ΔH	ΔS	ΔG	*Reaction Occurs*
<0	>0	<0	Always
>0	<0	>0	Never
<0	<0	<0 at low T	Only at low T
>0	>0	<0 at high T	Only at high T

The rest of this section explores the application of Gibbs free energy to various types of chemical reactions as a prelude to the next section on biological applications. In a solution, chemical work can be done by changing the numbers and types of components (reactants and products) within the system. In this case the change in the Gibbs free energy can be written as

$$\Delta G = \Sigma(\Delta G_i) = \Sigma(\mu_i \Delta n_i), \tag{13.5}$$

where the summation is over all the species $\{i\}$ in solution, n_i is the number of moles of species i, and μ_i is the Gibbs free energy per mole, known as the *chemical potential*, of species i.

In the special simple case of a phase equilibrium between two species, for example, water and ice, Equation (13.5) becomes

$$\Delta G = \mu_w \Delta n_w + \mu_i \Delta n_i. \tag{13.6}$$

Because $n_w + n_i =$ constant, we know that any change in the number of moles of one species is due to the opposite change in the other so that $\Delta n_w = -\Delta n_i$. It follows from Equation (13.6) that at equilibrium when $\Delta G = 0$, we must have

$$\mu_w = \mu_i,$$

so that the molar chemical potentials of both species must be equal at equilibrium.

Let's consider in some detail the thermodynamics of a general bimolecular chemical reaction

$$n_A A + n_B B \Leftrightarrow n_C C + n_D D, \tag{13.7}$$

where n_A is the relative number of moles of species A reacting with n_B moles of B to produce n_C moles of C and n_D moles of D. To proceed, we need to know that the chemical potential can be written for the ith ideal solution component as

$$\mu_i = \mu_i^0 + RT \ln(c_i), \tag{13.8}$$

where μ_i^0 is its chemical potential at some standard condition and c_i is its molar concentration. At equilibrium the total Gibbs free energy of the reactants must equal that of the products, resulting in $\Delta G = 0$ for the reaction, so that we can write

$$n_C \mu_C + n_D \mu_D = n_A \mu_A + n_B \mu_B.$$

Substituting expressions from Equation (13.8) with appropriate subscripts for each term, we have

$$(n_C \mu_C^0 + n_D \mu_D^0 - n_A \mu_A^0 - n_B \mu_B^0) + RT(n_C \ln(c_C) + n_D \ln(c_D) - n_A \ln(c_A) - n_B \ln(c_B)) = 0,$$

and after using the mathematical facts that $n \ln(c) = \ln(c^n)$ as well as that ln A + ln B = ln(AB), we find

$$\Delta G^0_{\text{total}} + RT \ln\left(\frac{{c_C}^{n_C} {c_D}^{n_D}}{{c_A}^{n_A} {c_B}^{n_B}}\right) = 0, \tag{13.9}$$

where G^0_{total} is the first term in parentheses in the previous equation, equal to the net standard free energy for the reaction. In this expression the c_is are now the equilibrium molar concentrations, although for clarity we do not label them differently. Defining the equilibrium constant for the reaction K_{eq} as the term in brackets we have that

$$\Delta G^0_{\text{total}} = -RT \ln K_{eq}. \tag{13.10}$$

Note the general form of the equilibrium constant, having its numerator equal to the product of the equilibrium molar concentrations of the reaction products, each raised to the appropriate relative number of moles (as in the balanced chemical reaction equation) and its denominator equal to the same relation for the reactants.

Let's pause to digest these important results. In Equation (13.10), we note that if $K_{\text{eq}} > 1$ then $\Delta G^0_{\text{total}} < 0$ and the reaction will proceed spontaneously under standard conditions, a so-called exothermic reaction. If $K_{\text{eq}} < 1$ then $\Delta G^0_{\text{total}} > 0$ and the

$$\begin{aligned} dG &= (TdS - W) + PdV - TdS \\ &= PdV - W, \end{aligned}$$

or

$$-dG = W - PdV.$$

We conclude that for a reversible isobaric, isothermal process the decrease in Gibbs free energy is equal to the "useful," non-PdV, work done by the system.

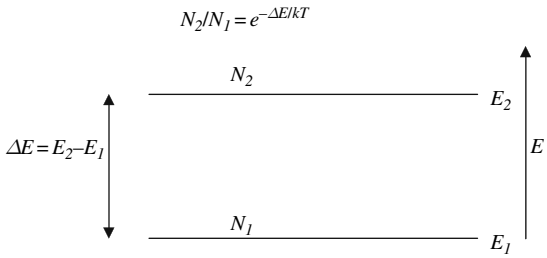


FIGURE 13.6 *The Boltzmann factor gives the relative populations of two energy levels with populations N and energies E.*

reaction cannot proceed spontaneously, but requires external energy in order to occur, a so-called endothermic reaction. In biology many reactions are *coupled reactions* in which energy from a spontaneous exothermic reaction may be used to drive an otherwise unallowed endothermic reaction. The hydrolysis of ATP to ADP is the most common such spontaneous reaction in cells with a value of $\Delta G^0{}_{\text{total}} = -7$ kcal/mole at standard conditions (25°C, pH 7; not those in a cell) and is used to "drive" many endothermic coupled reactions. We discuss some aspects of the thermodynamics of ATP hydrolysis in the next section.

A second important point about Equation (13.10) is that it can be solved for K_{eq}

$$K_{eq} = e^{-\frac{\Delta G_{total}{}^{o}}{RT}}, \tag{13.11}$$

so that a measurement of $\Delta G^0{}_{\text{total}}$ can be used to determine the equilibrium constant of the reaction. The term $e^{-\Delta G/RT}$, or when more commonly written in *per particle* instead of *per mole* form, $e^{-\Delta E/k_B T}$, is known as the *Boltzmann factor* and gives the relative populations of the two states separated by energy ΔE (Figure 13.6). Note that, as we saw in the last chapter, the term k_BT is an energy, equal at room temperature (20°C) to 4×10^{-21} J or 1/40eV (electron-volt, where 1 eV = 1.6×10^{-19} J). In the previous chapter we saw that k_BT is roughly the thermal energy of a gas molecule, so that the ratio in the exponent of the Boltzmann factor is comparing the energy difference between the two states to the thermal energy of a particle. When ΔE is large compared to thermal energies, the exponent is large and negative so that the population of the higher energy state is very small compared to that of the lower energy state. There is not enough thermal energy to excite reasonable numbers of particles to the higher energy state. On the other hand if ΔE is small compared to k_BT then the exponent is close to zero and the exponential is close to one, so that the populations of the two states are comparable because it is easy to make an upward energy transition because there is sufficient thermal energy available. We use the Boltzmann factor in later studies of atomic and molecular systems.

Our discussion has been based on equilibrium thermodynamics alone and as such does not give any information on times to reach equilibrium. Predictions can be made of whether reactions will occur spontaneously, but the rates of reactions cannot be determined from equilibrium thermodynamics alone. In concluding this section, we briefly consider some issues from reaction kinetics that concern the time-dependence of reactions. We focus on a simplified version of the bimolecular reaction given by Equation (13.7) in which two reactant molecules, A and B, produce two product molecules, C and D (so that all $n_i = 1$ in Equation (13.7)).

In a general way the steps of the reaction can be divided into three parts: the approach of A and B (often by diffusion), the reaction, and the separation of C and D. The free energy of interaction between molecules can be schematically represented as a function of the reaction coordinate, as shown in Figure 13.7, where the reaction coordinate is a parameter that indicates the progress of the reaction and so is related, but not necessarily proportional, to the elapsed time. The overall free energy change for the reaction is the net difference between the free energies of the final and initial states.

Typically in such a reaction, there will be an energy barrier, or *activation energy*, that needs to be overcome before the reaction products can be formed. This may be due to charge interactions or to steric effects requiring a more ordered arrangement of A and B before they can react. If this activation energy is small, then the "*rate-limiting step*" may be the simple coming together of A and B. In this case the reaction is known as *diffusion-controlled* (or *diffusion-limited*). With larger activation energy, the reaction is said to be *reaction-controlled.* In this case, remembering that the thermal energies of A and B are not equal but distributed about an average, only the more energetic molecules with sufficient energy to "climb" the

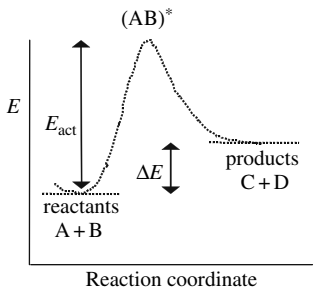


FIGURE 13.7 *An energy diagram for a general chemical reaction showing the activation energy E_{act} of the forward reaction and the overall free energy change ΔE. The particular reaction shown here is endothermic.*

energy barrier can interact and form an intermediate complex $(AB)^*$ that can then form products.

Many reaction-controlled processes in biology are modulated by enzymes, proteins that effectively lower the activation energy of reactions to enhance their completion in a process known as *catalysis*. Enzymes are highly specific, each having a unique active site at which binding to a specific macromolecule (substrate) occurs. Lowering of activation energies by enzymes may speed up that particular reaction by tremendous factors, often as much as 10^{14} times. Upon completion of the enzyme-assisted reaction, the enzyme molecules are released unchanged and can bind another substrate.

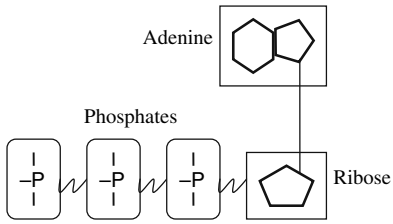


FIGURE 13.8 *Block diagram of the ATP (adenosine triphosphate) molecule with its high energy bonds (~).*

3. BIOLOGICAL APPLICATIONS OF STATISTICAL THERMODYNAMICS

In this section several examples of important biological processes are considered from a thermodynamic point of view. The energy-driving mechanisms of ATP hydrolysis and photosynthesis are first considered from an overall energy and molecular perspective. As they are important molecular processes in many facets of biology, we also briefly consider conformational transitions in macromolecules, including protein folding, helix–coil transitions in biopolymers and the self-assembly processes in polymerization.

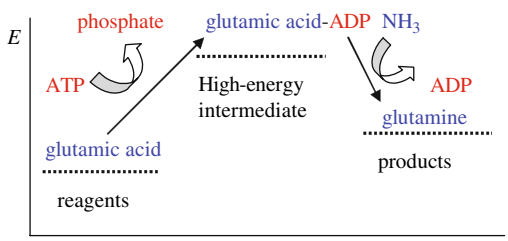
If the food we eat were to be simply burned, all its energy would go to heat. In order for our bodies to utilize some fraction of this energy, elaborate reactions occur that convert some of the energy stored in various foods into ATP. For example, each molecule of glucose, when completely oxidized, yields about 36 molecules of ATP, with an energy conversion efficiency of over 50%. Such an efficiency is much higher than that of manmade motors or engines with typical efficiencies of 10–20%. As you probably know, ATP is the predominant source of energy for chemical reactions in all living cells and is usually present at fairly high concentrations of 1–10 mM (where 1 mM $= 10^{-3}$ M).

The ATP molecule consists of the parts shown schematically in Figure 13.8: adenine with ribose attached and the three phosphate groups. Under physiological conditions, ATP is highly negatively charged and has divalent cations (Mg^{2+} or Ca^{2+}) bound. Hydrolysis (or splitting) of ATP involves the combining of a water molecule with the phosphate group farthest from the ribose to produce ADP and inorganic phosphate. The reaction releases a relatively large amount of energy; the farthest phosphate bond in ATP is said to be a high-energy phosphate bond. The precise total free energy change from the hydrolysis of ATP to ADP will depend on local concentrations of ATP, ADP, and phosphate but typical actual free energy changes in cells are quite large, ranging from $\Delta G = -11$ to -13 kcal/mole. This reaction is so favorable and likely to proceed spontaneously, that ATP must be constantly replenished in the mitochondria of cells. If allowed to reach thermal equilibrium, the cell would die. Rather, ATP concentration is maintained in a complex nonequilibrium steady-state reaction.

ATP plays an essential role in nearly all biosynthetic reactions, producing new protein (using most of the cell's ATP) as well as DNA, RNA, and polysaccharides in all cells. Each day an average adult hydrolyses as well as produces over 70 kg (roughly the person's weight) of ATP. The large free energy change of ATP hydrolysis can be linked with other reactions that have positive free energy changes, so that the coupled reactions become energetically feasible.

Let's consider an example reaction to illustrate the role of ATP in synthesizing glutamine, an amino acid. As with all such syntheses, the key to ATP's effect is the energetic coupling via a common intermediate. Figure 13.9 shows the free energy

FIGURE 13.9 *Free energy diagram for glutamine synthesis. The energy from ATP hydrolysis is used to form a high-energy intermediate from glutamic acid that subsequently combines with ammonia to form glutamine. The separate reaction #2 does not occur without energy input. Coupling of the two reactions #1 and #2 leads to an overall reaction that proceeds through the common intermediate with a net release of free energy.*



1 $\quad ATP + H_2O \longrightarrow ADP + phosphate$

2 $\quad glutamic\ acid + NH_3 \longrightarrow glutamine + H_2O$

1+2 $\quad ATP + glutamic\ acid + NH_3 \longrightarrow ADP + phosphate + glutamine$

changes associated with ATP hydrolysis and the unfavorable reaction forming glutamine from glutamic acid and ammonia. This latter reaction alone has a standard free energy change of $\Delta G^0 = +3.4$ kcal/mole and cannot proceed without a source of energy. Coupling with ATP hydrolysis to form a "high energy intermediate" allows the biosynthesis to occur with a net standard free energy change of $(-7 + 3.4 =) -3.6$ kcal/mole. In order to replace the macromolecular building blocks of the organism, ATP must be continually produced. All animals and most microorganisms rely on photosynthesis as their ultimate source of food.

Green, chlorophyll-containing, plants are the ultimate converters of energy supplied by the sun into oxygen and organic molecules that sustain life. In its most simplistic form, photosynthesis converts carbon dioxide and water to glucose and oxygen in an overall reaction

$$6CO_2 + 6H_2O \rightarrow C_6H_{12}O_6 + 6O_2.$$

The free energy change for this reaction is $\Delta G^0 = +686$ kcal/mole, so that clearly the process does not occur spontaneously, but must have an outside energy supply in the form of photons of light. Photosynthesis is a unique process that harvests photon energy into chemical energy. More than 100 sequential reaction steps have been elucidated in the overall reaction, each with a specific enzyme.

Briefly, the overall process can be divided into two major portions known as the light and dark reactions. The light reactions, requiring photons and unique to photosynthesis, first convert water to free oxygen, protons, and electrons. The protons are pumped across a membrane generating ATP, and the electrons bind to an enzyme (NADP) to be used in a subsequent coupled reaction. The dark reactions use the ATP and electron–donor enzyme NADP to convert carbon dioxide to glucose. For each carbon dioxide molecule 8 photons are needed for a total of 48 photons per glucose molecule. The efficiency of conversion of photon energy at the site of photon absorption, the reaction center, is about 20%, whereas the overall efficiency of photosynthesis is about 5% under optimal conditions. Uncovering the molecular details of photosynthesis is an active area of research involving lots of physics. For example, pulsed laser experiments carried out at very low temperatures have shown that the earliest steps in the direct absorption of a photon occur faster than 1 ps (10^{-12} s). Spectroscopy of various types has been essential in unraveling the kinetics and conformational changes that occur as the photon energy is distributed to various chemical bonds.

Finally, we consider some thermodynamic aspects of the conformations of macromolecules. As discussed in Section 5 of Chapter 3 there are certain biostructural motifs that are common in nature: the α-helix in proteins, the Watson–Crick double helix in DNA, or the self-association of identical protein molecules to form complex structures such as the filamentous polymer actin or smaller aggregates such as hemoglobin (Figure 13.10). Under certain conditions, macromolecules may spontaneously form these ordered conformations or aggregates from less well-ordered states of random coil or from isolated monomer subunits, respectively. The driving mechanisms are the detailed electrical bonds that form between portions of the macromolecules, or between individual subunits, stabilizing the overall structures. Even without that detailed electrical information, thermodynamic quantities can give some general information about the possible conformational reactions and some insight as to the mechanisms and stability of various ordered configurations of macromolecules.

Proteins in their native form have unique conformations that consist of regions of more (helix, β-sheet) or less (random coil) order. If a protein is mildly heated so enough thermal energy is added to break the weaker bonds that maintain the secondary conformation, but not so much as to break covalent bonds along the protein backbone, then the protein can lose its overall structure and become entirely random coil in a process known as denaturation. If cooled under controlled conditions, proteins will often spontaneously renature to form native, functioning protein molecules.

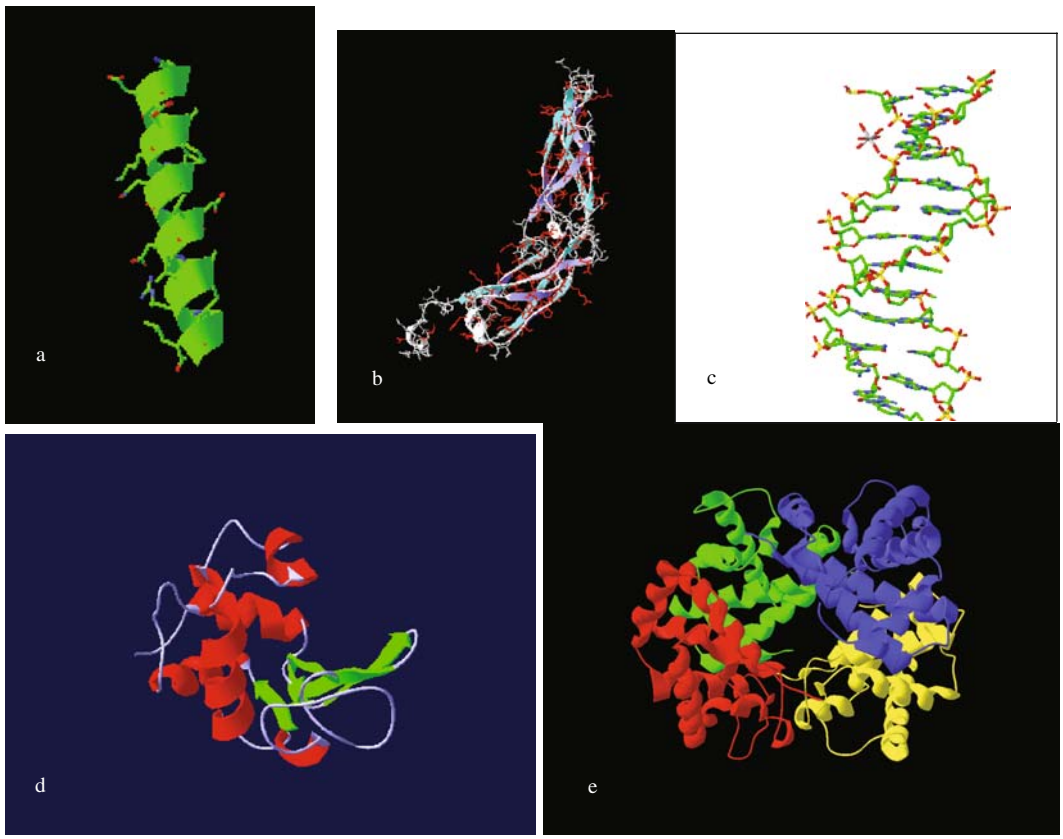


FIGURE 13.10 *Three structural motifs in biomolecules: (a) alpha helix, (b) beta-sheet, and (c) double helix. (d) The protein lysozyme showing regions of alpha helix* (red) *and beta sheet* (green) *as well as random coil, and (e) hemoglobin, composed of four identical subunits shown in colors.*

We can understand this behavior from some simple thermodynamic arguments. Comparing the denatured and native helical (for example) conformations, it is clear that the entropy of the denatured state is greater. This is due to the fact that the coil is a much more random structure with many more possible ways to distribute its energy and thus a much larger statistical weight Ω and entropy related through Equation (13.1). We can write this as $\Delta S_{\text{coil}} > 0$, with reference to the helix state. Furthermore, it is clear that in order to disrupt the secondary bonding to form the coil from the helix, heat must be input and so $\Delta H_{\text{coil}} > 0$ for the coil, again compared to the helix. Combining these, we see that

$$\Delta G_{\text{coil}} = \Delta H_{\text{coil}} - T\,\Delta S_{\text{coil}}$$

may be positive at low temperatures, but may become negative at a sufficiently high temperature (see Table 13.1). Thus, the helix is stable at lower temperatures whereas the coil is stable at higher temperatures.

Furthermore, we know that $K_{\text{eq}} = e^{-\Delta G_{\text{coil}}/RT}$, where

$$K_{\text{eq}} = \frac{c_{\text{coil}}}{c_{\text{helix}}},$$

with these concentrations representing the fraction of protein residues in each conformation. The transition from having most residues in the helix (small K_{eq} and therefore large $\Delta G_{\text{coil}} > 0$) to having most in the coil (large K_{eq} and therefore large $\Delta G_{\text{coil}} < 0$) will occur over a range of temperature as the protein is heated. Note that if both ΔH and ΔS are themselves large as well as positive, then whether their

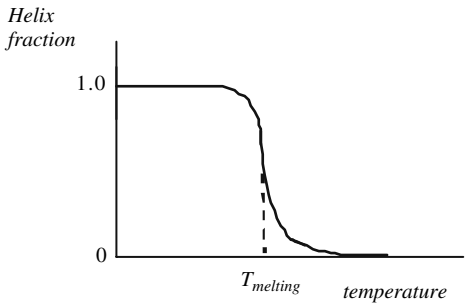


FIGURE 13.11 *Typical temperature dependence of the melting of a helical protein. The cooperative transition temperature is characterized by a relatively sharp decrease in the helix content of the protein.*

difference in the expression for ΔG ($= \Delta H - T\Delta S$) is positive or negative becomes a very sensitive function of T and the "melting transition" of a protein will occur over a narrow temperature range as is actually observed for most proteins (Figure 13.11). Because the values of ΔH and ΔS are modest for each residue's bonds, this sharp melting occurs as a result of a *cooperative transition* in which many residues melt simultaneously.

Similarly if the coil-to-helix transition is monitored, one discovers that this transition is also cooperative, meaning that after several energetically costly bonds are formed, subsequent bonding occurs with less energy required per bond. The large initial energy needed to form the several bonds that greatly restrict possible conformations of the backbone substantially decreases the entropy. Once that initial start is formed in the helix, additional neighboring bonds form rapidly with less energy per bond required.

For the case of subunit assembly in a protein or other biopolymer, there is a decrease in entropy as subunits form a larger structure. This is true because the overall translational and rotational motion of the subunits are coupled together and many side chains become immobilized as well, reducing the number of degrees of freedom and thereby increasing the order. A typical decrease in entropy of dimerization is about 0.1 kcal/mol-K, corresponding to about +30 kcal/mol of free energy (the term $-T\Delta S$, with $T \sim 300$ K) at room temperature. In order for dimerization to proceed spontaneously, there must be a source of free energy for the reaction so that the overall free energy change is negative. Most of this energy comes from hydrophobic interactions when water is excluded from the surface area of subunit contact. Because on dimerization less total protein surface is exposed to water, there is a decrease in this contribution to the free energy as discussed in Section 1 above. Estimates are that in a typical dimerization of a protein 10–20 nm^2 of surface area previously exposed to water becomes internalized within the dimer. At an average free energy change of about -2.5 kcal/mol/nm^2 of surface area, hydrophobic interactions result in a ΔG of -25 to -50 kcal/mol. In addition there are specific bonds (hydrogen, van der Waals) between the protein subunits causing the dimer to be stabilized. Many macromolecules can continue to add subunits spontaneously and rapidly to form a long polymer molecule. Included are such important molecules as DNA, RNA, and the proteins actin and tubulin.

CHAPTER SUMMARY

In treating macroscopic systems composed of large numbers of particles, statistical methods are used. A *microstate* is defined as a detailed specific state (one of an extremely large number) in which each atom in the system has a particular energy level. A *macrostate*, in contrast, is defined by the set of energy levels and the numbers of atoms in each level, the occupation numbers; this information defines the overall energy of the system, but, in general, there are many, many microstates that all produce the same macrostate. *Entropy*, S, is defined in terms of the statistical weight of the system Ω, which is a function that contains all the occupation number information, as

$$S = k_B \ln \Omega. \qquad (13.1)$$

Whereas overall energy conservation holds for an isolated system, various forms of energy have different degrees of "order," or "usefulness," or entropy. For example, such a system of particles with only thermal energy, in the form of random diffusive motions, is less ordered and less useful than the equivalent amount of energy in the form of overall translational kinetic energy. The system with only an overall translational energy will have lower entropy than the thermal system because such a translating system is much more ordered and there are very many fewer ways that the energy can be distributed over the possible macrostates. On the other hand, such a system will tend to thermalize, or randomize its motion over time, heading toward the thermal system, and thus increasing its entropy over time. This idea is contained in the second law of

thermodynamics, which states that the total entropy of a closed system always increases,

$$\Delta S \geq 0, \tag{13.2}$$

with $\Delta S = 0$ only in the special case of a reversible process. An alternate statement of this law is contained in

$$\Delta S \geq \frac{Q}{T}, \tag{13.3}$$

where Q is the heat input to the system at absolute temperature T.

Another thermodynamic variable that is particularly useful in open systems at constant pressure and temperature, conditions often occurring in biology, is the Gibbs free energy, G,

$$G = H - TS = U + PV - TS. \tag{13.4}$$

As an example of its utility, G can be related to the equilibrium constant of a chemical reaction, K_{eq},

$$\Delta G^0_{\text{total}} = -RT \ln K_{eq}. \tag{13.10}$$

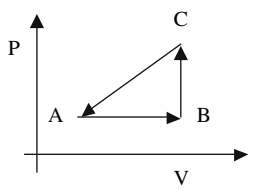
Given a set of energy levels in an atomic system, with energies E_i and populations N_i, the Boltzmann factor gives the relative populations of any two states, for example 1 and 2, as

$$\frac{N_2}{N_1} = e^{-(E_2 - E_1)/k_B T}.$$

In Section 3, we considered two specific applications of some of these ideas: coupled kinetic reactions in the hydrolysis of ATP and the helix–coil melting transition of a protein. Analysis of both of these involves studying the Gibbs free energy changes, resulting from both enthalpy and entropy changes.

QUESTIONS

1. The figure below shows a P–V diagram in which an ideal gas goes from state A to state A in a reversible cycle via the processes A→B, B→C, C→A. In each entry of the following table insert +, −, or 0 to indicate the sign of the associated quantity.



	ΔU	Q	W	ΔS
A→B				
B→C				
C→A				
Total				

2. In the following table check the boxes of those quantities that must be zero in the respective reversible process. Assume the system is an ideal gas.

	Isobaric	*Isothermal*	*Isochoric*	*Adiabatic*
ΔU				
ΔT				
ΔP				
ΔV				
ΔS				
Q				
W				

3. Please order the following from highest to lowest entropy: 1 kg of ice, water, and water vapor.

4. Discuss a colloquial statement of the second law: the energy available for useful work always decreases.

5. Find three examples of a system going from less ordered to more ordered and discuss why the second law of thermodynamics is not violated in each case.

6. Some cashiers arrange dollar bills to all face the same way, whereas others do not. Which pile of bills has more entropy?

7. Discuss the molecular basis of the hydrophobic effect. In particular which is the more fundamental process: the attraction of hydrophobic portions of a macromolecular structure, or the minimization of the disruption of hydrogen bonding in water?
8. Discuss why the Gibbs free energy is appropriately named "free."
9. Discuss the difference between an endothermic and an exothermic reaction. What state variable determines which one a particular reaction is?
10. Explain the difference between reaction and diffusion-controlled chemical processes.
11. What is the difference between the reversible melting of a biopolymer and its irreversible denaturation?
12. What does it mean for a transition in a macromolecule to be cooperative? Give an example.
13. What is the function of an enzyme?

MULTIPLE CHOICE QUESTIONS

1. Which of the following statements is false? The entropy of a closed system (a) is a measure of its disorder, (b) always increases unless the process is quasistatic, (c) is a measure of the dilution of internal energy among allowed microstates of the system, (d) is proportional to the statistical weight of the system.
2. Suppose there are three identical atoms each with energy levels given in Figure 13.3. If the total energy of the system is 3ε, the number of macrostates of the system is (a) 1, (b) 2, (c) 3, (d) 4.
3. In the previous question one of the macrostates is (1, 1, 1, 0) using the notation of Example 13.2. How many microstates correspond to this macrostate? (a) 1, (b) 2, (c) 3, (d) 6.
4. A hypothetical engine operates in a cycle taking in 10,000 J from a hot reservoir and 5000 J from a cold reservoir. In the cycle it performs 15,000 J of work. Such an engine (a) obeys both the first and second laws of thermodynamics, (b) obeys the first law but violates the second law of thermodynamics, (c) violates the first law but obeys the second law of thermodynamics, (d) violates both the first and second laws of thermodynamics.
5. The zeroth law of thermodynamics concerns bodies A, B, and C, and the relation "is in thermal equilibrium with." Suppose each of the following relations is substituted for "is in thermal equilibrium with." For which relation will the "zeroth law" fail? (a) "communicates via email with," (b) "is as tall as," (c) "works in the same building with," (assume one job for each), (d) "owns the same model car as" (assume one car for each).
6. Living cells constitute a low entropy state of matter. Living cells (a) violate the second law of thermodynamics, (b) can exist because they help increase the entropy of the rest of the universe, (c) are not subject to physical laws such as thermodynamics, (d) demonstrate that the laws of thermodynamics are incomplete.

PROBLEMS

1. At rest, our bodies generate heat at a rate of about 100 W. Calculate the minimum amount of entropy we generate in a day, neglecting the small entropy increase from eating.
2. What is the entropy change of a cube of water 1 cm on a side that freezes at 0°C?
3. Repeat the calculations of Example 13.1 for the case of six coins. Make a table showing the possible microstates and macrostates and find the probabilities of each macrostate.
4. The splitting of ATP can be schematically given as $\mathrm{ATP} + \mathrm{H_2O} \rightarrow \mathrm{ADP} + \mathrm{P}$. If the reaction has a $\Delta G = -7$ kcal/mole at 25°C, what is the equilibrium constant at that temperature?
5. If Equation (13.10) is solved for ($\ln K_{eq}$) and ($\Delta H - T\Delta S$) is substituted for ΔG, we can write that

$$\ln K_{eq} = \frac{-\Delta H}{RT} + \frac{\Delta S}{R}.$$

   Describe how you might use this equation to determine both ΔH and ΔS from a knowledge of K_{eq} as a function of temperature. Such a graphing procedure is known as a van't Hoff graph. What assumptions are involved in your analysis?
6. Suppose there are three identical atoms, each with energy levels shown in Figure 13.3. If the total energy of the system is 4ε, find all possible macrostates and the number of microstates for each of them. Use the notation of Example 13.2.
7. Re-do the previous problem for the case when the total energy of the three atoms is 6ε.

14 Electric Forces and Fields

In this chapter we begin our study of electromagnetism, one of the four fundamental interactions in nature. Aside from gravity, ultimately all of the forces that we are familiar with are due to electromagnetic interactions; pushes and pulls, normal, frictional, tension, compression, shear, and viscous forces are all electromagnetic in origin. Other forces that we learn about are also electromagnetic, including the historically diverse electric and magnetic forces as well as all the various chemical bonding forces. In fact, all of chemistry (other than nuclear chemistry) is basically electromagnetic in origin. Even more surprising is that light and other forms of (nonnuclear) radiation are electromagnetic in nature and can exert electromagnetic forces. Optics, the science of light, is thus also a branch of electromagnetism.

The basic laws of electromagnetism were developed over a 50 year span in the 19th century, culminating in Maxwell's four fundamental equations. Maxwell's equations are one of the most successful descriptions of our world, only requiring modification by quantum mechanics on the atomic distance scale. Aside from gravity, the other two fundamental forces in nature are nuclear forces that we do not experience directly in our daily lives. These are considered later in this book in connection with nuclear radiation and the fundamental structure of matter. In this and the next two chapters we turn our attention first to the nature of electricity, the electrical properties of matter, and methods used to study those properties.

1. ELECTRIC CHARGE AND CHARGE CONSERVATION

Humankind's first contact with electricity was through electrical storms and bolts of lightning hurled from the heavens with the power to kill or create fire (Figure 14.1). The Greeks discovered manmade static electricity, produced by friction, just as we know it today. Frizzy hair charged up by combing on a dry day and electrical sparks produced when touching a metal doorknob after walking on a thick carpet are common examples of static electricity buildup through friction. It is only in the 20th century that we have learned that these macroscopic phenomena are due to the elementary charged particles, electrons and protons, making up all atoms.

Our modern picture of matter, briefly introduced in Chapter 1, views atoms as composed of protons and neutrons within a central nucleus and electrons. Electric charge is a property of elementary particles that comes in two types, termed positive and negative, and in a quantized, or discrete, smallest possible unit. The quantum of electric charge is

$$e = 1.6 \times 10^{-19}\,\mathrm{C}$$

(the SI unit for electric charge is the coulomb, C, defined in Section 2 below) and is equal in magnitude to the electric charge of the electron or the proton. It is taken as

J. Newman, *Physics of the Life Sciences,* DOI: 10.1007/978-0-387-77259-2_14,

FIGURE 14.1 *Lightning strikes.*

positive so that the charge on the electron is $-e$. All known particles have been found to have electric charges that are multiples of $\pm e$.[1] Atoms have no net electrical charge, consisting of as many positively charged protons as negatively charged electrons and some number of neutral, or uncharged, neutrons.

The fact that there are two types of electric charge allows the electric force to be either attractive or repulsive. In contrast, there is only one type of mass and all masses attract each other via gravitational interaction. Electrical forces between like charges (either both positive or both negative) are repulsive, whereas those between unlike charges are attractive. In the next section we discuss the nature of the electrical force in more detail. Because protons are all positively charged, those in a nucleus (aside from hydrogen with its single proton) should repel one another so that the nucleus would be unstable. This argument compels one to search for another fundamental force that holds the nucleus together, the strong nuclear force, discussed later in this book.

Macroscopic matter is typically electrically neutral, being composed of neutral atoms and molecules. However, because the numbers of molecules are so large, even a relatively small fraction of charged atoms or molecules (known as *ions*) give an object a net charge and can lead to macroscopic electrical forces between charged objects. Often objects are charged by a transfer of electrons from another object so that one gains an excess of electrons and the other has an excess of protons. Furthermore, many neutral molecules have their centers of positive and negative charge offset (so-called *polar molecules*) in either a permanent fashion, as in water, or by inducing such a polarity through electrical interaction with other objects (Figure 14.2). In such cases, neutral molecules can interact electrically with net charges or even with other polar molecules, although the forces generated are weaker than those between charged molecules. The electrical properties of macroscopic objects are discussed in Section 3 below.

Among the pillars of modern science are the conservation laws of physics. We have already seen applications of the conservation of energy, linear momentum, and angular momentum in our discussions of mechanics.

Conservation of electric charge is another hallmark of science. It may be succinctly stated that the net electric charge in an isolated system remains constant.

Although apparently simple, it is a very powerful law that can be somewhat subtle as well. Its simplest form occurs in a system with a fixed population of elementary particles. In this case those particles remain unchanged. However, there are many systems in which the "fundamental" constituents may change identity and number.

As an example, although the proton and electron are stable particles, the isolated neutron decays to produce three other elementary particles (proton, electron, and antineutrino) in the following reaction

$$n^o \rightarrow p^{+1} + e^{-1} + \bar{\nu}^o,$$

where the superscripts indicate the electric charges. Isolated neutrons will decay by this reaction in a few minutes whereas those within a nucleus may be stable or decay on varying time scales. When a neutron within a nucleus decays, a new species of nucleus with one more proton and one fewer neutron forms in a process known as *beta-decay*. This process results in the ejection of a high-speed electron and antineutrino. Although

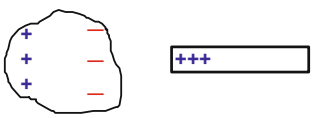


FIGURE 14.2 *The positively charged rod induces a separation of charges in the neutral object on the left.*

[1]Quarks, the theorized constituents of protons and other heavier elementary particles, have electric charge magnitudes of $e/3$ or $2e/3$ and are always found in combinations in nature resulting in integral charges.

this reaction is complex, it must satisfy a number of conservation laws, among them energy, momentum, and electric charge. In terms of electric charge, the original neutral neutron becomes three particles with electric charges +1, −1, and 0, so that the total final charge remains equal to zero. A second example is the production of matter from energy, in which a proton and an antiproton (negative antiparticle to the proton) annihilate to produce pure energy which then produces a set of pions; the initial zero electric charge is conserved even here in the production of matter since four positive and four negative pions are produced (Figure 14.3).

We see that charge conservation is basically a question of bookkeeping, maintaining the total net charge. Nature, the ultimate bookkeeper, seems to be exquisitely precise at conserving electric charge. At any time the total charge of the system remains constant, even if the numbers and cast of particles change. Conservation of electric charge has never been found wanting, no matter how complex the physical system may be.

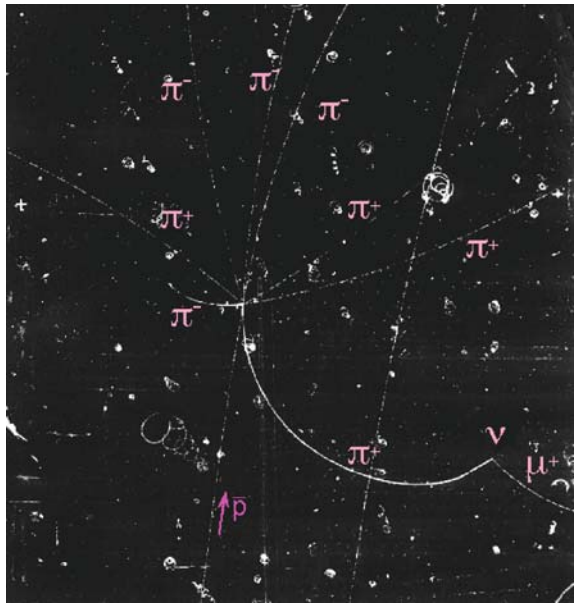


FIGURE 14.3 *Bubble chamber photo of the trail of an antiproton (labeled as $\bar{p}$) colliding with a stationary proton, annihilating each other to create pure energy which in turn created 8 pions (π). The chamber lies in a strong magnetic field that curves the oppositely charged particles in opposite directions. One of the pions subsequently decays into a muon and a neutrino which leaves no track.*

2. COULOMB'S LAW

The electrical force on a charged object may be determined from two pieces of knowledge. First, we need to know the fundamental law governing the force between any two charged particles, known as Coulomb's law. In addition, we need to appreciate the superposition principle that allows us to use the rules of vector addition to compute a net force on an object from individual forces from other charged particles based on Coulomb's law.

A charged particle (known as a point charge) exerts a force on a second point charge that is proportional to the product of their charges, inversely proportional to the square of their separation distance, and directed along the line joining the two particles,

$$\vec{F}_{1 \text{ on } 2} = k \frac{q_1 q_2}{r^2} \hat{r}, \tag{14.1}$$

where k is a constant of proportionality and $\hat{r}$ is a unit vector (a vector with a magnitude of one; remember that the special symbol ^ is used for unit vectors; you might want to review some basic ideas on vectors discussed in Chapter 5) pointing from particle 1 to particle 2 (Figure 14.4). Note that the sign of F changes from positive, if the charges are like (both negative or both positive), to negative, if the charges are unlike, indicating that the force is repulsive or attractive, respectively. Also remember that because of Newton's third law, the force of q_1 on q_2 is equal and opposite to that of q_2 on q_1, so that these two form an action–reaction pair of forces. The exponent on r is known to be very precisely 2; from careful experiments it has been determined to be 2.00 . . . out to 16 places after the decimal point, that is, to one part in 10^{16}.

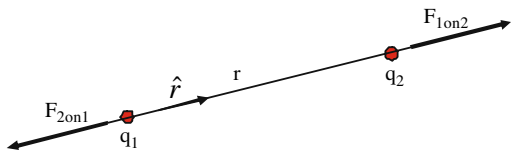
Coulombic forces are long-range forces, decreasing as $1/r^2$ the farther away the two interacting charges are, but in principle always remaining nonzero. We show in a discussion of charges in solution in Section 5 that in reality Coulombic forces do not extend infinitely far because there are always other nearby charges tending to shield them and effectively decrease their range. If the two charges are in a vacuum, the constant k is equal to

$$k = 9.0 \times 10^9 \ N\text{-}m^2/C^2,$$

but the constant varies in different media as we show.

Coulomb's law also applies to atomic systems even though quantum mechanics is needed to correctly describe the physics at those distances. As discussed above, the smallest electric charge found in nature is e, so that the

FIGURE 14.4 *The pair of equal and opposite Coulomb's law forces between two like point charges.*



force between a proton and an electron in an atom, with separation distance of 0.1 nm, is attractive with a magnitude given by

$$F = k\frac{e^2}{r^2} = 9 \times 10^9 \frac{(1.6 \times 10^{-19})^2}{(10^{-10})^2} = 2.3 \times 10^{-8}\, N.$$

Although this appears to be small, it is actually a relatively large force, as can be deduced by mentioning the recently measured force between a myosin and actin molecule (the major protein constituents of muscle) of several piconewtons ($10^{-12}\, N$), determined in a petri dish assay using a laser tweezers experimental technique (see Chapter 19).

Example 14.1 How much stronger is the electric force of a proton on an electron than the gravitational force between them?

Solution: In Equation (2.6), let M be the proton mass and m be the electron mass. In Equation (14.1), let $|Q| = |q| = e$. If we then divide Equation (14.1) by Equation (2.6) we get

$$\frac{F_{\text{electric, proton on electron}}}{F_{\text{gravity, proton on electron}}} = \frac{ke^2/r^2}{GMm/r^2} = \frac{ke^2}{GMm} = 2 \times 10^{39}.$$

(Plug in the values of k, e, G, M, and m to see that this is true.) This ratio is independent of the separation of the proton and the electron, because both the electric and gravitational forces depend on separation exactly the same way and the r^2's cancel in numerator and denominator. The electric force of one proton on one electron is about 10^{39} times greater than the gravitational force of the proton on the electron at any distance of separation.

As the previous example showed, the electrical force between the proton and electron is tremendously greater than their gravitational attraction, greater by a factor of about 2×10^{39} times. Whenever electrical forces are involved, gravitational forces can be completely neglected. It is only when objects are electrically neutral that it becomes necessary to include the gravitational force.

In order to simplify future equations, Coulomb's law is usually written in terms of another constant ε_0, the permittivity constant of the vacuum, where $k = 1/4\pi\varepsilon_0$ so that

$$\varepsilon_0 = 8.85 \times 10^{-12}\ \text{C}^2/N\text{-m}^2.$$

Coulomb's law can then also be written in the more common form,

$$\vec{F}_{1 \text{ on } 2} = \frac{1}{4\pi\varepsilon_0}\frac{q_1 q_2}{r^2}\,\hat{r}, \tag{14.2}$$

When there are more than two point charges involved in a system under study the superposition principle for forces allows one to find the net force on one point charge by adding up the individual vector forces acting on that charge. We can write this as a simple vector addition

$$\vec{F}_{\text{net}} = \sum \vec{F}_i, \tag{14.3}$$

where it is implied that the sum is over the forces due to all other charges present. Recall that in vector addition we do not just add the magnitudes of the forces algebraically. An example helps to illustrate this.

Example 14.2 Find the net force on a 4.0 μC charge at a corner of a square with 20 cm sides if the two neighboring corners have charges of −3.0 μC and 5.0 μC as shown in Figure 14.5.

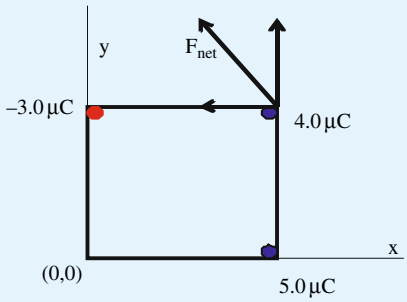


FIGURE 14.5 *Point charge arrangement for Example 14.2 showing the forces acting on the 4 μC charge.*

Solution: We first separately find the force on the 4 μC charge from each of the other two charges using Coulomb's law, keeping track of the direction of those two forces. The force from the −3 μC charge is attractive, directed along the negative x-axis, and of magnitude

$$(9 \times 10^9)(3 \times 10^{-6})(4 \times 10^{-6})/(0.2)^2 = 2.7 \text{ N}.$$

Similarly the force from the 5 μC charge is repulsive, directed along the positive y-axis, and of magnitude

$$(9 \times 10^9)(5 \times 10^{-6})(4 \times 10^{-6})/(0.2)^2 = 4.5 \text{ N}.$$

The net force is then given, in ordered pair notation, by

$$\vec{F}_{\text{net}} = (-2.7, 4.5) \text{ N},$$

so that its magnitude is

$$F_{\text{net}} = \sqrt{(2.7)^2 + (4.5)^2} = 5.2 \text{ N},$$

and it is directed at an angle of

$$\theta = \tan^{-1}(4.5/2.7) = 59^\circ$$

from the negative x-axis (or 121° from the x-axis).

To briefly review, the major steps in solving problems of this type are to first find the individual vector forces produced and then use the rules of vector addition to find the magnitude and direction of the net force, if needed.

As an example of the use of calculus to find the force on a point charge due to a charge distribution, let's calculate the force on a positive point charge a perpendicular distance d from a very long straight line of positive electric charge with a uniform charge per unit length, $\lambda = Q/L$, along the x-axis as shown in Figure 14.6. We divide the line of charge into infinitesimal elements of length dx with charge λdx and use Coulomb's law to write an expression for the force on the point charge from this element of charge. This force will be along the line joining the two charges. It is clear that there will be another element of charge symmetrically placed so that when we add its force on the point charge, the x-components will cancel and there will only remain a repulsive force along the perpendicular direction to the line of charge as shown. The net force on q from the pair of symmetrically placed line charge elements is

$$dF = 2\cos\theta \frac{1}{4\pi\varepsilon_0}\frac{q(\lambda dx)}{r^2}.$$

Substituting (d/r) for $\cos\theta$ and $[x^2 + d^2]^{1/2}$ for r, and integrating from 0 (we've already included the charges along the negative x-axis so we only integrate along the positive axis) to ∞, we have

$$F = \frac{1}{2\pi\varepsilon_0}\int_0^\infty \frac{\lambda q d}{[x^2 + d^2]^{3/2}}\,dx.$$

After a trigonometric substitution and a bit of work, the result of the integration is

$$F = \frac{\lambda q}{2\pi\varepsilon_0 d}.$$

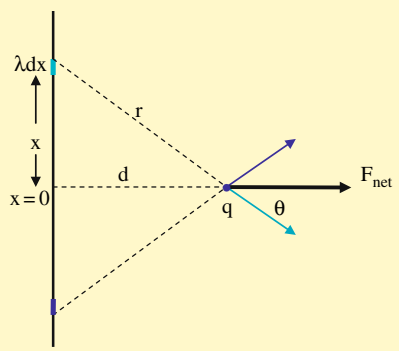


FIGURE 14.6 *Geometry for the boxed infinite line charge example.*

If a real extended object is charged by, for example, transfer of charge to its surface, then the distribution of the charge on the object will depend on its electrical characteristics. We study the basic differences in the electrical properties of materials in the next section. To find the electrical force between real charged objects, it is not immediately clear how to determine values to use for r in Equation (14.2). If the separation distance is much greater than the dimensions of the object, then we can treat the objects as points. With spherical objects charged so that the electrical charge distributes itself uniformly around the sphere (as we say, "in a spherically symmetric manner"), we can take the distance r to be the center-to-center distance regardless of the separation distance of the surfaces of the spheres. An example calculation for the force on a point charge from an extended object is given in the box. In Section 4 we show another method for such calculations.

3. CONDUCTORS AND INSULATORS

Electrical properties of materials are determined by their atomic structure. In particular, the nature of the binding of the outermost (*valence*) electrons of the atoms in the material defines its electrical interactions. Other atomic electrons closer to the nucleus do not take part in interatomic interactions. In a solid composed of an enormous number of identical atoms, the atoms or molecules are strongly interacting and are often arranged in a crystalline well-ordered array. We show in Chapter 25 that as a consequence of the quantum nature of the atomic interactions solids can be divided into three distinct classes based on their electrical properties.

In one class, known as electrical *conductors*, including metals such as copper, iron, and aluminum as the most common members, the outermost electrons of the atoms are not bound to any particular atom but are free to migrate about in the solid. Although the conductor as a whole remains electrically neutral, these "*free electrons*" can wander about under the influence of electric forces and give rise to the characteristic ability of conductors to allow a ready flow of electrons. In the absence of an externally applied electric force, these free electrons still migrate about in their local lattice, or array, of positive metal ions in a random diffusive motion so that the solid remains locally electrically neutral. When an external electric force is applied to a conductor, the electrons immediately respond throughout the conductor, making up an electric current, or flow of electrons, which we study in Chapter 16.

A second class of solids, known as electrical *insulators* or *dielectrics*, consists of materials whose outermost electrons are very tightly bound to individual atoms and are not at all free to move even under the influence of rather large forces. Common insulators include rubber, wood, glass, and most plastics. These are very poor conductors of electricity because the electrons are so tightly bound to the atoms of the solid lattice.

Usually materials that are good electrical conductors are also good thermal conductors and those that are good electrical insulators are also good thermal insulators. This is explained by the observation that motion of free electrons is the predominant mechanism for heat conduction (random or diffusive free electron motions) as well as electrical conduction (drift velocity of free electrons). Electrical insulators with few, if any, free electrons are also poor thermal conductors.

Air is also a good insulator, although under extreme conditions at which the electrical forces are very large, air molecules can become ionized, in a process known as *dielectric breakdown* (Figure 14.7). When this occurs the air becomes conducting and a spark jumps through the air between conducting surfaces, such as between your fingers and a metal doorknob on a dry day. Under the right atmospheric conditions, lightning may discharge by charge transfer to the Earth, a conductor with infinite storage capability. In the case of a doorknob the spark contains a relatively small total charge. Lightning often contains huge amounts of charge and is correspondingly much more dangerous. The ionized air is known as a *plasma*, a gas of ionized particles. Often plasma is considered a fourth state of matter (in addition to solids, liquids, and gases) because of its unusual properties.

Pure water is also a good insulator, because it has few ions to transport charge. The normal high conductivity of water is due to the presence of contaminating ions, usually salts and metal ions. In Section 5 we study the electrical properties of solutions to learn about the electrical forces that macromolecules experience.

A third class of solids, known as *semiconductors*, has mixed electrical properties, sometimes acting as a good insulator, but also capable of conducting electric currents. Silicon and germanium are the two most common semiconductor materials; these behave intrinsically as semiconductors. Today, nearly all electrical devices contain semiconductor materials, characterized by normally being insulators, but through the use of small controlling signals, able to become good conductors of electricity. Semiconductor "microchips" can be manufactured with specific desired properties by "doping" intrinsic semiconductor materials with small amounts of specific impurities designed to lead to the desired electrical performance. We study these in more detail in Chapter 25.

When an object has a net charge, either positive or negative, it has gained this charge by the flow of electrons. An excess of electrons on an object gives it a net negative charge, whereas a deficiency of electrons on that object gives it a net positive charge. The excess charge on an insulator remains locally where the charge was deposited, usually by contact with another charged object. On the other hand, the excess charge on a conductor adds to the free electron density and is rapidly distributed on the conductor, ending up on the surface of the conductor as we show in the next section. Most manmade electrical devices consist of layers of conductor, semiconductor, and insulator configured to perform specific functions. Perhaps the simplest is the electrical cord, consisting of copper conducting wire surrounded by a plastic or rubber layer. The copper wire is used because of its highly efficient transfer of free electrons along its length and the insulator functions to isolate the copper wire, not allowing it to come into contact with other conductors (including us!).

Note that the $1/d$ spatial dependence of this result means the force varies more slowly with distance than that between two point charges. In fact, the line of charge has an infinite charge and so the real question is why we get a finite answer for the force. This is due to a cancellation effect. Charges far from $x = 0$ contribute very weakly to the net result not only because they are farther away (the fundamental $1/r^2$ dependence for point charges), but also because they contribute very weakly to the net perpendicular component because the angle θ is so close to 90°.

FIGURE 14.7 *Dielectric breakdown of air around a Van de Graaf generator.*

4. ELECTRIC FIELDS

Coulomb's law is an example of a long-range force, one in which the interacting objects need not be in contact. Such forces involve *action at a distance*, as opposed to contact forces. (Actually all contact forces really involve action at a distance because, as was discussed in connection with friction, they are all due to electromagnetic forces; although very close together, these "contacts" actually involve distances that are large compared to atomic dimensions.)

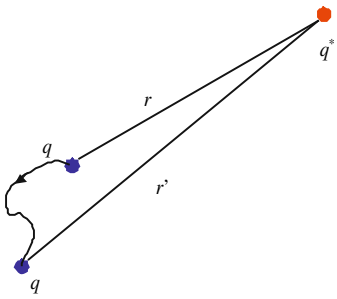


FIGURE 14.8 *How does charge q* learn that charge q has moved?*

A natural question to ask when long-range forces are at work in Coulomb's law is exactly how each charge learns about locations and values of other charges in order to experience a force. For example, given a point charge q^* that experiences a force due to another charge q a distance r away (Figure 14.8), suppose charge q moves to a larger distance r'. How will charge q^* learn of the change? Will q^* immediately experience a decrease in the electric force acting on it and a change in its direction?

Einstein's special theory of relativity (Chapter 24) tells us that no information signal can travel faster than the speed of light $c = 3 \times 10^8$ m/s (186,000 mi/s or 670 million miles per hour). Given this fact of nature, which is universally accepted in science, charge q^* will not learn of changes in the other charge's position until some finite time later, no matter how brief. The information actually propagates outward from charge q at the speed of light in the form of an *electric field*, defined below. Thus, the act of a static point charge q exerting a force on another static point charge q^* actually is a two-step process: first, q continually produces an electric field that travels outward at the speed of light; and second, q^* experiences a force by direct interaction with the electric field arriving at its location. Clearly the process is reciprocal, with q^* also producing an electric field that interacts with q directly.

As long as both charges are held at rest the situation is completely reciprocal with each charge interacting with the static electric field produced by the other charge. However, if one of the charges, say q, at time t rapidly moves to a new position (e.g., as in Figure 14.8), getting farther from charge q^*, it will immediately experience a smaller force in a different direction through interaction with the ever-present (not changing with time) local static electric field due to q^*, which is weaker farther from q^* and which is radially directed from q^*. On the other hand, q^* will not experience a decreased force until some time later when the information (field) travels at the speed of light from q the separation distance r' between the two charges (taking a delay time $\Delta t = r'/c$). The introduction of the electric field in the case of static charges may seem arbitrary and unnecessary, however, the electric field is a real physical quantity that can carry energy, momentum, and angular momentum.

By using the notion of a test point charge, taken by convention to be positive, we can introduce the definition of the electric field $\vec{E}$ at some point in space as

$$\vec{E} = \frac{\vec{F}}{q^*} \tag{14.4}$$

where $\vec{F}$ is the force on the test charge q^*. The electric field at the site of q^* is independent of the magnitude of the test charge, depending only on the charges producing $\vec{E}$ and their location with respect to q^*. In fact, the electric field exists whether or not there is a charge q^* at that location. From Equation (14.4) we see that units for E are N/C. The test charge is taken to have a vanishingly small electric charge so that it does not produce significant forces on those other charges that are producing the electric field. Although a real test charge may actually be used to probe the electric field, more often it is only a hypothetical construct used in the definition of the electric field. A real charge used in place of q^* would measure the same electric field only if it had a charge small enough so that no distortion of the source charges producing the electric field occurred.

The electric field of a point charge q at a distance r away may be found from Coulomb's law and the definition of $\vec{E}$ to be

$$\vec{E} = \frac{1}{q^*}\left(\frac{qq^*}{4\pi\varepsilon_0 r^2}\,\hat{r}\right) = \frac{q}{4\pi\varepsilon_0 r^2}\,\hat{r}, \tag{14.5}$$

where $\hat{r}$ is a unit vector along the outward radial direction from q. The choice of direction agrees with our previous definition in Equation (14.1) and ensures that if a positive test charge is placed at this position it will experience a repulsive or attractive force directed along $\hat{r}$ depending on whether q is positive or negative,

respectively. Note that the electric field is radially symmetric (has the same magnitude at any point on the surface of a sphere of radius r centered at charge q) as expected, because there is no preferred direction in space (Figure 14.9).

To find the net electric field produced by more than one point charge, we use the principle of superposition for vectors to simply add up the vector contributions

$$\vec{E}_{\text{net}} = \sum \vec{E}_{i,} \tag{14.6}$$

where $\vec{E}_i$ is the electric field at the observation point due to the ith point charge. An example helps to reinforce this idea.

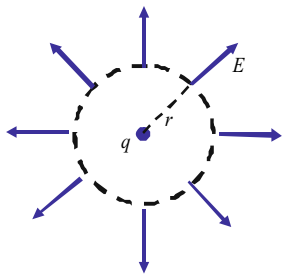


FIGURE 14.9 *The electric field of a point charge is spherically symmetric.*

Example 14.3 Let's calculate the electric field due to a pair of equal and opposite point charges at a point along the perpendicular bisector of the line joining the charges.

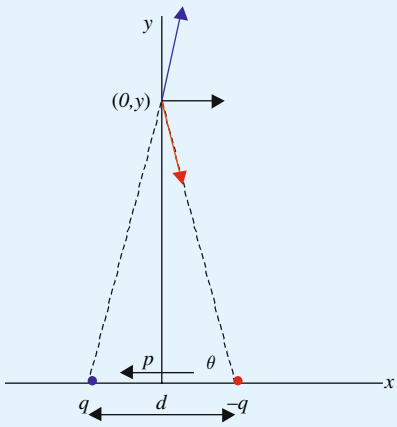


FIGURE 14.10 *Geometry for Example 14.3.*

Solution: We choose to place the two charges symmetrically along the x-axis a distance d apart and then to calculate the electric field at an arbitrary point along the y-axis as shown.

The magnitude of the electric field from each charge is the same and equal to

$$E = \frac{q}{4\pi\varepsilon_0[y^2 + d/2)^2]}\frac{1}{},$$

with the color-coded directions shown in the figure. From symmetry it is seen that the y-components cancel and the x-components add to give the resultant electric field (shown in black). The net electric field is then equal to the net x-component given by

$$E_{\text{net}} = 2E\cos\theta = 2E\frac{(d/2)}{[y^2 + (d/2)^2]^{1/2}} = \frac{q\,d}{4\pi\varepsilon_0}\frac{1}{[y^2 + (d/2)^2]^{3/2}},$$

where we have used the large triangle in Figure 14.10 to obtain an expression for $\cos\theta$.

(*Continued*)

Often we are interested in the case when the distance y is much larger than the charge separation d. In this limit, the equal and opposite charges are known as an electric dipole, and we can neglect the term $(d/2)$ compared to y in the denominator to find that

$$E_{\text{dipole}} = \frac{qd}{4\pi\varepsilon_0 y^3} \quad \text{or}$$

$$\vec{E}_{\text{dipole}} = \frac{-\vec{p}}{4\pi\varepsilon_0 y^3}, \quad \text{(along dipole perpendicular bisector)},$$

where $p = qd$ is defined as the electric dipole moment, with its direction taken as from – to + charge, along the $-x$ direction in this example. An electric dipole is then a pair of equal and opposite charges with very small separation distance compared to the distance to the observation point. Note that the electric field of the dipole decreases faster $(1/r^3)$ than that of a point charge $(1/r^2)$, as might be expected because of the partial cancellation effect of having opposite charges.

Example 14.4 Repeat the previous calculation, finding the electric field along the x-axis (Figure 14.11).

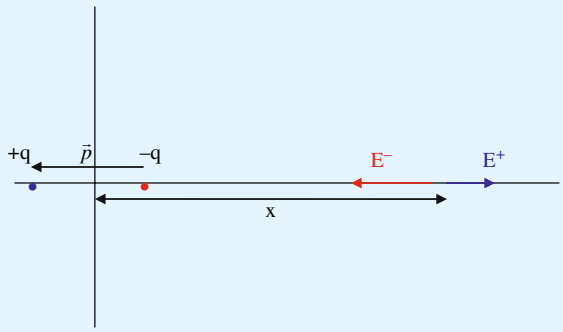


FIGURE 14.11 *Charges and field of Example 14.4.*

Solution: In this one-dimensional case we only have fields along the x-axis. The net result is along the x-axis and given by

$$E = \frac{1}{4\pi\varepsilon_0}\frac{q}{[x + (d/2)]^2} - \frac{1}{4\pi\varepsilon_0}\frac{q}{[x - (d/2)]^2}.$$

At this point if we look at the situation when $x >> d$, the dipole limit, then if we simply let $d = 0$ in the above expression, we find $E = 0$. Clearly E does go to zero, but we are interested in how it approaches zero and so we need to do some more mathematical manipulations. By factoring out the x^2 terms in both denominators, we can rewrite this expression as

$$E = \frac{q}{4\pi\varepsilon_0 x^2}[(1 + d/2x)^{-2} - (1 - d/2x)^{-2}].$$

In the dipole approximation with $d<<x$, we can expand each of the terms in the bracket using the binomial theorem: $(1 \mp \varepsilon)^{-n} = 1 \pm n\varepsilon \ldots$, valid when $\varepsilon << 1$, so that we have, to a good approximation (with $\varepsilon = d/2x$),

$$E_{\text{dipole}} = \frac{q}{4\pi\varepsilon_0 x^2}[(1 - d/x) - (1 + d/x)] = \frac{-1}{2\pi\varepsilon_0}\frac{qd}{x^3} \quad \text{or}$$

$$\vec{E}_{\text{dipole}} = \frac{1}{2\pi\varepsilon_0}\frac{\vec{p}}{x^3}. \quad \text{(along dipole axis)}$$

Note that in this case the electric field points along the dipole axis. We find the same $(1/x^3)$ spatial dependence here as in the previous example. In fact, the electric field due to an electric dipole varies as $(1/r^3)$ everywhere, as long as the dipole approximation $d<<r$ is true. As already mentioned, this more rapid decrease with x (or, in general, with r) than for a point charge is due to the near cancellation of electric fields by the two equal and opposite charges.

In order to find the net electric field produced by a continuous distribution of electric charge, the charged object is divided into small elements, each of which resembles a point charge. In place of a discrete summation of electric fields, as in Equation (14.6), a continuous summation, via calculus, must be done. As an example, we work out the electric field above an infinite plane of uniformly distributed electric charge. The surprising result of the boxed calculation is an important conclusion that is referred to again in the next chapter. The electric field above a uniform plane of charge, with charge per unit area $\sigma = Q/A$, is a constant, $\sigma/2\varepsilon_0$, directed perpendicular to the plane no matter how far above the plane. Thus, the plane of charge produces a constant electric field everywhere. Table 14.1 lists some formulas for the electric fields of several symmetric charge configurations.

Table 14.1 *Electric Fields of Various Geometries*

Geometry	*Parameters*	*E*
Point charge	Q	$\frac{1}{4\pi\varepsilon_0}\frac{Q}{r^2}$
Line charge (infinite)	$\lambda = Q/L$ r = perp. distance from line	$\frac{1}{2\pi\varepsilon_0}\frac{\lambda}{r}$
plane (infinite)	$\sigma = Q/A$	$\frac{\sigma}{2\varepsilon_0}$
sphere	*total* Q r = *distance from center with* $r >$ *sphere radius*	$\frac{1}{4\pi\varepsilon_0}\frac{Q}{r^2}$

Next, we discuss a method to view a mapping of the electric field in space. A topographical map, showing the elevations above sea level, is an example of a two-dimensional *scalar field.* At any point $\{x,y\}$ on the map a scalar, the elevation, is assigned. We could use a function $h(x,y)$ to describe this scalar field, where for each $\{x,y\}$ the function $h(x,y)$ assigns a height (Figure 14.13). An example of a three-dimensional scalar field might be a mapping of the temperature within a room. In this case a scalar is assigned to each point $\{x,y,z\}$ whose value might also be a function of time, perhaps varying differently at each point, so that a more complex function $T(x,y,z,t)$ might be used to map this scalar temperature field.

Here we calculate the electric field due to an infinite plane of charge. Consider the x–y plane to have a uniform positive charge per unit area $\sigma = Q/A$ and let's calculate the electric field along the z-axis at a distance d above the plane. We divide the plane into concentric rings of radius r and thickness dr (Figure 14.12). All the charge in each ring is the same distance from the point at which we calculate the electric field. The ring with radius r contains a charge

$$\sigma A = \sigma(2\pi r)dr.$$

By symmetry, the electric field due to the ring is along the z-axis; the x- and y-components cancel. The contribution of the ring to the vertical electric field at our observation point is

$$dE = \frac{1}{4\pi\varepsilon_0}\frac{2\pi\sigma r}{[r^2 + d^2]}\cos\theta\, dr.$$

Substituting

$$\cos\theta = \frac{d}{\sqrt{(r^2 + d^2)}}$$

and integrating over all r values, the total E field is given by

$$E = \frac{\sigma}{2\varepsilon_0}\int_0^\infty \frac{rd}{[r^2 + d^2]^{3/2}}\, dr.$$

The integral can be performed directly resulting in

$$E = \frac{\sigma}{2\varepsilon_0}.$$

This is a very surprising result, showing that the electric field is constant and independent of the height d above the plane.

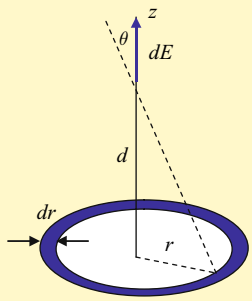


FIGURE 14.12 *Geometry for the calculation of the electric field of an infinite plane.*

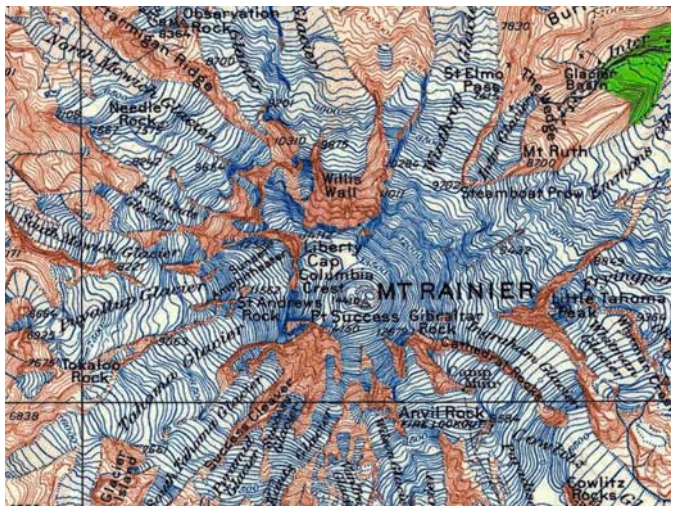


FIGURE 14.13 *A topographical* (topo) *map of Mount Rainier in the state of Washington.*

The electric field is an example of a *vector field.* At each point $\{x,y,z\}$ a vector $\vec{E}$ is assigned whose value may also depend on time, $\vec{E}(x, y, z, t)$. In the case of static charges, there will be no time-dependence and to each spatial point a constant vector is assigned. How can we pictorially represent a vector field in a way similar to that used for a scalar field, as in Figure 14.13? We have already used a mapping of a vector field when we discussed the steady flow of a fluid and used the notion of streamlines to map the velocity vector field. There, as here, we needed to represent not only the magnitudes of the vectors but also their directions. A representation known as *electric field lines* (streamlines in the context of fluid flow) can be used in which contours are drawn that are everywhere tangent to the vector directions. To convey information on the magnitudes of vectors, the density of lines drawn is made proportional to the local magnitude of the vectors in that region. Regions where electric field lines are dense correspond to strong electric fields, whereas regions devoid of lines of force correspond to weak or absent electric fields. For a point charge, electric field lines are therefore radial lines drawn outward from a positive charge and inward toward a negative charge. Electric field lines must always start and end on electric charges, the origins of the electric field. Two-dimensional maps for a few point charge distributions are shown in Figure 14.14.

Calculations of the electric field from a continuous distribution usually require more sophisticated mathematics, as in the boxed example above. In certain cases with sufficient symmetry, however, useful information about the electric field can be obtained from a symmetry argument. For example, for a long wire with a static positive uniform charge distributed along it (see the boxed example in Section 2 above and Figure 14.15), symmetry dictates that the electric field far from the ends of the wire must radiate outward from the wire as shown by the electric field lines. There can be no component of the electric field along the wire direction because there is no reason why the field would point one way or the other along the wire. We say that symmetry dictates that the field must lie in a plane transverse to the wire. Furthermore, in that plane there is also no preferred direction (we say there is azimuthal symmetry about the wire axis) so that the electric field can only depend on the perpendicular distance from the wire $r_\perp$ and not on the orientation around the wire. The only other parameter that the field can depend on is the linear charge density $\lambda = (Q/L)$, and not the charge Q, which is infinite for an infinite wire. Simply by noting the dimensions of E (given by $Q/\varepsilon_0 L^2$) and λ, one could surmise that the electric field magnitude must be proportional to $\lambda/\varepsilon_0 r_\perp$, in agreement with the boxed calculation in Section 2 above apart from constants (there we found $F = \lambda q/2\pi\varepsilon_0 d$ so that

$$E = F/q = \frac{\lambda}{2\pi\varepsilon_0 d}$$

where $d = r_\perp$). Symmetry arguments are powerful tools when the situation allows their application.

FIGURE 14.14 *Electric field mappings for* (left) *an electric dipole, or a pair of equal and opposite charges, and* (right) *three equally spaced co-linear charges of −4, +2, and +2 units from left to right.*

Thus far in our discussion of electric fields we have dealt with point charges and briefly with continuous charge distributions. We conclude this section with a discussion of the effect of a conducting metal object, charged or uncharged, on the nearby electric field; the case of insulating objects is taken up in the next chapter. We do not try to be rigorous, but rather try to motivate and explain general phenomena using specific examples.

Suppose first that an isolated solid metal object (a good electrical conductor) is given an excess electric charge. How does the excess charge distribute itself on the conductor? Will it spread uniformly throughout its volume? Uniformly over its surface? Or will it distribute itself in some more complex way? Remembering that a conductor has mobile free electrons, the excess free electrons will experience long-range repulsive forces and very rapidly move to reduce their interaction. To this end, they move to the surface of the conductor where they cannot escape; it can be shown that within the volume of a solid conductor there are no excess free electrons: there is zero net charge within a solid conductor. After reaching this *electrostatic equilibrium*, the distribution of charge on the surface is such that the electric field within the conductor is exactly zero. We can prove that this must be true by contradiction: if the electric field inside were not zero, free electrons would experience a net force and move, contradicting our assumption of equilibrium. These are general results: *the electric field and net charge inside any conductor after reaching electrostatic equilibrium are zero*.

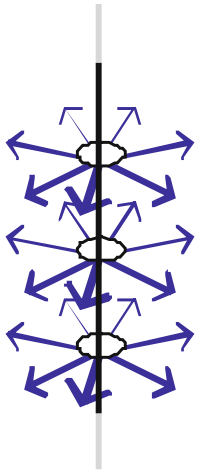


FIGURE 14.15 *By symmetry, the electric field from a long charged wire must be radially directed and depend only on the distance from the wire (away from the ends of the wire).*

If the object is both isolated and has sufficient symmetry (sphere, cylinder, large plane surface, etc.), then one can argue that any excess charge must be uniformly distributed over its surface. In general, *the electric field just outside the conducting surface must be perpendicular to the surface*. We again argue this last statement by contradiction: if there were a component of $\vec{E}$ parallel to the surface it would result in a net force on the surface charges along that direction parallel to the surface and therefore the assumed equilibrium could not exist. The outward force perpendicular to the surface is balanced by the attractive binding forces holding the charge on the surface, so that the charges remain in equilibrium. Any net charge on a conductor rapidly distributes itself so that the field inside is zero and the field outside is perpendicular to the surface (Figure 14.16). When the object has no symmetry, it turns out that the charge and external electric field tend to be greater where the curvature is greatest, that is, where the object has the smallest radius of curvature.

Suppose that an uncharged conductor is not isolated but lies in an external electric field produced by other charges with which we are not concerned. What can we say about the interaction of the field with the neutral conductor and about the conductor's effects on the external electric field? By the same arguments just made, at electrostatic equilibrium the electric field outside the conductor must be perpendicular to its surface and the field inside must be zero. But how has the electric field due to the external charges been modified by the presence of the uncharged conductor so as to result in zero electric field inside the metal? Even an uncharged conductor has many free electrons that can respond to the force produced by the external electric field. Rapidly these electrons will distribute themselves until they experience no net force; in doing so, they create an electric field just opposite to the external field within the volume of the conductor (Figure 14.17). At that point electrostatic equilibrium is reached and the particular stable arrangement of surface charges is just appropriate to cancel the electric field inside the conductor from the external charges. The electric field outside the conductor is modified by the presence of the conductor to assure that the field lines end on the conducting surface perpendicularly.

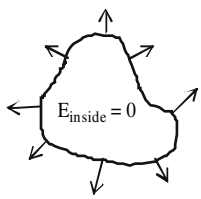


FIGURE 14.16 *A metal object with a net charge that distributes itself on the surface producing an external* E *field perpendicular to the surface but having zero internal* E *field.*

These properties of electrical conductors allow them to *electrically shield* their insides from any external electric fields. Electrical cables used for electronics applications are often made with braided metal sheaths that are used as electrical shields, protecting the internal signals from any undue influence from stray external electric fields. Indeed, the metal chassis (or case) around the major "chips" in computers and other electronic equipment is designed to do this same job.

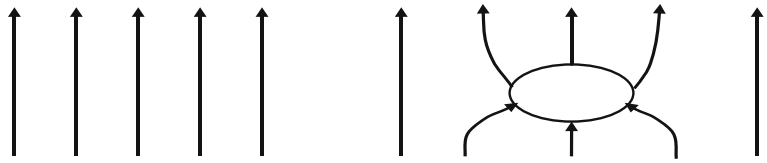


FIGURE 14.17 (left) *Original uniform external electric field in space.* (right) *Distortion of external field by an uncharged metal object so that the* E *field lines end perpendicularly on the metal. Induced charges on the metal surface cancel the electric field inside the metal.*

5. PRINCIPLES OF ELECTROPHORESIS; MACROMOLECULAR CHARGES IN SOLUTION

Electrophoresis is the forced migration of charged particles, usually macromolecules, in an electric field (Figure 14.18). If a macromolecule has a net charge q and a constant, uniform external electric field $\vec{E}$ is applied, there will be a net force $\vec{F}$ on the molecule given by $\vec{F} = q\vec{E}$. In general, the macromolecule will quickly accelerate and the electric force will be balanced very rapidly by a growing frictional force $-f\vec{v}$ due to collisions with solvent molecules. After reaching equilibrium, the molecule will migrate in the electric field with a constant velocity, obtained from setting the net force $q\vec{E} - f\vec{v}$ equal to zero and solving for the velocity,

$$\vec{v} = \frac{q\vec{E}}{f} \tag{14.7}$$

The electrophoretic mobility **U** is defined as the velocity normalized by the applied electric field, and using Equation (14.7) can be written as

$$\mathbf{U} = \frac{v}{E} = \frac{q}{f}. \tag{14.8}$$

Electrophoretic mobility is an intrinsic property of the macromolecule, depending only on its charge and frictional properties.

For a real macromolecule in solution, both the actual net charge q and the frictional factor f will be difficult to ascertain. If electrophoretic mobility were to be measured, one of these parameters would still need to be obtained independently before the other could be found from the above equation. This fact, and difficulties in generating a known uniform field locally at the site of the macromolecule, have made electrophoresis complex and little used as an analytical tool to learn about the electrical properties of macromolecules. However, there are a number of electrophoresis methods in use in most biomolecular laboratories. Before considering some of these techniques in a bit of detail in the next section, we need to gain a basic understanding of the charge on a macromolecule in solution.

Unlike isolated ions, such as Na^+ or Cl^-, that have a definite charge state, macromolecules have a variable net charge that depends on the pH of their local environment. Macromolecules such as proteins or nucleic acids, consist of many subunits, each with multiple ionizable charged groups that may be neutral, positive, or negative, depending on the pH. The term *zwitterion* or *polyelectrolyte* is used to describe such macromolecules with numerous charged groups (Figure 14.19). By adjusting the pH, the net charge on a macromolecule can thus be made positive, negative, or neutral. That particular pH at which the macromolecule is electrically neutral (having a net charge equal to zero) is called the *isoelectric point.* At pH values below the isoelectric point the macromolecule has a net positive charge whereas at a higher pH its net charge is negative.

Macromolecules are rarely suspended in pure water. Almost always they are found with salts, buffers, and often with many other small and large molecules. The

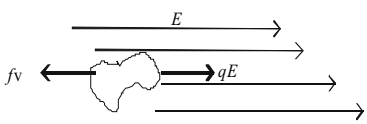


FIGURE 14.18 *Electric and viscous drag forces acting on a macromolecule.*

concentration of ions gives some measure of their effectiveness in electrical shielding; a better measure, however, is the *ionic strength* I, defined as

$$I = \frac{1}{2}\Sigma c_i z_i^2, \qquad (14.9)$$

where the sum is over all ionic species of concentration c_i and valence z_i.

It is important to realize that although the Coulomb force is long-range, as we have discussed, normally macromolecules in solution will be effectively electrically shielded unless at very low ionic strengths (Figure 14.20). Because of the electrical attraction of opposite charges, a charged macromolecule in solution will have large numbers of small ions of opposite charge, called *counterions*, surrounding each of its charged groups. These counterions form a charge cloud that tends to completely cancel the effects of the macromolecular charge beyond a certain characteristic distance, known as the *screening* (or *Debye*) *length.* A calculation of the screening length L_D finds

$$L_D = \left(\frac{\varepsilon_0 \kappa k_B T}{2e^2}\right)^{1/2} I^{-\frac{1}{2}}, \qquad (14.10)$$

where κ is a (dielectric) constant characteristic of the electrical properties of the solvent (water has $\kappa = 80$), k_B is Boltzmann's constant, and T is the absolute temperature. Table 14.2 gives the screening lengths for different concentrations of ions in water. Effectively, at ion concentrations above about 10–100 mM, the macromolecular charges are fully screened and there are no electrical interactions with other large molecules until they come within about 1 nm. At lower ion concentrations there may be longer-range electric interactions between macromolecules.

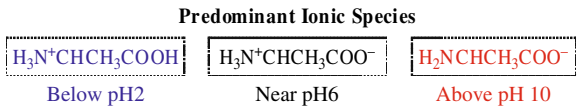


FIGURE 14.19 *The three different ionic forms of the amino acid alanine. Proteins, made from hundreds of amino acids, will have large numbers of variable electric charges, depending on the pH of the surroundings.*

Table 14.2 *Screening Lengths at Different Ionic Strengths of Solution*

Concentration (mM)	*Screening Length (nm) for Monovalent Ions*	*Screening Length (nm) for Divalent Ions*
0.1	30.4	17.6
1.0	9.6	5.6
10	3.0	1.8
100	1.0	0.6

6. MODERN ELECTROPHORESIS METHODS

There is a fundamental problem in using electrophoresis as described in the previous section. In order to maintain a buffer and solvent system with a typical salt concentration of 0.1 M, close to physiological, even a very modest electric field will create substantial heating of the solution, resulting in convection currents that would completely distort the controlled migration of macromolecules. We study this heating phenomenon when we study electric currents, but it is ultimately due to the transformation of kinetic energy of the charge carriers (ions or free electrons) into internal energy of the medium through collisions and it is a similar effect, for example, to that resulting in the heat generated by a toaster. Early in the history of electrophoresis, the answer to the heating problem was to reduce the ionic strength of the solution; but then, as we have seen, long-range interactions are possible and in some cases the macromolecules may not be stable under those conditions. Today almost all electrophoresis is carried out not in solution, but in gels, to avoid overall convection problems due to heating or vibrational disturbances.

One of the most important electrophoresis techniques is *SDS gel electrophoresis*, used to measure molecular weights of proteins. Because the conformations of

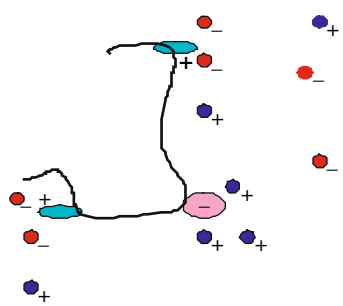


FIGURE 14.20 *A region of a macromolecule with its surrounding cloud of counterions. The concentration of counterions is usually many orders of magnitude greater than that of macromolecules.*

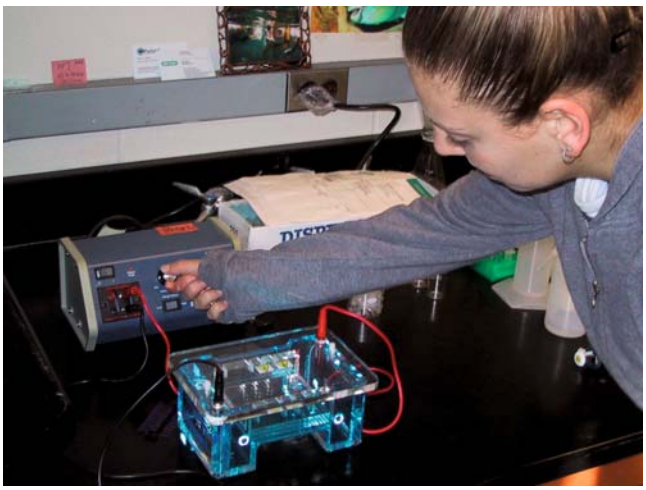


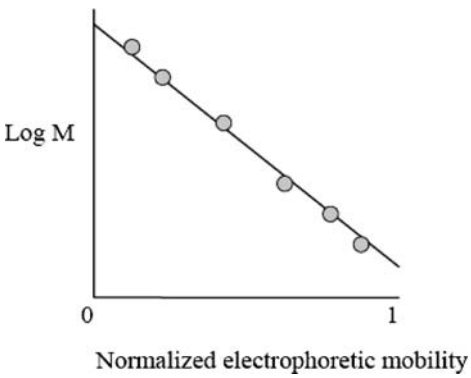
FIGURE 14.21 *Gel electrophoresis being set up to run. Plexiglass housing holds a slab gel connected to a power supply being adjusted.*

proteins are so diverse, substantial information about a protein would be required to know how the friction factor in Equation (14.8) is related to molecular weight. Instead, in this technique the proteins are first denatured so that they lose all of their secondary structure and become simply random coil backbone polymers. Then SDS (sodium dodecyl sulfate), a highly charged reagent that binds to all proteins with a very similar mass of SDS per unit length of protein backbone, and thus a very similar electric charge per unit length of protein, is added to saturate the protein. These highly charged SDS molecules exert strong internal repulsive forces that tend to stretch out the random coil protein into a rodlike shape. In essence, all proteins are made to look virtually the same: rods of the same diameter but with lengths that are proportional to the molecular weight of the protein.

The technique involves placing a small amount of such denatured SDS-protein mixture (with a colored dye or stain added so that one can see where the fastest migrating protein is located) at the top of a slab or tube of a gel (typically polyacrylamide), and turning on an electric field within the gel using electrodes attached to a power supply (Figure 14.21). The proteins and dye migrate down the gel at a constant rate that depends on the molecular weight of the protein with the smaller proteins migrating faster because they are less impeded by the gel. At a given concentration of gel material and given electric field strength, standards of known molecular weight are used to empirically construct a calibration curve of molecular weight versus electrophoretic mobility (basically determined from the distance traveled down the gel normalized between 0 and 1; see Figure 14.22). Molecular weights of unknown samples can be determined from their mobilities and such a calibration curve. Over a limited molecular weight range, the electrophoretic mobility of proteins is found to be proportional to the logarithm of their molecular weight, as shown in the figure. This technique, known as *SDS-PAGE* (polyacrylamide gel electrophoresis), can rapidly and cheaply measure molecular weights with an accuracy of about 5% and can also determine trace amounts of impurities in a sample. It is one of the most common tools in the study of proteins today.

Precisely how macromolecules move through the supporting gel material in gel electrophoresis is not well understood. Our description and the usual analysis of electrophoretic mobility are totally empirical. For very large macromolecules such as high molecular weight DNAs that tend to get stuck in the pores of even very dilute gels, it has been experimentally discovered that, by using a series of electric field pulses of short duration and varying direction, DNA migration can be enhanced. These efforts have led to an increased understanding of the migration of macromolecules in gels. Such knowledge is also applicable to the motion of macromolecules through networks of filamentous proteins within the cytoplasm of a cell and is leading to new insights on the dynamics of cells.

FIGURE 14.22 *Example of calibration plot for SDS-polyacrylamide gel electrophoresis.*



Another important gel electrophoresis method, using the ideas developed above on the polyelectrolyte nature of macromolecules, is *isoelectric focusing.* Native proteins migrate in an electric field through a gel in which a pH variation has been established. Proteins migrating in the gel will constantly vary their electric charge as the local pH changes until they arrive at the location corresponding to their isoelectric point (Figure 14.23). They remain there because, with their net charge equal to zero, they experience no force. The isoelectric point of a protein is an intrinsic property, therefore a detailed map of proteins separated according to isoelectric points can be obtained.

Often isoelectric focusing is combined with SDS-PAGE in two-dimensional gel electrophoresis. In this case, the native proteins are first run in a pH gradient gel slab along one direction. When completed, the electric field is set at 90° to its initial direction, a new gel slab saturated with SDS and denaturants is butted

against the original gel slab, and the proteins are made to migrate into the new gel. There they denature, acquire an SDS coat, and migrate according to molecular weight along the new direction. When complete there is a two-dimensional map of proteins with isoelectric point along one direction and, by calibration, molecular weight obtainable from the position along the other direction (Figure 14.24). Proteins with similar isoelectric points or with similar molecular weights can be further separated by this method as long as the other property is distinct. There are numerous other variations of these techniques in one or two dimensions in current use with new methods constantly being developed.

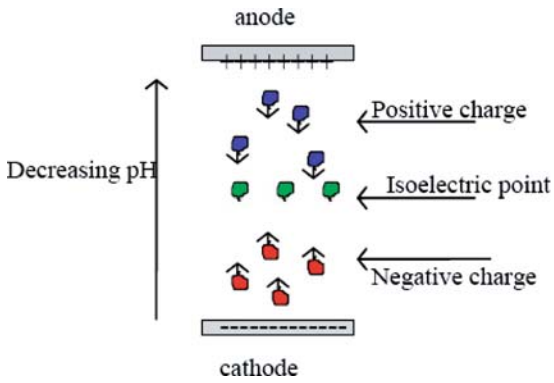


FIGURE 14.23 *Schematic of isoelectric focusing. Polyelectrolytes move until they reach their isoelectric point and have zero net charge. Remember at a pH below (above) the isoelectric point they are positively (negatively) charged.*

7. *Gauss's Law

This section is optional. Subsequent material does not depend on this section. Starred questions and problems at the end of the chapter refer to this optional section.

We've seen how electric charges create electric fields and in Section 4 of this chapter we saw, at least in principle, how to calculate the electric field from charge distributions based on the field produced by a point charge and the superposition principle. In this section we learn one of the fundamental principles of electricity, Gauss's law, which connects the average electric field on a closed surface (one that has an inside and an outside, or said differently, one that encloses some volume) to the net charge contained within that surface.

The easiest way to picture Gauss's law is to use the mapping of electric field lines. Suppose that there is no charge contained with some closed surface. Then any field lines that enter the surface must also leave the surface and no new lines can originate from within the surface, since there are no charges on which the field lines can end or begin. Any net charge contained within the closed surface can serve as endpoints for electric field lines, with positive charges generating new lines and negative charges ending field lines. Thus the net number of field lines crossing a closed surface is related to the enclosed net charge. To quantitatively discuss Gauss's law, we need to introduce the notion of the flux of a vector field.

Suppose there is a uniform electric field in a region of space, as shown in Figure 14.25. If a plane surface (shown here as an open surface, not surrounding any volume) lies with its normal making an angle θ with respect to the electric field lines, we define the electric flux Φ_E through the surface to be

$$\Phi_E = E_\perp A = EA_\perp = EA\cos\theta \tag{14.11}$$

where $E_\perp = E\cos\theta$ and $A_\perp = A\cos\theta$. Thus, in words, the electric flux is a measure of the number of electric field lines that cross the surface. Picture the field lines as arrows shot at a bulls-eye target. If the target directly faces the oncoming arrows

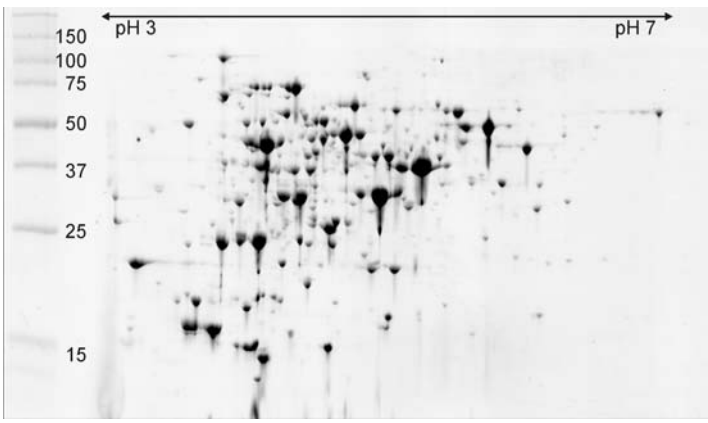


FIGURE 14.24 *Two-dimensional gel electrophoresis of the cytoplasmic proteins of a bacterium. The vertical scale is molecular weight (in kDa) and the horizontal scale is isoelectric point.*

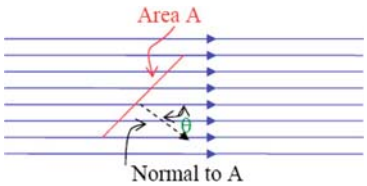


FIGURE 14.25 *The flux of a uniform* E *field.*

(so that $\theta = 0$) then the flux of arrows will be maximum. On the other hand, as the tilt angle θ increases towards 90°, less and less area presents itself as a target and the flux decreases toward zero (in the limit as θ goes to zero and the target is thin).

Now that we have a working definition of electric flux, we can state

Gauss's law, which relates the electric flux over a closed surface to the net charge contained within that surface as

$$\Phi_E = \frac{Q_{net,\ enclosed}}{\varepsilon_0}. \qquad (14.12)$$

Gauss's law is very generally true and it is one of the four basic relations of electromagnetism, known as Maxwell's equations,

discussed in Section 4 of Chapter 18. The calculation of the electric flux is very difficult in most cases, but can be greatly simplified if there is sufficient symmetry. In those cases, Gauss's law enables you to calculate the electric field produced by the enclosed charges. We look at several examples of the power of this law just below. Even in the absence of such simplifying symmetry, Gauss's law remains true and, with advanced mathematics, serves as the basis for solving many problems in electromagnetism.

Example 14.5 Calculate the electric field produced by a point charge q, using Gauss's law.

Solution: We put the point charge at the center of an (imaginary) spherical surface of radius r, called a Gaussian surface, shown in Figure 14.26. The surface is not actually present in the problem; we choose its shape and size and evaluate the electric flux over the surface in order to find an expression for the electric field at its surface, a distance r from the point charge. Because the single point charge of this problem suggests spherical symmetry, we picked a spherical surface.

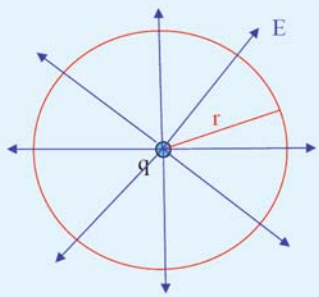


FIGURE 14.26 *Charge* q *surrounded by an imaginary Gaussian spherical surface of radius* r.

Because we know the electric field of a point charge points in the radial direction and depends only on the distance r from the point charge, the electric field will lie along the normal to the spherical surface and will be constant in magnitude on its surface, so that the electric flux can be written as $\Phi = EA$. Substituting this into Equation (14.12) and writing that the surface area of a sphere is $A = 4\ \pi r^2$, we find on solving for E, and writing it as a vector, that

$$\vec{E} = \frac{q}{4\pi\varepsilon_0 r^2}\,\hat{r},$$

in agreement with Equation (14.5).

Example 14.6 Calculate the electric field produced by a long thin wire with a uniform positive charge per unit length λ along the wire.

Solution: From the symmetry of the problem we know that the electric field will point radially away from the wire and will depend only on the distance r from the wire and not on the angle around the wire. To take advantage of this symmetry, we choose a Gaussian surface with the same symmetry, a cylinder centered on the wire with a radius r and some length L, as shown in Figure 14.27. We first need to evaluate the electric flux. But because E is constant on the cylindrical wall of our Gaussian cylinder, the contribution to the flux from the cylinder walls is just $\Phi = EA$, where A is the area of the cylinder wall, $A = 2\pi rL$. The circular end-caps on the rest of the closed cylindrical surface do not contribute to the flux because the E field is radially directed and the normals to the two end-caps are each perpendicular to this (think of an arrow shot at the end-caps: if they are oriented as shown the arrows cannot strike these targets.)

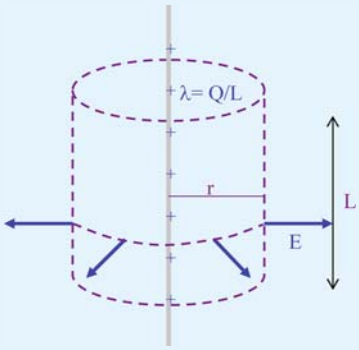


FIGURE 14.27 *An* (infinite) *line charge with a Gaussian cylinder setup for calculating Gauss's law.*

Therefore we have from Gauss's law that

$$\Phi_E = EA = E(2\pi rL) = \frac{Q_{\text{enclosed}}}{\varepsilon_0} = \frac{\lambda L}{\varepsilon_0}$$

or solving for E, we find that

$$E = \frac{\lambda}{2\pi\varepsilon_0 r},$$

where E points radially. This agrees with the formula quoted in Table 14.1 above.

Example 14.7 Find the electric field between two parallel plane metal plates with equal and opposite charges Q on them, as shown in Figure 14.28. This configuration is known as a parallel-plate capacitor.

Solution: From the planar symmetry, we expect the electric field to lie along the direction perpendicular to the planar surfaces, unless we get near the edges of the plates where the symmetry breaks down. We use a small Gaussian cylinder with cross-sectional area A oriented along the E field lines and with one end located within one of the metal plates. To calculate the electric flux we consider the three different portions of the cylindrical surface separately. The end-cap within the metal plate sees no electric field, because as we learned earlier in this chapter the electric field within metals is always zero in electrostatics, and therefore has no flux contribution. The cylindrical wall also contributes nothing to the electric flux because its normal is perpendicular to the electric field direction. The only contribution to the flux comes from the end-cap on the right with area A and with an electric field E pointing along its normal. Therefore the total electric flux is

$$\Phi_E = EA.$$

(*Continued*)

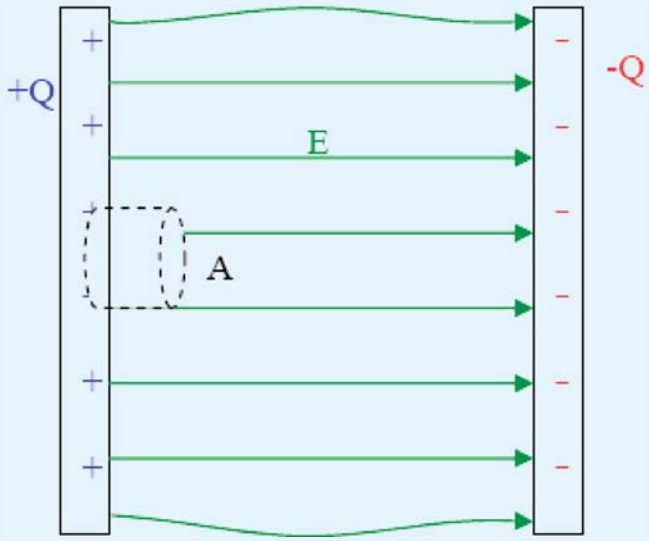


FIGURE 14.28 *Two parallel, flat metal plates (charged equal and opposite) shown with a small Gaussian cylinder in place to use Gauss's law to find* E *between the plates.*

Gauss's law says that this flux is proportional to the total charge enclosed within the Gaussian surface; this charge is equal to the surface charge density, σ (the charge per unit surface area on the plates) times the area A of the Gaussian cylinder end-cap. Then we have that

$$\Phi_E = EA = \frac{\sigma A}{\varepsilon_0},$$

so that we have

$$E = \frac{\sigma}{\varepsilon_0}.$$

This is a somewhat surprising result (but in agreement with the formula in Table 14.1), since it says that the E field is a constant between the plates and does not depend on where between the plates you look. On first glance you might expect the E field to depend on the distance from the plates in some way. This was further discussed in connection with the E field from an infinite plane of charge in Section 4.

CHAPTER SUMMARY

Electric charge can be either positive or negative, but comes in individual units, or quanta, in multiples of the charge on the electron or proton, with magnitude $e = 1.6 \times 10^{-19}$ C. In an isolated system, the total electric charge is conserved and remains constant in time.

The fundamental force law between two point electric charges, q_1 and q_2, separated by a distance r is given by Coulomb's law

$$\vec{F}_{1 \text{ on } 2} = \frac{1}{4\pi\varepsilon_0}\frac{q_1 q_2}{r^2}\hat{r}, \qquad (14.2)$$

where the unit vector lies along the line joining the two charges and is directed from 1 to 2 and the permittivity is $\varepsilon_0 = 8.85 \times 10^{-12}$ C^2/N-m^2.

Materials can be categorized by their electrical properties into conductors, such as metals, that have

"free electrons" able to move in response to electric forces, insulators (or dielectrics), such as wood or rubber, that do not conduct electricity under normal conditions and semiconductors, such as (doped) silicon, that allow for controlled conductivity in modern electronics.

The electric field is defined as the force per unit positive test charge q^*

$$\vec{E} = \frac{\vec{F}}{q^*}. \qquad (14.4)$$

Electric forces are an example, along with gravity, of "action at a distance," where electric charges experience electric forces without "contact." One way to explain this is to use the field concept, whereby all electric charges emit electric fields that travel at the speed of light and interact with other charges to produce electric forces. A point charge q produces an electric field at a distance r given by

$$\vec{E} = \frac{q}{4\pi\varepsilon_0 r^2}\hat{r}. \qquad (14.5)$$

Table 14.1 gives the electric field produced by a variety of other charge distributions.

Because conductors respond extremely rapidly to external fields, at electrostatic equilibrium the electric field inside any conductor vanishes, there can be no net charge inside any conductor, and the electric field just outside any conducting surface must point perpendicular to the surface.

Electrophoresis is a broad category of experimental methods involving forced migration of electrically charged macromolecules in electric fields. Modern methods use SDS polyacrylamide gel electrophoresis (PAGE) and isoelectric focusing to gain information on the molecular weight and electric charge properties, respectively, of the macromolecules. In the former method, the electrophoretic mobility, given by

$$\mathbf{U} = \frac{v}{E}, \qquad (14.8)$$

is related to the molecular weight of the migrating macromolecule, whereas in the latter method macromolecules are brought to an equilibrium location within a pH gradient where their net charge vanishes.

Gauss's law is one of the four fundamental Maxwell equations and relates the electric flux over a closed surface to the enclosed charge,

$$\Phi_E = \frac{Q_{\text{net, enclosed}}}{\varepsilon_0}, \qquad (14.12)$$

where the flux is defined as

$$\Phi_E = E_\perp A = EA_\perp = EA\cos(\theta). \qquad (14.11)$$

QUESTIONS

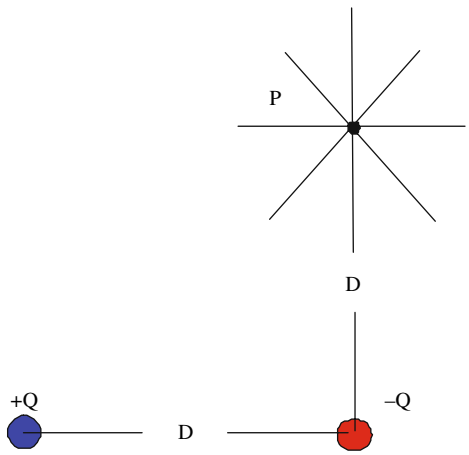
1. A system has a total net charge of $+15e$. If 20 protons and 5 electrons are removed what is the system charge?
2. A nucleus with 81 protons and 127 neutrons is observed to emit a beta particle (high-speed electron). How many protons and neutrons are left in the nucleus?
3. Two equal charges are a fixed distance apart. If a third charge of the same sign is placed at the midpoint of the line joining the two charges, is it in equilibrium? What happens if it is slightly displaced to one side along the line? What if it is slightly displaced off the line? Repeat these questions if the third charge is of opposite sign to the other two.
4. What would happen to the force between two point charges if
   (a) The distance between them was doubled?
   (b) The charge of one of them was halved?
   (c) The sign of both was changed?
   (d) The sign of one was changed?
   (e) The distance between them was doubled and the charge of one was halved?
5. Distinguish between the net charge on a conductor and its total number of free electrons.
6. Why do you expect good electrical conductors also to be good thermal conductors?
7. Two isolated charges are 1 m apart. If one of the charges "instantaneously" moves to a nearby location, how long will it take for the other charge to discover this?
8. What is the direction of the force on a positive point charge q close to a large plane sheet of positive charge? Does your answer depend on how far the charge is from the plane?
9. Why is it that you can sometimes generate static charge "shocks" when going to touch metal (such as a doorknob) when the humidity is low but not when it is high?
10. Give some examples of scalar fields; of vector fields.
11. Does the electric field of a spherical ball of charge exactly equal that of a point charge with the same total charge located at the center of the sphere? What about inside the spherical ball?

12. Why does the test charge, used in defining the electric field, need to be infinitely small in magnitude?

13. At a point in space there is an electric field, due to external charges, with magnitude E and pointing in the positive x-direction. A small charge having a magnitude of 1 μC experiences a force of 1 μN in the negative x-direction. Circle the letters of all of the following that is/are true.

(a) The charge must be positive.
(b) The charge must be negative.
(c) The mass of the charge must be 1 kg.
(d) The strength of E must be 1 N/C.
(e) A charge having a magnitude of 2 μC placed at the same point would experience a force of 1 μN because E due to the external charges doesn't change.
(f) A charge having a magnitude of 2 μC placed at the same point would experience a force of 2 μN because E due to the external charges changes to $2E$.
(g) A charge having a magnitude of 2 μC placed at the same point would experience a force of 2 μN because E due to the external charges doesn't change.

14. The figure shows two charges separated by a distance D. Point P is D to the right of the positive charge and D up from the negative charge. Draw an arrow with its tail at P and whose head points in the correct direction of the electric field due to $+Q$ and $-Q$. P is just a point; there's no charge there. Use the lines emanating from P as a guide. You can place your field vector along any one of the eight lines through P or somewhere between.



15. Consider three identical charges at the corners of an equilateral triangle. What is the electric field at the center? In the case of four identical charges at the corners of a square, what is E at the center? Can you generalize this for the E field at the center of an N-sided equilateral polygon with N identical charges at the corners?

16. Three parallel infinite lines of charge with the same linear charge density are located at the corners of a square. Find the direction of the electric field at the fourth corner.

17. A hollow charged conducting sphere of radius R and charge Q is centered at the origin. There is a positive point charge of charge q located at the origin as well as an infinite line of charge (with linear charge density λ) parallel to the x-axis at $y = 2R$. What is the electric field at the point $x = z = 0$, $y = R/2$? (Hint: First consider which charges produce an electric field at the observation point.)

18. Explain the apparent paradox that a charge inside a closed uncharged metal container produces an electric field outside the container that can interact with other external charges, but these charges do not produce an electric field inside the container (electrical shielding).

19. Why is ionic strength a better parameter than concentration to use for describing electrical properties of solutions?

20. Explain what the screening length means. In particular, why does it decrease as the ionic strength is increased?

21. Explain how SDS-PAGE is able to separate macromolecules based on molecular weight.

22. Why is it that SDS-PAGE can be performed on any macromolecule with the same electrode arrangement of negative electrode at the top (or start) and positive electrode at the bottom (or end) of the gel, regardless of the sign of the intrinsic charge of the macromolecule?

23. *Explain electric flux in words, discussing its dependence on all three variables in Equation (14.11).

24. *Why does the electric flux only depend on the total electric charge contained inside the Gaussian surface? In answering this think about the electric field lines produced by charges outside the surface versus inside the surface.

25. *Discuss the surprising result that the electric field produced between a pair of flat parallel metal plates with equal and opposite electric charge on them does not depend on location between the plates.

MULTIPLE CHOICE QUESTIONS

1. A proton is initially at rest at $x = -d$ and an electron is initially at rest at $x = +d$. At the same instant they are released. They subsequently (a) fly away from each other, (b) collide at $x = 0$, (c) collide close to $x = -d$, (d) collide close to $x = +d$.

2. Two equal positive charges are held in place, 2 cm apart. Where should a positive test charge be placed so that the test charge oscillates back and forth? (a) On the perpendicular bisector of the line connecting the first two charges. (b) On the line connecting the first two charges and between them. (c) On the line connecting the first two charges but not between them. (d) It is not possible to make such an oscillator.

3. The electric field at the midline between two infinite line charges with linear charge densities of λ and $-\lambda$ separated by a distance $2d$ is given by

(a) $\dfrac{\lambda}{2\pi\varepsilon_0 (2d)}$, (b) $\dfrac{\lambda}{2\pi\varepsilon_0 d}$, (c) $\dfrac{\lambda}{\pi\varepsilon_0 d}$, (d) 0.

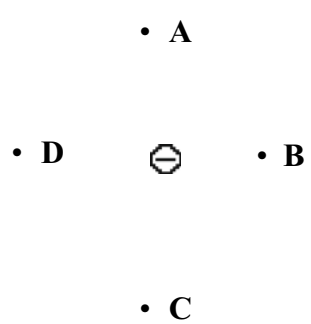
4. With three equal point charges Q at the corners of a square with sides of length L, the magnitude of the electric field at the fourth corner will be (a) equal to three times the electric field of the point charge Q, or $\dfrac{3Q}{4\pi\varepsilon_0 L^2}$, (b) less than three times the electric field of a single point charge but more than twice that value, so between $\dfrac{2Q}{4\pi\varepsilon_0 L^2}$ and $\dfrac{3Q}{4\pi\varepsilon_0 L^2}$, (c) less than $\dfrac{2Q}{4\pi\varepsilon_0 L^2}$, (d) more than $\dfrac{3Q}{4\pi\varepsilon_0 L^2}$.

5. The table below represents electric field values measured at different distances from some source. Which one of the following is most likely to be the source? (a) a sphere of charge, (b) a line of charge, (c) a dipole, (d) a sheet of charge.

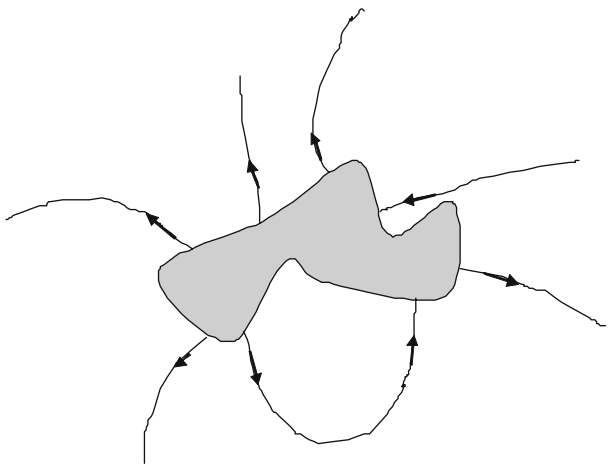
Distance (cm)	*Field strength (V/m)*
4	12.84
8	1.69
12	0.52

6. A solid sphere of copper has a small spherical bubble at its center. At first, the copper is electrically neutral. Then, at one instant the surface of the bubble is coated with some added charge (such as some extra electrons). After a few minutes (a) the added charge will still all be on the bubble's surface, (b) half of the added charge will be on the bubble's surface and half on the sphere's outer surface, (c) the added charge will all be on the sphere's outer surface, (d) the added charge will be uniformly distributed throughout the material between the bubble and the outer surface.

7. In the figure below a uniform external field points left to right. A negative spherical charge is placed in this field as shown. At which point is the total field (external plus sphere's) most likely to be zero? (a) A, (b) B, (c) C, (d) D.



8. Suppose that a picture of electric field lines is drawn following the convention that 2 field lines emerge from a small sphere with +2 pC of charge. In this picture there is an irregular closed surface, the interior of which is hidden, as shown to the right. The net amount of charge inside the closed surface must be (a) +8 pC, (b) +6 pC, (c) +4 pC, (d) −2 pC.



9. Which of the following is not a scalar field? (a) a mapping of the temperature of the human body, (b) a mapping of the topography of New York State, (c) a mapping of the water velocity in a stream, (d) a mapping of the mass distribution in our galaxy.

10. Which of the following is an incorrect symmetry argument about the external electric field of a charged spherical conductor? (a) It must point radially because at any observation point there is always a symmetric distribution of charge to cancel any components of E transverse to the radial direction; (b) it can only depend on the distance from the sphere center r because the sphere is uniform and an arbitrary rotation of the sphere cannot change the result, so the answer must be independent of the angles in spherical coordinates and can only depend on r; (c) it must decrease as $1/r^2$ because that is the spatial dependence of the electric field of a point charge; (d) it is proportional to the net charge on the conductor because the charge is distributed uniformly on the sphere.

11. Which of the following statements about a conductor is false. (a) The electric field inside is always zero; (b) just outside, the electrostatic field is perpendicular to its surface; (c) at equilibrium the net charge inside the conductor is zero; (d) a charge located within a hole in a conductor at equilibrium feels no force from charges outside the conductor.

12. A lump of copper is placed in a uniform external electric field E that points left to right. When the charges in the copper come into equilibrium the induced electric field inside the lump (a) is larger than E and points left to right, (b) is smaller than E and points

left to right, (c) is the same size as E and points right to left, (d) is zero, (e) none of the above.

13. The electrophoretic mobility of a macromolecule in a uniform electric field depends on all but which of the following? (a) The pH of the solution, (b) the isoelectric point of the macromolecule, (c) the electric field applied, (d) the frictional properties of the macromolecule.

14. SDS polyacrylamide gel electrophoresis can be used to measure the (a) electrophoretic mobility, (b) net charge, (c) isoelectric point, or (d) molecular weight of a macromolecule.

15. *In applying Gauss's law to a problem to find the electric field outside a thin spherical shell of radius R with electric charge distributed over its surface, the best choice for a Gaussian surface would be (a) a long cylinder of radius R, (b) a spherical shell of radius R, (c) a spherical shell of radius $r > R$, (d) a spherical shell of radius $r < R$.

16. *When using Gauss's law to solve for the electric field between two long concentric cylinders, of radii R_1 and $R_2 > R_1$, with equal and opposite electric charge on them, the appropriate Gaussian surface would be (a) a cylinder of radius r, such that $r > R_2$; (b) a sphere of radius r with $R_1 < r < R_2$; (c) a cylinder of radius r with $r < R_1$; (d) a cylinder of radius r with $R_1 < r < R_2$.

17. *In Example 14.7, when calculating the flux through the Gaussian surface why is the flux through the cylinder wall equal to zero? Is it because (a) the electric field is perpendicular to the normal to the cylinder, (b) the effective area of the cylinder wall is zero because the average direction of the normal to the surface cancels out, (c) the electric field is zero on that surface, (d) the angle θ between the electric field and the normal to the cylinder wall is zero?

PROBLEMS

1. How many electrons make up 1 C of electric charge? What is the mass of these electrons?

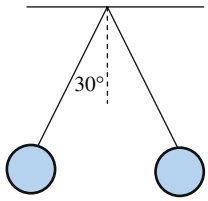
2. Estimate the number of electrons in the Earth. The Earth's mass is 6.0×10^{24} kg. Assume that for each electron there is one proton and on average one neutron.

3. How close must two protons be if the electric force between them is equal to the weight of either at the Earth's surface?

4. An electron ($m = 9.11 \times 10^{-31}$ kg) is suspended at rest in a uniform electric field of magnitude E. Take into account gravity at the Earth's surface, and determine the magnitude and direction of the electric field.

5. In a simple model of the hydrogen atom, the electron revolves in a circular orbit around the proton at a distance of 0.53×10^{-10} m. What is the speed of the electron in orbit?

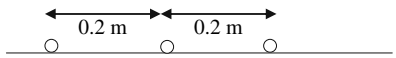
6. Consider an arrangement of two point charges $+Q$ and $-Q$ each of which has a mass m, placed on either end of a massless rod of length D. Suppose that the rod is fixed to a horizontal surface by a nail through its center and that the apparatus is subjected to a uniform electric field E parallel to the plane of the surface and perpendicular to the rod. What is the net torque on the system of rod and charges about the pivot point?



7. A large electroscope is made with "leaves" that are 50 cm long wires with 20 g spheres at the ends. When charged, nearly all the charge resides on the spheres. If the wires each make a 30° angle with the vertical as shown on the right, what total charge Q must have been applied to the electroscope and what is the tension in the wire? Ignore the mass of the wires.

8. Suppose that electrical attraction, rather than gravity, were responsible for holding the moon in orbit around the Earth. If equal and opposite charges Q were placed on the Earth and the moon, what value of Q would be needed so that the moon would stay in its present orbit? Potentially useful data: mass of Earth $= 5.98 \times 10^{24}$ kg, mass of Moon $= 7.35 \times 10^{22}$ kg, radius of orbit $= 3.84 \times 10^{8}$ m.

9. Three equal 2 μC charges are equally spaced 0.2 m apart along a line as shown. Find the net force on each of the charges.



10. Find the force on a 5 μC point charge located at a vertex on an equilateral triangle of 0.5 m sides if 10 μC point charges are located at the other two vertices.

11. Six equal charges are at the corners of a hexagon. What is the force on a seventh equal charge at the center of the hexagon? Suppose one of the six charges is removed. Find the force on the charge at the center. (Hint: As is often the case there is a hard way and an easier way to solve this. For the easier method, use superposition ideas to remove the sixth charge by adding an equal and opposite charge at its site.)

12. Equal and opposite 5 μC point charges are located at the points $y = \pm 0.5$ mm (+5 μC at $y = +0.5$ mm and −5 μC at $y = -0.5$ mm). Find the force acting on a 2 μC point charge when it is located at each of the following sites: (a) ($x = 1$ mm, $y = 0$); (b) ($x = 0$, $y = 1$ mm); (c) ($x = 0$, $y = -1$ mm).

13. According to the boxed calculation in the chapter, the force on a point charge a distance d from an infinite line of charge with charge per unit length λ is

$$F = \frac{\lambda q}{2\pi\varepsilon_0 d}.$$

Find the force on a 2 μC charge located 2 m from a line charge with a linear charge density of 0.2 μC/m. Compare this to the situation when the same 2 μC charge is 2 m away from a second point charge and find the value of the second charge that would give the same net force.

14. Find the electric field at the center of a square produced by four equal 2 μC charges located at its corners. Also for the same situation, find the electric field at the center if two of the neighboring charges at the corners are −2 μC and the other two charges are 2 μC.

15. A sphere of radius R contains a uniform distribution of total charge Q. What is the force on a point charge q located a distance $2R$ from its center?

16. Find the electric field at the midpoint between a pair of equal and opposite 5 μC charges separated by 3 m. Which way does it point?

17. Find the electric field at the center of an equilateral triangle of 0.5 m sides with charges at the corners of 5 μC, −10 μC, and −10 μC.

18. Two point charges, 5 μC and −8 μC are 1.2 m apart. Where should a third charge, equal to 5 μC, be placed to make the electric field at the midpoint between the first two charges equal to zero?

19. Three parallel infinite line charges with equal charge densities of 2 μC/m lie in a plane and are equally spaced by 0.5 m. Find the electric field along a line perpendicular to their plane through the middle line charge a distance of 2 m away.

20. Compare the electric field produced 10 cm away from either a 10 μC point charge, from a 10 m long line of charge with the same 10 μC total charge, and from a 10 m × 10 m plane with the same charge distributed uniformly. Assume the equations for an infinite line or plane apply.

21. A biological membrane can often be modeled as two closely spaced parallel planes with equal and opposite surface charge densities. We study this in detail in later chapters, but for now calculate the electric field within the membrane assuming the charge density on either plate is ± 0.1 μC/cm^2 and "vacuum" between the plates. We show later that this calculation is only about a factor of three too large when the vacuum is replaced by the lipid molecules actually present in the membrane. (Hint: See Table 14.1 and use superposition to find the net field from both planes.)

22. Certain fish are extremely sensitive to small electric fields, with sharks, rays, and eels able to detect electric fields as low as 7 μN/C. At a 1 m distance, what is the minimum charge these fish can detect (ignore charge screening)?

23. The electric field inside biological membranes is extremely high, roughly 1×10^7 N/m. If this electric field generated the only force on a sodium ion, what would its acceleration be?

24. What is the ionic strength and Debye screening length at room temperature (300 K) of the following aqueous solutions

(a) 0.15 M NaCl + 0.015 M $MgCl_2$
(b) 0.5 M $MgCl_2$ + 0.2 M KCl

25. A sphere with a 0.05 μC net charge on it undergoes electrophoresis in distilled water at 20°C due to a uniform 1 N/C electric field. If the sphere migrates at a speed of 1 cm/s find its radius. Reminder: the friction factor for a sphere is $f = 6\eta\pi R$.

26. *Given a spherical shell of radius R with total positive charge Q together with a positive charge q at its center, find the electric field both inside and outside the shell using Gauss's law.

27. *Two long concentric cylinders of radius R_1 and R_2, with $R_1 < R_2$, have equal and opposite charges per unit length, $\pm\lambda$, on them (with $+\lambda$ on the cylinder at R_1). Find the electric field in the following regions using Gauss's law: (a) $r < R_1$, (b) $R_1 < r < R_2$, and (c) $r > R_2$.

28. *Using Gauss's law find the electric field produced by a large planar sheet of electric charge with a charge per unit area equal to σ.

29. *Using the previous problem and the principle of superposition, find the electric field between two such planar sheets separated by distance d with equal and opposite charge densities, $\pm\sigma$. Check that your result agrees with Example 14.7.

30. *A spherical *conducting* shell of inner radius R_1 and outer radius R_2 has zero net charge. A point charge $+q$ lies at its center.

(a) Use Gauss's law for a Gaussian spherical surface of radius r, such that $R_1 < r < R_2$, to prove that there must be an induced charge of $-q$ on the inner metal surface. What is the charge density on this surface?
(b) What is then the charge on the outer surface of the conductor and what is the charge density on this surface?
(c) Use Gauss's law to find the electric field both inside and outside the conductor and show that you get the same result in the absence of the conductor. Note that any additional electric charges outside the conductor will not affect the electric field within the conductor. The region inside the conductor is said to be shielded from electric fields outside the conductor.

15 Electric Energy and Potential

In the last chapter we discussed the forces acting between electric charges. Electric fields were shown to be produced by all charges and electrical interactions between charges were shown to be mediated by these electric fields. As we've seen in our study of mechanics, conservation of energy principles can often be used to understand the interactions and dynamics of a system. In this chapter we introduce the concept of electric potential energy and electric potential, and apply these considerations to a variety of situations. The fundamental electric interactions in atomic, macroscopic, and macromolecular systems are each presented. Biological membranes are discussed in some detail, with emphasis on their ability to act as capacitors, energy storage devices. Membrane channels are introduced, focusing on sodium channels: how they work and how they are selective. We return to a more detailed description of the electrical properties of channels in the next chapter. This chapter concludes with a discussion of the mapping of the electric potential produced by various organs of the human body including muscles, heart, and brain (EMG, EKG, and EEG, respectively). These medical techniques are often used for diagnostic purposes.

1. ELECTRIC POTENTIAL ENERGY

The electric force is a conservative force. As we saw in Chapter 4, this means that the work done by the electric force in moving a particle (in this case, charged) between two points is independent of the path and depends only on the starting and ending locations. Furthermore, there is an electric potential energy function that we can write down, whose negative difference at those two locations is equal to the work done by the electrical forces

$$-(PE_{E,\text{final}} - PE_{E,\text{inital}}) = -\Delta PE_E = W. \tag{15.1}$$

Recall that two expressions we have used for potential energy functions in mechanics, gravitational (mgy) and spring potential energy ($\frac{1}{2}kx^2$), followed from the general definition of work and the particular form of the force. In a similar way, if Coulomb's law for the force due to a point charge q_1, on a second point charge q_2, separated by a distance r is substituted into the general definition of work (see the box below), one obtains the *electric potential energy* of the two point charges

$$PE_E(r) = \frac{q_1 q_2}{4\pi\varepsilon_0 r}. \tag{15.2}$$

J. Newman, *Physics of the Life Sciences*, DOI: 10.1007/978-0-387-77259-2_15,

Here we derive an expression for the electric potential energy between two point charges. We imagine that there is a point charge, say $q_1 > 0$, located at the origin and bring a second point charge, $q_2 > 0$, from infinitely far away where it does not feel any electric force to some distance r away from q_1. Because both charges are positive, there is a repulsive force between them and positive external work must be done to bring q_2 toward the charge at the origin. This work is equal and opposite to the (then, negative) work done on q_2 by the electric force from q_1. According to Equation (15.1) the change in potential energy will then be positive as might be expected, because if the external force is removed, the repulsive force will change the positive electric potential energy of q_2 into kinetic energy as it accelerates away from the origin.

From the general definition of work and Equation (15.1), the electric potential energy change is given by

$$\Delta PE = PE(r) - PE(\infty) = -\int_{\infty}^{r} [F\cos\theta]\mathrm{d}s$$

where F is the electric force on charge q_2 and θ is the angle between the force vector and the displacement vector $\mathrm{d}\vec{s}$. The path taken by the charge does not matter, therefore we choose it to be inward along the radial direction. In this case θ is equal to 180°, so that $\cos\theta$ is equal to -1, and the displacement $\mathrm{d}s$ is equal to $-\mathrm{d}r$. We substitute Coulomb's law for the force to find

$$\Delta PE = \frac{-q_1 q_2}{4\pi\varepsilon_o}\int_{\infty}^{r}\frac{1}{r^2}\mathrm{d}r.$$

Remembering that

$$\int\frac{1}{r^2}\mathrm{d}r = \frac{-1}{r},$$

we do the integration and evaluate the resulting expression at the limits to find that the potential energy at a distance r from the origin is given by Equation (15.2).

Note that, just as in the mechanical energy cases, we need to define the location of zero potential energy because only potential energy differences have meaning. For springs, the natural choice was to reference the spring potential energy to a zero value for an unstretched spring that exerts no force. For gravitational potential energy near the Earth's surface, we were free to define the location of zero potential as we chose because the gravitational force on a mass is constant in the approximation we used. For other more general situations using gravity, the zero of gravitational potential energy occurs when all masses are infinitely far apart so as not to be interacting. Similarly, in the case of electrical forces, when the charges are infinitely far apart ($r \to \infty$) they do not interact and it is therefore natural to choose this situation to correspond to zero electric potential energy. Equation (15.2) already satisfies this convention.

The electric potential energy for charges of like sign that repel one another is positive according to Equation (15.2), whereas for unlike charges that attract each other it is negative. Example plots for both cases are given in Figure 15.1. We recall that the negative of the slope of such a plot is equal to the force acting at position r. In this case, with one charge at the origin and the other at r, when $PE_E(r) > 0$ because the energy decreases with increasing r, the negative of the slope is always positive, consistent with a repulsive force acting. The steeper the curve is, the larger the force (and therefore acceleration) acting. We can imagine a charged particle sitting on the energy curve and falling down its hill with a decreasing acceleration (but still increasing velocity) as it moves toward larger r values. A charge projected toward the origin with some initial kinetic energy will travel up the PE_E hill as far as corresponds to the conversion of all its kinetic energy to potential before falling back down the energy hill.

Similarly, when $PE_E(r) < 0$, because increasing r leads to less negative, or increasing PE_E, the negative of the slope is itself negative, confirming that the force is attractive. A charged particle placed on this curve will also fall down the potential hill ever more rapidly (with increasing acceleration) as its distance from the other charge at the origin decreases. We discuss electric potential energies for other situations later in this chapter in connection with molecular bonding.

Having found an expression for the electric potential energy of a pair of point charges, we can write an expression for the total energy of this two-particle system. We include the kinetic energy of each particle, the electric potential energy, and any other mechanical potential energies, PE_{mech}, appropriate to the situation. The conservation of energy principle then states that

$$E = KE_1 + KE_2 + PE_{\text{mech}} + \frac{q_1 q_2}{4\pi\varepsilon_o r} = \text{constant}. \qquad (15.3)$$

As we have seen in applications in mechanics, energy conservation is a powerful concept that has a great degree of practical utility as well.

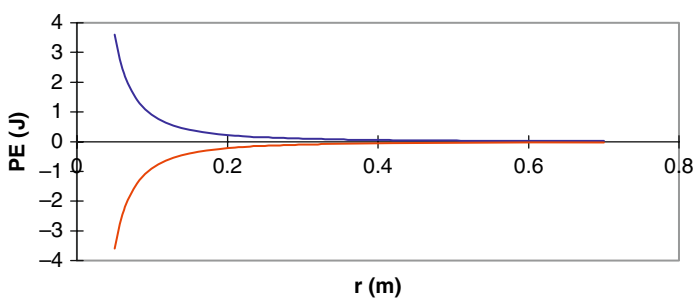


FIGURE 15.1 *Electric potential energy for two point charges of 1 μC magnitude with the upper curve for like sign charges and the lower curve for opposite charges.*

Example 15.1 Find an expression for the total energy of a hydrogen atom treating the electron as traveling in a circular orbit around the stationary proton. Find an answer in terms of only the radius of the circular orbit.

Solution: The total energy consists of the kinetic energy of the electron, traveling in a circle, and the electric potential energy of the electron–proton pair. We can write this as

$$E = \frac{1}{2}mv^2 + \frac{1}{4\pi\varepsilon_0}\frac{(+e)(-e)}{r}.$$

To express the velocity of the electron in terms of its orbital radius, we use the fact that the only force on the electron is the Coulomb force and this must supply the centripetal acceleration according to

$$F = \frac{1}{4\pi\varepsilon_o}\frac{e^2}{r^2} = m\frac{v^2}{r},$$

where both the force and centripetal acceleration are radially directed. Solving for mv^2 and substituting into the expression for the energy, we have

$$E = \frac{1}{2}\frac{1}{4\pi\varepsilon_o}\frac{e^2}{r} - \frac{1}{4\pi\varepsilon_o}\frac{e^2}{r} = -\frac{1}{8\pi\varepsilon_o}\frac{e^2}{r}.$$

This result says that the energy of a hydrogen atom is solely determined by the radius at which the electron orbits the proton. Note that the total energy is negative. This is the signature of a bound system, with the negative potential energy term dominating over the positive kinetic energy term. We show in Chapter 25 that although this is a correct statement, the electron cannot orbit the proton at any radius, but only at certain allowed radii. This fact of nature leads to a discrete set of allowed energy levels for the hydrogen atom from the above equation relating E to r, as first derived by Neils Bohr in 1913.

In our discussion, electric potential energy has been introduced as arising from a direct interaction between charges via the Coulomb force. However, as was discussed in the last chapter, charges experience electric forces by direct interaction with an electric field due to the other charges rather than by action at a distance interactions of charges. In the next section, we introduce the electric potential, an important concept that intrinsically accounts for electric fields.

2. ELECTRIC POTENTIAL

A charged particle q_o in an electric field $\vec{E}$ will experience a force equal to $q_o\vec{E}$. Associated with the interaction of the charge and the electric field is an electric potential energy. In the last section we saw the form of this potential energy if there is only one other point charge producing the electric field. In general, the electric potential energy will factor into a product of the charge q_o and a function that depends only on the other charges present and their distribution in space. This function therefore represents the electric potential energy per unit charge and is called the *electric potential* (or simply the *potential*), $V(r)$, where

$$V(r) = PE_E(r)/q_o. \tag{15.4}$$

Specifically, q_o is the charge located at the position at which the potential is being determined. The SI unit for electric potential is the volt, from Equation (15.4) given by

1 J/C = 1 volt (V). From our discussion you may correctly suspect that $V(r)$ is intimately related to the electric field produced by the other charges of the system; we show this connection shortly.

A very important unit for electric potential energy is the *electron volt* (eV), defined as the work done in moving an electronic charge through a potential difference of 1 V. From the charge on an electron, $e = 1.6 \times 10^{-19}$ C, we see that 1 eV = $(1.6 \times 10^{-19}\ \text{C}) \times (1\ \text{V}) = 1.6 \times 10^{-19}$ J. The electron volt is a very useful unit of energy in dealing with elementary particles such as electrons and protons since typical values are eV and awkward powers of 10^{-19} are not needed.

To find an equation for the electric potential produced by a single point charge at the origin we can use Equation (15.2) in which we arbitrarily assign q_2 to be the charge located at the origin, and q_1 to be a charge q_0 at an observation point a distance r away where we wish to evaluate V. Using Equation (15.4), V is found by dividing Equation (15.2) by the charge q_1 (= q_0). Because the label q_2 is arbitrary, we drop its subscript to find a general expression for the electric potential of a point charge located at the origin,

$$V(r) = \frac{q}{4\pi\varepsilon_0 r}. \tag{15.5}$$

The electric potential function of a point charge maps the potential energy per unit charge in space, so that if a charge q_0 were placed at position r the potential energy of the two-charge system would be $PE = q_0 V(r)$. Implicit in this is the zero-level of electric potential to be at infinite separation.

Note that the electric potential function of a point charge is defined everywhere in space and does not actually require another charge to interact with at a point in order to have a defined value at that point. Note the physical significance of *the electric potential at a point is the external work needed to move a unit positive charge from infinitely far away to that point along any path*. This is true because the change in electric potential energy equals the negative of the work done by the electric forces, which in turn is equal and opposite to the work done by external forces. So, for example, when you turn on your flashlight using two 1.5 V ($2 \times 1.5 = 3$ V total) batteries, each unit of charge (1 C) that moves through the light bulb from one side of the battery to the other has used 3 J worth of battery energy.

It may be helpful to discuss an analogy with gravitation in order to better appreciate the meaning of electric potential. If a gravitational potential function had been analogously defined as $PE_{grav}/m = gh$, we see that such a "gravitational potential" would correspond to the height function multiplied by the constant g. A roller coaster track would define this gravitational potential function by virtue of its height (Figure 15.2). An expression for the gravitational potential energy function of someone riding on the roller coaster could then be easily found by multiplying that function by her mass. We did not introduce such a gravitational potential previously because, in our constant g approximation near the Earth's surface, there would be no particular benefit. However, in the case of electricity with both positive and negative charges and with a spatially varying electric field, a mapping of the electric potential in space without regard for other interacting charges will be quite useful in the same way in which a mapping of the electric field was in the last chapter. Remember, however, that the electric potential is a scalar function, whereas the electric field is a vector quantity representing three functions, one for each vector component. A two-dimensional mapping of the scalar field representing the electric potential is similar to a topological map as discussed in the last chapter. In this case the height above a point in the plane represents the potential at that point. For the three-dimensional case, a scalar potential value is assigned to each point in space. These mappings can be visualized using color-coded computer methods, for example (see ahead to Figure 15.9). But, what is the relation between the electric field and the electric potential?

FIGURE 15.2 *Boomerang, Knott's Berry Farm, California: gravitational potential varies with height.*

To answer this question let's take the simple case of a constant, uniform electric field along the x-direction, reducing the problem to essentially one dimension. The force on a point charge q_o in such an electric field is $\vec{F} = q_o\vec{E}$ and the work done on q_o by the electric field in moving a distance Δx along the electric field direction is

$$W = F\Delta x = q_o E_x \Delta x.$$

Accordingly, the change in electric potential energy is $\Delta PE_E = -q_o E_x \Delta x$ so that the electric potential is given, in this simple case, by

$$\Delta V = \frac{\Delta PE_E}{q_o} = -E_x \Delta x \quad \text{(uniform } E\text{)}, \tag{15.6}$$

where Δx is positive when along the E field direction. This equation relates the constant electric field to the change in potential between two locations separated by Δx. If the potential function is known, then the electric field may be found from the relation

$$E_x = -\frac{\Delta V}{\Delta x}, \tag{15.7}$$

where, in more than one dimension, there are similar expressions for the y and z components of the electric field. We mention that in the two- or three-dimensional case, given a mapping of the potential, the direction of the electric field is along the direction of the steepest descent of the function; that is, at any given point the electric field will be along that direction corresponding to the most rapid decrease in potential.

It is also worth mentioning that Equation (15.7) shows that the electric field may be expressed in units of (V/m) in addition to the previously introduced equivalent units of (N/C), with 1 N/C = 1 V/m. The V/m is probably the more common unit for electric fields. Note that when Equation (15.7) is multiplied by a charge q_o its meaning becomes

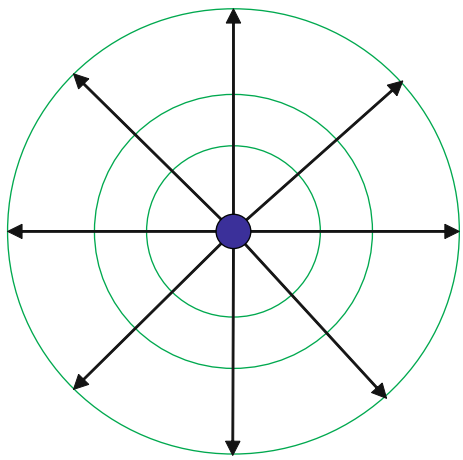
$$F_x = -\frac{\Delta PE_E}{\Delta x}, \tag{15.8}$$

recovering an equation we have seen previously (Equation (4.23)).

For a positive electric charge q_o, the positive work done by an electric field acting alone will tend to drive the charge toward lower electric potential. This is seen by the fact that the product of $W = F_x\Delta x = q_o E_x \Delta x = -q_o \Delta V > 0$, so that $\Delta V < 0$, and the charge will move down the potential hill. On the other hand, a negative charge will be attracted toward a higher potential because in that case with $q_o < 0$ we must have $\Delta V > 0$. Plots of electric potential have the same dependence on r as electric potential energy and are therefore quite similar to those in Figure 15.1. These statements concerning the directions of the forces acting on charges are generally true despite our assumption of a constant electric field. *Positive charges tend to move toward lower potentials, or down potential hills, whereas negative charges tend to move toward higher potentials, or up potential hills.*

Figure 15.3 shows a mapping of the electric field and electric potential of a point charge. Note that the potential is mapped as a series of, in this case spherical, contours of constant potential, known as *equipotential surfaces* (in three-dimensional space). No work is required to move a charge around on an equipotential surface because there is zero potential difference between all its points. Therefore, the electric field is always perpendicular to equipotential surfaces, as we saw in the previous chapter for the case of a conducting surface. This is true because if the electric field had a component parallel to an equipotential surface, there would then be a net force acting to do work on a charge moving on the surface and it could not have a constant potential. It is straightforward to map

FIGURE 15.3 *Radial E field vectors and spherical equipotential surfaces (circles in two dimensions) of a point charge.*



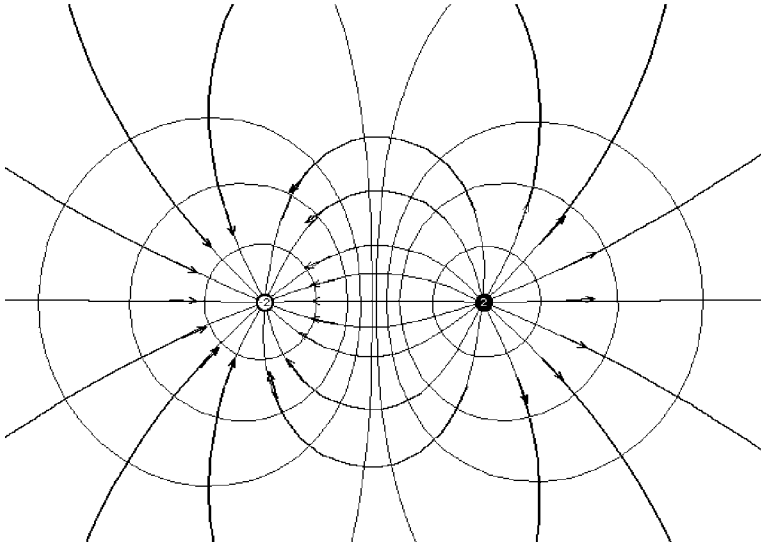


FIGURE 15.4 *Electric dipole field map with equipotentials. Note in this case the equipotential surfaces are not spheres, but are everywhere perpendicular to the electric field. Make sure you are clear on the difference between electric field lines and equipotentials. Which are which in the figure?*

equipotential surfaces once a mapping of the electric field is known. A surface is constructed that is everywhere perpendicular to the electric field lines (Figure 15.4).

An interesting example of an electrostatic potential in biology involves the honeybee. Coated with a fine layer of hair, the honeybee develops electrostatic charge when it flies, so that it actually can reach electrostatic potentials of several hundred volts. When the bee lands on a flower to drink nectar, pollen grains are electrostatically attracted to the fine hairs and will "jump" short distances through air from the electrostatic forces (see Figure 15.5). The honeybee then grooms itself and collects the adhered pollen in pollen sacs attached to its hind legs. Fortunately, not all of the pollen is collected for the bees to eat and the remaining pollen is able to pollinate other flowers as the bee visits them. It is also thought that the electrostatic voltage developed may help deliver pollen grains to the stigma of flowers by electrostatic attraction. (As an aside, for your information, recently there has been a precipitous decline in honeybee populations around the world. As yet the cause is unknown, although quite a number of factors have been surmised including virus infections, parasites, pesticide effects, nutritional issues, and other factors. Because honeybees pollinate about 90% of the fruit and vegetable crops in the United States alone, their declining numbers are having a major impact on the worldwide economy.)

FIGURE 15.5 *A honeybee with pollen grains adhering to its fine body hairs by electrostatic attraction.*

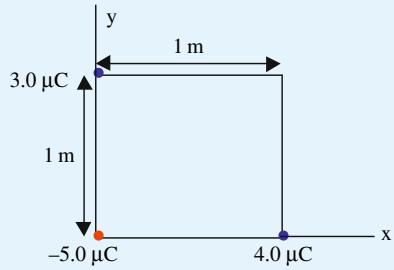
3. ELECTRIC DIPOLES AND CHARGE DISTRIBUTIONS

From the equation for the electric potential of a point charge (Equation (15.5)), we can find the electric potential of an arbitrary distribution of electric charge by generalization. If there are a number of individual point charges in the system (see Figure 15.6), the potential at some point in space, that we call the observation point, is simply the algebraic sum of the individual potentials due to each charge,

$$V = \frac{1}{4\pi\varepsilon_o}\Sigma\frac{q_i}{r_i}, \tag{15.9}$$

where r_i is the distance from the observation point to the ith charge, q_i. In this sum, one must be careful to include the sign of the electric charge. There is a clear advantage in calculating the net electric potential, a scalar quantity, over adding vector components of the electric field in order to find the net electric field. Because there is a direct connection between the two, it is almost always easier to find V first and then find $\vec{E}$ directly from V. A specific example helps to illustrate these ideas.

Example 15.2 Calculate the potential and the electric field at the empty corner of a square of 1 m sides when there are point charges at each of the other corners as shown.



Solution: We first calculate the electric potential at the empty corner of the square. Because potential is a scalar, we simply add the potential due to each charge, as in Equation (15.9), to find

$$V = \frac{1}{4\pi\varepsilon_o}\left[\frac{3\times10^{-6}}{1} + \frac{4\times10^{-6}}{1} - \frac{5\times10^{-6}}{\sqrt{2}}\right] = 3.1\times10^4\ \text{V}.$$

The factor $\sqrt{2}$ is the length of the diagonal of the square, the distance from the $-5\ \mu$C charge to the observation point. To find the electric field at the same point we must add the electric field vectors produced by each point charge at the observation point. This sum is given, in ordered pair notation, by

$$\vec{E} = \frac{1}{4\pi\varepsilon_\circ}\left[\left(\frac{3\times10^{-6}}{1^2},0\right) + \left(0,\frac{4\times10^{-6}}{1^2}\right) + \left(\frac{-5\times10^{-6}}{(1^2+1^2)}\cos 45, \frac{-5\times10^{-6}}{(1^2+1^2)}\sin 45\right)\right],$$

where the direction of the field from the $-5\ \mu$C charge is along the diagonal of the square toward the charge and we have taken its x- and y-components. Combining terms, the net electric field is

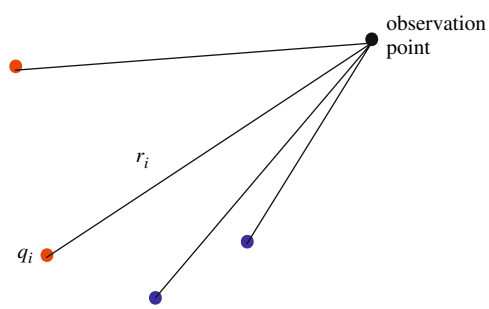
$$\vec{E} = (1.1\times10^4, 2.0\times10^4)(\text{V/m}).$$

In general, it is clearly easier to calculate scalar electric potentials than vector electric fields.

One particular arrangement of two charges that is of general significance is the *electric dipole* already studied in Examples 14.2 and 14.3. Its significance lies in the fact that even though it is electrically neutral, the separation of positive and negative charges allows it to produce an electric field and corresponding electric potential. Electric dipoles of two types occur in nature. A net separation of equal positive and negative charges may be permanent, as, for example in the important case of the water molecule (Figure 15.7). Even molecules that are electrically neutral and have no permanent dipole moment can, in the presence of an external electric field, form a dipole moment by a process known as electric polarization. The imposed electric field causes a separation of positive and negative charges in the otherwise neutral molecule leading to an induced dipole moment. This important process is discussed in more detail in the next section.

To calculate the electric potential of a dipole, we first specify a coordinate system and then use Equation (15.9) to add the individual

FIGURE 15.6 *Geometry to calculate potential from a distribution of point charges.*



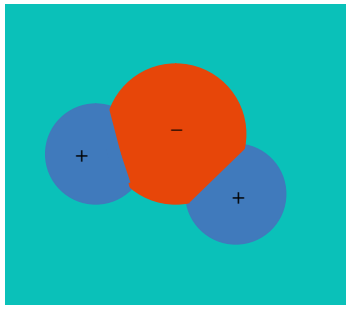


FIGURE 15.7 *Molecular structure of the water molecule. The red oxygen carries a partial negative charge, and the blue hydrogens each carry a partial positive charge so that there is a separation of the centers of positive and negative charge producing a permanent dipole moment for water.*

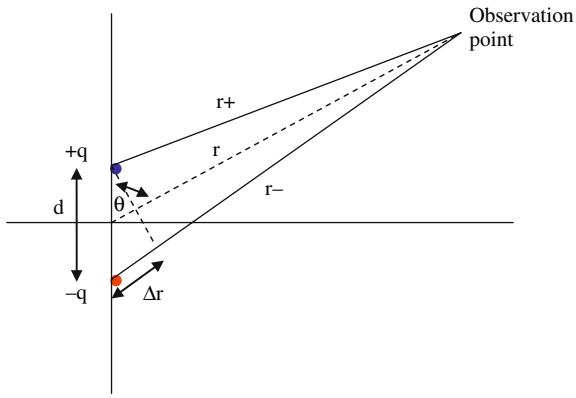


FIGURE 15.8 *Geometry for electric dipole calculation.*

potentials. If we choose the arrangement shown in Figure 15.8, we find the potential to be

$$V_{\text{dipole}} = \frac{1}{4\pi\varepsilon_o}\left[\frac{q}{r_+} - \frac{q}{r_-}\right], \tag{15.10}$$

where r_+ and r_- are the respective distances of the positive and negative charges to the observation point. If the observation point is much farther away than the size of the dipole d, so that with $r = r_- \sim r_+ + \Delta r$ as shown in Figure 15.8, then from the figure, we can write that

$$\left[\frac{1}{r_+} - \frac{1}{r_-}\right] = \frac{r_- - r_+}{r_+ r_-} = \frac{\Delta r}{r^2} = \frac{d\cos\theta}{r^2},$$

where θ is the angle between the vector $\vec{r}$ from the dipole center to the observation point and the dipole axis, chosen by a convention in which the axis points from negative to positive charge along the dipole. Substituting this into Equation (15.10) results in

$$V = \frac{qd\cos\theta}{4\pi\varepsilon_o r^2} = \frac{p\cos\theta}{4\pi\varepsilon_o r^2}, \tag{15.11}$$

where we have defined the *electric dipole moment* to be $p = qd$, equal to the magnitude of either charge times the charge separation distance.

The electric potential of a dipole differs from that of an isolated charge in two significant ways. First, the dipole potential decreases much faster with increasing distance, varying as $1/r^2$ whereas the potential of a point charge varies as $1/r$. This is to be expected because the net charge of the dipole is zero and the force on, or the interaction energy with, a charge at the observation point is expected to be substantially less than that due to a single charge q at the site of the dipole (see the example just below). Second, the dipole potential is no longer spherically symmetric, but has an angular dependence. This is also to be expected because the dipole has a symmetry axis defining a preferred direction in space.

Example 15.3 Calculate the electric potential and field of an electric dipole along its axis.

Solution: Using the notation of Figure 15.8 as applied to an observation point along the dipole axis, say the z-axis, we can write expressions for the electric potential and field of a dipole as

$$V = \frac{1}{4\pi\varepsilon_o}\left[\frac{q}{r_+} - \frac{q}{r_-}\right] = \frac{1}{4\pi\varepsilon_o}\left[\frac{q}{z-(d/2)} - \frac{q}{z+(d/2)}\right]$$
$$= \frac{q}{4\pi\varepsilon_o z}\left[\frac{1}{1-(d/2z)} - \frac{1}{1+(d/2z)}\right]$$

and

$$E = \frac{1}{4\pi\varepsilon_o}\left[\frac{q}{r_+^2} - \frac{q}{r_-^2}\right] = \frac{1}{4\pi\varepsilon_o}\left[\frac{q}{(z-d/2)^2} - \frac{q}{(z+d/2)^2}\right]$$
$$= \frac{q}{4\pi\varepsilon_o z^2}\left[\frac{1}{(1-d/2z)^2} - \frac{1}{(1+d/2z)^2}\right],$$

where $\vec{E}$ points along the z-axis. To proceed, we simplify the final term in the bracket of each expression using the binomial theorem when $x << 1$,

$$\frac{1}{(1 \pm x)^n} = 1 \mp nx \ldots,$$

to find

$$V = \frac{q}{4\pi\varepsilon_o z}\left[\{1+(d/2z)\} - \{1-(d/2z)\}\right] = \frac{qd}{4\pi\varepsilon_o z^2} = \frac{p}{4\pi\varepsilon_o z^2}$$

and

$$E = \frac{q}{4\pi\varepsilon_o z^2}\left[\{1+d/z\} - \{1-d/z\}\right] = \frac{qd}{2\pi\varepsilon_o z^3} = \frac{p}{2\pi\varepsilon_o z^3}.$$

We compare the z-dependence of these two expressions, per unit dipole moment, in Figure 15.9. Note the faster decrease in E with distance from the dipole, varying as $1/z^3$ versus the $1/z^2$ variation of V.

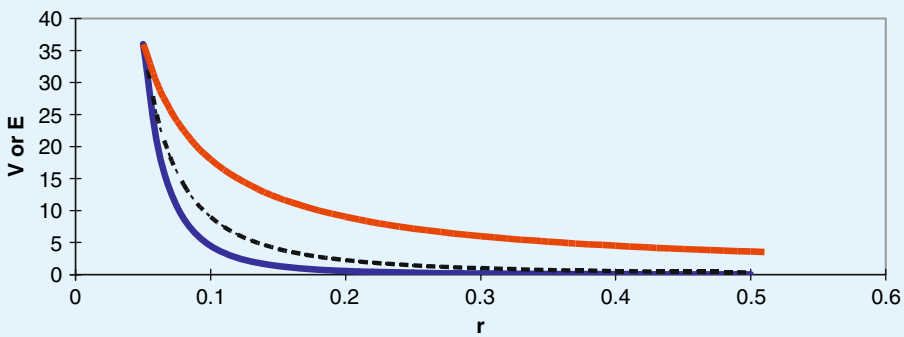


FIGURE 15.9 *Electric potential ($1/r^2$,* lower dashed line) *and field ($1/r^3$,* solid line in blue*) along the axis of a (unit) electric dipole. The plots have been normalized to coincide at the maximum value shown. Upper curve* (red) *has a 1/r dependence, for comparison.*

It is interesting to check that we can calculate the electric field for Example 15.3 directly from the expression for the electric potential using Equation (15.7). To find E_z we simply differentiate V with respect to z:

$$E_z = -\frac{dV}{dz} = -\left[-\frac{d}{dz}\left[\frac{p}{4\pi\varepsilon_o z^2}\right]\right]$$
$$= -\frac{(-2)p}{4\pi\varepsilon_o z^3} = \frac{p}{2\pi\varepsilon_o z^3},$$

in agreement with the separate and more difficult calculation in the example.

Continuous distributions of electric charge, in which the charge is found throughout a volume or on a surface, are obviously more common real-life examples of actual charge distributions than point charges. Most of these situations must be handled using numerical methods on a computer, but if there is sufficient symmetry in the geometry of an object on which the charge resides then analytical expressions for the potential can be obtained using calculus. One useful representation for the electric potential of a charge distribution is a potential map, very much like a topological map. An example is given in Figure 15.10 for a protein molecule. Such mappings are particularly useful for visualizing the potential in the neighborhood of a complex macromolecular surface that would be detected by a small ion or molecule.

4. ATOMIC AND MOLECULAR ELECTRICAL INTERACTIONS

Our current understanding of the electrical interactions between elementary constituents of matter comes from quantum mechanics, a subject we explore briefly toward the end of this book. One ultimate question in our fundamental understanding is why atoms are stable objects. Consisting of a positive nucleus and negative electrons that, according to Coulomb's law, should attract each other, they might be expected to be unstable and collapse. The negative potential energy curve of Figure 15.1 corresponds to this situation. An electron would be expected to "fall" down this potential energy hill to the nucleus at the origin. We show later how quantum mechanics addresses this fundamental question but for now we simply treat atoms as stable objects. As two atoms approach each other, once their electron clouds (for now, a vague term that indicates the rough size of an atom) overlap, there is a very strong repulsive force arising from quantum mechanical effects. This very strong repulsion is sometimes called a *hard-sphere repulsion* because it resembles the strong repulsive interaction between two billiard balls that prevents them from overlapping in space when they come into contact. As long as atoms do not overlap in space, we will have a reasonable degree of understanding of their electrical interactions by treating them as point charges and dipoles and ignoring quantum mechanics.

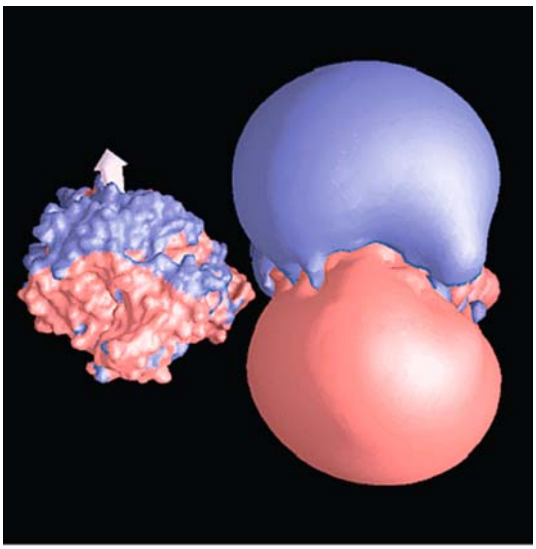


FIGURE 15.10 *The acetylcholine esterase molecule with two types of color coding. On the left, the surface is color-coded with positive* (blue) *and negative* (red) *charges (with the dipole moment shown as the white arrow), whereas on the right two equipotential surfaces are mapped, each corresponding to* k_BT *energy (GRASP modeling).*

Atomic distances are usually measured in angstroms (Å), where 1 Å = 0.1 nm. The smallest atom, hydrogen, has a diameter of about 1 Å, whereas the largest atoms are only several Å in diameter. If we calculate the magnitude of the electric potential energy due to the Coulomb interaction between two electrons, separated by a distance of 1 Å, we find from Equation (15.2),

$$PE = \frac{(1.6 \times 10^{-19})^2}{4\pi\varepsilon_o(10^{-10})} = 2.3 \times 10^{-19}\ J = 14\ \text{eV}.$$

(For comparison with bond strengths discussed in Section 5 of Chapter 12, this energy corresponds to

$$PE = \frac{2.3 \times 10^{-19}\ J/\text{bond} \cdot 6 \times 10^{23}\ \text{bonds/mol}}{4.18\ J/\text{cal}} = 33\ \text{kcal/mol},$$

about 4–5 times larger than the energy of the strongest atomic bonds that exist.)

We can classify the various types of electrical interactions possible between atoms or molecules. The strongest interactions are those due to direct *charge–charge interactions*, having a potential energy given by Equation (15.2), but with the permittivity of vacuum ε_o modified by the electrical properties of the medium in which the charges are immersed (discussed in the next section). With one charge at the origin, the potential energy of such interactions decreases with separation distance as $1/r$. Charge–charge interactions only occur between two

ionized atoms or molecules, both having net charge. Other types of electrical interactions are discussed in decreasing order of strength based on their dependence on separation distance.

The *charge–dipole interaction* occurs when one atom or molecule is charged and the other has a permanent dipole moment. According to Equation (15.4), the interaction potential energy should be given by the product of the charge and the dipole potential, given by Equation (15.11). In this case the potential energy decreases with separation distance as $1/r^2$ and is proportional to the product of the charge and dipole strength, also depending on the orientation of the dipole in space.

If both atoms or molecules have no net charge but are permanent dipoles then the *dipole–dipole interaction* occurs with an energy that varies as $1/r^3$ and depends on the two dipole strengths as well as their relative orientation in space. All of the above interactions can be either attractive or repulsive, resulting in potential energies that are either negative or positive, respectively.

When one of the atoms or molecules has both no net charge and no permanent dipole moment, it can still interact electrically with a charge on another atom or molecule. The charge creates an electric polarization (or separation of positive and negative centers of charge) of the neutral atom or molecule so that an induced dipole is formed (Figure 15.11). The *charge–induced dipole interaction* is always attractive because the induced dipole is always created with the opposite charge closest to the original isolated charge. The interaction dies away faster still with separation distance, varying as $1/r^4$.

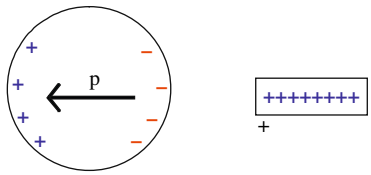


FIGURE 15.11 *A charged rod inducing a net dipole on a neutral sphere.*

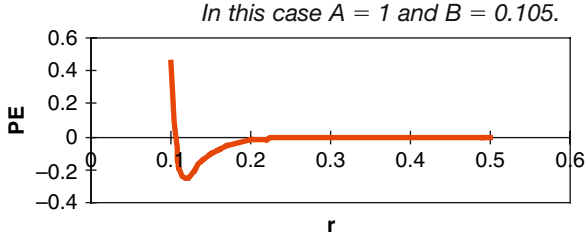
What is the situation when both atoms (or molecules) have neither a net charge nor a permanent dipole moment? Will they still interact electrically? All atoms are composed of a number of electrons and an equal number of protons in the nucleus. The time average of the electric dipole moment will be zero, because we have assumed no permanent dipole. However, over short time intervals there will be a nonzero rapidly varying dipole moment that can interact with a second neighboring neutral, nonpolar atom or molecule to induce a corresponding electric polarization and induced dipole moment. Known as the *dispersion interaction*, this interaction is always attractive, just as for the case for the charge-induced dipole interaction. Varying as $1/r^6$ it is the most rapidly decaying attractive force between atoms or molecules and is only significant for two molecules that are in very close proximity.

The total potential energy function for the interactions between two atoms or molecules is the sum of all the interaction energies. It will, of course, depend on the details of the particular atoms or molecules, but for many purposes can be accurately modeled by combining a positive (repulsive) hard-sphere potential energy function with a negative (attractive) longer-range potential energy function. One commonly used form for neutral nonpolar atoms or molecules is the Lennard–Jones or "6–12" potential function

$$PE(r) = 4A\left\{\left(\frac{B}{r}\right)^{12} - \left(\frac{B}{r}\right)^{6}\right\}, \qquad (15.12)$$

a plot of which is shown in Figure 15.12. This function displays most of the usual features seen in atomic or molecular systems. There is a very steep "repulsive wall" at the closest approach distances representing the hard-sphere repulsion. The minimum represents the equilibrium separation distance (at $2^{1/6}B$) for the two particles. Beyond the minimum, the slope becomes positive indicating an attractive force (recall Equation (15.8)) and there is a much less steep "attractive tail" that reaches a nearly neutral plateau beyond about $2B$. With the parameters A and B chosen for the particular system, this potential form is a generally useful approximation.

FIGURE 15.12 *The Lennard–Jones potential function often used to model atomic or molecular systems. In this case A = 1 and B = 0.105.*



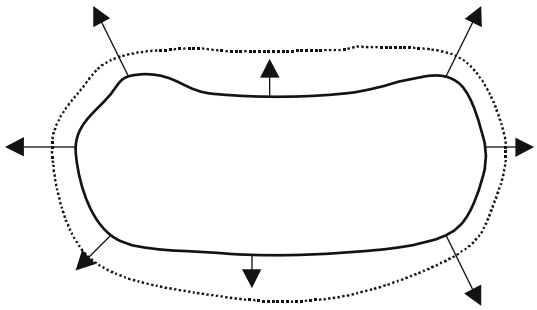


FIGURE 15.13 *Electric field and equipotential surface for a conductor. The electric field is greatest where the curvature is greatest; equipotential surfaces are bunched where the field is largest. The metal surface itself is an equipotential.*

5. STATIC ELECTRICAL PROPERTIES OF BULK MATTER

Having described the fundamental nature of conductors and insulators, let's examine and contrast some of their properties in the presence of an electric field. As we have seen in Section 4 of the previous chapter, at electrostatic equilibrium any excess charge on a conductor resides on its surface and the electric field inside a conductor is zero even when the conductor is placed in an external electric field. Furthermore, at equilibrium the electric field at the external surface of the conductor is always perpendicular to its surface. In the language of electric potential, the surface of a conductor is an equipotential (Figure 15.13). No work is required to move charges on the conductor's surface or throughout its interior as well, since all portions of the conductor are at the same potential.

In the case of an insulator in an electric field, charges are not free to migrate in response to the field. We can distinguish two types of insulators based on whether the molecules have a permanent dipole moment. In *polar dielectrics*, those with a permanent dipole moment, the dipoles will tend to align in the external electric field to some extent. This alignment is due to a torque on the dipole p from the interaction with the electric field E. In a uniform electric field each of the dipole charges q will experience the same force qE, resulting in equal but opposite forces on the dipole (known as a *couple*). The resulting torque on each charge about the dipole center is equal to (see Figure 15.14) $Fr_{\perp} = qE(d/2)\sin\theta$, so that the net torque is given by

$$\tau = qEd\sin\theta = pE\sin\theta, \tag{15.13}$$

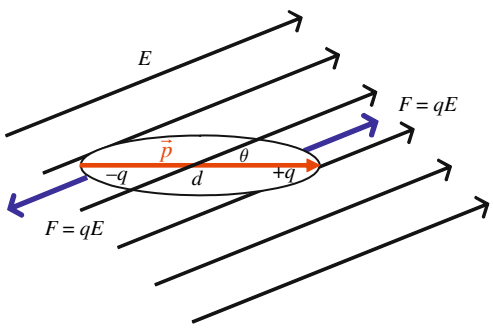
where d is the dipole length and θ is the angle between the dipole and the electric field. This torque will tend to align the dipole with its axis along the E field direction. However, they will not all completely align with the field because of thermal motions that tend to randomly orient the dipoles. Only if the external electric field is quite large and/or the temperature is sufficiently low will the dipole alignment be essentially complete.

A dipole in an electric field will have a potential energy corresponding to the work done by the torque in rotating the dipole. In a uniform electric field this potential energy can be shown to be

$$PE_p = -pE\cos\theta, \tag{15.14}$$

where θ is defined just as in Equation (15.13) to be the angle between $\vec{p}$ and $\vec{E}$ As expected, the lowest energy ($PE_p = -pE$) occurs when the dipole is oriented along the $\vec{E}$ field, a position of stable equilibrium, and the highest energy occurs when $\vec{p}$ and $\vec{E}$ point in opposite directions ($PE_p = pE$), a position of unstable equilibrium. When the dipole is oriented along the E field small perturbations in its orientation lead to a restoring torque as seen in Figure 15.14, but when the dipole is aligned oppositely to the field a small perturbation will lead to a large torque that tends to flip its orientation to line up with $\vec{E}$. These energy ideas are important in later discussions.

FIGURE 15.14 *A couple is exerted on a permanent dipole in a uniform E field.*



When *nonpolar dielectrics* are placed in an external electric field, the molecules become polarized, with their electrons shifting the center of charge away from that of the nuclei in the direction of E, producing an induced dipole moment (Figure 15.15). The extent of this polarization, and therefore the magnitude of the induced dipole moment, depends on the electrical characteristics of the particular molecules.

In any case, when a slab of dielectric, either polar or nonpolar, is placed in an electric field, the net result is to create surface charge layers on the slab as shown in Figure 15.16. There is no net charge throughout

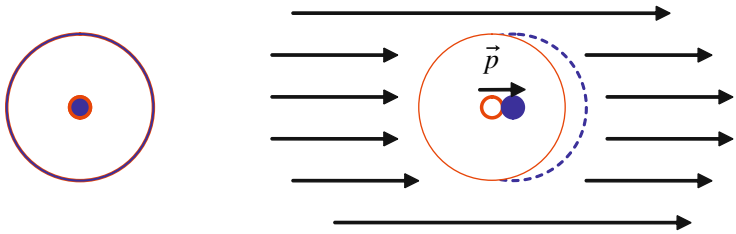


FIGURE 15.15 (left) *Nonpolar atom with centers of positive* (blue) *and negative* (red) *charge overlapping;* (right) *same atom in a uniform electric field, with center of negative charge shifted to the left creating an electric dipole along the electric field.*

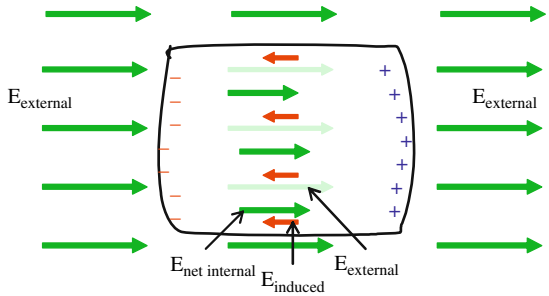


FIGURE 15.16 *The net internal E field is the superposition of the external field* (light green) *and the internal field due to the induced surface charges* (red)*. It is always reduced due to the shielding of the induced charges.*

the dielectric volume, but because of either the orientation of polar dielectric molecules or the induced dipole of nonpolar dielectrics, surface layers of charge are present. The net effect of these surface charges is to reduce the electric field within the dielectric through a partial shielding. Unlike in a conductor, where the free charges can move in response to an electric field and distribute themselves on the surface so as to cancel the electric field within the conductor, dipoles in a dielectric can only partially reduce the internal electric field. The extent of field reduction depends on the dielectric material and is characterized by the *dielectric constant* κ, a dimensionless number that indicates the factor by which the internal electric field is reduced compared to its value in vacuum

$$E = \frac{E_o}{\kappa}. \tag{15.15}$$

Table 15.1 lists some values for dielectric constants of various insulating materials. Note the extremely high dielectric constant of water indicating that water is a very good insulator. This seems contrary to common knowledge that, for example, it is dangerous to be in water during an electrical storm. The conductivity of water is due entirely to the ionic content of the water. Pure water itself is a very poor conductor of electricity.

Table 15.1 *Dielectric Constants of Some Insulating Materials*

Material	*Dielectric Constant*
Air	1.00054
Paper	~4
Pyrex glass	4.7
Rubber (Neoprene)	~7
Ethanol	25
Water	80

Recently scientists have developed methods to calculate accurate electric potentials near the surface of a macromolecule. This has been a significant advance in our understanding of the interplay of native structure and function and also in our ability to design synthetic new macromolecules not found in nature. Macromolecules are inherently highly charged structures immersed in an ionic environment, whether inside a cell or in a buffered solvent in a test tube. The charges on macromolecules, such as proteins or nucleic acids, play a major role in determining the native structure of the molecule as well as its functioning. Specific small molecules that bind to a macromolecule, known as ligands, may be recognized not only by their size and

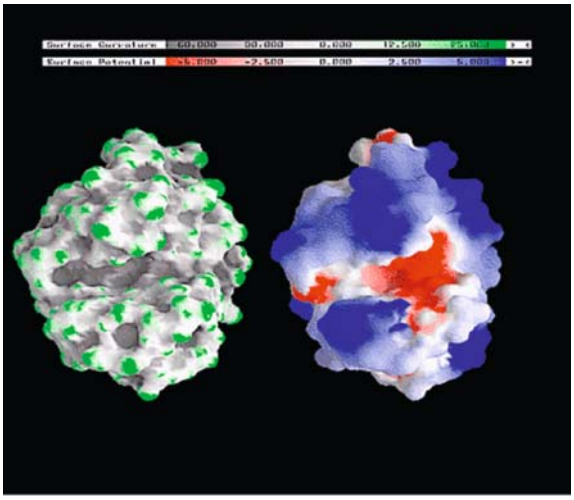


FIGURE 15.17 *Two color-coded images of the protein lysozyme. The left image is coded by curvature and shows a major binding cleft for polysaccharides, whereas the right image is coded by electrostatic charge and shows a highly negative* (red) *binding site in an otherwise positively charged* (blue) *lysozyme. Computer modeling has allowed these detailed pictures only in recent years.*

shape, but also through their charge interactions. Charge groups near an active site on an enzyme may play a role in regulating the binding rates of ligands.

These electric potential calculations require a detailed knowledge of the three-dimensional structure of a macromolecule, complete with the locations of all its atoms. A catalog of the complete structure of many proteins is rapidly growing and is available in computer databases. In one widely used calculational scheme, a cubic lattice grid of points (like three-dimensional graph paper) is chosen and values for charge density, dielectric constant, and ionic strength parameters are assigned to each lattice point. The surface of the macromolecule is usually taken as the so-called van der Waals (or hard-sphere) envelope of the surface atoms and a low internal dielectric constant (2 – 4) is chosen to represent a mean value whereas a large value (~80) is assigned to external lattice sites to represent the aqueous solvent. At this point the problem becomes a classical electrostatics calculation with a large set of point charges given at known locations. The qualitative presentation in Section 3 above is used in a more quantitative form to write down the mathematical problem and numerical methods have been developed in order to calculate the electric potential. Those methods have been greatly improved in recent years so that fairly rapid calculations can now be performed and these improvements have led to a rebirth in the application of electrostatics to the study of macromolecular interactions.

The three-dimensional mapping of the electric potential (see color-coded examples in Figures 15.10 and 15.17) reveals patterns of interaction energies that are not at all apparent from the three-dimensional structure of the macromolecule itself. Patterns of positive or negative potential can be seen over the surfaces of macromolecules and such specific potential features near an active site for binding of a ligand can give important information on the electrostatic interactions with the ligand. Studies of similar macromolecules can show the importance of various specific portions of those structures.

A general knowledge has been assembled on the electrostatic effects of various common structural elements found in proteins and nucleic acids and this body of knowledge has been extremely useful in de novo protein design, the planning and fabrication of new proteins not known to occur in nature. Since the mid 1980s several such proteins have been designed and made. So far they have not been designed with the idea of inventing important new macromolecules, but rather to test fundamental notions on the relationship between structure and functioning of macromolecules by designing simplified macromolecular "motifs". For example, a number of proteins have been created to act as membrane channels (see below) in order to test ideas on the minimum necessary characteristics of such proteins to allow a functioning channel. Knowledge gained in these endeavors will no doubt lead to the future development of new proteins able to perform specific biological functions in living tissue, perhaps replacing the function of defective proteins.

6. CAPACITORS AND MEMBRANES

The lipid bilayers of cell membranes can be electrically modeled as a sandwich consisting of two layers of a conductor (the plane of the polar lipid heads) separated by a dielectric layer (the hydrocarbon tails; see Figure 15.18). Such an electrical arrangement is known as a *capacitor*, or sometimes a condenser. When made of metals and insulators this is a common device for storing electric potential energy and is found in essentially all electronic devices, from telephones to computers. Surrounding a cell, the lipid bilayer provides a barrier to maintain a different internal environment of ions and macromolecules from the extracellular bathing fluid. Because of an unequal distribution of various ions between the inside and outside of all living cells, there is an electric potential difference across

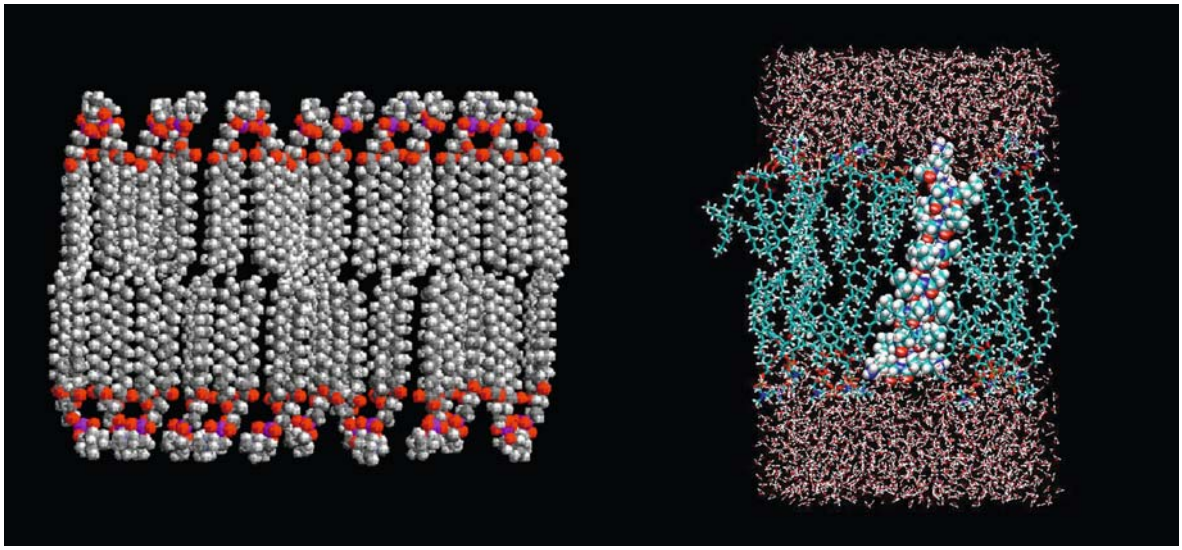


FIGURE 15.18 *Two models of a lipid bilayer with polar heads on the surface and hydrocarbon tails buried within. The image on the right also shows an α-helical transmembrane polypeptide.*

all cell membranes known as the *resting potential.* Its magnitude varies according to cell type, but the inside of cells is always negative with respect to the outside and the magnitude of the potential difference is roughly 100 mV or 0.1 V and is relatively constant.

Certain types of cells have evolved to respond to particular types of stimuli (electrical, chemical, or mechanical) all with the same basic signal, a transient change in the membrane potential (*depolarization* of the membrane), followed by a restoration of the resting potential (*repolarization*). Such cells include nerve, muscle, and sensory cells, all having a similar basic membrane structure. We first develop some concepts about the storage of charge on a generic capacitor before returning to consider the capacitance and charge properties of membranes.

As we have seen, the work done in assembling any array of electric charges results in an electric potential energy. A device used to store electric charge will also thereby store energy. Any array of conductors will serve this function and act as a capacitor, but several simple geometries using two conductors (known as the plates of the capacitor) usually separated by a dielectric are most often used. Figure 15.19 shows some examples of common capacitors used as electrical devices.

FIGURE 15.19 *Variously packaged common capacitors.*

Consider a parallel-plate capacitor shown in Figure 15.20. Such a capacitor is a prototype for all capacitors and even the electrical symbol for a capacitor $\left|\right|$ resembles a parallel-plate capacitor. If made with thin metal foil conductors and dielectric layers in between, the plane sheets can be rolled up to form a compact cylindrical device. In introducing capacitance, we first assume that there is no dielectric layer between the plates but just vacuum. Suppose equal and opposite charges $\pm Q$ are placed on the two plates. There will then be an electric field between the two plates and a corresponding potential difference that we denote by V. In general, the charge Q on either plate of a capacitor and the potential difference across the plates are proportional, defining the capacitance C by

$$Q = CV. \tag{15.16}$$

From the boxed calculation in Section 4 of Chapter 14, the electric field from a plane sheet of charge is $E = \sigma/2\varepsilon_o$, where σ is the charge per unit area (Q/A) on the sheet. For the parallel plate situation of Figure 15.20, the fields produced by each plate add in the space between the plates, but

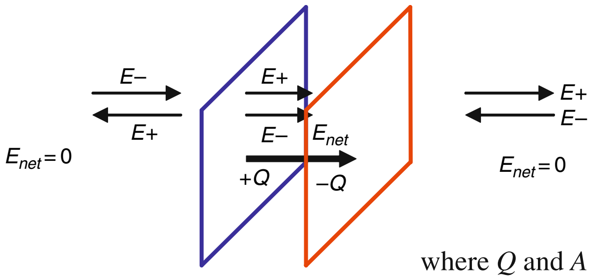


FIGURE 15.20 *A charged parallel plate capacitor, showing the cancellation of electric fields outside and net E field within the capacitor. The electric field from each plate is constant and points either away from the positive or toward the negative plate. Superposition of these electric fields leads to confinement of the electric field between the capacitor plates.*

cancel in the space outside the plates as shown. The electric field between the plates (away from the edges where boundary effects occur) is therefore constant and given by

$$E = \frac{Q}{\varepsilon_o A}, \tag{15.17}$$

where Q and A are the charge on an area of one of the plates. Because E is a constant, the potential difference between the plates is given by Equation (15.6) as

$$V = Ed = \frac{Qd}{\varepsilon_o A}, \tag{15.18}$$

where d is the plate separation. From Equations (15.16) and (15.18), we find that the capacitance of the parallel-plate capacitor is given by purely geometric factors as

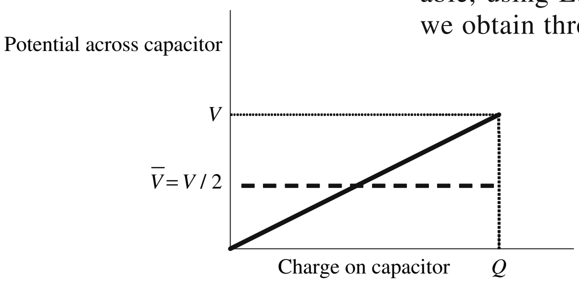
$$C = \frac{\varepsilon_o A}{d}. \quad \text{(parallel−plate C)}. \tag{15.19}$$

The fact that the capacitance depends entirely on geometry is a general result, regardless of the capacitor's shape. Units for capacitance are given by those of Q/V or 1 C/V = 1 farad (F). A farad is an enormous value for capacitance and units of pF to μF are common.

A charged capacitor not only stores charge, but also energy. For a parallel-plate capacitor we can calculate the stored energy from the following argument. Imagine the plates to be initially uncharged and the charging to occur by the transfer of electrons from one plate to the other in a process that results in equal but opposite final charges. After the plates have been partially charged and the potential is at some intermediate value V' between 0 and the final potential V, in order to transfer a small additional amount of charge Δq we need to do an amount of work equal to $\Delta q\, V'$. To transfer a total amount of charge Q, we cannot simply multiply the final charge and potential together because the potential changes in proportion to the amount of charge transferred. However, we can obtain the correct value for the work done by imagining that instead of continuously transferring charge, we transfer all of the charge Q through a constant potential difference that is equal to the average value during the actual process. Because the average potential is $V/2$ (see Figure 15.21), we find that the work done is $W = Q(V/2)$ The potential energy stored in the capacitor is then equal to

$$PE = \frac{1}{2} QV. \tag{15.20}$$

FIGURE 15.21 *The potential across the capacitor increases linearly with the charge on the plates. The same work (equal to the area under the diagonal line) is done in charging the plates continuously as would be done by transferring charge Q across the average potential of V/2 (area under the heavy dashed line).*

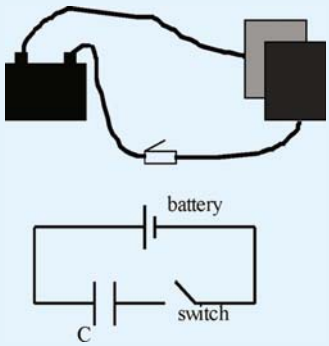


Because in the actual charging process both Q and V vary with time, it is often useful to rewrite Equation (15.20) in terms of the capacitance and only one variable, using Equation (15.16). Substituting for either Q or V in Equation (15.20), we obtain three equivalent forms for the stored energy of a capacitor,

$$PE = \frac{1}{2} QV = \frac{1}{2} CV^2 = \frac{1}{2}\frac{Q^2}{C}. \tag{15.21}$$

An example may help to clarify the appropriate use of these expressions.

Example 15.4 This is our first example of an electrical circuit. We want to find the charge on the 10 × 10 cm plates of a parallel-plate capacitor (shown below on the right) with a 1 mm air gap after it is connected to the terminals of a 12 V battery as shown.



Solution: The figure shows both the actual physical arrangement and the electrical circuit diagram used to represent the situation. Note that the symbol for a battery (with "+", longer line, and "−", shorter line, terminals) is somewhat similar to that for a capacitor (with equal length lines) and that the drawing of connecting wires is arbitrary as long as they have the same connections at their ends. The battery is a device that supplies a potential difference, or voltage, between its two terminals. When the switch shown in the diagram is closed, charge flows from the battery onto the capacitor plates until the voltage across the capacitor plates reaches the same value of 12 V that is across the battery terminals. At this point the two separate "halves" of the circuit, the left and right portions of the circuit diagram corresponding to the two physically separated metal parts of the circuit, divided by the air gap in the capacitor and the battery acid within the battery, are each equipotential surfaces and no further charge flows. The positive side of the circuit is at a potential of 12 V with respect to the 0 V of the negative side.

To determine how much charge flows, we must first calculate the capacitance of the parallel-plate capacitor using Equation (15.19). We find

$$C = \frac{\varepsilon_o A}{d} = \frac{8.85 \times 10^{-12}\,(0.1 \times 0.1)}{0.001} = 88\ pF.$$

The amount of charge on each plate of the capacitor is then found from the definition of capacitance, Equation (15.16), to be

$$Q = CV = 88 \times 10^{-12}\ F \times 12\ V = 1.1 \times 10^{-9}\ C,$$

with the plate attached to the positive battery terminal with +1.1 nC and the other plate with −1.1 nC of electric charge.

What does it mean that the work done in charging a capacitor is stored as potential energy? One view is that the energy is stored in the configuration of charges and that if the two capacitor plates are connected by a conductor, electrons on the negative plate will gain kinetic energy and rapidly flow to the other, positive, plate, thus neutralizing both plates. We discuss this further in the next chapter where we discuss the flow of electric charge.

Another equivalent, but perhaps more revealing, view is that the energy is stored in the electric field that is created between the capacitor plates. If we substitute for C and V from Equations (15.18) and (15.19), we can find an expression for the potential energy that depends only on E and the geometry of the plates

$$PE = \frac{1}{2}CV^2 = \frac{1}{2}\left(\frac{\varepsilon_o A}{d}\right)(Ed)^2 = \frac{1}{2}\varepsilon_o E^2 Ad. \tag{15.22}$$

The product Ad is just the volume between the capacitor plates that the electric field fills uniformly without extending outside the capacitor, thus Equation (15.22) states that there is an energy per unit volume, or energy density, stored in the electric field and given by

$$\frac{PE}{(\text{Vol.})} = \frac{1}{2}\varepsilon_o E^2. \tag{15.23}$$

This is a fundamental relationship for the energy stored in an electric field. Despite the rather specialized example used to derive this result, we show later that it is indeed a very general and important result that is not restricted to capacitors. The fact that there is energy in the electric field, and that the energy is proportional to the square of the field, leads us to many significant developments in electromagnetism.

If a dielectric material with dielectric constant κ fills the space between the plates, then, as we have seen, the internal electric field is reduced by the factor κ. With a given charge Q on the plates, the presence of the dielectric reduces the potential difference between the plates by the same factor κ (because according to Equation (15.6) $V \propto E$) so that the capacitance is thereby increased by the factor κ (because according to Equation (15.16) $C \propto 1/V$):

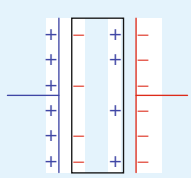
$$V = \frac{V_o}{\kappa};\ C = C_o\kappa, \tag{15.24}$$

where the initial values are those without the dielectric.

With a good insulator, capacitance values can be substantially increased. The increase in capacitance with a dielectric implies (from Equation (15.22)) that if the voltage across the capacitor is maintained constant by a battery, for example, then the stored energy and charge will both *increase* by a factor κ relative to the same capacitor without a dielectric. On the other hand, Equation (15.23) implies that because the electric field decreases by κ, E^2 should decrease by a factor of κ^2, whereas the term ε_o is multiplied by a factor of κ, so that the energy stored should *decrease* by a factor of κ relative to the same capacitor without a dielectric. The following example should help to clarify this apparent paradox.

Example 15.5 A 0.1 μF parallel-plate capacitor is charged by a 12 V battery and disconnected from the battery. A slab of dielectric with $\kappa = 4$ is then inserted to fill the gap in the capacitor. Find the charge on the capacitor plates and the voltage across the plates before and after inserting the dielectric. If the capacitor is then reconnected to the battery, how much more, if any, charge will flow onto the capacitor?

Solution: When connected to the battery, the capacitor will be charged to 12 V and will have $Q = CV = (0.1\ \mu\text{F})(12\ \text{V}) = 1.2\ \mu\text{C}$ of charge on each plate. After the capacitor is disconnected from the battery, this charge will remain on the capacitor. (We show in the next chapter that for a real (nonideal) capacitor, the charge will, in fact, slowly leak off, but we



ignore that here.) When the dielectric is inserted, the charge still remains on the capacitor, but the dielectric will have an induced layer of surface charge that will shield the charge on the metal plates (see the figure) and reduce the electric field and the potential within the capacitor by a factor of $(1/\kappa)$. Accordingly, the potential is reduced to $V' = 12/4 = 3$ V.

Equation (15.23) tells us there is a corresponding decrease in potential energy stored in the capacitor by a factor of $(1/\kappa)$. What happened to this energy? As the dielectric is inserted between the capacitor plates, the induced charges on the dielectric cause an attractive force pulling the slab into the gap between the capacitor plates. In terms of overall energy conservation, negative work has to be done on the dielectric by an external agent, using an external force to hold the slab back from accelerating into the gap, in order to position the dielectric within the capacitor. A careful calculation shows that this negative work just balances the decrease in stored potential energy.

If the capacitor is then reconnected to the battery, the potential across the plates will again rise to 12 V with the transfer of additional charge to the metal plates. The total charge on the plates is then given by the product of the voltage and the capacitance (now increased by a factor of κ), $Q_{\text{total}} = (12\ \text{V})(0.4\ \mu\text{F}) = 4.8\ \mu\text{C}$, so that an additional $(4.8 - 1.2) = 3.6\ \mu\text{C}$ of charge was transferred to the plates.

So, the resolution of the apparent paradox presented before this example is that the resulting energy stored depends on whether the capacitor has its voltage fixed, while attached to the battery, or has its charge fixed, when isolated. In the first case additional charge will flow onto the capacitor to maintain the voltage fixed at the battery value, whereas in the second case the field and voltage will decrease because of the dielectric screening.

The capacitance per unit area (specific capacitance) of cellular membranes was first determined in the 1920s to have a value of about 1 $\mu\text{F/cm}^2$. This number was used to estimate the thickness of the previously undetected cell membrane using the parallel-plate relation for capacitance (Equation (15.19) multiplied by the factor κ) $C/A = \kappa\varepsilon_o/d$. Using a value of $\kappa = 3$ (based on the knowledge that membranes contained lipids and that oils have a value of $\kappa \sim 3$) and the measured value for C/A, an estimate for the membrane thickness of $d \sim 3$ nm was obtained (you can verify this). Although today we know that most membranes are about 7.5 nm thick, this was the first such determination and indicated that the membrane thickness might correspond to the length of a macromolecule.

Assuming that the charge on a cell membrane is uniformly distributed, we can obtain an estimate of how much charge lies on a membrane. From Equation (15.15), by dividing both sides by the area of the membrane, we obtain

$$\frac{Q}{A} = \frac{C}{A}V. \tag{15.25}$$

If we take $V = 0.1$ V and a capacitance per unit area of 1 $\mu\text{F/cm}^2$, then we find a charge per unit area of 0.1 $\mu\text{C/cm}^2$. We can get a feeling for this charge density on the membrane by calculating the average spacing of the individual charges on the membrane surface. With x equal to the average separation between charges on the membrane surface, so that there is one charge per surface area x^2 (see Figure 15.22), we can find a value for x from

$$\frac{1\ \text{charge}}{x^2\ \text{cm}^2} = \frac{0.1 \times 10^{-6}\ \text{C/cm}^2}{1.6 \times 10^{-19}\ \text{C}} = 6.25 \times 10^{11}\ \text{charges/cm}^2\ ,$$

so that, solving for x, we find that there is one charge every $x = 13$ nm in a square array over the surface of the membrane.

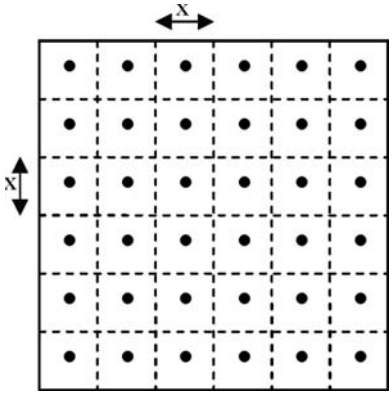


FIGURE 15.22 *A uniform surface charge model for a cell membrane with one charge centered in each box.*

We can also calculate the electric field inside the membrane from Equation (15.18). Substituting $V = 0.1$ V and $d = 3$ nm, together with a reduction in E by the factor $\kappa = 3$, we find that $E = 1.1 \times 10^5$ V/cm, an extremely high value. In fact, the largest possible E field in dry air is only 0.3×10^5 V/cm, with higher E fields in air causing dielectric breakdown. Such large E fields in membranes are responsible for relatively large forces on molecules within membranes, suggesting that by proper triggering, much energy can be released through interaction with the electric field.

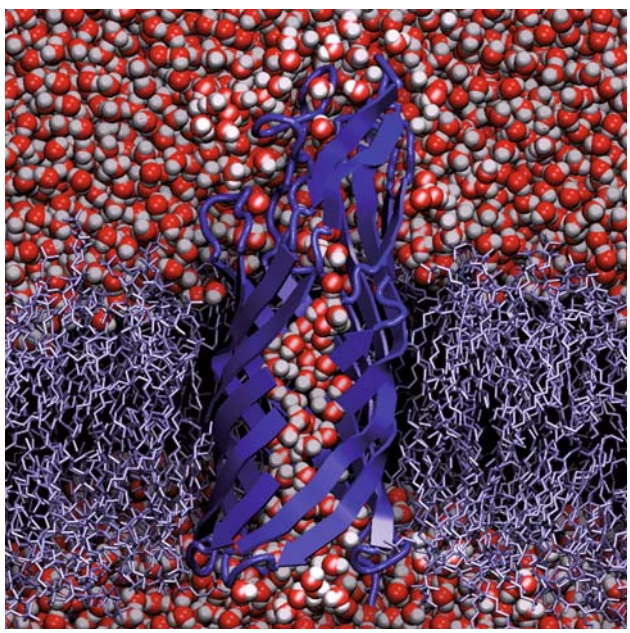
Although the membrane capacitance can be approximated by an expression for a parallel-plate capacitor, it should be pointed out that the electrical properties of a membrane are quite a bit more complex than an ideal capacitor. As we show in the next section and again in the next chapter, membranes do allow a flow of charge through specific pores known as channels. Furthermore, along membranes in large cells such as nerve or muscle, the properties of the membrane vary both spatially along the membrane and with time. Membranes are indeed far from passive conducting plates separated by an ideal insulator. They are dynamic structures with very complex electrical properties capable of rapidly changing the ionic environment of a cell, of transporting large macromolecules across the cell barrier, and of propagating electrical signals rapidly over long distances.

7. MEMBRANE CHANNELS: PART 1

Membrane channels are specific integral membrane protein/sugar/fatty acid complexes that act as pores designed to transport ions, water, or even macromolecules across a biological membrane (see Figure 15.23). Channels play a distinctive role in excitable cells, such as neurons and muscle cells, where they control the flow of ions and the subsequent generation of electrical signals. In this section, we learn some fundamentals of the general nature of channel structure and functioning in anticipation of a fuller discussion of the role of channels in nerve conduction in the next chapter.

There are probably hundreds of different specific channels in various types of mammalian cells. Although first studied and modeled by Hodgkin and Huxley in the early 1950s in ground-breaking experiments, channels have recently been studied using a large array of techniques including modern electrophysiology, biochemistry, and molecular biology. In a simple picture, channels can be said to exist in either of two states, open or closed, in which specific ions or small molecules can either pass through the channel "gate" or not. Control of the gating of a channel can be either by specific charges (*voltage-gating*) or by the binding of small molecules (*ligand-gating*). Ligand-gated channels include those for neurotransmitters and small proteins involved in other forms of cell signaling. Voltage-gated channels are present in nerve and muscle and we focus on these in our discussion.

FIGURE 15.23 *Molecular model of a membrane showing a channel in the form of a mostly helical protein that spans the membrane (shown in blue) allowing selected ions to enter or leave the cell.*



To give a more concrete idea of what a channel is and how it functions in a cell membrane, let's consider the sodium ion channel in some detail. The Na channel is formed by a complex of a single polypeptide chain of about 2000 amino acids with associated sugars and fatty acids. This single chain has four similar subunits, each composed of six helical portions, with each of these spanning the membrane so that the overall structure resembles that shown in Figure 15.24. The Na channel has been purified and shown to be functionally active when reconstituted in pure lipid membranes. In muscle, there are between 50 and 500 Na channels per μm^2 on the membrane surface. Each of these is normally closed, but can be opened by a change in the electric potential across the membrane. The open state is short-lived, lasting about 1 ms, during which time about 10^3 Na^+ ions flow into the cell through each channel from the higher Na^+ ion-rich extracellular medium. When the channel is open, the flow is highly selective for Na^+ ions, with potassium (K^+) ions some 11 times less likely to cross the Na channel.

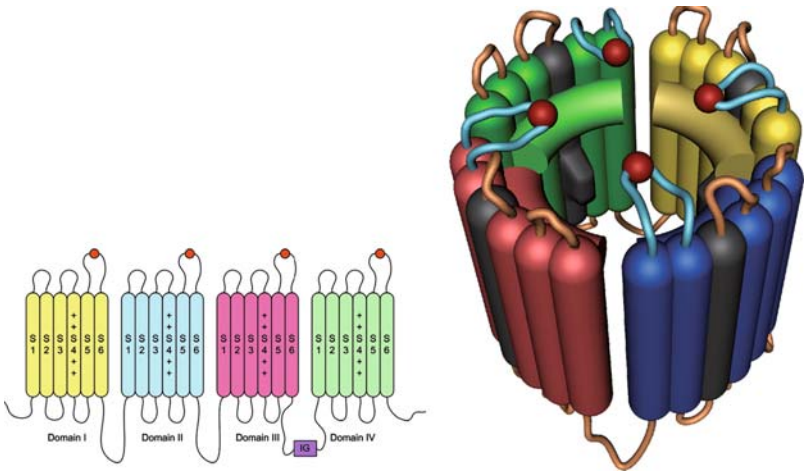


FIGURE 15.24 *Sodium channel:* (left) *schematic of alpha-helical sections spanning cell membrane;* (right) *molecular model of channel.*

From the vast array of questions that have been and are being asked about how these channels function, we consider two. How can the electric potential control the gating or flow of Na^+ ions through the channel? What allows the Na channel to be so selective in the transmission of ions? Although complete answers to these questions cannot yet be given, our knowledge has dramatically increased in the recent past.

Channels open in response to a stimulus detected by a sensor. In Na channels, the stimulus is an electric field near the channel, sensed by a collection of charges or dipoles on one particular helical section of each of the four subunits within the channel. It has been shown that there is a small movement of charge across the membrane just prior to the opening of a channel. These four to six gating charges move in response to the electric field stimulus and this interaction provides the needed energy to open the channel. Several specific models of voltage gating have been proposed that suggest different types of conformational changes in the channel helices spanning the membrane to explain the opening of a Na channel. Experiments with a large variety of monovalent ions and with various chemical blockers to prevent the channel from opening have shown that the Na channel has a pore of dimensions about 3 by 5 Å with its interior surrounded by a cluster of oxygen atoms. The size filtering of the pore coupled with the need to interact with the negative oxygen charge sites provides the specificity of the channel. Potassium (K^+) ions have a diameter of about 2.66 Å, whereas Na ions have diameters of about 1.9 Å. It is thought that the K^+ ions are associated with at least one water molecule and this would thereby prevent them from entering the Na channel simply based on size.

An obvious question that arises is how the potassium channel can then be even more specific, about 100 times more permeable to K^+ than to Na^+, given sodium's smaller size? The K channel is the narrowest channel known, excluding all ions larger than 3 Å in diameter. Smaller ions that could fit through the K channel as bare ions, such as lithium or sodium, do not enter because of the free energy cost to dehydrate these small ions. The potassium ion is able to shed its water molecules because it is able to interact very closely with the oxygens lining the K channel. Sodium and other small ions would not make such close contact with the oxygens, because their bare diameters are smaller than that of K^+, and the energy cost in shedding their water molecules is therefore too high.

We return to a somewhat detailed study of the cellular electrical properties controlled by channels in the next chapter.

8. ELECTRIC POTENTIAL MAPPING OF THE HUMAN BODY: HEART, MUSCLE, AND BRAIN

The human body uses a complex system of electrical signaling to control various life functions. A network of nerve cells provides both sensory input and motor control. Our brains are complex webs of neurons able to outperform the most sophisticated computers in even the simplest tasks of recognition. Muscles conduct electricity as well as generate force. The heart should be singled out as the most notable muscle in

the body, pumping blood by contraction of a series of muscles all controlled by the electrical activity of a pacemaker group of cells.

A number of medical technologies have been developed over many years to map the electrical activity of these various organs of the body. Here we briefly describe three such technologies to map the electric potentials from muscles, the heart, and the brain. The methods are known as *electromyography* (EMG of muscle), *electrocardiography* (ECG or EKG of the heart, the K rather than C appearing from the original Dutch), and *electroencephalography* (EEG of the brain). Although we have little fundamental knowledge that would enable us to directly interpret the complex time and spatial patterns of such electric potentials, doctors have many years of empirical data allowing these methods to be used as indicators of normal or abnormal behavior.

The fundamental principles of the three techniques are the same: the mapping in time and space of the surface electric potential corresponding to electric activity of the organ. When a resting nerve or muscle cell, with a membrane potential of about -100 mV relative to the external medium, is stimulated, a wave of depolarization spreads over the surface of the cell. The resting cell has no dipole moment, but while the cell is undergoing depolarization it can be electrically represented by a time-varying electric dipole moment that goes to zero after the resting membrane potential is restored in a process of repolarization. We study some details of the depolarization and repolarization processes in connection with nerve conduction in the next chapter. For now, it is clear that such changes will lead to local variations in electric potential. In trying to map these changes in electric potential, only in EMG can an electrode be directly used to measure local potentials. Otherwise, for the heart and brain surface electrodes must be used. Implicit in their use is the notion that the body is a very good conductor, so that potential changes measured, for example, in an EKG, between the ankle and the wrist, reflect the potential differences directly across the heart.

In EMG, the simplest of the three techniques discussed, either surface electrodes on the skin or a needle electrode inserted into a muscle record the time variations in electric potential. Needle electrodes can probe a single muscle fiber and give a characteristic time record of electric potential with variations of several mV observed (Figure 15.25). Such recordings of voluntary muscle activity can check for normal functioning of nerve stimulation of muscle. More detailed information can be obtained with external electrical stimulation of the muscle because an entire group of muscle fibers can be simultaneously activated. Measurements at a number of distances along a muscle can determine conduction velocities along the stimulating nerve. Although not as common as the EKG or EEG, the electromyogram can be more directly related to the depolarization of a single cell or small group of cells.

The heart is composed of many individual muscles contracting in a synchronous fashion controlled by the pacemaker or sinoatrial (SA) node, located in the right atrium. Triggering of the pacemaker cells roughly once per second stimulates a wave of depolarization down across both atria, leading to their contraction and pumping of blood into the ventricles (as discussed in Chapter 9). Following the atrial contraction and repolarization, another wave of depolarization is initiated by the atrioventricular (AV) node that lies between the two ventricles, leading to contraction of the ventricles and their subsequent repolarization. This entire sequence of events constitutes a heartbeat cycle and, just as in EMG, the waves of depolarization can be measured as surface electric potential changes, although in this case the potential waveform is quite complicated.

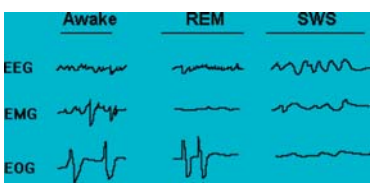


FIGURE 15.25 *Contemporaneous EEG and EMG recordings when awake, in rapid eye movement (REM) sleep, about 20% of the time for an adult, and when in slow wave sleep (SWS).*

In its simplest form, an EKG can be imagined to measure the electric potential due to the heart being represented as a single time-varying electric dipole moment $\vec{P}(t)$. In order to determine the value of $\vec{P}(t)$, three independent measurements must be made so as to determine the three vector components of the dipole moment as functions of time. Accordingly, there are three required surface electrodes that must be used in EKG. These are attached at both wrists and the left ankle. Additional electrodes, usually a total of 12, are used to assist in the analysis, but these are not fundamentally required. An EKG recording gives information on the time sequence of potentials and the characteristic peaks and valleys are labeled in a standard manner (Figure 15.26). Newer computer-interfaced instrumentation can obtain

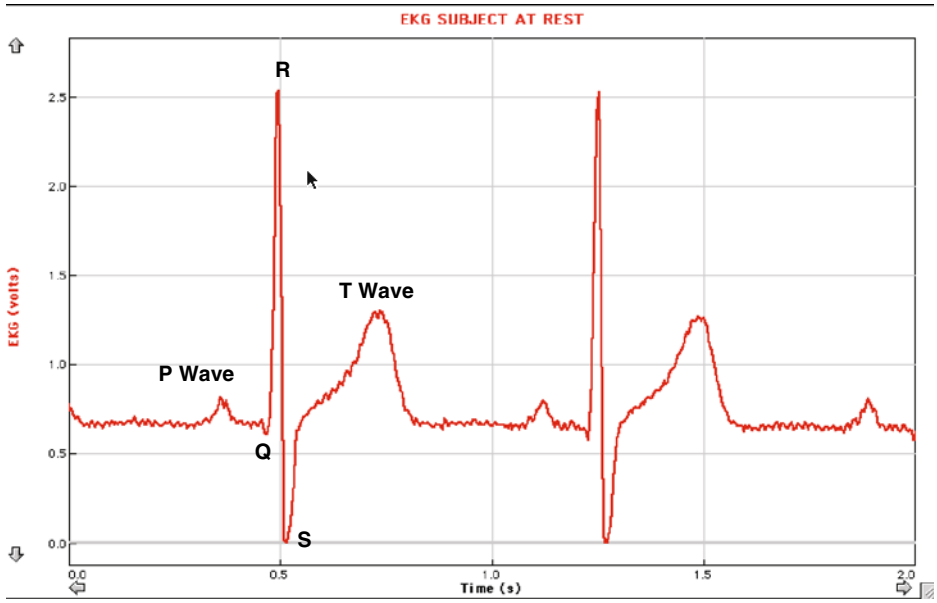


FIGURE 15.26 *A segment of an EKG signal showing the prominent features.*

high-quality data and analyze EKGs for the amplitudes, durations, and areas under the primary peaks. These are then used for diagnostic purposes, with computers even able to point out potential problems. As we have already remarked, despite a lack of knowledge to interpret these potential mappings in detail, simply by the huge number of recordings available, EKGs are very useful empirical tools for the diagnosis of various forms of heart disease.

Since the first detection of electrical signals from the brain in 1929, doctors and scientists have been recording such signals in the form of EEGs in order to learn about the electrical activity of the brain. These signals are much weaker than those from the heart or muscle, typically less than 0.1 mV, and the patterns of voltage signals recorded are much more complex as might be expected from much more asynchronous firings of neurons compared to the heart. Some characteristic wave trains can be associated with various activities or abnormalities, including different stages of sleep, epileptic seizures, and visual or auditory excitation (so-called evoked responses). Frequency analysis of the wave trains divides signals into four frequency bands ranging from slow Δ waves at 0.5–3.5 Hz common during sleep, θ waves in the range from 5 to 8 Hz common in newborns but indicating severe stress in adults, to normal α waves at 8–13 Hz from a relaxed brain and faster β waves at greater than 13 Hz from an alert brain (Figure 15.27).

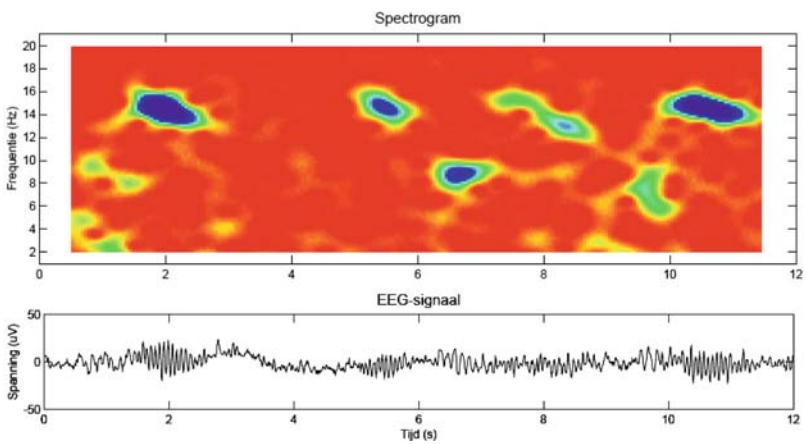


FIGURE 15.27 *Lower EEG trace analyzed in terms of its frequency content in the upper spectrogram. Note the microvolt scale for the EEG signal and also the mixture of different types of waves based on frequency content.*

A standard arrangement of 8–16 electrodes placed in a regular pattern around the head is used to record an EEG. Again, it should be emphasized that such measurements are not well understood, but are able to help in a diagnosis based simply on clinical studies of many individuals. In Chapter 18 we show another new technique for studying the electrical activity of the brain by measuring the very small magnetic fields generated by the brain. This technique, known as magnetoencephalography, or MEG, is better able to localize electrical activity within the brain and has recently led to very interesting results. The method requires some quite specialized equipment but is a growing area of research and potential clinical use.

CHAPTER SUMMARY

Electric potential energy is defined by the negative of the work done by the electric force. For two interacting point charges q_1 and q_2, separated by distance r, the PE of interaction is

$$PE_E(r) = \frac{q_1 q_2}{4\pi\varepsilon_0 r}. \qquad (15.2)$$

More commonly used is the electric potential V, at some point in space, defined by

$$V(r) = PE_E(r)/q_o, \qquad (15.4)$$

where q_o is a small, positive test charge imagined to be placed at the point of observation. The electric potential at a point is the external work needed to move a unit positive charge from infinitely far away to that point along any path. For a point charge q at the origin, the potential it produces a distance r away is given by

$$V(r) = \frac{q}{4\pi\varepsilon_o r}. \qquad (15.5)$$

When the electric field is uniform along some direction x, it is simply related to the variation in the electric potential along that direction as

$$E_x = -\frac{\Delta V}{\Delta x}. \qquad (15.7)$$

We can visualize variations in electric potential using equipotential mappings that show the surface contours with constant voltage. These surfaces must lie perpendicular to the electric field lines.

One particular arrangement of charges, a positive and equal negative charge q separated by some distance d, is particularly important. Known as an electric dipole, such a pair of charges gives rise to an electric potential given by

$$V = \frac{qd\cos\theta}{4\pi\varepsilon_o r^2} = \frac{p\cos\theta}{4\pi\varepsilon_o r^2}, \qquad (15.11)$$

where r is the distance from the dipole p to the observation point and θ is the angle between the dipole direction (taken as from the negative to positive charge) and the observation direction. Water molecules are the ubiquitous dipole in biology. Electric interactions between two charge distributions can be classified according to their energy dependence on the separation distance. Charge–charge interactions have a $1/r$ dependence, as given by Equation (15.1). Charge–dipole, dipole–dipole, and dipole–induced dipole interactions are each weaker than the previous listed interaction, dropping off faster with separation distance r. Even in the absence of permanent dipoles, molecules can interact through attractive dispersion forces, involving fluctuating induced dipole interactions.

When an electric dipole is placed in a uniform electric field it experiences no net force, but a net torque given by

$$\tau = qEd\sin\theta = pE\sin\theta, \qquad (15.13)$$

where θ is the angle between the dipole and the electric field. There is also a potential energy of this interaction given by

$$PE_p = -pE\cos\theta. \qquad (15.14)$$

A capacitor (with capacitance C) is an electrical device with two conductors oppositely charged (with

$\pm Q$ on either surface) with a potential difference V between them, where

$$Q = CV. \qquad (15.16)$$

In the case of two parallel conducting plates of area A separated by distance d, (e.g., a good model for the lipid membrane bilayer, consisting of two layers of polar, conducting "heads" separated by hydrocarbon, nonconducting chains) the capacitance is given by

$$C = \frac{\varepsilon_o A}{d}. \quad \text{(parallel-plate } C\text{)}. \qquad (15.19)$$

A capacitor stores electrical potential energy according to these equivalent expressions:

$$PE = \frac{1}{2}QV = \frac{1}{2}CV^2 = \frac{1}{2}\frac{Q^2}{C}. \qquad (15.21)$$

When the gap between the conducting plates of a capacitor is filled with a dielectric with dielectric constant κ, the electric field between the plates is reduced by κ,

$$E = \frac{E_o}{\kappa}, \qquad (15.15)$$

and, when the charge on the capacitor is fixed, the potential and capacitance are also changed according to

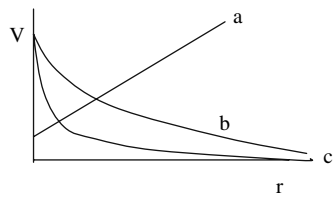
$$V = \frac{V_o}{\kappa};\ C = C_o\kappa. \qquad (15.24)$$

Thus, the addition of a dielectric increases the capacitance and the charge storing ability of a capacitor.

The chapter includes a few applications of these ideas in the study of membrane channels in biological membranes and in a variety of medical techniques to study the electrical activity of the heart (EKG), brain (EEG), and muscles (EMG).

QUESTIONS

1. Kinetic energy is always positive, whereas gravitational potential energy can be positive or negative depending on the choice of zero reference level. On the other hand, electric potential can also be positive or negative, but not because of the zero reference level (usually chosen at infinity) but rather because of the two types of electric charges. Discuss the physical significance of the sign of the electric potential energy and the shape of the curves in Figure 15.1.
2. In Example 15.1, the electron orbiting a proton in a hydrogen atom was found to have a negative energy. What would a positive electron energy imply?
3. Two point charges interact with each other. Discuss the sign of the electric potential at the location of the second charge (with respect to infinity) for all possible variations of the signs of the charges. That is, what will be the sign of the potential at the second charge in each of the four possible cases when the first charge is either positive or negative when the second charge is either positive or negative?
4. What is the change in kinetic energy, measured in both joules and electron volts, when a calcium ion Ca^{2+} is accelerated through a potential difference of 100 V.
5. Show that the SI units for electric field, either 1 N/C or 1 V/m, are equivalent.
6. A region of space has a uniform electric field present, say along the z-axis. How does the electric potential vary along the z-axis? If a negative charge is located at $z = 100$ units in this electric field and is moved to $z = 50$ units, does its electric potential increase or decrease? Does its electric potential energy increase or decrease? What direction is the force on the charge? Is the work done by the electric forces on the charge in changing its position positive or negative?
7. Sketch the equipotential surfaces around a long straight charged wire; around a charged sphere; around a charged plane; around a "point" electric dipole.
8. Explain why electric field lines are always perpendicular to equipotential surfaces.
9. In the graph shown of potential versus distance, which curve might represent the potential due to (i) a plane of charge, (ii) a point charge, (iii) a point dipole?



10. Represent atoms by small circles that can be charged, neutral, or have a permanent dipole moment represented by an arrow within the circle. Show in a sketch

how two such atoms interact in the following cases (show the direction of the force and/or torque on each atom): (i) one charged and one neutral atom; (ii) one charged and one polar molecule; (iii) and both dipolar molecules with parallel dipoles.

11. Suppose a uniform electric field exists along the x-axis and an electric dipole is oriented along the negative x-direction. Is it in equilibrium? What will happen if it wobbles slightly so that it orients slightly off the negative x-direction? Repeat this question if the dipole is oriented along the positive x-axis.

12. For a parallel plate capacitor, how does the capacitance change if (i) the plate areas are doubled, (ii) the voltage across the capacitor is halved, (iii) the plate separation is halved, (iv) the material between the plates is changed to one with twice the dielectric constant, (v) the charge on the capacitor is doubled?

13. The equation for the energy stored in a capacitor, $PE = 1/2\ CV^2$, implies that if a dielectric is inserted in an air-spaced capacitor, the capacitance increases by a factor of κ so that the stored energy should also increase by that factor. However, our other equation for the stored energy per unit volume says that $PE/V = 1/2\varepsilon_o E^2$ and with a dielectric inserted, ε_o becomes $\varepsilon_o\kappa$ and E decreases by a factor of κ so that the stored energy should decrease by a factor of κ. Explain the resolution of this apparent paradox. (See Example 15.5.)

14. The electric field within a biological membrane is as large as is physically possible. Discuss this statement.

15. Why is it more appropriate to discuss charge per unit area and capacitance per unit area for biological membranes than simply charge and capacitance?

16. What factors allow channels to be so specific to certain ions?

17. Contrast ligand-controlled gating with voltage-controlled gating of membrane channels.

18. What are the similarities and differences among the three techniques EMG, EKG, and EEG? Do a bit of further research into these techniques using the library or Internet resources.

MULTIPLE CHOICE QUESTIONS

1. The fundamental dimensions (mass, length, time, Q-charge) of electric potential are (a) MLT^{-2}, (b) ML^2T^{-2}, (c) $ML^2T^{-2}Q^{-1}$, (d) QT^{-1}.

2. An electron travels through free space from point A, which is at + 100 V, to point B, which is at + 200 V. The kinetic energy of the electron during this trip (a) stays constant, (b) increases by 1.6×10^{-17} J, (c) decreases by 1.6×10^{-17} J, (d) decreases by 100 V.

3. A sphere of copper has a radius of 10 cm. The sphere is in equilibrium, and the electric potential at one point on the surface of the sphere is known to be + 100 V. Which one of the following is true? The electric potential at the center of the sphere (a) is infinity, (b) is zero, (c) is + 100 V, (d) cannot be determined from the information given.

Questions 4 through 7 refer to two 1 μC point charges that are 2 m apart.

4. The energy that went into assembling these two charges is (a) 0.0045 J, (b) 0.009 J, (c) 0.0023 J, (d) 4500 J.

5. The net force on each charge is (a) 0.0045 N, (b) 0.0023 N, (c) 0.009 N, (d) 0.0045 N.

6. The magnitude of the electric field at each charge is (a) 2250 V/m, (b) 0.0023 N/C, (c) 4500 N/C, (d) 1125 V/m.

7. The potential at each charge due to the other charge is (a) 2250 V, (b) 9000 V, (c) 1125 V, (d) 4500 V.

8. Points A, B, and C are on the same line and are 0.01 m apart. The electrostatic potential at A is +5 V, at B is +4 V, and at C is +3 V. The component of the electric field at B along the line AC is closest to (a) +4 V, (b) −1 V, (c) 100 V/m pointing toward A, (d) 100 V/m pointing toward C.

9. The equipotential surfaces around a long straight wire with a uniform charge/length are concentric (a) spheres, (b) cylinders, (c) donuts, (d) planes.

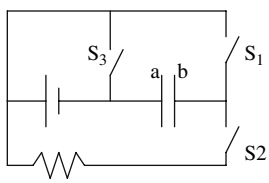
10. An electric dipole oriented along the x-axis sees a uniform electric field along the y-axis. The dipole will experience which of the following? (a) A net force along the y-axis and a torque orienting the dipole along the y-axis, (b) 0 net torque and a force along the y-axis, (c) 0 net force and a torque orienting the dipole along the y-axis, (d) 0 net force and a torque orienting the dipole along the z-axis.

11. The resting potential of a cell membrane is roughly (a) 120 V, (b) 1 V, (c) 0.1 V, (d) 1 mV.

12. The capacitance value of a capacitor depends on (a) the applied voltage, (b) the net charge on either of its plates, (c) geometrical factors, (d) all of the above.

13. The energy stored in a capacitor depends on (a) only the charge on either plate, (b) only the applied voltage, (c) only its capacitance, (d) any pair of the previous quantities.

Questions 14–16 refer to the following circuit diagram in which the battery has 6 V:



14. When the switch S_1 has been closed for a long time (and S_2 and S_3 remain open), the voltage across the capacitor is read on a perfect voltmeter to be $V_a - V_b$ = (a) 3 V, (b) 6 V, (c) −6 V, (d) −3 V.

15. If then switch S_3 is closed (with S_2 remaining open), the voltage across the capacitor immediately reads $V_a - V_b$ = (a) −6 V, (b) 6 V, (c) 0 V, (d) 3 V.

16. If instead of closing switch S_3 in Question 15 after S_1 had been closed a long time, S_2 were closed, the voltage $V_a - V_b$ immediately reads (a) -6 V, (b) 6 V, (c) 0 V, (d) -3 V.
17. A parallel-plate capacitor is attached to a 12 V battery. A slab of dielectric constant 3 is then inserted between its plates, filling the gap, while the capacitor is still connected to the battery. Which of the following occurs: (a) the voltage drops by a factor of 3, (b) the electric field between the plates drops by a factor of 3, (c) the stored energy drops by a factor of 9, (d) the stored energy increases by a factor of 3.
18. If the capacitor of the previous question is first disconnected from the battery and then has the same dielectric slab inserted, then (a) the voltage will drop by a factor of 3, (b) the stored energy will increase by a factor of 3, (c) the electric field will increase by a factor of 3, (d) the capacitance will increase by a factor of 9.
19. Suppose a lump of glass is placed in a uniform external electric field E that points left to right (see Chapter 14, multiple choice question 12). When the charges in the glass come into equilibrium the total electric field inside the lump (a) is larger than E and points left to right, (b) is smaller than E and points left to right, (c) is the same size as E and points right to left, (d) is zero, (e) none of the above.

PROBLEMS

1. Calculate the electric potential energy of three equal 3 μC point charges at the vertices of an equilateral triangle with 5 cm sides.
2. Imagine assembling four equal charges, one at a time, and putting them at the corners of a square. Find the total work done to assemble these if the charges are each 5 μC and the square has 25 cm sides.
3. Equal and opposite $\pm$ 10 μC charges lie along the x-axis with the + charge at $x = 0.1$ m and the − charge at $x = -0.1$ m. Find (a) the electric potential at the origin; (b) the electric field at the origin; (c) the work required to bring a third +10 μC charge from far away to the origin. Repeat all three parts if now all charges are +10 μC.
4. There is a 10 N/C uniform electric field along the x-axis.
   (a) If the potential at the origin is −15 V, what is the potential at $x = 10$ m?
   (b) Where is the potential zero?
5. A pair of equal and opposite 1 μC charges lies along the y-axis symmetrically about the origin, each a distance 0.01 m away, creating a dipole pointing along the y-axis. Find the following.
   (a) The electric field 10 m away along the y-axis
   (b) The potential at the same location as in part (a)
   (c) The electric field 10 m away along the x-axis
   (d) The potential at the same location as in part (c)
6. A 20 C-m electric dipole is located at the origin and points along the y-axis. Find the electric potential at the following locations.
   (a) $x = z = 0$, $y = 1$ m.
   (b) $y = z = 0$, $x = 1$ m.
   (c) $z = 0$, $x = y = 1$ m.
7. What is the torque on a 0.01 C-m electric dipole located between, and oriented parallel to the plates of, a 10 μF air-spaced parallel-plate capacitor with plates of 0.1 m^2 area connected to a 6 V battery?
8. Repeat the previous problem if the dipole is oriented at a 45° angle to the electric field. Also, what is the potential energy of the dipole?
9. Construct a graph of the electric potential energy of a dipole as a function of its angle with respect to the electric field. Where are the equilibrium points and which is a stable and which an unstable equilibrium?
10. What is the maximum torque on a 0.5 C-m electric dipole in a 10 N/C uniform electric field?
11. A lightning flash transfers 5.0 C of charge and 30 MJ of energy to the Earth from a cloud. What potential difference existed between the clouds and the ground?
12. In lightning storms, the potential difference between the Earth and the bottom of the thunderclouds can be as high as 50 MV. The bottoms of the thunderclouds are typically 1.0 mile above the Earth, and can have an area of 25 mi^2. If we model the Earth–cloud system as a huge capacitor, what are the capacitance of the Earth–cloud system, the charge stored in the "capacitor," and the energy stored in the "capacitor?"
13. An uncharged 20 μF capacitor is connected to a power supply set to 10 V. How much charge flows onto the capacitor plates? If the power supply voltage is then increased to 30 V, how much more charge flows onto the plates?
14. A 0.01 μF capacitor is to be constructed by rolling two strips of 10 cm wide metal foil with a 1 mm thick paper layer (dielectric constant = 4) sandwich into a cylinder. How long must the strips be?
15. A capacitor is formed between two metal plates, each 10 × 10 cm, separated by 1 mm.
   (a) What is its capacitance?
   (b) When connected to a 10 V battery how much charge flows onto the plates?
   (c) What is the net electric field between the plates?
   (d) What is the total force on each plate? (Hint: Find the electric field acting on each plate. The force per unit area is given by the product of the electric field acting on one plate and its charge per unit area.)
16. A 100 pF parallel-plate capacitor with a 0.5 mm plate spacing is charged by a 12 V battery.
   (a) What is the electric field between the plates?
   (b) What is the total energy stored in the capacitor? Find this in two ways: directly from the given information and also from the result of part (a).
17. An air-spaced parallel-plate capacitor is connected to a 12 V battery.

(a) If 36 μC of charge flows onto the plates, find its capacitance.
(b) If a slab of pyrex glass is inserted between the plates filling the gap, find the new capacitance of the capacitor.
(c) If the capacitor remained attached to the battery when the glass was inserted, what charge is now on the plates?

18. An air-spaced parallel-plate capacitor has an initial charge of 0.05 μC after being connected to a 10 V battery.
(a) What is the total energy stored between the plates of the capacitor?
(b) If the battery is disconnected and the plate separation is tripled to 0.3 mm, what is the electric field before and after the plate separation change?
(c) What is the final voltage across the plates and the final energy stored?
(d) Calculate the work done in pulling the plates apart. Does this fully account for the energy change in part (b)?

19. Suppose that a biological membrane is "doped" with excess surface charges so that there is the equivalent of one charge every 5 nm in a square array on each surface, with positive charge on one surface and negative charge on the other. If the resting membrane voltage is 100 mV, find the specific capacitance, the capacitance per unit area. (Hint: See the discussion at the end of Section 7.)

20. Find the electric field inside a 10 μF parallel-plate capacitor when connected to a 6 V battery if the gap between the capacitor plates is filled with air. Repeat your calculation if the gap is filled with paper with a dielectric constant of 4.

21. Suppose a biological membrane with a specific capacitance of 1 $\mu F/cm^2$ has a resting surface charge density of 0.1 $\mu C/cm^2$. Also suppose there are 50 sodium channels per μm^2 and that when each opens for 1 ms 1000 Na^+ ions flow through the channel. Find the membrane voltage 1 ms after 10% of these channels open, assuming no other changes occur during this time.

22. In a 100 μm^2 area of a muscle membrane having a density of sodium channels of 50 per μm^2 of surface area, when the sodium channels open there is a rapid flow of 1000 ions per channel across the membrane. Assuming a 100 mV resting potential, all the channels opening at once, and a membrane capacitance of 1 $\mu F/cm^2$, find the voltage change across this area of membrane due solely to the sodium ion flow.

23. The immediate cause of many deaths is ventricular fibrillation, an uncoordinated quivering of the heart as opposed to proper beating. An electric current discharged to the chest can cause momentary paralysis of the heart muscle, after which the heart will sometimes start organized beating again. A *defibrillator* is a device that applies a strong electric shock to the chest over a time interval of a few milliseconds. Assume that an energy of 300 J is to be delivered from the defibrillator, having a 30.0 μF capacitance. To what potential difference must the defibrillator be charged?

24. An alpha particle (which contains 2 protons and 2 neutrons) passes through the region of electron orbits in a gold atom, moving directly toward the gold nucleus, which has 79 protons and 118 neutrons. The alpha particle slows and then comes to a momentary rest, at a center-to-center separation $r = 9.23 \times 10^{-15}$ m before it begins to move back along its original path. (This technique is called *Rutherford Backscattering Spectroscopy* and the alpha particles are usually accelerated using a particle accelerator.)
(a) What was the initial kinetic energy of the alpha particle when it was initially far away, external to the gold atom? (Hint: Assume that the gold atom does not move because it is much more massive than the alpha particle.)
(b) Given the kinetic energy in part (a), through what potential difference was the alpha particle accelerated?
(c) How much work was done on the alpha particle in accelerating it through the potential difference in part (b)?
(d) When using a 1.1 MV tandem electrostatic accelerator, alpha particles are accelerated two times in succession (hence the tandem) and interact with the nucleus of a gold atom. Supposing that the alpha particles reach an energy of 3.3 MeV using the accelerator, what will be the minimum center-to-center separation of the alpha particle and the gold nucleus?

16

Electric Current and Cell Membranes

Thus far in our study of electricity, we have essentially confined our attention to electrostatics, or the study of stationary charges. Here and in the next three chapters we show some of the new phenomena that arise when charges move. We begin this chapter by generalizing our discussion to allow the flow of electric charges, known as an electric current, and we give a semiempirical derivation of Ohm's law. Electrical measurement methods and devices are described as an application of Ohm's law. More realistic models for a capacitor are then developed in a continued study of cell membranes in which electric charge can passively leak across the membrane. We give an overview of nerve structure and functioning and the spatial and temporal properties of the neuron membrane potential are detailed for both the quiescent and active states. The chapter concludes with a discussion of the electrical properties of individual ion channels as the underlying basis for membrane currents.

1. ELECTRIC CURRENT AND RESISTANCE

Although we have introduced the topic of membrane channels in the last chapter, we have not discussed the consequences of channels on the electrical properties of membranes. Membranes act as capacitors, storing charge and electric potential energy, but because of "leakage" of charge through channels, membranes are not the ideal capacitors treated in the last chapter. In order to discuss more realistic models for membrane electrical properties we first need to introduce some concepts related to the flow of electric charge.

Figure 16.1 shows a conducting wire attached at time zero between the plates of a previously charged air-spaced capacitor. Before the wire is connected we have already seen that there is an electric field between the plates of the capacitor, but no charge flows because the air is a good electrical insulator. As soon as the wire is connected, there will be an electric field in the wire that will drive the free electrons toward the positive capacitor plate, discharging the capacitor. The *electric current* in the wire is defined as the time rate of flow of charge along the wire

$$I = \frac{\Delta Q}{\Delta t}, \tag{16.1}$$

where the direction of the current is chosen by convention as opposite to the flow of the electrons. Thus, the electric current flows from the positive to negative plates of the capacitor in our example. The SI unit for electric current is the ampere (A), given by Equation (16.1) as 1 C/s = 1 A.

In our example, all of the net charge will travel through the wire very rapidly, resulting in a final uncharged capacitor. Clearly the electric current flowing in the wire is not constant in this situation because as the charge drains off the capacitor

J. Newman, *Physics of the Life Sciences*, DOI: 10.1007/978-0-387-77259-2_16,

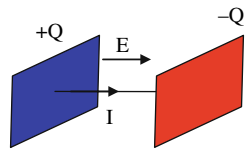


FIGURE 16.1 *Two charged conducting plates connected by a conducting wire at time zero.*

plates, the electric field that drives the electric charges decreases. If the initial charge on each capacitor plate was 1 μC and the flow of charge is complete within 1 μs, then the average electric current flowing is given by Equation (16.1) as $I = 1\ \mu C/1\ \mu s = 1\ C/s = 1\ A$. But clearly the current is not constant over this 1 μs, decreasing continuously as the charge is drained from the capacitor plates. We show below how to find the actual time dependent current flowing in this simple electric circuit.

Unlike the electric fields of previous chapters, the electric field driving the charges through the wire is not an electrostatic field. In fact, as we have seen, electrostatic fields cannot exist within a conductor. The electric field that drives the electric current, on the other hand, does exist within the conductor and is responsible for pushing the charge making up the current. This example illustrates that without a source of energy to maintain net charge on the plates of the capacitor, both the electric field in the wire and the current flow rapidly decrease to zero.

After charging the capacitor in Figure 16.1, we can think of the discharging of the capacitor as the conversion of electric potential energy to the kinetic energy of the electrons in the wire connecting the plates. As we show at the end of this section, the kinetic energy of the free electrons making up the current is then converted into heat via collisions with the metal atoms of the wire. The discharging of the capacitor occurs rapidly and therefore there is only a pulse of electric current in this case. In order to maintain a flow of electric charge, an external source of energy per unit charge, traditionally called an emf (pronounced "ee em eff," and short for the misnomer—electromagnetic force—because it is not really a force), is needed in the form of a battery or power supply.

The simple electric circuit shown in Figure 16.2 (left) consists of a battery with a uniform wire connected between its terminals. If the battery were simply a capacitor as in Figure 16.1, the initially separated positive and negative charges would quickly cancel each other out as charge flows along the wire and there would be no further change. Batteries convert chemical energy to electrical energy to continually maintain a separation of charge and supply a fixed voltage between their terminals. This is shown on the right side of the figure where the varying voltage is shown as it might be measured around the circuit, with the battery increasing the voltage each time around. A very good analogy is the flow of water due to gravity where the potential energy decreases as the water flows down hill and can only be restored by a pump of some kind, playing the role of the battery, to increase the height and thus the potential energy of the water. In our case the uniform wire of length L has a constant potential V between its ends, resulting, as we show, in a constant current flow along the wire. The constant flow of current is produced by a uniform electric field in the wire maintained by the battery and given by $E = V/L$. Electric field lines begin on the positive (+) terminal and end on the negative (−) terminal of the battery as long as the wire has no sharp bends and is smooth.

We can understand the origin of the constant current in this case by considering a microscopic picture of a collection of free electrons in the conducting wire and the forces acting on them. In the absence of an external electric field, the thermal energy of the free electrons causes them to diffuse about in a random walk traveling at very high speeds of about 10^6 m/s and making random collisions with the atoms of the metal wire (see the discussion in Chapter 2). The average velocity of the electrons, as opposed to their high speed, is zero in this case and there is no net flow of charge, therefore no electric current. When an electric field is applied, superimposed on its high-speed random walk motion, a free electron will experience an acceleration (in the direction opposite to the electric field because of the negative electric charge) given by

$$a = \frac{F}{m} = \frac{eE}{m}, \tag{16.2}$$

where e and m are the charge and mass of the electron. This acceleration lasts until the electron makes a collision with a metal atom causing it to veer off in another random direction at high speed, accelerating again according to Equation (16.2).

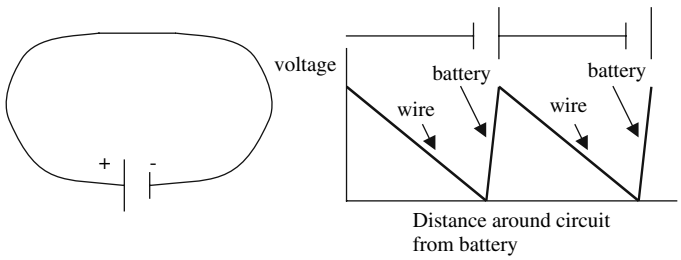


FIGURE 16.2 (left) *A battery with its terminals connected by a uniform wire.* (right) *Voltage as a function of distance around the circuit showing the decreasing voltage in the wires and the boost in voltage across the battery from chemical energy every time around the circuit loop.*

The mean time between collisions, τ, is so short that the electrons only acquire a very slow *drift velocity* of about 10^{-3} m/s given by

$$v_{\text{drift}} = at = \frac{eE}{m}\tau = \frac{eV\tau}{mL}. \tag{16.3}$$

If the number of free electrons per unit volume, or number density, in the wire is n and the wire has a cross-sectional area A, then the net free charge in a short length of the wire l is $\Delta Q = nAle$ (Figure 16.3). To find the current in the wire, we must divide ΔQ by the time required for all of that charge to move a distance l down the wire, $\Delta t = l/v_{\text{drift}}$, to find

$$I = \frac{\Delta Q}{\Delta t} = \frac{nAle}{(l/v_{\text{drift}})} = nAev_{\text{drift}}. \tag{16.4}$$

Substituting from Equation (16.3), the electric current is

$$I = \frac{ne^2\tau A}{mL}V. \tag{16.5}$$

Defining the *conductivity* σ of the wire, an intrinsic property of the material, to be

$$\sigma = \frac{ne^2\tau}{m},$$

we can rewrite Equation (16.5) as

$$I = \sigma\frac{A}{L}V = GV, \tag{16.6}$$

where G is known as the *conductance*.

Solving for V, this can be rewritten in terms of the *resistance* R

$$V = IR, \tag{16.7}$$

where

$$R = \frac{1}{G} = \frac{1}{\sigma}\frac{L}{A} = \rho\frac{L}{A}.$$

The *resistivity* of the material ρ is given by the inverse of the conductivity,

$$\rho = \frac{1}{\sigma},$$

both intrinsic parameters. This definition is made in analogy with the equality between the resistance and the inverse of the conductance

$$R = \frac{1}{G},$$

except that both of these quantities are dependent on the size and shape of the material, so that they are extrinsic parameters, unlike the intrinsic parameters depending only on the nature of the material and not on any geometric parameters.

We conclude that the current flowing in a conducting wire is proportional to the potential difference applied between the ends of the wire. This linearity of current with applied voltage (Equation (16.7)) is known as *Ohm's law*. A plot of the current through a wire as a function of the voltage across the wire is shown in curve A of Figure 16.4. The linear plot is characteristic of an ohmic (or linear) circuit element. Another equivalent statement of Ohm's law is that the resistivity of the material remains a constant, independent of the applied voltage.

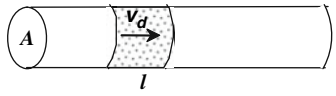


FIGURE 16.3 *Free charge in a wire of cross-sectional area A and length l traveling with a drift velocity v_d.*

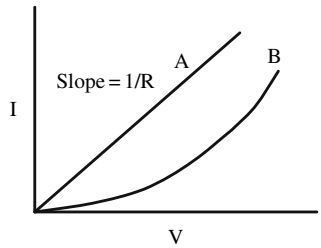


FIGURE 16.4 *The I–V curve for an ohmic circuit element (A) and a semiconductor diode (B).*

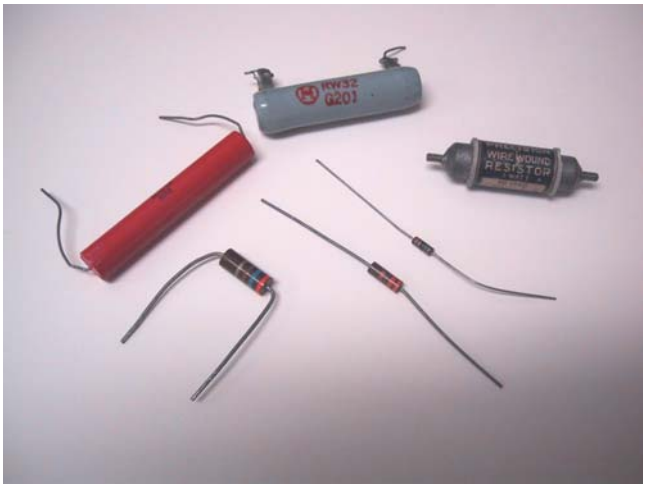


FIGURE 16.5 *An assortment of resistors.*

The SI unit for resistance is the ohm (Ω), where 1 V/A = 1 Ω (read as 1 ohm). Units for resistivity are then given as Ω-m and for conductivity as $(\Omega\text{-m})^{-1}$. The unit for conductance, the reciprocal of resistance, is the Ω^{-1} which is also known as the siemens (S). Table 16.1 lists some values for resistivity of various materials. A wire made from a metal will have a very low resistance value. For example, a 1 m length of 1 mm diameter copper wire has a resistance of only 0.02 Ω. Simple devices known as resistors (shown in Figure 16.5) are manufactured to have various resistance values. The symbol —/\/\/\— is used to represent a resistor in a schematic or circuit diagram such as the one shown in Figure 16.6. Connecting wires have negligible resistance, so that their length and shape are usually not important in a circuit diagram or in the actual circuit itself.

Table 16.1 *Resistivities of Various Materials (20°C)*

Material	*Resistivity, ρ (Ω · m)*
Conductors	
Aluminum	2.8×10^{-8}
Copper	1.7×10^{-8}
Iron	$10. \times 10^{-8}$
Mercury	$96. \times 10^{-8}$
Silver	1.6×10^{-8}
Tungsten	5.6×10^{-8}
Ionic materials	
Water (distilled)	$\sim 2 \times 10^{5}$
Fresh water	$\sim 5 \times 10^{2}$
Sea water	~0.3
Cytoplasm	~0.5
Fatty tissue	~15
Semiconductors	
Germanium	~0.5
Silicon	$\sim 2. \times 10^{3}$
Insulators	
Air (dry)	4×10^{13}
Glass	$10^{10} - 10^{14}$
Rubber	$10^{13} - 10^{16}$

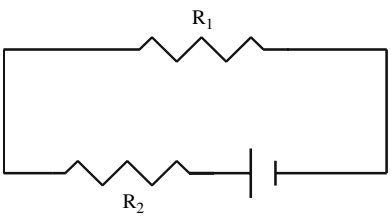


FIGURE 16.6 *A simple circuit diagram showing a battery connected to two resistors, one wired after the other.*

Example 16.1 How much electric current flows through water contained in an insulating tube 10 cm long and 5 cm in diameter when a 100 V potential difference is applied across the ends of the tube using electrodes inserted at either end? Ignore any complications from the metal electrode–water contact and do the calculation using the three entries in Table 16.1 for different purities of water.

Solution: The current that will flow is given from Ohm's law by $I = V/R$, where R is the resistance between the two electrodes supplying the 100 V potential difference. Using the relation between resistivity and resistance, and the dimensions of the water tube, we find that

$$R = \rho L/A = \rho \frac{0.1}{\pi(0.05/2)^2} = 51\rho.$$

Corresponding values are then, for distilled water, $R = 0.1 \times 10^{12}\ \Omega$ and $I = 1.0$ nA; for fresh water, $R = 2.5 \times 10^{8}\ \Omega$ and $I = 0.4\ \mu$A; and for sea water, $R = 0.15\ \text{M}\Omega$ and $I = 0.67$ mA. The huge increase in current of almost a factor of one million is due to the increase in ion content of the sea water versus fresh water versus distilled water.

Ohm's law is not a fundamental law on par, for example, with Newton's laws. It is a heuristically derived statement that the current and voltage are proportional in a conductor. Many electrical components, such as diodes, transistors, operational amplifiers, and the like, do not satisfy Ohm's law and are known as nonlinear devices (e.g., curve B in Figure 16.4). In fact, most if not all electronic devices have both resistors and nonlinear circuit elements in them.

Next we briefly consider the general topic of electrical energy and power. In the simple circuit of Figure 16.2, the battery terminals are maintained at a constant potential difference by chemical energy with the positive terminal at potential V_{battery} with respect to the negative terminal at $V = 0$. When the wire of length L is connected between the terminals, an electric current flows from the positive to negative terminal. If we plot the electric potential as a function of position along the wire (Figure 16.7), we see that it decreases linearly from the battery voltage at the positive terminal to zero at the negative terminal of the battery. A (positive) charge ΔQ flowing from the positive to negative battery terminal flows down this potential hill so that the decrease in electric potential energy is

$$\Delta \text{PE}_E = \Delta Q\, V. \tag{16.8}$$

Because in a time Δt, the charge flowing in the wire is $\Delta Q = I\,\Delta t$, the rate at which electric energy is lost is given by the *electric power P*,

$$P = \frac{\Delta \text{PE}_E}{\Delta t} = \frac{\Delta Q}{\Delta t} V = IV. \tag{16.9}$$

The SI unit for electric power is the Watt, just as for all other powers, as can be verified by substituting units for IV, $1\ \text{A} \times 1\ \text{V} = 1\ \text{CV/s} = 1\ \text{J/s} = 1\ \text{W}$.

If we examine the flow of energy in this example, stored chemical energy of the battery is used to maintain a constant potential difference between the battery terminals. This constant V produces a constant E field within the wire that, in turn, maintains a constant drift velocity for the charges. Thus, the kinetic energy of the charges remains constant along the wire, although energy is continually lost through collisions. As charge flows along the wire and down the potential hill of Figure 16.7, the potential energy loss at a rate P appears as thermal energy of the wire causing a temperature increase. This transfer of energy occurs through the collisions with the array of metal atoms in the wire while the drift velocity is maintained by the constant electric field using energy supplied by the battery. The electrical energy is said to be lost because the entire process is irreversible. As we have seen in our study of thermodynamics, a loss of potential energy of any kind to heat cannot be a truly reversible process.

Other expressions can be obtained for the power in terms of the resistance of the wire in our example. Using Ohm's law, Equation (16.7), to eliminate either V or I, we obtain

$$P = IV = I^2 R = \frac{V^2}{R}. \tag{16.10}$$

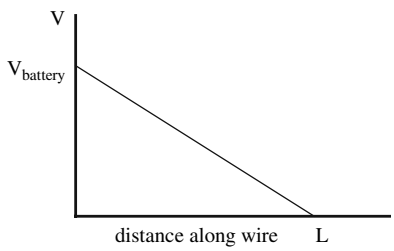


FIGURE 16.7 *The voltage, measured with respect to the negative terminal of the battery, along the uniform wire of length L in Figure 16.2.*

This conversion of electrical energy to thermal energy in a resistor is known as *Joule heating*. It is beneficially used in devices such as toasters, electric ovens, and heaters, but is a major source of energy loss in most other electrical devices. Excess heating can also be a fire hazard in poorly designed or defective house electrical wiring.

Example 16.2 Calculate the power consumption for the three situations in Example 16.1. Also, find the rate at which the water temperature increases if no heat is lost to the surroundings.

Solution: The power calculation is straightforward using, for example, V^2/R, to find powers of 0.1 μW (distilled water), 40 μW (fresh water), and 67 mW (sea water). If none of the input power is lost, it is all converted to heat in the water. The water temperature will rise at a rate determined from

$$\frac{\Delta Q}{\Delta t} = mc\frac{\Delta T}{\Delta t} = P,$$

where P is the I^2R Joule heating. The volume of water is given by $\pi r^2 L = (3.14)(.025)^2(0.1) = 2.0 \times 10^{-4}$ m^3, so that the mass of the water is about 0.2 kg, roughly independent of the salt concentration. Using a specific heat of 4180 J/(kg°C), we find rates of temperature increase of 1.2×10^{-10}°C/s (for distilled water), 4.8×10^{-8}°C/s (for fresh water), and 8.0×10^{-5}°C/s (for sea water). These heating rates are quite negligible, taking several hours to heat the sea water 1°C. However, if the tube length is decreased by a factor of 10 and the tube diameter is increased by a factor of 10, then the resistance will decrease by a factor of 1000, and both the current and power will increase by that same factor. In this case the heating is appreciable, increasing the sea water temperature by about 5°C/min.

2. OHM'S LAW APPLICATIONS AND ELECTRICAL MEASUREMENTS

Now that we have learned about electric current and resistance as well as potential, in this section we learn how to measure these in actual circuits and how to analyze some basic circuits. There are three common types of electric meters, often packaged in a multipurpose device known as a multimeter. By flipping a switch this device can measure current (as an ammeter), voltage (as a voltmeter), or resistance (as an ohmmeter). Although today these devices consist of complex semiconductor components, the fundamental principles of the devices can be more simply explained. Given a simple circuit consisting of a battery and resistor as shown in Figure 16.8, how can one use a multimeter to measure the current in the circuit, the voltages across the battery or resistor, and the resistance value of the resistor?

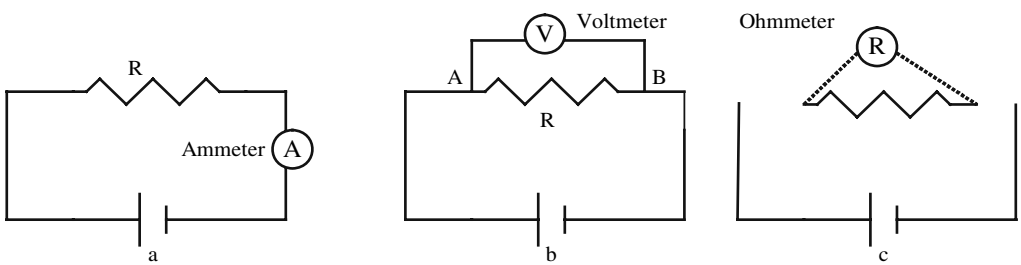


FIGURE 16.8 *Measurement of (a) the current through R, with an ammeter inserted into the circuit in series with R; (b) the voltage across R, with a voltmeter in parallel with R; or (c) the value of the resistance R itself, with an ohmmeter after removing the resistor from the circuit, as shown above.*

Any electrical measuring device has its own internal resistance that must be designed to minimize the impact of the presence of the meter on the electrical properties being measured. To measure the current in the circuit of Figure 16.8a, the multimeter must be set to act as an ammeter and be inserted into the circuit by "breaking" a wire (actually by replacing the one wire between the resistor and battery with two wires) and inserting the meter "in series" with the resistor. Being "in series" means that the same current must flow through the ammeter as flows through the resistor; there is no other path for the current to follow. However, the presence of the ammeter, with its internal resistance, affects the total resistance in the circuit and thereby the current. We would like to "analyze" this circuit; that is, we would like to write the equations that allow us to predict the current the ammeter would measure for given values of the battery voltage, resistance, and ammeter resistance.

There is a very general method to analyze circuits, even very complex ones, known as *Kirchoff's loop equation*. In this analysis, starting at an arbitrary point in the circuit diagram, one mentally "travels" around a closed loop, adding and subtracting the potential increases and decreases algebraically as the loop is traversed. The sum must add to zero because on returning to the starting position, the potential has that same starting value and thus *the potential difference around any closed loop must be zero*. In using the loop method, care is needed in choosing the proper algebraic sign for the potential difference across each circuit element. For batteries the potential increases when going from the − to + terminal across the device, whereas for resistors, the potential drops in going across the resistor in the direction of the current flow according to Ohm's law. Whichever direction one chooses to go mentally around a loop, a consistent set of potential differences must be summed to zero for the loop method to work properly. Let's continue with our analysis of Figure 16.8a; below we show the benefit of the loop equation in more complex circuit analysis.

Starting at the negative battery terminal (side with the shorter line in the symbol), we mentally "travel" around the loop clockwise (our arbitrary choice) adding and subtracting the appropriate voltages using Kirchhoff's loop equation for circuit (a) in the figure to obtain

$$V - IR - IR_{\text{ammeter}} = 0,$$

or

$$V = IR_{\text{equiv}},$$

where

$$R_{\text{equiv}} = R + R_{\text{ammeter}}. \qquad \text{(resistors in series)} \tag{16.11}$$

In this equation, V is positive because we are "traveling" from the − to + terminal, and the IR voltages across resistors are both decreases (drops), taken as negative, because we are "traveling" around in the direction of the actual current flow from the + terminal of the battery. Our answer for this circuit is actually an example of a general result when any two (or more) resistors are connected in series:

> *The equivalent resistance of resistors in series is the sum of their individual resistances.*

It also suggests that for an ammeter to have a negligible effect on the current in the original circuit, it must have a very small resistance, certainly negligible compared to the resistance in the circuit. Modern ammeters have a very low resistance, typically less than 1 Ω. Given values for V and R, the equation above predicts the measured ammeter current.

In order to measure the voltage across any component in a circuit, a multimeter is set to act as a voltmeter and needs to have its terminals connected across that circuit element as shown in Figure 16.8b to measure the voltage across the resistor. The voltmeter resistance is said to be "in parallel" with resistor R because both elements have the same potential difference across them. However, the current flowing out of

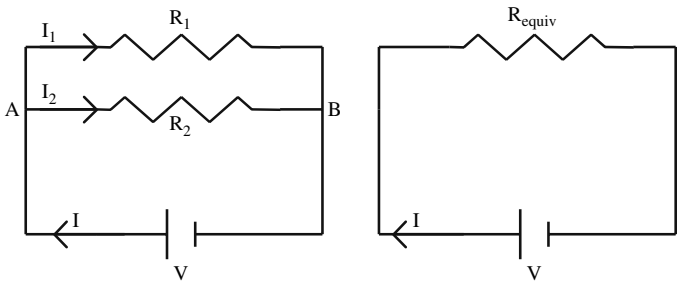


FIGURE 16.9 *A simple circuit with two resistors in parallel.*

the positive terminal of the battery, when arriving at point A in the figure, divides with part of the current flowing through each "branch" of the circuit later to recombine at point B. This is our first example of a multiloop circuit, one in which the same current does not flow through all the circuit elements, and we digress further to show how it can be analyzed.

Consider the circuit shown on the left in Figure 16.9, similar to that of Figure 16.8b because the voltmeter is represented by a resistor in parallel with the original resistor. Using Kirchhoff's loop equation to analyze this circuit, we can write down several equations depending on the chosen loop:

$$\begin{aligned} V - I_1 R_1 &= 0, \text{ clockwise around the outer loop, starting from B;} \\ V - I_2 R_2 &= 0, \text{ clockwise around the lower loop, from B;} \\ I_2 R_2 - I_1 R_1 &= 0, \text{ clockwise around the upper loop, from B.} \end{aligned} \tag{16.12}$$

Clearly these equations are not all independent, because, for example, subtracting the second from the first results in the third. An additional independent equation can be obtained by noting that at points A and B (branch points) where the current divides, by conservation of electric charge we must have that

$$I = I_1 + I_2, \tag{16.13}$$

where I is the current from the battery (see Figure 16.9 left). This is an example of a second more general rule, known as *Kirchoff's junction rule*, which states that *at a branch point (or junction) where several wires come together, the total current entering the branch point must equal the total current leaving that point*. Clearly this is a consequence of the general law of conservation of electric charge. Solving the first two equations of Equations (16.12) for each of the currents I_1 and I_2 and substituting into Equation (16.13), we can write that

$$I = \frac{V}{R_{\text{equiv}}} = \frac{V}{R_1} + \frac{V}{R_2},$$

where R_{equiv} is the single equivalent resistor that, when connected across the same battery voltage V will cause the same current I to flow from the battery (see the right side of Figure 16.9), so that $V = I\,R_{\text{equiv}}$. Dividing by V, we obtain

$$\frac{1}{R_{\text{equiv}}} = \frac{1}{R_1} + \frac{1}{R_2}, \qquad \text{(resistors in parallel)} \tag{16.14}$$

showing the general rule for resistors in parallel:

> *The inverse of the equivalent resistance of resistors in parallel is the sum of the inverses of individual resistances.*

Returning to the measurement of the voltage across a resistor in the circuit of Figure 16.8b, by putting the voltmeter in parallel with the resistor the equivalent resistance seen by the battery will change (actually, it will always decrease; can you show this from Equation (16.14)?) and therefore so will the current flowing out of the battery (it will always increase in such a circuit). The excess current will be drawn into the voltmeter loop of the circuit. The battery current is entirely determined by the "load", or equivalent resistance, on the battery from $V = IR_{\text{equiv}}$. To avoid changing the battery current significantly, the voltmeter must have a very high resistance, so that it draws negligible current and the equivalent resistance is essentially that of the circuit, R. Modern voltmeters have resistances of about 10 MΩ (1 MΩ = 10^6 Ω).

A multimeter can also function as an ohmmeter when directly connected to both sides of (across) a resistor which has been removed from the circuit, as in Figure 16.8c. By using an internal battery to send a known current through the resistor and by measuring the voltage across the resistor, the ohmmeter directly measures its resistance.

Example 16.3 Find the current that flows through each of the resistors shown in the circuit of Figure 16.10. Also determine the power generated in each resistor.

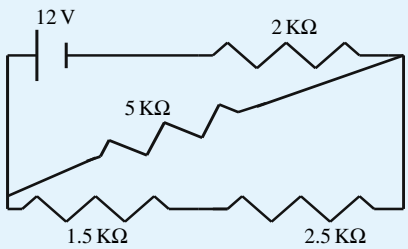
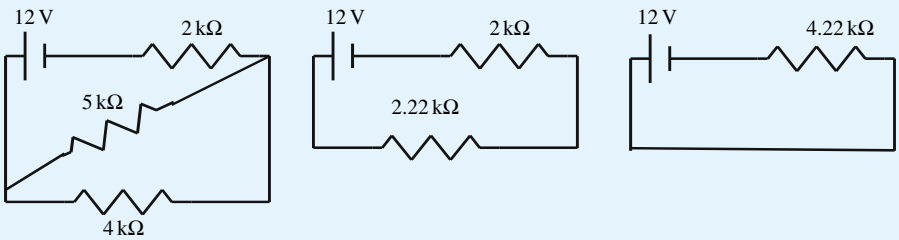


FIGURE 16.10 *Circuit for Example 16.3. Which resistors are in series or parallel with the others?*

Solution: In solving circuit analysis problems it is important to first take a careful look at the "lay of the land" or the circuit's basic "topology." In this example, the 12 V battery is the only source of current in the circuit and so it sends current out of its + terminal that then divides at the lower left branch point, some traveling through the 5 kΩ resistor and the rest traveling through the 1.5 kΩ and 2.5 kΩ resistors, which are in series with each other. The currents in these two branches (the 5 kΩ branch and the (1.5 kΩ + 2.5 kΩ) = 4 kΩ branch) recombine in the upper right corner and their sum, the net battery current, then travels through the 2 kΩ resistor and returns to the − terminal of the battery. It is very important for you to be able to understand and eventually generate this type of qualitative analysis before going to equations in order to find values for the currents.

With the understanding of the previous paragraph, we can solve this problem in a simple straightforward manner, by finding the total equivalent resistance in the circuit from the following: (1) first, the 1.5 kΩ and 2.5 kΩ are in series and together have a net resistance of 4 kΩ shown on the left below; (2) then the 4 kΩ and 5 kΩ are in parallel with each other (do you see why?), so that their equivalent resistance R is given by $1/R = 1/4\text{k}\Omega + 1/5\text{k}\Omega$, giving $R = 2.22\ \text{k}\Omega$, shown in the middle below; (3) then the 2.22 kΩ and the 2 kΩ are in series with each other yielding a net resistance in the circuit of 4.22 kΩ, shown on the right.



The circuit on the right tells us that the current out of the battery is just $I = (12\ V/4.22\ \text{k}\Omega) = 2.84 \times 10^{-3}\ \text{A} = 2.84\ \text{mA}$. All of this current passes through the 2 kΩ resistor because it is in series with the battery, but each of the other resistors only gets part of this current. To find how the current divides, we can work backwards in the set of figures just above. The current divides at the branch point so that the voltages across the 5 kΩ and equivalent 4k resistor (see the left figure above) are equal because the two branch points have a fixed potential V between them whether we "travel" through the 5 kΩ or 4 kΩ resistor. This implies that

$$V = (I_{5\text{k}}\ 5\ \text{k}\Omega) = (I_{4\text{k}}\ 4\ \text{k}\Omega)$$

(*Continued*)

so that $I_{5k}/I_{4k} = 4/5$. But we know that the total current, $I_{5k} + I_{4k}$, is 2.84 mA, so that we can find the individual currents from either the previous two equations with their two unknown currents, or from the following simple argument. By dividing the total current in 9 parts (based on the ratio equation above using $9 = 4 + 5$) we note that (4/9) of the total current, or 1.26 mA, flows through the 5 kΩ and (5/9) of the total current, or 1.58 mA, flows through the 4 kΩ equivalent resistor. Finally returning to the original circuit, each of the 1.5 kΩ and 2.5 kΩ resistors have $I_{4k} = 1.58$ mA flowing through them. You should check that these results are consistent and add up properly; follow each current around the original circuit and check Kirchoff's junction rule.

We finish this problem by noting that the power generated in each resistor is given by $P = I^2R$, so that if we know the values of the currents and resistors we can simply compute these values to be $P_{2k} = 0.016$ W, $P_{5k} = 0.0079$ W, $P_{1.5k} = 0.0037$ W, and $P_{2.5k} = 0.0062$ W. Note that the power supplied by the battery, given by $P = I_{total}V = 0.034$ W is equal to the total power dissipated in all the resistors. Check this yourself !

The preceding example was solved by simply using the rules for combining various resistors in series and parallel. There are more complex circuits where this type of analysis is not possible and Kirchoff's loop equation must be used. The next example has such a circuit.

Example 16.4 Find the current flowing though each resistor of the following circuit.

Solution: In this case, because of the second battery in the circuit we cannot simply combine resistors in series and parallel but must use Kirchoff's loop equation. Using the set of labeled currents, which can be chosen arbitrarily as long as they are consistent, we can write down two loop equations to allow us to solve for the two unknown currents labeled I_1 and I_2 in the figure. We have already implicitly used the junction equation in choosing the sum of the two currents from the batteries as the current in the central branch of the circuit. Follow the currents to the right junction point and check that they are self-consistent there as well. We need only choose two of the three possible loops: the top, bottom, or outer loops, but for practice we write all three down and then only use two of them to solve for I_1 and I_2.

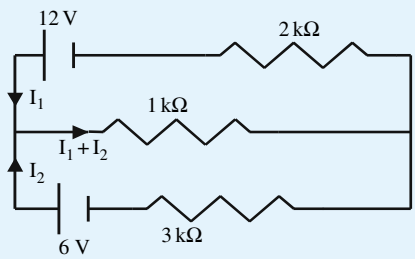


FIGURE 16.11 *Multiloop circuit for Example 16.4. Do you see why these resistors are not in series or parallel with each other?*

First around the outer loop, starting arbitrarily at the lower left corner and going clockwise, we have

$$-12\text{ V} + I_1(2\text{ k}\Omega) - I_2(3\text{ k}\Omega) + 6\text{ V} = 0.$$

Make sure you understand why the signs are as they are (these are not arbitrary). Around the top loop, starting at the upper left corner and still going clockwise (note: the direction is arbitrary, but it is perhaps a good idea always to "travel" around loops the same way to help reduce mistakes)

$$-12\text{ V} + I_1(2\text{ k}\Omega) + (I_1 + I_2)(1\text{ k}\Omega) = 0.$$

Finally, although not needed, around the bottom loop, again clockwise from the lower left corner,

$$-(I_1 + I_2)(1\ \text{k}\Omega) - I_2(3\ \text{k}\Omega) + 6\ \text{V} = 0.$$

Now, picking any two of these three equations, we need to do the algebra to solve for the two unknowns. We find that $I_1 = 3.82$ mA and $I_2 = 0.55$ mA. Check this for yourself.

We can also consider simple electrical circuits that have two capacitors C_1 and C_2 connected either in series or in parallel to a battery as shown in Figure 16.12. As just studied in the case of resistors, there will be a single equivalent capacitor that, when connected to the same battery, will produce the same resulting final state: the same charge will flow from the battery, storing the same amount of potential energy as in the original situation with two capacitors. In the next section we show the effects of having both resistors and capacitors in the same circuit, but first we complete this section by calculating the equivalent capacitance corresponding to those equations for the equivalent resistance of series and parallel resistor combinations, Equations (16.11) and (16.14).

Consider the case of two capacitors in series as shown on the left in Figure 16.12. Using the fact that the voltage across a capacitor is proportional to the charge on it, we have that $V_1 = Q_1/C_1$ and $V_2 = Q_2/C_2$, where the charges are those on each capacitor. Now, consider the portion of the circuit outlined in the dotted lines. This section of the circuit is completely isolated electrically and if it was originally neutral must remain so. Therefore the net negative charge on the right plate of C_1 and the net positive charge on the left plate of C_2 must add to zero, proving that $Q_1 = Q_2$. Then, using the loop equation, the voltage V across the battery is equal to the sum of the voltages V_1 and V_2 across each capacitor and we have that

$$V = V_1 + V_2 = \frac{Q}{C_1} + \frac{Q}{C_2}, \tag{16.15}$$

where Q is the common charge on each capacitor. The battery supplies positive charge Q to the left plate of C_1 which then induces an equal negative charge on its adjoining right plate, resulting in an equal and opposite positive charge at the left plate of C_2 and an induced equal negative charge on its right plate. We show in the next section that this "charging" of the capacitors when first connected to a battery takes some finite time, depending on the stray electrical resistance of the circuit. Finally, we see that if we replace the two capacitors by a single equivalent capacitor with capacitance C, that in order to have the same charge stored on this capacitor we require that

$$V = \frac{Q}{C} = \frac{Q}{C_1} + \frac{Q}{C_2} \quad or \quad \frac{1}{C} = \frac{1}{C_1} + \frac{1}{C_2}. \qquad \text{(capacitors in series)} \tag{16.16}$$

Capacitors in series combine reciprocally, just as resistors in parallel do according to Equation (16.14).

Using a similar analysis for capacitors in parallel, we see from the right-hand portion of Figure 16.12 that we now have that the total charge Q supplied by the battery is the sum of the charges on both capacitors: $Q = Q_1 + Q_2$. From this, we can write

$$Q = Q_1 + Q_2 = C_1V_1 + C_2V_2, \tag{16.17}$$

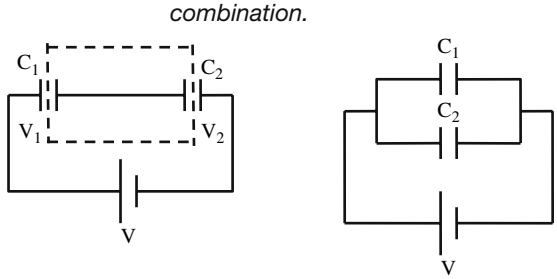


FIGURE 16.12 *Two capacitors in a* (left) *simple series or* (right) *parallel combination.*

Starting from Equation (16.19), and substituting from the definition of

$$I = -\frac{dQ}{dt}$$

(the minus sign is needed to make the current positive because it is equal to the time rate of decrease of the capacitor charge), the equation becomes

$$R\frac{dQ}{dt} + \frac{Q}{C} = 0.$$

Rewriting, we have

$$\frac{dQ}{Q} = -\frac{dt}{RC}.$$

Integrating both sides of this equation from $t = 0$ to time t and from $Q(t = 0) = Q_0$ to a value of $Q(t)$, written simply as Q, we find

$$\int_{Q_0}^{Q} \frac{dQ}{Q} = -\frac{1}{RC}\int_0^t dt',$$

so that

$$\log Q - \log Q_0 = \log\frac{Q}{Q_0} = -\frac{t}{RC}.$$

Taking the antilog of both sides, remembering that these logarithms are to the base e, we find

$$\frac{Q}{Q_0} = e^{-\frac{t}{RC}},$$

or Equation (16.20a). To then find the current as a function of time, we again use its definition, so that

$$I = -\frac{dQ}{dt} = -Q_0\frac{d(e^{-\frac{t}{RC}})}{dt} = \frac{Q_0}{RC}e^{-\frac{t}{RC}},$$

or Equation (16.20b). This same procedure can be used to analyze any electrical circuit consisting of batteries, capacitors, and resistors via the loop equation.

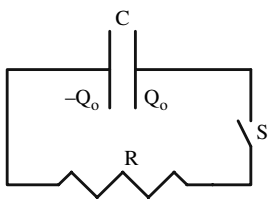


FIGURE 16.13 *An RC series circuit with the capacitor initially charged before closing the switch S connected to the resistor.*

and again replacing the two capacitors with a single capacitor C and noting that the voltages across each capacitor are the same because they are in parallel ($V_1 = V_2 = V$), we find

$$Q = CV = C_1V + C_2V \quad or$$

$$C = C_1 + C_2. \qquad \text{(capacitors in parallel)} \qquad (16.18)$$

Remember that, just as for resistors, these results for combining two capacitors in series or parallel can easily be generalized to larger arrays of capacitors using the same tools as in the above discussion. Circuits with only resistors or only capacitors present are ideals. In the next section we turn to a presentation of more realistic circuits with both resistors and capacitors present. Such circuits are more realistic because there is always a small amount of resistance (in the conducting wires themselves) or stray capacitance (between different conducting surfaces) present in any circuit regardless of whether an actual resistor or capacitor device is present in the circuit. We approach this topic using a model for cell membranes.

3. MEMBRANE ELECTRICAL CURRENTS

In the last chapter membranes were considered as ideal capacitors with a specific capacitance (capacitance per unit area) of about 1 $\mu F/cm^2$. This turns out to be a very good approximation for a pure phospholipid bilayer which has an extremely high resistivity of about 10^{15} Ω-cm, comparable to a very good insulator. The very high equivalent resistance prevents charge from crossing the lipid region and maintains the stored charge as if the bilayer were an ideal capacitor. However, as discussed in the last chapter, biological membranes are full of proteins that act as channels allowing ionic currents to flow across a membrane.

The simplest model, or equivalent circuit, for a biological membrane in the resting state is shown in Figure 16.13 and is known as an *RC series circuit*. For now, we ignore how the equivalent capacitor was charged (to a voltage $V_0 = Q_0/C$) and we imagine that at time zero the switch S is closed (corresponding to the membrane channels opening), discharging the capacitor. The capacitor does not discharge instantaneously, but follows a time course that depends on the values of R and C. The resistance R represents the effective resistance to current flow across the membrane and is discussed further below.

To analyze this circuit, we use Kirchhoff's loop method, discussed in the last section. Let's write a loop equation for the circuit in Figure 16.13 after the switch is closed and a path is provided for current flow. When the switch is closed current will flow from the $+Q_0$ side of the capacitor clockwise around the circuit. Starting at the switch S and mentally going clockwise around the loop, we find

$$-IR + \frac{Q}{C} = 0. \qquad (16.19)$$

Because both Q and I vary with time, it turns out that we need calculus to solve this equation (see box) to find that the charge on the capacitor and the current through the resistor are given by

$$Q = Q_0\, e^{-\frac{t}{RC}}, \qquad (16.20a)$$

$$I = I_0\, e^{-\frac{t}{RC}}, \qquad (16.20b)$$

where Q_0 is the initial charge on the capacitor and I_0 is the initial current when the switch is closed and given by $I_0 = Q_0/RC$.

The results obtained in Equations (16.20) are shown in Figure 16.14 with the charge and current plotted as functions of time. Because the voltage across the capacitor is proportional to the charge ($V = Q/C$) and the voltage across the resistor is also proportional to the current ($V = IR$), these voltages follow the same time courses as Q and I, respectively. The key parameter in these results is the product RC, which has units of time and is known as the *RC time constant* $\tau = RC$. Its value determines the rate at which the discharging of the capacitor occurs, with the charge, current, or voltage across either R or C dropping to $(1/e) = 0.37$ of its initial value in a time $\tau = RC$ (see Figure 16.14).

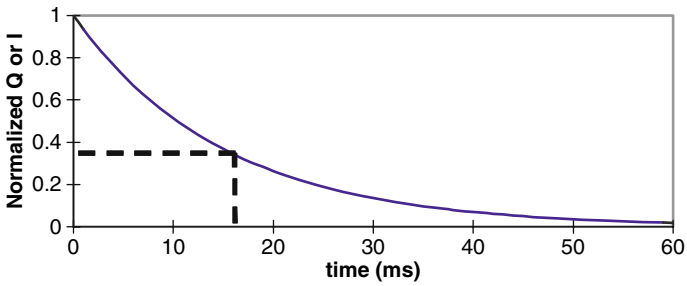


FIGURE 16.14 *Normalized capacitor charge or electric current in an RC circuit (Equations (16.20a) and (16.20b), normalized to their initial values) for a $\tau = 15$ ms RC time constant. The voltages across the capacitor and resistor also follow the same time course. Dashed lines indicate that at $t = \tau$, the normalized Q or I has decreased to (1/e) = 0.37 of its starting value of 1.0.*

All electrical devices and complete circuits have some associated capacitance as well as resistance. In high-speed electrical applications, such as computers, the RC time constant sets fundamental limits on the speed at which a circuit can change its voltage. Computers use voltage as information, with a high or low voltage representing a bit of information, either a 1 or a 0, and calculations are done by electronic arithmetic that changes bits rapidly. Consequently, increasing the processing speed of a computer depends heavily on reducing the associated capacitance of the fundamental electronic device building blocks of the microprocessor.

Example 16.5 In the simple RC circuit of Figure 16.13, the 10 μF capacitor is initially charged to 60 μC. When the switch is closed, an initial current of 0.3 mA is measured in the circuit. Find the charge on the capacitor and the current in the circuit after 0.6 s.

Solution: To learn the time course of the current and charge, we need to first find the value of the resistance in the circuit. When the switch is first closed, the initial voltage across the resistor is the full initial voltage V_0 across the capacitor. Because the initial charge on the capacitor is 60 μC, the initial voltage is $V_0 = Q_0/C = 60\ \mu C/10\ \mu F = 6$ V. This voltage, on closing the switch, immediately produces the given initial current flow $I_0 = 0.3$ mA. From Ohm's law $R = V/I$, so knowing the initial current we can solve for $R = (6\text{ V})/(0.0003\text{ A}) = 20\ k\Omega$. Now, knowing $RC = (20\ k\Omega)(10\ \mu F) = 0.2$ s, we can use Equations (16.20) to find the charge and current after 0.6 s, equal to three time constants. Substituting that $(t/RC) = 3$, we find that the exponential is given by $e^{-3} = 0.05$, so that after 0.6 s there will remain only 0.05 times the initial charge and current. Our answers then are that after 0.6 s there remain $(60\ \mu C)(0.05) = 3\ \mu C$ of charge and the current is $(0.3\text{ mA})(0.05) = 15\ \mu A$.

Let's now apply some of these ideas to a biological membrane where we are particularly interested in the transverse currents across the membrane. For membranes in the resting state, RC time constants range from 10 μs to 1 s. In dealing with membranes it is useful to discuss the electrical properties of a 1 cm^2 area; these are known as the specific capacitance C/A, and specific resistance RA. Defined in this way the product of the specific capacitance and specific resistance $(C/A)(RA) = RC$ is still equal to the time constant. From $R = \rho L/A$, we have that $RA = \rho L$ in units of Ω-cm^2. Using the value quoted for the membrane specific capacitance C/A, in the previous chapter of 1 μF/cm^2, the different time constants correspond to different values for the specific resistance $RA = \rho L$ of 10 to 10^6 Ω-cm^2. The broad range of values for the resistivity indicates a large variability in both the numbers of channels per unit area and in the average number of open channels in the resting state in different cells.

We now want to get some estimate of the numbers of charges flowing through each open channel that make up the membrane current. Using a value of 0.1 V for the resting potential, we determined in the last chapter that a typical value for the surface charge density Q_0/A is about 0.1 μC/cm^2. Because 1 mol of a monovalent ion corresponds to a charge of $\mathcal{F} = N_A e = 6 \times 10^{23} \times 1.6 \times 10^{-19} \approx 10^5$ C/mol, where $\mathcal{F}$ is known as the Faraday constant, we can find the number of moles corresponding to a charge Q_0 per unit area. If due to monovalent ions, the surface charge density corresponds to

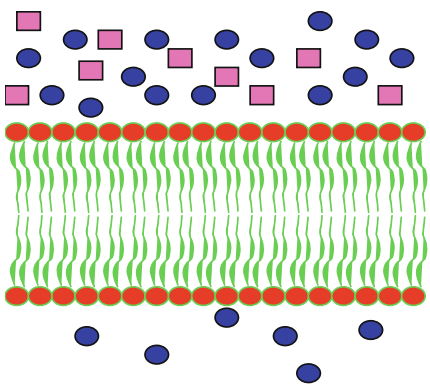
$$(10^{-7}\ \text{C/cm}^2)/\mathcal{F} = 10^{-12}\ \text{mol/cm}^2 = 1\ \text{pmol/cm}^2.$$

If we approximate the average current density (current per unit area) by dividing the charge density value by the time constant, we find a current density I/A of 100 μA/cm^2 using a 1 ms time constant. This corresponds to the flow of 1 nmol of ions/cm^2/s. Using a value of about 10 channels/μm^2 (or 10^9 channels/cm^2) for the surface density of channels, the ratio of I/A (10^{-4} A/cm^2) to channels/cm^2 gives a value for the current in a single channel of about 0.1 pA, corresponding to the flow of about 10^{-18} mol of ions/s. This means that each channel carries about 600,000 ions/s or about 600 ions in the 1 ms time constant. Measured values for a variety of single channels give currents of this magnitude or 10–100 times larger (see Section 6). Note that the number of ions flowing across the membrane is insignificant in terms of the total concentrations of ions both in the cytoplasm and extracellular medium, so that the ion concentrations in these media remain essentially constant.

Thus far in our discussion we have ignored the membrane charging mechanism, or in the language of equivalent circuit diagrams, we have ignored a source of energy, a battery or power supply. What is the origin of the membrane resting potential? We show that the selective permeability of the membrane to various ions, controlled by the channels, is the source of this potential.

Suppose first that there are only K^+ channels in a membrane so that, to a good approximation, only those ions can cross the membrane barrier. If we start with an excess of KCl on one side of the membrane, the K^+ will reach an equilibrium across the membrane in which there is no net flow of ions even though the K^+ concentration is not equal on both sides of the membrane. Why is this? Clearly in the absence of any electrical effects, diffusion alone would tend to drive the K^+ concentration to the same final value on both sides of the membrane. However, despite this diffusional driving force, electrical attractive forces due to the presence of the excess (negative Cl^-) ions, which cannot cross the membrane, balance this tendency toward a uniform concentration at equilibrium (Figure 16.15).

FIGURE 16.15 *Portion of a membrane (with channels not shown) permeable only to K^+* (blue) *showing that even at equilibrium, the concentration of K^+ is higher on the side with Cl^-* (pink) *due to electrical forces.*



From an equilibrium equation similar to that of the discussion of Figure 13.6 in Chapter 13, we can write that

$$\frac{c_0}{c_i} = \mathrm{e}^{-\frac{PE_0 - PE_i}{RT}}, \tag{16.21}$$

where R is the molar gas constant, and the c's and PE's are molar concentrations and potential energies, respectively, of the K^+ on the outside (o) and inside (i) of the membrane. Writing that $\mathrm{PE}_0 - \mathrm{PE}_i = N_A q \Delta V = z\mathcal{F}(V_0 - V_i) = z\mathcal{F}V_K$, where N_A is Avogadro's number, z is the valence or number of charges per ion (so that $z\mathcal{F}$ is the charge of a mole of ions), and V_K is the equilibrium membrane potential due to potassium ions. Solving for V_K by taking the natural logarithm of Equation (16.21), we have

$$V_K = \frac{RT}{z\mathcal{F}} \log\left(\frac{c_0}{c_i}\right). \tag{16.22}$$

Equation (16.22) is known as the *Nernst equation* and determines the equilibrium membrane potential contribution from the imbalance of a particular ion, known as the *Nernst potential*. Table 16.2 gives typical concentrations and Nernst potentials for Na^+, K^+, Ca^{2+}, and Cl^-.

Table 16.2 *Typical Ion Concentrations and Nernst Potentials (Mammalian Skeletal Muscle)*

Ion	*Typical Internal Concentration (mM)**	*Typical External Concentration (mM)*	*Nernst Potential (mV)*
Na^+	12	145	+67
K^+	155	4	−98
Ca^{2+}	10^{-4}	1.5	+129
Cl^-	4	120	−90

* 1 mM = 10^{-3} M = 10^{-3} mol/L.

The Nernst potential represents the equilibrium situation for a particular ion species. If the transmembrane potential is equal to the Nernst potential for some ion species "A," V_A, then there will be no net flow of A across the membrane even if the membrane has a high conductivity for A. No net flow does not mean that the channels do not allow any ion flow, but rather that the inward and outward flows of ion A are equal. If the transmembrane potential is higher or lower than the Nernst potential then there will be a net flow of A one way or the other across the membrane with the ionic current proportional to the difference between the actual potential and the Nernst potential for that ion

$$I_A = G_A\,(V - V_A), \tag{16.23}$$

where G_A is the A ion conductance and V is the actual transmembrane potential. If only the one ion species can cross the membrane, then the membrane potential will equilibrate at the Nernst potential for that ion. In the resting state, open K^+ channels dominate and the resting potential is close to the equilibrium potential for K^+, −0.1 V. This behavior is identical to that expected if there were a battery in series with a resistor for each ion species. These separate batteries across the membrane function when their corresponding channels are open, corresponding to when their series resistance decreases.

At this point in our discussion we can present a more realistic circuit diagram for a membrane than a simple *RC* circuit. In the membranes of the axons of neurons, Na^+ and K^+ channels dominate, and Hodgkin and Huxley proposed the equivalent circuit shown in Figure 16.16. The arrows through the resistors in the figure indicate conductances that can vary with time as the ionic channels are made to open or close (known as gated channels). Only Na^+ and K^+ channels are explicitly indicated with a net leakage conductance representing other net ion flows. Before we study some of the electrical properties of neurons and this equivalent circuit representation in Section 5, we first give a more qualitative overview of the structure and functioning of neurons and the ways in which their electrical properties have been studied.

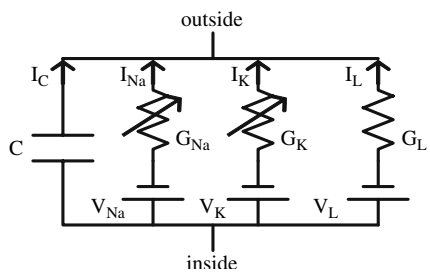


FIGURE 16.16 *The Hodgkin–Huxley equivalent circuit for an axon membrane. The batteries represent the specific ion Nernst potentials (L = leakage, representing the small contribution from other ions), producing specific ion currents as shown. The total membrane current is given by the sum of the four currents listed with the capacitor current equal to (from Q = CV)*

$$I_C = C\frac{\Delta V}{\Delta t},$$

where V is the voltage across the membrane.

4. OVERVIEW OF NERVE STRUCTURE AND FUNCTION; MEASUREMENT TECHNIQUES

The human nervous system consists of some 10^{11} nerve cells, or neurons, each one making an average of over 1000 interconnections. On an individual level we have a reasonable understanding of the functioning of a single nerve cell,

but we have precious little knowledge of the larger-scale, or more global functioning, of our nervous system. Three main ways to categorize nerve cells include whether they are part of the central (brain + spinal cord) or peripheral (all else) nervous systems, part of the autonomic (connections with involuntary muscles and internal organs) or somatic (peripheral connections to voluntary muscles and surface sensors) nervous systems, or whether they are afferent (so-called sensory neurons, carrying information from the peripheral to the central nervous system) or efferent (so-called motor neurons, carrying information in the opposite direction). There are many different types of neurons, however, they all have common features and are believed to function in a very similar manner.

Neurons are single cells with a cell body containing a nucleus and usually a single long thin structure, the axon, which may be more than 1 m in length. There are also several shorter processes, known as the dendrites, radiating away from the cell body (Figure 16.17). Cell bodies tend to be clustered together in regions connected by bundles of axons. At the far end of the axon are the terminal endings.

Nerve cells conduct an electrical signal called the action potential, or nerve impulse, discussed in detail in the next section. These signals are very similar in all nerves, traveling from the dendritic end to the terminal bundle end at speeds of up

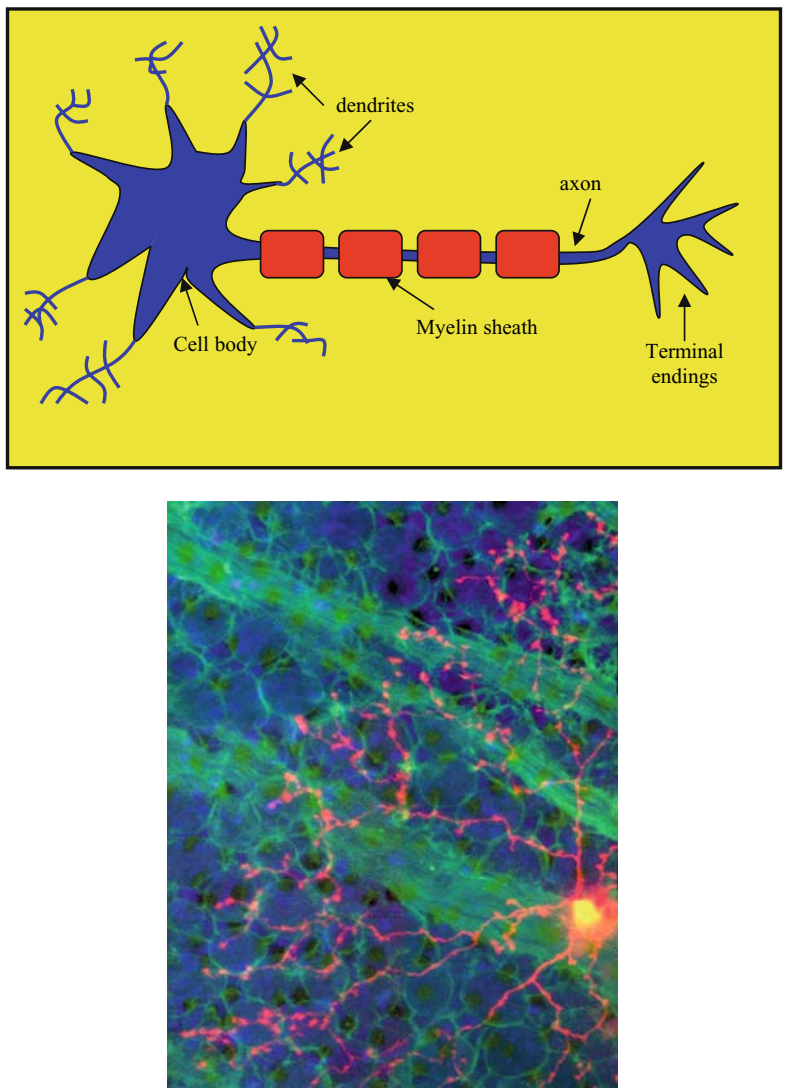


FIGURE 16.17 *Structure of the neuron* (top) *schematic;* bottom *multiphoton scanning microscopy view of nerve bundles* (green) *and a retinal "starburst" cell* (red) *found in visual processing network.*

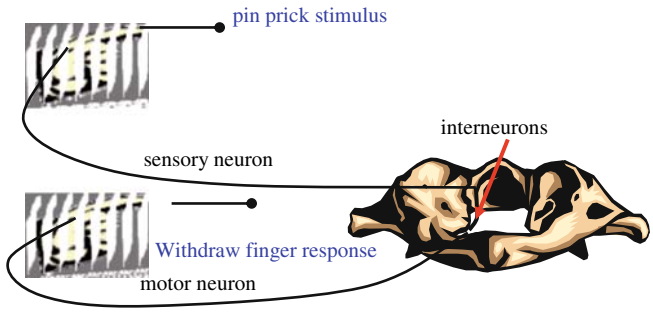
to 100 m/s. Usually each neuron is electrically isolated from the next and signals are passed on to the next cell chemically. This occurs through the release of a neurotransmitter from synaptic vesicles at the terminal endings. These chemicals diffuse across the synapse, a small cleft between the terminal endings of one neuron and the dendrites of the next, and are detected by membrane receptors on the dendrites to provoke an electrical response. Receptors are membrane bound proteins that, on binding neurotransmitters either directly (through so-called ligand-gated channels) or indirectly through open ion channels, cause a membrane depolarization and a continuation of the action potential. In certain neurons direct electrical connections between neighboring cells occur via "gap junctions," pores connecting two neighboring cells that allow the direct passage of very small molecules. These are commonly found in embryo tissue and are believed to provide a means for cell–cell communication in undeveloped tissue. In nerve cells, however, gap junctions do not allow as great a variety of control mechanisms as chemical synapses do, and are therefore relatively rare.

It is useful to describe the overall circuitry involved in a simple reflex response. At a minimum such a response requires four cells. The knee jerk reflex is well known as a simple reflex involving a muscle fiber, a receptor transducer cell, a sensory neuron, and a motor neuron. When a doctor taps the patellar tendon near the knee, the attached muscle is stretched. A stretch receptor senses this and produces an electrical response that is carried by an action potential along a sensory neuron to the spinal cord. There a reflex response is generated as an action potential in a motor neuron returning to the same muscle fiber. Arrival of this action potential generates a sequence of chemical steps that result in the contraction of the muscle, and the knee jerk response. A similar sequence of events occurs when you respond to a pinprick on your finger (Figure 16.18). Of course this is a simplistic view, and there are other neural connections that allow control over the sensory and motor signals from the central nervous system as well, but it serves to give a picture of the overall circuitry in a simple reflex.

Electrical properties of individual neurons can be studied in living tissue using inserted microelectrodes. Most of the early research work was done using the giant axon from a squid, a particularly large cell with an axon of about 1 mm in diameter. The electrode is a glass capillary tube containing a conducting salt solution and a metal wire electrode. Electrodes are used both to measure membrane voltages (with the wire inside the tube connected to a sensitive voltmeter) and to inject small amounts of current (with the wire attached to a power supply). Usually the microelectrode is set to zero potential in the extracellular medium and, when inserted through the membrane into the cell, reads the resting membrane potential, typically a small (0.1 V) negative voltage with respect to the outside. When used to study a nerve impulse, often current is applied through a second electrode as a stimulus and subsequent changes in potential are measured. Alternatively, a constant voltage step change could be applied, fixing the membrane potential, and the changes in current flow across the membrane measured. This method is known as the *voltage-clamp* technique.

On first thought, one might guess that the membrane could be voltage-clamped by connecting an ideal battery across its thickness. The battery would supply whatever current was needed to offset the membrane currents in order to maintain a fixed membrane potential. This is, however, not quite true because the battery terminals cannot be "attached" to the membrane and there are unpredictable junction potentials at the metal–solution boundary due to contact resistance that would vary with the current flow. Only the metal electrodes would be voltage-clamped, not the membrane itself. Instead, voltage-clamping involves using an *electric feedback loop* to continually inject small currents in order to maintain a fixed potential.

FIGURE 16.18 *A simple reflex circuit.*



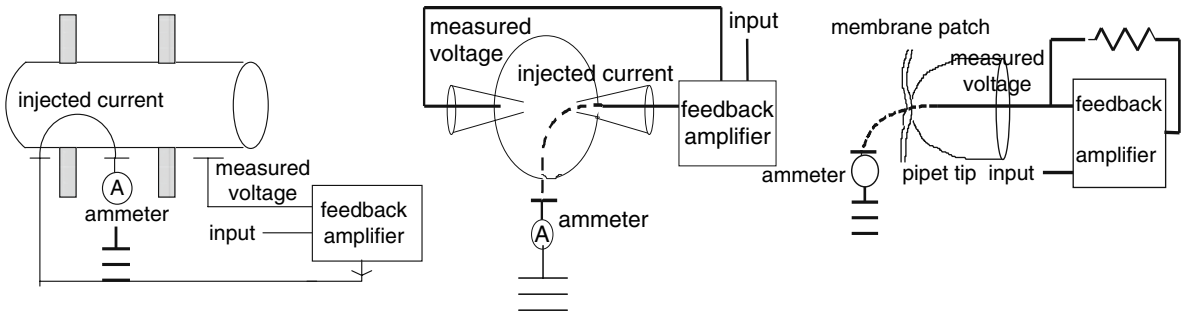


FIGURE 16.19 *Three types of voltage-clamps. From left to right: gap method with insulating dividers, double electrode method for cells, patch-clamp method for pieces (patches) of membrane.*

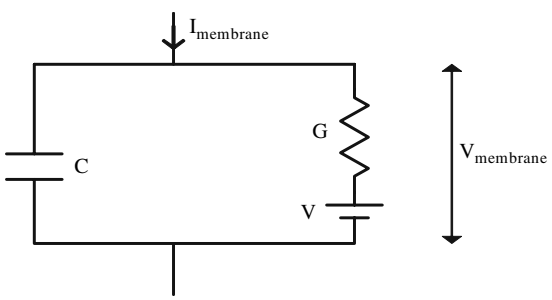
Figure 16.19 shows three examples of voltage-clamp circuitry using feedback loops. In each method, the membrane potentials are "*space-clamped*" in such a way as to have no spatial variation of potential. In two of these methods two electrodes are used, with one measuring the potential relative to a reference voltage set at the desired level. This voltage difference signal is then used to inject a current through the second electrode to reduce the difference signal and maintain the voltage clamp. Such a procedure is an example of negative feedback, in which an "error signal" is sent back to the source and used to make small corrections so as to restore a desired value of a variable. The space-clamping is achieved by either using long intracellular electrodes or by using a small membrane area isolated by either applying insulators in gaps dividing the membrane or by a patch-clamp arrangement. *Patch-clamping*, developed in 1976, uses a micron-diameter pipette tip pressed against an intact cell with some suction applied to form a very tight seal on a microscopic area of membrane so that the resistance between the inside and outside solutions is many GΩ (1 GΩ = 10^9 Ω). Patch-clamping has led to a 100-fold increase in the sensitivity of membrane current measurements (see Section 6 below).

5. ELECTRICAL PROPERTIES OF NEURONS

When several electrodes are used to probe the spatial pattern of normal membrane potentials it is found that small cells have membrane electric potentials that are constant over their entire surface whereas larger cells, such as neurons, can have potentials that vary spatially as well as temporally. Although a small cell's membrane can be reasonably modeled by a simple single-loop circuit diagram, Figure 16.20, in which the membrane voltage and current values depend on time, but not on spatial location (a so-called *lumped-parameter model*), neurons cannot.

Modeling the electrical properties of a neuron requires a so-called *distributed-parameter network*. The simplest scheme for a neuron that leads to some useful results is a *linear cable model* shown in Figure 16.21. This ribbon of repeated circuit elements is characterized by a set of parameters that vary along the length x. Here the inner and outer conductors represent the intracellular and extracellular fluid. Each section of length Δx along the cable has per-unit-length values of membrane capacitance c_M, conductivity g_M, transverse (inner to outer) current I_m, and inner and outer longitudinal resistance r_i and r_0, as well as inner and outer values for longitudinal current along the axon I_i and I_0, and voltage difference across the membrane V_M. The model was first developed to represent an electrical cable (hence the name) that leaks some current transversely across the insulation between the two co-axial conductors. Although the mathematics of this model is complex, it is based on a

FIGURE 16.20 *Equivalent circuit for the membrane of a small cell with no spatial variation in its electrical parameters.*



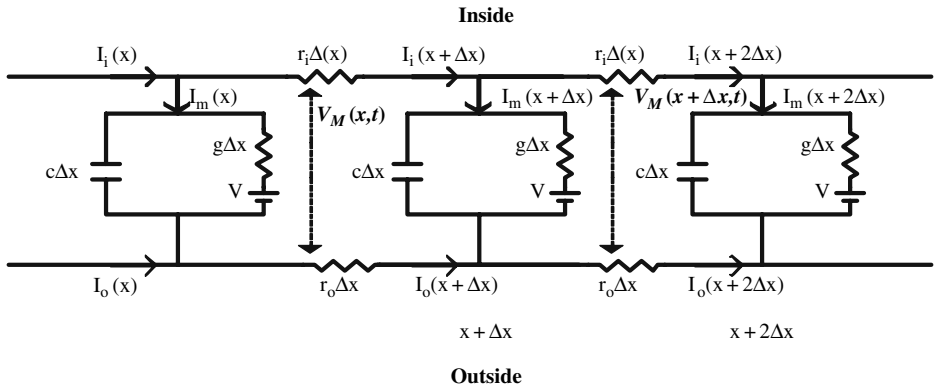


FIGURE 16.21 *A cable model for the electrical properties of the membrane of a nerve axon. There are two parallel conductors along the inner and outer surfaces with repeated transmembrane circuit elements representing the local current-voltage characteristics that vary with position.*

straightforward application of Kirchhoff's rules. Here we are content with showing a few of the model's predictions.

Two parameters of the model are needed: the RC ($= C/G$) time constant, given by

$$\tau_M = \frac{c_M}{g_M}, \tag{16.24}$$

and the space constant given by

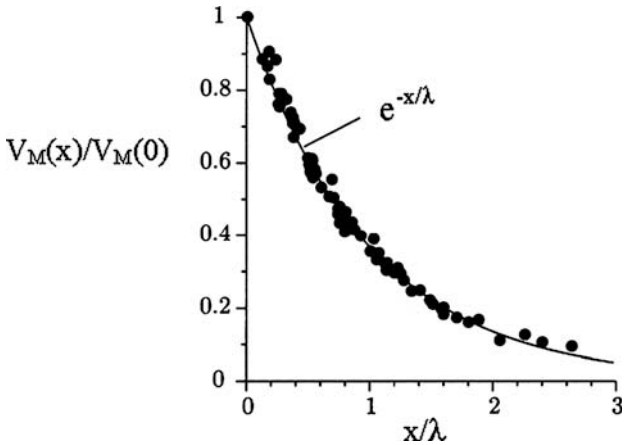
$$\lambda = \frac{1}{\sqrt{(r_i + r_0)g_M}}. \tag{16.25}$$

Note how the units work out in Equation (16.24), with both c_M and g_M per-unit-length constants so that their ratio has time units, whereas in Equation (16.25) the per-unit-length constants combine to give λ units of distance. The time constant is a property solely of the membrane with typical values of several ms, whereas the space constant depends also on the cell dimensions and geometry and has typical values of several mm. If a steady electric current is applied at one point ($x = 0$) along a neuron, the membrane voltage difference V_M from the resting potential decreases exponentially along the axon in either direction according to

$$V_M = V_M(0)\mathrm{e}^{-\frac{|x|}{\lambda}}, \tag{16.26}$$

as shown in Figure 16.22. After a brief initial time when the current is applied, this result is time-independent because current is continually injected by the electrode to achieve a steady state.

FIGURE 16.22 *The spatial variation in the membrane voltage from measurements along axons stimulated by a small current from an electrode at x = 0.*



If, on the other hand, a short pulse of current is injected into an axon at $x = 0$ at time zero, the model can be used to calculate the voltage response as a function of both position and time. This situation corresponds to a typical stimulation of a nerve or muscle membrane. Results for this model are plotted in two ways in Figure 16.23. On the left the spatial variation of the voltage response is shown for several different times (different curves). At increasing times the response spreads out from $x = 0$, decreasing in amplitude at $x = 0$, but increasing in amplitude at other locations for a brief time. This is perhaps better shown in the figure on the right where the time-dependence is plotted at several different distances from $x = 0$ (given in units of λ). The voltage rises and then falls with an exponential tail. The peak can be seen to move to farther locations at later times, but with a rapidly decreasing amplitude. If

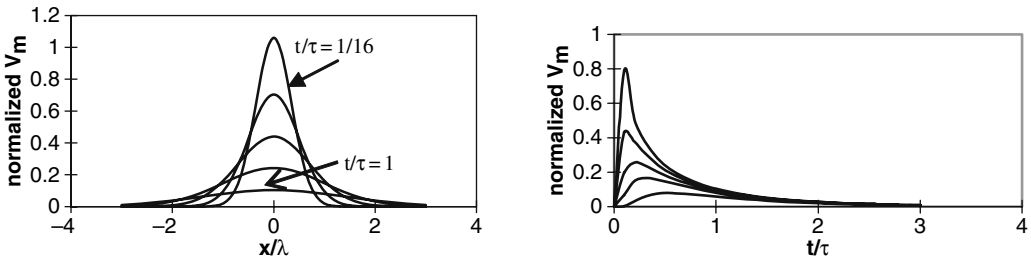


FIGURE 16.23 (left) *Spatial dependence of spreading membrane voltage at various times (decreasing voltage curves at x = 0 correspond to t/τ = 1/16, 1/8, 1/4, 1/2, and 1);* (right) *Time-dependence of membrane potential at various distances from the stimulus at x = 0 (decreasing peak voltage curves are at x/λ = 0.5, 0.75, 1.0, 1.5 and 2).*

the potential changes are below a threshold value, this will be the only response of the membrane, a localized brief signal. Data on so-called miniature end-plate potentials, due to spontaneously released neurotransmitters, are accurately modeled by the cable model. On the other hand, if the potentials exceed a threshold value, then a totally different type of behavior is observed: a nonlinear nerve pulse is initiated.

A nerve pulse, or *action potential*, is an all-or-nothing propagating potential wave that is the basis of all neural communication. The *Hodgkin–Huxley (H-H) model* is a generalization of the cable model in which the cross-membrane elements of the cable are spelled out in detail. In place of a single conductance channel, H-H uses three such paths, for K^+, Na^+, and for other leakage currents, with the conductances for Na^+ and K^+ given as variable conductances (shown with arrows through their equivalent resistor values in Figure 16.16). This latter change makes the entire problem nonlinear because the conductances for Na^+ and K^+ are now themselves functions of both membrane voltage and time. From Equation (16.23), we see that the ionic currents will now depend on the membrane voltage in some nonlinear way (with the exponent of V_M not equal to 1).

The crux of the H-H model is the specification of the conductances G_{Na} and G_K. Hodgkin and Huxley obtained these functions by fitting data from space-clamped measurements (eliminating the x-dependence, or the cable properties) that were also voltage-clamped, allowing direct measurement of membrane currents. Individual membrane currents due to Na^+ and K^+ were measured by a number of methods, including radioactive labeling of the salt ions, or using channel blockers, specific chemicals that block, or shut off, only one type of ion channel. From numerous measurements of currents at specific membrane voltages, plots of the conductances of each type of channel as functions of potential were obtained. With empirical equations for these conductances, the H-H model can account for all of the features of an action potential.

FIGURE 16.24 *Membrane voltage changes during an action potential* (bold)*, together with sodium and potassium ion conductances across the membrane.*

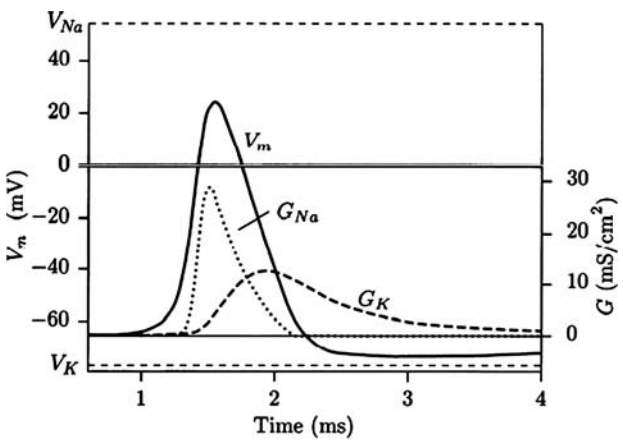


Figure 16.24 shows the time-dependence of an action potential and the associated ionic conductances. The Na^+ conductance increases after a time delay relative to the potential, peaks with the potential, and then falls off more rapidly. Again relative to the potential, the K^+ conductance rises more slowly and peaks after the fall of the potential. Although the H-H model was developed under space and voltage-clamped conditions, it can explain a large number of distinguishing features of an action potential, including: (1) an all-or-nothing response, with a threshold value of membrane current, in which a fixed pulse shape propagates down an axon at a constant speed; (2) an absolute refractory period of time after the action potential during which a second action potential cannot be elicited; (3) a relative refractory period of time during which a second action potential can only be elicited by an elevated current level substantially beyond a lower threshold; (4) a specific strength-duration relation giving the threshold current for

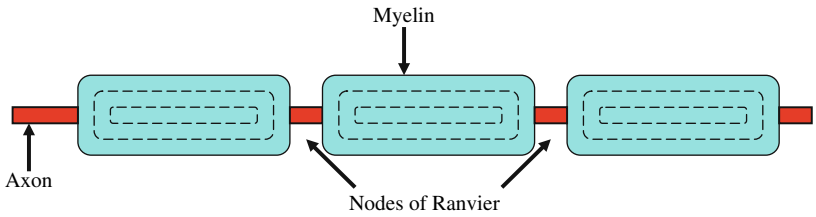


FIGURE 16.25 *Myelin sheath surrounding the axon with regularly spaced nodes of Ranvier.*

different duration current pulses; and (5) accommodation, in which the membrane adjusts to a sufficiently slow increase in current without producing an action potential.

The H-H model was developed through studies of the giant squid axon, however, most vertebrate neurons have a quite different structure that greatly modifies the nerve conduction mechanism. All neurons are found in association with other nonneural cells, known as supporting cells. In the central nervous system these are mostly glial cells and in the peripheral nervous system they are predominantly Schwann cells. Most of these supporting cells provide a myelin sheath that surrounds the axons of neurons in segments, or internodes, that have narrow gaps, known as nodes of Ranvier, at regular intervals. The internodes are roughly 1 mm long, some 1000 times longer than the nodes (Figure 16.25), and substantially change the electrical properties of nerve conduction. The myelin sheath provides a highly insulating layer effectively reducing membrane currents, which are fairly well confined to the nodes where there is a much higher density of channels than in the internodes. Myelin also greatly increases the space constant λ so that the membrane potential changes occurring at one node spread over many nearby nodes. Thus, a membrane potential depolarization occurring at one node caused by local membrane currents will rapidly appear at nearby nodes triggering membrane currents there as well. This type of signal propagation is known as *saltatory conduction* (from the Latin for "to jump" and having nothing directly to do with salt) because the membrane currents are triggered only at the nodes and not in a continuous fashion along the axon. Action potentials generated by saltatory conduction travel at much faster speeds (up to 100 m/s versus 20 m/s in squid giant axons) and myelinated neurons also have much smaller diameters (20 μm versus 0.5 mm in squid giant axons). A number of neuromuscular diseases, including multiple sclerosis (MS), affect the myelin around axons.

6. MEMBRANE CHANNELS: PART II

In Part I of our discussion of membrane channels in the previous chapter, we focused on the control and selectivity of voltage-gated ion channels. Now that we have learned something about electrical circuits, we return to membrane channels and discuss patch-clamp measurements of the electric currents through single channels. Patch-clamping was mentioned at the end of Section 4 above as a means of space-clamping, or electrically isolating a patch of a cell membrane. Four types of patch-clamps can be distinguished for use in recording the electrical activity of single channels (Figure 16.26).

A micropipette tip pushed up against a cell membrane provides an initial low resistance seal of about 50 MΩ. By applying suction, a gigaseal (GΩ seal) is then obtained where the patch is isolated from its surroundings by a huge resistance value; this configuration is known as the *cell-attached mode* (A in Figure 16.26) and is useful for studying voltage-gated channels or channels controlled by extracellular molecules supplied by the pipette. If the pipette tip is withdrawn pulling on the membrane, the membrane will rupture at the pipette tip edge producing an inside-out patch mode (B) if done in air or in the absence of divalent cations. By then immersing the inside-out patch in an external solution, one can control the ion content on the "cytoplasmic" side of the membrane.

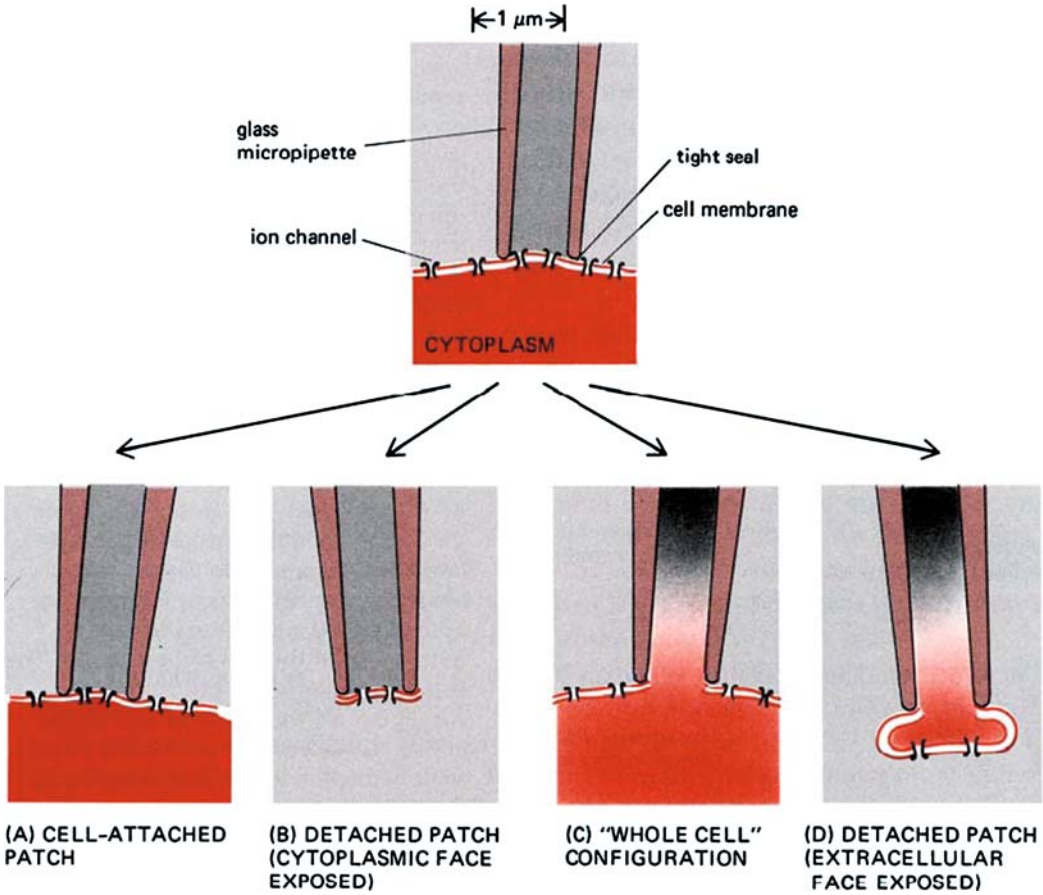
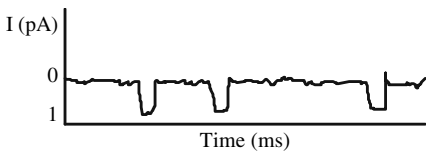


FIGURE 16.26 *Types of patch-clamps.*

Alternatively, after forming a gigaseal, an additional pulse of suction or voltage will open the membrane, exposing the cytosol to the pipette contents. This allows whole-cell recording (C) to occur in which the pipette can introduce ions or chemicals or even proteins into the cell. Small cells can be studied quite well using this method. For larger cells, pulling on the attached membrane, by surface tension, leads to the formation of an outside-out patch (D), with the extracellular face of the membrane able to be immersed in an external solution.

Each of these methods results in a patch of membrane (possibly the entire small cell) as a boundary between two controlled solutions, one within the pipette and one external. At that point current flow across the membrane can be monitored by electrodes and amplifier circuitry. With only a few channels per square micron of membrane area, the sensitivity of the electronics is such that single channel recordings can be made in which the flow of typically 5 picoamperes (1 pA = 10^{-12} A) of current lasting typically 1 ms can be measured. Such a flow corresponds to about 30,000 monovalent ions through a single channel in the membrane. Our ability to measure such small currents accurately hinged on the development of field-effect transistor (FET) amplifiers which have very low noise characteristics.

FIGURE 16.27 *Single channel recording from a patch-clamp. Each step in current is the opening and closing of a single ion channel.*



Based on the macroscopic sodium channel currents measured for large membrane surfaces discussed in the last section, one might guess that the single channel Na^+ current recording would be just a miniature version of that continuous curve in time. However, what is found in a single channel measurement is totally different. The current instead comes in individual discrete, rapid bursts of charge flow (Figure 16.27). These pulses are spaced close together at times corresponding to a large macroscopic current and farther apart, on average, during smaller macroscopic currents. Effectively, with a large number of identical channels, the current pulses add together to give the continuous macroscopic current curve. An alternative way to consider this

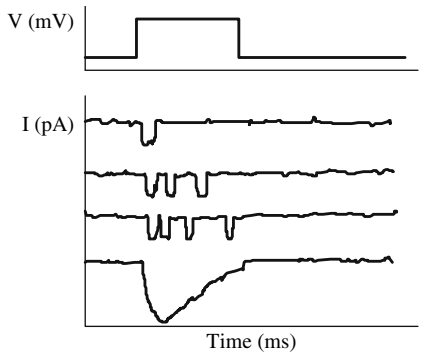


FIGURE 16.28 *The upper voltage represents an applied depolarizing voltage clamp, and the three single channel recordings, top three I versus t curves, indicate a series of repeated measurements of current in synch with the applied voltage pulse. When many repeated single channel recordings are summed (bottom curve) the macroscopic aggregate current that results is found to be the same as when measurements are made over a larger surface area to give a macroscopic current signal directly in a single measurement.*

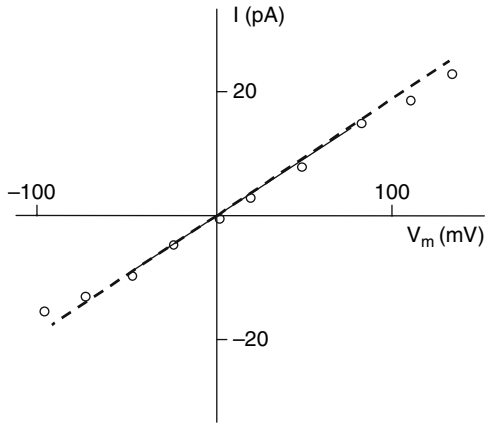


FIGURE 16.29 *Current-voltage relations for a single K^+ channel. The conductance of the channel can be obtained from the slope of the dashed line.*

is that repeated depolarization of the same channel, under identical conditions, will elicit a seemingly random, different response each time; but when those responses are summed, the average single channel current is found to mimic the macroscopic current observed from a single simultaneous measurement on large numbers of such channels (Figure 16.28).

The conductance of a single channel can be determined by measuring the channel current as a function of the membrane potential difference from the equilibrium potential for that ion species (see Equation (16.23)). Figure 16.29 shows an example of data from a K^+ channel. The slope is the conductance for that channel under the experimental conditions.

A simple model of ionic channels can be developed in which the channels exist in only two possible states, closed (C) with zero conductance and open (O) with a constant conductance. Figure 16.30 shows a hypothetical energy diagram for this model. Note that the energy levels might well depend on the membrane voltage. According to equilibrium thermodynamics, the ratio of the number of open to closed channels is given by the Boltzmann factor as

$$\frac{N_0}{N_C} = \mathrm{e}^{-(E_0 - E_C)/k_B T}. \tag{16.27}$$

The probability that a channel is open, P_0, is given by the ratio

$$P_0 = \frac{N_0}{N_0 + N_C} = \frac{\dfrac{N_0}{N_C}}{\dfrac{N_0}{N_C} + 1} = \frac{\mathrm{e}^{-(E_0 - E_C)/k_B T}}{\mathrm{e}^{-(E_0 - E_C)/k_B T} + 1} = \frac{1}{1 + \mathrm{e}^{(E_0 - E_C)/k_B T}}. \tag{16.28}$$

A plot of P_0 versus $(E_0 - E_C)/k_B T$ is shown in Figure 16.31. For large negative values of the abscissa, corresponding to a higher closed than open state energy, with the difference large compared to thermal energies, all channels are open. In the opposite limit of large positive values, all channels are closed. When the open and closed energy values are equal, 50% of the channels are open. The energy levels of both states change in response to the membrane potential.

In a further refinement of this model, various mechanistic models of the energies of the two states can be assumed. For example, one simple scheme is to imagine a gating molecule acting as a dipole and either spanning the channel, so that the channel is closed, or not, so that the channel opens. Thus the rotation of a dipole—triggered by electrical

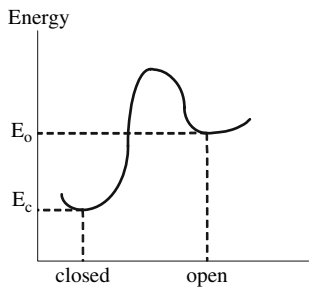


FIGURE 16.30 *Energy diagram for two-state model of an ion channel.*

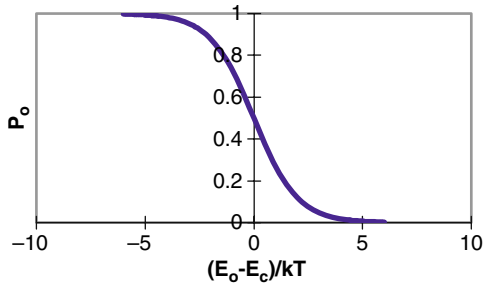


FIGURE 16.31 *Probability of channel being open as a function of the energy difference.*

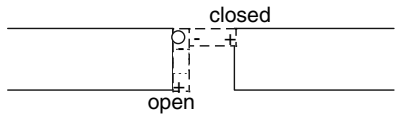


FIGURE 16.32 *A model for a two-state dipole control of channel opening and closing.*

forces—controls the conductance of the channel (Figure 16.32). The interaction energy can then be written in terms of the dipole and the membrane potential and an analysis and comparison with data can lead to an estimate of the valence of the gating charge ze on the dipole. Hodgkin and Huxley's work showed that z is about 6 for the sodium channel, so that six positive charges are needed to shift from the cytosolic to the extracellular side of the membrane in order to give the observed voltage-dependence for the gating. Equivalently 6 negative charges can shift across the membrane in the opposite direction, or 12 charges could shift halfway across, and so on. Although some features of the H-H model can be recovered from this simple model, multistate channel models, with additional parameters, have also been developed.

CHAPTER SUMMARY

When placed in an electric field directed along a wire, the free electrons in the conductor move randomly about at thermal velocities (~10^6 m/s) while drifting along the wire at very low speeds (~mm/s). The drift velocity produces a net flow of charge Q, making up an electric current I, defined as

$$I = \frac{\Delta Q}{\Delta t}. \qquad (16.4)$$

In a conductor, the current is proportional to the applied potential difference V through Ohm's law

$$V = IR, \qquad (16.7)$$

where R is the electrical resistance of the conductor. As current flows through a conductor, electrical energy is lost thermal energy through joule heating at a rate given by

$$P = IV = I^2R = \frac{V^2}{R}. \qquad (16.10)$$

Electrical circuits can be analyzed using two fundamental rules: Kirchoff's loop equation, stating that the net voltage difference around any closed loop in a circuit is zero, and the junction rule, stating that at any branch point in a circuit the total current into the branch point must equal the total current flowing out. In simple circuits with either just two resistors or two capacitors, we can develop the following rules for finding net R and net C values:

$$R_{\text{equiv}} = R_1 + R_2 \qquad \text{(resistors in series)}$$

$$\frac{1}{R_{\text{equiv}}} = \frac{1}{R_1} + \frac{1}{R_2}, \qquad \text{(resistors in parallel)} \qquad (16.14)$$

$$\frac{1}{C} = \frac{1}{C_1} + \frac{1}{C_2} \qquad \text{(capacitors in series)} \qquad (16.16)$$

$$C = C_1 + C_2 \qquad \text{(capacitors in parallel).} \qquad (16.18)$$

In circuits with series R and C elements, an analysis finds that when discharging the capacitor the charge on the capacitor and the current in the circuit both decrease exponentially according to

$$Q = Q_0\, e^{-\frac{t}{RC}}, \qquad (16.20a)$$

$$I = I_0 \, e^{-\frac{t}{RC}}, \qquad (16.20b)$$

An analysis of the potential difference across a membrane with an imbalance in the concentration of ions on both sides of the membrane leads to an equation, the Nernst equation, for the potential difference across the membrane due to the difference in ion concentration on the inside (c_i) versus outside (c_0)

$$V_K = \frac{RT}{z\mathcal{F}} \log\left(\frac{c_0}{c_i}\right) \qquad (16.22)$$

where K is the example of potassium ions, R is the molar gas constant, z is the ion valence and $\mathcal{F}$ is the Faraday constant. Using this basic idea, nerve impulses can be modeled as combinations of such potential differences that have different time behaviors in the Hodgkin–Huxley and other models of nerve conduction.

To test these models, special techniques have been developed to measure electrical properties of membranes in order to relate them to their structure. In particular, patch-clamping techniques allow scientists to study the electrical properties of single special channels or pores, made from individual proteins embedded in biological membranes, by measuring pulses of current flow corresponding to the opening of a single one of these membrane channels. The time-varying voltage signal seen from earlier voltage clamp measurements agrees with both patch clamp measurements averaging over large numbers of such pores, and with time-averaging single current pulse measurements after repeated channel openings, as seen, for example, in Figure 16.28.

QUESTIONS

1. Some mistakes students make as they are learning about circuits involve using wrong language to describe situations, leading to or caused by conceptual misunderstandings. For example, it is fairly common to hear students say that current flows across a resistor, or that voltage flows around the circuit, or to ask what the voltage of the resistor is. What is wrong with each of these statements?
2. Explain how it is that when you turn on an electric light switch, the light comes on immediately even though the electrons making up the electric current travel at very slow speeds of only about mm/s. Develop an analogy with water coming out of a full hose when the valve is first opened.
3. If conductors cannot have electrostatic fields within them, what is the mechanism that produces the force on electrons within a conducting wire in a circuit when there is an electric current flowing?
4. If free electrons in a conducting wire experience a net force due to the electric field in the wire, why don't they accelerate continuously instead of traveling along with a constant average velocity?
5. Explain the difference among resistivity, resistance, conductivity, and conductance. Which are intrinsic properties of a material and which depend on its size and shape?
6. If a resistor of resistance R is connected to a battery of voltage V, the equation for the power dissipated in the resistor, $P = I^2R$, implies that a larger resistor will dissipate more energy and get hotter than a smaller resistor. This is not true. Explain why not.
7. A homeowner keeps losing electric power during a hot summer evening due to blown 20 A fuses. After replacing the blown fuse several times and having the same problem, he decides to use a 30 A fuse so it won't blow with the same electrical devices on. Why is this a bad idea?
8. Copper wires covered with rubber-based insulation are commonly used in household electrical wiring. These wires come in different gauges, corresponding to different diameters of the copper wire, where increasing gauge corresponds to decreasing diameter. Calculate the resistance of 100 m length of 14 gauge wire (1.63 mm diameter) and of 10 gauge wire (2.59 mm diameter). According to National Electric Code standards, the maximum current capacities of these two wires are 15 A and 25 A. Which can carry more current?
9. Check that the SI units of the product of R and C are seconds; verify that specific resistance times specific capacitance also has units of seconds.
10. Two physics students are each measuring the RC time constant of a simple series RC circuit. One of them sets the initial voltage on the capacitor to 10 V and measures the time for the voltage to drop to 5 V. The second student, using the same circuit, sets the initial voltage to 20 V and measures the time for the voltage to drop to 10 V. Will these times be the same? Why?
11. In circuit analysis, Kirchhoff's loop equation is often equated with conservation of energy, whereas Kirchhoff's branching equation for currents is often equated with conservation of electric charge. Discuss this statement.

12. Explain in words what an equivalent resistor means when replacing some collection of resistors in a circuit by an equivalent resistor.

13. Explain why when using a multimeter as a voltmeter its two wire leads can be simply put in parallel, or across, the circuit element whose voltage is to be measured, but when used as an ammeter this cannot be done, but rather a wire leading to that circuit element must be "broken" so that the ammeter can be inserted in series with it. Discuss this in words and in terms of Kirchhoff's equations.

14. A flashlight bulb acts as a small resistance when connected to a battery. If two identical bulbs are connected in parallel to the same battery will they be brighter, dimmer, or the same brightness as when a single bulb is connected to that same battery? Repeat this when the two bulbs are placed in series across the same battery.

15. In the previous question, does the battery supply more, less, or the same current with two bulbs in parallel as when a single bulb is connected to the battery? Answer this when the two bulbs are placed in series across the battery.

16. Given an unlimited supply of 100 Ω resistors, how could you arrange a network of them to have an equivalent resistance of 150 Ω? Of 75 Ω?

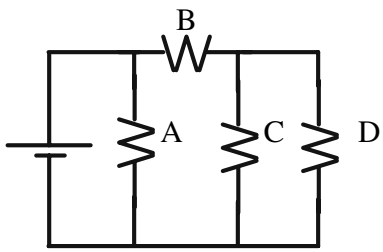
17. Discuss the meaning of the two parameters of the linear cable model of a neuron, the space and time constants. What do they tell us?

18. In the cable model, discuss in words the function of each of the circuit elements.

19. What is the fundamental goal of a patch-clamp?

20. Can you think of any other physical processes like membrane channel current that appear continuous on one level, but are actually made up of discrete small packets on a finer level?

Questions 21 and 22 refer to: Consider the circuit to the right. The battery is a perfect source of emf. Treat A, B, C, and D as bulbs of equal resistance.



21. Rank order the brightness of the bulbs in the circuit shown.

22. Fill in the following table with "S" for same brightness, "D" for dimmer, "B" for brighter, and "O" for goes out. "U" means that bulb is "unscrewed" for that situation.

Situation	*A*	*B*	*C*	*D*
1	U			
2		U		
3			U	
4				U

MULTIPLE CHOICE QUESTIONS

1. One end of a resistor in a simple circuit is at +5 V; the other is at +3 V. Which one of the following is true? (a) Electrons must be entering the resistor at the +5 V end and leaving at the +3 V end. (b) Electrons leaving the resistor have a higher kinetic energy than electrons entering the resistor. (c) Electrons at the +3 V end have a higher electric potential energy than at the +5 V end. (d) A current of 2 A must be flowing through the resistor.

2. A pure parallel combination of resistors has an equivalent (or effective) resistance of 2 Ω. Which one of the following is true? (a) The sum of the individual resistances is 2 Ω. (b) The sum of the reciprocals of the individual resistances is 2 Ω. (c) Each of the individual resistances is greater than 2 Ω. (d) Each of the individual resistances is smaller than 2 Ω.

3. The statement, "The current in a resistor is directly proportional to the potential difference across the resistor," is known as (a) Coulomb's law, (b) Gauss's law, (c) Ohm's law, (d) Ampere's law.

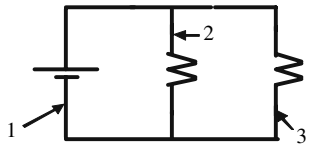
4. The electrical resistance of a long piece of wire is R. The wire is stretched to be twice as long and, because the wire's volume doesn't change, its cross-sectional area is halved. The electrical resistance of the stretched wire is (a) $R/2$, (b) R, (c) $2R$, (d) $4R$.

5. A steady current flows through a resistor. An electron in the current flow enters the resistor at the resistor's +5 V end and leaves at the resistor's +10 V end. Which one of the following is true? (KE = kinetic energy, PE_E = electric potential energy.) (a) $\Delta KE = 0$, $\Delta\ PE_E < 0$, (b) $\Delta KE < 0$, $\Delta\ PE_E = 0$, (c) $\Delta KE > 0$, $\Delta\ PE_E < 0$, (d) $\Delta KE = 0$, $\Delta\ PE_E > 0$.

6. A voltmeter is used to read the potential difference across the poles of a battery. The battery is rated at 20 V. The battery is connected in series to a switch, an ammeter, and a resistor. When the switch is open, the ammeter reads 0.0 A and the voltmeter reads 20.0 V. When the switch is closed the ammeter reads 1.0 A and the voltmeter reads 19.0 V. Which one of the following is most likely to be the explanation for this result? (a) The ammeter has too little resistance. (b) The voltmeter has too much resistance. (c) The battery has an internal resistance of 1.0 Ω. (d) The resistor in the circuit has a resistance of 20.0 Ω.

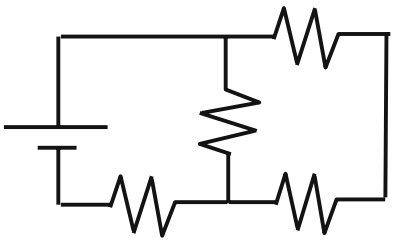
7. A battery has an emf of 10 V and an internal resistance of 1 Ω. When the battery is connected to a combination of resistors, a perfect ammeter reads a current of 1 A leaving the positive pole of the battery.

At the same time, a perfect voltmeter placed across the poles of the battery will read a potential difference of (a) 11 V, (b) 10 V, (c) 9 V, (d) 1 V.



8. Two resistors are in parallel as shown in the figure above. When the current at point 1 is 0.15 A and the current at point 2 is 0.05 A, what is the current at point 3? (a) 0.15 A, (b) 0.10 A, (c) 0.05 A, (d) between 0.05 A and 0.10 A.

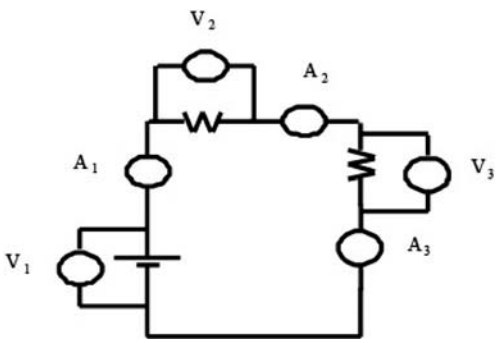
9. Two resistors are connected to an ideal battery in series. Resistor 1 has a potential difference across it of 10 V and resistor 2 has a potential difference across it of 20 V. Now, the two resistors are connected to the same battery in parallel. The potential difference across resistor 1 (a) is now 10 V, (b) is now 20 V, (c) is now 30 V, (d) cannot be calculated because the resistances aren't given.



Questions 10 and 11 refer to the figure above

10. What is the equivalent resistance of the circuit shown in the figure to the right? Each resistor is 1 Ω. (a) 4 Ω, (b) 1.67 Ω, (c) 0.60 Ω, (d) 0.25 Ω.

11. In the circuit shown if the battery supplies 3 V and each resistor is 1 Ω, what is the current through the resistor in the middle branch? (a) 3 A, (b) 2.4 A, (c) 1.2 A, (d) 0.6 A.



Questions 12–15 refer to the figure above. The A's are perfect ammeters, the V's are perfect voltmeters, the battery is a perfect source of emf, and the resistors are equal. V_2 reads 5 V and A_3 reads 1 A.

12. V_1 must read (a) 0 V, (b) 5 V, (c) 10 V, (d) some value that depends on the actual emf of the battery and the actual resistances.

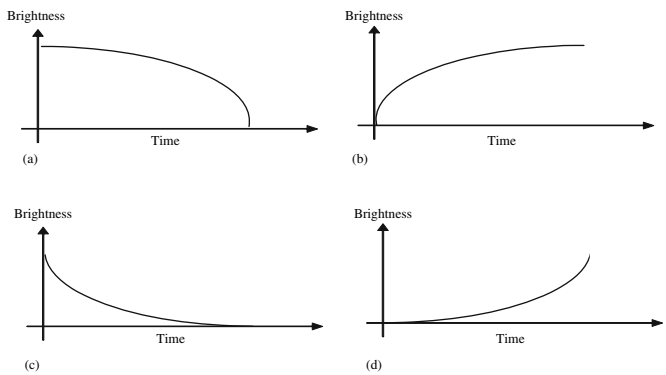
13. V_3 must read (a) 0 V, (b) 5 V, (c) 10 V, (d) some value that depends on the actual emf of the battery and the actual resistances.

14. A_1 must read (a) 1 A, (b) 2 A, (c) 3 A, (d) some value that depends on the actual emf of the battery and the actual resistances.

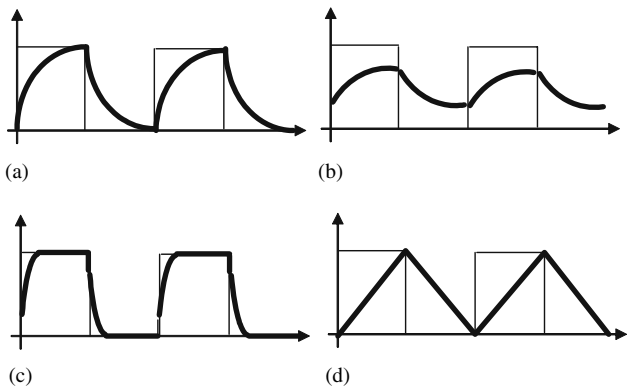
15. A_2 must read (a) 1 A, (b) 2 A, (c) 3 A, (d) some value that depends on the actual emf of the battery and the actual resistances.

16. Two light bulbs, one rated at 50 W and a second rated at 100 W, are both supposed to be connected to a 110 V source of emf. Which one of the following is true? The 50 W bulb has (a) twice the resistance as the 100 W bulb, (b) four times the resistance of the 100 W bulb, (c) half as much resistance as the 100 W bulb, (d) one quarter as much resistance as the 100 W bulb.

17. A bulb (i.e., a resistor) is connected in series to a switch, a battery, and an uncharged capacitor. At $t = 0$, the switch is closed. Which of the following best describes the brightness of the bulb as a function of time?



18. Which of the following best describes the potential difference across a capacitor that is connected in series to a resistor and a source of emf that is sequentially $+V$ for time T, then 0 for time T, and so on, when T is small compared with RC? Vertical axes are potential difference, horizontal axes are time. The lighter plots are the emf, the bolder plots are the capacitor voltage.



PROBLEMS

1. What is the equivalent current in a solution of monovalent ions flowing through a capillary tube such that 1 mM of ions leaves the tube each second.

2. A capacitor with a charge of 5 μC has its terminals shorted by a metal wire so that the charge flows off within 2 μs. What is the average current flowing during that time?

3. What is the average current when all the sodium channels on a 100 μm^2 patch of muscle membrane open together for 1 ms? Assume a density of 50 sodium channels per μm^2 of surface and a flow rate of 1000 ions per ms through each channel.

4. Calculate the conductance and the resistance of a 10 m length of 14 gauge copper wire, which has a diameter of 1.63 mm. If this wire is connected directly to the terminals of a 12 V dc power supply, shorting it, how much current will flow assuming the power supply can deliver an unlimited amount of current?

5. A 1 cm^3 cube of gold ($\rho = 1.61 \times 10^{-8}$ Ωm) is drawn out into a uniform cylinder of 20 m length. What is its electrical resistance?

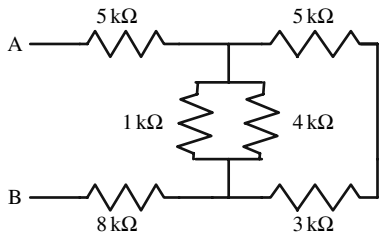
6. Two 100 m 14 gauge wires (1.63 mm diameters), one of copper and one of aluminum, are soldered together and the 200 m wire is then connected to a 6 V dc power supply with unlimited current.
(a) How much current flows in the wire?
(b) What is the potential across each 100 m section of wire?
(c) How much power is developed in each section of wire?

7. A 1000 W heater runs from a 100 V dc power supply.
(a) How much current flows in its heating cable wire?
(b) What is the resistance of the wire?

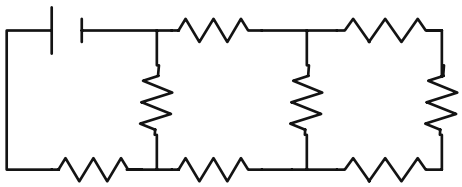
8. An electric eel, found in the rivers of Brazil, can discharge lethal currents of 1 A at 400 V. How much power does the eel generate?

9. An immersible heater coil is to be designed to heat an insulated container with 4 liters of distilled water from 20° to 50°C in less than 30 min.
(a) How much energy must be input to heat the water to this temperature?
(b) To heat the water, what minimum power must be supplied?
(c) If a 12 V power supply is to be used, what minimum current must flow in the heating coil?
(d) What must be the total resistance of the heating coil? Is this a maximum or minimum resistance to heat the water in 30 min or less?

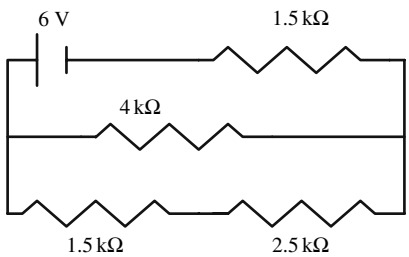
10. Determine the equivalent resistance between points A and B in the following circuit.



11. Given the network of equal 1 kΩ resistors shown below, compute its equivalent resistance and the current drawn from the 12 V power supply. (Hint: Combine resistors in stages using the simple rules for series and parallel combinations of resistors.)

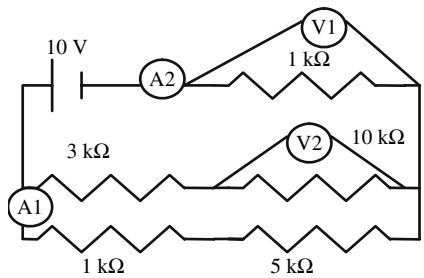


12. Analyze the circuit shown below to find the currents flowing through and the power generated in each resistor.

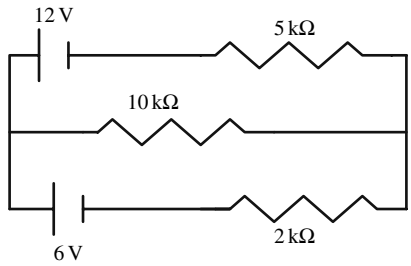


13. A single 1 MΩ resistor is connected across a power supply. An ammeter is inserted to measure the current out of the battery. If a voltmeter with 10 MΩ resistance is used to measure the voltage across the resistor, what will be the percent change in the current reading on the ammeter when the voltmeter is connected across the resistor?

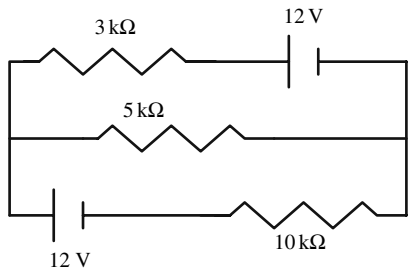
14. Find the reading that each of the (ideal) meters would have in the following circuit.



15. Analyze the following circuit to find the current flowing through the 10 kΩ resistor.



16. Find the current in the central branch of the following circuit.



17. *RC* time constants can be easily estimated by measuring the time (known as the half-time) for the capacitor voltage to decrease to half of some arbitrary starting value when discharging through a resistor. From Equation (16.12a), the voltage across the capacitor will vary as

$$V(t) = V_0\, e^{-\frac{t}{RC}}.$$

Show how a single measurement of the half-time can be used to determine the *RC* time constant. (Hint: Substitute $V(t) = V_0/2$.)

18. A 100 μF capacitor wired in a simple series *RC* circuit is initially charged to 10 μC and then discharged through a 10 kΩ resistor.
(a) What is the time constant of the circuit?
(b) What is the initial current that flows?
(c) How much charge is left on the capacitor after 1 time constant?
(d) What is the current after 1 time constant?
(e) How much charge is left on the capacitor after 3 time constants have elapsed and what current is flowing then?

19. A simple *RC* series circuit has a 100 μF capacitor.
(a) If the time constant is 50 s, what is the value of the resistor?
(b) Suppose that a second identical resistor is inserted in series with the first. What is the new time constant of the circuit?
(c) Suppose the second identical resistor is placed in parallel with the first resistor, still connected to the capacitor. What is the new time constant in this case?

20. Consider a defibrillator, acting as a 32 μF capacitor and a 47 kΩ resistor in a series *RC* circuit. The circuitry in this system applies 5000 V to the *RC* circuit to charge it.
(a) What is the time constant of this circuit?
(b) What is the maximum charge on the capacitor?
(c) What is the maximum current in the circuit during the charging process?
(d) What are the charge and current as functions of time?
(e) How much energy is stored in the capacitor when it is fully charged?

21. We've seen that the Earth's atmosphere is able to act as a capacitor, with the ground and the clouds acting as plates with an air gap in between. Under certain circumstances air can be made to conduct, so that electric charge can flow from the clouds to the ground in what we call a lightning bolt. Assuming that the clouds are distributed around the entire Earth at a fixed distance of 5000 m above the ground of area $4\pi R^2_{\text{Earth}}$, where $R_{\text{Earth}} = 6400$ km, the resistance of the air between the clouds and the ground is calculated to be $R = 300\ \Omega$.
(a) Assume that the charge is distributed spherically, so that $V = k(Q/r)$ and therefore ΔV is the difference in potential between the lower plate (the Earth's surface) and the upper plate (the clouds). In addition, assume that in a typical day, 5×10^5 C of charge is spread over the surface of the Earth. What is the potential difference between the clouds and the ground?
(b) What is the capacitance of the Earth–cloud capacitor?
(c) If the charge on the clouds is discharged through the air, what is the capacitive time constant for this discharge?
(d) How many lightning strikes does this amount of charge correspond to if each lightning strike contains about 25 C of charge?
(e) Approximately how long would it take the Earth–cloud capacitor to discharge to 0.1% of its initial charge?
(f) Assuming that the charge is immediately replenished as soon as the discharge process ends, approximately how many lightning bolts are there per day?

22. Fill in all the steps in the calculation of the number of ions crossing a membrane channel when it opens (see Section 16.3 following Example 16.5). Now, using those same numbers, calculate the total number of moles of charge crossing the membrane when the

membrane of a spherical cell with a radius of 10 μm completely depolarizes. If this charge were all K^+ leaving the cell, calculate the fraction of the K^+ present in the cell interior that crosses the membrane when it depolarizes. (Hint: You need to calculate the number of moles of K^+ inside the spherical cell and the number on the surface of the membrane, all of which is assumed to cross the membrane when it depolarizes; see Table 16.2.)

23. Check the values of the Nernst potential in Table 16.2 using Equation (16.22).

24. Show that Equation (16.28) leads to the plot shown in Figure 16.31. In particular, show that the P_0 values for $E_0 = E_C$ and for large and small values of the difference $(E_0 - E_C)$ are correct.

17 Magnetic Fields

Imagine that you are a student in a summer program at a major research institution. The training involves learning how to do different types of microscopy and you are studying a sample of bacteria under the light microscope. You notice that some of the bacteria always seem to migrate toward the same side of the microscope slide and you get curious and wonder what could be causing them to choose that direction. Remembering chemotaxis, the directed motility response to chemicals, and thinking that perhaps there is a variation in oxygen level or something of that kind, you check that the cover slip is on correctly and that there is no preferred direction. After puzzling a bit you decide to make up another slide but even then you still observe the same type of bacteria all clumped toward the same side of the microscope.

You decide to rotate the microscope around on the table and find that the bacteria always appear to clump toward the same side of the room, independent of the orientation of the microscope. Now you are truly puzzled. After a number of repeated observations, you begin to wonder about some strange external force. But what could it be? Gravity acts vertically and the bacteria are confined to the horizontal slide, and furthermore always go in the same direction, independent of the microscope orientation. You try taking the microscope into another room where you observe the same phenomenon, and you notice that they always move in the same direction. You try a third room, where there happens to be a small stirring bar magnet lying nearby on the table. Now the bacteria move in a different direction, but still move in that same direction regardless of the orientation of the microscope. By chance, the magnet gets moved and you discover that the bacteria clump in the direction of the magnet. You pick up the magnet and, by moving it about, can control the direction in which the bacteria move. You have discovered *magnetotactic bacteria*.

Something similar to this story actually occurred in 1975 when the first magnetic bacteria were discovered. These bacteria actually have microscopic permanent magnets embedded in them that steer the bacteria toward the Earth's magnetic north. Figure 17.1 shows an electron microscope image of such a bacterium with the tiny crystalline magnets visible. One can only speculate on the significance of the magnets for the bacteria. Causing them to swim toward magnetic north in the northern hemisphere means that they swim not only toward the north, as a horizontally held compass points, but also vertically downward at an angle toward the North Pole through the Earth. The magnetic force then causes them to swim beneath the murky waters of ponds and lakes to the muddy bottom where they are found to thrive. Other animals, including bees and some types of fish and birds, and some algae also contain microscopic magnetic particles, although their function in orientation or sensory input is not totally clear.

Permanent magnets, like those that hold pictures on a refrigerator door, are familiar to us as objects that attract and are attracted to other magnetic materials, such as iron. This actually occurs through the production of a magnetic field and the interaction of a magnet or magnetic material with this field. In this chapter we show the equivalence of such a magnetic field to one produced by moving electric charges or electric currents, including

J. Newman, *Physics of the Life Sciences*, DOI: 10.1007/978-0-387-77259-2_17,

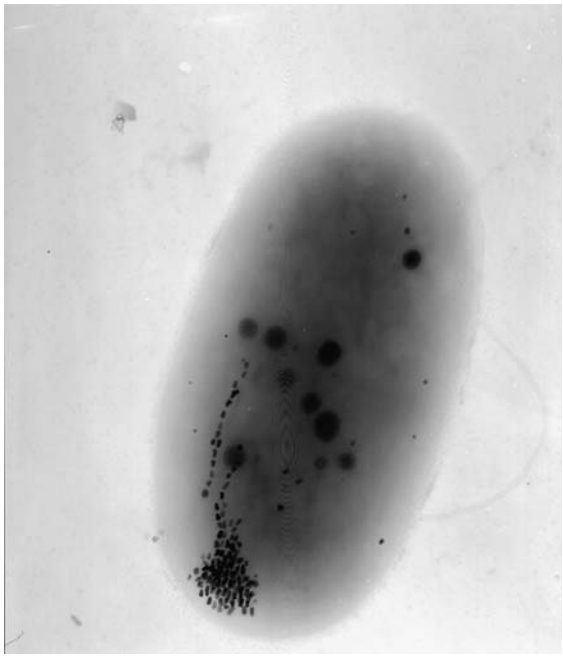


FIGURE 17.1 *A magnetic bacterium with chains of small magnetic particles, called magnetosomes, made from magnetite and each about 100 nm in size.*

those of the neurons in our brain. We first learn the basic force law governing the interaction of moving electric charges with a magnetic field and some ideas on the interaction energy. Magnetic forces on macroscopic objects, such as magnets or wires with electric current flowing, are shown to have the same origin as magnetic forces on atoms.

1. MAGNETIC FIELDS AND FORCES

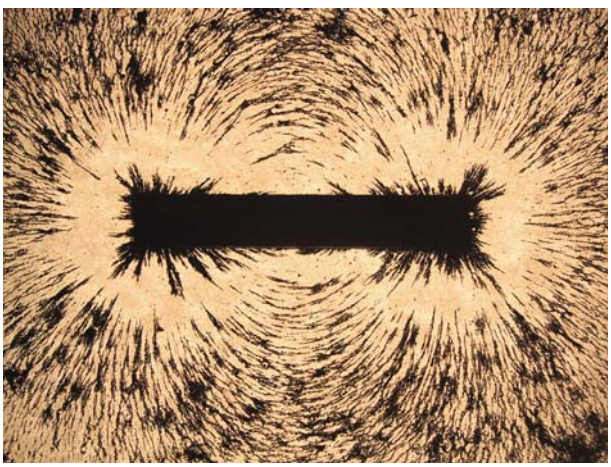
Our fascination as children with bar and horseshoe magnets is no doubt similar to the earliest human experiences with magnetite, the naturally occurring magnetic mineral Fe_3O_4, or with iron. The physical basis of the familiar attraction or repulsion of two magnets (Figure 17.2) brought near each other is complex and resides in the atomic structure of the materials. A compass needle, itself a small magnet, can be used to determine the presence, direction, and even strength of a magnetic field. Thus, a compass needle can be thought of as the magnetic analog of a test charge for electric fields. We show later that a compass needle experiences a torque proportional to the magnetic field strength causing it to orient along the magnetic field. In Section 4 below, we discuss the generation of a magnetic field by electric currents, whether macroscopic currents in a wire or microscopic currents in a neuron or in a piece of magnetic material. Here we first discuss the interaction of a magnetic field with electric charges.

Suppose there is a uniform magnetic field B in a region of space. Let's discuss the magnetic force on an electric charge q in this region of space. First of all it is found that if the charge is at rest there is no magnetic force acting at all. Furthermore, if the charge does move, but has its velocity along the magnetic field direction (determined, e.g., by a compass needle) then there is still no magnetic force on the charge. If, however, it moves perpendicular to the magnetic field direction with velocity v, then there is a magnetic force F_M that acts on the charge and is given by

$$F_M = qvB \qquad (v \perp B). \tag{17.1}$$

The direction of the magnetic force is found to be perpendicular to the plane in which the velocity and magnetic field lie as shown in Figure 17.3. To remember which of the two possible directions is correct for the magnetic force (up or down from the plane of $\vec{v}$ and $\vec{B}$) we use the *right-hand rule* that tells us to consider the set of three vectors $\vec{v}$, $\vec{B}$, and $\vec{F}_M$ (in that order corresponding to x-, y-, and z-axes) as making up a right-handed coordinate system as shown in Figure 17.4. Thus, curling the fingers of your right hand from $\vec{v}$ toward $\vec{B}$ results in your thumb pointing along the appropriate direction for $\vec{F}_M$ as in Figure 17.3. With a little practice, this rule is useful for finding the direction of the magnetic force if the directions of the magnetic field and charge velocity are known.

FIGURE 17.2 *Iron filings mapping out the magnetic field of a bar magnet. A compass needle placed near the magnet will orient along the magnetic field lines.*



The SI unit for magnetic field is the tesla (T) where, from Equation (17.1), 1 T = 1 N-s/C-m. One tesla is a fairly large magnetic field (a large magnet for an MRI—medical resonance imaging—machine may have a magnetic field of several tesla) considering that the Earth's magnetic field is only about 0.5×10^{-4} T. A smaller unit, the gauss (G), with $1\ \text{G} = 10^{-4}$ T, is often used for magnetic fields, so that the Earth's magnetic field is approximately 0.5 G.

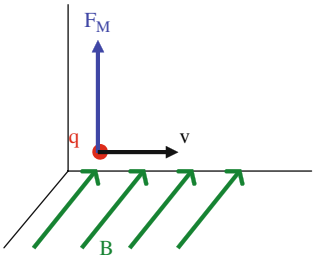


FIGURE 17.3 *Charge* q *moving with velocity* v *perpendicular to a uniform* B *field. The charge experiences a force* F_M*, given by Equation (17.1), perpendicular to the plane containing* B *and* v *and in the direction given by the right-hand rule.*

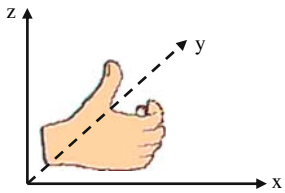


FIGURE 17.4 *The right-hand rule: curl the fingers of your right hand from* x *to* y *and your thumb points along* z*. This rule works for any three vectors that follow the right-hand rule, where the order of the vectors corresponds to the first* (x) *and second* (y) *vectors resulting in the third* (z)*.*

What will be the trajectory of this charge under the influence of the magnetic force? Because the force is perpendicular to the velocity vector, we know from our discussions of centripetal force in mechanics that the speed of the particle will not change but that its direction will change. As it does, the magnetic force continually remains perpendicular to the velocity vector and so the charge will move in a closed circular path just as in the case of a centripetal force. When an object is tied to a string and swung in a circle, the string provides the centripetal force that is always directed toward the circle's center and is perpendicular to the tangential-directed velocity. In the present case, the magnetic field supplies the force needed to steer the charge in a circle. We can further analyze the motion of the charge by using Newton's second law and our knowledge of centripetal acceleration to write

$$F_M = qvB = ma = \frac{mv^2}{r}, \tag{17.2}$$

where m is the mass of the particle and r is the radius of the circle. Solving this equation for the ratio q/m, an intrinsic property of the particle, we find

$$\frac{q}{m} = \frac{v}{rB}, \tag{17.3}$$

indicating that for a particle of given charge-to-mass ratio with a velocity perpendicular to a constant uniform magnetic field, the particle will move in a circular orbit with a radius proportional to its velocity.

At this point in our discussion we can understand the basis for a *mass spectrometer*, a device used to determine the relative masses and abundances of ions. As shown in Figure 17.5 the material to be analyzed is first vaporized and ionized by stripping an electron from each atom. The positive ions are then accelerated through a fixed potential difference V, so that they obtain a kinetic energy equal to eV. They then pass through a hole and enter a region in which there is a uniform magnetic field perpendicular to their velocity. As we have just seen, they will travel in a circular arc under the influence of the magnetic force. Those ions traveling in a circle of particular radius r will pass through a second hole and be detected.

We can use this knowledge to find the ion's mass. Because, from $eV = \frac{1}{2}mv^2$ we have that

$$v = \sqrt{\frac{2eV}{m}},$$

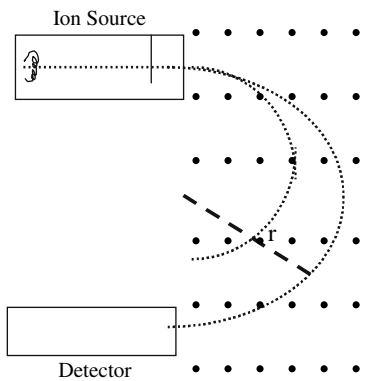


FIGURE 17.5 *Schematic of a mass spectrometer. Positive ions travel in circular paths with radii depending on their charge-to-mass ratio. The* B *field is out of the paper as shown by the dots representing arrow tips coming out towards you.*

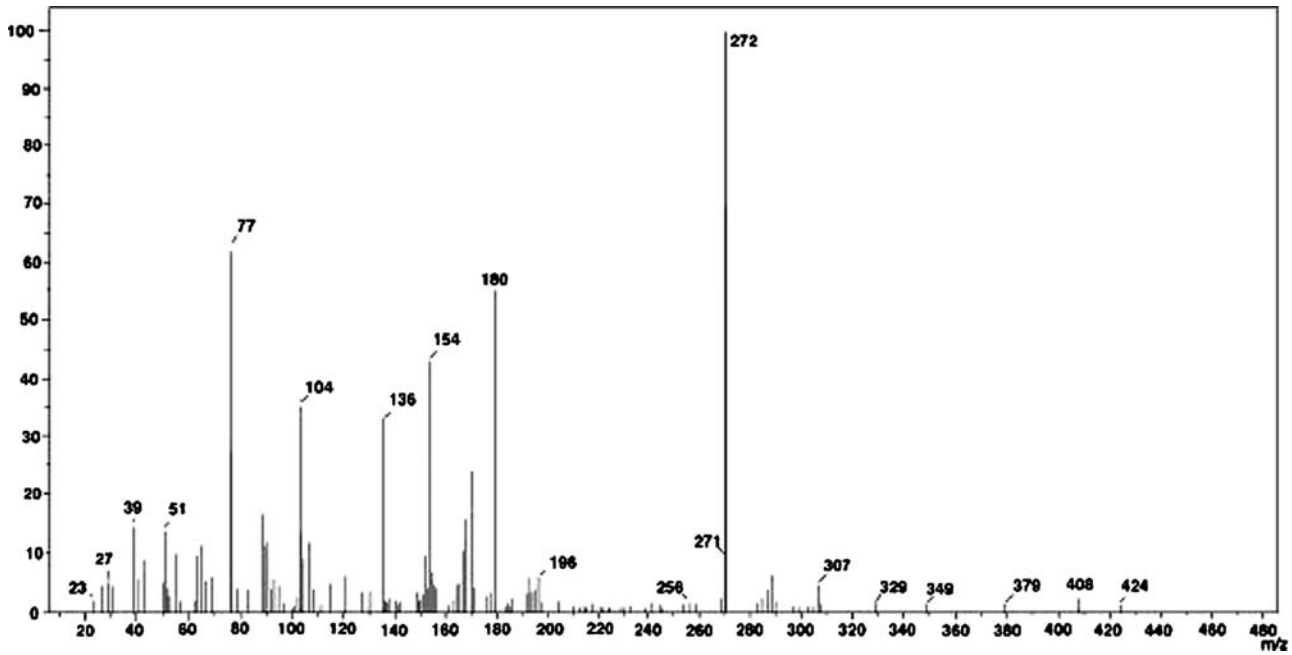


FIGURE 17.6 *Mass spectrograph showing relative abundance of different detected ions.*

we can substitute this expression into Equation (17.3) (with $q = e$) to solve for m after a bit of algebra,

$$m = \left(\frac{er^2}{2V}\right)B^2. \tag{17.4}$$

With V and r fixed, B can be varied and the masses of the detected ions determined. By measuring the number of detected ions per unit time, the relative abundances of the ions can also be found. Plotting the detector output signal as a function of B^2 shows peaks appearing at intervals corresponding to the atomic mass of the ions (Figure 17.6).

Mass spectrometers can be used to separate different *isotopes*, atoms with the same numbers of protons but different numbers of neutrons, of a material. All large hospitals have mass spectrometers used for a variety of purposes including identification of respiratory gases (those inhaled and exhaled during a diagnostic test) and anesthesia gases as well as various isotopes used for radiation therapy purposes.

Returning to our discussion of the force on an ion in a magnetic field, what happens if the electric charge has a velocity in an arbitrary direction in the region of a uniform magnetic field? From our special cases, we can conclude that if the velocity vector is written as the sum of its component parallel and perpendicular to $\vec{B}$, the parallel component will be unaffected, whereas the perpendicular component will result in a magnetic force given by Equation (17.2). Writing this mathematically, the perpendicular component of velocity can be written as $v \sin\theta$, where (see Figure 17.7) θ is the angle between the vectors $\vec{v}$ and $\vec{B}$, so that in general

$$F_M = qvB \sin\theta. \tag{17.5}$$

We can conclude from this that for a given speed, the magnetic force on the ion will be greatest when the velocity is perpendicular to the magnetic field.

With an arbitrary velocity, the net effect on the particle's motion is to produce a helical trajectory about the magnetic field direction as if the particle moved along the path of a stretched spring oriented along the field (Figures 17.8 and 17.9). This is so because the axial velocity along the field is constant, and the perpendicular component is turned about the field direction by the magnetic force, resulting in a helical motion. Circular or helical orbits are characteristic of the motion of charged particles in magnetic fields.

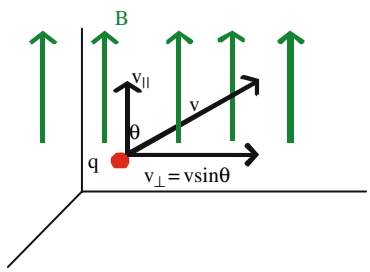


FIGURE 17.7 *Charge q moving in a uniform magnetic field at an arbitrary angle θ with respect to the field. The parallel component of $\vec{v}$ is unaffected, but the perpendicular component will be turned by a magnetic force.*

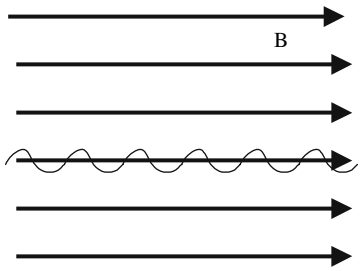


FIGURE 17.8 *In general a charged particle in a uniform magnetic field travels in a helix around the B field.*

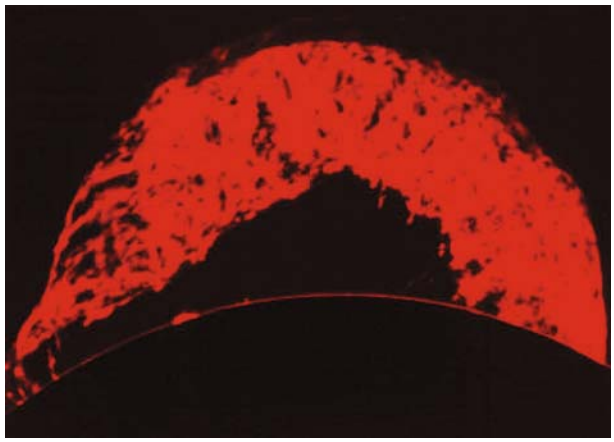
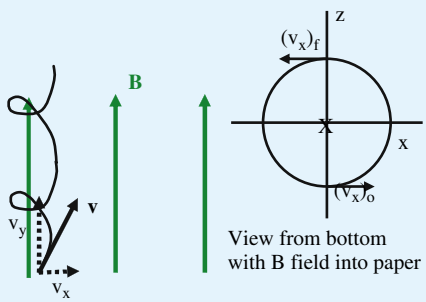


FIGURE 17.9 *Solar flares are due to charged particles moving in the magnetic field of the sun. The orbits are many Earth diameters with the charged particles spiraling around the magnetic field lines.*

Example 17.1 A proton enters a region where there is a uniform 0.5 T magnetic field along the y-axis between $y = 0$ and $y = 20$ cm. If the velocity of the proton has x- and y-components of $(4 \times 10^5, 6 \times 10^5)$ (m/s), find (a) how long the proton takes to travel through the B field region; (b) describe its trajectory; (c) find the point where it emerges from the B field and its velocity at that point.



Solution: (a) The y-component of velocity remains unchanged and so the proton takes a time

$$t = \frac{y}{v_y} = \frac{0.2}{6 \times 10^5} = 0.33\ \mu\text{s}$$

to emerge from the B field region. (b) Although it keeps the constant y velocity, its initial x velocity results in a centripetal force (ev_xB) when in the B field region that steers the proton in a circle in the x–z plane with the constant tangential speed of v_x (make sure you see why). Thus the overall motion of the proton is helical with the radius of the circle it travels in given from Equation (17.3) as

$$r = \frac{v_x m}{Be} = 8.4\ \text{mm}.$$

(c) We know the trajectory of the proton will bring it more or less straight through the B field, undergoing a small circular motion in the x–z plane as it travels along

(*Continued*)

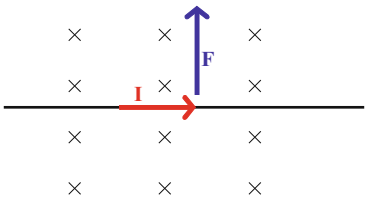


FIGURE 17.10 *A current carrying wire in a uniform magnetic field (directed into the page; the crosses representing tails of arrows) experiences a force given by Equation (17.6).*

the y-axis at 6×10^5 m/s. To find how many revolutions the proton makes around the B field we calculate its total circumferential distance traveled as $(v_x\, t) = 0.13$ m, so that dividing this by the circumference, $2\pi r$, we find that the proton has made 2.5 circular trips in the $x - z$ plane. Its net displacement from the start when it entered the B field region (see above figure) is then $\Delta x = 0$, $\Delta y = 20$ cm, and $\Delta z = 2r = 16.8$ mm. It emerges with its y-velocity unchanged and its x-velocity turned around because it has made a net 1/2 turn in the $x - z$ plane, so that its x-, y-, z-components of velocity are $(-4 \times 10^5, 6 \times 10^5, 0)$ (m/s). After it leaves this region it will maintain those constant velocity components.

We have seen that the magnetic force on a moving charged particle is always perpendicular to its velocity. Because this is generally true and because the velocity is always directed along the instantaneous displacement of the charge, *magnetic forces can never do any work*. This follows from the general definition of work because it is only the component of the force along the direction of the displacement that can do work. In the case of magnetic forces, this component is always zero and so, in general, no work can ever be done by magnetic forces. Magnetic forces only steer a charge's velocity vector, but do not change its magnitude. Thus, the charge's kinetic energy does not change and from the work–energy theorem we can also conclude that no work is done by magnetic forces.

Thus far in our discussion of the magnetic force on charges we have only considered isolated charged particles. When charges flow through a conductor that is in a region where there is a magnetic field, these charges will also experience a magnetic force. As shown in Figure 17.10, if a wire with a current flowing is oriented perpendicular to a magnetic field, then according to Equation (17.1) the moving charges will experience a force transverse to the wire as well as to the direction of the B field. The actual moving charges are free electrons (going to the left in Figure 17.10), experiencing an upward force as shown (using Equation (17.1) and remembering that the electrons have negative charge). As the electrons move in response to this force they attract the positive charges of the wire so that the entire wire feels an upward force as shown. From now on we analyze magnetic forces assuming that the charge carriers are positive and moving along the current direction. Check that you get the same direction for the magnetic force from Equation (17.1) in this case. If the wire is free to move then the kinetic energy the wire gains as it moves cannot be due to the magnetic force, because this force can never do any work; the wire's increased kinetic energy comes from the work done by the electric field from the electrons pulling on the positive charges of the wire. Of course, such an electric force will only exist if there is a current flowing in the wire to produce the charge separation.

We can adapt Equation (17.1) to the case of a current I flowing along a wire perpendicular to a uniform B field to find the net magnetic force on the wire. In a length L of the wire, the total charge Q making up the electric current flowing with velocity v_{drift} can be found. Because the current I is given by the total charge Q divided by the time Δt for the charge to flow a distance L, $\Delta t = L/v_{\text{drift}}$ we can write that $Q = I\Delta t = IL/v_{\text{drift}}$, so that Equation (17.1) becomes

$$F_M = ILB. \qquad (L \perp B) \qquad (17.6)$$

If we use the same right-hand rule as used for Equation (17.1), curling the fingers of your right hand from the direction of L (taken as the direction of the current flow) toward B, your thumb determines the direction of the magnetic force (see Figure 17.4).

In this section we have seen that moving electric charges, whether making up a current in a wire or any other type of configuration can experience a magnetic force due to interaction with a magnetic field. The force law for this interaction is more complex, however, than that for the electric field interaction because it not only depends on the charge and field magnitude, but also on the magnitude of the velocity vector and its orientation with respect to the field (or equivalently the current and its orientation). Furthermore, we have seen that this interaction, in general, can do no work on charges. In the next two sections we look at two specific important applications of the magnetic force.

2. TORQUE AND FORCE ON A MAGNETIC DIPOLE

At the end of the last section we considered the magnetic force on a straight current-carrying wire in a uniform magnetic field. Another important geometry of current flow, the *current loop*, is worthy of its own discussion. A current loop is a generic term for a simple circuit with a single closed loop, regardless of the exact trajectory of the current. Its importance lies not only in actual conducting wire circuits, but also in its use as a model for understanding the magnetic properties of matter through atomic electron current loops.

In Section 4 we show that a current loop, or in fact any current carrying wire, generates its own characteristic magnetic field. Here we wish to examine the forces acting on a current loop placed in an external uniform magnetic field. Consider the rectangular current loop in Figure 17.11 lying in a region of uniform B field as shown. In this orientation, the two edges that are parallel to the magnetic field have no force acting on them, whereas the other two edges perpendicular to the B field each have a force on them given by Equation (17.6). Because the current direction is opposite in those two wire segments, the corresponding forces act in opposite directions to create a couple (the torque due to equal and opposite forces) about the horizontal axis shown in the figure. There is no net force acting on the loop but the net torque acting will tend to produce a rotation of the loop as shown.

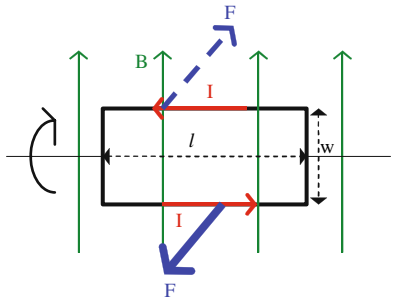


FIGURE 17.11 *A current loop in a uniform magnetic field. The two forces shown are perpendicular to the plane of the paper as determined by the right-hand rule for Equation (17.6).*

Using the dimensions of the loop shown, we can calculate the net torque acting on the current loop about its central axis in the orientation shown in Figure 17.11 to be

$$\tau = I\ell B\frac{w}{2} + I\ell B\frac{w}{2} = I\ell wB = IAB, \tag{17.7}$$

where $w/2$ is the lever arm and $A = \ell w$ is the area of the loop. If the loop is able to rotate, the couple will produce a rotation of the loop about the axis of rotation as shown. Equation (17.7) gives the maximum torque acting on the loop because, as can be seen in the side view shown in Figure 17.12, the lever arm distance changes with the orientation of the loop. With θ equal to the angle between the B field and the normal to the plane of the loop, the lever arm can be written as

$$r_\perp = \frac{w}{2}\sin\theta,$$

so that in general the torque on a current loop in a uniform B field becomes a function of the rotation angle

$$\tau = \mu B\sin\theta, \tag{17.8}$$

where we have introduced the *magnetic dipole moment* $\mu = IA$.

The magnetic dipole moment is a vector quantity, just as is the electric dipole moment, and we choose its direction to be perpendicular to the plane of the current loop. A simple second right-hand rule indicates which of the two directions perpendicular to the current loop plane is correct: if the fingers of your right hand are curled along the direction of current flow in a wire loop, your thumb will point in the proper direction of the magnetic dipole moment. Of course, if the current direction reverses so does the direction of the magnetic dipole moment, in accord with this right-hand rule. Note that if, instead of a single loop, we have a circuit with a tightly wound helical loop of N turns, we can replace this with N identical loops each having the same area and current so that the magnetic dipole moment of the circuit is $\mu = NIA$. Also note that Equation (17.8) is very similar to the equation for the torque on an electric dipole moment in an electric field (Equation (15.13))

$$\tau = pE\sin\theta,$$

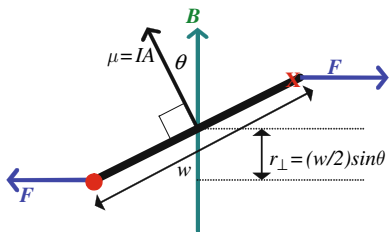


FIGURE 17.12 *Side view of a current loop in a uniform magnetic field. The normal to the loop makes an angle θ with respect to the* B *field. The two forces that produce a net torque are shown with the moment arm* $r_\perp$ *as well as the magnetic dipole moment* $\mu = IA$ *along the normal to the loop. The net torque tends to align the magnetic dipole moment with the magnetic field.*

where the electric dipole moment p and electric field have their analogs in the magnetic dipole moment and magnetic field. We conclude that although a current loop in a uniform magnetic field will experience no net force it will feel a torque tending to align the magnetic dipole moment with the magnetic field direction just as an electric dipole tends to align along an electric field.

Example 17.2 A circular coil of radius 15 cm and made of 100 turns has a resistance of 100 Ω and is attached to a 12 V battery with light flexible wires through a switch. If the normal to the coil is oriented at 45° to a uniform 2 T magnetic field, find the torque on the coil when the switch is closed.

Solution: When the switch is closed the current in the coil is $I = V/R = 0.12$ A. The magnetic dipole moment of the coil is $\mu = NIA = 0.85\ \text{Am}^2$ (equivalent units are Nm/T or J/T). The net torque on the coil is then $\tau = \mu B \sin 45° = 1.2$ Nm.

Our discussion thus far in this section has been limited to uniform magnetic fields. Qualitatively, it is easy to see that if the magnetic field varies in magnitude with distance at the current loop then the loop will feel a net force. Because the force on each current segment is given by Equation (17.6) and the B field in that equation will be different at the two sides of the loop, the forces producing the torque shown in Figure 17.12 will be different in this case. The same will apply if the B field changes direction across the loop. In a nonuniform magnetic field there will be a net force on a current loop, as well as a torque tending to align the magnetic dipole moment with the magnetic field. In the next section we show that the first experiment to detect the intrinsic spin of the electron involved sending a beam of atoms through a nonuniform magnetic field in which a net force acted on them that depended on the orientation of the intrinsic magnetic dipole moment of their electrons.

Before concluding this section we mention that there is also an effective energy associated with the interaction of a magnetic dipole moment with an external magnetic field. From the general equation for work, we found that the work done by a net torque in rotating an object by a small angle is (see Equation (7.20))

$$\Delta W = \tau \Delta\theta.$$

In the case of a magnetic dipole this becomes $\Delta W = \mu B \sin\theta\, \Delta\theta$. For a given magnetic dipole moment in a uniform magnetic field B, this expression leads to a potential energy given by

$$PE_\mu = -\mu B \cos\theta, \tag{17.9}$$

analogous to the expression for the potential energy of an electric dipole in an electric field (see Equation (15.14)). We return to Equation (17.9) in the next section as well as in our discussions of magnetic resonance techniques in the next chapter where we show its fundamental role. It is important to reiterate here that although the expressions for the force, torque, and potential energy of a magnetic dipole involve the magnetic field, magnetic fields do no work. As we discussed in the last section, it is always an electric field that must be responsible for any work and therefore for any changes in mechanical energy.

3. THE STERN–GERLACH EXPERIMENT AND ELECTRON SPIN

Electron spin was first demonstrated experimentally by Stern and Gerlach in 1922 in an experiment in which a beam of silver atoms was passed through a nonuniform magnetic field and then detected. According to Equation (17.9), the magnetic dipole moment μ of an atom interacts with the magnetic field with an interaction energy

$$PE_\mu = -\mu B \cos\theta,$$

where the magnetic dipole moment is proportional to the total angular momentum of the atom. At the time of this experiment the angular momentum of atoms was thought to consist entirely of angular momentum of the electrons orbiting around the nucleus. If the nonuniform B field is along the z-axis then there will be a force on the atom according to the general relation

$$F_z = -\frac{\Delta PE}{\Delta z}$$

(see Equation (15.8)). In our case if B varies in the z-direction, then there will be a net force given by

$$F_z = \mu \cos\theta \frac{\Delta B}{\Delta z} = \mu_z \frac{\Delta B}{\Delta z}$$

which is only present if the B field is nonuniform.

Now classically we might expect that, since the orientation of the atoms in the beam is random, the magnetic dipole moment would have no preferred direction and the narrow initial beam would be spread out symmetrically in the z-direction as shown in Figure 17.13a. We show in Chapter 25 that, in fact, the magnetic dipole moment cannot point in any direction, but is spatially quantized to point only in directions that result in certain discrete values of the z-component of the dipole moment. At the time of this experiment the predictions of quantum mechanics were that for atoms with a net angular momentum, the z-component of its value μ_z, takes on values that are small integer multiples of a basic amount, say A, so that $\mu_z = nA$, where n, an integer, might, for example, equal 1, 0, or -1 (having a positive, zero, or negative component along the z-axis), or perhaps in another case $n = 3, 2, 1, 0, -1, -2, -3$. Quantum mechanical predictions always had an odd number of possible values for μ_z centered on zero. A beam of these atoms, after passing through the nonuniform magnetic field, would separate into an odd number of beams, each with a different z-component of μ, with one beam, the $n = 0$ beam, undeflected.

Of course, with the magnetic field turned off, the beam of atoms would travel straight through without any deflection. Also, if the atoms had no orbital angular momentum, so that it was thought $\mu = 0$, it was expected that the entire beam would pass through with no deflection. However, when the experiment was done with silver atoms that had no orbital angular momentum, or a few years later when the experiment was repeated with ground state hydrogen atoms with a single electron with no orbital angular momentum, the result was as depicted in Figure 17.13b. Rather than no deflected beam as expected (or an odd number of beams centered around an undeflected beam as would be expected if the atoms actually did have some orbital angular momentum) only two distinct components were detected with each deflected in the opposite direction and no undeflected beam. This result was interpreted in terms of two spatially quantized components of an intrinsic magnetic moment of the electron, known as electron spin. The Stern–Gerlach experiment was an early dramatic confirmation of the spin hypothesis, discussed further in Chapter 25.

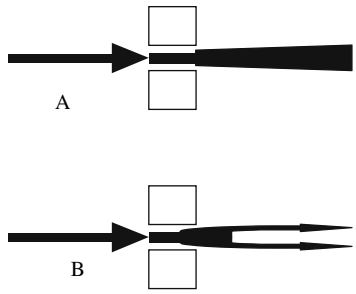


FIGURE 17.13 *Schematic of the Stern–Gerlach experiment. An atomic beam passes through a nonuniform magnetic field. In* A*, the classically expected continuously spread beam, a result not observed; in* B *the actual observed splitting of the atom beam with no orbital angular momentum into two components, explainable only by introducing electron spin.*

4. PRODUCING B FIELDS

In our discussion of electric charge, the fundamental quantities were the individual electric charges which produce electric fields. These could be positioned to form electric dipoles or more complex arrangements of charges. Despite much effort looking for an expected symmetry between magnetism and electricity in this regard, individual magnetic charges, or *magnetic monopoles*, have never been found in nature. The magnetic dipole moment introduced in Section 2 is the most elementary magnetic quantity known. Effective "circulating currents" of electrons in an atom produce a magnetic dipole moment, as do the individual charged particles of the proton and electron due to their intrinsic "spin". We show that such elementary

magnetic dipoles are ultimately responsible for permanent magnets. Compass needles, for example, behave as magnetic dipole moments, orienting along the magnetic field direction because of a torque on them according to Equation (17.8).

The lack of symmetry between electricity and magnetism has the important consequence that magnetic fields are not simply produced by the presence of magnetic charge as electric fields are produced by electric charges. Instead, all magnetic fields require the motion of electric charges, either as macroscopic or microscopic currents. There is, however, a natural symmetry between the form of the interactions of electric charges with electric and magnetic fields and before discussing how to produce B fields, we digress a bit in order to mention this.

Earlier in this book we learned that electric charges produce electric fields that in turn can interact with other electric charges. We show shortly that moving electric charges produce magnetic fields that, as we have just seen, can interact only with other moving electric charges. These are very powerful statements. With some thought about the relative nature of velocities, we can reach a very significant conclusion. Electric charges that are at rest as seen by one observer will be in motion as seen by a second observer moving relative to the first. Therefore, given our statements about electric and magnetic fields, the first observer at rest with respect to the charges will measure only electric fields resulting in forces on other stationary charges, whereas the second observer will see the same charges to be moving and producing magnetic fields and feeling magnetic forces from other B fields as well.

When this apparent paradox is carefully considered, it leads to the undeniable conclusion that electric and magnetic fields are manifestations of the same underlying phenomenon, one that we call the *electromagnetic field*. Different observers may detect different combinations of electric and magnetic fields but the theory of electromagnetism makes consistent predictions for all of the observables, including the complete description of the trajectories of all the charges, as it must in order to be a correct theory. We return to some of these ideas in the next chapter, but first turn to a discussion of the production of magnetic fields.

Consider a long straight wire with a constant current I flowing along it as shown in Figure 17.14. Experimentally it is found that there is a magnetic field produced by the current. This finding, first discovered in 1820 by Oersted, initiated the linkage known as *electromagnetism* between the previously separate subjects of electricity and magnetism. The magnetic field was found to depend on the magnitude and direction of the current and on the perpendicular distance r from the wire. The B field is proportional to the current in the wire and inversely proportional to r so that

$$B = \frac{\mu_0}{2\pi}\frac{I}{r}, \tag{17.10}$$

where $\mu_0/2\pi$ is the constant of proportionality. The constant μ_0 is known as the permeability of the vacuum and is exactly equal to $4\pi \times 10^{-7}$ T · m/A. Do not confuse it with a magnetic dipole moment. It is the fundamental constant of magnetism, playing the role of ε_0 for electricity. The magnetic field is found to point tangentially along circles centered on the wire lying in a transverse plane. A compass held near the wire would have its needle always pointing perpendicular to a radial line from the wire to the compass giving the tangential direction for B as shown in the figure.

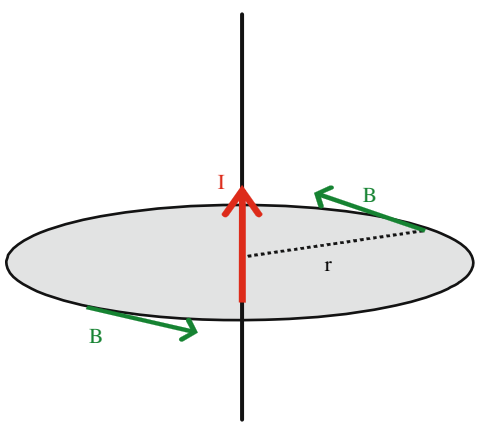


FIGURE 17.14 *The magnetic field of a long straight current-carrying wire.*

Magnetic field mappings can be constructed in a similar way to the electric field mappings of Chapter 14. There we used the notion of a small positive "test charge" to sample the electric field at various points in space to determine the electric field. The magnitude and direction of $\vec{E}$ at a point in space were obtained in principle from the magnitude and direction of the force on the test charge at that position. By "moving" the test charge around, the electric field lines could be mapped out. In an analogous way we can imagine using a small "test compass" to map the magnetic field lines. The magnitude and direction of the magnetic field at a point can be obtained in

principle from the magnitude of the torque on the compass needle and its orientation, respectively, when it is placed at that point. For the long straight wire just discussed the $\vec{B}$ field mapping consists of concentric circles centered on the wire as shown in Figure 17.15.

The direction of the $\vec{B}$ field around the circles is determined by the direction of the current flow. The same right-hand rule as for the magnetic dipole moment indicates the proper direction of the magnetic field: put the thumb of your right hand along the direction of the current and your fingers curl in the proper direction of the magnetic field lines around the wire. Reversing the current direction reverses the clockwise/counterclockwise nature of the B field circles.

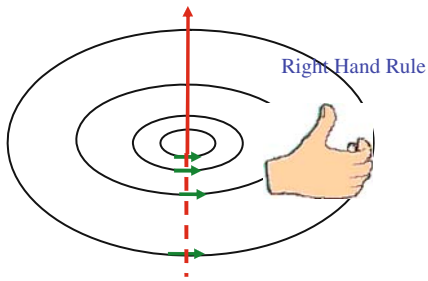
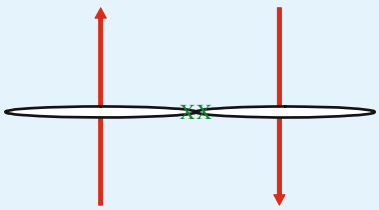


FIGURE 17.15 *The magnetic field of a long straight current-carrying wire is along the tangent to concentric circles in the direction given by the right-hand rule.*

Example 17.3 Two long parallel wires a distance of 20 cm apart carry equal and opposite currents of 10 A. Find the magnetic field midway between the wires.



Solution: Each wire produces a magnetic field with a magnitude of $B = \mu_0 I/2\pi r$. Using the right-hand rule the wire on the left produces a B field into the paper at the midpoint shown, as does the wire on the right. Therefore the net B field is the sum of the two magnitudes and is equal to

$$B = \frac{\mu_0 I}{\pi r} = \frac{4\pi \times 10^{-7}(10)}{\pi(0.1)} = 4 \times 10^{-5}\ T.$$

In mapping electric fields we saw that the electric field lines either started or ended on electric charges. Because there are no magnetic charges, magnetic field lines must be closed curves. In adding magnetic fields from different currents, the same principle of superposition applies as in the case of electric fields. For example, the net magnetic field due to two long straight equal current-carrying wires with the current flowing in opposite directions, the example just worked out midway between the wires, is shown in Figure 17.16.

Another interesting example is the magnetic field produced by a current-carrying circular loop. A cross-section through a diameter of the loop gives a magnetic field mapping similar to that of Figure 17.16. Close to the wire, the wire appears to be nearly straight and this nearby current dominates to produce a magnetic field nearly circular around the wire. Along the axis of the loop the field from the entire loop adds to produce a more uniform field that can be enhanced by looping the wire many times, each loop increasing the field near the center. Several coiled geometries are commonly used. One of these, the *solenoid* or helical coil of wire, is used to produce a large uniform magnetic field along its axis (Figure 17.17).

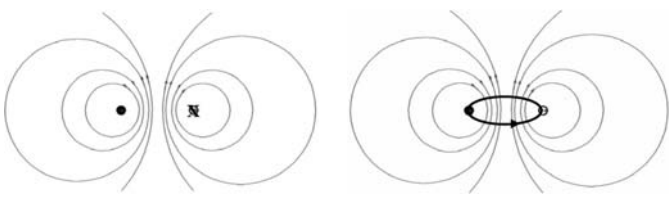


FIGURE 17.16 (left) *Magnetic field lines for two long straight current-carrying wires perpendicular to the plane of the page each carrying equal and opposite currents.* (right) *Same mapping for a coil of current-carrying wire.*

Permanent magnets produce their magnetic fields from the net magnetic dipole moment of their constituent atoms. As we saw in the last section, all atoms have orbiting electrons that constitute atomic current loops with a magnetic dipole moment. In addition the intrinsic electron spin produces a spin magnetic dipole moment.

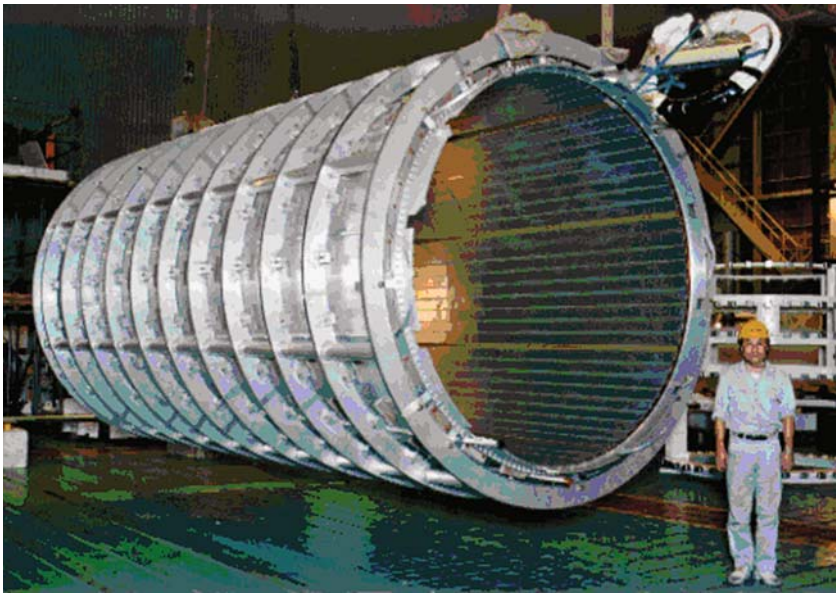


FIGURE 17.17 *A very large solenoid to be used to steer elementary particles at the CERN accelerator in Geneva, Switzerland. The section seen here is wrapped with 8 km of superconducting wire to create a 2 T magnetic field along the central axis.*

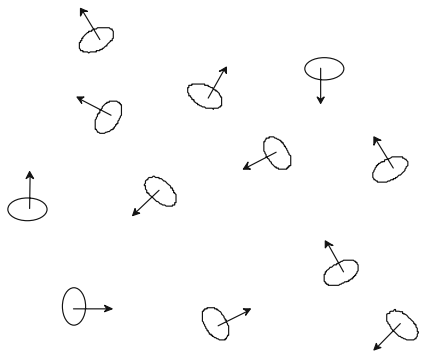


FIGURE 17.18 *A random array of atomic magnetic moments resulting in no net dipole moment.*

Usually these magnetic moments are randomly oriented and, on average, produce no net magnetic dipole moment and hence no macroscopic magnetic field (Figure 17.18). *Ferromagnetic materials*, including iron, nickel, and cobalt, have long-range interactions between their magnetic dipole moments causing large numbers to align in *magnetic domains*, small but macroscopic regions with dimensions of 0.01–0.1 mm. If many of these domains are aligned, as in permanent magnets, there is a large effective magnetic dipole moment produced that creates an external magnetic field similar to one produced by a current loop or solenoid (compare left and right in Figure 17.19). The two poles of a magnet, the north N and south S, are labeled so that the external magnetic field lines go from north to south (Figure 17.19; and the internal field lines complete closed loops and run from south to north).

Microcrystals of magnetite found in various bacteria, plants, and animals are single domains with a characteristic size of about 50 nm. Domains smaller than about 40 nm would have magnetic dipoles too small to be effective due to thermal motions, whereas domains larger than about 80 nm would tend to divide into multiple domains. By aligning a collection of 20 or so single domains in a magnetic filament, the magnetic dipole moment of magnetotactic bacteria is sufficiently strong to act very much like a compass needle to orient the bacteria's swimming direction, as mentioned at the beginning of this chapter.

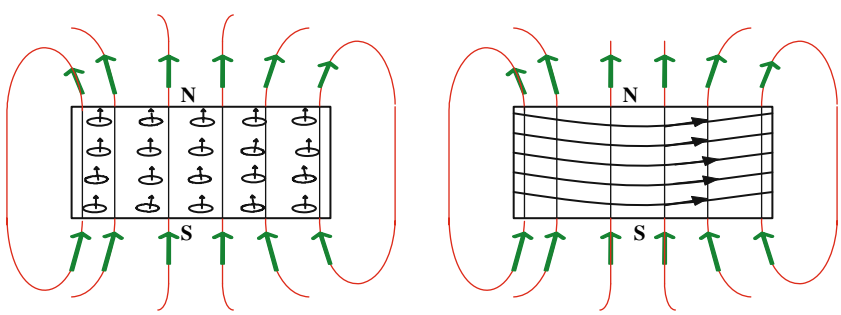


FIGURE 17.19 (left) *Aligned ferromagnetic domains in a magnet with its effective magnetic field;* (right) *effective surface winding currents producing the same magnetic field as the magnet.*

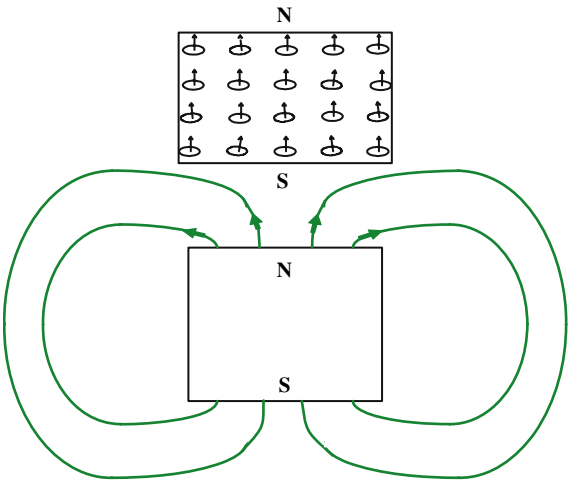


FIGURE 17.20 *A strong permanent magnet, with its dipole magnetic field, causing the alignment of the ferromagnetic domains of a second material, above. The induced magnetic field is such as to create a south pole near the permanent north pole of the magnet, producing a net attraction of the magnets due to the interaction of the induced dipoles with the nonuniform B field of the magnet.*

FIGURE 17.21 *An electromagnet on a crane used to pick up scrap metal.*

Ferromagnetic materials with unaligned domains can be magnetized by placing them in a strong magnetic field that serves to increase the extent of domain alignment and to increase the size and strength of the net magnetic domain. These materials then show *induced magnetism*. The induced magnetism occurs through the same mechanism that produces an aligning torque on a current loop in a magnetic field. Figure 17.20 shows a permanent magnet, with its external dipole field, causing the alignment of the effective magnetic dipole moments of the magnetic domains of a second ferromagnet. It can be seen that the induced magnetism is such as to produce the opposite magnetic pole nearest the magnet so that there is a net attractive force. The detailed reason for the attractive force hinges on a nonuniform B field and the discussion in Section 3. It is this phenomenon that explains the ability of a permanent magnet to pick up paper clips or to "stick" to a refrigerator door containing iron. Induced magnetism does not occur in nonmagnetic materials such as aluminum, copper, ceramics, or plastics and thus magnets will not "stick" to these materials.

Ferromagnetic materials can be demagnetized by heating them above a characteristic temperature, known as the Curie point, or Curie temperature. For example, iron loses its ferromagnetic properties above a temperature of 1040 K.

An *electromagnet* makes use of induced magnetism to create very large magnetic fields. A solenoid wrapped around an iron rod can be used to create an axial magnetic field down the bore of the rod that aligns magnetic domains, enhancing the magnetic field typically by a factor of several thousand. One common geometry has two such solenoids with iron cores and with current windings in the same direction placed in close proximity with a small gap between the magnet "pole faces". The strong axial magnetic field generated in the gap is then nearly the same as the internal field in the iron core. The strongest electromagnets are limited to producing B fields of about 0.5 T, with joule heating of the coils the limiting factor in these magnets (Figure 17.21). To produce larger B fields requires the use of wires that are superconducting, having effectively no resistance to current flow and thus no joule heating. Superconducting magnets are routinely used in MRI machines, for example, where magnetic fields of several T are used.

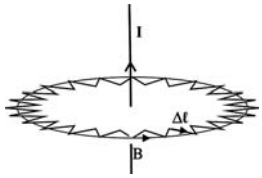


FIGURE 17.22 *A long current-carrying straight wire with a concentric Amperian loop composed of many short segments and the* B *field directed circumferentially.*

5. *AMPERE'S LAW

This section is optional. Subsequent material does not depend on this section. Starred questions and problems at the end of the chapter refer to this optional section.

We've just seen in the last section that magnetic fields can be produced by a long straight current-carrying wire according to Equation (17.10). This result is limited to that particular configuration. What is the general law that governs the magnetic field produced by some arbitrary configuration of electric current? It is given by Ampere's law, the magnetic analog of Gauss's law, studied in Chapter 14. There we saw that the flux of the electric field over some Gaussian surface is related to the total charge enclosed within the Gaussian surface. Electric field lines start and stop only on electric charges; but there are no magnetic charges (monopoles) found in nature. This statement gives rise to another fundamental law: Gauss's law for magnetic fields, in which the flux of B over a Gaussian surface is equal to zero. Because there are no magnetic charges, magnetic fields arise from electric currents and Ampere's law takes a different, but analogous form.

In place of the flux of E, Ampere's law involves the "circulation" of B around an "Amperian loop." We choose a closed path, or Amperian loop, constructed of many short segments of length $\Delta\ell$ as shown in Figure 17.22. The circulation of B is calculated by forming the products $B_{\parallel}\Delta\ell$, where is the component of B parallel to the segment $\Delta\ell$, and adding these up all around the closed path

$$circulation = \Sigma B_{\parallel}\Delta\ell.$$

Ampere's law states that the circulation around a closed curve is proportional to the net current that passes through, or is enclosed by, the Amperian loop

$$\Sigma B_{\parallel}\Delta\ell = \mu_0 I_{\text{enclosed}}. \tag{17.11}$$

The following two examples illustrate how to apply Ampere's law to calculate the magnetic field produced by currents. Just as for Gauss's law, Ampere's law is generally true (as long as the currents are steady currents; Ampere's law is not universally true, but needs some modification in the case of time-varying electric currents, made by Maxwell, as we show in Chapter 18). However, in order to be able to calculate B fields directly from Ampere's law, there must be sufficient symmetry as we now show.

Example 17.4 Calculate the B field produced by a long straight wire-carrying current I.

Solution: This is the same problem as discussed in the previous section and we know that the solution should be given by Equation (17.10), so that this is a good check for us. Refer back to Figure 17.14 for a sketch of the situation. By the symmetry of the problem we know that the B field lies in concentric circles around the wire with current I. Therefore let us choose an Amperian loop that is a circle, centered on the wire, of radius r. Then we know that the B field is constant in magnitude on the Amperian loop and the circulation is given by $circulation = \Sigma B_{\parallel}\Delta\ell = B(r)\,2\pi r$, where B is already tangent to the circle all the way around and has a constant value that depends only on the radius r, namely $B(r)$, and the sum of all the $\Delta\ell$ segments is just the circumference of the circle. Then, from Ampere's law, we have

$$B(r)\,2\pi r = \mu_0 I,$$

or solving for B we have

$$B(r) = \frac{\mu_0 I}{2\pi r},$$

in agreement with Equation (17.10).

Example 17.5 Calculate the B field within a solenoid of radius R carrying current I made from N turns of wire and having a total length L as shown in Figure 17.23.

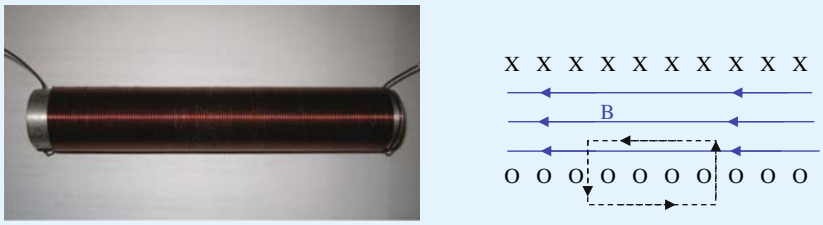


FIGURE 17.23 (left) *A solenoid made from winding wire around a metal core;* (right) *a longitudinal cross-section of the center portion of a solenoid showing the cut windings with current flowing around from the bottom to the top (counterclockwise, as viewed from the* left*), producing a uniform* B *field to the left. The rectangular Amperian loop is used to evaluate Ampere's law in order to find the* B *field inside the solenoid.*

Solution: With the coils wound tightly, the magnetic field inside the solenoid will be parallel to its axis except at the ends whereas the magnetic field outside the solenoid will be very weak, again except near the ends. We choose an Amperian loop in the form of a rectangle, shown in Figure 17.23 (on the right) and evaluate the circulation around the loop by adding the contributions from each side. The segment outside the solenoid contributes a negligible amount because the B field is extremely small. Similarly the two segments that lie perpendicular to the inside B field do not contribute because there is no parallel component of B along these. Only the segment of length ℓ inside the solenoid contributes an amount

$$\text{circulation} = B\ell$$

so that Ampere's law becomes

$$B\ell = \mu_0 (N\ell/L)I,$$

where the term in parentheses represents the number of turns of the solenoid contained within the Amperian loop. Each turn of the coil contributes a term I to the total enclosed current, therefore we need to insert the total enclosed current on the right-hand side of Ampere's law. Solving this equation for B, we find that

$$B = \mu_0 (N/L)I = \mu_0 nI,$$

where $n = N/L$ is the number of turns of the solenoid per unit length. Note that the length of the Amperian loop does not and cannot enter the final answer, because the Amperian loop is invented by us and not a part of the original problem. We see that the magnetic field inside the solenoid is uniform (this is approximate for a finite length solenoid) and only depends on the current in the coil and the number of turns per unit length, or the current per unit length. This result is somewhat similar to that for the E field within a capacitor, $E = \sigma/\varepsilon_0$, which only depends on the charge per unit area on the plates.

CHAPTER SUMMARY

The magnetic force on a charge q moving with a velocity v in a direction making an angle θ with a magnetic field B is given by

$$F_M = qvB \sin\theta. \qquad (17.5)$$

Note that in order to feel a magnetic force the charge must be moving and have a component of velocity perpendicular to the magnetic field. If the charge is moving perpendicular to $\vec{B}$ then its orbit will be a circle around the B field direction (its orbit will lie in a plane perpendicular to $\vec{B}$). In general the orbit will be a helix with its axis along the B field direction and with this axial velocity component remaining constant. The mass spectrometer uses magnetic forces to measure the mass m of small ions by causing them to orbit in circles of radius r after acceleration through a potential difference V according to

$$m = \left(\frac{er^2}{2V}\right)B^2. \qquad (17.4)$$

A current (I) carrying wire of length L lying perpendicular to a magnetic field will also experience a net force given by

$$F_M = ILB. \qquad (L \perp B). \qquad (17.6)$$

The magnetic analog of the electric dipole is a small current loop, constituting a magnetic dipole moment given by

$$\mu = IA,$$

where A is the area of the loop. If such a loop is placed in a uniform magnetic field it will not experience any net force, but will feel a torque given by

$$\tau = \mu B \sin\theta, \qquad (17.8)$$

where the angle θ lies between the magnetic field and the normal to the area of the loop. Such a loop will also have a magnetic potential energy given by

$$PE_\mu = -\mu B \cos\theta. \qquad (17.9)$$

If the magnetic dipole lies in a nonuniform magnetic field then there will also be a net force that will deflect the dipole along the direction of the field variation. This effect was used in the Stern–Gerlach experiment in which a beam of atoms with no orbital angular momentum, and hence no expected magnetic dipole moment, was deflected into two separated beams by a spatially varying magnetic field. Such a deflection was the first experimental evidence for electron spin.

Magnetic fields are produced by electric currents; there are no magnetic charges (monopoles). A long straight current-carrying wire produces a magnetic field a distance r away given by

$$B = \frac{\mu_0}{2\pi}\frac{I}{r}, \qquad (17.10)$$

where μ_o is the magnetic permeability of the vacuum. The field is directed in circles around the wire with a direction given by a right-hand rule. A solenoid can be used to produce a uniform magnetic field throughout its interior, in a similar way to that by which a capacitor produces a uniform electric field between its plates. Permanent magnets, or ferromagnets, are produced by microscopic circulating currents in ordered domains.

Ampere's law relates the circulation of the magnetic field around a closed Amperian loop to the total enclosed current that passes through the loop:

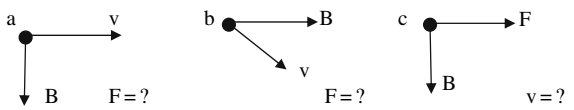
$$\Sigma B_\parallel \Delta\ell = \mu_0 I_{\text{enclosed}}. \qquad (17.11)$$

Although Ampere's law is generally true for steady currents, solving it for the B field produced by current distributions is difficult without sufficient symmetry. In general Ampere's law needs to be modified for time-varying currents in order to be universally true. We show this in the next chapter as one of Maxwell's equations.

QUESTIONS

1. What trajectories of motion can a charged particle have in the region of space where there is a uniform magnetic field? Consider a charge moving parallel, perpendicular, and at an arbitrary direction to the field.
2. Discuss the following statement. Because the magnetic force is proportional to the strength of the magnetic field, it is always the case that a larger magnetic field will produce a larger acceleration of a charged particle.

3. What is wrong with the following argument? Because the work done by a force is the product of the force and the displacement and because the magnetic force is proportional to the magnetic field, the larger the magnetic field is, the greater the work done on a charged particle moving in the field.
4. A positively charged particle sits at rest at the origin. If only a uniform magnetic field is applied in the x-direction, describe the motion of the particle. If only an electric field is applied along the y-direction, describe the particle's motion. If both fields are simultaneously applied with the directions as given, describe the particle's motion.
5. Describe the motion of a positively charged particle at the origin under each of the following circumstances: (a) the particle is initially moving along the x-axis and a uniform magnetic field lies along the y-axis; (b) the particle is initially moving along the x-axis and a uniform magnetic field lies along the x-axis; and (c) the particle is initially moving in the $x - y$ plane at a 45° angle between the positive axes and a uniform magnetic field lies along the x-axis.
6. A mass spectrometer is to be used to separate hydrogen from deuterium (containing an extra particle, a neutron, with essentially the same mass as a proton, both in the nucleus). What will be the ratio of the radii of curvature at which the two are detected? Which will have the larger radius of orbit?
7. Find the direction of the missing vector in the following diagrams for a positively charged particle.



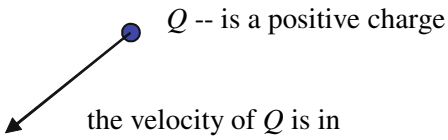
8. A long straight horizontal wire oriented along the N–S direction has a constant current flowing towards the north. What is the direction of the magnetic force on the wire due to the Earth's magnetic field, remembering that there will be a vertical component of the Earth's field? Given that the Earth's magnetic field is typicaly about 0.5×10^{-4} T is the wire much influenced by this force?
9. A square loop of wire forms a current loop. When placed in a uniform magnetic field the four sides of the loop will, in general, experience a magnetic force. Describe the net force and net torque on such a current loop oriented in the $x - y$ plane with the current flowing clockwise as viewed from above when placed in a uniform magnetic field that is oriented along (a) the z-axis or (b) the x-axis. Which situation has the greater magnetic potential energy?
10. A current loop is oriented with its normal along the direction of a uniform B field throughout the region. Compare two such loops with current flowing in opposite directions. Is there a torque acting on either loop? Which one is in stable equilibrium?
11. Why does the Stern–Gerlach experiment require the atom beam to travel through a nonuniform magnetic field?
12. A long straight wire has a constant current flowing along it. A small square current loop, also with a constant current flowing through it, sits in the magnetic field of the straight wire. At what orientation of the loop is the torque on it a maximum? A minimum? At what orientation of the loop is its potential energy a maximum? A minimum? Draw appropriate sketches for each part.
13. Find the direction of the magnetic field produced by a long straight vertical wire with current flowing up at the following points in a horizontal plane. Indicate your answers using N–S and E–W directions as appropriate. (a) A point east of the wire; (b) a point north of the wire; (c) a point 45° southeast of the wire.
14. What is the difference between ferromagnetism and paramagnetism. Which type of magnetism is the kind that holds refrigerator magnets up?
15. What is the purpose of the iron core in an electromagnet?
16. Magnetic bacteria in waters in the northern hemisphere swim toward magnetic north, which has a downward component, causing the bacteria to swim toward the murky bottom where they can feed. What magnetic bacteria behavior would you expect to find if you traveled to the southern hemisphere?
17. The electric field of a long straight line of charge with linear charge density λ is

$$\vec{E} = \frac{\lambda}{2\pi\varepsilon_0 r}\hat{r}.$$

If these charges move along a wire, making up an electric current, we have seen that a magnetic field is produced with a magnitude given by Equation (17.10) and oriented in circles around the wire. Contrast these two expressions, discussing similarities and differences.
18. *In electrostatics, electric fields must start and end on electric charges. What do you expect the circulation of the electric field to equal?
19. *Discuss the differences between the magnetic field produced by a current traveling in a single circular loop versus the field produced by a solenoid of the same radius.

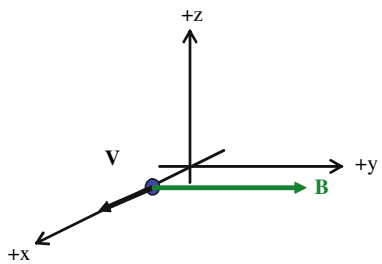
MULTIPLE CHOICE QUESTIONS

1. In which direction is the magnetic force on Q due to the current I in the figure shown? The force points (a) into the page, (b) out of the page, (c) in the same direction as the velocity, toward the wire, (d) to the "northwest" (vertical component up, horizontal component to the left).



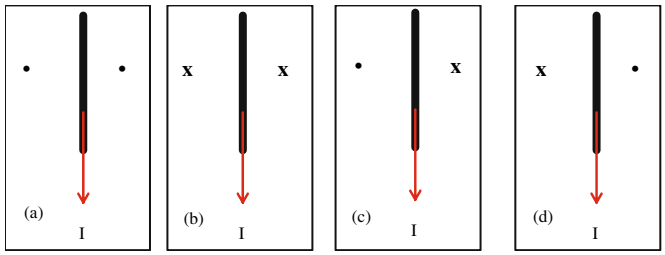

the current I in a long straight wire is coming out of the page

2. The figure below shows a proton ($q = +e$) at one instant with velocity V pointing along the positive x-axis. The proton is moving through a magnetic field B that points along the positive y-axis. At the instant shown the magnetic force on the proton (a) points in the positive x-direction, (b) points in the positive y-direction, (c) points in the positive z-direction, (d) is zero.



3. Suppose that the proton in the previous question were an electron ($q = -e$) instead. Which one of the following would then be true? (a) The magnetic force on the electron would have the same magnitude and direction as on the proton. (b) The magnetic force on the electron would have the same magnitude but the opposite direction as on the proton. (c) The magnetic force on the electron would be much smaller than on the proton because the mass of the electron is much smaller than that of the proton. (d) The magnetic force would still be zero.

4. With our usual convention that the symbols **x** and • mean "into the page" and "out of the page," respectively, which of the following pictures best depicts the direction of the magnetic field in the plane of the paper due to the associated wire?



5. A negatively charged particle traveling east in a horizontal plane enters a region where there is a uniform magnetic field pointing vertically down. The initial force on the particle points (a) north, (b) south, (c) east, (d) west.

6. A uniform 1 T magnetic field is directed vertically downward in a region of space. A proton traveling at 10^5 m/s in a horizontal plane and aimed northward enters this region. The initial acceleration of the proton is (a) 1.6×10^{-14} m/s^2 westward, (b) 9.6×10^{12} m/s^2 eastward, (c) 1.6×10^{-14} m/s^2 eastward, (d) 9.6×10^{12} m/s^2 westward.

7. A proton is spiraling around 1 T uniform magnetic field lines along the x-axis. If the x-velocity of the proton is 10^5 m/s and the radius of the circular projection of the trajectory in the $y - z$ plane is 1 mm, the average velocity of the proton is (a) 1.4×10^5 m/s, (b) 1.96×10^5 m/s, (c) 10^5 m/s, (d) 0 m/s.

8. A 5 m long straight wire carries a current of 5 A directed north (ignore the rest of the circuit). The magnitude of the force on this section of wire in a region where a uniform 5 T magnetic field points south is (a) 125 N, (b) 25 N, (c) 0 N, (d) depends on how the current completes the circuit.

9. A uniform magnetic field of 1.5 T pointing north extends over a large region of space. A small ball carrying charge +1 pC travels east at 10^3 m/s. After the charge has traveled 0.01 m in this field the work done on it by the magnetic force (a) is zero, (b) is 1.5×10^{-9} J, (c) is 1.5×10^{-11} J, (d) cannot be calculated because the path is circular.

Questions 10 and 11 refer to a current loop lying in the $x - y$ plane and having a magnetic dipole moment of 0.1 A-m^2 along the z-axis. A uniform 2 T magnetic field lies in the $x - z$ plane at a 30° angle to the x-axis.

10. The net torque on the current loop has a magnitude (a) 0.2 Nm, (b) 0.1 Nm, (c) 0.17 Nm, (d) 0 Nm.

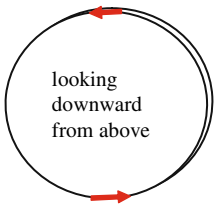
11. The potential energy of the current loop is (a) −0.2 J, (b) 0.2 J, (c) −0.17 J, (d) −0.1 J.

12. Two long straight wires are parallel to each other, a distance of 4 m apart, and each carries a current of 5 A but in opposite directions. The magnetic field along the midline between the wires is (a) 0 T, (b) 10^{-6} T, (c) 5×10^{-7} T, (d) 2.5×10^{-7} T.

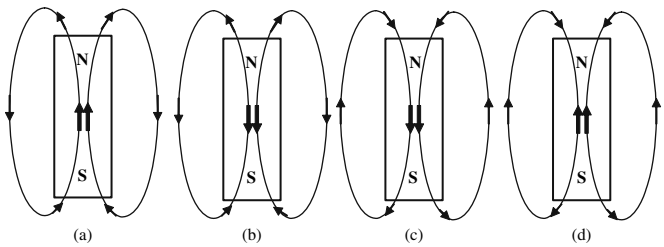
13. Suppose in the previous question the two wires have their currents in the same direction, say vertically upwards on the paper. The magnetic field at a point 1 m to the right of the wire on the left (a) is zero, (b) points out of the paper, (c) points up along the wire, (d) points into the paper.

14. The figure below shows two circular loops of the same diameter, each carrying a current circulating in the same direction. They are held in place so that their faces are parallel to each other. Their centers lie on a line that is perpendicular to both faces. The magnetic force on the upper loop due to the lower loop tries to

make the (a) upper loop smaller and pull it downward, (b) upper loop smaller and push it upward, (c) upper loop larger and push it upward, (d) upper loop larger and pull it downward.



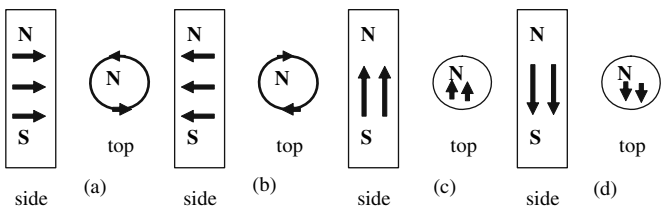
15. Which one of the following best illustrates magnetic field lines associated with the bar magnets shown?



16. Which one of the following is true? (a) The end of a compass labeled N is a magnetic north pole and the Earth's geographic north pole is also a magnetic north pole. (b) The end of a compass labeled N is a magnetic north pole and the Earth's geographic north pole is a magnetic south pole. (c) The end of a compass labeled N is a magnetic south pole and the Earth's geographic north pole is also a magnetic south pole. (d) The end of a compass labeled N is a magnetic south pole and the Earth's geographic north pole is a magnetic north pole.

17. The magnetic field of a bar magnet comes from (a) magnetic monopoles sprinkled throughout the magnet, (b) one end being positively charged, the other negatively, (c) atomic currents running up and down the length of the magnet in its interior, (d) atomic currents circulating around the outside surface of the magnet.

18. The figures show a cylindrical bar magnet viewed looking at a side and looking down on the top. The magnetic field is produced by electrons moving coherently together over the surface of the magnet. Which picture shows the proper direction of the electron flow?



19. *Three long parallel vertical wires each carry current I, with two having current traveling upwards and one downwards. What is the total circulation of B around a closed curve containing all three wires? Is it (a) $\mu_0 I$, (b) $2\mu_0 I$, (c) $3\mu_0 I$, (d) 0?

20. *Suppose a current I is traveling down a long thin-walled cylindrical shell of radius R and you would like to find the B field outside the shell using Ampere's law. A good choice for an Amperian loop would be (a) a sphere of radius $r > R$, (b) a cylinder of radius $r > R$ and length L, (c) a circle of radius $r > R$, (d) a square with sides of length $r > R$.

PROBLEMS

1. How fast must an electron travel in an extremely large magnetic field (30 T) so that the force on it will be as large as the force on a single myosin muscle protein from the chemical energy of one ATP molecule, 3 pN or 3×10^{-12} N? This should indicate to you that even large constant magnetic fields can exert only negligible forces on the atoms in our body. In fact, people are routinely exposed to such large DC magnetic fields in magnetic resonance imaging (see the next chapter) without any ill effects.

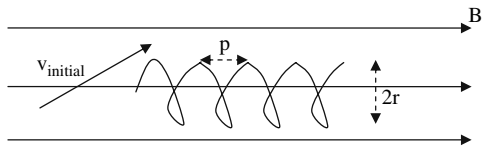
2. A proton moving at a constant velocity of 10^6 m/s enters a region of uniform magnetic field perpendicular to its velocity. If the magnetic field is 5 T, find the force on the proton. What will the force be if the proton's velocity makes a 45° angle with the field direction?

3. How fast must a proton be traveling to be steered in a 0.02 m radius circle by a 6 T uniform magnetic field?

4. An electron initially traveling at a speed of 10^5 m/s enters a region where there is a uniform magnetic field of 2 T oriented 45° to its initial velocity. Quantitatively describe the trajectory of the electron, including its path, the orbit radius and average velocity of the electron.

5. A beam of electrons is accelerated from rest through a potential difference of 100 V. What uniform magnetic field perpendicular to the beam is needed to steer it in a circle of 5 cm radius?

6. An electron enters a uniform magnetic field of magnitude $B = 0.5$ T at a 45° angle to the magnetic field's direction. What is the radius r and the pitch p (distance between loops as shown below) of the electron's helical path assuming its speed is 0.05c?



7. Suppose that in a charge-to-mass ratio experiment electrons are accelerated from rest through a potential difference of 200 V and then travel through a region

of magnetic field along a curved path due to a magnetic force exerted on them. The radius of the path is measured to be 7.5 cm.

(a) What is the velocity of the electrons as they leave the "accelerator"?

(b) What is the magnitude of the magnetic field, assuming that it is perpendicular to the beam?

(c) What is the angular velocity of the electrons?

(d) What are the frequency and period of orbit of the electrons?

8. A mass spectrometer can detect the different isotopes of an ionized element. If Zn^{2+} ions are accelerated through a 10 kV potential and enter a 10 T magnetic field region, calculate the different radii for the isotopes ^{64}Zn and ^{66}Zn, where the numbers refer to the atomic weight in atomic mass units.

9. Fill in the details in the derivation of Equation (17.4).

10. A 0.5 m length of rigid metal rod is connected, through two springs (with 10 N/m spring constants) perpendicular to the rod at either end, to a circuit that supplies 2.5 A of current through the wire. If the rod is perpendicular to a uniform 2 T magnetic field, find the distance each spring is stretched. What force is the origin of the work done to stretch the springs?

11. A 5 cm diameter circular loop of wire carries a 2 A current. When placed in a 0.5 T magnetic field, find the maximum and minimum torque that can act on the loop and sketch the loop orientation relative to the B field for each case.

12. What is the magnetic potential energy of a 2 cm diameter circular loop carrying a 1A current when placed in a 2 T uniform magnetic field after reaching its stable equilibrium position?

13. A solenoid consists of 4000 closely wound turns with a circular cross-section of 4 cm diameter. If a 0.5 A current flows through the solenoid, calculate its magnetic dipole moment.

14. A beam of neutral hydrogen atoms with a speed of 10^4 m/s along the x-axis enters a 1 m long region where there is an inhomogeneous magnetic field pointing in the z-direction but with a gradient (variation) in the z-direction given by 0.01 T/m. Given the atom's magnetic dipole moment of 9.27×10^{-24} J/T, compute the following.

(a) The acceleration of the atoms while in the magnetic field region.

(b) The possible deflections in the z-direction of the atoms as they pass through the magnetic field region, depending on whether the atom's magnetic moment is along the $+z$ or $-z$ direction.

15. A long straight vertical wire carries an 8 A current upwards. Find the magnetic field (magnitude and direction) at the following points in a horizontal plane relative to an origin at the wire: (a) 2 m east of the wire and (b) 4 m south of the wire.

16. Two long vertical wires are separated by 2 m. The one on the left carries a current of 8 A upward and the other carries a current of 5 A downward.

(a) Find the magnetic field at the midway point between the wires.

(b) Where is the location of the line at which the magnetic field vanishes?

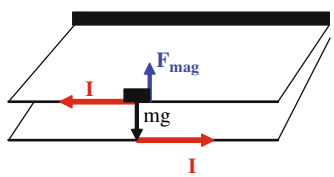
17. Three long parallel wires each carry a 2 A current with two of the currents up and one down. The wires pass through the corners of an equilateral triangle with 0.2 m sides. Find the magnetic field (magnitude and direction) at the center of the triangle.

18. Two long parallel wires separated by a distance d each carry a constant electric current I. Find an expression for the magnitude of the force per unit length of wire (F/L) on each wire from the magnetic field of the other wire. Also find the direction of the force on each wire if the two currents are either parallel or antiparallel.

19. A current balance is a device that has two parallel rigid wires carrying the same current in opposite directions. The bottom wire is fixed and the other one is attached in such a way that it can pivot in response to a force from the second wire (see figure). First the pivot is adjusted so the top wire is in equilibrium with no current flowing, and then the current is turned on. By adding external weight to the top wire it can be kept at its equilibrium separation distance and the magnetic force between the wires can be determined. This device can be used to calibrate current by direct force measurement.

(a) Write down the B field produced by the bottom fixed wire (assuming it to be infinite) and determine that it will produce an upward force on the top current carrying wire.

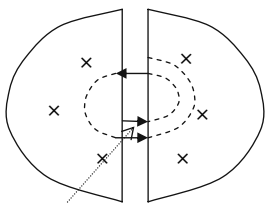
(b) Compute the force on the 40 cm long top wire if both currents are equal to 10 A and the separation distance is 0.5 cm, and thereby determine the mass needed to be added to the top rod to keep it at that separation distance. Note that these forces are not large.



20. The magnetic field produced by the ionic currents of the heart tissue is about 5×10^{-11} T, about a million times weaker than the Earth's magnetic field. (These weak fields can be directly recorded using modern technology in a magnetocardiogram. Special detectors known as superconducting quantum interference devices or SQUIDs, invented in 1970, can measure both constant and alternating magnetic fields even 1000-fold weaker

than that of the heart.) How far would you have to be from a long straight wire carrying a 1 A current to produce the same magnetic field as the heart, assuming it is the constant value given above?

21. A *cyclotron* consists of large magnets (called dees because they are in the shape of the letter D) with a small gap between them as shown in the figure. An accelerating voltage is applied across the gap and charged particles, such as protons, are accelerated across the gap and then enter a region where a uniform magnetic field steers them in a semicircle to return to the gap. The accelerating voltage polarity is then reversed and so the particle accelerates further, returning across the gap and entering the opposite region of magnet field, where it is steered around in a semicircle again. This process is repeated many times to accelerate the particle to high speeds in a relatively small region of space.



(a) First show that if the particle of mass m and charge q has a speed v and the uniform magnetic field is B, then it will travel in a semicircle of radius $r = mv/qB$.

(b) Then show that the particle will travel in the semicircle in a time $t = \pi m/eB$ that is independent of the radius of the orbit. This allows the cyclotron to have a constant frequency of oscillation of the accelerating voltage given by $f = 1/(2t)$ as long as the particle energy is below about 50 MeV. Beyond this relativity effects occur and the time does vary with particle velocity or radius of orbit.

22. A cyclotron is sometimes used for carbon dating. Carbon-14 and carbon-12 ions are obtained from a sample of the material to be dated and are accelerated in the cyclotron. If the cyclotron has a magnetic field of magnitude 2.40 T, what is the difference in cyclotron frequencies for the two ions?

23. *Rail guns*, like those, for example, in the movie *Eraser*, have been suggested for launching projectiles into space without chemical rockets and for ground-to-air antimissile weapons of war. A tabletop model rail gun consists of two long, parallel, horizontal metal rails 3.50 cm apart. A projectile of 5.0 g mass is placed on a metal bar that rests across the two rails, is originally at rest at the midpoint of the rails, and is free to slide without friction. When the switch is closed, electric current is established in the circuit consisting of the rails and the bar, both having low electric resistance. Suppose that the battery used in the rail gun produces a current of 30 A clockwise, as viewed from above.

(a) Find the magnitude and direction of the magnetic field at the midpoint of the bar immediately after the switch is closed, assuming that the magnetic field is due to two long straight wires.

(b) The magnetic field varies along the bar in such a way that as you move closer to either wire B increases. Assuming that the average effective B field along the length of the bar is 10 times larger than the field at the midpoint, what is the magnitude and direction of the force on the bar?

(c) What is the acceleration of the bar when it is in motion? Is it constant?

(d) What is the velocity of the bar after it has traveled 1.0 m to the end of the rails?

(e) Suppose that instead of the velocity in part (d), you wanted a velocity that was larger by a factor of 50. What current would you need to produce this velocity, everything else being the same?

24. *Current I travels along the axis of a long cylindrical shell of radius R, with a negligible thickness. Use Ampere's law to find the B field both inside and outside the shell.

25. *Suppose that the current I traveling down a long straight (nonmagnetic) solid cylindrical wire of radius R is uniformly distributed throughout the cross-section of the wire, so that the current density is $I/\pi R^2$. Find the magnetic field using Ampere's law both inside and outside of the wire as a function of r, the distance from the wire's axis.

26. *A solenoid of 25 cm length is wound with 10,000 turns of wire and has a radius of 3 cm. Find the magnetic field inside the solenoid when a current of 2 A is flowing.

27. *Three long straight parallel wires each carry a current $I = 1.5$ A, two in the same direction and one in the opposite direction. In a transverse plane, the wires lie at the vertices of an equilateral triangle of side $L = 10$ cm. Using Ampere's law and the principle of superposition, start from scratch and find the magnetic field at the center of the triangle.

18

Electromagnetic Induction and Radiation

In this chapter we generalize our discussion of magnetic fields in the previous chapter to include time-varying magnetic fields. A new phenomenon arises, known as electromagnetic induction, in which a time-varying magnetic field actually produces an electric field. This is described by a new fundamental law of electromagnetism known as Faraday's law. nuclear magnetic resonance (NMR) is discussed in some detail as an application of this new physics. We also discuss the use of NMR techniques to allow medical imaging within the human body, a tremendously growing and beneficial method to image soft body tissue in three dimensions with remarkably high resolution. As NMR imaging developed and became clinically used, the medical community quickly decided to change the name of the method to Magnetic Resonance Imaging (MRI), dropping the term nuclear, present in NMR because of the role of nuclear magnetic moments but having the incorrect connotation of nuclear energy and radiation. An analogous technique using Electron Spin Resonance (ESR) is also briefly discussed.

A final major piece of electromagnetism, the fact that changing electric fields can produce magnetic fields, allows us to qualitatively summarize in words the complete theory of electromagnetism, given by a set of four equations known as Maxwell's equations. We show how these equations lead to the production of electromagnetic radiation. At the outset it should be made clear that electromagnetic radiation is quite distinct from nuclear radiation which has its source in the nuclei of atoms. There are many different forms of electromagnetic radiation, including visible, ultraviolet, and infrared light as well as microwaves, x-rays, radio waves, and so on, all comprising the entire electromagnetic spectrum discussed in the next chapter. All of these have some common features that we study in this and the next chapter.

1. ELECTROMAGNETIC INDUCTION AND FARADAY'S LAW

If a current-carrying coil is placed in a uniform magnetic field we have seen that it will feel a torque tending to orient the coil with its normal (the direction of its magnetic dipole moment) along the field direction. Suppose that a wire forming a complete (or closed) circuit, but with no battery or power supply, is placed in a uniform magnetic field. Because there is no current source in the circuit, no current flows and the loop experiences no torque. Remarkably, however, if an external torque is applied to make the loop rotate so that the direction of its normal relative to the B field changes, a current will flow in the wire while it is rotating (Figure 18.1). This is an example of the phenomenon of *electromagnetic induction*; the current that flows in this situation is known as an *induced current*.

Consider another example in which we have an isolated loop of wire. If a permanent bar magnet is held near the loop and moved toward it, an induced current will flow in the wire while the magnet is moving (Figure 18.2). Once the magnet stops,

J. Newman, *Physics of the Life Sciences*, DOI: 10.1007/978-0-387-77259-2_18,

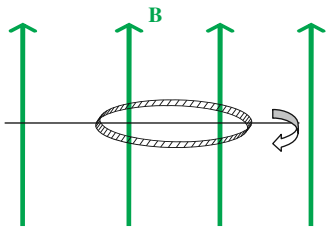


FIGURE 18.1 *A loop of wire in a uniform* B *field. When rotated, for example, about the horizontal axis shown, an induced current is produced as long as the loop is changing its orientation with respect to* B.

there is no longer any induced current. If the magnet is moved away from the loop, the induced current flows in the reverse direction, but again only while the magnet is moving. Alternatively, if the same loop of wire sits near a solenoid, a tightly coiled helix of wire as shown in Figure 18.3, when the switch is closed sending current from the battery through the solenoid, there will be a brief pulse of induced current in the loop. If the switch is opened up, there will also be a brief pulse of induced current flowing in the opposite direction while the solenoid current drops to zero.

What is the fundamental connection in these three seemingly different examples that leads to an induced current in a wire loop physically isolated from a battery or power supply? In the latter examples of a bar magnet moving closer to or farther from the loop or in a solenoid having its magnetic field turned on or off, the common feature is a changing magnetic field at the loop. However, the first example of a rotating loop takes place in a uniform magnetic field, so the general phenomenon cannot be due simply to a changing magnetic field magnitude at the loop.

Experiments show that three factors affect whether electromagnetic induction occurs in a circuit. There must be a time variation in either the local magnetic field at the circuit, in the orientation of the magnetic field relative to the plane of the loop, in the area A of the loop itself, or in some combination of these three. The link between these factors comes from the general condition that there must be a change in time of the quantity known as the *magnetic flux* Φ_B, defined as

$$\Phi_B = \overline{B}_\perp A = \overline{B} A \cos\theta, \tag{18.1}$$

where $\overline{B}_\perp$ is the component of the average magnetic field $\overline{B}$ perpendicular to the face of the wire loop and θ is the angle between the magnetic field and the normal to the loop (Figure 18.4). Magnetic flux is defined in exact parallel to electric flux introduced in connection with Gauss's law in Section 7 of Chapter 14 (Equation 14.11). Thus Φ_B can vary with time from changes in any of the three factors making up its definition in Equation (18.1).

Before considering magnetic flux in more detail, we mention that if in any of these examples we were to replace our single isolated loop with an isolated tightly wound coil of N turns, we can approximate this situation by imagining that we replace the coil with N identical stacked loops. This is commonly done for reasons that are clear shortly. In that case, the total magnetic flux through the isolated coil when placed in the same B field is simply N times the flux through a single loop given by Equation (18.1).

We can think of magnetic flux as a measure of the number of magnetic field lines that cross the area of the loop. This number is affected by any of the three factors discussed above. Suppose first that our loop is oriented with its normal parallel to the field lines (thus $\theta = 0$ in Figure 18.4). The stronger the magnetic field is, the denser the field lines and therefore the more will cross the area of the loop, and thus the greater the magnetic flux. If the loop is rotated (increasing θ), then just as in the case of shooting arrows towards a bulls-eye target that is rotated so that the projected area seen by the archer is decreased, the magnetic flux will decrease. When the loop is oriented at $\theta = 90°$, the projected area is zero and no magnetic lines can cross the loop so that the magnetic flux is zero. So a variation in the angle θ clearly affects the magnetic flux, as does the actual area A of the loop itself. From the three examples above, we've seen that there must be a time variation in at least one of these three variables appearing in the magnetic flux to produce electromagnetic induction.

We can now write the general statement of the law of electromagnetic induction, known as *Faraday's law*, named after Michael Faraday for his discoveries in the 1830s.

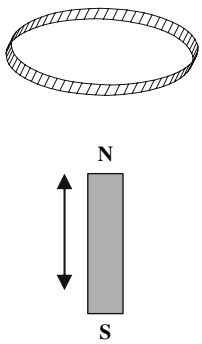


FIGURE 18.2 *As a bar magnet is moved relative to the loop, an induced current flows.*

Faraday's Law relates the average induced emf in terms of the time rate of change of the total magnetic flux through a coil

$$\varepsilon = -\frac{\Delta \Phi_B}{\Delta t}. \tag{18.2}$$

Recall that the emf ε is the "driving energy per unit charge" supplied by a source such as a battery or power supply. Although it is related to the potential V, having the same units, we give it a different name and symbol because it has its own energy source, for example, chemical energy for a battery. Recall also that the emf produces an average electric field E within the conducting material whose magnitude is given by

$$\overline{E} = \frac{\varepsilon}{\Delta x}, \tag{18.3}$$

where Δx is the length of wire (similar to Equation (15.7)). Whenever there is a changing magnetic flux in a coil, Faraday's law states that the coil will act as a battery, producing an induced emf and generating an electric field within the wire that will produce a charge flow.

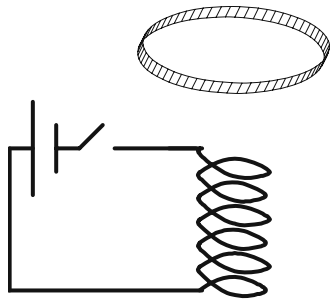


FIGURE 18.3 *When the switch is closed, so that the solenoid has a current that changes from zero to a constant final value, an induced current will flow in the isolated loop during the brief time while the solenoid current changes.*

Thus far we have introduced Faraday's law by discussing currents flowing in a circuit and noting that a time-varying current, which produces a time-varying B field, results in an induced emf which in turn produces an E field. Because electric charges that move at constant velocity produce steady currents, time-varying currents are produced by accelerating electric charges. On a more fundamental level, we can state that *accelerating electric charges, producing time-varying magnetic fields, will also generate electric fields*. Electric fields generated in this way are fundamentally different from electrostatic fields. Electrostatic fields must start and stop on stationary or constant velocity electric charges and they never form closed loops. Electric fields produced by accelerating charges, whether making up an electric current in a wire of a circuit or in empty space, form closed loops. We have seen that this is true for closed loops of wire where the electric field acts continuously around the wire to produce the induced current, but it also holds in empty space. In Section 4 of this chapter we show that as an important consequence of this, accelerating electric charges produce electromagnetic radiation. Faraday's law is one of the fundamental laws of electromagnetism.

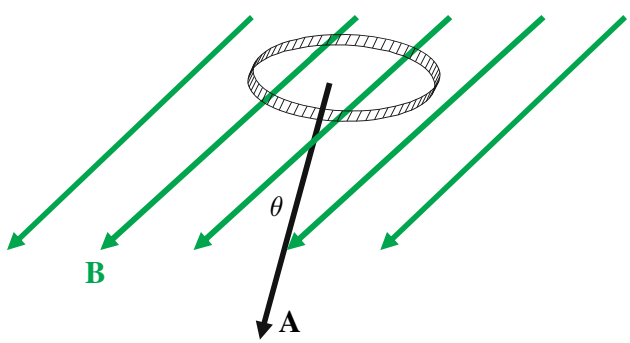
In the examples we mentioned above, depending on the sign of the change in the magnetic flux (whether the bar magnet moves toward or away from the loop, or whether the magnetic field from the solenoid is increasing or decreasing when the switch is opened or closed) the induced current will flow one way or the other. The negative sign in Equation (18.2) indicates the proper direction for the induced emf and current flow. Its meaning is that

> *the induced emf is always of a polarity such as to oppose the change of magnetic flux that created it.*

This statement on the polarity of the induced emf is known as *Lenz's law*.

Lenz's law has its foundation in the conservation of energy principle. Consider the solenoid example of Figure 18.5. When the switch is closed, there is a brief time during which the current increases from zero to a final value. Because the windings are right-handed, the current flows clockwise (viewed from below) so as to produce an increasing magnetic field upwards along the solenoid axis in the figure as shown. During that brief time while the B field also increases (in tandem with the current increase) to its final value, Faraday's law tells us that an induced emf is produced in the upper coil, leading to an induced current flow in that isolated solenoid circuit. Lenz's law then states that the magnetic field produced by this induced current will be in a direction leading to the creation of a downward B field to oppose the increasing upward B field from the first solenoid. The downward B field produces a downward magnetic flux that tends to cancel the increasing upward magnetic flux from the B field of the lower circuit. Be clear that these effects only last for the very short time during which the lower circuit's current is changing when the switch is closed. To produce a downward B

FIGURE 18.4 *A coil of area* A *with its normal oriented at an angle* θ *with respect to a uniform* B *field. Only when either* A, θ, *or* B *changes at the loop will there be an induced current flowing in the coil.*



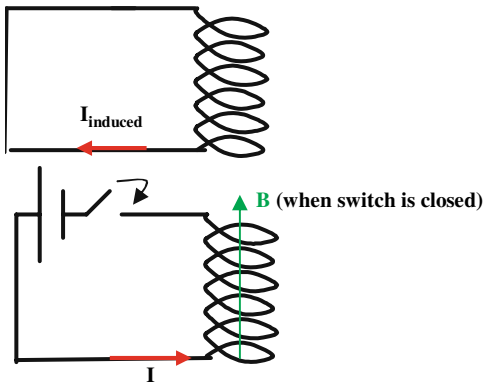


FIGURE 18.5 *The lower coil* (with right-handed windings) *produces an increasing* B *field in the direction shown for a brief time when the switch is closed. This changing magnetic flux at the upper solenoid coil* (also with right-handed windings) *produces a brief induced current in the direction shown. Once the* B *field reaches its final value, the induced current vanishes.*

field, the induced current must flow around the upper loop with the opposite polarity as shown.

As a proof of this, consider what would happen if, in fact, the induced current actually flowed in the other direction (counterclockwise around the upper loop). Then this changing induced current in the upper solenoid would create its own changing local upwards B field that would add even further to the increasing upward magnetic field. This would produce yet a greater change in magnetic flux at the upper solenoid, causing a further induced current in the same direction as that produced by the lower circuit. The resulting change would then lead to a further flux change at the upper coil, and so on. This process of positive feedback would lead to a runaway increase of energy with no apparent source, violating conservation of energy. We conclude that the induced current must flow as Lenz's law states to provide a negative, rather than positive, feedback and to conserve energy. This discussion provides a proof of Lenz's law by contradiction that is generally applicable.

Example 18.1 What happens when the circular coil with N loops shown in Figure 18.6, in a uniform magnetic field B aligned with the loop axis, is suddenly pulled at two points along a diameter so that the coil collapses to a linear array of wires in a time T, as shown in the right side of the figure.

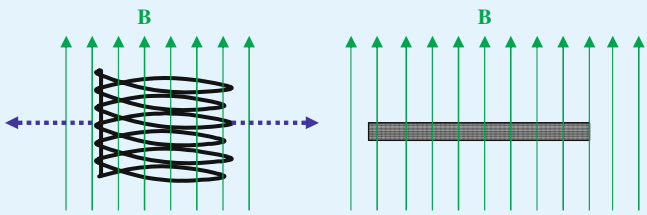


FIGURE 18.6 *A coil of* N *turns, initially with a uniform* B *field along its perpendicular axis* (left)*, is stretched along the dotted lines collapsing it to a linear dimension in a time T* (right)*. While the area is changing, there is an induced emf in the coil.*

Solution: Initially there is a net flux through the coil given by $\Phi_B = NBA$, where A is the cross-sectional area of one circular loop. The factor N appears because each turn in the coil contributes to the total flux. Because the flux is constant, there is no induced current in the coil. During the time T while the coil is stretched, shrinking its area to zero, the flux changes from its initial value to zero and there is a time variation of the magnetic flux. The average emf during this time interval has a magnitude given by $\varepsilon = NBA/T$, and only exists during the time T, after which the flux is zero and no longer varies with time. The direction of induced current flow is such as to maintain the original upward magnetic flux through the coil and, opposing the decrease, given a right-handed coil, is therefore clockwise around the collapsing coil as viewed from below.

Consider one application of this new physics: the basis of an electric AC (alternating current) generator. Shown schematically in Figure 18.7, a simple AC generator consists of a coil of wire, with N turns of area A, made to rotate at a uniform angular velocity ω in a region of uniform magnetic field B. As the coil rotates, the magnetic flux varies (co-)sinusoidally (because of the variation of cos θ = cos ωt

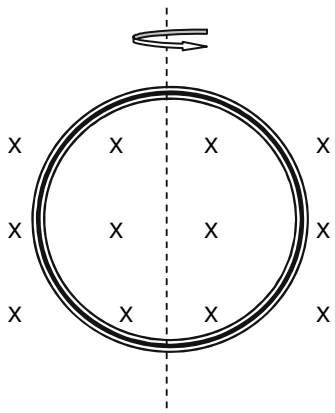


FIGURE 18.7 (left) *Diagram of a simple AC generator, with a coil rotating in a uniform magnetic field;* (right) *photo of an AC generator useful when the power is otherwise out.*

factor in Equation (18.1)) and so there is a corresponding continuous variation in the induced emf. Since the magnetic flux varies as

$$\Phi_B = NBA \cos \omega t,$$

the induced emf will be given by

$$\varepsilon = -\frac{\Delta \Phi_B}{\Delta t} = -NBA\frac{\Delta \cos \omega t}{\Delta t} = NBA\ \omega \sin \omega t = \varepsilon_{\max} \sin \omega t$$

(the middle mathematical step involves calculus—the differentiation of the cosine term, resulting in the product—$\omega \sin \omega t$). As the coil rotates the induced emf varies sinusoidally and is called an AC voltage. In the United States residential AC voltages are supplied at a frequency of 60 Hz and with a magnitude measured by an AC voltmeter of 120 V.

Example 18.2 A long straight wire lies along the x-axis and carries a constant 5 A current. Two identical 100 turn circular coils of 10 mm diameter and 500 Ω resistance each lie 5 m from the wire in the $x - y$ plane (with their normals perpendicular to the $x - y$ plane; see Figure 18.8), the first along the positive and the second along the negative y-axis. Find the average induced current in the appropriate coil(s) under the following circumstances. (a) If the first coil is rotated around an axis through its center parallel to the x-axis by 90° in a time of 10 ms. (b) If the second coil is continuously spun around an axis through its center along the y-axis direction at a rate of 4 turns/s. (c) If both coils are left as originally given and the current in the wire doubles in a time of 10 ms. In each case give the magnitude, direction, and duration of the average induced current. Assume that the flux is uniform over each small coil area.

Solution: (a) The magnetic field produced by the current in the straight wire is directed out of the paper at the location of the first (upper) loop and so the magnetic flux is maximal at the start. When the coil rotates by 90° the magnetic flux goes to zero in 10 ms and so the magnitude of the average induced emf is given by

$$\varepsilon = \frac{\Delta(NBA \cos\theta)}{\Delta t} = \frac{NBA}{t},$$

(*Continued*)

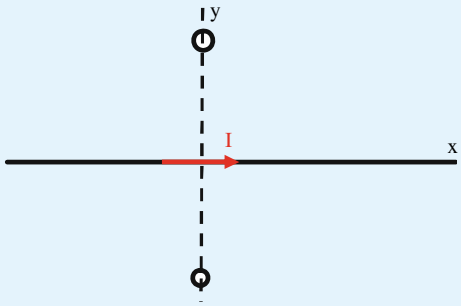


FIGURE 18.8 *Diagram for Example 18.2.*

where B is the field due to the straight wire, $B = \mu_o I/2\pi r$, r is the distance from the straight wire to the upper coil, N the number of turns of the coil, A the coil area, and t the rotation time. Putting in the numbers, we have that $B = 2 \times 10^{-7}$ T and $\varepsilon = 0.16\ \mu$V. From this we can find that the average induced current that flows is given by $\varepsilon/R = 310$ pA. To find the direction of the induced current note that as the coil is rotated about a diameter in the x-axis direction and the flux out of the paper decreases, the induced current is produced to oppose that decrease, so as to produce its own magnetic flux out of the paper. This means that the induced current flows around the loop in a counterclockwise direction, persisting only for the 10 ms during which the coil is rotated. There is no induced current in the second coil in this case because its magnetic flux doesn't change.

(b) In this case the B field at the bottom loop has the same magnitude as in part (a) but is directed into the paper at its location. As the coil spins about a diameter along the y-axis, the induced emf varies sinusoidally just as in the generator example and is given by $\varepsilon = NBA\ \omega \sin \omega t$, where $\omega = 2\pi \cdot 4$ rad/s. We find the induced current is $I = \varepsilon/R = 7.9 \sin(8\pi t)$ (in nA) and alternates direction as the coil rotates. The average induced current is actually zero in this part because the average of the sine curve is zero. In this case the first coil has no induced current.

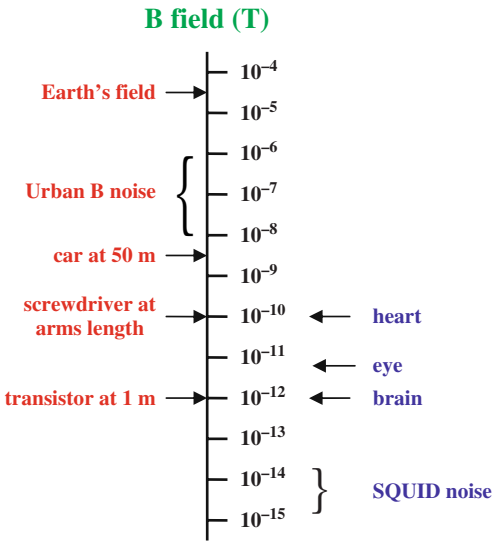
(c) In this case the flux through each coil changes due to the change in the B field from the wire and we have

$$\varepsilon = \frac{\Delta(NBA\cos\theta)}{\Delta t} = \frac{N\Delta BA}{t},$$

where the B field now doubles in a time $t = 10$ ms, so that $\varepsilon = 0.16\ \mu$V for both coils and the induced currents are both 310 pA. These are the same numbers as in part (a) because the magnitude of the flux change is the same as in that part. Using a similar argument as in part (a) the first coil will have a clockwise induced current, and the second will have a counterclockwise induced current. (Do you see why?)

Another application of Faraday's law is in the detection of time-varying magnetic fields. An isolated small coil, known as a search coil, connected to a very sensitive ammeter can be used to detect a time-varying magnetic field. Any time-varying magnetic flux at the search coil will result in an induced current that is detected by the ammeter. This basic technology is used in nuclear magnetic resonance measurements discussed in the next section. To measure the intrinsic weak magnetic fields of the human body requires a much more sensitive means of detection. Electric currents in the human brain produce local magnetic fields of only about 10^{-12} T. These should be compared to the Earth's static magnetic field of 10^{-4} T or even the spontaneous fluctuations in the Earth's magnetic field of 10^{-7} T (Table 18.1). Magnetic field noise due to electrical power lines can be as large as the Earth's magnetic field, making it extremely difficult to detect the small B fields produced in the brain.

Table 18.1 *Weak Magnetic Fields*



Measurement of the extremely weak magnetic field due to brain activity (*magnetoencephalography* or *MEG*) has been made possible by an exquisitely sensitive detector based on superconductivity known as a *superconducting quantum interference device* or *SQUID* (Figure 18.9). These devices operate at liquid helium temperatures (4.2 K) and can detect magnetic field fluxes as low as 10^{-15} T m^2. Even with our ability to measure such small magnetic fields, there is the issue of how to avoid the overwhelmingly larger local fields due to the Earth or stray power lines mentioned above. This is accomplished using detector coils arranged to be sensitive only to magnetic fields that vary rapidly with position in space (said differently, with non-constant spatial gradients) and are known as gradiometers (Figure 18.10; note that in the figure if both loops see the same time-varying B field, there is no net induced current; only if the two loops see different time-varying B fields will there be a detected current). With this arrangement, spatially slowly varying stray magnetic fields are not

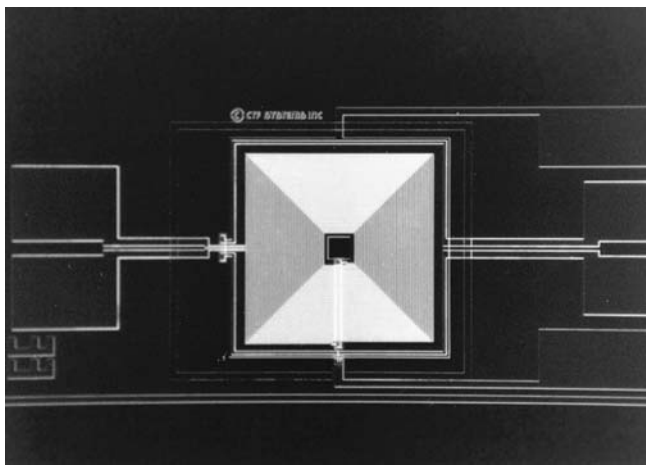


FIGURE 18.9 *A SQUID detector, several cm on a side, made using thin film technology.*

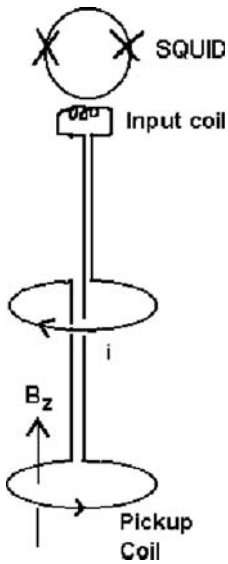


FIGURE 18.10 *Gradiometer with two coils that cancel out distant spatially constant B fields. Note the direction of the induced current flow in the two coils. Do you see how it works? See text.*

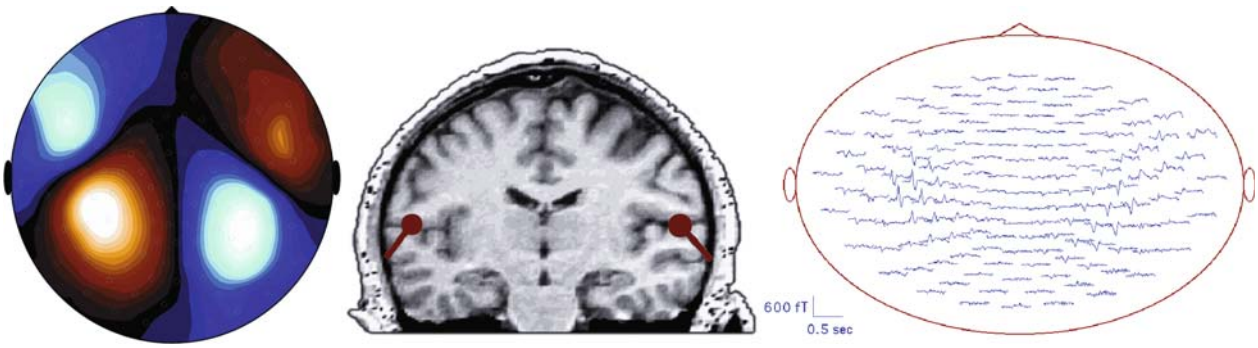


FIGURE 18.11 *MEG false color recording* (left) *of brain response to hearing pure tone,* (center) *superimposed on MRI cross-section of the brain.* (right) *mapping of the MEG signal used to generate the false color recording.*

detected and only the local fields that vary rapidly in space, even though they are very weak, are recorded. MEG recordings use arrays of over 100 SQUID detectors around the human head to measure local time-varying magnetic fields generated from the whole head. Figure 18.11 shows an example of some MEG brain mappings.

2. NUCLEAR MAGNETIC RESONANCE (NMR)

Nuclear magnetic resonance originated in 1946 in independent experiments by two groups of physicists in the United States. In the early 1950s the technique expanded to studying the structure of small organic molecules and there was an explosion of activity by organic chemists using NMR. By the mid-1960s the technique had been improved to the point where the first larger biologically important molecules were studied. Since that time the technique has undergone several huge leaps in progress and is now routinely used to study the detailed structure and dynamics of all sizes of biological molecules. Perhaps even more important is its application in medical diagnosis, using imaging methods, known as magnetic resonance imaging (MRI), discussed in the next section.

Our discussion of NMR begins with a brief description of a few properties of atomic nuclei. Remember that nuclei are extremely small ($\sim 10^{-15}$ m) composite structures containing protons and neutrons. Nuclei with an odd number of either protons and/or neutrons possess the property of *nuclear spin*, an intrinsic angular momentum. As a consequence of having both electric charge and nuclear spin these nuclei also have intrinsic magnetic dipole moments. One can think of these magnetic dipole moments as arising from the spinning electric charge of the nucleus. Some biologically important nuclei that have a magnetic moment, so that they can be studied using NMR, include ^{1}H, ^{13}C, ^{15}N, ^{17}O, and ^{31}P (the numbers indicate the total number of protons and neutrons in the nucleus). For now, we limit ourselves to the most common nucleus, that of the hydrogen atom, consisting of a single proton.

Quantum mechanics predicts that the magnetic moment of a nucleus is given by

$$\mu = \gamma_n S_n, \tag{18.4}$$

where S_n is the nuclear spin angular momentum and γ_n is called the gyromagnetic ratio, a quantity that is usually positive but whose value depends on the particular nucleus as well as its environment. The nuclear spin angular momentum, for a single unpaired spin, is given by

$$S_n = \frac{\sqrt{3}}{4\pi} h,$$

where h is Planck's constant, $h = 6.63 \times 10^{-34}$ J-s. For the isolated proton the gyromagnetic ratio is equal to $\gamma_p = 2.68 \times 10^{8}\ \text{s}^{-1}\ \text{T}^{-1}$. Its exact value depends on the

local environment of the proton (neighboring bonds and charges) and it is this variation that allows NMR to identify hydrogen protons attached to different atoms.

When placed in an external magnetic field, the nuclear magnetic dipole moment will interact with the field with an interaction energy given by

$$\mathrm{PE}_{\mu} = -\mu B \cos\theta$$

(see Equation (17.9)). For a nucleus of spin $\frac{1}{2}$ such as a proton, quantum mechanics also predicts that its magnetic dipole moment vector is not free to point in any direction at all, but must align so as to have a fixed component of nuclear spin along the magnetic field, as mentioned for the electron in connection with the Stern–Gerlach experiment in the last chapter. Figure 18.12 shows the possible orientations of the proton magnetic dipole moment when in an external magnetic field. The two different orientations are commonly referred to as having their spins either parallel (spin up) or antiparallel (spin down) to the field, even though the spins do not actually point in those directions but can orient anyway along the cones shown so as to have the same z-component.

Accordingly, there are two possible energy levels for a proton in a magnetic field given as

$$E = -\mu B \cos\theta = \pm\mu_z B, \tag{18.5}$$

where we have chosen to orient the z-axis along the external magnetic field and $\pm\mu_z$ are the possible components of the magnetic dipole moment along the z-direction for the spin "up" and spin "down" nuclei. We can schematically draw this situation using an energy level diagram (Figure 18.13) in which the original energy of the nucleus (single proton) in the absence of an external field splits into two different energy levels in the presence of an external magnetic field. The difference in energy between the two energy levels is

$$\Delta E = \mu_z B - (-\mu_z B) = 2\mu_z B, \tag{18.6}$$

and depends only on the external field and the local environment of the proton through the parameter γ_n of Equation (18.4).

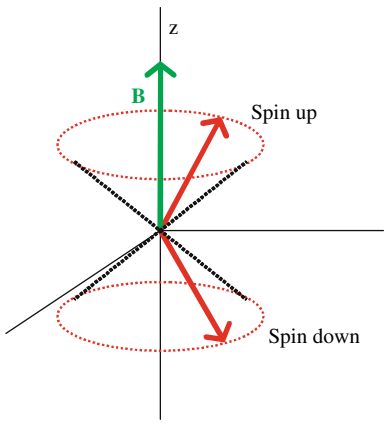


FIGURE 18.12 *The two possible orientations of a spin ½ nucleus in a magnetic field along the z-axis. The z-components (and so the cone half-angle of about 55°) are determined, although the spins may have any orientation along the cones and are classically pictured as rotating about the vertical* B *field direction somewhat like a spinning top.*

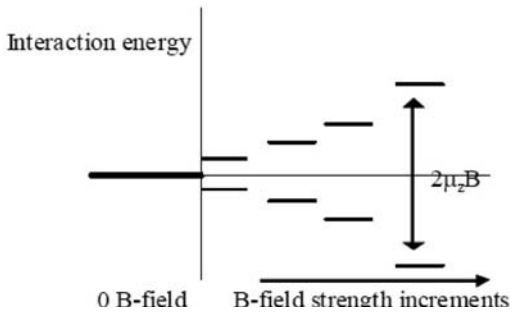


FIGURE 18.13 *Energy level diagram for a proton in a magnetic field.*

Now that we've introduced some concepts about the nucleus (N) and its interaction with a magnetic field (M) we're now ready to turn to the resonance (R) phenomenon and to the basis for NMR. Putting some numbers into Equation (18.6) for the energy difference between the two orientations of the proton spin, using a strong magnetic field of 1 T, the difference is found to be ~2×10^{-7} eV, an extremely small energy difference. It is so small even compared to $k_B T$ at room temperature, 2.5×10^{-2} eV, that according to the Boltzmann factor for the ratio of the numbers of nuclei in the two different states,

$$\frac{n_+}{n_-} = e^{-\frac{\Delta E}{k_B T}} = e^{-\frac{2\times 10^{-7}}{2.5\times 10^{-2}}} = 0.999992, \tag{18.7}$$

there will be nearly the same number of nuclei with their spins in either orientation. The difference in numbers amounts to only a few nuclei more per million in the lower energy state than in the upper energy state.

NMR involves adding electromagnetic energy to the sample in quantum packets (photons), each with an energy corresponding precisely to ΔE (a resonant condition), and measuring the net absorption of such photons. These very low energy photons have a frequency given by $E = hf$ (as we show in the next chapter; h is Planck's constant introduced above in connection with nuclear spin), corresponding to radio frequency (RF) radiation with typical frequencies of several hundred MHz. The resonant condition can then be written as

$$\Delta E = 2\mu_z B = hf. \tag{18.8}$$

For a given type of nucleus in a given local environment, the component of magnetic moment is determined. By either fixing B and varying f, or by fixing f and varying B, a resonance condition can be achieved at which there will be a net absorption of energy causing the nuclei in the lower energy state to flip their spins and jump to the higher energy state. In NMR machines of this type (continuous wave machines), the RF frequency is fixed and continuously applied and the B field magnitude is varied by small amounts while scanning through resonance conditions.

Albert Einstein showed that the same RF photons that can be absorbed and cause a nuclear spin flip to a higher energy state, can also, with equal probability, stimulate a nucleus already in the higher energy state to drop to the lower energy state and emit a second RF photon also with energy ΔE. If there were equal numbers of nuclei in each of the two states, there would be no net absorption of RF photons because there would, on average, be as many absorbed as emitted. Therefore the net absorption of energy is due only to the fact that there are slightly more nuclei in the lower energy state than in the upper energy state. Because this population difference is so small, the NMR signal is correspondingly very small. Although this finding of Einstein's makes NMR more difficult than it might be otherwise, we show later in Chapter 25 that this same prediction of Einstein is a key ingredient in the functioning of the laser.

How is this net absorption of RF radiation detected? After the small amount of RF energy has been absorbed and the RF excitation signal is turned off, the sample returns to thermal equilibrium by emitting radio frequency energy of this same frequency. This occurs as the net magnetic dipole moment created by the absorption of the excitation energy by the sample relaxes back to the lower energy states. The changing magnetic dipole moment creates a changing average magnetic field at the detector or search coil that, according to Faraday's law, induces a current in the search coil (Figure 18.14). After amplification, the induced current detected as a function of the applied magnetic field represents the NMR spectrum that is then used to understand properties of the sample.

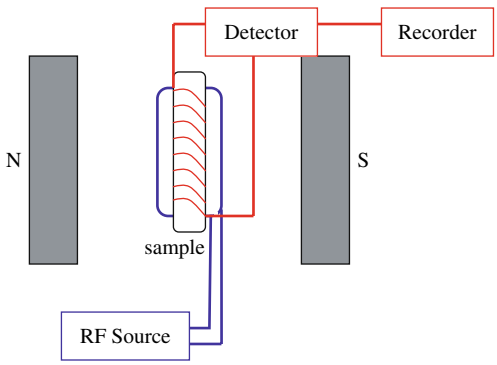


FIGURE 18.14 *Block diagram of an NMR spectrometer. Note the two separate coils, the RF input coil* (in blue) *and the detector search coil* (in red) *wound around the sample tube.*

Figure 18.15 shows a simple NMR spectrum recorded from a sample of small organic molecules. Some of the features of the spectrum include the number and position of the peaks, the area under each peak and the "line shape" and structure within a peak. In the rest of this section we indicate the kinds of information contained in each of these features and how NMR can be used to learn about the structure and function of more complicated biomolecules.

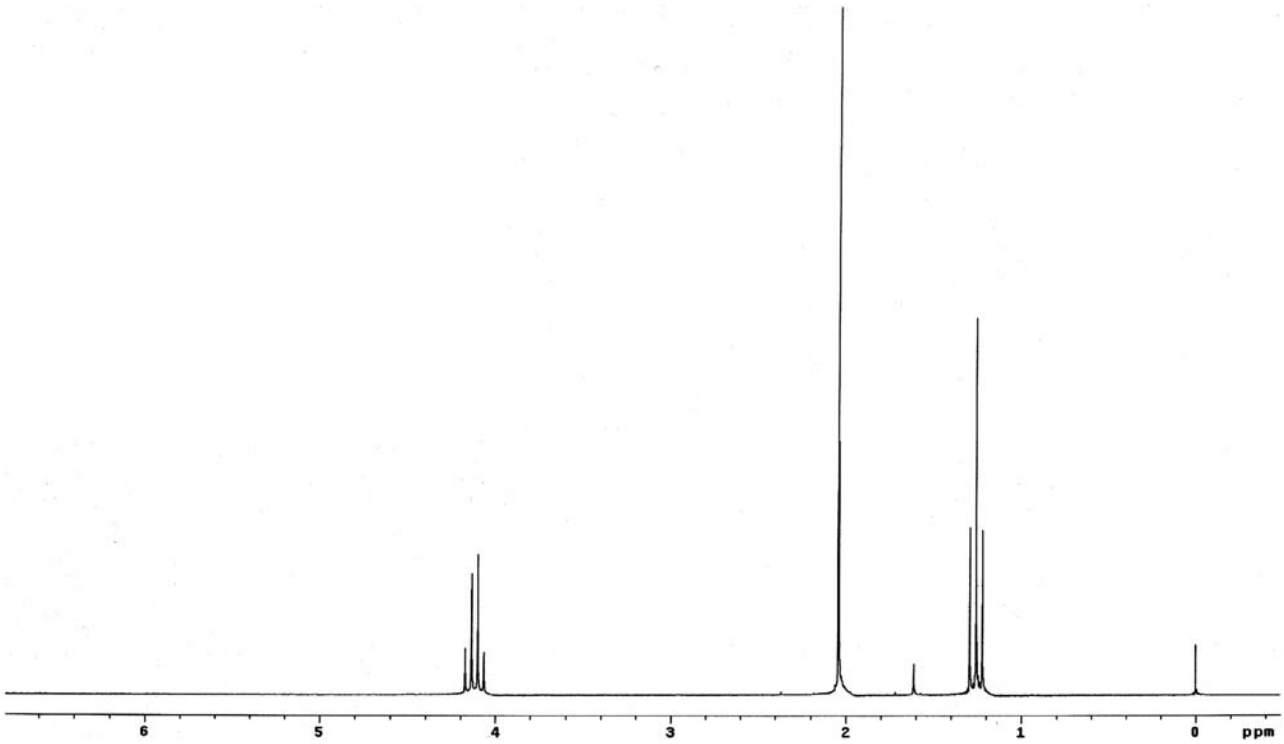


FIGURE 18.15 *The proton NMR spectrum of ethyl acetate with eight protons in three different environments. Zero ppm is defined by the peak due to TMS and the other small peak is due to a small amount of contaminating water.*

Even though the resonance signal detected in proton NMR measurements is entirely due to the energy differences between nuclear spin states ΔE, there is not just a single peak due to all the protons. Each different local environment in which a hydrogen nucleus resides experiences a slightly different magnetic field due to local screening effects of the nearby magnetic dipoles. For example, delocalized electrons in benzene rings or other ringlike organic structures moving under the influence of the large external magnetic field produce their own small local magnetic fields. From the discussion in the last section, we know that Lenz's law tells us that these induced currents produce magnetic fields generally in the opposite direction, a phenomenon known as *diamagnetism*. The extent of this shielding depends on the ring orientation relative to the proton location.

The primary feature of any NMR spectrum is the position of a peak, measured as the *chemical shift* δ, a dimensionless parameter with respect to some reference position,

$$\delta = \frac{f_{\text{sample}} - f_{\text{ref}}}{f_{\text{ext}}} \cdot 10^6, \tag{18.9}$$

where f_{ext} is the frequency of the applied RF radiation (from Equation (18.8)), f_{ref} and f_{sample} are the measured reference and sample signal frequencies, and the factor of 10^6 gives δ in parts-per-million (or ppm), the most appropriate scale because the frequency shifts are so small due to the tiny differences in nuclear energy levels. One commonly used reference material is $(CH_3)_4Si$ (tetra methyl silane or TMS) because it is chemically inert and has 12 equivalent protons giving a single strong peak.

Chemical shifts and their origin is the key science in NMR. The "art" of NMR is in interpreting the chemical shifts of the very large number of peaks found in the spectra of complex macromolecules and being able to associate particular peaks with specific hydrogen nuclei.

The relative strengths of various peaks, measured best by the areas under the peaks, are proportional to the relative numbers of equivalent nuclei (protons in our case). Thus in simple spectra of ethanol, CH_3CH_2OH, there should be three peaks

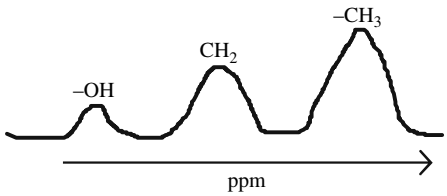


FIGURE 18.16 *Low resolution NMR spectrum of ethanol showing the 3:2:1 ratio of peak areas.*

(at low resolution) with areas in the ratio of 3:2:1 corresponding to the three local environments for H (Figure 18.16).

When spectra are taken at higher resolution, the detailed widths and structure of peaks, as well as their positions, reveal much information. The widths of peaks are fundamentally determined by the so-called "relaxation mechanisms" that return a nucleus to the lower energy state after excitation by an RF photon. These widths can give information about the local environment and can be clues as to the location of certain nuclei in a larger macromolecule. Furthermore, what appears at low resolution to be a single peak, may at higher resolution be "split" into multiple peaks, often in characteristic sets. For example, in the case of acetaldehyde, CH_3COH, whether the lone OH proton has spin "up" or "down" will produce a different local environment at the site of the other three equivalent protons. Therefore, in a population of such molecules, the low-resolution single peak for the CH_3 (methyl) group is split into two roughly equal peaks in a higher-resolution spectrum. Similarly, the methyl group protons can have a variety of possible relative spin states with either zero, one, two or all three protons with spin "up". If you count up the various possibilities, the probabilities for these corresponding spin states are in the ratio of 1:3:3:1 (see Figure 18.17). The resulting single low-resolution OH proton peak is split into the commonly observed methyl quartet of peaks with areas in the ratio of 1:3:3:1. These spin effect splittings of low-resolution lines are limited to protons attached to nearby covalently bound groups because the spin orientation-dependent magnetic fields are so weak.

There are a number of problems in both obtaining high-quality NMR data and in interpreting those data from large macromolecules. All biological materials are water based and the protons in the water hydrogens far outnumber other hydrogens in macromolecules. The NMR peak from water is so large that it typically covers up all other peaks in a large range of chemical shifts. This masking of other peaks can be avoided by either working in D_2O, also known as heavy water, where D is the isotope of hydrogen known as deuterium with one proton and one neutron in the nucleus, or most commonly nowadays by using a variant technique known as Fourier transform NMR (see below). Deuterium has an NMR signal but it is at a quite different resonant frequency and does not mask the proton peaks.

Another measurement problem is the fact that the signals are usually quite weak and difficult to distinguish from background noise (Figure 18.18). By averaging many repeated measurements, in a general method known as *signal averaging*, the ever-present but random noise is averaged out while the signal remains and can be better distinguished. With macromolecules the number of peaks can be extremely large (thousands) and in order to have peaks fairly well separated, very large magnetic fields are required. The larger the external magnetic field is, the higher the RF frequency needed for resonance conditions according to Equation (18.8), and the larger the corresponding frequency shift of characteristic peaks, and therefore the better the resolution. Currently many NMR machines use RF frequencies of over 600 MHz with corresponding external magnetic fields of over 14 T. Such high magnetic fields require a superconducting magnet because of high I^2R power losses and heating in the electromagnets.

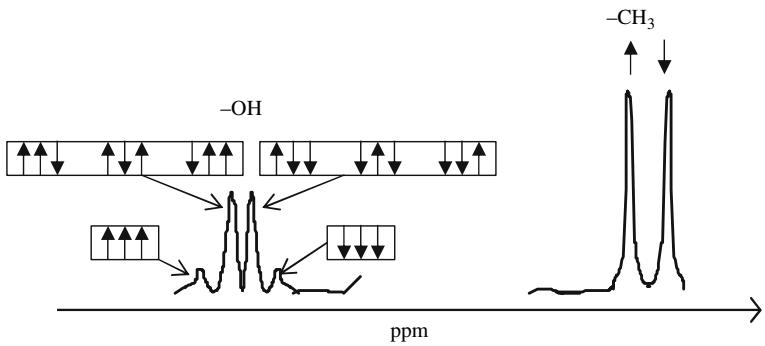


FIGURE 18.17 *Possible spin states and high-resolution NMR spectra of acetaldehyde showing the spin splittings, where the methyl peak is split into two by the –OH proton and the –OH peak is split into a methyl quartet. The various spin-splittings are labeled for each peak. The total area under the CH_3 peaks is three times that of the –OH peak total area.*

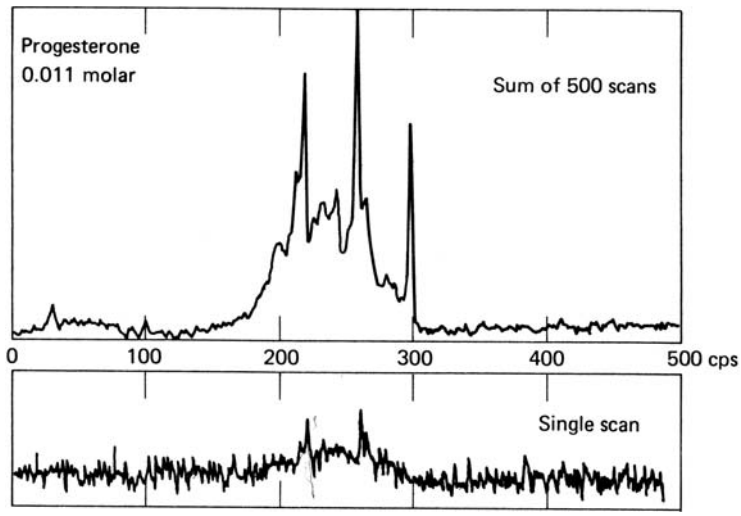


FIGURE 18.18 (bottom) *Single proton NMR scan of a sample of the female sex hormone, progesterone;* (top, *with reduced amplitude) the sum of 500 scans showing the vast improvement in the signal-to-noise ratio*

Ultimately, the objective of NMR spectroscopy of biological macromolecules is to learn details of the structure and dynamics that will aid in understanding their functioning. Since all proteins and nucleic acids are built from a relatively small number of building blocks, either amino acids or nucleotides, the results of detailed NMR studies of these smaller molecules, as well as model helical, or for protein β sheets, have been tabulated. These have been helpful in identifying some peaks that are more easily distinguished.

Most current NMR work uses *Fourier transform (FT) NMR*, a variation of the method discussed thus far. In this technique one or several RF pulses are used, rather than a continuous wave, to excite the hydrogen protons and monitor their relaxation. The basis of the technique can be understood most easily when a single pulse is used (one-dimensional FT-NMR). The external magnetic field is kept constant, and the single RF pulse can be shown to be equivalent to the sum of a large range of different frequencies of RF radiation centered about the "carrier frequency" of the pulse. Thus, in this method, in place of varying the RF frequency monotonically as a function of time, a large range of different frequencies are simultaneously applied to the sample and, by proper detection and analysis (involving the mathematical manipulation known as Fourier transforms) the entire NMR spectrum can be obtained from the single pulse (Figure 18.19).

FT-NMR methods have become extremely sophisticated, using sequences of pulses applied with varied delay times between the pulses (multidimensional NMR) to better resolve differences in the response of protons in different environments. These methods have become almost routine in determining the structure of small proteins, and can now be used on solutions of proteins even up to molecular weights of about 25,000 to resolve the positions of their atoms with a resolution of about 2 Å, comparable to that obtained using x-ray diffraction methods on crystals of proteins.

In concluding this section, we briefly discuss the related technique of *electron spin resonance (ESR)*, sometimes also known as electron paramagnetic resonance (EPR). Some materials that are not ferromagnetic exhibit a weaker form of magnetism known as *paramagnetism*. This effect occurs in atoms, molecules, or ions that have a net magnetic dipole moment usually due to an unpaired electron, such as O_2, Cu^{2+}, Mn^{2+}, and other transition or rare earth ions. Most other atoms and molecules have paired electrons so that the spin and orbital angular momentum add to zero magnetic dipole moment. Paramagnetic materials tend to align their magnetic dipole moments in an external field producing a weak magnetism that, unlike ferromagnetic materials, does not persist when the external field is removed. The extent of the alignment of the individual magnetic dipole moments depends on the strength of the B field because the interaction energy, given by $\text{PE} = -\mu B \cos\theta$, competes with the thermal energy of the atom or ion. It is the Boltzmann factor, $e^{-\Delta U/k_B T} = e^{\mu B \cos\theta / k_B T}$, that determines the extent of the alignment of the magnetic dipole moment and the overall net magnetic moment of the material.

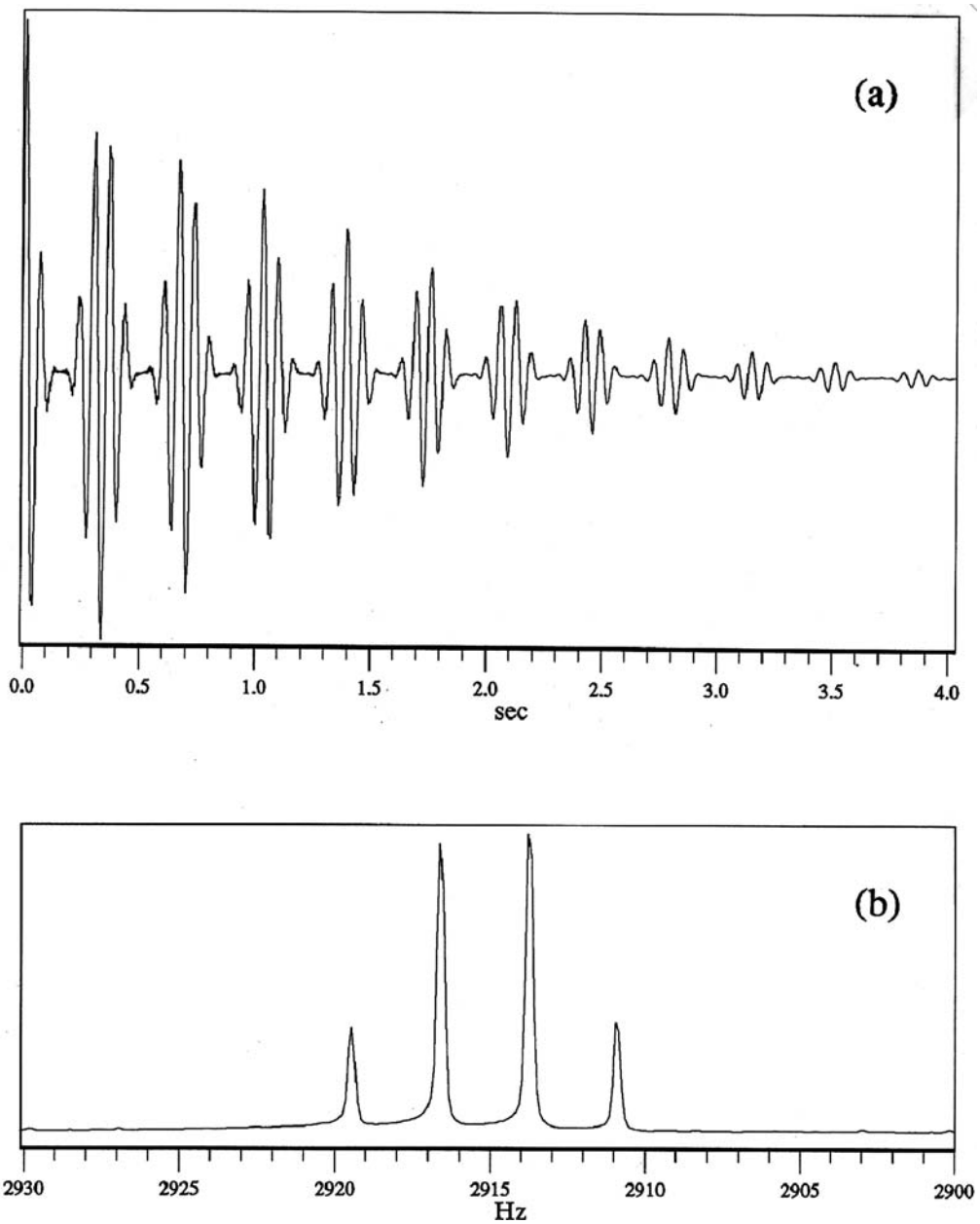


FIGURE 18.19 (a) *The direct FT NMR signal from acetaldehyde, CH_3CHO, and* (b) *a portion of its spectrum, obtained by taking the Fourier transform of* (a). *Note the methyl quartet structure for the OH proton.*

Basically very similar to NMR, ESR requires an unpaired electron in an atom, whose magnetic dipole moment generates a signal when placed in an external magnetic field. The electron magnetic dipole moment is given by an equation similar to Equation (18.4),

$$\mu_e = \gamma_e S_e, \tag{18.10}$$

where the electron's gyromagnetic ratio γ_e is about 2000 times larger than the proton's, as are the interaction energies with the magnetic field. Just as in Equation (18.4), there will be a resonance condition but now the photon frequencies must be several thousand times greater, corresponding to microwave frequencies of about 10 GHz.

In biological studies, ESR can be used to study macromolecules containing transition metal complexes with unpaired electrons such as iron and copper that occur in such interesting native macromolecules as hemoglobin (with Fe) and cytochrome oxidase (with Cu). Alternatively, because most macromolecules do not contain unpaired electrons,

a so-called "spin label," or small free radical with unpaired electron, can be attached to a macromolecule at a specific site. ESR clearly does not give the same kind of detailed structural information as NMR because the signal comes only from unpaired electrons, usually a single site on a macromolecule. The method has, however, been widely used to study conformational changes at important sites on a macromolecule, often binding sites for small ligands, and to probe motions of macromolecules, especially those bound to membranes.

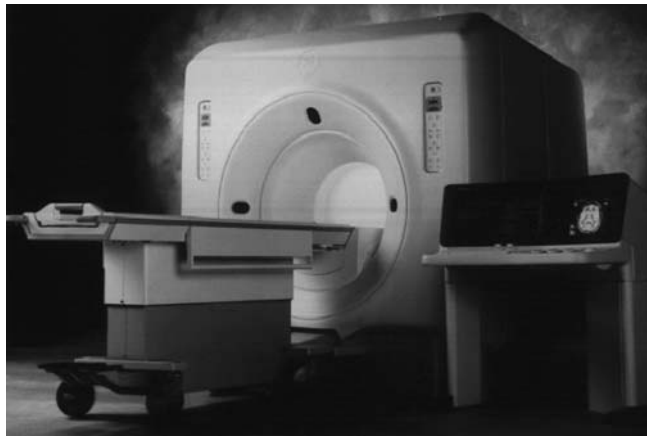


FIGURE 18.20 *An MRI machine used for whole-body medical imaging.*

3. MAGNETIC RESONANCE IMAGING

In the previous section we discussed the physical basis for NMR and its application to the study of biomolecules in solution. The samples in those types of studies are small (with volumes of ~1 cm^3) and the magnetic field must be extremely uniform over the small sample volume in order to have a consistent resonance condition throughout the sample. In this section we show how to apply the same NMR principles to allow imaging of large regions of the human body in a technique known as magnetic resonance imaging (MRI).

There are two main new considerations that need to be discussed. First, how is the spatial information encoded in the data in order to obtain images of cross-sections through the body? Second, what is responsible for the contrast seen in these images? Clearly the magnet configuration must be very different to allow a person to be in the strong magnetic field needed to align nuclear spins throughout a large region of the body. Very large gaps between the magnetic poles (~1 m) are needed for a person to lie in the magnetic field (Figure 18.20). The magnetic field must be very large to give high resolution and must also be controlled extremely well in order to be able to do the spatial imaging, as we show. The magnets used in MRI are exclusively superconducting magnets, electromagnets that use high-efficiency superconducting current coils. Special wire materials are used that must be kept at very low temperatures (typically liquid helium temperatures of $-269°C$) in order to be superconducting, essentially eliminating the I^2R heating and allowing very high currents and correspondingly high magnetic fields to be maintained.

Let's first consider how the spatial information can be encoded. If the magnetic field were completely uniform over the entire area of the body to be imaged, there would be no way to spatially distinguish the origin of the signal. Instead, magnetic field gradients (varying linearly in a particular direction at a typical rate of 10^{-2} T/m or 1 G/cm) are used so that the resonance condition, Equation (18.8), will vary along that direction, say the z-direction, according to the local magnetic field value. If a selective RF pulse with a carrier frequency matching those resonant frequencies of protons within a particular slice or plane perpendicular to the z-direction is used, then only those protons will be detected (Figure 18.21). In essence, the z-position of a free

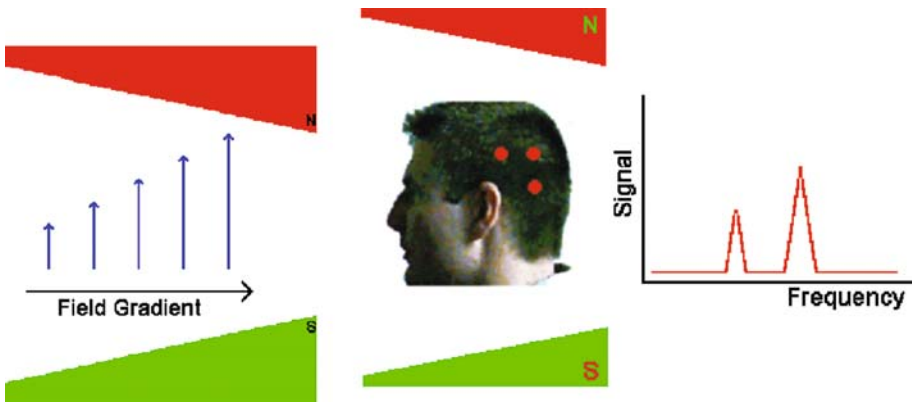


FIGURE 18.21 (left) *Field gradient established by gradient coil;* (right) *signal detected if there were only three equivalent "proton centers" in the patient's head* (shown in red)*; note that only two peaks are seen because of the variation in resonance position along the field gradient, one with twice the integrated intensity of the other.*

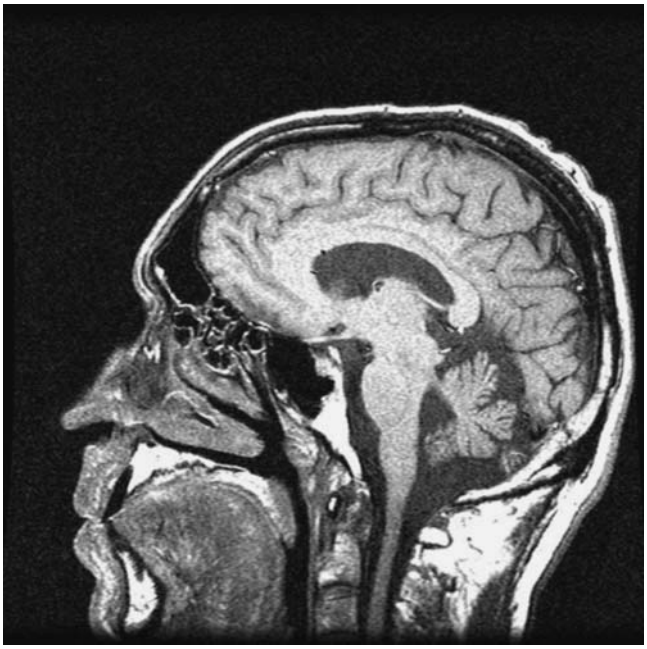


FIGURE 18.22 *An MRI image of a cross-section through the human head.*

proton is encoded in a resonant frequency that is proportional to z. For a constant magnetic field gradient, the thickness of the slice depends on the equivalent range of frequencies in the pulse, typically a few Hz. The longer the RF pulse is, the narrower the slice detected because a longer pulse more closely resembles a pure sine curve which would match the resonance condition at a very narrow slice.

Once the slice selection has been achieved and the protons in a particular slice transverse to the field gradient z-direction are aligned, subsequent field gradients in the x- and y-direction are applied consecutively, each for varying times. This procedure, after being applied n different times for the x-gradient and n different times for the y-gradient and performing a Fourier analysis, yields an image with n^2 datapoints, or pixels, in the transverse plane. The limit on n, and hence on the ultimate spatial resolution, depends on factors such as the signal detection sensitivity as well as the linearity and stability of the magnetic field gradients, patient movement during the typical several minutes needed for a scan, and other artifacts. The overall RF pulse sequence and field gradient sequence is repeated many times using signal averaging to reduce the noise and give a larger signal-to-noise ratio.

A remaining major question is what is responsible for the high degree of image contrast seen in MRI images (Figure 18.22). We have seen that protons in different local magnetic fields will have different resonances yielding different chemical shifts in an NMR frequency spectrum. Because the predominant form of protons in the body is in the water content (roughly 70% of the body is water) one way to distinguish different structural features is simply in the local water content (Table 18.2). For soft tissue these differences are relatively small and do not give sufficient contrast alone. Another signature of NMR spectra is the relaxation time corresponding to the return of spin states to an equilibrium population mentioned in the last section. There are two characteristic times, the spin–lattice relaxation time T_1 and spin–spin relaxation time T_2. The spin–lattice relaxation involves the transfer of energy to neighboring molecules and the spin–spin interaction involves the emission of an RF photon by a proton as its spin flips to a lower energy state and the subsequent absorption of the photon by a neighboring nuclear spin causing its excitation. These characteristic times vary substantially depending on the type of tissue and are the primary source of image contrast. Thus in the false color MRI images commonly seen, the different colors usually refer to differences in relaxation times.

Table 18.2 *Water Content of Normal Human Tissue*

Tissue	*% water*
Brain (white matter)	84
Kidney	81
Myocardium	80
Skeletal muscle	79
Brain (gray matter)	72
Liver	71
Nerve	56
Bone (cortex)	12
Teeth	10

MRI methods have steadily advanced in sophistication, both in magnet design (open high field strength magnets avoid the claustrophobia problems some

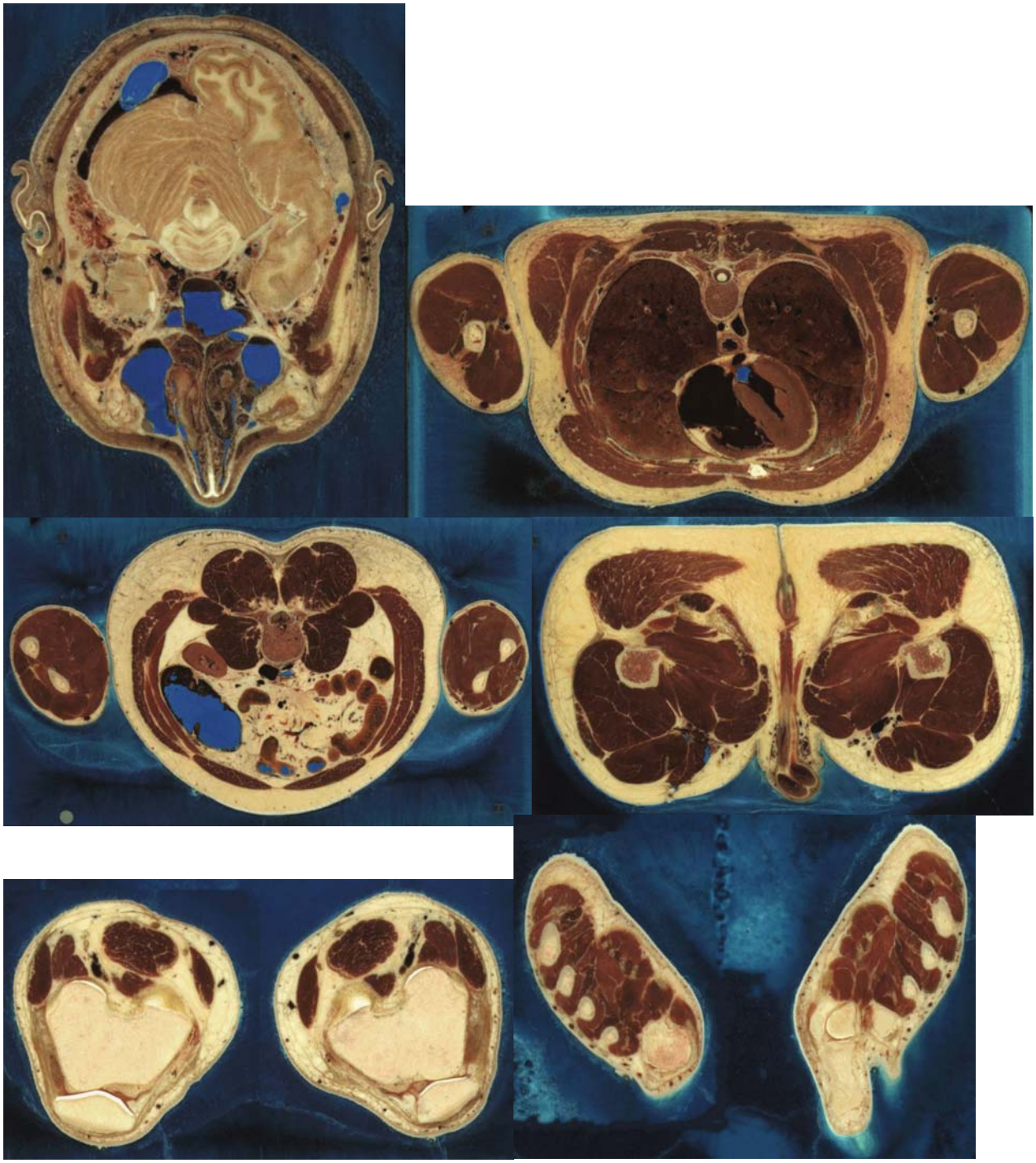


FIGURE 18.23 *Separate MRI scans of* (from left) *head, thorax, abdomen, upper thigh, knee, and feet.*

patients have while maintaining the rigorous requirements on uniformity and stability of fields) and in pulse sequences and data analysis methods so that multislice high-resolution images at any orientation can be obtained rapidly (Figure 18.23).

Interesting and significant variations on MRI include the use of other nuclei to monitor, in real-time, changes in specific biomolecules within the body. For example, the real-time MR imaging of ^{31}P nuclear spins present in ATP can monitor the splitting of ATP because the phosphate environments change leading to an NMR signature change. Such studies can be used diagnostically to check for proper metabolic functioning within the human body.

4. MAXWELL'S EQUATIONS; ELECTROMAGNETIC RADIATION

In our discussions of electricity and magnetism, we have seen that electric charges produce electric fields and that electric currents, whether they be in a circuit or in circulating atomic charges, produce magnetic fields. Faraday's law also revealed that changing magnetic fields (or more generally magnetic fluxes) can produce electric fields as well. It was first proposed by Maxwell, on theoretical grounds of symmetry, that changing electric fields can also produce magnetic fields by an induction process similar to that of Faraday's law. The mathematical relations between electric and magnetic fields and their sources, electric charges and currents, as well as their variations in space and time are known as Maxwell's equations.

Consisting of a set of four fundamental equations, Maxwell's equations represent one of the most successful theories of all science, more successful than even Newton's laws or the law of gravitation. Although Maxwell's equations were first published in 1873 and Einstein's special theory of relativity was not published until 1905, Maxwell's equations proved to be relativistically correct. Today, Maxwell's equations still stand without change as the fundamental explanation of all electromagnetic phenomena, requiring only a quantum mechanical synthesis in order to explain those same phenomena on an atomic distance scale.

We summarize Maxwell's equations in words. The mathematical statement of the equations, involving calculus, does not provide further illumination at the level of our presentation. However, it should be mentioned that particular real-life problems, such as magnet design for MRI machines, are approached by the direct mathematical solution of Maxwell's equations subject to particular "boundary conditions" imposed by the geometric spatial boundaries and time constraints of the problem. Usually computers are used to generate numerical solutions, although in certain idealized situations analytic solutions can be obtained.

In words, two of Maxwell's equations involve relating the fields to their sources of charge in relations known as Gauss's laws. Gauss's law for electric fields connects the electric field to electric charges and holds not only under electrostatics, where it yields Coulomb's law, but also quite generally (see Section 7 in Chapter 14). A second similar law for the magnetic field includes the fact that there is no magnetic "charge" (no magnetic monopoles) and predicts that magnetic field lines will form closed curves. Maxwell's other two equations connect a changing magnetic flux with an induced electric field (Faraday's law), and a current or a changing electric flux (see below) with an induced magnetic field. It should be made clear that Maxwell's equations have their roots in many years of experimentation as well as in less complete theories by others (including Ampere's law studied in Section 5 of the previous chapter). Maxwell's major contributions were his completion of the content of the last-mentioned law of induced magnetic fields, his synthesis of these results in a minimal set of four equations, and the predictions he then made based on these equations. Perhaps the most important of these predictions is the production of electromagnetic (EM) radiation.

First, recall from Chapter 14 that electric field lines start on positive charges and terminate on negative ones. Electric fields due to charges that are at rest are called "static" fields. Static fields never loop around back onto themselves to form closed curves. In Chapter 17 we found that moving charges also make magnetic fields. Magnetic field lines never start or stop; they always loop around in closed curves. Thus, if the source charges are at rest or moving with constant speed, they make starting and stopping E and looping B. In the first section of this chapter we learned that it is possible to make E-lines that loop around by having B change in time. That is, E that loops around and is perpendicular to B can arise when the source charges accelerate.

Maxwell's great contribution to electromagnetism was to conjecture that if time-changing B could make E, maybe time-changing E could make B. Here's his idea. Consider a capacitor with vacuum between its plates. Connect the capacitor to a battery and begin to charge it. Around the wire coming into the capacitor there is looping B as the current flows. Around the wire leaving the capacitor there is also looping B. But, no

charge moves between the plates of the capacitor, so between the plates there is no current. Before Maxwell, it was thought that there was no B between the capacitor plates. The magnetic field around the wires, it was thought, abruptly stopped at the capacitor plates. But there's E between the plates and as the capacitor charges that E changes in time. So perhaps, Maxwell argued, B doesn't just abruptly stop at the plates. Perhaps the same kind of looping B around the wires is found between the plates looping around changing E-lines. It turns out that he was right.

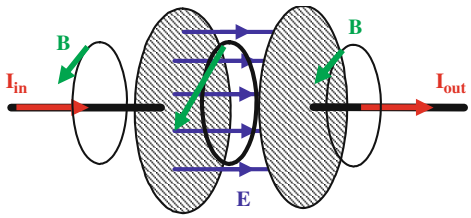


FIGURE 18.24 *A parallel-plate capacitor being charged, showing the B-field around the wires and within the capacitor due to the changing E-field.*

Maxwell's idea is depicted in Figure 18.24. The plates of the capacitor are circular disks. The current in equals the current out as the capacitor is charging. The E-field points from the positive plate to the negative, from left to right in the figure, and is increasing in this scenario. B curls around the increasing E as one's right fingers curl around one's right-hand thumb when the thumb points in the direction of changing E (as around the current in the wires).

As a result of Maxwell's insight we have this mantra: "time changing B makes curling E, time changing E makes curling B." That's how "EM radiation" is made. A good analogy to study first is what happens when you shake one end of a long string. The instant you accelerate one end of the string, material near your hand begins to get an upward velocity (see Figure 18.25). But, because of inertia (the mass of the string), that velocity can't get to the other end of the string instantaneously. What happens is that the upward velocity makes some slope in the string near your hand. Then that new slope makes some upward velocity out ahead of it. Then that new velocity makes some newer slope. Then the newer slope makes newer velocity. And so on. Figure 18.25 shows more precisely what happens.

The hand on the left accelerates the end of the string (originally horizontal and at rest) for a brief time and then continues upward with constant speed. During the acceleration a kink forms in the string. The kink travels away from the hand at a fixed speed: the speed of a wave in the string. In front of the kink the string is not moving. Behind the kink the string moves upward with constant speed. Down through the kink the upward velocity of string material changes from a maximum in the back to not much in the front. It takes time for the information that the hand has accelerated the left end of the string to travel to the right end. The fact that the information doesn't travel instantaneously is because of Newton's second law: a finite force produces a finite acceleration when there is mass to be moved. We saw in Chapter 10 that the speed of the wave is $v = \sqrt{T/\mu}$, where T is tension in the string and μ is mass per unit length.

Essentially the same thing happens with electric and magnetic fields. If you accelerate a charge upward, the E-line attached to the charge begins to get some upward velocity. This time changing E makes a curling B (that is also changing in time). The time-changing B then makes curling E (that is also changing in time). And on and on. In fact the picture is similar to Figure 18.25, except instead of a string there is an E-line. The information can't propagate infinitely fast because E and B obey a kind of Newton's second law. In this analogy between strings and electromagnetism, B corresponds to slope in the string and E corresponds to upward velocity in the string. The role of tension is played by $1/k_M$ (where $k_M = \mu_0/4\pi$ is the magnetic force constant equal to 1×10^{-7} N-s^2/C^2) and the role of mass density is played by $1/k_E$ (where $k_E = 1/4\pi\varepsilon_0$ is the electric force constant equal to 9×10^9 N-m^2/C^2). These two initially quite independent concepts, that of electric and magnetic fields, are united through Maxwell's equations to predict the speed of electromagnetic radiation. Thus, the wave speed for a traveling EM disturbance is

$$c = \sqrt{\frac{k_E}{k_M}} = \sqrt{\frac{1}{\mu_0\varepsilon_0}}, \qquad (18.11)$$

which comes out to be 3×10^8 m/s, the speed of light in vacuum. Amazingly, Maxwell's equations led to the proposal that all electromagnetic radiation travels at the speed of light, a statement that was subsequently shown to be true.

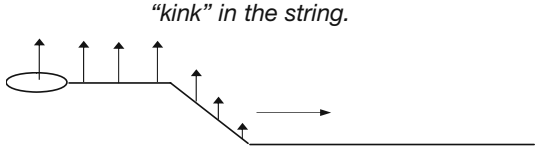


FIGURE 18.25 *A string being driven on the left to produce a traveling "kink" in the string.*

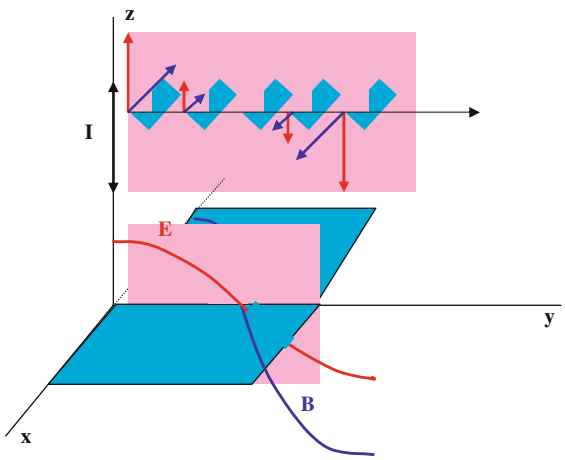


FIGURE 18.26 *An oscillating current along the vertical z-axis produces a periodic electromagnetic wave. In this figure we focus on that part of the wave that travels out along the y-axis and show the time-varying B (along the x-axis; blue arrows) and time-varying E field (along the z-axis; red arrows), perpendicular to each other and to their direction of travel (to the right in the figure). The lower graph plots the coupled E and B fields at one instant of time.*

You shouldn't take the relation between EM radiation and strings literally. They share lots of common features, but they are also very different. If Figure 18.25 represents an *E*-line, the *B* that travels along with the *E* kink is not the slope of the line, but rather is a separate field perpendicular to *E*, in and out of the page. (These waves are called "electromagnetic" because *B* always tags along with *E*; or is it vice versa? See Figure 18.26.) Waves on a string are disturbances in the string from equilibrium. Similarly, EM waves are disturbances in the background electric and magnetic fields throughout space. Waves on a string require a tangible body, the string. Electromagnetic waves are perfectly happy to travel through a vacuum. Electromagnetic waves don't need anything to ripple through.

For both strings and electromagnetism, the disturbance is initiated by acceleration, acceleration of a bit of the string or acceleration of charge. If the acceleration is periodic, the disturbance is periodic as well. In that event, it is possible to generate periodic traveling EM waves and, if there are boundaries, standing EM waves as well. It is customary to speak of the "electromagnetic spectrum" (see Section 4 in the next chapter) that is related to the wave relation $c = \lambda f$. Periodic waves have a wavelength and a frequency the product of which is the wave speed, c in the case of EM radiation. EM waves of different f, or equivalently, different λ, are often given different names, although they are the same thing. These names are historical and usually have to do with the mechanism by which the radiation is produced and detected. The phenomenon we call "light" is periodic EM radiation with λ between about 400 nm and 700 nm, where the size of an atom is about 0.1 nm. (The corresponding frequencies are 7.5×10^{14} Hz and 4.3×10^{14} Hz, respectively.) Light is EM radiation whose wavelength is a few thousand atoms long. We go into the phenomenon of color in more depth later, but for now it is sufficient to say that a single frequency of light corresponds to a "pure color" (just as a single frequency of sound corresponds to a pure tone). Light that consists of a broad range of frequencies mixed together is called "white light."

When EM radiation "impinges" on matter, the electric field in the wave causes—or "induces"—charges in the matter to accelerate. When charges accelerate they produce "induced" EM radiation. As a result, *the total EM field one detects at any point is a superposition of the incident fields due to the original sources and these induced fields*. This idea is the entire basis for all of optics, a topic that we treat in a great deal more depth later. But first, in the next chapter we continue our study of electromagnetic radiation, learning about the different types of EM radiation and their description as waves.

CHAPTER SUMMARY

Faraday's law relates a changing magnetic flux,

$$\Phi_B = \overline{B}_\perp A = \overline{B} A \cos\theta, \qquad (18.1)$$

to an induced emf, ε, according to

$$\varepsilon = -\frac{\Delta\Phi_B}{\Delta t}. \qquad (18.2)$$

The flux can change in several ways: B itself may be time-dependent and/or the area A of the circuit may change, and/or the orientation of the circuit may change with respect to the direction of B. In any case, when the magnetic flux changes with time, there will be an induced emf in the circuit. Its polarity is governed by Lenz's law: the induced emf is always of a polarity such as to oppose the change of magnetic flux that created it.

Applications of Faraday's law are numerous and diverse. Time-varying electric currents in nerve cells of the brain can be detected through changes in magnetic flux through SQUID (superconducting quantum interference device) detectors in order to map brain activity in MEG (magnetoencephalography) recordings. NMR (nuclear magnetic resonance) signals are also detected using Faraday's law. NMR involves causing nuclear spins to flip to excited states in a

strong magnetic field by added resonant energy and then watching the relaxation back to the ground state. The resonance condition

$$\Delta E = 2\mu_z B = hf, \qquad (18.8)$$

must be satisfied, where the energy difference ΔE is between the two spin states in the applied B field and the frequency f is the photon frequency needed to cause the transition. NMR looks at the frequency of emitted photons from the relaxation process or the decay times of these relaxations. Different local environments of the nuclei lead to slightly different (few ppm) frequencies and so different nuclei in different portions of a molecule can be distinguished and studied. By applying B fields that are spatially varied, NMR can be used to produce images in a technique known as MRI (magnetic resonance imaging). MRI technology makes use of the different relaxation times for different tissue to produce ~1 mm resolution images of cross-sections, with any desired orientation, of the human body and has been a powerful medical diagnostic tool.

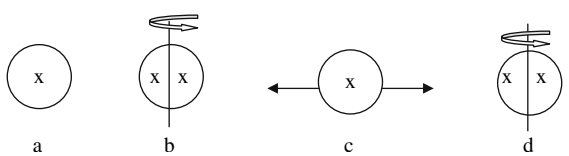
Section 4 of this chapter summarizes our fundamental knowledge of electricity and magnetism in a discussion of Maxwell's four equations. Two of these are Gauss's laws, for the electric field (see Section 7 of Chapter 14) and for the magnetic field (where the magnetic flux over any closed surface must equal zero because there are no magnetic charges). These are both related to field mappings, where electric field lines must start or stop on electric charges and magnetic field lines form only closed contours. Faraday's law and a modified form of Ampere's law constitute the other two Maxwell equations. These relate either a changing magnetic field with an induced electric field (Faraday's law), or a changing electric field with an induced magnetic field. Electromagnetic radiation is a direct consequence of Maxwell's equations and a number of general properties of EM radiation follow from the equations, including the fact that it travels at the speed of light, given by

$$c = \sqrt{\frac{1}{\mu_0 \varepsilon_0}}, \qquad (18.11)$$

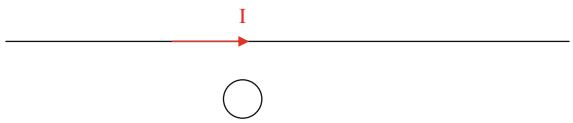
and that EM radiation consists of oscillating E and B fields that are mutually perpendicular and lie in a plane transverse to the direction of propagation.

QUESTIONS

1. Define magnetic flux, clearly distinguishing the three different factors that can affect its value. Give an example for each of the three different ways in which the flux can change.
2. Discuss the statement that "A changing magnetic field produces an electric field" in light of Faraday's law.
3. If there is a changing magnetic flux through a coil, an emf is produced that leads to an induced electric field within the coil. How can this be, since we we learned earlier when we studied statics that there cannot be any electric field within a conductor?
4. Find the direction of the induced current, if any, in the circular coil shown in each of the following situations, pictured below. In (a) the coil is in a region where the B field is increasing into the paper, indicated by the tail of the vector x; in (b) the B field is constant in magnitude and initially oriented into the paper but is rotating about a vertical axis; in (c) the B field is constant and the coil is stretched by pulling it along the horizontal; in (d) the B field is constant in magnitude but rotates as in part (b) but so does the coil rotate with the field.

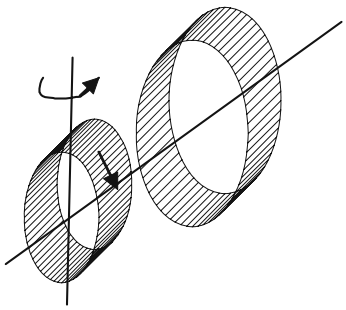


5. Suppose that there is a small coil lying near a long straight current-carrying wire as shown in the sketch below. Find the direction of the induced current, if any, in the coil under the following circumstances. If there is an ambiguity, indicate why and the possible answers. In (a) the coil remains stationary, but the current in the wire increases; (b) the current is constant, but the coil moves downward; (c) the current is constant and the loop remains stationary; (d) the current decreases and the coil moves downward; (e) the current decreases and the coil moves upward.

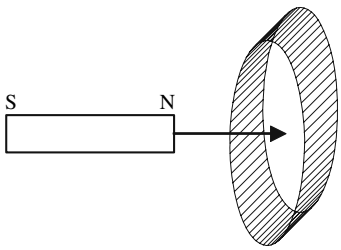


6. Two parallel coils lie along the same axis as shown with both wrapped in the same sense. Find the direction of the induced current in the larger coil if
   (a) The current in the smaller coil (direction shown) is increasing.
   (b) The smaller coil, with constant current, is moving away from the larger coil.
   (c) The larger coil is moving toward the smaller coil, having a constant current.

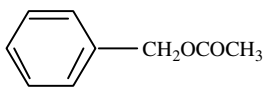
(d) The smaller coil, with constant current, is rotating counterclockwise around a vertical axis as shown.



7. Suppose the two coils of the previous problem each have N turns. Why does the induced emf increase by a factor of N^2 over that when each coil is a single turn?
8. A bar magnet is thrust toward a coil as shown. In what direction is the induced current as the magnet approaches the coil at constant speed? As it recedes from the magnet after passing through?



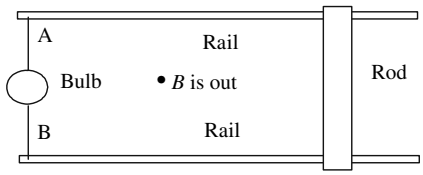
9. Show that the gyromagnetic ratio in Equation (18.4) has units of $s^{-1}\,T^{-1}$.
10. In order to have an NMR signal, nuclei must have an odd number of protons or neutrons or both. Why is this so?
11. Why is it the case that NMR machines with larger magnetic fields must operate at higher radio frequencies?
12. Discuss the expected relative intensities for the three lines observed in the low-resolution proton NMR spectrum of benzyl acetate, shown below, corresponding to C_5H_5, CH_2, and CH_3.



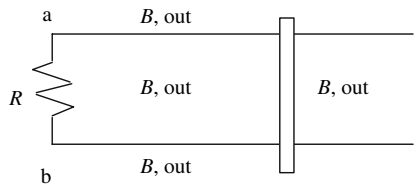
13. Why are NMR chemical shifts commonly measured in parts per million?
14. According to Equation (18.7), do you expect the NMR signal to increase or decrease with increasing temperature?
15. Explain how spatial imaging is obtained in MRI.
16. Describe some similarities and differences between NMR and ESR.

MULTIPLE CHOICE QUESTIONS

1. Which one of the following does not involve electromagnetic induction? (a) A wire is dragged through a constant magnetic field. (b) A compass points toward the Earth's north geographic pole. (c) A coil of wire is rotated in the field of a permanent magnet. (d) A permanent magnet is dropped vertically into an aluminum tube.
2. The figure shows a conducting rod riding along two horizontal, conducting rails. The rails, in turn, are connected by a light bulb. A uniform magnetic field pointing out of the page exists in the region between the rails. At one instant, the rod is moving toward the light bulb. Current flows through the light bulb and the rod experiences a magnetic force. Which one of the following is true? (a) Current goes through the bulb from A to B and the force on the rod is to the right. (b) Current goes through the bulb from A to B and the force on the rod is to the left. (c) Current goes through the bulb from B to A and the force on the rod is to the right. (d) Current goes through the bulb from B to A and the force on the rod is to the left.



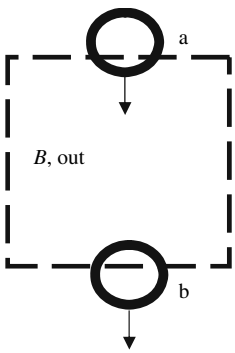
Questions 3–4 refer to: The figure shows a conducting rod sliding over two bare wires. The wires are in a horizontal plane (ignore gravity) and are connected through the resistor R. A uniform magnetic field points out of the page everywhere. Friction between the rod and the wires is negligible. At the instant shown, the rod is moving from right to left, toward R, with a speed v. (A hand got the rod started before the instant shown, but is no longer in contact with it.)



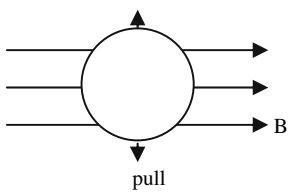
3. Which one of the following is true at the instant shown? (a) Current is flowing through R from a to b. (b) Current is flowing through R from b to a. (c) There is no current flowing through R. (d) The rod experiences no net force.
4. Which of the following best describes the magnitude of the current in R subsequent to the instant shown? (a) The magnitude increases because the rod speeds up. (b) The magnitude decreases because resistance always decreases current over time. (c) The magnitude decreases because the rod slows down. (d) The magnitude remains constant because the rod travels with constant speed.
5. A (metal) car is traveling due north in the United States where the Earth's magnetic field points both northward and vertically downward. As seen by the driver, the induced emf causes (a) the right-hand side of the car to be positively charged and the left-hand

side to be negatively charged, (b) the right-hand side of the car to be negatively charged and the left-hand side to be positively charged, (c) the top of the car to be positively charged and the bottom to be negatively charged, (d) the top of the car to be negatively charged and the bottom to be positively charged.

6. A bar magnet is dropped into a vertical aluminum tube. Suppose the north pole of the magnet is pointing down. You look down the tube from the top (south pole end of the magnet pointing up at you). Which one of the following is true at any instant, as viewed by you, while the magnet is inside the tube? (a) Current circulates around the tube in a counterclockwise fashion below the magnet, and the magnet experiences a magnetic force pointing up. (b) Current circulates around the tube in a clockwise fashion below the magnet, and the magnet experiences a magnetic force pointing up. (c) Current circulates around the tube in a counterclockwise fashion below the magnet, and the magnet experiences a magnetic force pointing down. (d) Current circulates around the tube in a clockwise fashion below the magnet, and the magnet experiences a magnetic force pointing down.
7. The figure shows a copper ring dropped vertically into a region of magnetic field. In which direction is the force on the ring due to electromagnetic induction? (a) Up at a, up at b. (b) Up at a, down at b. (c) Down at a, up at b. (d) Down at a, down at b.



8. A uniform horizontal magnetic field B points north. The average induced emf measured when a circular wire of radius R oriented in a vertical plane along the N–S direction as shown is stretched vertically until it collapses to a vertical straight wire in a time t is (a) $\pi R^2 B/t$, (b) B/t, (c) $2\pi R^2 B/t$, (d) 0.



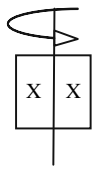
9. A coil of wire consisting of 5 turns is placed in a uniform external magnetic field that is changing in strength at a constant rate. As a result, an electrical potential difference of 10 V is induced from one end of the coil to the other. The coil is replaced by one with the same orientation as the first but with 10 turns. The induced potential difference in the second coil is (a) 5 V, (b) 10 V, (c) 15 V, (d) 20 V.
10. Which of the following nuclei does not give an NMR signal? (a) ^{13}C, (b) ^{19}F, (c) ^{40}Ca, (d) ^{31}P.
11. If a ^{13}C nucleus, when put in a uniform magnetic field, has two energy levels E_1 and E_2 resulting from its nuclear spin pointing either up or down, the resonant frequency will be (a) E_1/h, (b) E_2/h, (c) E_1E_2/h, (d) $(E_2 - E_1)/h$.
12. If a low-resolution NMR single peak splits, at higher resolution, into four peaks with areas in the ratio of 1:3:3:1, we can usually conclude that (a) the species representing the peak is the methyl group, (b) there is a nearby OH group, (c) there is a nearby methyl group, (d) the species representing the peak is OH.
13. The stronger the magnetic field is in an NMR experiment, (a) the stronger the NMR signal, (b) the greater the magnetic moment of the nucleus. (c) the greater the chemical shift, (d) the greater the resonant frequency.
14. By using a linear magnetic field gradient along a given direction in MRI, the resonance frequency is (a) the same throughout a slice along that direction, (b) varies linearly with distance away from the gradient line in a plane perpendicular to the gradient direction, (c) varies linearly along the gradient direction, (d) varies quadratically with distance along the gradient line because magnetic energy varies as B^2.
15. An electromagnetic wave is said to be transverse because (a) E is perpendicular to B, (b) E points in the direction of propagation, (c) E equals Bc, (d) both E and B are perpendicular to the direction of propagation.
16. The speed of an electromagnetic wave in vacuum is determined by the electric and magnetic force constants through the relation

$$\text{(a)}\sqrt{k_E k_M},\ \text{(b)}\sqrt{\frac{1}{k_E k_M}},\ \text{(c)}\sqrt{\frac{k_E}{k_M}},\ \text{(d)}\sqrt{\frac{k_M}{k_E}}.$$

PROBLEMS

1. Suppose a small 1 cm radius coil with 100 turns is used to try to measure the time-varying magnetic field produced by neuronal electric currents in the brain. If the normal component of this field varies by 0.5×10^{-12} T over 0.1 s, find the average induced emf in the coil during this time.
2. A long straight wire lying along the x-axis carries a constant 5 A current along the positive x direction. A small 5 mm diameter loop lies in the x–y plane centered at $y = 2$ m. If the loop is suddenly stretched so that its area shrinks to zero in 0.2 s, find the average induced emf in the coil assuming the B field is constant over its area. Draw a sketch showing the direction of the induced current.

3. A square coil of 100 turns with 2 cm sides lies in a uniform 2 T magnetic field. If the coil is made to rotate at a frequency of 60 Hz about an axis through its center and parallel to one side, as shown, write an expression for the induced emf as a function of time and find the maximum emf generated in the coil.



4. A 5 cm radius circular coil lies in a region with a uniform magnetic field perpendicular to its surface. If the magnetic field varies with time t according to $B(t) = 0.1 + 0.05\,t$ for $0 \leq t \leq 100$ s, with B measured in tesla and t measured in seconds, find the induced emf during the 100 s interval.

5. Give an order of magnitude estimate, based on Faraday's law, of the maximum induced emf detected by a search coil with a 0.2 m diameter 1 cm away from a long neuron which carries an average current of 10 pA switched on in 1 ms.

6. A helicopter has blades of length 2.5 m, extending out from a central hub and rotating at 4.00 rev/s. If the vertical component of the Earth's magnetic field is 50.0 μT, what is the emf induced between the blade tip and the center hub?

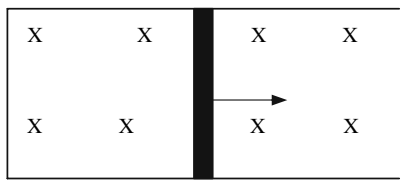
7. A Boeing 737 has a wingspan of approximately 40 m (120 ft). Suppose that a 737 is flying horizontally where the downward component of the Earth's magnetic field is 50 μT. At what speed would the 737 have to fly in order for there to exist a 1.5 V potential difference across its wingtips? Is this a reasonable speed for a 737?

8. To monitor the breathing of a hospital patient, a thin belt (a 200-turn coil) is placed around the patient's chest. Suppose that the belt has a radius of 20 cm and when the patient inhales, the belt expands to a radius of 20.5 cm. The magnitude of the Earth's magnetic field is 50.0 μT and makes an angle of 50° with the normal to the coil. Assuming that a patient takes 1.80 s to inhale, find the average induced emf in the coil during this time. What is the induced emf when the person exhales over the same time interval. What would a voltage versus time trace look like on a monitor screen as the person inhales and exhales?

9. A lightning bolt strikes the ground 200 m from a 100-turn coil. Suppose that the radius of the coil is 0.8 m and that the current carried by the lightning bolt is 6.0 MA and falls to zero in 10.5 ms. What is the induced emf in the coil if it is oriented with its normal along the magnetic field direction? If the wire has a cross-sectional area of 7.85×10^{-7} m^2 (diameter of wire is 1 mm) and it is made out of copper ($\rho = 1.7 \times 10^{-8}\ \Omega$m), what is the magnitude and direction of the induced current in the wire?

10. A 4 mm diameter circular coil of 25 turns and total resistance 0.001 Ω lies in the x–y plane at a distance of 3 m from a long straight current-carrying wire along the x-axis. If the current in the wire is increasing at a rate of 0.2 A/s, find the induced current in the coil and give its direction in a sketch. (Assume the magnetic flux is uniform over the area of the coil.)

11. A 20 cm long conducting rod completes the circuit shown through which there is a constant uniform 1.2 T magnetic field. If the rod, having essentially all of the electrical resistance of the circuit, $R = 100\ \Omega$, is free to slide along the track without friction and is pulled at a speed of 2 m/s to the right, find the average electric field in the rod.



12. In the previous problem, what force is needed to pull the rod at the constant speed of 2 m/s to the right?

13. If you wanted to produce a sinusoidal 20 V peak-to-peak signal in a 10 cm diameter pick-up coil with 100 turns sitting in a 1.2 T uniform magnetic field, at what angular velocity would you have to spin the coil?

14. Induced emf measurements can be used to measure the speed of a conducting fluid such as sea water. If a 20 cm inner diameter nonconducting pipe has sea water flowing through it at a flow rate of 10 gal/min and a uniform magnetic field of 0.05 T is applied transversely across the pipe, find the induced emf across a diameter. (Hint: Look at Problem 11 above.)

15. Suppose a proton NMR resonance peak occurs at a frequency of 453 MHz when measured at a temperature of 300 K. Find the average number of protons with spin-down versus spin-up per every million protons in the sample.

16. Calculate the difference in spin state energy levels in joules and in eV for a mole of protons in a 14 T magnetic field using the free proton gyromagnetic ratio, γ_p, of $2.68 \times 10^8\ \text{s}^{-1}\text{T}^{-1}$ and the fact that the proton magnetic moment is equal to $\mu_z = \gamma_p(h/4\pi)$. What is the expected proton resonance frequency? How many more protons, per million, will have their spins up than down?

17. Given the gyromagnetic ratios for protons and for C^{13} of 2.68×10^8 and $0.67 \times 10^8\ \text{s}^{-1}\text{T}^{-1}$, respectively, calculate the resonant frequencies, in MHz, for these two nuclei at a magnetic field strength of 1.41 T.

18. In MRI, suppose the static magnetic field is 0.5 T and a field gradient of 0.5×10^{-4} T/cm is applied across a person's head that is being imaged. If protons are being imaged, by what percent does the resonant frequency for protons in identical environments vary across the 15 cm width of the person's head?

19 Electromagnetic Waves

In this chapter we begin to investigate the broad area of physics concerned with electromagnetic radiation and waves. The next four chapters discuss various aspects of optics, the science of light. Here we first introduce electromagnetic waves and some of their properties including their structure, energy, and momentum. Laser (or optical) tweezers is an exciting new technique that allows manipulation of microscopic structures or of individual macromolecules even within living cells. We introduce the technique based on the momentum contained in an electromagnetic wave, and show that laser tweezers represents a novel rapidly growing experimental technique. A brief discussion of photons, the elementary quanta of electromagnetism, and the notion of wave–particle duality are given in order to understand the basis for a large array of spectroscopic techniques using the various portions of the electromagnetic spectrum.

1. ELECTROMAGNETIC WAVES

Electromagnetic (EM) radiation is created whenever charges accelerate. This occurs, for example, when time-varying currents run up and down the transmitter of a radio station or when atoms bounce around inside a fluorescent light bulb. The "news" that acceleration has occurred travels outward at the speed of light c. Figure 19.1 is a picture of a sphere of radius r surrounding a negatively charged electron. At time t, the radiation E-field is measured everywhere on the surface of the sphere. That field (a few of the vectors are shown in red) is due to the acceleration of the electron at a time earlier than t. The relevant earlier acceleration is shown in the figure (in black). The time at which the fields shown were created is the present time t minus the time necessary for the radiation to travel a distance r; that is, $t - r/c$. At that earlier time the electron was at the center of our sphere.

The electric field radiated by the electron in Figure 19.1 has a magnitude given by

$$E_{\text{rad}} = \frac{e}{4\pi\varepsilon_0 c^2 r}\sin(\theta)a\left(t - \frac{r}{c}\right), \tag{19.1}$$

where e is magnitude of the electronic charge, θ is the smallest angle between the acceleration direction and the line connecting the charge to the point of observation, and $a\,(t - r/c)$ is the value of the charge's acceleration at the time r/c before the present time t. Because of the θ dependence, the magnitude of the field is a maximum on the equator of the sphere (where $\sin(90°) = 1$) and zero at the poles (where $\sin(0°) = \sin(180°) = 0$). In other words, if you look at a charge directly along its line of acceleration you don't see any radiation; the maximum radiation is observed at right angles to the acceleration. The radiation E-field vectors are tangent to the sphere everywhere and point as shown. They are always perpendicular to the direction of propagation of the radiation.

The E_{rad} field of Equation (19.1) decreases with distance from the source as $1/r$ and not as the usual $1/r^2$ dependence of the electrostatic field. We can demonstrate that this

J. Newman, *Physics of the Life Sciences*, DOI: 10.1007/978-0-387-77259-2_19,

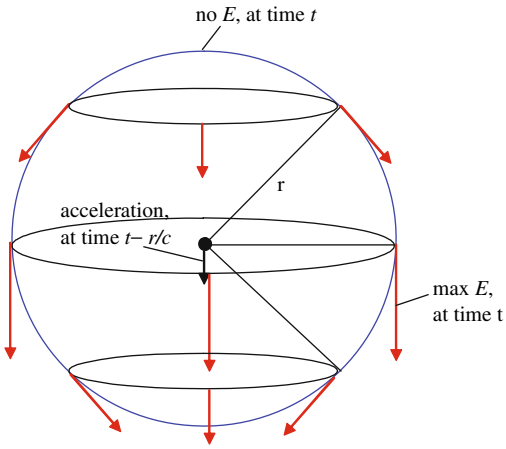


FIGURE 19.1 *A map of the E-field on a sphere of radius r due to an accelerating electron shown at an earlier time, t − r/c, located at the center.*

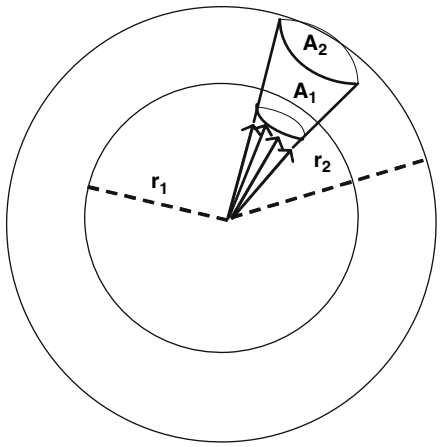
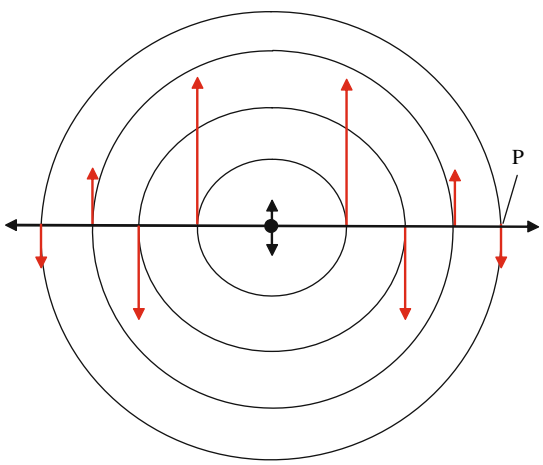


FIGURE 19.2 *A point source radiating at the center of two concentric spheres showing that the energy passing through A_1 and A_2 per second must be the same and thus that the intensity must decrease as $1/r^2$.*

must be true from the following geometric argument. The oscillating electron supplies energy at a certain constant rate so that the power P carried by the radiation is constant. Remember that power is proportional to the intensity, which is itself proportional to the square of the field. As this energy is carried away by the spherical radiation wave traveling at a constant speed c, the total amount of energy crossing any spherical surface per second must be the same. As shown in Figure 19.2, because the surface area of a sphere increases with the square of the radius (remember that the surface area of a sphere of radius r is given by $A = 4\pi r^2$), the total energy crossing a spherical surface per second at two different radii r_1 and r_2 can only be equal if the intensity decreases as $1/r^2$. This follows because if the intensities I_1 and I_2 represent those at radii r_1 and r_2, then we must have $P = I_1 4\pi {r_1}^2 = I_2 4\pi {r_2}^2$, so that $I \propto 1/r^2$. From this, we can conclude that the E-field then must vary as $\sqrt{1/r^2} = 1/r$ in agreement with the above expression for E_{rad}.

Often electrons have heavy, positively charged, protons nearby (as in an atomic nucleus, e.g.). The electric field causing the electrons to accelerate will also cause the protons to accelerate, but because the protons are 2000 times more massive, their motion usually can be ignored. In that case, the total electric field near a neutral glob of matter equals the static Coulomb fields of the protons plus the static Coulomb fields of the electrons plus the radiation fields of the electrons. The static fields approximately cancel out, leaving just the radiation fields. That's why we can't measure the static electric field of a galaxy 10 billion light years from Earth, but can detect its radiation field just fine (as light, with our eyes; oh, okay, and maybe a telescope).

FIGURE 19.3 *A series of spherical waves of radiation emanating from an oscillating electron at the center. Oscillating E-fields with decreasing amplitude with distance from the center are shown.*



If the electron in Figure 19.1 oscillates up and down periodically, it emits a continuous series of concentric spheres of radiation, each with radius expanding at the speed of light. Some of these spheres are shown at one instant in Figure 19.3. A few E-vectors are also shown. E is large near the oscillating electron and smaller farther away. Suppose the period of oscillation of the electron is T seconds. The concentric spheres shown are generated every $T/2$ seconds and the distance between them is $cT/2$. If one sits at a fixed point in space, such as P in the figure, the sequence of E's that pass through there will vary sinusoidally in time with a fixed amplitude. This passing wave of E is transverse. Furthermore, if we look at E at any instant over a small planar patch that is tangent to any one of the spheres of radiation, the value of E will be about the same everywhere in the patch. The size of such a patch of uniform E will be small in close to the radiating electron and

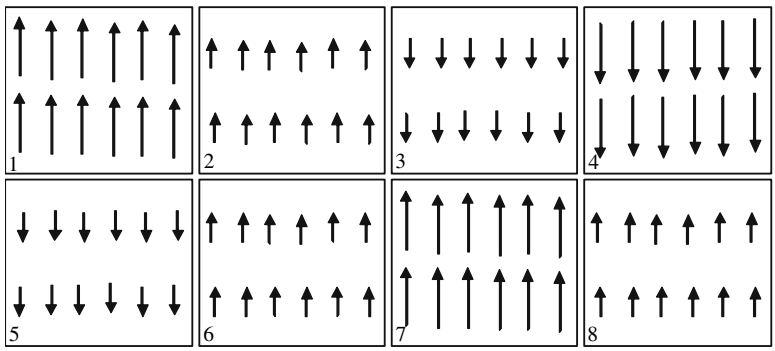


FIGURE 19.4 *Series of time-panels of the electric field at point P in Figure 19.3 (see text).*

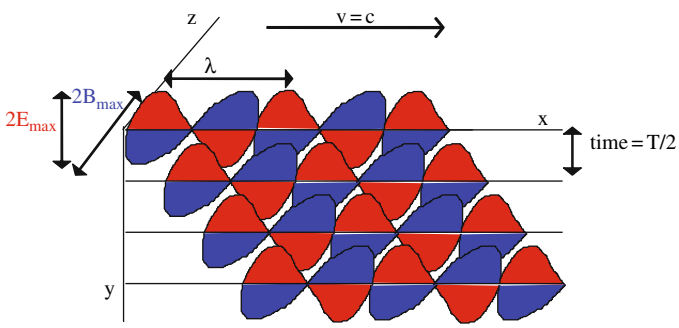
will be larger the farther out one goes. A set of *E*-vectors that are all the same at one instant over a plane is called an "electric field plane wave."

Figure 19.4 depicts a series of time-lapse photographs of the electric field, produced by the radiating electron in Figure 19.3, over a small planar patch that is perpendicular to the direction of propagation of the radiation. Read the panels in cartoon fashion: left to right, top to bottom. Thus, at first *E* is large and pointing up, then small and up, then small and down, then large and down, then small and down, then small and up, then large and up (one complete cycle after the first panel), and so on. The period of the oscillation equals "six panels" in this strip. The cartoon keeps running along in the same way, over and over. Because the electron in Figure 19.3 is always oscillating vertically, the *E*-fields in the panels of this strip are also always pointing vertically because we are at point P on the sphere's equator (see Figure 19.1). Such a plane wave is said to be "linearly polarized." More typically, the source of electromagnetic radiation is a large number of electrons that tumble about irregularly. Usually, such electrons will vibrate in concert with each other for only a short time. The cartoon for radiation from a collection of tumbling electrons will have panels where *E* is vertically oriented for a while, then oriented at some other angle, then oriented at another angle, and so on, with no connection between panels. Such a plane wave is said to be "unpolarized."

Because time-varying *E* makes *B*, there is a magnetic field that travels along with *E*. When *E* is large, so is *B*. When *E* is zero, so is *B*. The magnetic field is perpendicular to *E* and both are perpendicular to the direction of propagation. The direction of *B* at any moment can be determined by the following usual right-hand rule equivalent to $(\vec{E},\vec{B},\vec{v})$ forming a right-handed coordinate system such as (x, y, z): place the thumb of your right hand in the direction of propagation of the wave and your extended fingers in the direction of *E*; curl your right-hand fingers 90° to point in the direction of *B*.

When far from the source, these fields can be described as plane waves shown schematically in Figure 19.5. Here an electromagnetic plane wave is shown to be composed of oscillating electric and magnetic fields traveling along the *x*-axis. Both $\vec{E}$ and $\vec{B}$ are found to lie in a transverse plane, perpendicular to the *x*-direction along which the wave travels. Furthermore, $\vec{E}$ and $\vec{B}$ are also perpendicular to each other and oscillate together, in phase with each other, as the wave travels along at speed *c*. We can write each of the magnitudes of the fields in the form of traveling waves (as in Section 3 of Chapter 10)

FIGURE 19.5 *A portion of a traveling electromagnetic wave with in-phase mutually perpendicular E (*red, *along y-axis) and B (*blue, *along z-axis) fields. The wave is shown moving toward the right along the x-axis at four different times separated by T/2, the time for the wave to move a distance λ/2, where λ/T = c.*



$$E(x,t) = E_{\max}\sin(kx - \omega t),$$
$$B(x,t) = B_{\max}\sin(kx - \omega t), \qquad (19.2)$$

where, as before,

$$k = \frac{2\pi}{\lambda} \text{ and } \omega = \frac{2\pi}{T},$$

with λ the wavelength and *T* the period of the wave ($T = 1/f$ where *f* is the frequency of oscillation of the wave). The speed of the wave is given by $v = c = \omega/k$, (equivalent to $c = \lambda f$;

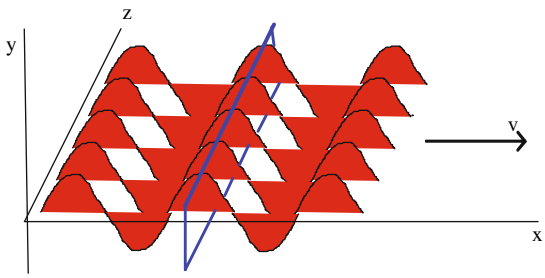


FIGURE 19.6 *A plane electromagnetic wave (showing only the E-field) traveling along the x-axis with its common plane wavefront highlighted. This entire wave is shown at the same time, unlike the previous figure, which is a series of four different snapshots in time.*

see Equation (10.10)). As a consequence of Maxwell's equations, the values of E_{max} and B_{max} are related to each other as

$$\frac{E_{max}}{B_{max}} = c. \tag{19.3}$$

We show shortly that this expression leads to the fact that both the $\vec{E}$ and $\vec{B}$ fields carry the same contribution to the total energy of the wave.

Remember that Figure 19.5 is like four snapshots of the fields (if they were somehow made visible) traveling along the *x*-axis. As time ticks on, the wave shown will move along the *x*-axis at speed *c* with the $\vec{E}$ and $\vec{B}$ fields oscillating with period *T* at any particular *x* location, as in Figure 19.4 for the *E*-field. The wave also has some spatial extent in the transverse plane (not shown). A *plane wave* has a flat or plane wavefront (the locus of all points at which $\vec{E}$ is in phase; e.g., all the crests of the $\vec{E}$ wave). For this type of idealized wave, the amplitudes in Equation (19.2) are constants, not varying with the distance the wave has traveled (Figure 19.6).

Recall that at the end of the previous chapter we made an analogy between EM waves and waves on a string. There we loosely associated the velocity of the string with *E* and the slope of the string with *B*. The correct analogies are that *E* corresponds to transverse velocity and *B* corresponds to stretch of the string. On a string, velocity is associated with kinetic energy and stretch is associated with potential energy. For a traveling wave on a string these energies are at a maximum together and they travel at the wave speed. The same is true of EM waves. The *E* part of the energy and the *B* part are in phase, having maxima together and zeroes together, and they both travel at the speed of light.

We have seen that there is an energy density associated with an electrostatic field given by Equation (15.22),

$$\frac{PE}{V} = \frac{1}{2}\varepsilon_0 E^2.$$

It can be shown that energy can also be stored in a magnetostatic field and that the energy density associated with *B* is given by

$$\frac{PE}{V} = \frac{1}{2}\frac{B^2}{\mu_0}.$$

Given this, it should not be surprising that an electromagnetic wave made from oscillating *E* and *B* fields also has an associated energy density given by

$$\frac{PE}{V} = \frac{1}{2}\left(\varepsilon_0 E^2 + \frac{B^2}{\mu_0}\right), \tag{19.4}$$

where the *E* and *B* fields vary sinusoidally. We can rewrite this expression using Equation (19.3) to substitute for $B = E/c$ and the fact that $\mu_0 = 1/(\varepsilon_0 c^2)$ (from Equation (18.11)), where we have also used the fact that *E* and *B* are in phase so that Equation (19.3) holds not only for the maximum values but at all times. We then have that $B^2/\mu_0 = \varepsilon_0 E^2$, so the energy density can be rewritten as

$$\frac{PE}{V} = \frac{1}{2}\left(\varepsilon_0 E^2 + \varepsilon_0 E^2\right) = \varepsilon_0 E^2. \tag{19.5}$$

Because each of the terms in the bracket in Equation (19.5) is equal, the electric and magnetic fields each contribute equally to the total energy density of the EM wave.

As an EM wave moves with speed *c* in vacuum, it carries energy. If we imagine such a wave traveling along the *x*-axis (Figure 19.7), then in a time Δt the wave will move a distance of $c\,\Delta t$. In that time a volume of the wave equal to $Ac\,\Delta t$ will sweep

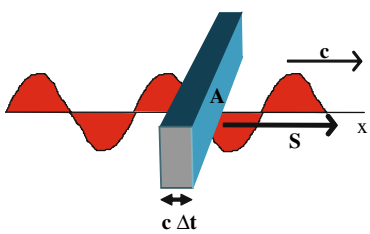


FIGURE 19.7 *An EM wave carries energy per unit area per unit time according to Equation (19.6).*

through a cross-sectional area A perpendicular to the wave velocity. The energy transported by the wave in time Δt through the area A is then

$$\left(\frac{\text{PE}}{V}\right)Ac\Delta t = \varepsilon_0 E^2 Ac\Delta t,$$

so that the EM wave energy per unit time per unit area, S, is equal to

$$S = c\frac{\text{PE}}{V} = c\varepsilon_0 E^2. \tag{19.6}$$

$\vec{S}$ is known as the *Poynting vector*. It points in the direction of travel of the wave, and is measured in units of J/s/m^2, or W/m^2. Because E is taken as a sinusoidal function, S varies with time as well.

The average value of S represents the intensity I of the wave, or the mean energy flow per unit time per unit cross-sectional area. Intensity is important because that's what the eye detects when the EM radiation is in the visible range. Because E is given by Equation (19.2) and the average of the function $\sin^2(x)$ over one period is equal to ½, we have that

$$I = \frac{1}{2}\varepsilon_0 c E_{\text{max}}^2. \tag{19.7}$$

The intensity of EM waves is a measurable quantity; detectors can measure the amount of energy per unit time and per unit area that reach them. On the other hand, the Poynting vector fluctuates with time, often much too fast to be detected directly. For example, visible light has a frequency of about 10^{15} Hz. In order to detect such rapid fluctuations a time resolution of about $1/10^{15}$ s = 1 fs, is required and this is just at the edge of our current abilities.

Example 19.1 An EM plane wave traveling along the x-axis has an effective cross-sectional area of 1.5 cm^2, a maximum electric field of 1500 N/C and a frequency of 4×10^{15} Hz. Find each of the following quantities: its maximum B field, energy density, an expression for the Poynting vector, the intensity of the wave, and the energy striking a 0.5 cm^2 area with its normal along the x-axis in 10 s.

Solution: We first find the amplitude of the magnetic field from Equation (19.3) to be $B = E/c = 5 \times 10^{-6}$ T. These values then allow us to calculate the energy density, from Equation (19.4) to be 2×10^{-5} J/m^3. Alternatively we can use Equation (19.5) directly to find the same result. The Poynting vector then has an amplitude given by Equation (19.6) to be $S_{\text{max}} = (\text{PE}/V)c = 6000$ W/m^2 along the x-axis and varies at the very high frequency of 4×10^{15} Hz. It can be written as $\vec{S} = 6000 \sin(2\pi \cdot 4 \cdot 10^{-15}t)$ and is directed along the x-axis. Its average value over time is the intensity given by Equation (19.7) as $I = {}^1/_2 S_{\text{max}} = 3000$ W/m^2. Finally, if this wave strikes the given surface, the power reaching the surface is just $P = IA = 0.15$ W, so that in 10 s the energy absorbed will be 1.5 J.

Example 19.2 If the maximum intensity of an EM wave is 1000 W/m^2 (about what it is in sunlight reaching the Earth), what is the maximum E?

Solution: Solving for E from Equation (19.7), we find $E = \sqrt{2I/\varepsilon_o c} = 870$ V/m, or about 9 V/cm. Such a field drives as much current through a cm of skin as a 9 V battery!

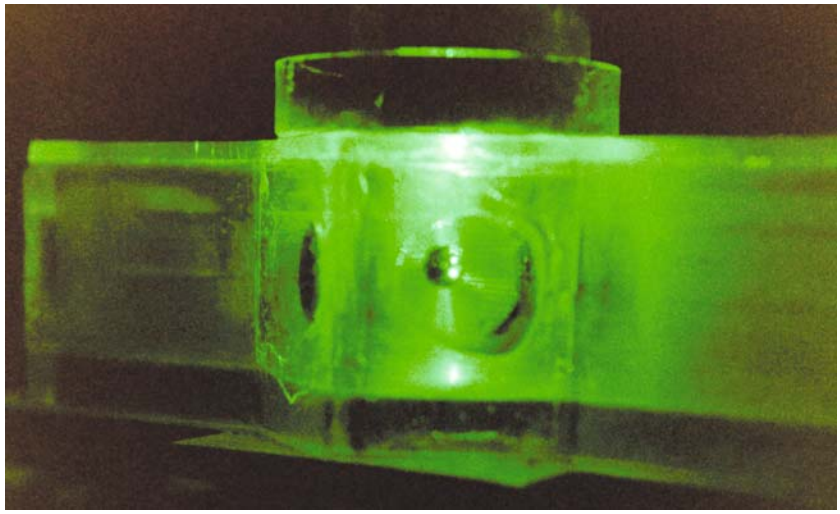
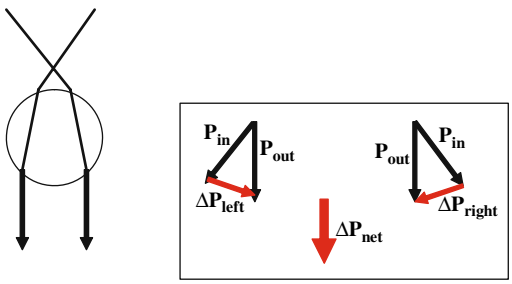


FIGURE 19.8 *Optical levitation of a drop of glycerol at the center of the chamber by a vertical green laser beam.*

When an electric radiation field encounters a charge, it makes the charge jiggle with the same frequency as the radiation. A jiggling charge is accelerating, so it radiates as well, emitting "induced" radiation with the same frequency as the "incident" radiation. This induced radiation travels outward in all directions. The total E observed at any point in space is the vector sum of all E's, the incident E's as well as all induced E's. This is just the superposition principle for electric fields that we've discussed previously. We return to this in Chapter 22.

In addition to carrying energy, an electromagnetic wave also carries linear momentum, and hence can exert a force. Although the amount of momentum or force is usually small compared to ordinary forces we experience, the force generated by an intense light beam, for example, from a laser, is enough to provide an upward force on small particles to balance their weight and suspend them in air. The pressure exerted by EM waves is known as *radiation pressure*. Figure 19.8 shows a small drop of glycerol being suspended in water by the radiation pressure of a laser beam. The possibility of "trapping" micron-sized spheres was first demonstrated in the early 1970s. In the next section we discuss a new technique that uses radiation pressure to allow the direct manipulation of microscopic objects.

FIGURE 19.9 *The refraction of a focused laser beam passing through a transparent sphere. The magnitude of the light beam's momentum depends on its color and its intensity. For fixed color and intensity only the direction of the light momentum changes on refraction. The insert shows the change in momentum of the extreme light rays shown. The symmetric situation results in a net change in momentum of the laser beam along its propagation axis, so that there is a reaction force upwards on the sphere, toward the focus point.*



2. LASER TWEEZERS

First conceived and developed in the mid-1980s by Ashkin and colleagues at AT&T Bell Laboratories, laser, or optical, tweezers is a method of using radiation pressure to trap atoms, molecules, or larger particles. In applications with the simplest possible arrangement using a single laser beam, particles with sizes in the range of several hundred microns down to about 25 nm can be "trapped" and moved about using the radiation pressure of the EM radiation. How does radiation pressure trap such particles?

If a plane electromagnetic wave is incident on a particle, the radiation pressure on the particle would be such as to propel it along the direction of the beam. This is due to the fact that the reflected wave results in a net decrease in forward momentum of the wave. Conservation of momentum for the system composed of the EM wave and the particle then dictates that the particle must sustain a forward momentum. This process is responsible for the suspension of micron-sized spheres in gravity as in Figure 19.8 when the beam intensity is adjusted so as to just balance the sphere's weight. A higher intensity beam would propel the sphere upwards, whereas a lower intensity beam would allow the sphere to fall but at a reduced acceleration compared to g. This analysis does not as yet explain how a laser or optical tweezers can trap a particle.

Consider a transparent sphere with dimensions large or comparable to the wavelength of the EM wave that impinges on it. We show in the next

chapter that at an interface between different materials, EM waves are bent, or refracted, as shown in Figure 19.9. This process is studied in detail later; for now, all we need to know is that the arrows drawn to represent the direction of travel of the wave will be bent at each surface. Refraction of an EM wave is the basis for laser tweezers. In Figure 19.9, the EM wave is shown coming to a focus above the sphere and two rays are drawn to typify the path of the wave. As shown in the insert, if the change in momentum of the wave is determined for this situation, the net change is in the forward direction, so that there is an equal and opposite change in the sphere's momentum resulting in a net force on the sphere in the opposite direction, toward the focus point. Similarly if the focus point is below the sphere, there will be a restoring force due to the radiation pressure directed again toward the focus point. These longitudinal forces directed toward the focus point act to stabilize or "trap" the particle longitudinally.

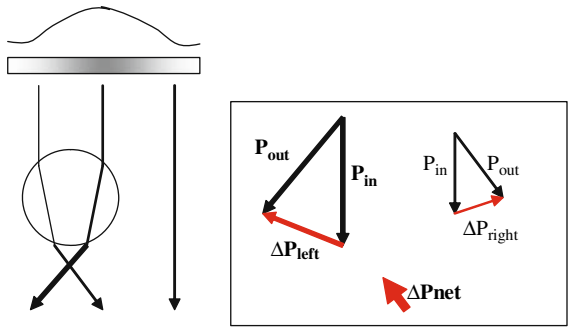


FIGURE 19.10 *Similar to the previous figure, but with the laser beam having a typical transverse intensity profile as shown and with the sphere off-center and above the focus point. The insert shows the momentum change of the extreme laser beam rays, now no longer symmetric. This analysis shows that, because the intensities are not symmetric, there will be a net change in the light momentum as shown and so the sphere feels an oppositely directed reaction force not only along the beam toward the focus point, but also transversely toward the beam axis.*

Figure 19.10 shows the same arrangement with the sphere off-center but with the beam having a variation in intensity across its cross-sectional area. A similar momentum change analysis, shown in the insert, reveals that in addition to a force toward the focus point, in this case downward, there will also be a transverse force directed toward the more intense portion of the beam along its axis. In a real single beam laser tweezers arrangement, the beam will have its maximum intensity at the center and the sphere would be trapped transversely to lie along the center and at the focus point.

Typical forces capable of being exerted are in the pN (10^{-12} N) range. In order to move the trapped particle about, either the laser beam itself or the sample, sitting on a microscope stage, is moved. An experimental station uses a good quality inverted microscope with an optical port for the laser as shown in Figure 19.11. Usually near-infrared

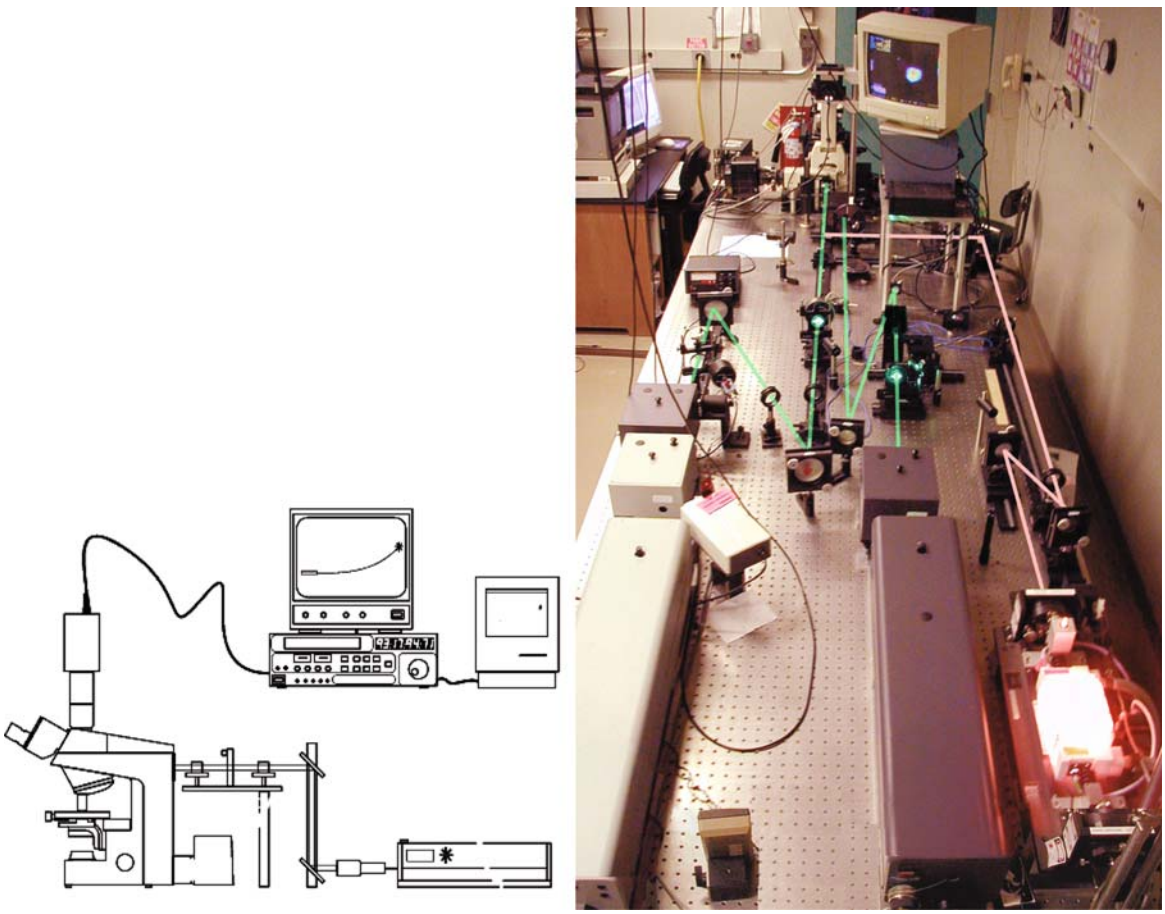


FIGURE 19.11 (left) *Schematic of a basic laser tweezers experimental setup;* (right) *Laser tweezers experimental station with microscope at far end of optical table.*

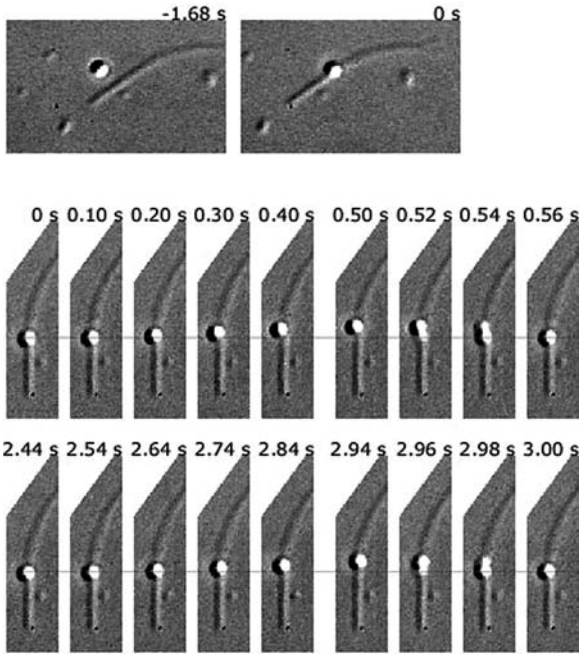


FIGURE 19.12 *Time series showing a plastic sphere with a motor protein attached that is trapped in laser tweezers. The motor protein is driving the sphere upwards on the axoneme but the trapping force is just greater and able to keep the sphere trapped although it wobbles about the trap center (red line).*

laser light with a wavelength of about 1 μm is used with biological samples. Although the beam is invisible to the human eye and therefore needs to be detected with an infrared-sensitive CCD camera for recording, its use usually avoids the problems of light absorption and subsequent heating of samples that can occur when using visible light. The sample can be directly viewed through the usual microscope eyepiece using a standard visible light source of the microscope. Care must be exercised to ensure that the laser beam is not directed on the sample when viewing by eye because the beam is invisible and, for a sufficiently intense laser beam, can cause damage to the eye. Laser beams have a cross-sectional intensity profile that is bell-shaped with a maximum in the center and therefore automatically act to trap particles in the transverse direction according to the above discussion.

Optical tweezers of biological samples takes place in an aqueous solvent, whether a solution in which the biomolecules of interest are placed or the cytoplasm of a cell. As the trapped object is moved, there will be an immediate viscous drag force acting that will balance the trapping force, so that the object will move at a constant velocity. This Stokes' drag force (see Equation (9.6))

$$F_f = 6\pi\eta r\, v$$

can be used to calibrate the trapping force achieved at a particular laser intensity and beam geometry. By measuring the maximum velocity with which a micron-sized plastic sphere of known radius can be dragged by the laser tweezers, the trapping force can be determined by its balance with the Stokes' force because the viscosity and sphere radius are known. In this way, the maximum applied trapping force can be found as a function of the laser intensity.

If a sphere is attached to one end of a linear macromolecule such as DNA or a filamentous protein such as actin and the other end of the macromolecule is immobilized, laser tweezers can be used to stretch the macromolecule a given distance and measure the minimum applied trapping force at which the sphere just "pops out" of the trap (Figure 19.12). Under this condition, the applied trapping force is just equal to the force the macromolecule is exerting on the sphere, allowing a measurement of the elastic force exerted by the macromolecule.

There are already many applications for which laser tweezers have been used. Individual motile (swimming) organisms, such as E. coli bacteria, have been trapped by laser tweezers while they continue to live normally with flagella beating (Figure 19.13). Similarly laser tweezers can be used to manipulate subcellular organelles within a living cell. The IR laser passes through the cell membrane and can trap large organelles or structures, such as individual chromosomes, that can then be moved about. It can also be used to exert forces directly on a cell membrane (Figure 19.14). More quantitative measurements of forces have been made with laser tweezers as mentioned in the last paragraph. In this way the stiffness and breaking strength of such molecules can be measured. Time-resolved measurements on trapped micron-sized plastic spheres attached to the ends of an actin filament or a microtubule have been made to study the forces generated when single molecules of either myosin or kinesin move along the respective filaments (Figure 19.15). These motility assays have recently achieved measurements at subpiconewton force and nanometer displacement resolutions with a time resolution of about 1 ms, allowing the results of single molecule interactions to be studied in great detail.

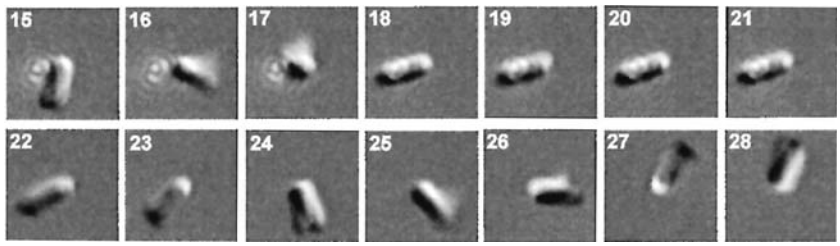


FIGURE 19.13 *Images of an E. coli bacterium rotating into a focused laser trap in frame 18 and remaining trapped for four frames before shutting the trap, releasing the bacterium.*

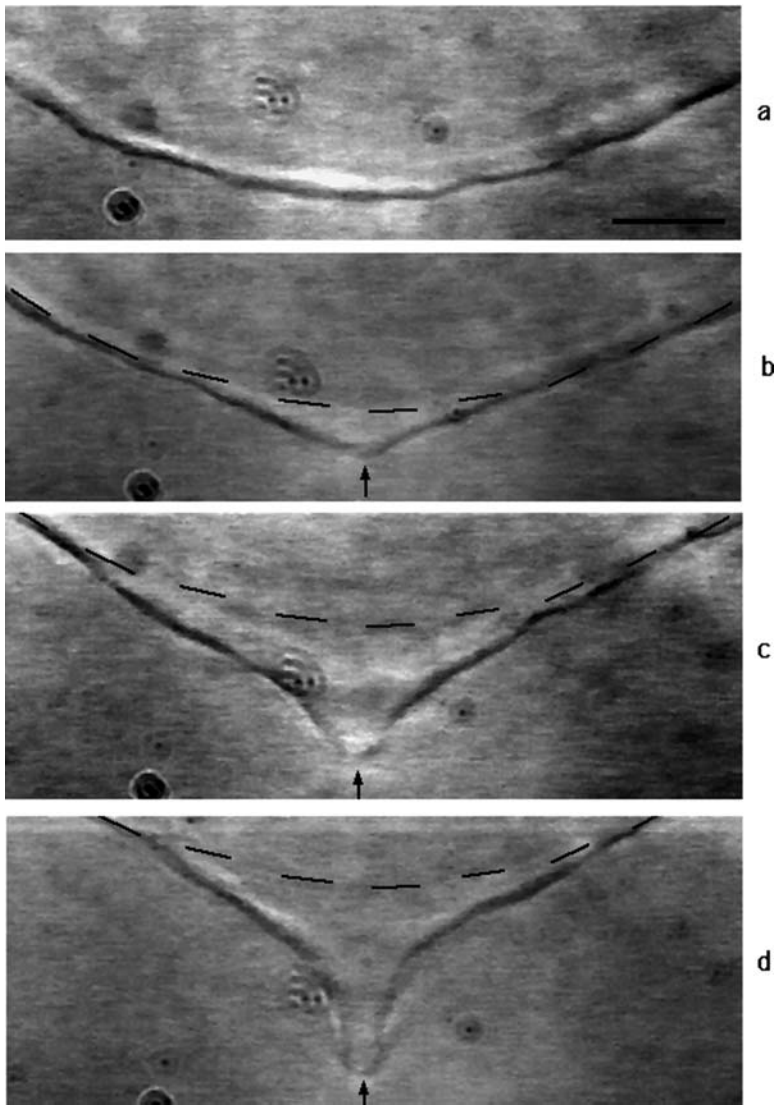


FIGURE 19.14 *Laser tweezers applied at the arrow exert mechanical forces on a membrane, stretching it from its undeformed contour* (dashed curve).

3. POLARIZATION

We have seen that electromagnetic waves are transverse waves with their electric and magnetic fields both lying in the transverse plane (perpendicular to the wave velocity) and also perpendicular to each other. The direction of the electric field is known as the polarization direction of the wave. If, as the wave propagates, the electric field remains along the same direction in space, the wave is said to be *linearly polarized*. Sometimes, in this case, the wave may be referred to as vertically or horizontally polarized if the $\vec{E}$ field points in either of those directions.

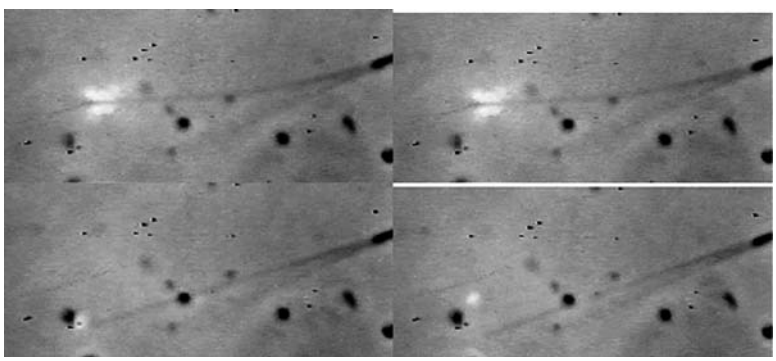


FIGURE 19.15 *A time series of phase contrast images of a microtubule, deflected by laser tweezers (visible as a set of four bright spots), returning to its unbent position. The images are taken at* $t = -0.04$ *s* (upper left; *the laser is switched off at* $t = 0$), t = *0.12 s* (top right); $t = 0.52$ *s* (bottom left); *and* $t = 1.0$ *s* (bottom right). *Laser tweezers can give information about the forces and displacements involved and can be used to micromanipulate individual macromolecules.*

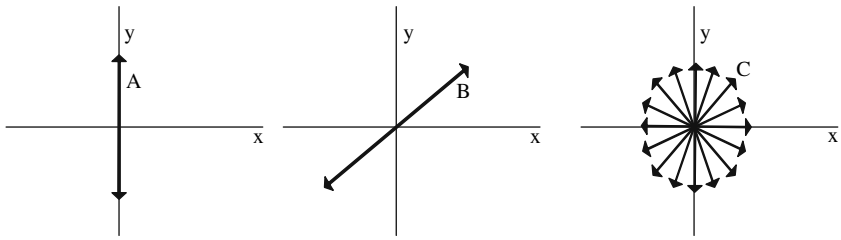


FIGURE 19.16 *Three types of EM polarization indicated by doubled-pointed arrows in the transverse plane: A, linearly polarized in the vertical direction; B, linearly polarized at 45° to the vertical; C, unpolarized, schematically drawn, where the electric field randomly orients as a function of time. Note that the EM wave is traveling along the z-axis in all cases.*

In more general terms, because $\vec{E}$ is confined to the transverse plane, there are two independent orthogonal directions possible, let us say the x- and y-axes as shown in Figure 19.16. Depending on the source of the EM radiation, the electric field may or may not have a definite polarization direction. For example, light from the sun, from a flame, or from incandescent or fluorescent light bulbs has its origin in the independent motions of huge numbers of atoms or molecules and has no particular polarization direction. Such light is said to be *unpolarized*, meaning that the superposition of the $\vec{E}$ field directions from all of the individual sources of light (atoms/molecules) leads to a random orientation for $\vec{E}$ as a function of time.

Various methods can be used to change the polarization properties of EM radiation. We discuss some of these in more detail in our discussions of optics in the next chapters. Here we illustrate one particular method, the use of an absorption polarizer such as a sheet of Polaroid, for producing a linearly polarized light wave from unpolarized light. Polaroid sheets contain long chains of organic molecules that are preferentially oriented in one direction. When incident unpolarized light falls on such a sheet, the oscillating electric field component along the chain direction is preferentially absorbed because, simply put, the electrons are able to move along that direction and take up some of the energy corresponding to light polarized along the chains. In contrast, the electric field polarized perpendicular to the chains is not able to interact strongly with electrons because they have more limited mobility in the transverse direction and this electric field polarization passes directly through the otherwise transparent sheet. The net effect is that after passing through a Polaroid sheet, unpolarized light becomes linearly polarized along the Polaroid axis, which is perpendicular to the organic chain axis, as shown in Figure 19.17.

Other forms of EM radiation behave quite similarly, although the nature of the polarizer device will be different. For example, an unpolarized microwave beam (with a wavelength of several cm) can be polarized by passing it through a set of parallel wires, such as the metal baking rack used in a conventional oven. The polarization axis in this case is perpendicular to the wires for the same reason as in Polaroid film: electrons are better able to absorb the energy of the microwaves along the axis of the wire, leaving the transmitted microwaves preferentially polarized perpendicular to the wires.

Polaroid sunglasses function just as described above, having their polarization axis vertical. What is the advantage of Polaroid sunglasses over others that simply attenuate the total intensity regardless of the polarization direction? As was mentioned above, sunlight is unpolarized and so there would be no benefit in preferentially blocking one polarization direction over another in looking directly at sunlight. However, when unpolarized light reflects from a surface, more of the horizontally than vertically polarized light is reflected and this is commonly seen in the form of "glare". As long as you hold your head upright, Polaroid sunglasses are quite effective in blocking this glare (Figure 19.18).

The polarization properties of a wave may be investigated using polarizers placed in the path of the wave. If we choose to orient a first polarizer with its axis vertical then regardless of the polarization of the incident wave, only vertically polarized waves will be transmitted through the polarizer, assuming an ideal polarizer. If the incident wave was unpolarized, then because on average horizontal and vertical polarizations are present in equal amounts, half the incident intensity will pass through the first polarizer. If the initial wave were already vertically polarized then all of its intensity would be transmitted

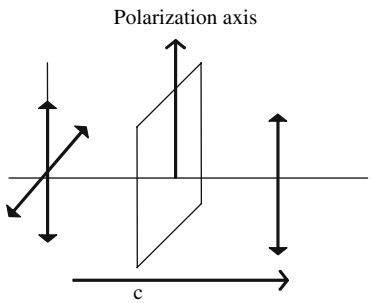


FIGURE 19.17 *Vertically polarized light is obtained from an unpolarized light beam, traveling to the* right*, that is incident on a sheet of Polaroid oriented with its transmission axis oriented vertically.*

whereas, on the other hand, if it were initially horizontally polarized no light would be transmitted.

Suppose that a second polarizer, often called an *analyzer*, is now placed in the path of the vertically polarized wave transmitted by the first polarizer. We can treat the vertically polarized wave as a superposition of two linearly polarized waves, one parallel and one perpendicular to the analyzer's axis, making an angle θ with that of the polarizer (see Figure 19.19). Only the parallel component will be transmitted through the analyzer. If the incident electric field on the analyzer is E_0, its transmitted component along the analyzer's axis, is

$$E_t = E_0 \cos\theta. \tag{19.8}$$

Clearly if the analyzer is rotated around, there will be two positions ($\theta = \pm 90°$) at which the transmitted field vanishes and the polarizers are then said to be "crossed." Because, according to Equation (19.7), the intensity is proportional to the square of the electric field, the transmitted intensity I_t is related to the incident intensity I_0 by

$$I_t = I_0 \cos^2\theta. \tag{19.9}$$

Equation (19.9) is sometimes known as the law of Malus.

FIGURE 19.18 *A Polaroid screen (on the* right*) used to block glare on a computer monitor.*

Example 19.3 A vertically polarized 0.15 W laser beam is incident on a polarizer with its transmission axis 45° to the vertical. The beam then passes through a second polarizer with its transmission axis horizontal. What is the intensity of the transmitted beam? What would be the intensity of the transmitted beam if the second polarizer were removed? What would it be with the second polarizer in place if the first polarizer were removed?

Solution: Using Malus' law the intensity passing through the first polarizer will be $I_1 = 0.15 \cos^2 45 = 0.075$ W. Then the intensity passing through the second polarizer will be $I_2 = I_1 \cos^2 45 = 0.038$ W, because the second polarizer's transmission axis makes a 45° angle with that of the first. If the second polarizer is removed, the transmitted intensity will simply be $I_1 = 0.075$ W, whereas if the first polarizer is removed there will be no transmitted intensity because the angle between the vertically polarized incident beam and the horizontal second polarizer is 90°.

In our discussions the transverse electric field vector was pictured in general as two orthogonal components along a pair of axes in the transverse plane. For unpolarized light each of these components varies randomly in time and, on average, has equal amplitude. For linearly polarized light along some arbitrary direction, the components of $\vec{E}$ oscillate

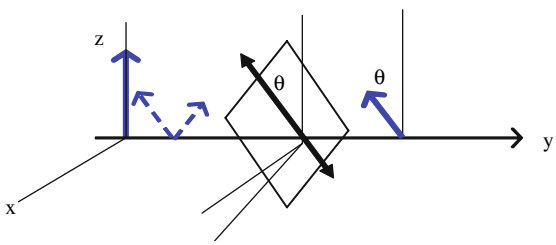


FIGURE 19.19 *Vertically polarized (along z) wave traveling to the right incident on an analyzing polarizer in the transverse plane oriented with its transmission axis at angle θ from the vertical. The original E field can be decomposed into the two components (shown as* dashed lines*) along the polarizer axes. Only the* portion of E *along the transmission axis of the polarizer (in* black*) is then transmitted.*

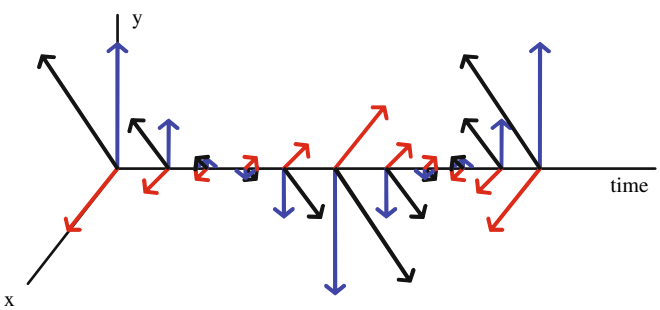


FIGURE 19.20 *In phase x-* (red) *and y-* (blue) *components of an electric field add to give linearly polarized light* (black)*, as shown in this time sequence covering one period of oscillation.*

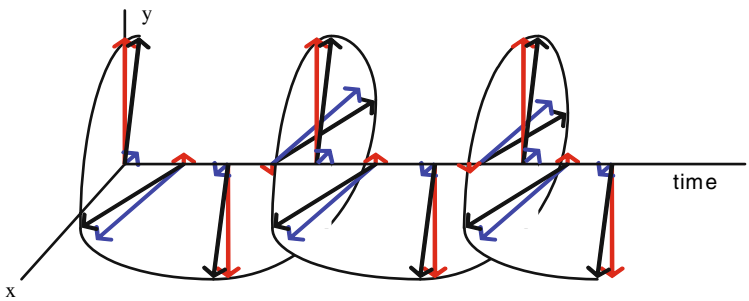


FIGURE 19.21 *Circularly polarized light with the E_x (blue) and E_y (red) components 90° out of phase so that the net E vector (black) tip traces out a circle in the transverse plane (or a helix in space, as shown).*

in phase and add to give a resultant along a fixed linear direction as shown in Figure 19.20. Different relative magnitudes of the two components of $\vec{E}$ result in linearly polarized light at different angles. For example, the larger the red component is compared to the blue, the closer the resultant E points to the x-axis.

Another interesting type of polarization is known as circular (elliptical) polarization, where the $\vec{E}$ field moves around the transverse plane in a circle (ellipse). We can think of this type of polarization as arising from x- and y-components of $\vec{E}$ that are 90° out of phase as shown in Figure 19.21. Starting with the y-component at a maximum and the x-component equal to zero, as the y-component decreases the x-component increases; the x-component reaches a maximum when the y-component vanishes; and so on. With both x and y-components equal in magnitude, the $\vec{E}$ field traces out a circular pattern in the transverse plane. In our example, the $\vec{E}$ field traces out a counterclockwise directed circle in the transverse plane as viewed from the right. If the x- and y-components are unequal, elliptical polarization occurs. Note that the actual path of $\vec{E}$ is a circular (or elliptical) helix in space as the wave moves along (as shown in Figure 19.21). These types of polarization are important in discussing the spectroscopic technique of circular dichroism (CD) in Chapter 23.

4. THE ELECTROMAGNETIC SPECTRUM

In our general discussion of electromagnetic waves in Section 1, we have seen that accelerating electric charges produce electromagnetic radiation in the form of transverse traveling waves of E and B. There we introduced the notion of a frequency and wavelength, connected through their product with the speed of light $c = f\lambda$, for the traveling wave expressions for E and B. The range of possible wavelengths, or frequencies, is enormous. Figure 19.22 shows the electromagnetic spectrum with wavelengths, frequencies, and

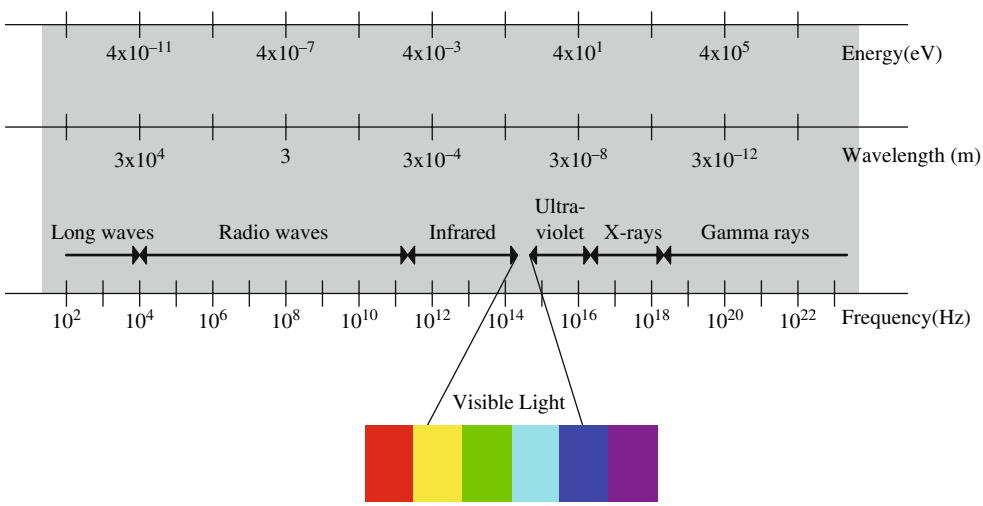


FIGURE 19.22 *The electromagnetic spectrum. The logarithmic scales for frequency, wavelength, and energy are shown together with the names of the various general regions.*

corresponding energies given using logarithmic scales. Electromagnetic radiation can arise through a large number of different types of processes, all having to do with accelerating electric charges. The broad categories of electric waves, radio waves, microwaves, infrared, visible, ultraviolet, x-ray, and gamma rays are used to distinguish the various parts of the EM spectrum that are produced in different ways. All electromagnetic waves are similar, in our classical wave picture, in having transverse electric and magnetic fields. They differ greatly not only in how they are produced but also in how they interact with matter.

We show in the next section that EM radiation, although for some considerations can be thought of as a classical wave, is actually composed of individual "wave-particles," known as photons. These elemental quanta of energy have associated frequencies and wavelengths. The energy carried in each photon is proportional to its frequency. We saw this briefly in Equation (18.8) of the previous chapter in connection with RF photons in NMR. This proportionality is the origin of the energy scale labeled in Figure 19.22. Higher-frequency EM radiation corresponds to higher-energy photons.

The lowest-energy, lowest-frequency radiation is produced by simple alternating current circuits in which the electric current is made to oscillate in time at a given frequency. Higher-frequency oscillations of current along an antenna result in radio or TV signals. Associated frequencies range from about 10 kHz to about 1 GHz (10^9 Hz). Interestingly, radio signals can also arise from nuclear magnetic dipole transitions, as was discussed in connection with NMR (nuclear magnetic resonance) and MRI (magnetic resonance imaging) in the last chapter. The highest frequencies obtainable in oscillating electric circuits, up to about 10^{11} Hz, are associated with microwaves (including radar). We have seen that microwave radiation is also emitted in the phenomenon of ESR (electron spin resonance) in paramagnetic materials.

Still higher frequency radiation has its origin in various energy transitions in atoms, molecules, or nuclei. When an atom or molecule makes a transition from a higher energy state to a lower one, often the energy difference is emitted in the form of a photon. In macroscopic systems the number of emitted photons is enormous and constitutes electromagnetic radiation. The higher the frequency of radiation is, the larger the energy difference between the two transition energy levels, or states, involved. The lowest such energy transitions occur in different spin states within the nucleus (NMR transitions; radio frequency radiation) and in paramagnetic electron spin transitions (ESR; microwave radiation) as mentioned. Higher-frequency radiation, such as infrared, visible, ultraviolet, or x-rays, is produced by transitions between various electron energy levels, whether closely spaced rotational or vibrational, or further spaced electronic states (these are discussed in Section 6 below and also in Chapter 25). The highest-energy, highest-frequency radiation, gamma rays, consists of high-energy photons resulting from energy transitions of the protons and neutrons within the nucleus.

It is also very instructive to examine the wavelengths of the various forms of electromagnetic radiation as shown in Figure 19.22. Note that longer wavelengths correspond to lower energy radiation. This is due to the inverse relation between frequency and wavelength and is explained further in the next section in connection with photon energies. Visible light is but a very narrow window of the EM spectrum, although the only one visible to the human eye. Other animals have visual receptors that extend out into the infrared or ultraviolet regions. We discuss the functioning of the eye in Chapter 21. Nonvisible radiation interacts with our bodies in various ways. We feel warmth from infrared radiation, get sunburned from damage that ultraviolet radiation causes to our skin, and receive small amounts of molecular damage from x-rays at the doctor or dentist and from naturally occurring gamma radiation. In Section 6 we survey the interactions of various forms of radiation with matter.

5. THE QUANTUM THEORY OF RADIATION: CONCEPTS

Our discussion of electromagnetic radiation has thus far been in terms of electromagnetic waves produced by accelerating charges. Maxwell's equations predict such waves and give a rather full account of their properties. In the early 1900s and

culminating in the 1920s and 1930s, a more complete theory of radiation was developed incorporating particlelike properties of radiation as well as wavelike properties. These additional properties cannot be accounted for by classical physics, but have their basis in *quantum mechanics*, our best theory of the microscopic world.

Our understanding of radiation is now based on a picture in which there is a fundamental quantum of radiation, known as the *photon*. Introduced by Einstein, the photon has both particlelike and wavelike properties. Photons carry discrete amounts of energy and momentum that can be localized in space, just like a particle. The energy of a photon is related to its frequency, a wavelike property, by

$$E = hf, \tag{19.10}$$

where h is a new fundamental constant of nature known as Planck's constant, and has the value $h = 6.63 \times 10^{-34}\ \text{J} \cdot \text{s}$. Because $f = c/\lambda$, we can also write Equation (19.10) as $E = hc/\lambda$. The theory of *special relativity* (which we study briefly in Chapter 24) shows us that the energy and momentum of a photon are related as $E = pc$. From this we can conclude that the photon's momentum depends only on its wavelength, another wavelike property, as

$$p = \frac{h}{\lambda.} \tag{19.11}$$

You should definitely be surprised that photons, with no mass, can have momentum, a property that up until now we have associated with a particle with mass. We show in Chapter 24 that these two fundamental equations for the energy and momentum of a photon can explain a number of phenomena that cannot be explained by a theory based solely on classical electromagnetic waves.

How are we to picture radiation having both wave and particle properties? We are accustomed to thinking of waves as extended disturbances and of particles as "pointlike" objects. Up until now these have been mutually exclusive concepts. Particles do not have wave properties and waves do not have particle properties. Quantum mechanics turns those ideas upside down as we show. For now, it is sufficient to qualitatively introduce the notion of a *wave packet*. Figure 19.23 shows a schematic representation of a wave packet, a wave that has a limited extent in space. Although not a pure frequency, a wave packet can change its spatial dimension in response to interactions with the external world. The greater the spatial extent, the closer the frequency content is to a pure single frequency (the limit being a perfect sine wave of infinite extent). In this way a single photon can behave more like an extended wave or like a particle depending on its spatial extent.

The intensity of a classical wave (being the energy per unit time per unit area and proportional to the square of the amplitude of the wave) does not correspond to a property of a single photon, but rather to the number of photons per second, each carrying a particular energy given by Equation (19.10). A more intense beam of light of a single color contains more photons per second traveling in the beam. For example, in a 1 W beam of laser light at a visible green wavelength of 514.5 nm there are many, many photons per second.

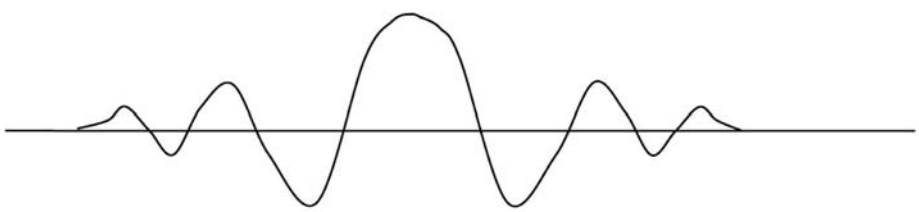


FIGURE 19.23 *A wave packet, localized in space but able to change its spatial extent on interaction with its environment.*

Example 19.4 Calculate the number of photons per second in a 1 W beam from an argon ion laser with a wavelength of 514.5 nm.

Solution: Since each green photon carries an energy given by Equation (19.10) as $E_{photon} = hf = hc/\lambda = (6.6 \times 10^{-34})(3 \times 10^{8})/(514.5 \times 10^{-9}) = 3.8 \times 10^{-19}$ J, the number of photons in the 1 W beam is given by

$$N = \frac{1 \text{ J/s}}{3.8 \times 10^{-19} \text{ J/photon}} = 2.6 \times 10^{18} \text{ photons/s},$$

a very large number indeed.

Because there are so many photons at the usual ambient levels of light, we are not usually aware of the discrete nature of light or of the interactions of a single photon.

Processes that rely on a single photon to have sufficient energy to cause an event to occur will be sensitive to the frequency of the radiation and not to the intensity of the beam. For example, in the photoelectric effect, a process involved in the photodetection of light and discussed in Chapter 24, a minimum energy threshold must be exceeded in order for the detection process (absorption of the photon with the ejection of an electron) to occur. A high-intensity beam of photons with subthreshold energies (huge numbers of photons, each with not enough energy) will not cause the emission of any electrons, whereas another beam of higher-frequency photons but with very low intensity may allow the detection of single photons (Figure 19.24).

We have stressed the common features of electromagnetic radiation up to this point. What most distinguishes the different forms of photons, aside from their method of production, is their interaction with matter. In the next section we discuss the large variety of such interactions and some of their consequences.

6. THE INTERACTION OF RADIATION WITH MATTER; A PRIMER ON SPECTROSCOPY

When electromagnetic radiation is incident on matter some fraction of the photons will usually be transmitted, passing through a sufficiently thin piece of matter without interacting, and the rest will be absorbed and will interact with the material. The particular type of interaction will depend on the energy of the photon and the structure of the matter. We have already discussed some of the interactions of radio and microwave radiation in connection with NMR and ESR. In this section we focus on photons of higher energies, especially the infrared, visible, and ultraviolet portions of the EM spectrum.

The energy level diagram is the crucial indicator of the type of possible interactions with radiation. Figure 19.25 shows a schematic diagram of typical energy levels with closely spaced rotational and vibrational energy levels, with energy differences of 0.01–0.1 eV. These correspond to overall rotational motions of molecules or to the relative positional vibrations of atoms in a molecule and are the lowest energy transitions other than nuclear or electron spin flip energy differences. Larger energy level spacings, with energy differences of about 1 eV, are due to the various electronic states (related to the valence electron's mean distance from the nucleus), and the even larger binding energies of the inner electrons.

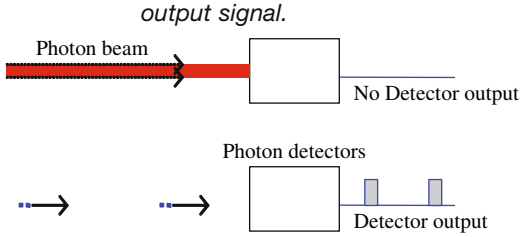


FIGURE 19.24 *An intense photon beam with* red *photons, all with energy too low to be detected, produces no output from a photon detector, whereas individual more energetic* blue *photons each can be detected and produce an output signal.*

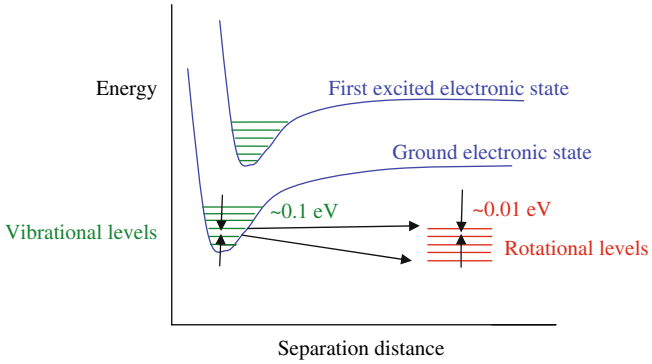


FIGURE 19.25 *Typical energy level diagram of an atom or molecule showing two electronic states and associated vibrational and rotational energy levels.*

Transitions of electrons from the lower "ground state" to higher energy levels, with a change in energy equal to ΔE, can occur upon the absorption of a photon with an energy, frequency, and wavelength such that

$$E_{\text{photon}} = hf = \frac{hc}{\lambda} = \Delta E. \tag{19.12}$$

Table 19.1 shows the general correspondence between the type of EM radiation and the atomic or molecular transitions produced. In general, the shorter the photon wavelength, the more energy imparted to the atom or molecule that absorbs the photon.

Table 19.1 *Types of Atomic or Molecular Transitions Produced by EM Radiation*

EM Radiation (Increasing Energy)	*Energy Level Transitions*
Microwave	Rotational
Infrared	Molecular vibrational
Visible and Ultraviolet	Valence electronic
X-ray	Inner electronic

Two general types of interactions can be distinguished based on whether ionizing or nonionizing radiation is used. Ionizing radiation consists of x-rays or gamma rays and results in the ejection of one or more electrons; we study such processes in Chapter 26. Nonionizing radiation results in a large variety of possible interactions and there is a correspondingly large number of spectroscopic techniques used to probe the radiation–matter interactions in order to learn something about the structure of the sample.

Let's imagine that we perform a spectroscopy experiment in which a beam of monochromatic (literally, single color, but this term is used to mean a uniform frequency or wavelength for nonvisible light where we don't use the concept of color) photons is directed on a biomolecular sample to study the nature of the absorbed or reemitted radiation. Figure 19.26 shows a typical experimental setup for spectroscopy. Although the wavelength of the incident light is varied, the appropriate detector, monitoring either the transmitted or scattered light, records a *spectrum* showing the variation in the measured parameter as a function of incident wavelength. Depending on the type of interaction, we discuss the information one can obtain from such measurements. Our presentation is by no means a complete summary of the various types of spectroscopy.

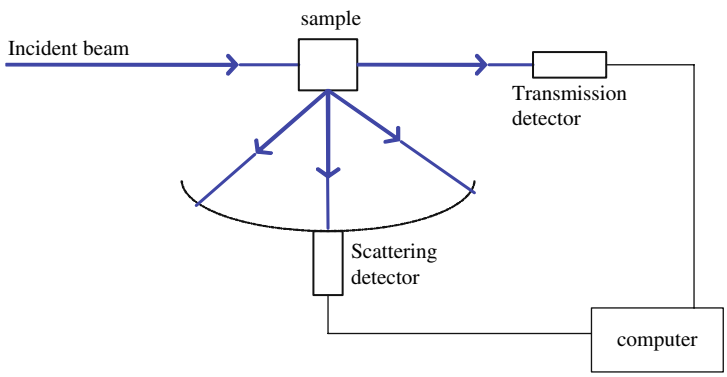


FIGURE 19.26 *A scattering experiment in which both the transmitted and scattered (typically at 90°) EM radiation is detected. In some techniques the incident wavelength of light is continuously changed and the detected signal is recorded as a function of wavelength to produce a spectrum, whereas in others the incident wavelength is fixed and the detected signal is analyzed for wavelength content to produce a spectrum.*

With infrared, visible, or ultraviolet light incident on a sample, some small fraction of the incident photons will be absorbed and can be detected using *absorption spectroscopy*. Most of the absorbed energy that has caused various transitions to occur ends up heating the solution via collisions of the molecules. If the incident intensity on the sample is I_0 then the intensity detected after traveling a distance x through the sample with molar concentration c of absorbing molecules will be

$$I = I_0\, e^{-\varepsilon x c}, \tag{19.13}$$

where ε is the molar extinction coefficient, dependent on the wavelength of the incident EM wave. When rewritten taking the logarithm using base 10 as $\log(I/I_0) = -\varepsilon x c \log(e) = -\varepsilon x c/2.3$, we can introduce the absorbance A, from the Beer–Lambert law

$$A = \varepsilon x c/2.3 = \log(I_0/I). \tag{19.14}$$

When the distance x is taken to be 1 cm, the standard path length of special optical cells used in spectroscopy, then A is commonly called the optical density. An optical density of 2 means that only $10^{-2} = 0.01$ or 1% of the light will be transmitted. Remember that ε, and thus A, depends on the wavelength. At the wavelength corresponding to a maximum absorbance, A is known as the extinction coefficient, and, once known, can be routinely used to determine the concentration of a solution of absorbing molecules after a measurement of the absorbance.

Example 19.5 A 1:10 dilution of a DNA solution is pipetted into a 1 cm path length quartz (non-uv absorbing) cuvette which is then put in an absorption spectrometer to determine its optical density at 260 nm. Measurements find that the absorbance is 0.24 OD. If the extinction coefficient is known to be 6600 $M^{-1}\,cm^{-1}$, what is the molar concentration of the original DNA sample and what fraction of the incident light would be transmitted through a cuvette with the original DNA solution in it?

Solution: According to Equation (19.14) the molar concentration c is given by $c = 2.3A/\varepsilon x = (2.3)(0.24)/(6600)(1)\ M = 8.4\ \mu M$. Then because the measured sample was a 1:10 dilution, the original concentration of the sample was 84 μM. If the original DNA solution had been used directly the measured optical density would be, barring any nonlinear effects, 2.4 OD. Using Equation (19.14), we find the ratio $I_o/I = 10^{2.4} = 250$, so that only 0.4% of the intensity is transmitted.

In *infrared spectroscopy*, the infrared photon energies are just sufficient to excite vibrational excitations of various covalently attached atoms. In general, IR radiation will heat such samples because the internal energy is increased by the absorption of these photons. Imagining that the bonds correspond to springs (see Chapter 4), when the photon frequency corresponds to the spring resonant frequency there will be a large increase in the absorption of photons. Because the "effective spring constant" of a bond is specific to the type of atoms bound, IR spectra can be used to "fingerprint" the sample molecules. Methyl (C–H), carbonyl (C = O), and amide (N–H) bonds, for example, each have characteristic absorption energies that depend also on the local environment of the atoms. Because of the large number of similar absorption energies in large macromolecules such as DNA, IR spectra tend to be broad superpositions of many peaks. One major difficulty is that water, the universal solvent in biology, is a strong absorber of IR radiation and, because it is present in

relatively enormous quantity, masks the absorption peaks due to other molecules. The use of D_2O or other methods have overcome this problem. Much of the IR work with larger macromolecules compares two otherwise identical samples when one variable of the environment, for example, the pH, has been changed. The information obtained from such difference measurements can be used to learn about conformational changes that have occurred under these conditions.

Ultraviolet-visible (uv-vis) absorption spectroscopy probes valence electron excitations. It is precisely these interactions that determine the "chemistry" of the material in that these valence electrons make up the chemical bonds between atoms. Photon energies are sufficient to excite the stronger double and triple bonds and the strongest contributors to the absorption are ring-structures commonly found in biomolecules such as the amino acids tryptophan and tyrosine or the bases of nucleic acids (see Figure 19.27). This technique is routinely used in biological research to measure the concentration of molecules because the absorption is usually proportional to concentration. Measurements are very sensitive to the overall conformation of molecules and can be used to monitor such changes as the melting of DNA (Figure 19.28). Again, difference measurements are commonly used to study the specific effects of a particular perturbation on the sample.

Of the absorbed visible photons, some small fraction will be re-emitted (scattered) by the molecules with nearly unchanged photon energy (*elastic scattering*), but redirected spatially. If the scattering molecules of molecular weight M are small compared to the wavelength of light, then the scattering will be uniform in all directions (isotropic). In this case, the intensity of the scattered light will depend on the wavelength of the incident light λ in a characteristic and very strong way (inverse fourth power of λ)

$$I_{\text{scatter}} \propto \frac{I_{\text{incident}}\, Mc}{\lambda^4}, \tag{19.15}$$

where c is the mass concentration (kg/m^3) of the scatterers. Light-scattering measurements, usually performed using monochromatic incident light, can be used to determine the molecular weight of the scatterers. For larger molecules, the scattering is no longer isotropic, and the spatial dependence of the scattered intensity can be used to give information on the size and shape of such molecules.

Lord Rayleigh, in the 1870s, first discovered the strong λ^{-4} dependence of scattered light and was able to answer the age-old fundamental question: Why is the sky

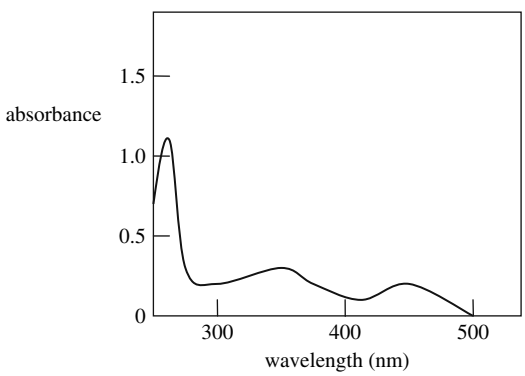


FIGURE 19.27 *Example uv-vis spectrum from small ringlike molecules showing broad characteristic peaks.*

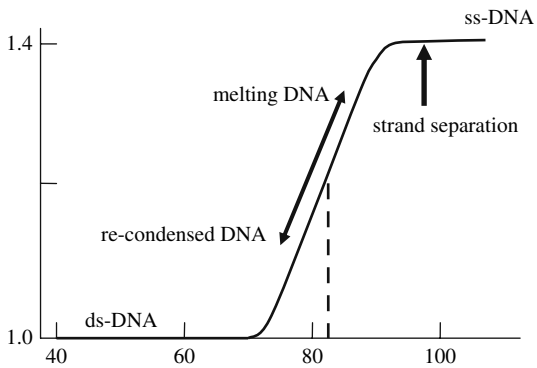


FIGURE 19.28 *An absorbance melting profile of DNA showing likely conformations in each region. The absorbance maximum at $\lambda = 260$ nm is monitored as the temperature of the DNA sample is changed. As the double-stranded DNA is heated it opens up in definite stages before irreversible strand separation occurs. Prior to this the DNA will spontaneously reassemble to a completely functional state when cooled under controlled conditions.*

blue? The answer lies in Equation (19.15) and the fact that when we look up at the sky, the light that we see is sunlight that has been scattered from small gas and dust particles in the atmosphere. Of all the visible colors that our eyes are able to see, shades of blue have the shortest wavelength and, according to Equation (19.15), are therefore scattered with greater intensities. Hence, the sky appears blue. The same argument also explains the brilliant colors of a sunrise or sunset. In those cases we are looking toward the sun and the blue light is predominantly scattered out of the sunlight headed toward our eyes, leaving reds and oranges to reach our eyes directly (Figure 19.29).

Some still smaller fraction of the absorbed visible light will be re-emitted as photons with a wavelength and energy different from the incident light. After a molecule absorbs a visible photon, thereby exciting its valence electron to a higher energy state, often the excited electron will lose some energy to heat via collisions with the solvent in a very short time (10^{-12} s). Subsequent emission of a photon (typically within a time of 10^{-9} s) requires, by conservation of energy, the photon to have a lower energy than the incident photon, and therefore a longer wavelength (recall that $E_{\text{photon}} = hc/\lambda$). Thus, for example, if a solution of macromolecules has blue light shining on it, some small fraction of the light will emerge, let's say, red. This process is known as *fluorescence* and although the fluorescent intensity is very small compared to the incident intensity it can be detected quite easily because of its different color. Fluorescence also occurs when ultraviolet light is absorbed and visible fluorescent photons are emitted. The detergent industry even adds fluorescent "brighteners" in order to enhance the appearance of clothing from extra visible light emitted from uv fluorescence.

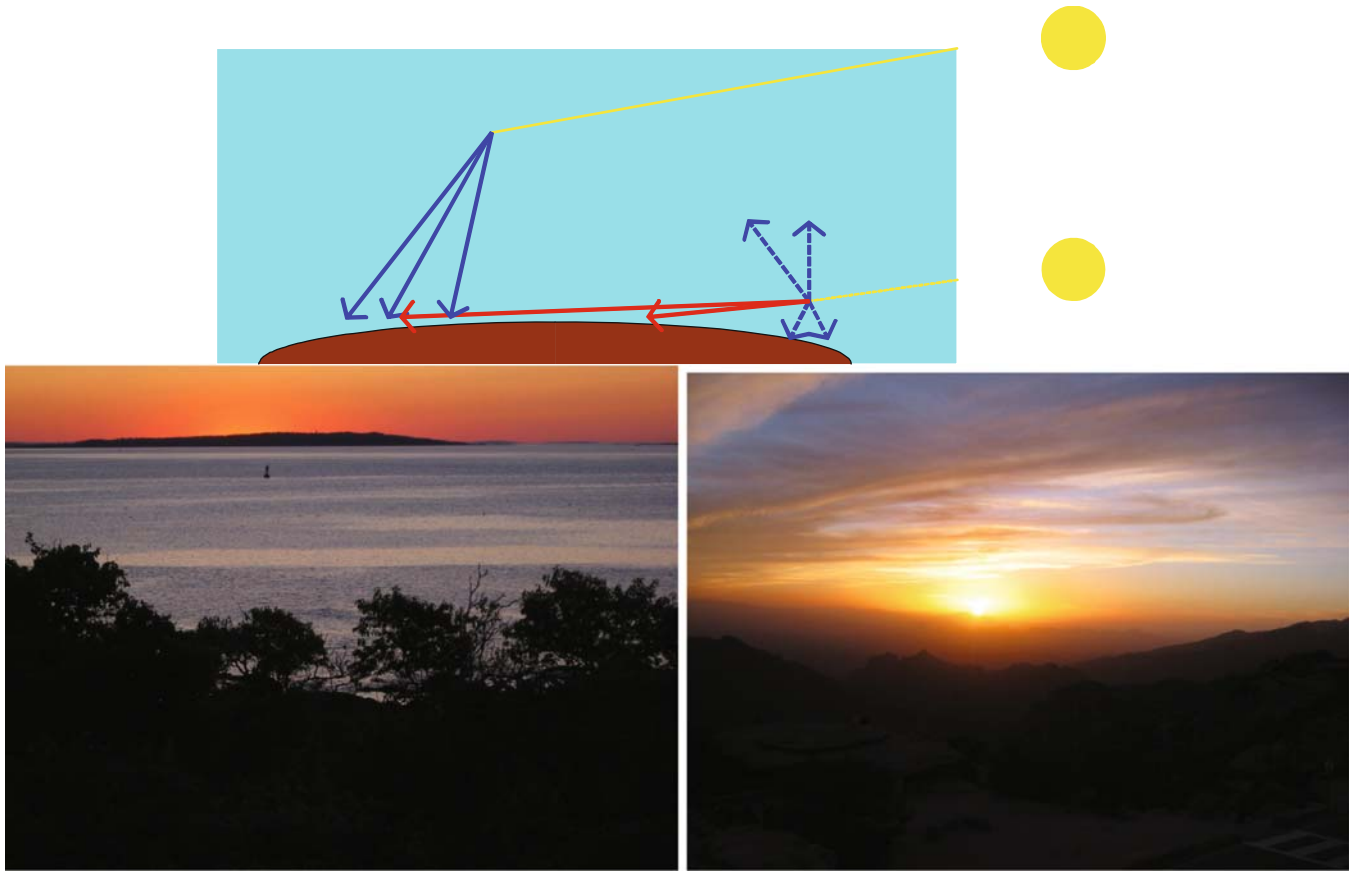


FIGURE 19.29 (top) *The shorter wavelengths* (blues) *of sunlight are preferentially scattered by the upper atmosphere so that the sky will appear* blue *to an observer looking up. When the sun is on the horizon, and the* blue *light is scattered out, the remaining* reds *and* oranges *give the wonderful colors of sunsets and sunrises.* (bottom) *Sunrise over the coast in Maine and sunset over Tuscon, Arizona.*

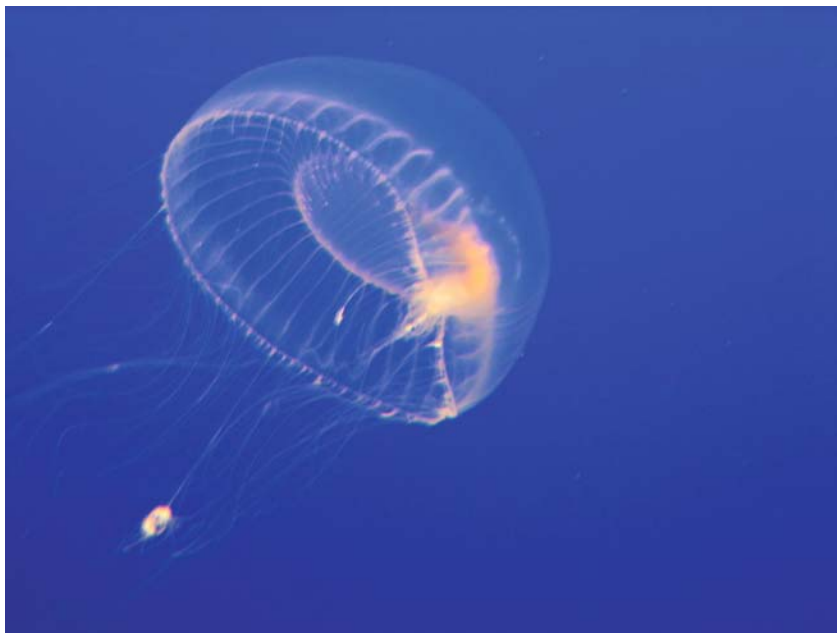


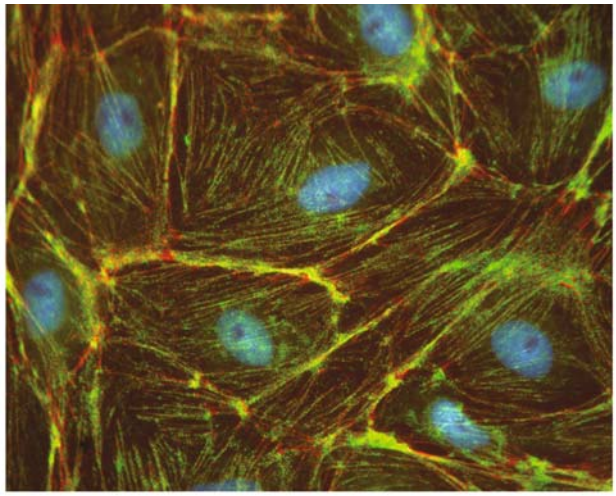
FIGURE 19.30 *A lovely phosphorescent jellyfish at the Monterey Aquarium.*

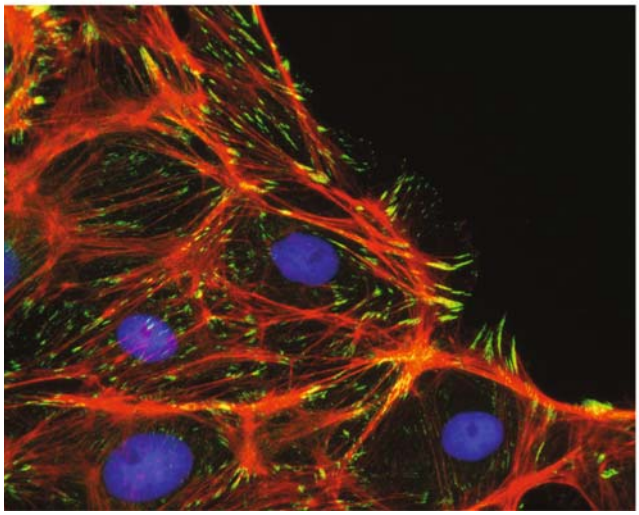
In some few types of molecular systems, the excited electron finds itself in a very long-lived state as far as molecular times go, roughly 10^{-3} s. After this long time a photon is emitted in the process of *phosphorescence*. This is the way in which fireflies generate their mysterious light using the chemical compound luciferase. Roughly half of all types of jellyfish also emit phosphorescence, also known as bioluminescence (Figure 19.30).

Fluorescence spectroscopy can use "intrinsic" fluorescent groups, known as chromophores, if they are present (in proteins tryptophan is a good chromophore) or "extrinsic" fluorescent molecules, or fluors, if attached to the molecule of interest. In either case, fluorescent light is usually detected at a 90° angle to the incident beam direction using some sort of color filter to only detect the fluorescent photons. The quantum yield, defined as the ratio of the number of fluorescent photons emitted to the total number of photons absorbed, is usually very small and quite sensitive to the local environment of the chromophore, including the pH, temperature, neighboring chemical groups, and concentration. Therefore, the fluorescence intensity can be a measure of the local conformation of the biomolecule and measurements can often be used to detect the binding of ligands, small molecules with specific attachment sites, or the polymerization of monomer proteins, such as actin, to form long threadlike filaments.

Aside from its use in spectroscopy, fluorescence is also quite useful in imaging molecules in microscopy. By labeling specific molecules with a fluorescent dye and modifying a microscope with a color filter device, striking images of the arrangement of molecules otherwise too small or dilute to see are possible using *fluorescence microscopy* (Figure 19.31). We return to this topic in the next chapter in a more detailed discussion of microscopy.

Often spectroscopic techniques can study not only the intensity of the scattered or absorbed radiation, but also its polarization and time-dependence. Polarization information can be quite useful in learning about the rotational motion of macromolecules. In fluorescence polarization measurements, if a brief pulse (~ns duration) of intense vertically polarized light is incident on a sample, then after absorbing a photon, chromophores will be able to rotate before emitting a fluorescent photon. In doing so, the polarization of the emerging fluorescence will have a horizontal component as well as a vertical component. By analyzing the time-dependence of these two independent intensity components in the subsequent fluorescent burst of emission, the rotational timescales of motion of the macromolecules can be determined. Several other important types of polarization measurement techniques are discussed in Chapter 23.





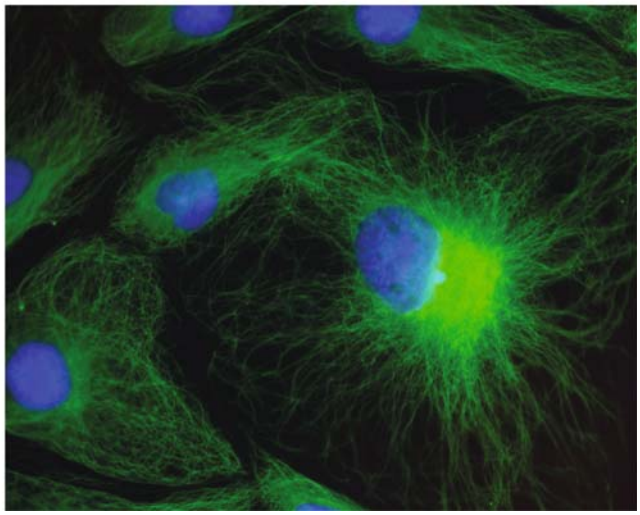


FIGURE 19.31 *Three examples of multiply labeled epithelial cells using fluorescence microscopy.* (top) *Nonmuscle myosin* (green)*, alpha–actinin* (red) *and the nucleus* (blue)*;* (lower left) *with actin* (red)*, vinculin (an adhesion plaque protein;* green*), and the nucleus* (blue)*;* (lower right) *microtubules* (green) *and nucleus* (blue)*.*

CHAPTER SUMMARY

Electromagnetic radiation is produced by accelerating electric charges and consists of coupled traveling E and B waves, perpendicular to each other and to the direction of propagation, and that decrease in amplitude as they travel outward as $1/r$. The waves have the form, in one dimension,

$$E(x,t) = E_{\max} \sin(kx - \omega t),$$
$$B(x,t) = B_{\max} \sin(kx - \omega t), \qquad (19.2)$$

with

$$\frac{E_{\max}}{B_{\max}} = c. \qquad (19.3)$$

The energy density of the EM radiation is given by

$$\frac{PE}{V} = \frac{1}{2}\left(\varepsilon_0 E^2 + \frac{B^2}{\mu_0}\right) = \varepsilon_0 E^2, \quad (19.4, 5)$$

(*Continued*)

where the two terms have equal energy. The Poynting vector is the energy per unit time per unit area and is given by

$$S = c\frac{PE}{V} = c\varepsilon_0 E^2, \qquad (19.6)$$

and points in the direction of propagation of the wave. The average value of S is given by the intensity,

$$I = \frac{1}{2}\varepsilon_0 cE_{\max}^2. \qquad (19.7)$$

Laser tweezers is a technique that traps small transparent objects at the focal point of a laser with a trapping force proportional to the laser beam intensity. Careful calibration can then measure the force acting on the individual object. A wide range of biological applications have been developed using laser tweezers, now capable of studying the forces generated by individual macromolecules.

In EM waves, the direction of the E field is known as the polarization direction. If the E field is confined to a linear direction, the wave is said to be linearly polarized. Random orientation of E in a wave is known as an unpolarized EM wave. The E field can also be made to sweep in a circle or an ellipse as the wave propagates in circularly or elliptically polarized beams. The law of Malus gives the intensity transmitted through an analyzing polarizer in terms of the incident intensity I_0 after passing through a first polarizer with an angle θ between the two polarizers,

$$I_t = I_0\cos^2\theta. \qquad (19.9)$$

Photons are the elementary constituents of electromagnetic radiation and each photon has an energy and a momentum given by

$$E = hf, \qquad (19.10)$$

$$p = \frac{h}{\lambda}. \qquad (19.11)$$

Photons with different frequencies (or wavelengths) constitute electromagnetic radiation from different parts of the electromagnetic spectrum (see Figure 19.22). The study of the interactions of electromagnetic radiation with matter is called spectroscopy. Nonionizing radiation can be absorbed according to the Beer–Lambert law

$$A = \log(I_0/I), \qquad (19.14)$$

where A is the absorbance and I and I_0 are the transmitted and incident intensity, respectively. Absorption can only occur if the incident photon energy matches a possible energy transition in the absorbing atom or molecule. Scattering may also occur, whereby the incident photon energy is absorbed and re-emitted, either at the same frequency (elastic scattering) or at a lower frequency (e.g., fluorescence). In elastic scattering, the scattered intensity varies as the inverse fourth power of the wavelength and this strong dependence on wavelength is responsible for the blue color of the sky and for the bright red/orange colors of sunrises/sunsets.

QUESTIONS

1. Why must the electric and magnetic fields of a spherical wave vary as $1/r$? Fill in the details of the energy argument of Figure 19.2.
2. Compare electromagnetic waves with waves on a string. Give as full an accounting as you can of how they are similar and how they are different.
3. Even though the magnetic field of an electromagnetic wave is a factor of c weaker than the electric field, the energy contained in the magnetic field is the same as that of the electric field. Show how this follows from the definitions of the field energies.
4. What is the distinction between the Poynting vector and the intensity of an electromagnetic wave?
5. Explain in words how a focused laser beam can supply a longitudinal force in the direction of the focal point on a microscopic particle. See Figure 19.9.
6. If a laser beam has a transverse intensity profile decreasing symmetrically from a maximum in the center of the beam, explain in words how the focused beam can supply a transverse force toward the center of the beam on a microscopic particle. See Figure 19.10.
7. Assuming that micron-sized spheres can be attached to biological molecules and these can be then trapped in laser tweezers, suggest several experiments that can be done to study cellular and macromolecular properties.
8. Distinguish between unpolarized and polarized electromagnetic waves. Because all electromagnetic waves are transverse, aren't they all polarized?

9. How does Polaroid film work to polarize light? If unpolarized light is incident on a Polaroid sheet, what fraction of the incident intensity is transmitted?
10. If vertically polarized light passes through a perfect polarizer and no light is transmitted, what can you conclude?
11. Unpolarized light passes through two consecutive polarizers with axes oriented at 45° from each other. What fraction of the incident intensity is transmitted through the second polarizer?
12. Explain why circularly polarized light is equivalent to two linearly polarized electric fields at right angles to each other and 90° out of phase. Describe the phase relations that produce a clockwise or counterclockwise circularly polarized plane wave as viewed from in front of the wave watching it approach. Using your two arms, kept at right angles to each other, devise a way to move your arms to simulate the electric fields of right or left circularly polarized light.
13. The visible spectrum of light ranges from around 400 nm violet to 750 nm red. Which color photon has more energy? More momentum? Does having more momentum mean that those photons travel faster?
14. A pure sine wave has infinite spatial extent and a single frequency. A square pulse has a finite spatial size and is composed of an infinite range of different frequencies. These two examples are limits for a wave packet. Discuss this idea.
15. What type of radiation would you expect to be emitted when a molecule makes a transition between neighboring rotational states with energy spacings of about 0.01 eV? Between neighboring vibrational levels with energy differences of about 0.1 eV? Between electronic states with energy differences of about 1 eV?
16. What fraction of the incident light is transmitted through a sample with an optical density of 1.0? Of 2.0? Of 1.5?
17. Can a fluorescent dye that absorbs strongly in the green appear to be blue? Red? Yellow?
18. Discuss the difference between inelastic and elastic scattering. Can there be absorption with elastic scattering? Can there be fluorescence with elastic scattering?

MULTIPLE CHOICE QUESTIONS

1. In electromagnetic radiation (a) the electric and magnetic fields are both parallel to the direction of propagation, (b) the electric field is perpendicular to and the magnetic field is parallel to the direction of propagation, (c) the electric field is parallel to and the magnetic field is perpendicular to the direction of propagation, (d) the electric and magnetic fields are both perpendicular to the direction of propagation.
2. Which of these is not an example of the "electromagnetic spectrum"? (a) X-rays used by your dentist. (b) γ-rays used to prevent food spoilage. (c) Microwaves used to boil water. (d) Ultrasonic waves used to image a fetus.
3. A TV wave has a wavelength of about 1 m. The frequency of such a wave is (a) 1 Hz, (b) 3 Hz, (c) 300 MHz, (d) 3×10^{15} Hz.
4. The wavelength of violet light is about (a) 400 m, (b) 400 cm, (c) 400 μm, (d) 400 nm.
5. Electromagnetic waves emitted from the wiring in your house due to AC currents have a wavelength of about 5 times (a) 10^{-10} m, (b) 10^{-2} m, (c) 10^{3} m, (d) 10^{6} m.
6. The magnetic field inside a solenoid is proportional to the current flowing through the solenoid. Suppose the current through a solenoid is doubled, by how much is the energy associated with the magnetic field in the solenoid changed? (a) It remains the same. (b) It is doubled. (c) It is increased by a factor of four. (d) It is halved.
7. One light source is four times more intense than a second light source. The maximum electric field in the light from the first source is (a) four times greater than that from the second, (b) two times greater than that from the second, (c) one half as great as that from the second, (d) one quarter as great as that from the second.
8. The intensity of an extremely bright laser is 10^{7} W/m^2, about 10,000 times brighter than sunlight. The average electric field in sunlight is roughly 10^{3} V/m. The average electric field in the bright laser is about (a) 10^{7} V/m, (b) 10^{5} V/m, (c) 10^{3} V/m, (d) 10 V/m.
9. The sun emits electromagnetic radiation uniformly in all directions. Given the distance from the sun to the Earth of about 150×10^{6} km, and the mean radius of the Earth of about 6400 km, the fraction of the sun's radiation that falls on the Earth is (a) $(6400/150 \times 10^{6})$, (b) $(6400/150 \times 10^{6})^2$, (c) $(6400/300 \times 10^{6})^2$, (d) $(1/2)(6400/300 \times 10^{6})^2$.
10. The central 1 cm diameter portion of a 10 W laser beam with a diameter of 2 cm is focused down to a 100 μm diameter spot size. The intensity at the focal point is (a) 1.3×10^{9}, (b) 6.4×10^{8}, (c) 3.2×10^{8}, (d) 8.0×10^{7} W/m^2. (Assume the laser beam has a uniform intensity across its cross-sectional area.)
11. A small plastic sphere is initially located below and to the left of the center of a focused laser beam. The force on this sphere is directed (a) down and to the right, (b) up and to the left, (c) up and to the right, (d) down and to the left.
12. A typical force capable of being exerted by a focused laser beam in a laser tweezers experiment is (a) 10^{-2} N, (b) 10^{-7} N, (c) 10^{-12} N, (d) 10^{-17} N.
13. A plane wave of light propagating along the positive x-axis is linearly polarized along the z-axis. A polarizing sheet parallel to the $y - z$ plane is at $x = 1$ m. The transmission (or polarizing) axis of the sheet is parallel to the y-axis. A second polarizing sheet is placed parallel to the $y - z$ plane at $x = 0.5$ m. A light detector is placed at $x = 1.1$ m. The detector

will receive (a) maximum light when the transmission axis of the second sheet makes a 45° angle with respect to the *y*-axis, (b) maximum light when the transmission axis of the second sheet is parallel to the *z*-axis, (c) maximum light when the transmission axis of the second sheet is parallel to the *y*-axis, (d) no light no matter what is the orientation of the transmission axis of the second sheet.

14. If a laser beam with intensity I_0 is incident on a polarizer and the transmitted intensity is determined to be also approximately I_0, we can conclude that (a) the initial beam was unpolarized, (b) the initial beam was polarized perpendicular to the polarizer's transmission axis, (c) the initial beam was circularly polarized, (d) the initial beam was polarized parallel to the polarizer's transmission axis.

15. If a laser beam with intensity I_0 is incident on a polarizer and the transmitted intensity is determined to be approximately $I_0/2$, we can conclude that the initial beam (a) must have been unpolarized, (b) may have been unpolarized, circularly polarized, or linearly polarized at 45° to the transmission axis of the polarizer, (c) must have been circularly polarized, (d) must have been linearly polarized at 45° to the transmission axis of the polarizer.

16. The number of photons per second emitted by a 1000 W, 100 MHz FM radio station is about (a) 10^3, (b) 10^8, (c) 10^{11}, (d) 10^{28}.

17. A 440 nm photon has an energy of (a) 4.5×10^{-19} J, (b) 1.5×10^{-27} J, (c) 4.5×10^{-17} J, (d) 3.8×10^{-19} J.

18. A sample with an optical density of 2.5 transmits (a) 8, (b) 300, (c) 3, (d) 0.3 percent of the incident light.

19. The midday sky appears blue because (a) our eyes see blue better than other colors making up white light; (b) molecules in the sky reflect back light from the oceans which are blue; (c) molecules in the sky re-emit blue light more than other colors visible to our eyes; (d) hydrogen molecules in the sky have a strong absorption peak in the blue.

PROBLEMS

1. If a plane electromagnetic wave has a maximum magnetic field amplitude of 2×10^{-7} T, find the peak value of the electric field of the wave.

2. A plane electromagnetic wave is traveling along the *x*-axis. If the electric field of the wave has a maximum value of 2×10^{-4} N/C and lies along the *y*-axis, find the wave's maximum magnetic field and its direction.

3. Given that the peak value of the magnetic field of an electromagnetic plane wave is 5×10^{-7} T, find the intensity of the wave.

4. If the maximum trapping force from a laser tweezers on a 1 μm radius spherical particle in water at 20°C is 10^{-12} N, find the minimum flow velocity of the water, in mm/s, that will just free the particle from the trap.

5. If in the previous problem the calculated flow velocity of the water is doubled, by what factor must the intensity of the laser beam be increased to just maintain the trap?

6. A vertically polarized beam of light passes through a sheet of Polaroid with its transmission axis at a 30° angle to the vertical. What fraction of the incident intensity emerges from the Polaroid? If this beam is then incident on a second sheet of Polaroid with its transmission axis along the vertical, what fraction of the beam's original intensity is transmitted?

7. Unpolarized light of intensity I_0 passes through a Polaroid with its transmission axis vertically oriented. What intensity emerges? If the transmitted light passes through a second Polaroid sheet with its transmission axis 60° to the vertical what fraction of the original incident light intensity I_0 emerges?

8. Three polarizers arranged in series are each oriented at 30° from the previous one. If an unpolarized light beam travels through the three polarizers and emerges with an intensity of 0.2 W/m^2, what was the intensity of the beam incident on the first polarizer?

9. A circularly polarized light beam is incident on a vertically oriented polarizer. If the incident beam has an intensity I_0, describe the transmitted beam intensity and polarization.

10. Two polarizers are in series with one another with an angle of 45° between their axes. If a circularly polarized beam with an intensity of 0.8 W/m^2 is incident on the first polarizer, oriented vertically, describe the intensity and polarization of the beam transmitted through the second polarizer. What happens if the two polarizers are interchanged keeping their axes' orientation fixed? Does the transmitted intensity change? Does the polarization direction of the transmitted beam change?

11. Suppose a point source of light generates 60 W. Four meters away there is a light detector that is 75% efficient and has a detector area of 10 cm^2 oriented with its normal directed at the point source. What power will the detector record?

12. Suppose a 50 kW radio station, operating at a frequency of 106.5 MHz, emits EM waves uniformly in all directions.
(a) What is the wavelength of the radio waves?
(b) How much energy per second crosses a 1.0 m^2 area 100 m from the transmitting antenna?
(c) What is the maximum value of the electric field at this point, assuming the station is operating at full power?
(d) What voltage is induced in a 1.0 m long vertical car antenna at this distance?

13. Our nearest star (*Proxima Centari*) is 4.2 light-years away (1 light-year equals the distance light travels in a year). How far away is *Proxima Centari* in meters?

14. The Andromeda Galaxy (our nearest neighbor galaxy) is approximately 2 million light-years away (see the previous problem). How far away is Andromeda in astronomical units? 1 AU = 1.5×10^8 km.

15. The human eye is most sensitive to light having a wavelength of 5.50×10^{-7} m, which is in the green–yellow region of the visible electromagnetic spectrum. What is the frequency of this light?

16. Suppose that a laser pointer is rated at 3 mW. If the pointer produces a 2 mm diameter spot on a screen, what is the radiation pressure exerted on the screen?

17. A laser pulse lasting 10^{-9} s from a Neodymium–YAG laser has an energy of 5 J. If the wavelength of the laser is 1.06 μm, how many photons are in the pulse? What is the power of the laser pulse? If the pulses are repeated at a rate of 10 Hz (10 pulses/s), what is the average power output of the laser?

18. Calculate the frequency range and photon energies (in eV) of visible light if the wavelength range is 400–750 nm.

19. If the wave packet of a photon contains frequencies in the range $6.8–7.1 \times 10^{14}$ Hz, what is the average wavelength and ± wavelength uncertainty spread of the photon?

20. Suppose an atom has two energy levels at −3.40 eV and −1.51 eV. If an electron makes a transition from the upper to the lower of these levels, what is the wavelength of the emitted photon?

21. A 1 ns laser pulse from a neodymium–YAG laser, at 532 nm, contains 5 J.
(a) Find the energy and momentum of each photon.
(b) How many photons are in the 1 ns pulse of laser light?
(c) If all the photons are completely absorbed by an object, what is the force exerted on the object by the laser light?

22. The laser light pulses from the previous problem are passed through a device known as a frequency doubler that basically combines two pulses into a single one with twice the frequency.
(a) Find the wavelength, energy, and momentum of the frequency doubled pulses.
(b) Assuming 100% efficiency, how many of these photons are in the 1 ns pulse?
(c) If all of these photons are absorbed by the same object as in the previous problem, compare the force exerted on the object by the frequency doubled pulse compared to the original pulse.

23. A solution of a purified protein is put in a 1 cm path length quartz optical cell and into a uv spectrophotometer. The optical density measured at a 260 nm wavelength is 0.05. If the molar extinction coefficient of this protein is 349 M/cm at 260 nm find the molar concentration of the protein and the % of uv light transmitted by the solution.

24. If the incident intensity on a 0.01 M protein solution in a 1 cm optical cell is reduced in the transmitted beam by a factor of 1500, what is the molar extinction coefficient of the protein at the measuring wavelength?

25. A detector measures the elastic scattering from a gas in a sealed glass cell. If the incident light intensity is halved and the wavelength of the light used is reduced by 20%, find the percent change in the scattered intensity. Does the amount of scattered light increase or decrease with the 50% reduction in incident intensity at the new wavelength?

20 Geometrical Optics

In our ordinary experiences, light seems to behave as if it travels in straight lines until it strikes an object. Shadows cast by objects, the beam of light from a flashlight or a car's headlights, the bright rays of sunlight through a clearing in the clouds, and the pencil-like laser light beams used in light shows all tell us that this is true (Figure 20.1). On the other hand, sound does not cast a "shadow;" you can be heard around sharp corners without a straight path to the listener. Radio and television waves also can be received without straight line paths to the radio or TV station antenna and as we have seen these are also forms of electromagnetic radiation just as is light. What distinguishes light in its ability to travel in straight lines is its small wavelength of about 5×10^{-7} m, much smaller than any characteristic dimension of a typical object in its path. Sound, as well as radio and television EM waves, have wavelengths with macroscopic dimensions.

In this chapter we consider the properties of light that can be understood based on geometrical optics, in which light is treated as traveling in a straight line path in a uniform medium. We show that only at a boundary between two media with different optical properties (defined below) will light be deflected from its straight trajectory. Spherical mirrors are discussed as a method for imaging objects. Fiber optics is discussed as a "device" used to steer light. In the next chapter further applications of geometrical optics are discussed including lenses, the structure and function of the eye as well as the use of eyeglasses to correct vision problems, and a discussion of magnifying glasses and the compound microscope used routinely in biology.

1. OPTICAL PROPERTIES OF MATTER

In our discussions of electricity and magnetism we have seen that there are two fundamental parameters that describe the electric and magnetic interactions in space, the permittivity ε_0 and permeability μ_0 of the vacuum. These two parameters together determine the speed of light in vacuum. When light interacts with a material medium, the electromagnetic fields interact with charges and atomic currents and the fundamental parameters are modified by the presence of the medium. The effects of the material medium can be taken into account by modifying the two fundamental parameters to values ε and μ, characteristic of the material. Note that the dielectric constant κ introduced in Chapter 15 is just equal to the ratio $\varepsilon/\varepsilon_0$. For isotropic materials these will be constants whose values will determine the speed of light in the medium

$$v = \frac{1}{\sqrt{\varepsilon\mu}}, \tag{20.1}$$

J. Newman, *Physics of the Life Sciences*, DOI: 10.1007/978-0-387-77259-2_20,

FIGURE 20.1 (left) *Geometrical optics through a cloud: sun's rays travel in straight lines.* (right) *Geometric optics by laser light.*

in a relation similar to that defining the speed of light in vacuum c. Because, in general, $\varepsilon \geq \varepsilon_0$ and $\mu \geq \mu_0$ we have that the speed of light in a material medium is always less than c. For most materials μ is very close to μ_0, and it is the permittivity ε that really determines the speed of light. We introduce the *index of refraction* of the medium n to be

$$n = \sqrt{\frac{\varepsilon\mu}{\varepsilon_0\mu_0}} \approx \sqrt{\kappa}, \tag{20.2}$$

so that the speed of light in the medium is given by

$$v = \frac{c}{n}. \tag{20.3}$$

Table 20.1 lists indices of refraction for some common materials that are transparent to visible light.

Table 20.1 *Refractive Indices of Materials*[a]

Material (20°C Unless Specified)	*Refractive Index*
Diamond	2.42
Glass (crown)	1.52
Benzene	1.50
Quartz (fused)	1.46
Water	1.33
Air (1 atm, 0°C)	1.0003

[a]Measured at a wavelength of 589 nm (yellow sodium light).

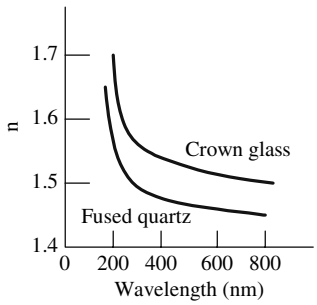


FIGURE 20.2 *Dispersion curves for two different transparent materials.*

Unlike the permittivity and permeability of the vacuum, those of a material medium are dependent on interactions with electromagnetic (EM) radiation and are therefore dependent on the frequency (or wavelength) of the light. This phenomenon is known as *dispersion* and will account for some of the experimental findings discussed below when using white light. White light is a generic term used to describe a broad mixture of light waves of different colors (frequencies or wavelengths) that can be quite different in its intensity mix of colors from different sources. Figure 20.2 shows typical dispersion curves (index of refraction versus wavelength) for two different transparent materials.

As an aside, we briefly mention what occurs when the material is not transparent to EM radiation. Near wavelengths at which EM radiation is absorbed by the material and thus no longer transparent, the index of refraction rises with increasing wavelength in an atypical behavior known as *anomalous dispersion*. This is due to the phenomenon of resonance, where the radiation frequency matches a natural absorption frequency of the material and can be readily absorbed. In our discussion of absorption spectroscopy, we saw that it is this resonance phenomenon that accounts for the absorption. Thus, absorption and dispersion are very strongly coupled together. Figure 20.3 shows that for any material, at whatever frequency of radiation, where there is an absorption peak, there is a corresponding anomalous dispersion. In the rest of our discussion in this chapter we limit ourselves to situations in which absorption of radiation is negligible.

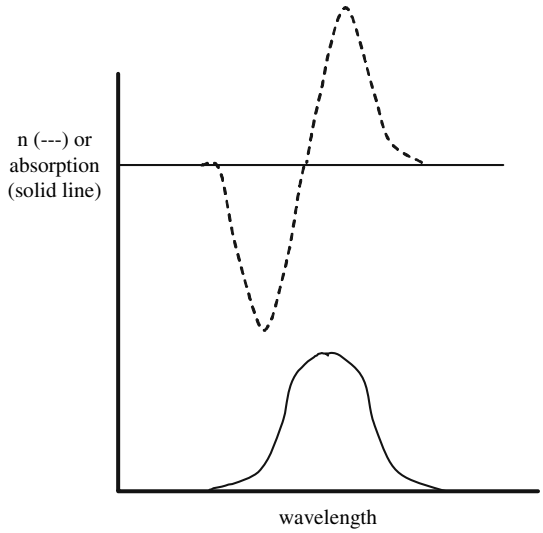


FIGURE 20.3 *The connection between anomalous dispersion* (dotted line) *and absorption* (solid line).

2. LIGHT AT AN INTERFACE

When light strikes the interface between two different optically transparent media it is partially reflected and partially transmitted into the second medium. Because light will travel in straight lines within a uniform material, we can follow its path by tracing rays that are representative of the light beam. For a plane light wave, the wavefronts lie in the transverse plane and the rays are all parallel to the propagation direction. If the light wave is a spherical wave, emanating from a point source, the wavefronts are spherical and although the rays are still perpendicular to the wavefronts, they are diverging as shown in Figure 20.4.

If the interface between the two transparent media is a smooth plane surface, an incident plane wave with parallel rays will undergo *specular reflection* from the surface with the reflected rays remaining parallel. As shown in Figure 20.5, if the incident ray makes an angle of incidence θ_i with the normal, a line drawn perpendicular to the surface, then the reflected ray will leave the surface with an angle of reflection θ_r equal to the angle of incidence and will lie in the plane defined by the incident ray and the normal, known as the plane of incidence. This is known as the *law of reflection* and was already briefly discussed in Chapter 11 in connection with sound waves. It can be written as

$$\theta_r = \theta_i, \tag{20.4}$$

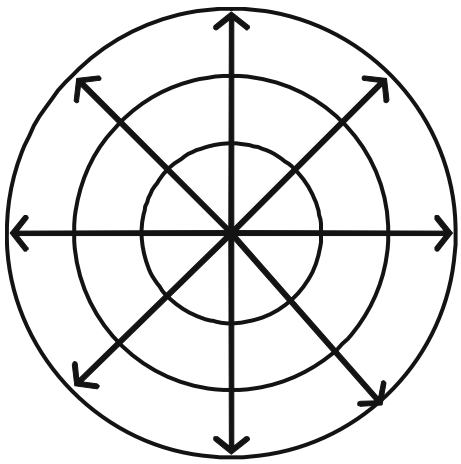


FIGURE 20.4 *Spherical wavefronts diverging from a point source with radially directed rays.*

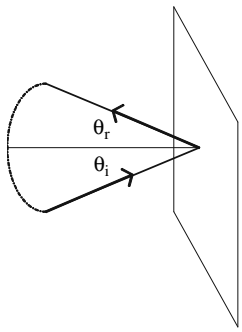


FIGURE 20.5 *The law of reflection at a plane interface.*

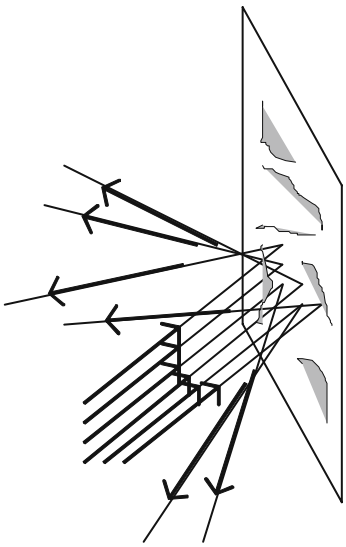


FIGURE 20.6 *A diffuse reflection from a rough surface.*

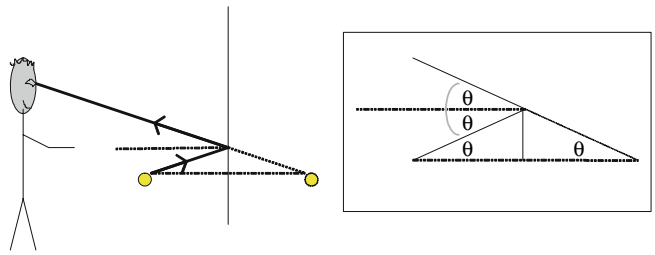


FIGURE 20.7 *Imaging a point source in a plane mirror. The insert shows the geometry that proves the object and virtual image are equal distances from the mirror.*

with the proviso that the reflected ray lies in the plane of incidence. Note that so far we've neglected the transmitted portion of the light; we return to this below.

If the surface is a rough plane, then the reflected light no longer maintains its spatial regularity and is said to undergo *diffuse reflection* (Figure 20.6). Whereas specular reflection occurs from mirrors, windows, or high-gloss surfaces, diffuse reflection occurs from dull unpolished surfaces. The key distinguishing feature of these two types of reflections is the formation of an image only on specular reflection.

Imagine first that there is a point source of light in front of a plane mirror as shown in Figure 20.7. At a particular location of your eye, there will be rays of light that will reflect from the mirror, reaching your eye, and appearing to diverge from a point source image located behind the mirror. This type of image is called a *virtual image* because the light only gives the illusion of emanating from behind the mirror. We soon show that a *real image* is one through which light actually passes and one for which light could be captured on a viewing screen placed at the image. Using a simple geometric argument, shown in the figure and based on the law of reflection, we can determine that the *image distance*, defined as the distance from the image to the mirror surface, is equal to the *object distance*, defined as the distance from the object to the mirror surface. Thus, the image appears to be behind the mirror the same distance as the object actually lies from the mirror.

If the source of light is an extended object, either self-luminous such as a real light bulb or any other object itself reflecting light from its surface, then the same analysis holds point by point and a virtual image of the object will be created behind the mirror. For a plane mirror the image will be equal in size to the object, erect, as far behind the mirror as the object is in front, but will appear left–right reversed as shown in Figure 20.8. Leonardo da Vinci wrote all his scientific notebooks using a left–right reversal as if in code; they can be clearly read by simply imaging them in a mirror (Figure 20.9).

Now let's consider the portion of the light that enters the transparent second medium if the interface is a plane surface. Figure 20.10 shows a set of rays, with their associated wavefront, incident on a plane interface, with only the portion of the light that enters the second medium drawn. Some of the light will also be reflected as already discussed. If we assume that the second medium has a larger index of refraction than the first, so that the speed of light in the second medium is slower than in the first, then the rays will bend toward the normal, as shown in the figure. This is so because during the time it takes for the wavefront at the ray to the right in the figure (point B) to reach the interface (point B'), the wavefront at the ray on the left (point A)

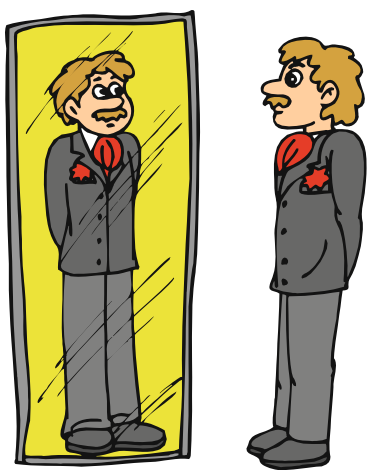


FIGURE 20.8 *A person and his virtual image in a plane mirror. Note that the flower on the man is on his left side, and it is on the right side of the virtual man in the mirror. What's wrong with the perspective of this cartoon?*

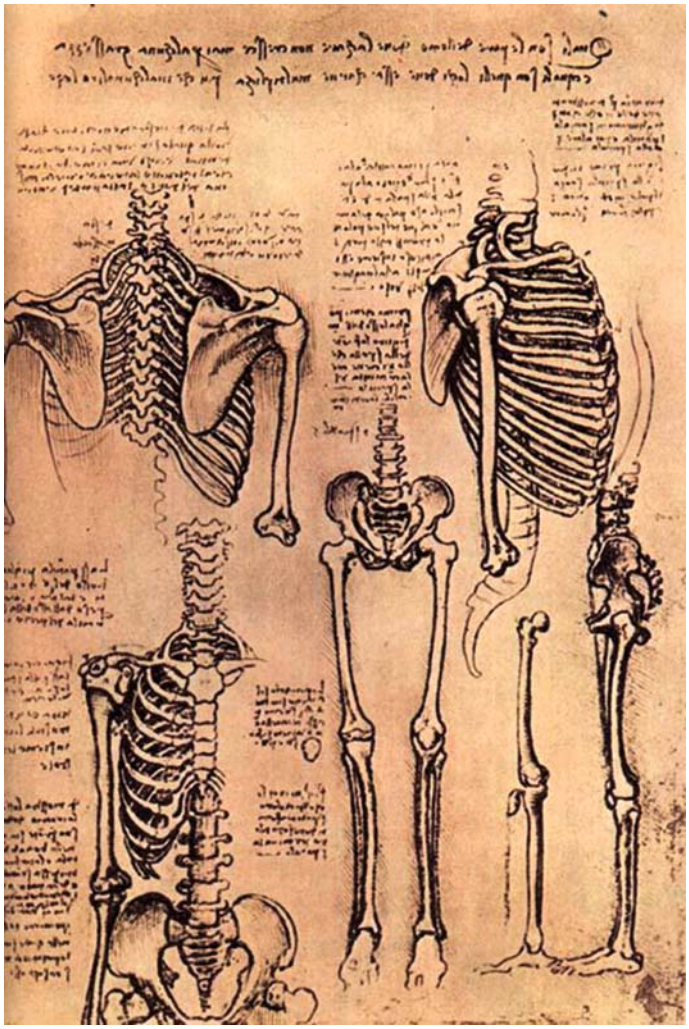


FIGURE 20.9 *Drawing of human skeletons by Leonardo da Vinci. Note the mirror-image writing throughout.*

will travel a shorter distance (to point A′). Because of the slowing of the rays, when the incident light is at any nonzero incidence angle, it must bend or refract.

The angle of refraction (the angle between the refracted ray and the normal to the interface pointing into the second medium) can be determined from *Snell's law* (also simply known as the *law of refraction*)

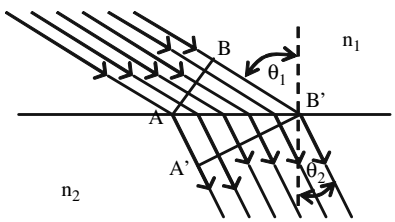


FIGURE 20.10 *Refraction of a plane wave at a planar interface. The second medium has a larger index of refraction, slowing the rays and bending them toward the normal* $-\theta_2 < \theta_1$ *because* $n_2 > n_1$.

A general principle in optics, known as Fermat's principle, can be used to derive the laws of reflection and refraction. Fermat's principle states that in traveling between any two points, light will always take the path that requires the least time. We now use this principle, together with Figure 20.11a and b, to derive these two laws. In both cases, we want light to travel between points A and B in the shortest time. Points A and B are fixed at distances y_A and y_B from the interface and at an x separation of L. We wish to find the point C for which the least time is required for light to go from A to B.

Using the notation of Figure 22.11a, the time for light to travel from A to B is given as

$$t = \frac{s_A + s_B}{v} = \frac{\sqrt{x^2 + y_A^2} + \sqrt{(L-x)^2 + y_B^2}}{v}.$$

To minimize the time, we take dt/dx and set it equal to zero, dropping v because it is a constant,

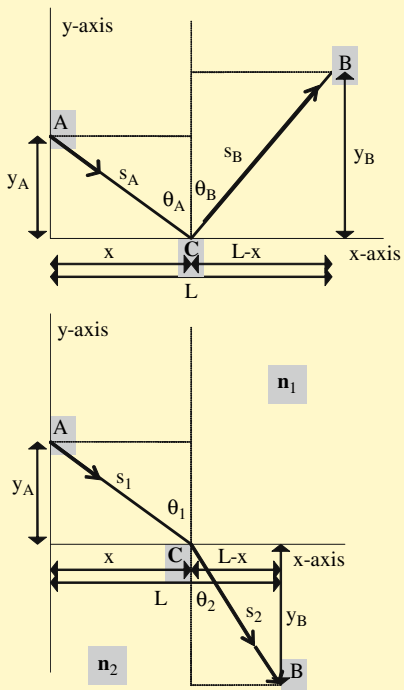


FIGURE 20.11 ***(a)*** (top) *Geometry for proof of the law of reflection* (see box)*;* ***(b)*** (Bottom) *Same for law of refraction.*

(*Continued*)

$$\frac{dt}{dx} = \frac{x}{\sqrt{x^2 + y_A^2}} - \frac{(L - x)}{\sqrt{(L - x)^2 + y_B^2}} = 0$$

Using the definitions of sin θ_A and sin θ_B, this becomes

$$\sin \theta_A = \sin \theta_B \quad \text{or} \quad \theta_A = \theta_B,$$

thus proving the law of reflection.

To prove the law of refraction, we consider Figure 20.11b and write the time for travel between A and B as

$$t = \frac{s_1}{v_1} + \frac{s_2}{v_2}$$

Writing s_1 and s_2 in identical ways as above, but using $v_1 = c/n_1$ and $v_2 = c/n_2$, we proceed in the same way to find

$$\frac{dt}{dx} = \frac{1}{c}\frac{d}{dx}\left(n_1\sqrt{x^2 + y_A^2} + n_2\sqrt{(L - x)^2 + y_B^2}\right) = 0\,,$$

resulting, after performing the derivatives just as above, in the law of refraction,

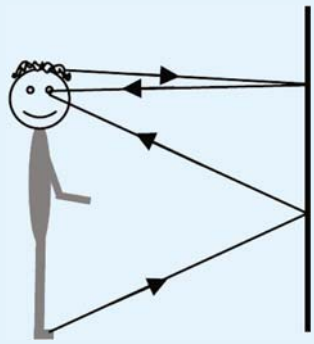
$$n_1 \sin \theta_1 = n_2 \sin \theta_2.$$

$$n_1 \sin \theta_1 = n_2 \sin \theta_2, \tag{20.5}$$

where θ_1 and θ_2 are the angles of incidence and refraction and n_1 and n_2 are the corresponding indices of refraction of the two media. A proof of Snell's law, as well as the law of reflection, is given in the box. As we argued above, Equation (20.5) predicts that if $n_2 > n_1$ then $\theta_2 < \theta_1$, and so the refracted ray will bend toward the normal. If $n_2 < n_1$, so that the speed of light in the second media increases, Snell's law predicts that the refracted light will bend away from the normal.

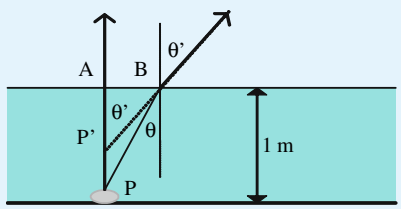
Example 20.1 Find the minimum height of a plane mirror needed for a person of height H to see her entire self in the mirror.

Solution: In order to see your feet in the mirror, a raytracing diagram shows that a ray needs to emanate from your feet and hit the mirror at a height midway between your feet and your eyes, because the angles of incidence and reflection are then equal. Similarly to see the top of your head, the mirror must extend at least half of the distance between the top of your head and your eye level. Adding these distances together, a mirror must be half of your height, and positioned as discussed, for you to be able to see yourself fully, regardless of the distance you stand from the mirror. Try it by masking off a full length mirror.



Example 20.2 How far below the surface of a lake will a pebble actually 1 m below the surface appear to a person viewing it from above?

Solution: A ray of light emanating from the pebble will be refracted at the water surface and bend away from the normal as shown in the figure (because $n_{water} > n_{air}$), and the vertical ray will pass through the surface remaining vertical. Because of the refraction, the pebble appears closer to the surface than it really is. The image of the pebble can be found from the diagram using the small angle approximation that $\sin\theta \sim \theta$ because we are viewing from above at small angles from the normal. We write the law of refraction in this approximation as $n_{air}\,\theta' = n_{water}\,\theta$, so that from Table 20.1 $\theta = (1/1.33)\,\theta'$. When the refracted ray reaches the person's eye, his brain will extrapolate backwards and view the pebble as lying at point P′. To then calculate the distance AP′ of the image below the water surface, we use the fact that AB = AP′ tan θ' = AP tan θ. Using the small angle approximation tan $\theta \sim \theta$, we have that AP′ = AP (θ/θ') = AP (1/1.33) = 0.75 m.



When light is incident on an interface between two different transparent media, in general some will be reflected and some refracted, as was mentioned above. Although we have separately discussed the laws of reflection and refraction, calculating the relative amounts of the reflected and refracted beam intensities at a boundary requires a more sophisticated analysis from Maxwell's equations. The result for a beam at normal incidence on a plane surface coming from medium 1 to medium 2 is given by the following relation for the fraction of the incident intensity I_o that is reflected

$$\frac{I_r}{I_o} = \left(\frac{n_2 - n_1}{n_2 + n_1}\right)^2. \tag{20.6}$$

The fraction of the incident intensity that is transmitted is then equal to $1 - I_r/I_o$, so that the total fraction adds to 1 (remember that we have neglected any absorption of light).

Using the example of a light beam in air striking a plane piece of glass (with $n = 1.5$) normal to its surface, we find from Equation (20.6) that 4% of the incident intensity is reflected

$$\frac{I_r}{I_o} = \left(\frac{1.5 - 1}{1.5 + 1}\right)^2 = \left(\frac{0.5}{2.5}\right)^2 = \frac{1}{25} = 4\%.$$

Figure 20.12 shows the results of a more difficult calculation of the fraction of unpolarized light that is reflected from a glass surface as a function of incidence angle. We discuss the case of polarized light in the next chapter. As the incidence angle approaches 90°, the reflected intensity approaches the total incident intensity. We conclude that light striking a plane glass surface at a grazing angle is totally reflected; thus, at grazing incidence the glass surface acts like a mirror. Try this out yourself with a window pane or other glass surface!

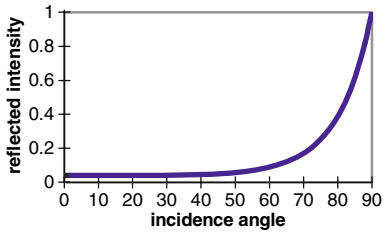


FIGURE 20.12 *The fraction of unpolarized light intensity in air reflected from a glass (n = 1.5) surface as a function of the incidence angle.*

3. SPHERICAL MIRRORS

Not all mirrors are plane mirrors discussed in the previous section. The most common type of curved mirror is the spherical mirror, typically made from a section of a spherical shell of glass that is polished and coated with a highly reflective metal coating on the back of the viewing side, to prevent the coating from getting scratched or damaged. Spherical mirrors come in two types, depending on which side of the spherical surface faces the light. Concave mirrors are made to reflect light from the side facing the center of the spherical surface (the "inside" or cave side) whereas convex mirrors are made to reflect light from the other "outside" surface. Examples of each are fairly common in our everyday life. Concave mirrors (see Figure 20.13) are used as makeup or cosmetic mirrors to produce an enlarged image. Convex mirrors are used as passenger side-view mirrors in cars, giving a wider viewing range, or as security mirrors in stores.

Spherical mirrors work in the same fundamental way that plane mirrors do to produce images based on the law of reflection. As shown in Figure 20.14 for a cross-section through a concave mirror, consider light that is traveling parallel to the so-called principal axis (shown in blue), going through the center of curvature C (the point located an equal distance R—the radius of curvature—from all parts of the mirror surface) and the center of the mirror. All such light rays will reflect from the mirror surface according to the law of reflection and converge at the focal point F of the mirror, a distance f from its surface along the principal axis, as shown. If the mirror dimensions are small compared to the radius of curvature, then the convergence of the reflected light will be "tight" and the light will converge to a focal point. If the mirror is larger, then the focal point will be smeared out a bit due to light farther from the principal axis getting focused to a shorter focal length, an effect known as spherical aberration and discussed more in the next chapter for lenses. If the surface of the mirror were parabolic, rather than spherical, then the reflections would truly focus at

FIGURE 20.13 (left) *Andrea's inverted image in a large concave mirror;* (right) *A Victorian bear and its image, with the bear located closer to the mirror than its focal point* (see below) *giving an enlarged upright image.*

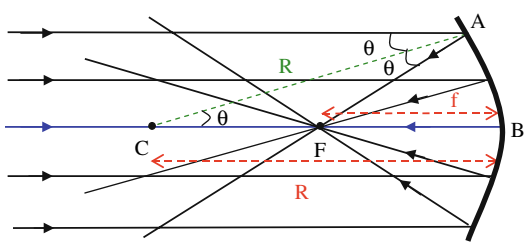
the same focal point no matter how large the mirror. In fact, although more difficult and therefore more expensive to manufacture, parabolic mirrors are used in a number of more critical applications, including solar collectors to trap sunlight for energy conversion, reflecting telescopes, and car headlights (which work by sending light in a reverse path from the focal point to a beam of light emerging parallel to the principal axis of the mirror).

Using the geometry shown in Figure 20.14 based on the law of reflection, we show in the caption that the focal length for a concave spherical mirror is given by

$$f = \frac{R}{2}, \tag{20.7}$$

where the focal length is measured from the mirror reflecting surface as shown. Thus, the radius of curvature directly determines the focal length of the mirror, the key parameter we show that determines image formation in the mirror.

FIGURE 20.14 *Light rays parallel to the principal axis (in blue) reflecting from a concave mirror and converging at the focal point F. The dotted green line indicates a radius, normal to the mirror, with the upper ray reflecting according to the law of reflection and passing through the focal point. A similar construction is implied for the other rays. Note that the triangle CAF is isosceles so that CF = AF. In the case that the angle θ is small, then AF = BF and we have that CF = BF so that Equation (20.7) follows.*



3.1. IMAGE FORMATION

An object located some finite distance from a spherical concave mirror will produce an image in the mirror. In order to locate the position and size of the image we can use a process known as *raytracing*, in which we follow three special rays that emanate from the object. Take a look at Figure 20.15 and note the object, represented by an upright arrow, located at distance d_o, the object distance, from the mirror surface. If we draw a ray (#1) from the tip of the arrow going parallel to the principal axis, we know that it will reflect through the focal point as shown in red. A second ray (#2) from the arrow tip that goes directly through the focal point will reflect from the mirror parallel to the principal axis (drawn in blue). We know this because it is just the time-reversed process from ray #1. The intersection of these two rays will be the location of the image of the arrow tip (or top of the object) in the mirror. We can confirm this with a third special ray (#3 drawn in green) that appears to emanate from C, the center of curvature, and which will reflect directly back on itself because the radial line is perpendicular to the mirror surface. The three rays cross at a common point, the image location, a distance d_i from the mirror. In fact, then, all rays that leave the arrow tip and reflect from the mirror will converge to the same image location; our three special rays are simply chosen because of the ease of this construction to locate the image.

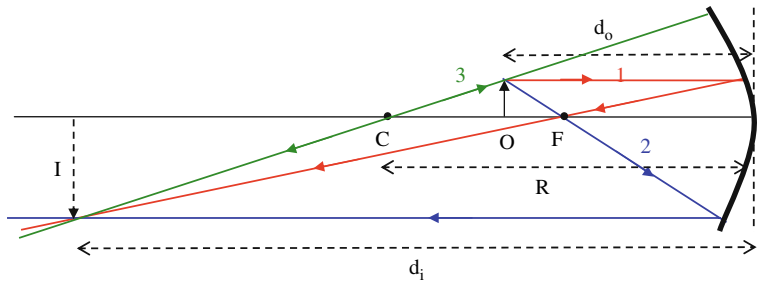


FIGURE 20.15 *Raytracing construction for a concave mirror. From the object at O we can construct three special rays: (1) a ray parallel to the principal axis reflects through the focal point; (2) a ray through the focal point reflects parallel to the principal axis; (3) a ray from the object that appears to have come from the center of curvature C will reflect back on itself. The common intersection of the reflected rays is the location of the image of the object at I.*

3.2. MIRROR EQUATION

If we examine the diagram shown in Figure 20.16, we can derive a quantitative relationship between d_o, d_i, and f known as the mirror equation. In this figure you can see the same mirror, object, of height h_o, and image, of height h_i, shown in Figure 20.15, together with ray #3 from that figure, as well as an additional ray (let's call it ray #4, in purple) which is directed at the point where the principal axis meets the mirror and reflects making equal angles with the axis to also reach the image location. Ray #4 makes up the hypotenuse of the two colored, dotted, similar triangles, giving us

$$\frac{h_i}{h_o} = \frac{d_i}{d_o}.$$

Ray #3 also forms the hypotenuse of two similar triangles with a common vertex at point C (one shown in black hatched lines) and with opposite sides given by the heights of the object and image. From these similar triangles we have that

$$\frac{h_i}{h_o} = \frac{d_i - R}{R - d_o}.$$

Setting these two expressions for the transverse magnification, $m = h_i/h_o$, equal to each other we have

$$\frac{d_i}{d_o} = \frac{d_i - R}{R - d_o}.$$

If we substitute $R = 2f$ and cross-multiply we have

$$(d_i)(2f - d_o) = (d_i - 2f)(d_o).$$

One more step of simplification gives us

$$d_i f - d_i d_o + f d_o = 0,$$

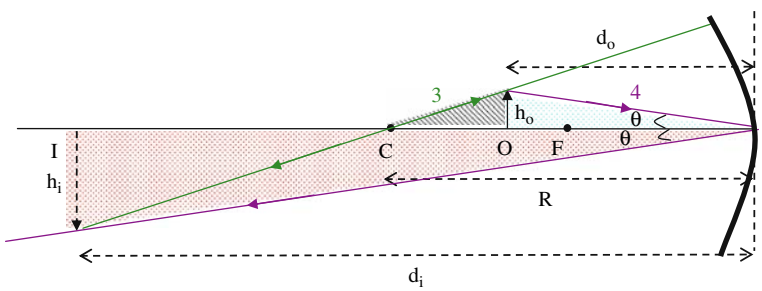


FIGURE 20.16 *Geometry used to derive the mirror equation. The blue and red dotted triangles are similar as are the black hatched triangle and the red triangle with vertex at C and opposite side at the image I.*

and if we divide by the product $(d_o d_i f)$, we find the *mirror equation*

$$\frac{1}{d_o} + \frac{1}{d_i} = \frac{1}{f}. \tag{20.8}$$

We also find from the first relation above that the magnification $m = h_i/h_o$ is given by

$$m = -\frac{d_i}{d_o}, \tag{20.9}$$

where the negative sign is inserted according to a convention to indicate that the image is inverted (upside down). Now that we have the mirror equation, we can determine the location (and magnification) of the image of an object in a concave mirror.

Example 20.3 A concave mirror has a radius of curvature of 25 cm. A 2 cm tall object is placed 20 cm from the mirror along its axis. Find the location of the image and its size.

Solution: Using $d_o = 20$ cm and $f = R/2 = 12.5$ cm and solving the mirror equation for the image distance gives

$$\frac{1}{d_i} = \frac{1}{f} - \frac{1}{d_o} = \frac{1}{12.5} - \frac{1}{20} = 0.03\ \text{cm}^{-1} \text{ so that } d_i = 33.3\ \text{cm}.$$

The magnification, according to Equation (20.9) is $m = h_i/h_o = -d_i/d_o = -1.67$, so that h_i, the size of the image, is $h_i = -1.67\ h_o = -(1.67)(2\ \text{cm}) = -3.33$ cm. Here, the minus sign indicates that the image is upside down (inverted from the upright object). The raytracing diagram would be similar to that shown in Figure 20.15.

Using the mirror equation to solve for the image distance when we are given the object distance and focal length, we have that

$$\frac{1}{d_i} = \frac{1}{f} - \frac{1}{d_o} = \frac{d_o - f}{f d_o} \quad or \quad d_i = \frac{f d_o}{d_o - f}.$$

If we imagine that, for a given f, we change the position of the object d_o, then clearly as long as $d_o > f$, the value for d_i will be positive. What does it mean when $d_o \leq f$ so that d_i is infinite or negative? Let's examine this question using a ray diagram for such a situation in Figure 20.17.

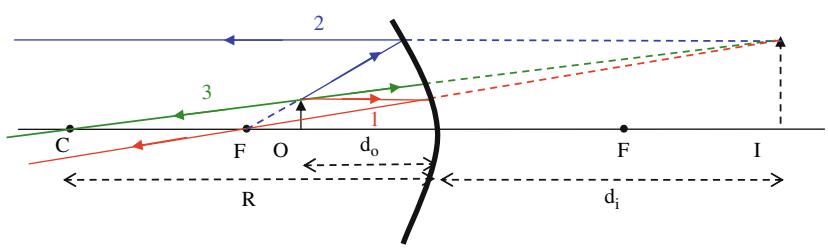


FIGURE 20.17 *Raytracing diagram for the case when* $d_o < f$, *resulting in a virtual image.*

Raytracing is similar in this case with three special rays drawn: ray (#1) from the tip of the arrow object going parallel to the principal axis reflects through the focal point as shown in red; a second ray (#2) from the arrow tip that seems to emanate from the focal point will reflect from the mirror parallel to the principal axis as shown in blue; a third ray (#3 drawn in green) that appears to emanate from C, the center of curvature will reflect directly back on itself. The three rays do not ever cross at a common point, and so there is no "real" image. Instead, if you were to view the light from the left side, it would appear to be diverging from an image located a distance d_i from the mirror but behind it! Such an image is called a virtual image. You could not put a piece of paper there to see the light actually form the image on the paper. It is similar to the virtual image seen in a plane mirror. The image is, in fact, erect (not inverted) and magnified.

In order to be able to use the mirror equation in this and other cases, we adopt the set of sign conventions shown in Table 20.2. When used consistently, these allow us to not only find the location of the image for both concave and convex mirrors, but also to find the magnification and whether the image is erect or inverted. Another two examples help us to see how these are applied.

Table 20.2 *Sign Conventions for Mirror Equation*

Quantity	*Positive When*	*Negative When*
Focal length, f	Concave mirror	Convex mirror
Object distance, d_o	Real object (usual case)	*Virtual object (rarely)
Image distance, d_i	Real image (located in front of mirror)	Virtual image (located behind mirror)
Magnification, m	Erect	Inverted

*In some optical systems, the image from one piece of optics can serve as the object for the next mirror; in this case the object can sometimes be found located behind the mirror.

Example 20.4 A 1 cm tall object is located 10 cm from a concave mirror with a radius of curvature of 40 cm. Characterize the image produced by the mirror.

Solution: Because the focal length is $R/2 = 20$ cm, we use the mirror equation to find the image distance

$$\frac{1}{d_i} = \frac{1}{f} - \frac{1}{d_o} = \frac{1}{20} - \frac{1}{10} = -\frac{1}{20}$$

so that, on inversion, we have $d_i = -20$ cm. The negative sign indicates that the image is virtual and located behind the mirror. We find a magnification of $m = -d_i/d_o = -(-20)/(10) = 2$, so that the image is erect and 2 cm tall.

Example 20.5 A passenger side-view car mirror is convex with a radius of curvature of 150 cm. If a car that is viewed in the mirror is actually 20 m away, describe the image in the mirror.

Solution: In our case, we can use the mirror equation with $f = -R/2 = -75$ cm and $d_o = 20$ m to find that

(*Continued*)

$$\frac{1}{d_i} = \frac{1}{f} - \frac{1}{d_o} = -\frac{1}{75} - \frac{1}{20} = -0.063\ \text{cm}^{-1} \quad \text{thus} \quad d_i = -15.8\ \text{m}.$$

Therefore, the image is virtual and has a magnification of $m = -d_i/d_o = -(-15.8)/20 = 0.79$. Convex lenses are used in side-view car mirrors to get a wide field of view. As we've seen, they will actually make the object appear to be smaller, which our eyes/brain translate to mean that the car is farther away. All convex car mirrors have the emblem, "Objects in mirror are closer than they appear," written on them and you need to be aware of this fact.

4. OPTICAL FIBERS AND THEIR APPLICATIONS IN MEDICINE

Consider the law of refraction for an incident beam emanating from medium 1 with a larger index of refraction (glass, e.g.) and incident on a plane boundary with medium 2 of lower index of refraction (air, e.g.). In this case the refracted beam is bent away from the normal as it passes from the glass into the air. Because $\sin\theta$ has a maximum value of 1, and because n_1/n_2 is larger than 1, we see that Snell's law

$$\sin\theta_2 = \frac{n_1}{n_2}\sin\theta_1 \tag{20.10}$$

does not allow the full range of values for θ_1, the angle of incidence. The maximum incidence angle allowed, which we call the *critical angle* θ_c, is given by setting $\sin\theta_2$ in Equation (20.10) equal to 1 to find that

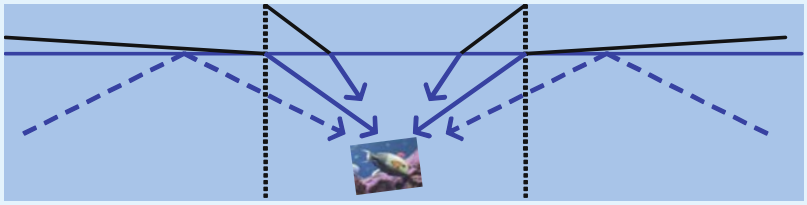
$$\sin\theta_c = \frac{n_2}{n_1}. \tag{20.11}$$

Larger incidence angles do not allow a solution for an angle of refraction in Snell's law. What does this mean?

If we imagine increasing the angle of incidence from 0°, since the refracted beam is bent away from the normal, just at the critical angle, the angle of refraction is 90° and the beam does not really enter the air. At larger angles still, there is no refracted beam. This, of necessity, implies that all of the intensity of the incident beam is reflected back into the glass in a process known as *total internal reflection* (Figure 20.18). For the glass–air interface, the critical angle using $n_{\text{glass}} = 1.5$ is 42°.

Example 20.6 A fish sees the world above the smooth water surface confined within a circular viewing disk above it. Calculate the angular spread of this disk.

Solution: Light from above the water surface is refracted toward the normal on entering the water. A ray that just grazes the water surface at $\theta \sim 90°$ will be refracted in the water to an angle θ', given by $\sin\theta' = \sin 90\ (1/1.33) = 0.75$, or $\theta' = 48.6°$. Therefore a fish (or you) looking up at the calm surface of the water from below will see the entire outside world within a circle of light making an angle of 48.6° with the normal. Therefore we see that light from the outside



world visible to the fish will form a cone with an apex angle of 48.6° at the fish. Light seen by the fish at larger angles will originate from within the water and consist of reflections from the water surface as shown in the figure.

Although there is no energy propagated into the second medium at incident angles at or above the critical angle, the electromagnetic fields do penetrate some small distance into the second medium. These fields make up what is called an *evanescent wave*, one that rapidly decreases in intensity with increasing depth into the second medium. One practical application of an evanescent wave is to illuminate only a thin layer of molecules in solution near an interface (Figure 20.19). In this way, either by simple microscopy or by fluorescence methods, those molecules near the interface can be selected for viewing. We return to this technique in Chapter 22.

Consider a solid cylinder of glass, as shown in Figure 20.20, into which a beam of light has been aimed. As long as the beam strikes the cylinder walls with an angle of incidence greater than the critical angle, the beam will continue to travel down the cylinder even if the cylinder is bent. Such a tube is called a light pipe and is able to bend light around curves although with tubes of macroscopic dimensions, the loss in intensity of light is substantial and these devices have not had much practical use.

Optical fibers are slender capillaries of glass or plastic material with a solid core, with diameters as small as about 10 μm, surrounded by a concentric layer of lower refractive index material, known as the *cladding*. As shown in Figure 20.21, these fibers function in the same way as light pipes but are much more efficient at confining the light intensity. Optical fibers can maintain about 10% of the incident intensity even after a fiber length of 30 mi. The cladding is used to provide a highly efficient mechanism for total internal reflection.

Optical fibers are used extensively in telecommunications to carry encoded light signals for audio for telephone, audio/video for television, and high-speed information transfer for computers. Optical fibers currently can carry at least

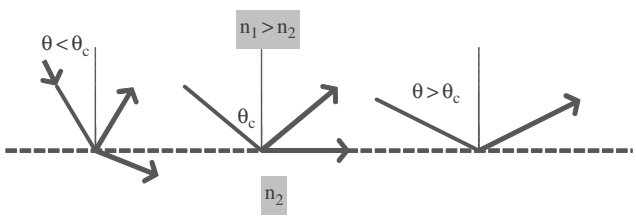


FIGURE 20.18 *Incident rays at angles less than, equal to, or greater* (from left to right) *than the critical angle.*

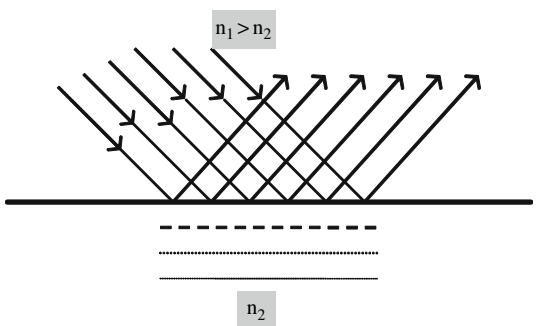


FIGURE 20.19 *A beam of light incident at an angle slightly larger than the critical angle, showing evanescent waves entering the second medium with decreasing intensity. These waves only penetrate microscopic distances and do not propagate into the second medium. We show in Chapter 22 that evanescent waves are useful in fluorescence microscopy of surface phenomena.*

FIGURE 20.20 *Total internal reflection in a light pipe.*

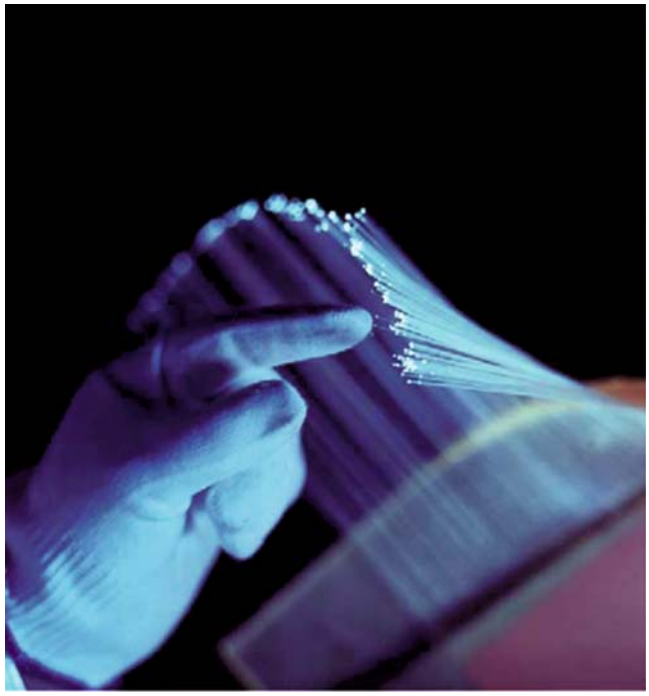


FIGURE 20.21 *Glass optical fibers transmitting laser light.*

1000 times more bits per second of digital information than electrical wires and in principle can carry orders of magnitude more. Because of this and the huge reduction in size and cost of such cabling, all new cabling for communications is done using fiber optics.

The technique known as wavelength division multiplexing (WDM) is used to increase the information content sent over a fiber optic manyfold. Multiple independent lasers using different wavelengths of light are coupled using special optical devices and the light is sent down a single common fiber optic. Each laser signal is modulated to produce pulse trains of light containing an enormous amount of information and each wavelength propagates down the fiber essentially independently of the others. Using similar optical devices at the receiving end, each wavelength can be independently split from the others to have its information decoded.

Although a single fiber cannot generate an image of a source of light, bundling of many fibers together in a coherent bundle, one that maintains the same relative positions of the fibers at each end, allows an image to be transmitted. In medicine, optical fibers are used as tools in two major ways: viewing internal features of the body and internal laser surgery or therapy. By the insertion of a flexible fiber optic bundle of many individual fibers, but still with a total diameter comparable to that of a hypodermic needle, light can be "injected" into remote internal areas of the human body and the reflected light can be captured in other fibers within the bundle and transmitted out of the body for viewing (Figures 20.22 and 20.23). This device is called an endoscope and versions exist for viewing within the lungs, GI tract, or the blood vessels and heart. Indeed these devices have allowed our first views within a living body. Movies of internal body parts using this technology are now common on television science programs.

In some small number of medical procedures using fiber optics, laser light can perform internal surgery. Various types of surgeries have been done on a limited basis, including the shattering of kidney stones, destruction of tumors, and laser angioplasty to remove plaque buildup in blood vessels. Many of these surgeries are still under development and are not without problems, but can sometimes provide a useful alternative. Medical and other applications of lasers are discussed in Chapter 25.

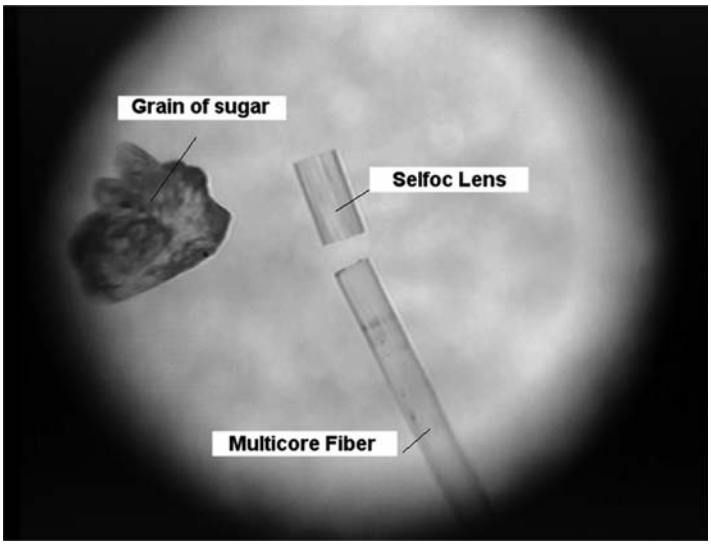


FIGURE 20.22 *Elements of an endoscope with a 200 μm diameter fiber.*

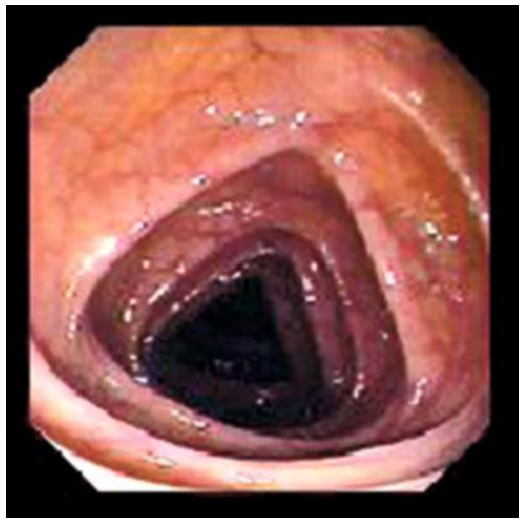


FIGURE 20.23 *View of normal colon using endoscopy.*

CHAPTER SUMMARY

Light travels through the vacuum at speed c, but travels through a transparent medium at a slower speed, given by

$$v = \frac{c}{n}, \tag{20.3}$$

where n is the material's index of refraction.

When light strikes a boundary between two media with different indices of refraction, n_1 and n_2, some of the light is reflected and the remainder is transmitted. The laws of reflection and refraction govern the angles at which the reflected and refracted rays appear:

$$\theta_r = \theta_i, \tag{20.4}$$

$$n_1 \sin\theta_1 = n_2 \sin\theta_2 . \tag{20.5}$$

Spherical mirrors produce an image located at d_i of an object located at d_o according to the mirror equation,

$$\frac{1}{d_o} + \frac{1}{d_i} = \frac{1}{f}, \tag{20.8}$$

where the focal length f is related to the radius of curvature of the mirror R by

$$f = \frac{R}{2}. \tag{20.7}$$

The magnification of the image is given by

$$m = \frac{h_i}{h_o} = -\frac{d_i}{d_o}, \tag{20.9}$$

where h_i and h_o are the heights of the image and object, respectively. These equations can be used for all spherical mirrors, as long as the sign conventions of Table 20.2 are followed.

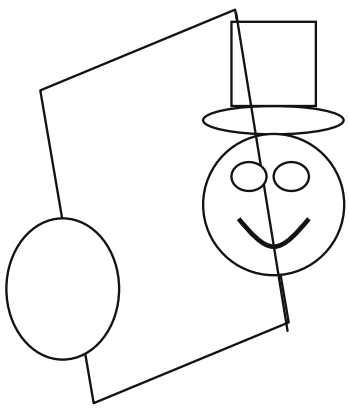
In the case of refraction from a larger index medium n_1 to a lower index medium n_2, there is a critical angle θ_c above which there is no transmitted light (all the light is reflected). This is known as total internal reflection and the critical angle is given by

$$\sin\theta_c = \frac{n_2}{n_1}. \tag{20.8}$$

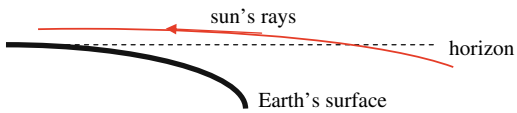
This is the basis of optical fibers, extremely thin glass filaments (core) surrounded by a thin glass layer (cladding) with lower index of refraction. Optical fiber applications range from communications, replacing electrical wires, to industrial and medical methods that have revolutionized our capabilities.

QUESTIONS

1. Does the speed of light in fused quartz increase or decrease as the wavelength of the light increases? Find the ratio of the speed of light in quartz at 800 nm to that at 400 nm.
2. Rank the following media in terms of increasing speed of yellow light through them: crown glass, water, diamond, and air.
3. Explain why a spherical wavefront emanating from a point source appears as a plane wave at a distant small detector.
4. Suppose that light emitted from a long straight wire is initially in phase as it leaves the wire. What will be the shape of the wavefront from such a wire?
5. Give a geometric argument to explain why a plane mirror always produces virtual images of the same size as the object.
6. In a wonderful demonstration using a plane mirror, the demonstrator and the observer each put their forehead and nose up against opposite edges of a large free-standing plane mirror. The back of the mirror has a partition blocking any viewing so that they can only see each other with their one eye in front of the mirrored surface. What will they each see in the mirror when they look at the other person? Now, for the trick. The demonstrator wears a hat which he holds onto with "both" hands. Then he can mysteriously make the hat rise off his head by lifting it with his "invisible" hand, the one behind the mirror. Try this out with a friend.



7. An analogy that is often qualitatively used to explain the law of refraction is a formation of soldiers marching in straight lines at a uniform rate that approaches a stream at an angle. As the first soldiers reach the stream they slow down, whereas the soldiers on the ground still maintain their same marching cadence and speed. Making the analogy of wavefronts to rows of soldiers, show that Figure 20.10 and Equation (20.5) are appropriate (with $n \propto 1/v$) and that the "wavelength" of the formation will scale with $1/v$ as well.
8. The diagram shows that it is possible to see the sun after it has set below the horizon. Discuss why this is so. (Hint: This is related to the fact that the atmosphere gets less dense with increasing altitude and so the index of refraction decreases with increasing altitude as well.)



9. Example 20.2 showed that the apparent depth of objects under water is foreshortened, as seen in air. What do you expect you will see if you view an object in air from under water? Will its "depth" in air be foreshortened or lengthened?
10. Noting the fact that 4% of light incident on a glass–air boundary is reflected, what fraction of the incident light on an air-spaced double-paned glass window with no special optical coatings is reflected?
11. Consider the graph of Figure 20.12 which indicates the fraction of light reflected from a surface as a function of incidence angle. Assume that water, ice, and the surface of mineral rock, or asphalt, all behave optically like that of glass, at least approximately. Explain how it is that when looking toward a distant but approaching automobile at night it may appear that there is a double set of headlights. As the car nears, the appearance of doubling goes away.
12. When the sun sets and the air is particularly uniform in temperature and humidity it is occasionally possible to see the "green flash" of the sun just as it sets. The last bit of sunlight seen on the horizon changes color from reddish orange to green. Discuss the reason for this in light of the fact that the blue and violet components of the sunlight have been scattered out of the light that reaches our eyes. Consider how the different colors in the light reaching our eyes are refracted.
13. Is a virtual image formed in a mirror, spherical or plane, located behind or in front of the mirror?
14. Consider a concave mirror with a focal length f and the following different ranges of distance for an object to be imaged. State whether the image is upright or inverted and whether it is enlarged or reduced in size: $d_o < f$, $f < d_o < 2f$, and $d_o > 2f$.
15. A small object is imaged in a large spherical mirror and the image appears blurred due to spherical aberration. What can be done to improve the quality of the image?
16. Why is the central core of an optical fiber surrounded by cladding? What properties should cladding have to be effective?

MULTIPLE CHOICE QUESTIONS

1. The image formed in a plane mirror is (a) right–left, up–down and in–out reversed, (b) right–left and in–out reversed, but up–down not reversed, (c) right–left reversed but in–out and up–down not reversed, (d) none of the above.

2. Two plane mirrors are stood vertically making a right angle between them. How many images of an object close to and in front of the mirrors can be seen? (Hint: Raytrace a picture of the situation.) (a) 1, (b) 2, (c) 3, (d) 4.

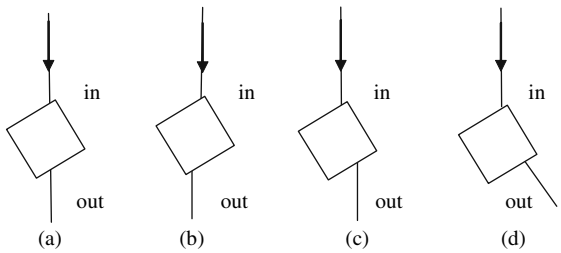
3. Headlights from a car illuminate the road and surroundings for the driver at night due to (a) refraction, (b) specular reflection, (c) diffuse reflection, (d) aberrations.

4. If, in the previous question, the road is wet, it will be harder to see at night because, relative to the situation with a dry roadway (a) there is more refraction, (b) there are more aberrations, (c) there is more specular reflection, (d) there is more diffuse reflection.

5. When looking down into the smooth surface of a clear lake you can see underwater fish at small angles of incidence but if you look out at larger angles of incidence, you will see reflections off the surface. This is because when you look out at larger angles, (a) the fish are farther away, (b) there is more reflected light, (c) there are more diffuse reflections, (d) there is less refracted light.

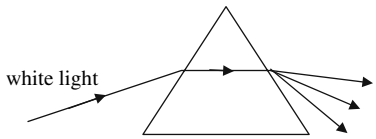
6. A pencil immersed in a glass of water appears to be bent. If outside the water the pencil makes an angle of 30° with the normal to the water surface, what is the apparent "bending angle" of the pencil at the water surface? (a) 22°, (b) 42°, (c) 12°, (d) 8°.

7. When a laser beam is aimed onto the face of a transparent glass cube, which of the following best illustrates the direction of the beam after it emerged from the cube?



8. A ray of light enters a cube of glass making an angle of 36° with respect to the normal of the entrance face. The index of refraction of the glass is 1.52. The ray emerges from the opposite face with an angle relative to the normal equal to (a) 23°, (b) 36°, (c) 54°, (d) 67°.

9. Based on Snell's law and Figure 20.2, if a beam of white light is refracted by a glass prism as shown, what will be the order of the "rainbow" of colors starting with the most refracted? It will be (a) red, yellow, blue, (b) blue, yellow, red, (c) red, blue, yellow, (d) blue, red, yellow.



10. If a beam of light refracts at a plane boundary between two different transparent media, what happens to the angle of refraction as the wavelength of the incident light increases? It will (a) increase, (b) decrease, (c) remain the same, (d) it depends on the incident and transmission media.

11. When light strikes a plane boundary between two media with different refractive indices which of the following cannot occur? (a) There is a reflected beam, but no transmitted beam, (b) there is a transmitted beam but no reflected beam, (c) there are both transmitted and reflected beams, (d) the speed of the light increases on entering the second medium.

12. Two identical beams of light traveling through water strike either crown glass or diamond at normal incidence. If the incident intensity of each beam is 10 mW, how much more light is reflected from the diamond? (a) 0.8 mW, (b) 1.9 mW, (c) 0.52 mW, (d) 2.3 mW.

13. When an object is placed 5 cm from a concave spherical mirror with a radius of curvature of 8 cm, the image formed will be (a) real and erect, (b) real and inverted, (c) virtual and erect, (d) virtual and inverted.

14. Where should an object be placed in front of a concave spherical mirror so that the image is at the same location as the object? At a distance equal to (a) f, (b) $2f$, (c) $3f$, (d) $f/2$.

15. To produce an image of an object in a concave spherical mirror at infinity, the object should be placed at (a) the center of curvature, (b) just up against the mirror on the principal axis, (c) at the focal point, (d) none of the above choices is correct.

16. Which of the following is a false statement about raytracing in a spherical mirror? (a) A ray from the object through the focal point will reflect parallel to the axis, (b) a ray from the object through the center of curvature will reflect on itself, (c) a ray parallel to the axis will reflect through the center of curvature, (d) a ray from the object directed to the point where the principal axis meets the mirror will reflect at an equal angle to the axis.

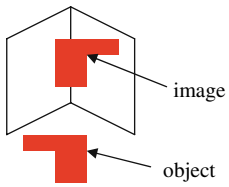
17. When a physician snakes an endoscope down a patient's esophagus she is able to see into the patient's stomach. Which of the following is most directly related to the optics of this process? (a) Brewster's angle, (b) dispersion, (c) total internal reflection, (d) diffraction.

18. At the plane boundary between two transparent media, as a ray of light approaches the boundary, from the medium of higher refractive index, at increasingly greater angles from 0° to approaching the critical angle (a) the reflected light decreases and

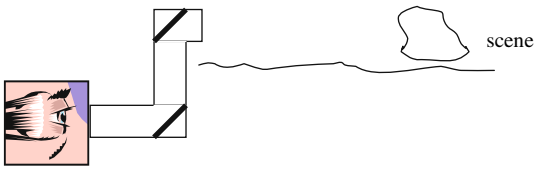
the transmitted light increases, (b) both the reflected and transmitted light decrease, (c) the reflected light increases and the transmitted light decreases, (d) both the reflected and transmitted light increase.

PROBLEMS

1. If you set two plane mirrors at a right angle with respect to each other, as shown, you will see yourself in three images. Two are direct images in either mirror and have left–right reversals, but the image in the corner is just as others see you, without a left–right reversal. Show how this occurs by drawing a ray diagram.



2. Describe how a periscope works (see figure) to provide an image of a scene just as you would see it with your eyes if they had a direct view of the scene.

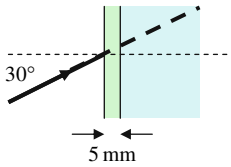


3. The two plane mirrors of the first problem are rotated toward one another so that the angle between them is 60°. The images of an object will form symmetric patterns. This is the basis of a kaleidoscope. If an object is put along the bisecting line between the mirrors, show in a diagram where the images (some are images of images) are located. Generalize this idea if the angle between the mirrors is made to be $360°/n$, where n is a small integer.

4. You are standing 2 m from a plane mirror. If your eyes are 8 cm apart, what angle apart do your eyes appear to you when you see them in the mirror?

5. Two different observers, standing at different distances from a plane mirror (and slightly offset so that they can each see the other), each hold an object in their hands. The observer 4 m from the mirror carries a 10 cm tall object and the observer 2 m from the mirror carries a 5 cm tall object.
(a) Find the angle subtended by each object as seen by each observer.
(b) Show that the ratio of the angle subtended by the taller object as seen by the nearer observer to that subtended by the shorter object as seen by the farther observer is the same as the direct ratio of the object heights.
(c) Show also that this ratio is independent of where the observers are located as long as they are much farther from the mirror than the object height distances.

6. A narrow pencil of light strikes the side of a rectangular fish tank at an angle of 30° below the horizontal as shown.
(a) What angle does the light ray make with the horizontal in the glass, assuming a 1.55 index of refraction?
(b) What angle does it make in the water?
(c) If the glass wall is 5 mm thick, by what distance is the exit spot inside the glass wall displaced from the location at which the incident beam is aimed?



7. Consider a slab of material with thickness $t = 2$ cm with an index of refraction 1.5 and where the upper and lower surfaces are parallel to each other. If light is incident at an angle θ with respect to the normal to the surface, show that the beam leaves parallel to itself on the other side of the slab at the same angle. If $\theta = 30°$, find the displacement (the perpendicular distance shifted) of the emerging beam from its incident direction.

8. In the previous question, if the light has a frequency of 88.3 MHz what is the speed of light in the medium, the wavelength of light in the air and in the medium, and how long does it take the wave to traverse the medium?

9. A concave spherical mirror is used by a dentist to produce an enlarged image of a tooth. If the radius of curvature of the mirror is 2.0 cm, how close is the mirror to the tooth when the image appears triple the size of the tooth? Is the image erect or inverted? Real or virtual?

10. A 3.5 cm tall object is placed 20 cm in front of a concave mirror. A real image forms that is 7 cm tall. Where is the image located? Is it erect or inverted? What is the radius of curvature of the mirror?

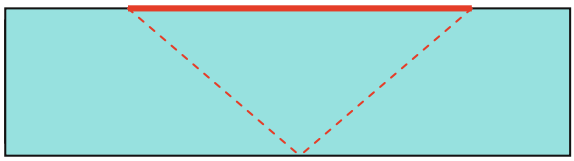
11. A 0.25 m diameter convex mirror is mounted high on the wall of a store as a security mirror. If a person is 5.0 m away from the mirror which has a radius of curvature of 2.0 m, find the height of the image formed by the mirror if the person is 1.5 m tall. Will the full height of the person be visible in the mirror if their feet are imaged at the mirror's bottom edge?

12. A concave makeup mirror has a radius of curvature of 25 cm. How close to the mirror should a young woman's nose be in order to see an image enlarged by three times? Draw a ray diagram to illustrate this after you find the answer.

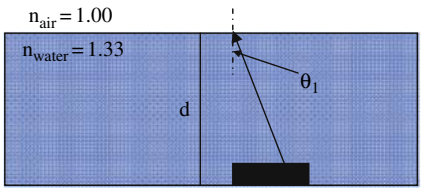
13. A small coin is placed a distance of 3 cm from a concave mirror with a 15 cm radius of curvature. Describe the image that is formed, finding its location, magnification, and whether it is erect or inverted and real or virtual.

14. A car that is 10 m from a convex side-view car mirror, with a 150 cm radius of curvature, is imaged in the mirror. Find the image location, magnification, and whether it is erect or inverted and real or virtual.

15. A small bright light is at the bottom of a large 8 ft deep swimming pool illuminating the surface of the pool. Show that the illuminated region, as seen from above, will be a circle of light and find its radius.



16. Where does the image of the box appear to be located, given the object is located a depth d below the surface and θ_1 is the angle of incidence of the ray with respect to the normal to the water–air surface?



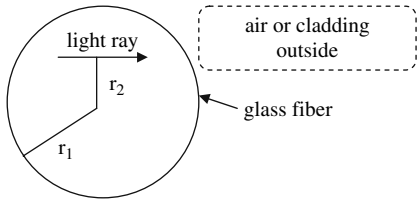
What is the apparent depth d' of the box in terms of the actual depth d, and the indices of refraction, n_{water} and n_{air}. (Hint: Use the fact that for small angles $\sin\theta \sim \tan\theta$, and use Snell's law.)

17. What is the angle θ such that the light rays will be totally internally reflected assuming that the pipe has an index of refraction of $n_{pipe} = 1.30$ and is surrounded on all sides by air.

18. Consider a light ray inside a cylinder of glass of refractive index n. The glass cylinder is in air. The direction of the ray is perpendicular to the cylinder axis. With r_2 equal to the distance of the light ray path from the center of the cylinder,

(a) Show that if $r_2 \geq r_1/n$, the ray will not escape the glass, but will follow a closed and broken polygonal path within due to total internal reflection.

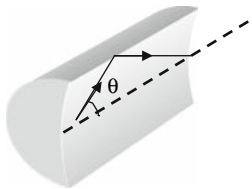
(b) Suppose this glass cylinder is a core cylinder, surrounded by a cladding wrapper having a refractive index smaller than that of the core. Now show that the ray within the core is trapped if $r_2 \geq r_1 (n_{cladding}/n_{core})$.



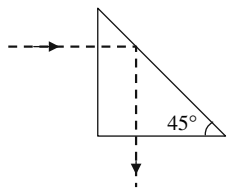
19. The diagram shows a cut lengthwise along an optical fiber. Such a cut defines a "meridional plane" of the fiber cylinder and a light ray confined to this plane is called a meridional ray.

(a) Find the maximum angle that a light ray can make with the axis of the fiber and still be confined to the fiber via total internal reflection if the fiber has refractive index of 1.4 and is surrounded by air.

(b) uppose the fiber is composed of core and cladding, with refractive indices, $n_{core} = 1.400$ and $n_{cladding} = 1.386$, a difference of 1%. Now find the maximum angle with the axis for a ray confined to the core.



20. A piece of glass in the shape of a right triangular prism has a beam of light enter normal to one face as shown. What is the minimum refractive index of the glass so that light entering one end will round the bend via total internal reflection? (There is no mirror coating on the bevel!)

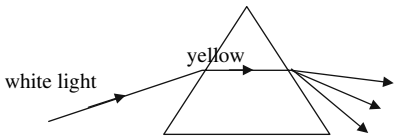


21. Refractive index is dependent on the wavelength and hence the color of light, increasing in magnitude from red to blue-violet. Typically, the refractive index given for a particular material refers to a wavelength value near the middle of the visible spectrum.

(a) Suppose white light is incident on an equilateral prism made from crown glass as shown. This

means that all colors, and hence all wavelengths in the incident beam are traveling in the same direction and strike the prism at the same angle. If refraction occurs such that yellow light (590 nm) travels along the path exactly parallel to the bottom face of the prism, determine the directions of red (630 nm) and blue (490 nm) light within the prism. (Estimate n values from Figure 20.2.)

(b) Sketch what will happen to these colors at the second interface, where the light exits the prism glass. Specifically, show that the colors do not recombine but exit in a range of angular directions, resulting in a "rainbow" of color.

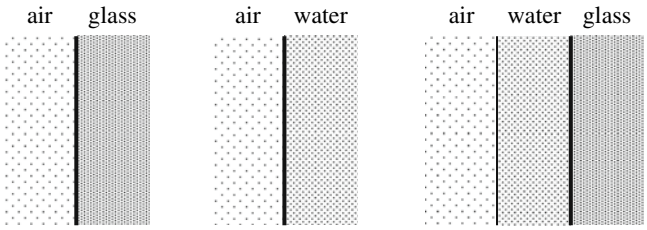


22. A beam of light in air consisting of two fairly pure colors with wavelengths of 400 and 550 nm is incident on a plane glass surface at an incident angle of 30°. Using n(400 nm) = 1.53 and n(550 nm) = 1.51, find

(a) The refraction angles of each color

(b) The separation of the two colored beams after traveling a depth (measured along the normal) of 1 m into the glass

23. At every interface, there is reflection and transmission according to the amount of change in refractive index at the boundary. Consider an air–glass interface, an air–water interface, and an air–water–glass interface. Calculate the percentage of light that makes it through each interface arrangement for incident light normal to the surface. (For the case of the air–water–glass interface, consider only a single pass, i.e., don't worry about light that is reflected from the second interface and then back again from the first interface, some of which now will make it through the second interface at the second incidence.)



24. A major problem with larger diameter fibers is the difference in travel times of rays along a fiber. In traveling a distance d, the shortest time is that of the axial beam $t_1 = d/v$, and the longest time t_2 is that of a ray bouncing back and forth along the fiber just at the critical angle. Compute the time difference between these two rays for a 1.5 index fiber that is 10 km long and 100 μm in diameter, surrounded by 1.49 index cladding. This effectively limits the frequency of a signal that can be transmitted without significant degradation in larger diameter fibers. Small diameter (~10 μm diameter) single-mode fibers, in which the light travels as a wave and not as a geometrical ray, overcome this problem.

25. A second major problem in communications with fiber optics is attenuation of signal along a fiber due to scattering and other losses. Over recent years there has been a truly tremendous improvement in the transmission of optical glass fibers of about 100 orders of magnitude. At near-IR wavelengths the fiber losses can now be kept very low, at under −0.4 dB/km (look back to Chapter 11 on sound for a discussion of dB; because this is a loss in intensity, it will be negative). What percent of the incident signal is lost after traveling 50 km? Fiber amplifiers are used to boost the signal periodically for longer distance communication over fibers.

21 Optical Lenses and Devices

In the last chapter we learned the fundamental laws that tell us how light behaves when traveling through a transparent material. Here we apply those to the most important case of optical lenses, both human-made, for use in optical devices, and naturally occurring, as in the eye. After learning the basics of imaging using a single simple lens, we show how the human eye exquisitely functions to allow us to see in color at extremely high resolution. Two human-made optical devices, the magnifying glass and the compound microscope, are then examined. These can be fairly well understood with only the tools of geometric optics that we have learned. Understanding the huge arsenal of new optical microscopies studied in Chapter 23 requires knowledge of wave optics, presented in the next chapter. There we introduce the wave nature of light and some of its major consequences.

1. OPTICAL LENSES

In the previous chapter, after introducing reflection and refraction we focused on the phenomena of imaging from reflections in spherical mirrors and total internal reflection. Here we turn to the portion of the light incident on a transparent glass surface that is transmitted. Lenses are perhaps the most important of optical devices. For a glass lens, we know that only about 4% of the incident light at near normal incidence (in the paraxial approximation, with light traveling nearly along the optic axis) will be reflected at each boundary with air and so most of the light will be transmitted after being refracted by the curved surfaces. Special optical antireflection coatings can even increase the transmitted light closer to 100%. Armed with the law of refraction, we can repeat an analysis similar to that of the last chapter for reflected light to trace the refracted rays of light and to describe the characteristics of the images formed by a lens.

Lenses are usually made either of glass or clear plastic and are ground and polished to have spherical surfaces. Several varieties of two basic forms of lenses exist: *converging lenses*, those thicker at the center than at the edges, and *diverging lenses*, those thinner at the center than at the edges (Figure 21.1). Occasionally lenses are made with cylindrical or even other surface contours for special purposes; these are not discussed here. Many common lenses, especially those in cameras, are not simple single pieces of glass, but are *compound lenses* made by cementing many individual lenses together with a transparent glue that has a similar index of refraction to that of the glass. We discuss these later, focusing first on simple lenses, those that have negligible thickness compared to their diameter. These are known as *thin lenses* and the equations we introduce are limited to such lenses.

Let's first consider a double convex lens shown in Figure 21.2. Incident rays parallel to the optic axis will be bent by refraction at both surfaces of the lens toward the optic axis (Figure 21.2a). For a thin lens there is a common point, the focal point F,

J. Newman, *Physics of the Life Sciences*, DOI: 10.1007/978-0-387-77259-2_21,

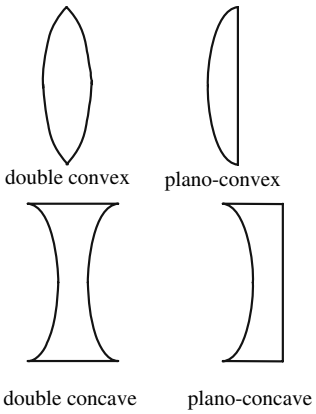


FIGURE 21.1 *Converging* (upper) *and diverging* (lower) *lenses.*

at which all these rays cross the optic axis, the straight line passing through the lens center and traveling perpendicular to both surfaces. The distance f from this point to the center of the lens is called the *focal length* of the lens and is the same distance to either side of the lens. That is, if the lens is rotated 180° about a vertical axis it will focus light at the same point. The focal length of a thin lens can be shown to be related to its radii of curvature R_1 and R_2 on each of its sides and its index of refraction n by the *lens-maker's equation*

$$\frac{1}{f} = (n-1)\left(\frac{1}{R_1} + \frac{1}{R_2}\right). \tag{21.1}$$

Note that this equation defines a single focal length for a lens, regardless of the side facing the incident light, even if the two radii of curvature are different. In this equation the radii are taken as positive if the surface is convex and negative if concave (discussed below). Note that a plane surface has an infinite radius of curvature.

Example 21.1 A plano-convex lens of refractive index 1.52 has a radius of curvature of 5 cm. First find its focal length. Suppose that a second plano-convex lens with the same index of refraction is cemented to the first along their planar faces. What radius of curvature is needed on this second lens to produce a net focal length 1/4 the value of the focal length of the first lens?

Solution: According to Equation (21.1) the focal length of the first lens is $[0.52/5]^{-1} = 9.6$ cm. The second lens must have a radius of curvature R, such that $(9.6/4) = [0.52(1/5 + 1/R)]^{-1}$. Solving for R, we find that $R = 1.7$ cm.

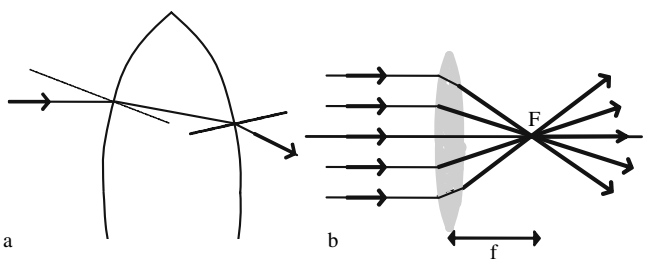
For a double convex (or any converging) lens, if an object O is to be imaged in the lens, we can trace three characteristic rays to locate the image. For the object shown in Figure 21.3, those three rays from the object arrowhead are: (1) a ray parallel to the optic axis that will be focused through the focal point on the far side of the lens (in red); (2) a ray passing through the focal point on the same side of the lens that, on passing through the lens, will be refracted to lie parallel to the optic axis (in blue); and (3) a ray passing through the center of the lens that will continue undeviated (this occurs because both sides of the lens are essentially parallel; the negligible thickness of the lens eliminates any parallel displacement of the ray; shown in green). The image of the arrow tip representing the object can be determined by the common point at which these three rays cross (of course, any two of these will cross at the image point). Then the entire image I is known since the ray along the optic axis passes straight through and images the tail of the arrow. Note that in this case the image is upside down, or inverted, and smaller in size.

In place of raytracing we can derive an equation that will allow us to find the image location as well as its lateral magnification. With the various distances defined in Figure 21.4, the derivation assumes paraxial optics (with incoming rays making small angles with the optic axis). There are two sets of similar right triangles. The first set (shown in green hatched lines) consists of one formed with the object height and distance as legs (with hypotenuse OC) and the other with the image height and distance as legs (and hypotenuse IC). From the similarity of these we have

$$\frac{h}{h'} = \frac{s}{s'}, \tag{21.2}$$

and a second set of similar triangles (one shown in red hatched lines and the other having F as a common vertex) yields

FIGURE 21.2 *(a) Refraction at convex lens surfaces tends to bend light toward the optic axis. (b) The focal point of a double convex thin lens.*



$$\frac{h}{h'} = \frac{f}{s' - f}. \tag{21.3}$$

Combining the two equations, we have that $s/s' = f/(s' - f)$, and after cross-multiplying and dividing both sides by the product $(s\, s'\, f)$ and simplifying the result (try it!), we find the *lens equation*

$$\frac{1}{s} + \frac{1}{s'} = \frac{1}{f}. \tag{21.4}$$

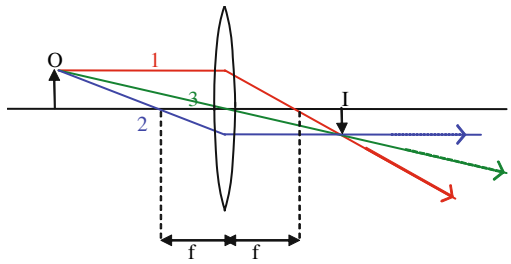


FIGURE 21.3 *Raytracing with a thin converging lens. The image lies at the intersection of the three numbered rays (see text for explanation of the construction).*

We can also invert Equation (21.2) to find an expression for the lateral magnification m

$$m = \frac{h'}{h} = -\frac{s'}{s}, \tag{21.5}$$

where the minus sign was introduced so that an inverted image has a negative magnification, and an erect image has a positive magnification.

In the case worked out above using ray diagrams, the object distance is greater than the focal length and we can see from Equation (21.4) that then because $1/s < 1/f$, the image distance s' is positive, indicating that the image is real and that the image will be inverted (because $m < 0$ in that case). If $s' > s$ the image will be magnified, whereas if $s' < s$ the image will be smaller than the object. It is easy to see that the dividing line occurs when $s = 2f$, because then $s' = 2f$ as well and the image will be "life-size," but still inverted. If $s > 2f$, then $f < s' < 2f$ and the image will be smaller, whereas if $f < s < 2f$, the image will be magnified.

The reciprocal of the focal length for a lens is called its *power* P, where

$$P = \frac{1}{f}. \tag{21.6}$$

Lens power is measured in reciprocal meters which are called diopters (D). Thus, the shorter the focal length of a lens is, the stronger its power. The diopter unit is mainly used in coding eyeglass lenses, a topic we return to in the next section.

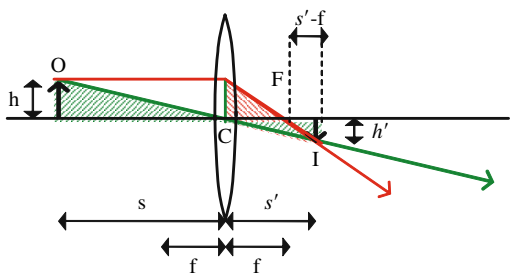
At this point, if we generalize Equations (21.4) and (21.5) using a set of sign rules, these equations are then valid for all thin lenses no matter what the configuration. These rules are given in Table 21.1. We illustrate their use with a few examples.

Table 21.1 *Sign Conventions for Thin Lenses*

Quantity	*Convention*
s	+ If object in front* of lens
	− If object behind lens
s'	+ If image behind lens
	− If image in front of lens
h, h'	+ If erect
	− If inverted
R_1, R_2	+ If surface is convex
	− If surface is concave
f	+ If converging
	− If diverging

* Front and back are with respect to incident light; that is, the front of the lens faces the incident light.

FIGURE 21.4 *Geometry to derive lens equation and magnification (Equations (21.4) and (21.5)).*



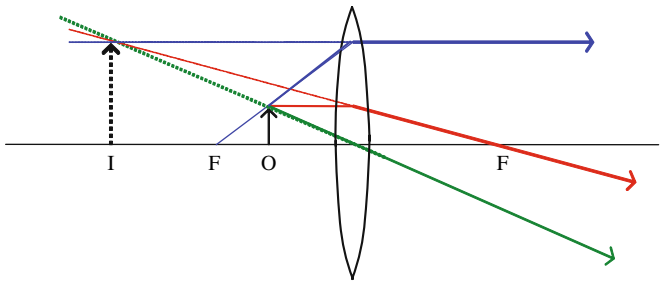


FIGURE 21.5 *Raytracing when the object O is within one focal length of a converging lens. The image I is virtual, erect, and magnified.*

First consider the situation in Figure 21.5 where the object is closer to the lens than one focal length. In this case, Equation (21.4) predicts that s' will be negative because $1/s > 1/f$. What does this mean in terms of the image? The figure shows raytracing in this case. It is clear that the focused rays do not converge and that therefore there is no real image, no place at which a screen can be put to see an image. On the other hand, a viewer on the far side of the lens from the object looking back through the lens will see the rays appear to emanate from a (virtual) image behind the lens, to be larger than the object, and to be erect. That the image is larger and erect follows from Equation (21.5) because s' is greater than s and is negative, making $m > +1$ (note the agreement with our sign conventions in Table 21.1). As the object approaches the focal point, the virtual image recedes to larger distances and is magnified to ever greater size. You may recognize this application of a lens as a magnifying glass (Figure 21.6). We discuss this situation further after we take a look at the eye in Section 3.

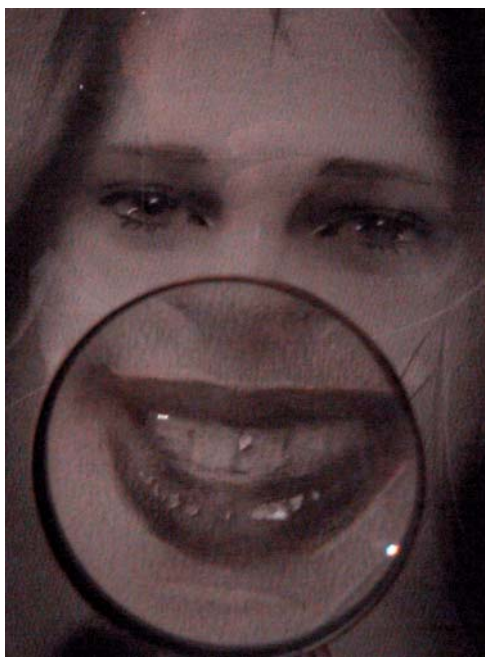
As a second application of the lens equation consider the diverging lens shown in Figure 21.7. According to the lens-maker equation, because the radii of curvature are both taken as negative, so is the focal length. Therefore, no matter where an object is placed on the left of the lens in the figure, according to the lens equation, the image will always be virtual with the object appearing smaller and erect. This is so because with $1/s > 0$, when subtracted from $-1/|f|$, we have that

$$\frac{1}{s'} = -\frac{1}{|f|} - \frac{1}{s}$$

so that we must always have $s' < 0$ and $|s'| < s$. Figure 21.7 shows the raytracing diagram for one such situation.

As a final case, we examine the problem of where to put a converging lens in order to image an object on a screen when the total object to screen distance is fixed at a distance L. In that case $(s + s') = L$, and we must find the possible lens locations, or the possible individual s and s' values. Writing the lens equation as

FIGURE 21.6 *A happy magnifying glass.*



$$\frac{1}{s} + \frac{1}{s'} = \frac{s + s'}{ss'} = \frac{1}{f}, \quad or \quad ss' = fL,$$

we need to solve for possible s and s' values. Substituting for $s' = (L - s)$ into the above, we have

$$s(L - s) = fL \quad or \quad s^2 - sL + fL = 0.$$

Solving the quadratic equation, there are two possible solutions given by

$$s = \frac{L \pm \sqrt{L^2 - 4fL}}{2} \quad or \quad s = \frac{L}{2} \pm \frac{L}{2}\sqrt{1 - \frac{4f}{L}},$$

as long as $f < L/4$ or $L > 4f$. To each of these values of s, let's call them s_+ and s_-, both of which are positive, there corresponds a value of s' $(s' = L - s)$ and a magnification $m = -(s'/s)$. Because the two values s_+ and s_- add up to L, it is clear that the two solutions are $s = s_+$ and $s' = s_-$ on the one hand and $s = s_-$ and $s' = s_+$ on the other. In the first case, the lens is closer to the screen than to the object and the magnification is less than 1. The image will be inverted and reduced in size and, of

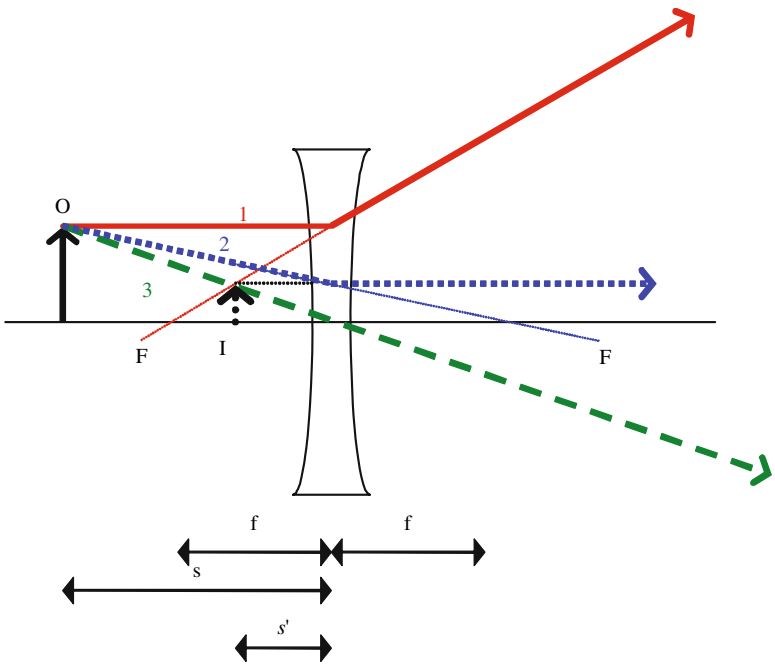


FIGURE 21.7 *Raytracing for a diverging lens to form the image I of an object O. Ray 1 is parallel to the optic axis and diverges as if it originated at the focal point; ray 2 is aimed at the focal point on the other side of the lens and emerges parallel to the axis; ray 3 passes undeviated through the lens center. The virtual image is found at the extrapolated location where these rays cross.*

course, real because it is actually formed on a screen. In the second case, the lens is closer to the object and the real image is enlarged but still inverted (Figure 21.8). With a handheld magnifying glass this is an easy and interesting experiment to try.

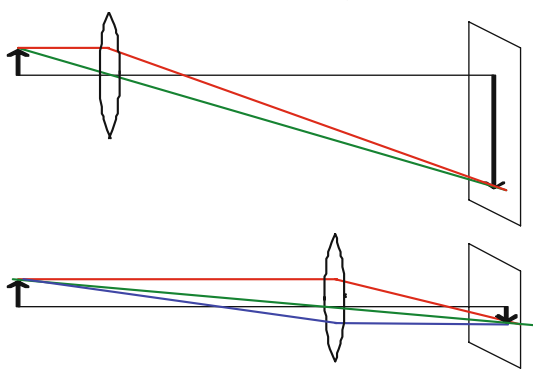
Example 21.2 A 0.2 cm tall object lies 10 cm from a 25 cm focal length magnifying glass. Find the location, magnification, and size of the image. Is it erect or inverted? Real or virtual?

Solution: By direct substitution into the lens Equation (21.4), using $s = 10$ cm and $f = 25$ cm, we find that $s' = -16.7$ cm. Because $s' < 0$, we know that the image is on the same side of the lens as the object and therefore a virtual image. The magnification is given by $m = -s'/s = 1.7$, so that the object appears to be $(0.2)(1.7) = 0.34$ cm tall and erect.

As was mentioned above, many lenses are compound or thick lenses composed of multiple lenses cemented together, whereas other optical systems may consist of multiple individual thin lenses. Situations in which there are multiple thin lenses can be handled using the formalism of this section. One begins by finding the image of the object in the first lens (closest to object) and then simply treats this image as the object to be imaged by the second lens, and so on. Using the same consistent sign convention given in Table 21.1, such problems can be analyzed without any new concepts. For example, if the two (thin) lenses are in contact, then we can derive a simple formula for the overall focal length of the combination as follows. Writing the lens equation for the first lens we have that

$$\frac{1}{s_1} + \frac{1}{s_1'} = \frac{1}{f_1},$$

FIGURE 21.8 *The two solutions to imaging an object on a screen a fixed distance away from an object.*



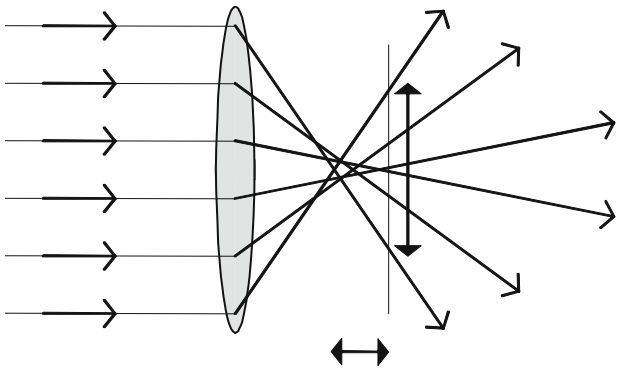


FIGURE 21.9 *Spherical aberration. Parallel rays far off-axis will focus at different distances along the optic axis* (horizontal double-headed arrow)*. The image in the paraxial focal plane, shown by the vertical line, will therefore be blurred laterally* (vertical double-headed arrow).

with a similar equation for the second lens,

$$\frac{1}{s_2} + \frac{1}{s_2'} = \frac{1}{f_2}.$$

But, using the first image as the object for the second lens means that $s_2 = -s_1'$, so that on adding the two equations together we find that

$$\frac{1}{s_1} + \frac{1}{s_2'} = \frac{1}{f_1} + \frac{1}{f_2}.$$

This equation can be interpreted as treating the two lenses in contact with each other as a single lens with a combined focal length f given by

$$\frac{1}{f_1} + \frac{1}{f_2} = \frac{1}{f}. \tag{21.7}$$

Most lenses in optical devices are, in fact, compound lenses designed to compensate for aberrations and so Equation (21.7) tells how to find the net focal length of the compound lens. We use this idea to analyze the compound microscope in the next section. But what is the purpose of cementing multiple lenses together?

All lenses suffer from various defects in the quality of the image they produce. Collectively these are termed lens aberrations. We can distinguish two classes of aberrations: monochromatic, those involving a single color, and chromatic, due to the dispersion of the lens material, refracting different wavelengths (colors) differently due to a variation in index of refraction. There are five major monochromatic aberrations, all of which distort the imaging of a single point of the object to a single point of the image. One of these, spherical aberration, is simply due to the spherical lens curvature giving rise to distortion when incident rays are far from the optic axis. In particular, as shown in Figure 21.9, parallel rays from a distant point source arriving at different distances from the optic axis will be imaged at slightly different points on the axis, resulting in a blurred image of the point object. Limiting the accepted rays to paraxial rays close to the optic axis with a stop, or aperture, can reduce spherical aberration. The four other monochromatic aberrations have to do with off-axis imaging and examples of their images are shown in Figure 21.10.

1. Coma with the comet Hyakutake showing its shape.

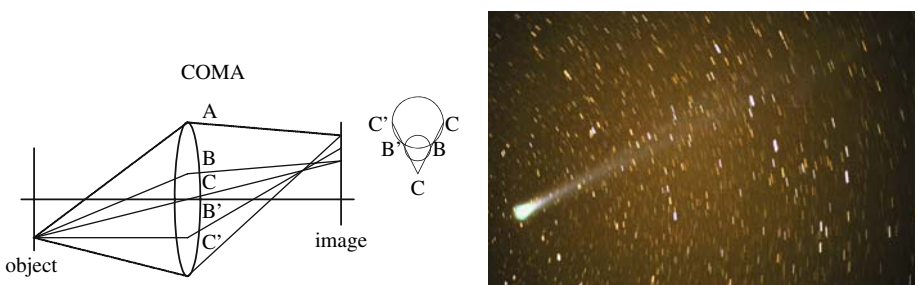
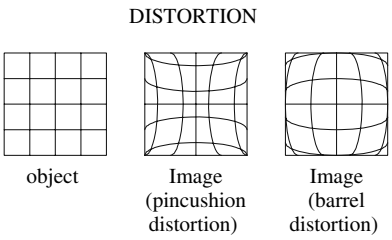
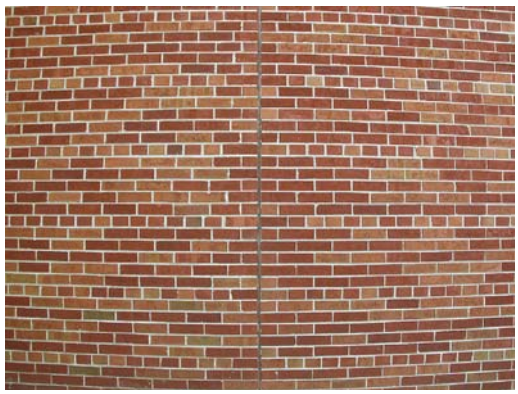


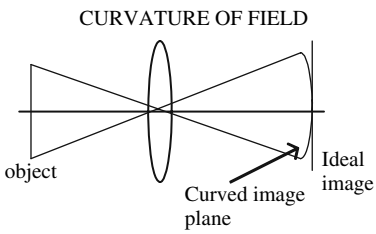
FIGURE 21.10 *Continued*

2. Distortion with an example of barrel distortion from a wide-angle camera shot of a brick wall.





3. Curvature of field with the painting "Anna's Bedroom" by Scott Kahn illustrating this aberration.



4. Astigmatism with a painting illustrating this aberration.

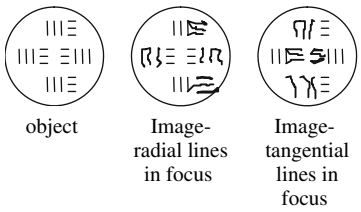


FIGURE 21.10 *The four other monochromatic aberrations, other than spherical aberration, with schematics and example photos.*

Chromatic aberrations due to the dispersion of glass result in different focal points for each color (Figure 21.11). The shorter wavelength light (violet end of the visible spectrum) experiences a larger index of refraction and is therefore refracted more and brought to a closer focal point. Compound lenses made from a converging and a diverging lens of different index of refraction glasses are designed to minimize chromatic aberration and are known as achromatic lenses (Figure 21.12). The longer optical path

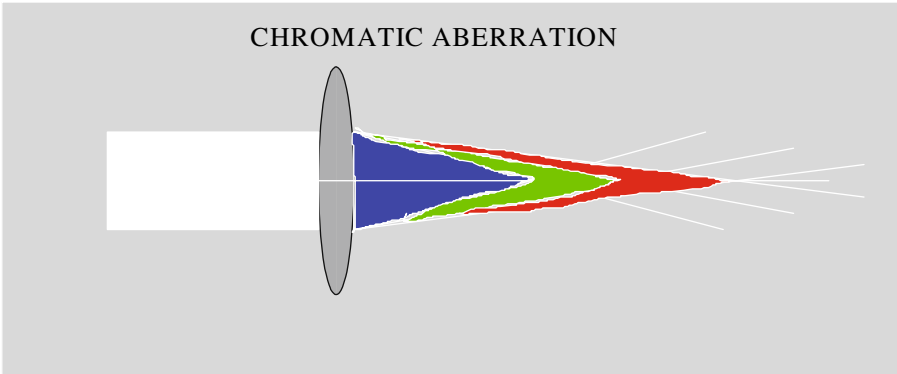


FIGURE 21.11 *Chromatic aberration, showing the effect of dispersion on the focusing of different colored light in a white light beam. Shorter wavelengths experience greater refraction due to the higher index of refraction.*

of the longer wavelength light in the diverging lens seen in the figure compensates for the smaller index of refraction at these wavelengths and brings the various colors to a common focus. All good quality cameras are made with achromatic lenses, the better ones having very thick multiple lenses to minimize other distortions as well.

2. THE HUMAN EYE

Eyes are our visual window to the world. They are complex structures that are capable of very high resolution, are extremely sensitive to light, capable even of detecting single photons, can give color and depth perception, and adjust to focus on objects from as close as 10 cm, in young eyes, to "infinity." In this section we first describe the structure of the eye with the aim of relating its anatomy to its functioning, as well as to some diseases. Then we focus on the retina and the visual pigment to describe in some detail the actual transduction of photon energy to an electrical response in the optic nerve.

A schematic cross-section of the human eye is shown in Figure 21.13. Starting from the outside, the external covering of the eye consists of three layers. Most of the outermost layer is the sclera, the white of the eye, a tough fibrous layer containing nerve endings but no blood vessels. The sclera covers about 85% of the eyeball, roughly a 2.5 cm sphere, but the front portion consists of a transparent 12 mm diameter cornea with an index of refraction of about 1.38. The *cornea* is the most refracting surface in the eye, with the largest index transition from $n = 1$ in air. The next two inner layers are the choroid, filled with pigments and blood vessels, and the *retina*, the site of photon detection. Neither of these layers extends into the cornea region (see Figure 21.13).

At the front end of the eye behind the cornea is a liquid-filled chamber with the aqueous humor, which is continually drained and replaced, bounded also by the lens and the iris. A buildup of pressure in this region can produce a condition known as glaucoma, which can lead to blindness. The *lens* is a double convex lens made from a crystalline array of 25% protein and 10% lipids, having an index of refraction of about 1.42. It is one of the few parts of our bodies that are preserved without any turnover of their cells. With age, or disease, the lens loses its perfect crystallinity and develops defects that scatter light. These are known as cataracts and, when sufficiently large, can adversely affect vision by "clouding" the eye, or scattering light just as if you tried to view the world through a thin layer of milk. The shape of the lens is controlled by ciliary muscles that can change its focusing ability in a process known as *accommodation.* Normally, without any shape change of the lens, we can focus on objects from about 20 feet to infinity. This ability is due to the finite thickness of the photon detection region allowing light

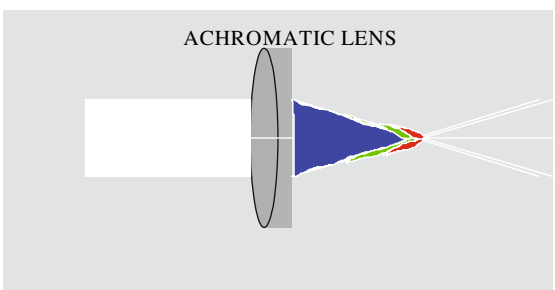


FIGURE 21.12 *Using an achromatic lens (compound lens corrected for chromatic aberration) there is much less chromatic blur.*

to be focused at slightly different distances (see below). To see objects closer than that distance, the eye cannot remain relaxed, but the lens must change shape, thickening to give a tighter focus.

The iris serves as an adjustable aperture and is pigmented, giving the eye its characteristic coloring. The central opening, known as the pupil, is the photon entrance path. Filling the eyeball is a gel-like material, the vitreous humor, which is more or less permanent. Six pairs of muscles control the movement of the eyeball in its socket, allowing us to focus images of interest on the highest acuity region of the retina, the fovea. This region of the retina, also known as the macula, has only cones, the photon receptor cells responsible for color vision, with each cone having a direct connection to a different optic neuron, or nerve cell. The macula therefore has the highest spatial resolution on the retina; outside the fovea there are roughly 10 cones per neuron connection or 125 rods, the other type of photon receptor, per neuron. These neurons collect in the optic disc, creating a blind spot with no visual pigment, and lead to the optic nerve bundle.

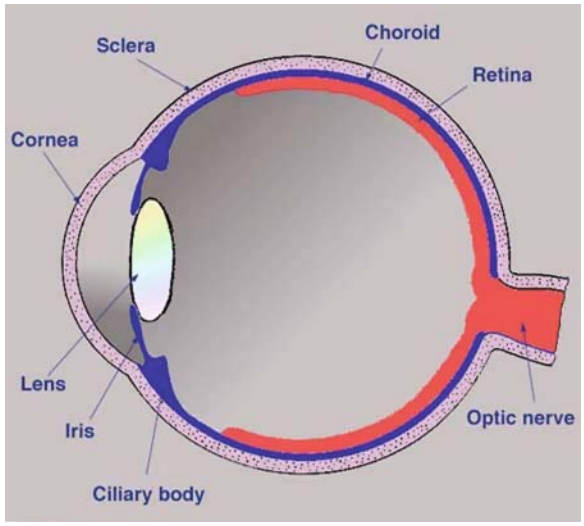


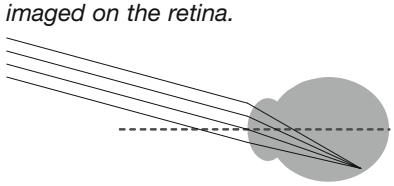
FIGURE 21.13 *The human eye in cross-section.*

Before we consider the photon detection process on the retina in more detail, consider the overall optical arrangement of the eye. As shown in Figure 21.14, parallel rays of light (from an object at "infinity") are focused by the relaxed lens to a point on the retina some 2 cm behind the lens, in fact at the central fovea which lies on the visual axis. An object at a large but finite distance is focused onto the retina as an inverted image. Our brain interprets this inverted image as erect. The total equivalent lens of the eye is a thick lens system, composed of the cornea, aqueous humor, and lens, subject to all of the aberrations mentioned in the last section. One function of the iris is to reduce the aperture size to limit incoming rays to be paraxial, thus reducing aberrations.

Often the eyeball is either elongated along the visual axis (myopia, or nearsightedness) or shortened in that direction (hyperopia, or farsightedness). A third defect in which the cornea is not spherical, but oval in shape, having different focusing properties along two different orthogonal directions, is known as astigmatism. All three of these defects can be corrected for by placing lenses in front of the eye (either as eyeglasses or contact lenses). Figure 21.15 shows ray diagrams for each of these, together with their corrections through the use of a lens. In myopia (top), parallel rays are brought to a focus in front of the retina (hence the name nearsightedness), blurring the image on the retina. By using a diverging lens, the image can be formed on the retina. The worse the myopia, the greater the power of the corrective lens needed. In hyperopia (middle), parallel rays would be focused behind the retina (far-sightedness) and so have not yet converged to form a clear image on the retina. A converging lens will move the focal point onto the retina. Again, the worse the eye's vision is, the stronger power corrective lenses needed. Inexpensive "reading glasses," simply matched converging plastic lenses, are sold in various diopter ratings for several dollars and are usually adequate for reading purposes. Astigmatism (bottom) produces a distortion in imaging so that two perpendicular lines cannot be both brought into focus. A cylindrical lens that focuses light along only one axis can correct this defect.

Example 21.3 Suppose the focal length of a person's eye is 3.0 cm when fully relaxed (looking at a distant object). If the person's retina is 3.3 cm behind the eye lens (a nearsighted eye compared to the normal distance of 3.0 cm), what must be the focal length of the corrective lenses so that this person can see "objects at infinity?"

FIGURE 21.14 *A distant object imaged on the retina.*



Solution: Using the thin lens equation with $s = \infty$ and $s' = 3.3\,\text{cm}$, we find an effective focal length of 3.3 cm needed. Because the effective focal length of such a two-lens system (the lens of the eye and the corrective lenses) is given, from Equation (21.7), by

$$\frac{1}{f_{\text{effective}}} = \frac{1}{f_{\text{lens}}} + \frac{1}{f_{\text{eye}}},$$

we can find the focal length of the needed lens to be

$$f_{\text{lens}} = \left(\frac{1}{f_{\text{effective}}} - \frac{1}{f_{\text{eye}}}\right)^{-1} = \left(\frac{1}{3.3} - \frac{1}{3.0}\right)^{-1} = -33.0\ \text{cm}$$

(or in diopters, $(1/-.33\ \text{m}) = -3\ \text{D}$).

There are some interesting points to be made about nearsightedness. For example, if the near point of the normal eye is 25 cm (s_{min}) and the distance to the retina is 3 cm (s'), then the minimum focal length of the normal eye is 2.7 cm (from

$$\frac{1}{f} = \frac{1}{s} + \frac{1}{s'}\Big).$$

Normal eyes produce an inverted real image on the retina that is $m = s'/s = 3\,\text{cm}/25\,\text{cm} = 0.12$ times as large as the object. Let's assume the nearsighted person in Example 21.3 has the same minimum focal length. Then, without corrective lenses, at closest focus

$$\frac{1}{2.7} = \frac{1}{s_{\text{min}}} + \frac{1}{3.3},$$

which leads to $s_{\text{min}} = 14.9$ cm. Nearsighted people can see clearly closer to their eyes than normal-sighted persons; they have a smaller *near point*, which is the closest distance an object can be brought into focus, without corrective glasses. The size of the real image for this case is 3.3 cm/14.9 cm = 0.22, almost twice as large as that of the normal-sighted person! Nearsighted people who are stamp or coin collectors have a great advantage over normal-sighted persons; they often don't need magnifying glasses to see fine details.

Unfortunately for the nearsighted, this close vision advantage vanishes when they put on corrective lenses. Suppose the person above is wearing the −3 D corrective lenses designed for distance vision and is looking at an object 25 cm away. The −3 D lens turns out to create a virtual image 14 cm in front of the lens that serves as the object for the eye's lens, using

$$s' = \left(\frac{1}{f} - \frac{1}{s}\right)^{-1} = \left(-3 - \frac{1}{0.25}\right)^{-1} = -0.14\ \text{m}.$$

But, that is approximately the closest the bare eye lens can focus, therefore with corrective lenses on, the person can no longer see as close as without them. (An object 14 cm away would form a virtual image only 10 cm in front of the glasses, too close to be in focus.) Note also that the virtual image formed by the corrective lenses is smaller than the object by a factor of 14 cm/25 cm = 0.56. This reduction cancels the magnification advantage the nearsighted person enjoyed, too. In fact, because the corrective lens is diverging, it always forms a reduced size virtual image in front of

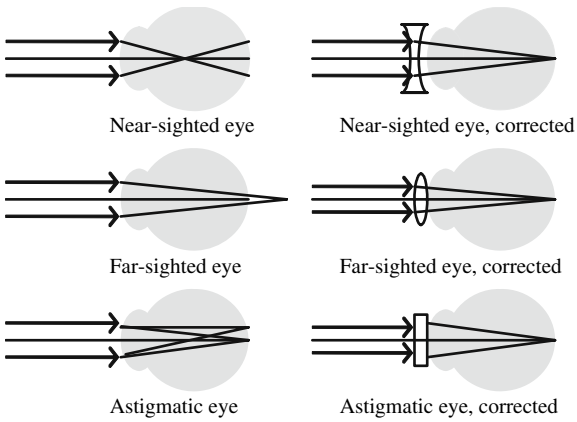


FIGURE 21.15 *Three common focusing problems with the eye and their correction with eyeglasses.*

the real object. The eye uses this image as its object and so nearsighted people with corrective lenses always perceive objects to be both closer to them and smaller, compared with what a normal person sees.

As the eye ages, the lens shape becomes more resistant to change by the ciliary muscles and so the eye cannot accommodate as well to bring close objects into focus. This weakening of accommodation is known as presbyopia. In the young eye, the near point can be as short as 7 cm. With age, the near point moves farther away due to presbyopia, having a mean of about 1 m for a 60 year old individual. Producing a similar effect as hyperopia, this can be corrected with a converging lens. People with myopia will often see an improvement in their vision with age due to presbyopia and will often require bifocal lenses with the lower portion made converging for reading or close vision and the upper portion made diverging for distance vision.

The structure of the retina is shown in cross-section in Figure 21.16. Note the striking fact that incident light must travel through the network of nerve cells (retinal ganglion cells) before reaching the photoreceptors, lying partly immersed in a layer of pigmented cells. Fortunately these cells are transparent, but only about 50% of the light that strikes the cornea gets to the retina, and only about 20% of that gets to the light detecting cells. These cells, the *rods* and *cones*, permanent and not replaced over time, are, however, 100% efficient. Light that is not absorbed by the photoreceptors

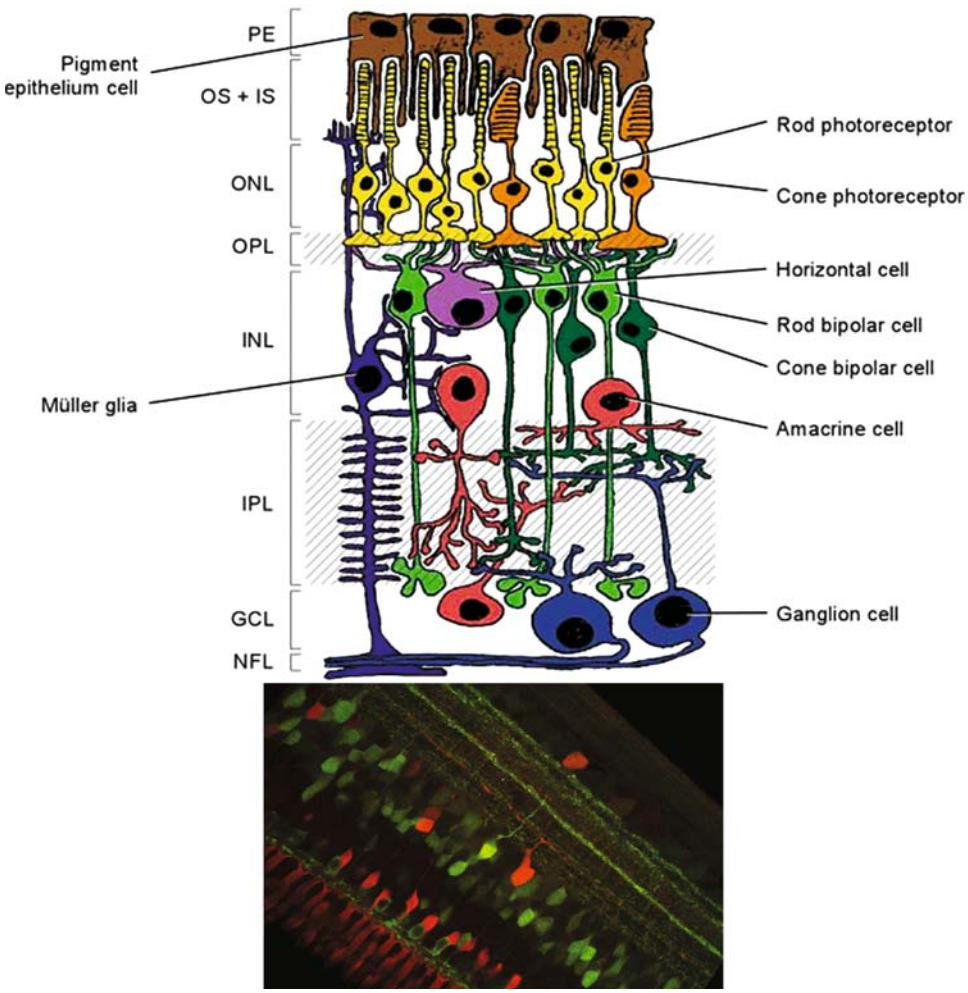


FIGURE 21.16 *The human retina.* (top) *Light passes through a network of nerve cells from the bottom of this drawing before being detected by the rods and cones.* (bottom) *Fluorescence microscopy image of the retina.*

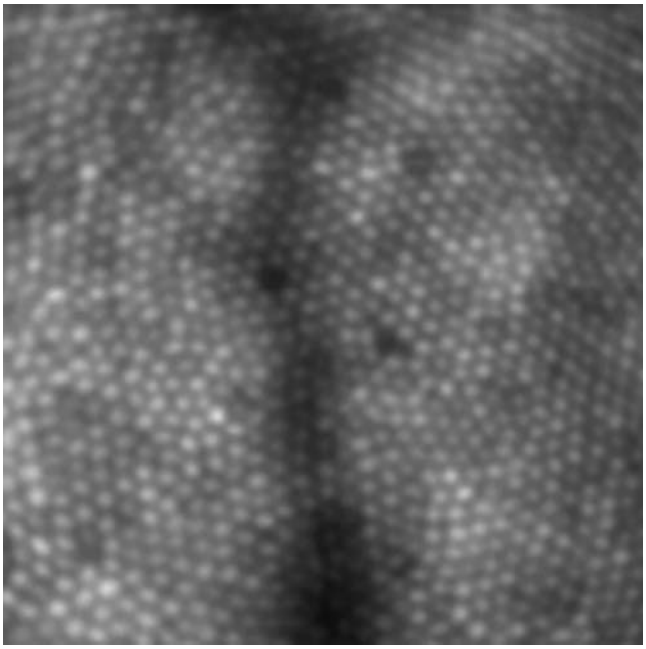


FIGURE 21.17 *Hexagonally packed cone cells at the fovea section of the retina in a living human eye. The image is taken at a location about 300 μm from the foveal center (which is equal to about 1 degree of visual angle) and it is about 150 μm across. The small spots are single cone photoreceptors which, at this location are separated by about 5 μm. The dark shadow is that of a blood vessel which runs above the photoreceptor layer.*

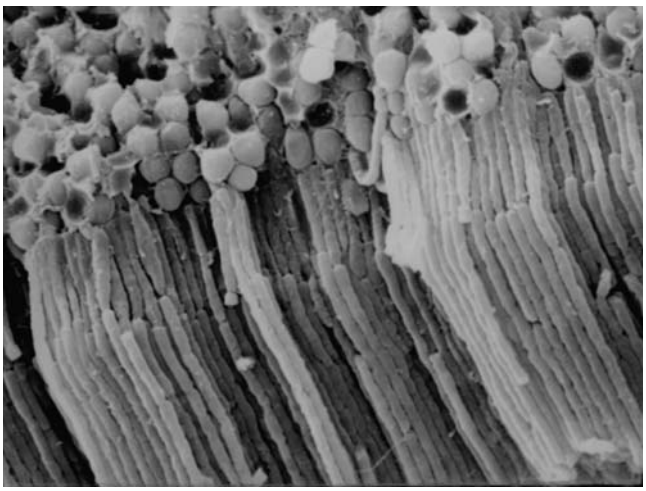
is subsequently absorbed by a layer of pigmented cells to prevent stray reflections of light within the retina. There are about 125 million rods and 7 million cones on the retina distributed such that only cones are at the central fovea, where vision is most acute (Figure 21.17). The rods and cones are named by their shapes, but are somewhat similar in overall structure (Figure 21.18).

There is an inner segment, filled with mitochondria to supply the energy needed for the light transduction, through which the light must also pass. The retina consumes the greatest amount of oxygen per unit weight of any tissue in the body. Inner and outer segments are connected through a thin cilialike portion containing microtubules. Light transduction takes place in the outer segments, the rod outer segments having been studied much more thoroughly than those of the cone. They are each 20 μm long and 2 μm in diameter and contain stacks of rod discs that are membranes containing the visual pigment *rhodopsin*, with about 10^5 rhodopsin molecules per μm^2 of surface area.

Rhodopsin consists of two parts: a protein portion with 348 amino acids, known as opsin, and a smaller hydrocarbon part $C_{20}H_{28}O$, a derivative of vitamin A known as retinal, the light-absorbing portion (Figure 21.19). The structure of retinal is shown in Figure 21.20. There are two possible sterioisomers, or conformations, of retinal: 11-cis-retinal found in the dark and all-trans-retinal that nearly instantaneously forms after the absorption of a single photon. With the advent of femtosecond (10^{-15} s) laser pulses, this first step in the vision process, the isomerization of retinal, has been found to occur within about 500 femtoseconds. Subsequent to this initial conformational change there is a sequence of conformational steps, discovered using pulsed laser spectroscopy, and other events that lead to an eventual electrical signal at the neuron. Each photon absorbed by a rhodopsin leads to the hydrolysis of over 100,000 molecules of cyclic GMP, the crucial signaling molecule in the subsequent transduction. The reduction in cyclic GMP, needed to keep Na^+ channels open in the rod membrane, causes the eventual polarization of the membrane and electrical signal.

The electrical signal that is sent from the retina to the brain over the optic nerve is not simply the sum of all rod and cone firings. Somehow the activity of the many rod and cone interconnections "preprocesses" information about the light falling on them so that a significant part of "seeing" occurs prior to what goes on in the visual cortex of the brain. It takes time to perform the preprocessing of visual information in the rod and cone networks, perhaps 0.1–0.2 s, a time that matches the response time of our nervous system.

FIGURE 21.18 *Cone cells of the human retina.*



We do not see the instantaneous values of the electric fields of the EM light waves, which vary about 10^{15} times per second. Rather we see the effects of electric fields that have been averaged over many cycles of oscillation. In fact, the signal sent from the eye to the brain is not directly related to E fields, but rather to the average intensity, proportional to E^2 averaged over many cycles. Now, rods and cones may detect EM intensity, but not equally well at all frequencies. The retinal molecule acts like a damped oscillator. When driven by the oscillating E field of the light, these molecules vibrate resonantly at different driving frequencies. There are three kinds of cone

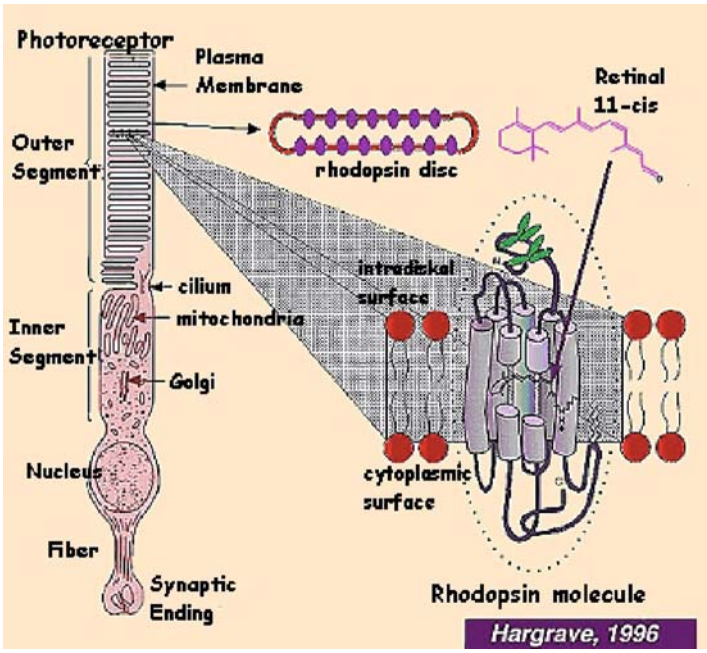


FIGURE 21.19 *A rod cell on left with a model of the seven-helix rhodopsin with the front two helices of opsin cut away to show the retinal molecule at the active site.*

cells (with three different resonant frequencies close to "red," "green," and "blue") and one kind of rod cell, with a resonant peak response between "green" and "blue."

The absorption spectrum of rhodopsin shown in Figure 21.21 indicates that dark-adapted rods are most sensitive (have their strongest absorption) in the green-blue at 500 nm. In strong light this shifts slightly to 550 nm, but with a single absorption spectrum rods are not able to distinguish all of the different colors we can see in bright light. Cones are much less sensitive to light (the figure does not show this because the absorption peaks are all normalized); in fact the cones effectively "turn off" in dim light. In dim light there is little distinction between colors; everything appears gray.

In bright light, the cones take over. They have three different types of visual pigments, each having a maximum absorption at a different visible wavelength, corresponding to a different color. It has long been known that (almost) any light color can be represented as a sum of three "primary" light colors: red (R), green (G), and blue (B). The RGB system is the basis for color TV, for example. On a color TV screen, each pixel is divided into an R, a G, and a B subpixel. Three electron beams sweep rapidly across the screen lighting each pixel with a certain amount of R, of G, and of B.

Clearly it is tempting to explain the RGB color system in terms of the three different cone cells. Undoubtedly, there is some connection between the two, but the connection must be fairly subtle, and, as yet, is still not worked out. One reason for this situation is that the "R" cone cells actually have their response maximum at a frequency that is closer to yellow than to red. A second reason is

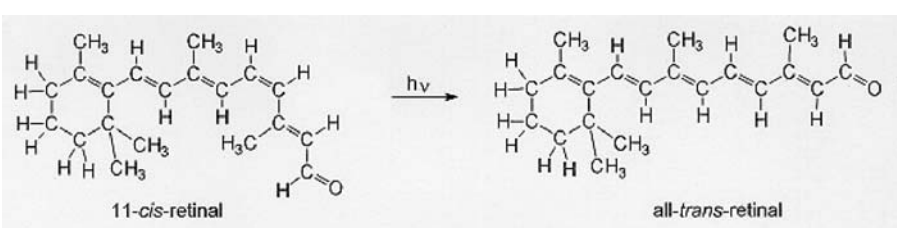


FIGURE 21.20 *The chemical structure of retinal, the light-sensitive portion of rhodopsin.*

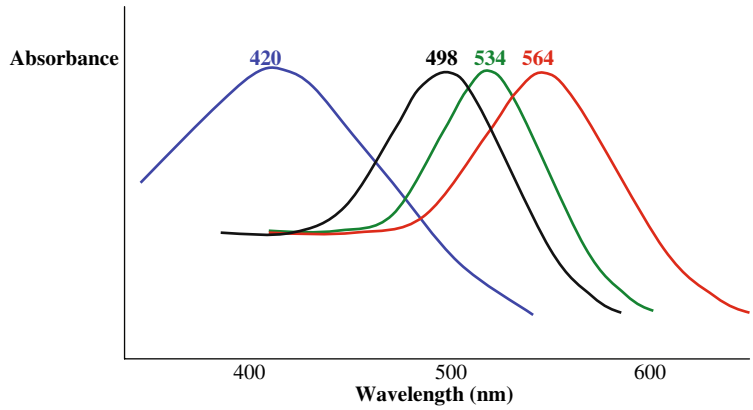


FIGURE 21.21 *The absorption spectrum of the* "red," "green," *and* "blue" *cones and of the rods (shown in* black*).*

that some people only have two of the three cone cells, yet many of them seem to perceive color about as well as people with the normal distribution of cells. Somehow, their brains fill in the missing information. Finally, in people who are red–green color blind, all three cells appear to be present. So, the final word is not in yet.

3. OPTICAL DEVICES: THE MAGNIFYING GLASS AND OPTICAL MICROSCOPE

Optical devices abound in our technologically oriented world, from the supermarket optical scanner to the sophisticated digital video camera. Nearly all of these devices incorporate multiple lenses, except for the simple magnifying glass which we study just below. As we have seen, to analyze imaging problems with multiple lenses we straightforwardly treat the image from the first lens as the object for the second lens, and so on. Similarly the overall magnification is the product of the magnifications from each lens in the combination. We show an example of this in the optical microscope below.

Recall that the near point is the closest distance that an object can be placed from the eye and remain in focus. If you want to see an object in more detail, you must bring it even closer to your eyes. In that way the image of the object on the retina will be enlarged. This allows more detail to be seen, because the image is then spread out over more detection sites increasing the spatial resolution on the retina, limited ultimately by the density of nerve cell connections. The unaided human eye thus has a limited ability to see detail because the near point limits our ability to bring objects as close as we might like to increase the size of a focused image on the retina. Figure 21.22a shows a small object at the near point and the image formed on the retina. If we take the near point to be 25 cm, a typical value, then the angle θ subtended by the object is equal to $\theta = h/25$ cm as shown in the diagram.

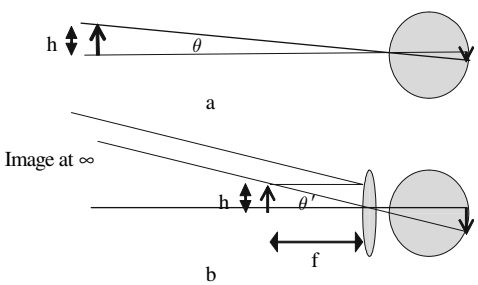


FIGURE 21.22 *(a) An object at the near point of the eye subtending an angle* θ*. (b) An object viewed through a magnifying glass when placed at its focal point, now subtending an angle* θ′*.*

To increase the image size even further on the retina, but still have the image in focus, a magnifying glass (convex lens) is needed. The converging lens increases the focusing ability of the lens of our eye and allows us to bring the object closer to our eye while still keeping the image in focus. The angular magnification, or magnifying power, is defined in terms of the increased angle subtended by the object as compared to that at the near point of the unaided eye (see Figure 21.22)

$$m_\theta = \frac{\theta'}{\theta}. \tag{21.8}$$

Using a magnifying glass with a relaxed eye focused at infinity and with the eye directly behind the magnifier, a virtual image at infinity is formed when the object is placed at the close focal point of the magnifier. Under these conditions $\theta' = h/f$ as shown in the diagram and we can write the angular magnification as

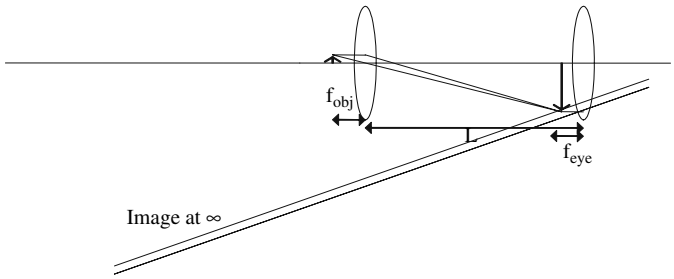


FIGURE 21.23 *Optics of a simple compound microscope. The object is placed just outside the focal point of the objective lens so that an enlarged, inverted real image is formed just inside the focal point of the eyepiece lens. Its image is a further enlarged virtual image viewed at infinity by a relaxed eye.*

$$m_\theta = \frac{\left(\frac{h}{f}\right)}{\left(\frac{h}{25\text{ cm}}\right)} = \frac{25\text{ cm}}{f}. \tag{21.9}$$

The smaller the focal length of the lens, the greater is the magnification seen by the eye. The exact position of the object relative to the focal point of the magnifier turns out to be unimportant, only changing the magnification by a small amount. The eye accommodates these small changes to make the image clear on the retina resulting in a small increase in magnification. The maximum magnification occurs when the image viewed through the magnifying glass occurs at the near point of the eye.

To obtain higher magnification still, a compound microscope can be used. Figure 21.23 shows a schematic drawing of such a microscope with two lenses, an eyepiece that functions as a magnifying glass and an objective lens that further magnifies the object. The overall magnification is the product of that produced by each lens. The object is placed just outside the focal point of the objective, $s \sim f_{obj}$, so that an inverted real image is formed with a lateral magnification of

$$m_{obj} = \frac{s'}{s} = \frac{s'}{f_{obj}},$$

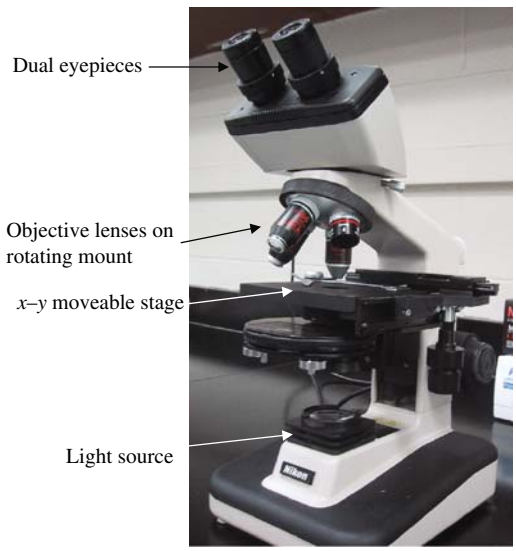
from Equation (21.5). This image then acts as the object for the eyepiece, adjusted to place the final virtual image at infinity, so that the eye can be relaxed as it views the image. In this case, we can write that $s' = (L - f_{eye})$, where L is the distance between the lenses, as shown in the diagram. The overall magnification compared to that at the near point with the unaided eye is then

$$m = m_{obj}\, m_{eye} = \left(\frac{L - f_{eye}}{f_{obj}}\right)\left(\frac{25\text{ cm}}{f_{eye}}\right) \approx \frac{25\text{ cm} \cdot L}{f_{obj} f_{eye}}, \tag{21.10}$$

because f_{eye} is generally much smaller than L, where all distances are given in cm. Usually the short focal length lenses of a microscope are compound lenses designed to eliminate aberrations. Magnifications of over 1000× are readily obtained.

Figure 21.24 shows a photo of a basic compound microscope with its three main features: the built-in light source or condenser, providing a uniform brightness with the use of lenses; a stage, designed to securely hold the sample and usually to move it about in the horizontal plane; and the barrel of the microscope, holding the lenses and usually allowing different objectives with different focal lengths to be used. Microscopy has developed substantially in the recent past. The use of high-sensitivity video cameras and electronic image-processing techniques, as well as the development of several new types of microscopy discussed in Chapter 24, have broadened the versatility of the microscope in studying fundamental processes in biology.

FIGURE 21.24 *A basic compound microscope.*



CHAPTER SUMMARY

Thin optical lenses have a focal length that is given by the lens-maker formula

$$\frac{1}{f} = (n - 1)\left(\frac{1}{R_1} + \frac{1}{R_2}\right), \qquad (21.1)$$

where n is the index of refraction of the lens material and R_1 and R_2 are the radii of curvature of the two faces of the lens. An object located a distance s from the lens will produce an image through a thin lens of focal length f at a distance s' given by the lens equation,

$$\frac{1}{s} + \frac{1}{s'} = \frac{1}{f}. \qquad (21.4)$$

The lateral magnification of the image is

$$m = \frac{h'}{h} = -\frac{s'}{s}. \qquad (21.5)$$

These three equations work under a wide variety of conditions if one uses the sign conventions of Table 21.1.

If two thin lenses are placed in contact, the overall focal length of the pair is given by

$$\frac{1}{f_1} + \frac{1}{f_2} = \frac{1}{f}. \qquad (21.7)$$

The structure, optics, and detection properties of the eye are discussed in Section 2 of this chapter with a discussion of some common vision disorders and their correction with lenses.

A simple magnifying glass of focal length f will produce a virtual image with a magnification given by

$$m_\theta = \frac{25\text{ cm}}{f}, \qquad (21.9)$$

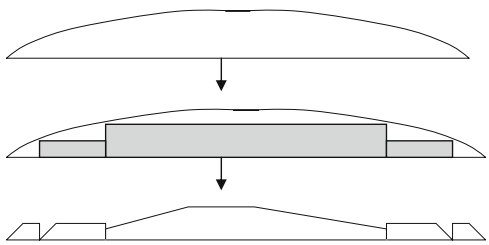
where 25 cm is taken as the near point of the eye. Similarly, the overall magnification of a compound microscope, made from an eyepiece and an objective lens, is given by

$$m = m_{\text{obj}}\, m_{\text{eye}} \approx \frac{25\text{ cm} \cdot L}{f_{\text{obj}} f_{\text{eye}}}, \qquad (21.10)$$

where L is the lens separation distance.

QUESTIONS

1. Distinguish carefully between a converging and diverging lens, a thin and a thick lens, and a real and a virtual image. Which combinations do not exist? For example, is there a thin diverging lens that forms a real image?
2. Carefully define all the symbols in the lens-maker equation. In a thin double convex lens made from $n = 1.5$ glass with both sides having 25 cm radii of curvature, what is the predicted focal length?
3. Review the raytracing algorithm for finding the image of a (real) object through a thin lens. Distinguish the four cases of a converging lens with object closer or farther than the focal length, and a diverging lens with the object closer or farther than a focal length. Make a sketch of an example of each case. Classify each case according to whether the image is (erect or inverted), (real or virtual), (magnified more or less than 1×), and (closer or farther than a focal length from the lens).
4. Consider the object–convex lens–real image configuration. If one places one's eye at the image location, one will not see the image; however, if one moves the eye farther back, away from both the lens and the image location, the image will become visible. Why is this?
5. Compare the (real) image of a person formed by a converging lens to the (virtual) image of that person in a plane mirror. In which is up/down reversed, left/right reversed, and the magnification possibly changed?
6. In applications where a large light source is to be focused, such as a lighthouse, stage lighting in a theater, or an overhead projector, both a large diameter and a thick lens are needed. Fresnel, realizing that the refraction occurs at the glass surface, designed a lens (now called a Fresnel lens) that keeps the large curvature and lens size, but collapses the lens down to a nearly planar lens by removing the glass interior. This overcomes the problem of weight, bulk, and cost of such a glass lens. Based on the figure below explain how this lens works.



7. What is the physical origin of chromatic aberration in a thin lens? Of spherical aberration? Which rays are brought to focus closer to the lens: blue or red? paraxial (those parallel to and close to the optic axis) or off-axis rays?
8. Why does closing down the iris reduce aberrations of the lens of the eye?
9. One way to correct the eye for myopia is with laser surgery called photorefractive keratectomy (PRK) in which the cornea is reshaped so that light from a distance object focuses on the retina instead of in front of the retina. What do you think the laser procedure does to the cornea? This surgery will not correct for the age-related presbyopia that leads to the need for reading glasses.
10. Discuss the origin of presbyopia and the need for reading glasses as one ages.
11. Which photoreceptors, rods or cones, give us our most acute vision? Our color vision? Our night vision?
12. Explain in basic terms why the magnification of a microscope (or any two-lens system) is equal to the product of the magnifications of the objective and eyepiece.

MULTIPLE CHOICE QUESTIONS

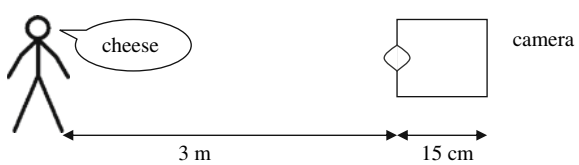
1. A light-emitting object is 10 cm from a thin lens. An upright virtual image is formed 20 cm behind the object. Which of the following is true? The lens has a focal length of (a) +6.7 cm, (b) +15 cm, (c) −15 cm, (d) −20 cm.
2. The distance from the eye's lens to the retina for a given person is 3.0 cm. This person clearly sees an object 27 cm in front of his eye. The focal length of the eye's lens in this case is (a) +2.7 cm, (b) +3.0 cm, (c) +3.4 cm, (d) −3.4 cm.
3. A 10 cm tall object 25 cm from a converging lens has its real image 50 cm from the lens. The object appears to be (a) 20 cm tall and erect, (b) 5 cm tall and inverted, (c) 5 cm tall and erect, (d) 20 cm tall and inverted.
4. A light source with an arrow pointing up is placed at the zero mark on an optical bench. A convex lens of unknown focal length is placed with its center at the 30 cm mark on the bench. A focused image appears on a collector when placed at the 60 cm mark on the bench and nowhere else. What must be true about the image? It is (a) real and inverted, (b) real and upright, (c) virtual and inverted, (d) virtual and upright.
5. The focal length of the lens in the previous question must be (a) −15 cm, (b) +15 cm, (c) +30 cm, (d) +60 cm.
6. Suppose a concave lens is inserted at the 15 cm mark on the bench in question 4. What would you have to do to the collector to find a focused image now? (a) Leave it at 60 cm. (b) Move it to some position between the 30 cm mark and the 60 cm mark. (c) Move it to a position farther out than the 60 cm mark. (d) You can't move the collector anywhere to get a focused image because no real image will be formed by this arrangement of lenses.
7. In drawing a ray diagram for a converging lens with an object farther away than a focal length which of the following is not a correct ray to draw: (a) a ray parallel to the optic axis deflects at the lens to appear to come from the near focal point, (b) a ray parallel to the optic axis deflects at the lens to go through the focal point on the far side of the lens, (c) a ray going through the near focal point deflects at the lens and emerges from the lens parallel to the optic axis, (d) a ray straight through the lens center.
8. A plano-convex lens of crown glass with a radius of curvature of 50 cm has a focal length of (a) 1.0 m, (b) 0.5 m, (c) 2.0 m, (d) −1.0 m.
9. A double concave lens of crown glass with radii of curvature magnitudes of 25 cm and 50 cm has a focal length of (a) 0.33 m, (b) −3 m, (c) −0.33 m, (d) −33 m.
10. Chromatic aberration in lenses is due to (a) dispersion, (b) interference, (c) total internal reflection, (d) varying degrees of absorption of different colors of light.
11. Which of the following best describes nearsightedness? The lens in the eye produces an image of an object (a) 5 m away that would form behind the retina, (b) 5 m away that forms in front of the retina, (c) 10 cm away that would form behind the retina, (d) 10 cm away that forms in front of the retina.
12. Without corrective lenses a woman can see an object clearly no closer than 0.5 m from her face. With corrective lenses she can see the object clearly as close as 0.1 m from her face. When the object is at 0.1 m her corrective lenses must form a (a) real image 0.5 m in front of her face, (b) real image 0.1 m in front of her face, (c) virtual image 0.5 m in front of her face, (d) virtual image 0.1 m in front of her face.
13. A prescription for corrective lenses reads −5 D for each lens. These corrective lenses are (a) diverging with focal length equal to 5 m, (b) diverging with focal length equal to 0.2 m, (c) converging with focal length equal to 5 m, (d) converging with focal length equal to 0.2 m.
14. It is essentially impossible for humans to react to a visual stimulus in less than 0.1 s. This is because that is about how long it takes (a) light to pass from the lens of the eye to the retina, (b) light to make one complete cycle of oscillation, (c) molecules in the cones to complete one cycle of vibration when they are excited by light, (d) firing from the retinal cells to be processed before being sent to the brain.
15. In very dim light you can see an object better by looking at it out of the corner of your eye than straight on. That is because (a) cones are more concentrated on

the periphery of the retina than rods and cones function better in dim light, (b) cones are more concentrated on the periphery of the retina than rods and rods function better in dim light, (c) rods are more concentrated on the periphery of the retina than cones and cones function better in dim light, (d) rods are more concentrated on the periphery of the retina than cones and rods function better in dim light.

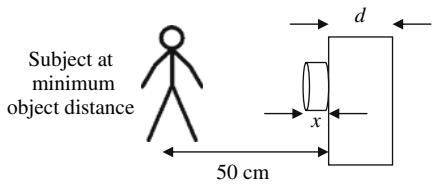
16. Color vision is due to (a) the varying sensitivity of different rhodopsins to different wavelengths of light, (b) three different kinds of rod cells, (c) the dispersion of the lens of the eye, (d) the pigmented epithelial cells.
17. A magnifying glass produces an image that is (a) upright and real, (b) inverted and real, (c) upright and virtual, (d) inverted and virtual.
18. Light entering the eye passes through the following layers in the order (a) lens, cornea, ganglion cells, retina, (b) cornea, lens, ganglion cells, retina, (c) lens, cornea, retina, ganglion cells, (d) cornea, lens, retina, ganglion cells.

PROBLEMS

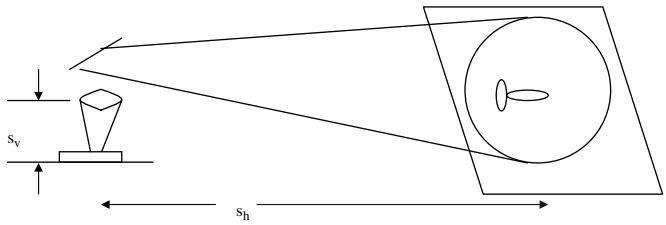
1. Sunlight can be focused on the ground by a lens when it is held 24 cm above the surface. What is the power of the lens?
2. A -5.0 diopter lens is used to magnify an insect when held 12 cm away. Describe the type of image, its position, and lateral magnification. Draw a ray diagram sketch.
3. A pinhole can function as a "lens". Consider a box with a very small hole in one side. The hole admits light from any and all points outside to the inside but from any one point outside only the ray directed at the hole can enter the box. Show with a diagram the image of the object obtained by the light admitted by the pinhole. What is the magnification? Such boxes can be used as cameras, provided they are mounted on a rock-solid surface. Film exposures are typically many seconds or minutes and the image quality can be superb.
4. An old-fashioned box camera has a fixed lens and a depth of 15 cm. Suppose the camera lens is designed to be optimal for taking photographs of objects 3 m away. What is the focal length of the lens?



5. A camera has a lens with adjustable position. The camera depth $d = 4$ cm. Determine the focal length of the lens and the necessary allowable extension of the lens x, in order that the camera be able to take sharp photographs of objects positioned anywhere from 50 cm to infinity, measured from the front surface of the camera body.

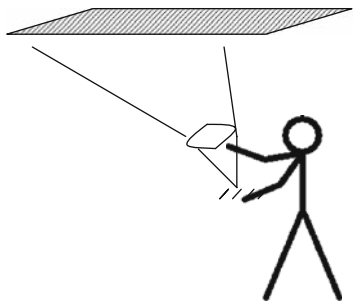


6. How far from its infinity setting must a 35 mm lens be moved so that it produces a sharp image of an object 3 m away?
7. Suppose the image of a creature swimming in a dish of water is projected onto a screen. The distance from the dish to the lens is 36 cm and the screen is 4.5 m away from the mirror and lens. (The mirror simply redirects the light from the lens to the screen. Treat the distance between the lens and mirror as small enough so that it can be neglected in the specification of the image distance. Thus, object distance s_v is 36 cm and image distance s_h is 4.5 m.)



   (a) What is the focal length of the projection lens?
   (b) If the creature swims at 1 cm per second in the dish, how fast does the image move on the screen?
8. A 35 mm film slide projector has a projection lens with a focal length of 135 mm.
   (a) Where should the slide be placed if the projection screen is 3 m away from the projection lens?
   (b) What is the magnification?
   (c) Now for the hard part (judging by the difficulty that even well-educated lecturers have with this one in practice). If we want to get a true (upright, nonreversed) image of the slide on the screen, how should the slide be placed into the projector? Should it be flipped upside down? Should it be reversed left and right? Should the slide be inserted backwards? (What does "backwards" mean here?)
9. A quick and easy way to get an approximate determination of the focal length of a convex lens is to measure the distance from the lens to an image of a light or other bright source some distance away. Suppose one has a lens that in fact has a focal length of 10 cm.

A fluorescent ceiling light with a grill cover is 1.5 m above the lens and a student is able to see an image of the lighting fixture grill on the back of his hand. He declares that the focal length of the lens is equal to the distance between lens and hand. Calculate the actual hand–lens separation distance for a sharp image and show that the error in the value of the focal length determined this way is less than 10%. Error = [(focal length − distance measured)/focal length] × 100%.



10. Consider a convex lens of focal length 20 cm. Calculate the image distance for each of the following object distances: ∞, 4 m, 2 m, 1 m, 80 cm, 60 cm, 40 cm, 20 cm.

11. It is often necessary to convey or relay the image of an object while keeping the image size unchanged (unit magnification). Consider an object for which an image of the same size is desired, exactly 1 m away.
(a) What is the focal length and the location of the single lens that will accomplish this?
(b) The image of the object, using one lens, will be inverted. A relay system using two identical convex lenses will invert the inversion, yielding an upright image. Design such a system for the same initial situation as in the previous part.

12. If a certain microscope has an eyepiece with 12× magnification and it is desired to view a specimen with an overall magnification of 60×, what is the power of the objective that must be used?

13. The magnification of a compound microscope can be slightly improved if the final image is not at infinity but rather at the near point of the eye. Derive the formula for the magnification under this condition. This arrangement tends to produce eye strain because the iris must be under tension to have the lens of the eye constantly focusing at the near point.

14. Tom Cruise catches a reporter shooting pictures of his daughter at his home. He claims the reporter was trespassing. To prove his point, he gives as evidence the film the police took from the reporter. His daughter's height of 0.62 m is 2.89 mm high on the film, and the focal length of the camera lens that the police seized was 210 mm. How far away from the baby was the reporter standing? Could the reporter be trespassing?

22 Wave Optics

In our discussions of geometric optics we completely ignored the fact that light can also behave as an electromagnetic wave. The wave packet picture of photons (introduced in Section 5 of Chapter 19) is compatible with treating light as a collection of particlelike photons that follow the rules of geometric optics only as long as the objects with which light interacts (mirrors, lenses, apertures, surfaces, etc.) are large compared to the wavelength of the light. This justifies our treatment of light thus far as traveling in straight lines except on refraction or reflection at the boundary between two media. Under other conditions, the wave packet will change its spatial extent on interacting with smaller objects and exhibit wavelike properties, some of which are discussed in this chapter.

We begin this chapter by re-examining, in the context of light waves, some concepts introduced earlier in Chapters 10 and 11 for traveling mechanical or sound waves. A major idea is the principle of superposition, as applied to waves overlapping in space and time, which leads to interference effects. Two waves of equal amplitude that are in phase will add constructively to produce a net wave with an amplitude twice as large whereas such waves that are 180° out of phase will add destructively to completely cancel each other. Another general property of a wave is diffraction, or bending, that occurs at an obstacle. These effects are studied for a variety of important geometries and their fundamental implications in limiting resolution in optics are discussed. The next chapter discusses a variety of applications in imaging that stem from these ideas.

1. DIFFRACTION AND INTERFERENCE OF LIGHT

1.1. PRELIMINARIES

We have seen that a monochromatic traveling plane light wave can be represented by a sine wave with a well-defined frequency and amplitude moving with a constant velocity v equal to c/n. Having defined the frequency, the wavelength is dependent on the medium and is given by

$$\lambda = \frac{v}{f} = \frac{c}{nf}, \tag{22.1}$$

and the intensity of the wave is proportional to the square of its amplitude. In a plane wave, all points on the wavefront are in phase and oscillate together (Figure 22.1). Such idealized plane waves can be fairly well represented by a laser beam, as we show in Chapter 25.

When light crosses a boundary between two different optical media, its frequency remains the same but its speed changes and therefore the wavelength of the

J. Newman, *Physics of the Life Sciences*, DOI: 10.1007/978-0-387-77259-2_22,

FIGURE 22.1 *(Nearly) plane waves in the Mediterranean along the Sicilian coast.*

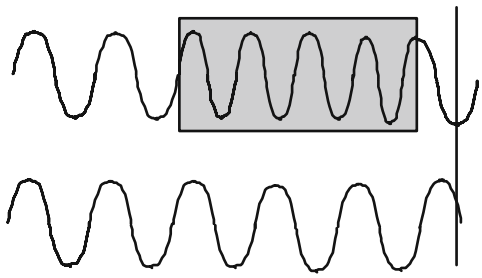


FIGURE 22.2 *Two initially in-phase light waves traveling to the right, showing the optical path difference when one wave passes through a different optical medium with a larger index of refraction*

light will change. As a consequence of this, light traveling through different media will be shifted in relative phase leading to some interesting phenomena. As shown in Figure 22.2, when one beam of light passes through a slab of material with higher index of refraction its wavelength is shortened and after emerging from the slab with the original wavelength, its phase has been shifted with respect to a second reference beam. In traveling some distance d in a particular medium, the number of wavelengths of light within that distance is given by $d/\lambda = nd/\lambda_o$, where λ_o is the vacuum wavelength of the light. Thus for a given λ_o of light, different beams traveling through different media will be shifted in phase by amounts related to the product nd for each media. Each beam arriving at the viewer has traveled a different *optical path*, defined as the distance traveled multiplied by the corresponding index of refraction

$$Optical\ Path = \sum n_i\, d_i, \tag{22.2}$$

where the sum is over the different distances d_i the beam travels in different media with indices n_i to obtain the net optical path traveled. This optical path will determine the relative phase of EM waves traveling through different media. In the case shown in the figure, as long as the two waves originate in phase, their relative phase at the viewer will be determined solely by differences in optical paths. These differences will create optical consequences at the viewer that can be used to learn something about the media through which the beams have traveled as we show shortly.

But how can we ensure that the initial beams are in phase? One way is to generate a plane wave beam of light and then split it into two sources with a beamsplitter device or simply with two apertures. As shown in Figure 22.3, a real extended source of light can, with the use of a small aperture, be used to generate light that has a wavefront that oscillates in phase. Light emanating from the aperture can be imagined to spread in space according to an idea due to Huygens and known as *Huygens' construction*. The method consists of imagining each point on the wavefront as a source of spherical wavelets that spread out in the forward direction at the speed of the wave; the new wavefront consists of the envelope, or tangent, to all the wavelets. This construction for the small aperture is shown to generate a spherical wave. At some large distance from the aperture, the wavefront is approximately plane and this method can be used to obtain a plane wave. Alternatively, a laser beam can be used because, as mentioned above, most laser beams behave as plane waves.

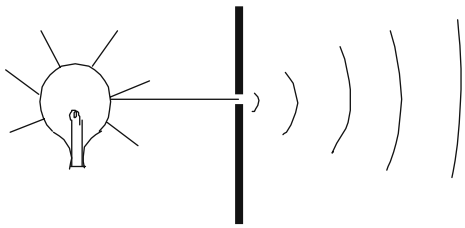


FIGURE 22.3 *An extended source* (light bulb) *and a pinhole aperture used to generate a plane wave far from the source.*

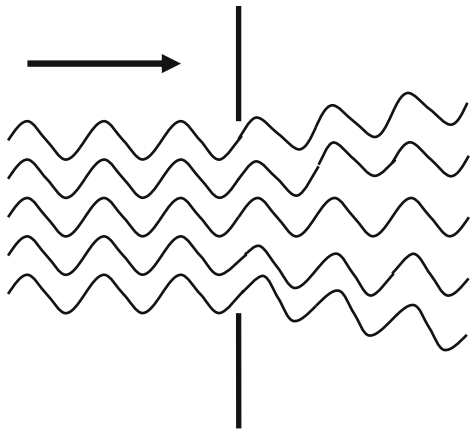


FIGURE 22.4 *Diffraction of a plane wave at an aperture.*

1.2. DIFFRACTION

When a wave meets an obstacle, or hole in an opaque material, and is partially obstructed, Huygens' construction can be useful. If the opening is very large compared to the wavelength of light, the unobstructed portions of a plane wave continue on as a plane wave, with some bending of the light at the wall edges. This bending is known as *diffraction* and occurs with all types of waves. As the opening gets smaller and comparable to the wavelength, the diffraction increases and the wave is spread much more. This is shown schematically in Figure 22.4.

Diffraction of sound is obvious to us: we can hear around corners. Sound has a wavelength comparable to the dimensions of objects around us and thus is easily diffracted through large angles. Visible light, with its wavelength of about 500 nm, is only diffracted near sharp edges or at small openings with micron-sized dimensions. As long as diffraction effects are negligible, light can be understood in terms of geometrical optics. In Section 2 we show some of the effects of diffraction on light transmitted through different slits and in Section 3 we consider the fundamental limitations on resolution due to diffraction.

1.3. INTERFERENCE

As a first example of interference, consider the situation shown in Figure 22.5 in which a light beam is incident on a *thin film* coating a second medium with a different index of refraction. To be specific, suppose that the beam is incident from air onto an organic film (e.g., an oil or gasoline) with index of refraction n and thickness t, coating the surface of water. At each surface there will be a division of the incident intensity into a reflected and refracted beam. If the incident light makes a small angle with the normal then there will be a series of multiple reflections, as shown in the diagram, resulting in the overlap of the reflected beams as seen by a person viewing from the air (we ignore the transmitted beams in what follows).

If we approximate the beams as normal to the surface then those labeled 1 and 2 have an optical path difference given by $O.P.D. = n\,(2t)$. What is the effect of this extra path difference on the relative phase of the two beams at the viewer? If the two beams that we are comparing were as in the situation in Figure 22.2, the optical path difference divided by the wavelength of light in air would be the number of wavelengths shifted between the two beams. If this number were equal to 1 then the two beams would again be in phase, but if it were equal to 1/2 then the two beams would be exactly out of phase.

In the case of reflection from a surface there is an additional effect that enters. If the light is reflecting from a medium with a larger index of refraction, as is the case

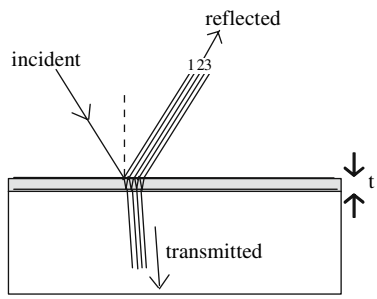


FIGURE 22.5 *Multiply reflected and transmitted light from a thin film.*

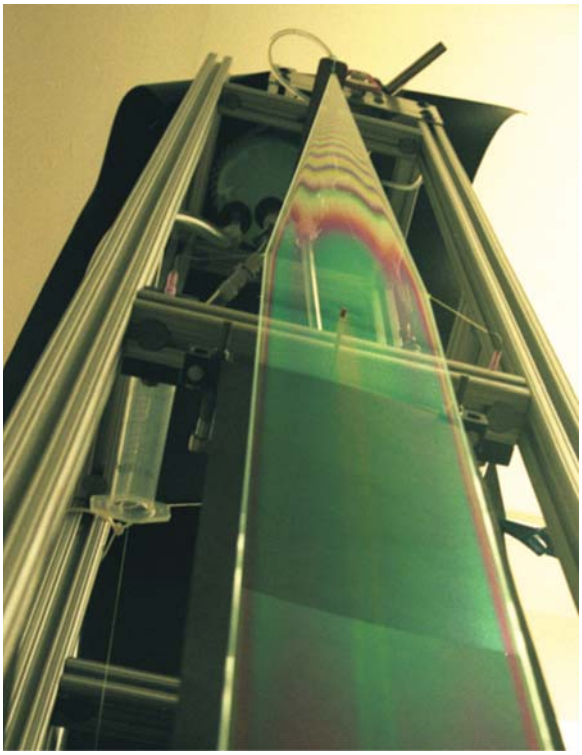


FIGURE 22.6 *Huge soap film showing reflection fringes, used to study flow in thin films.*

from air to the organic film, then there is an additional phase change of 180°, or π radians. Note that this is the same effect we see when a wave on a string is reflected from a fixed end (see the discussion of Figure 10.16). In our thin film example, because the film has a larger index than the lower water layer and air, only ray 1 will have the additional π phase shift; the other reflected beams arise from reflection at the second surface between the film with greater index of refraction and the water or air, and therefore no additional phase shift occurs on reflection. When the optical path difference ($2nt$) is just equal to an integer number of wavelengths of the light, then with the additional half-wave shift, the multiply reflected beams will all be out of phase with ray 1 at the viewer, a condition known as *destructive interference.* Using our expression for the optical path difference, we can write this condition as

$$2nt = m\lambda, \quad \text{(destructive interference)}, \tag{22.3}$$

where λ is the wavelength of light in the incident medium (air) and $m = 0, 1, 2, \ldots$ is an integer known as the order number. When $m = 0$, even with an extremely thin layer of index n material, there is still destructive interference (see below). Although the waves will arrive out of phase, the cancellation will not be complete in general because the intensities of the multiply reflected beams are not equal. With two beams of equal intensity 180° out of phase, complete cancellation would occur leaving no light at all.

If, on the other hand, the optical path difference is equal to a half-integer multiple of the wavelength, then with the added half-wave shift on reflection, the two most intense waves, labeled 1 and 2, will undergo constructive interference

$$2nt = \left(m + \frac{1}{2}\right)\lambda. \quad \text{(constructive interference)}. \tag{22.4}$$

In the case of an oil film on water or a soap film in air illuminated by white light, the reflected light reveals a set of brightly colored *fringes* as shown in Figure 22.6. The different colors arise from constructive interference at the corresponding wavelengths due to variations in film thickness.

Example 22.1 White light illuminates a film of 325 nm optical thickness with index of refraction $n = 1.5$, corresponding to a real thickness of 217 nm. Which visible wavelengths will appear intense on reflection and which will be absent?

Solution: According to Equation (22.3), destructive interference will occur for $\lambda = 2nt = 2(325) = 650$ nm, corresponding to a shade of red light. This is the only visible color that will be totally absent; the next wavelength at which there is total destructive interference is at $\lambda = nt = 325$ nm which is in the ultraviolet. According to Equation (22.4), constructive interference will occur for $2nt = 3\lambda/2$, resulting in $\lambda = 430$ nm and corresponding to violet light, as the only visible wavelength that will be intensified by constructive interference. Other values of m in Equation (22.4) do not lead to visible wavelengths. The resulting overall reflection will make the film appear bluish. Reflected colors can therefore be used as an indicator of film thickness.

Very thin films (with $t < \lambda/10$) will appear to be black because there is a negligible optical path difference and the half-wave shift of the primary reflected

beam tends to cancel all the other reflected light from the second surface where there is no half-wave shift. Artificial membranes made from lipids, formed in a manner similar to soap films, appear black when their thickness corresponds to that of a bilayer and this color change is used as a signature of film thickness.

The ideas just presented are the basis for *reflection-interference microscopy.* This microscope technique can be used to visualize cell-to-surface contacts. A typical sample (Figure 22.7) is a suspension of cells covered with a glass cover slip above which lies immersion oil with an index of refraction matching that of glass. The indices of refraction of the three media involved, the cover glass, n_g, the solvent, n_s, and the cell, n_c, satisfy $n_s < n_c < n_g$, therefore interference on reflection can be used to discern cell–surface adhesion sites. Remembering that an extra π phase shift occurs on reflection from a larger index medium, reflection at the solvent–cell interface produces a phase shift of π radians. In addition to that phase shift, there will be an optical path difference ($2n_s t$) between the reflections at the two surfaces from the solvent layer of variable thickness t (see Figure 22.7). At sites of cell attachment where this layer of solvent is vanishingly thin, there is no additional optical path difference and these sites will appear dark on reflection. In this method, total internal reflection is used to illuminate a very thin layer of the sample using the evanescent waves (see the discussion in Chapter 20 concerning Figure 20.14) and only the attached cells in a very thin surface layer will then be seen and magnified, often using fluorescent light from attached dyes (Figure 22.8).

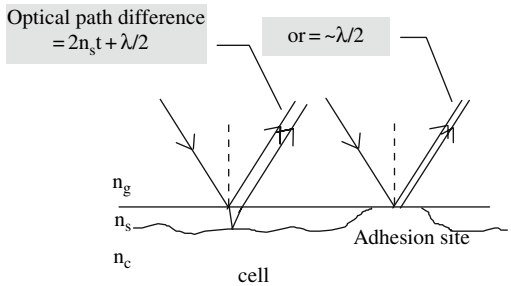


FIGURE 22.7 *A cell, solvent, glass cover slip (from bottom up) sample in a reflection-interference microscope.*

Aside from the beautiful patterns of colorful fringes appearing on reflection from thin films (Figure 22.9), these films can be used in a number of ways to perform optical tasks. Thin coatings on lenses can be used to reduce troublesome reflections in optical systems such as compound lenses with many optical surfaces routinely used for camera and microscope lenses. Such *nonreflective coatings* can be detected by the characteristic faint bluish color of reflected light. Special multilayered coatings on a glass surface can be used as an *interference filter* to select a very limited wavelength (with a range of less than 1 nm) region to be transmitted. Interference filters are particularly useful in spectroscopy. They are also used in laser safety goggles to block only the narrow wavelength range of laser light.

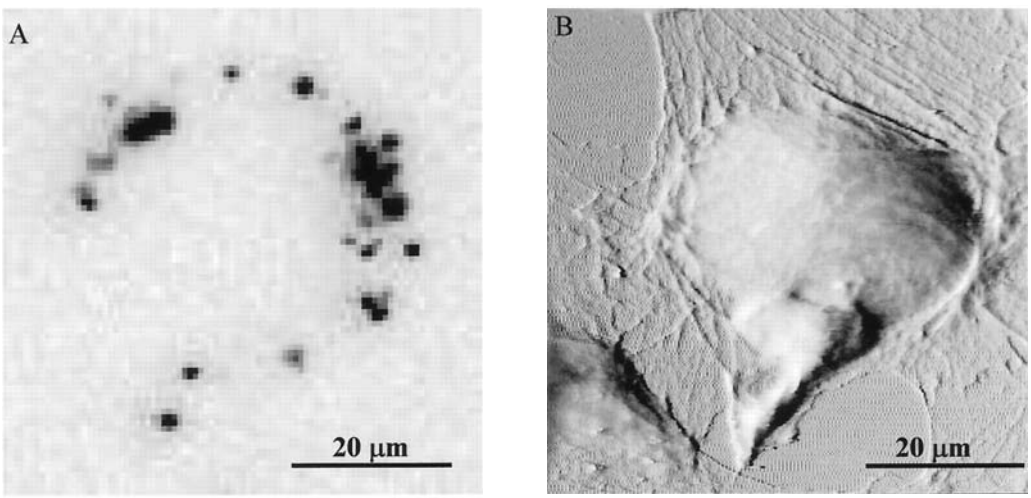


FIGURE 22.8 (left) *Total internal reflection fluorescence microscopy image of a fluorescently labeled cell surface showing focal contacts (darker regions are closer to glass slide with lightest imaged regions about 85 nm from glass).* (right) *Atomic force microscopy (see Chapter 8) image of the same cell, but the opposite surface seen in the left image. The cytoskeleton is visible here showing its contacts with the glass surface.*

FIGURE 22.9 *Thin film interference photo from the kitchen sink.*

2. SINGLE-, DOUBLE-, AND MULTIPLE-SLITS AND INTERFEROMETERS

2.1. SIMPLE INTERFERENCE WITH A DOUBLE-SLIT

Consider the experiment sketched in Figure 22.10 in which a light wave is incident on two closely spaced slits, an arrangement known as *Young's double-slit experiment*. Light passing through either slit will be diffracted, spreading out in waves that overlap on a screen some relatively large distance away. If the light waves at the two slits begin in phase, when the two waves arrive and overlap at some point on the screen their phase difference is simply related to the difference in the optical paths they traveled from either slit. In this case because the medium is simply air, the optical paths are the physical distances traveled and we can calculate the path difference using the geometry shown in the inset in the figure.

At positions on the screen where the optical path difference (shown in the figure to be $d \sin \theta$) is equal to an integral number of wavelengths of light, there will be *constructive interference* according to

$$d \sin \theta = m\lambda, \quad \text{(constructive interference)}, \tag{22.5}$$

where d is the slit separation, θ is the angle measured at the slits between the optic axis and the point on the screen, and m is the integer order number. A value of $m = 0$ corresponds to the point on the optic axis at the screen where clearly, by symmetry, the path difference is zero and we expect the two arriving waves to be in phase; this is known as the central maximum. There is also a corresponding condition for *destructive interference* given by

$$d \sin \theta = \left(m + \frac{1}{2}\right)\lambda. \quad \text{(destructive interference)}. \tag{22.6}$$

In this case, because the incident intensity at the two slits is equal, the destructive interference is complete and this equation gives the set of locations that are completely dark.

The light pattern on the screen will consist of a set of alternating bright and dark fringes (slit-shaped regions) spaced periodically and symmetrically about the optic axis. The first fringe to either side of the central maximum is known as the first-order maximum, the second bright fringe to either side is the second-order maximum, and so on. Because the distance to the screen, D, is very large compared to the slit spacing, then for small angles from the optic axis, the positions y at which bright fringes occur on the screen measured from the optic axis $y = 0$ are given by $\tan \theta \approx \sin \theta = y/D$ (see Figure 22.11). Substituting from Equation (22.5), the fringes are equally spaced along the screen at the positions $y = m\lambda D/d$.

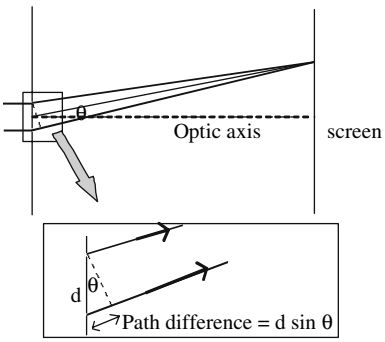


FIGURE 22.10 *Young's double-slit experiment. The inset shows the geometry to calculate the optical path difference, the extra distance traveled by the lower beam. Note that the diffraction angle θ shown in the main diagram is the same angle shown in the triangle in the insert.*

Example 22.2 A double-slit pattern is observed on a screen 5 m away from the slits, which are separated by 0.05 mm. If the first bright fringe is observed to lie a distance of 4.6 cm from the optic axis on which the central interference maximum lies, find the wavelength of the light producing the pattern.

Solution: The angle that the observation point (bright fringe) makes with the optic axis is given by $\sin\theta \approx \tan\theta = 0.046 \text{ m}/5 \text{ m} = 0.0092$. Substituting this expression into Equation (22.5) for interference maxima and choosing $m = 1$ for the first-order, we find $\lambda = d \sin \theta = 4.6 \times 10^{-7}$ m = 460 nm, a blue color.

Let's consider the light pattern on the screen in a bit more detail. If $E_1 = E_o\cos(\omega t)$ is the oscillating electric field from slit 1 at a point on the screen, then at a point of constructive interference the electric field from the second slit will be $E_2 = E_o\cos(\omega t + m2\pi) = E_o\cos(\omega t)$, because the phase difference must be a multiple of 2π. Furthermore, because the intensity on the screen is proportional to $E_{net}^2 = (E_1 + E_2)^2$, we have

$$I_{constr} = 4I_o, \tag{22.7}$$

where $I_o = E_o^2$ is the intensity at the screen produced by either slit (this could be directly measured by covering one of the slits). Similarly, at an interference minimum where E_1 and E_2 are 180° out of phase so that $E_{net} = (E_1 - E_2) = 0$, we have

$$I_{destr} = 0. \tag{22.8}$$

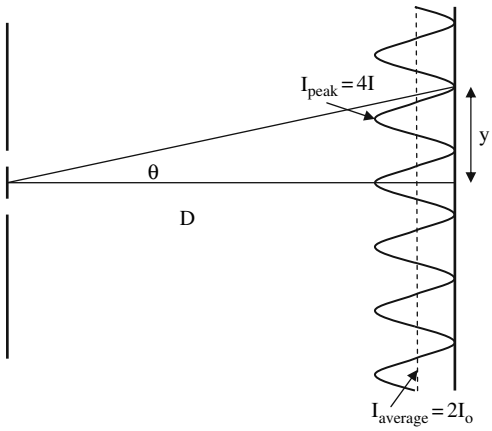


FIGURE 22.11 *Intensity pattern and geometry of the double-slit experiment (ignoring diffraction effects discussed below).*

Figure 22.11 shows the predicted distribution of the intensity along the screen produced by interference, given, for small θ, by $I = 4I_o\cos^2(\pi dy/D\lambda)$. Note that although the maxima have four times the intensity of a single slit, the average value of the intensity (easily found as the value of I about which the curve is symmetric, the dashed line in the figure) is just equal to twice the intensity of a single slit or the total intensity of the incident light passing beyond both slits. This must be the case according to conservation of energy with the total energy separately passing through each slit adding up to the total detected at the screen. The effect of interference is to spatially redistribute the intensity of light in a characteristic interference pattern of light.

In order to observe this double-slit interference pattern the two incident light beams must have the same frequency and be in phase at the two slits (or at least have a definite time-independent phase relation); two such beams are said to be *coherent.* If two completely incoherent beams, those with no definite phase relation (see below), were used, one at each slit, no interference pattern would be observed and the intensity at the screen would simply be the sum of the individual intensities of each light beam according to

$$I = I_1 + I_2. \qquad \text{(incoherent light).} \tag{22.9}$$

What determines whether light is coherent or incoherent? Actually there are various gradations in the coherency (or degree of phase integrity) of light in both time and space. Light can have a definite phase relation over various spatial distances across the transverse direction of a beam (leading to a well-defined wavefront and known as spatial coherence) as well as over various periods of time corresponding to definite phase relations over different distances along its direction of travel (leading to a well-defined wave shape and known as temporal coherence). For example, light from different portions of the heated filament of an incandescent light bulb has no definite phase relationship because the electrons generating the light at different locations do not interact with each other and emit their light in an uncorrelated manner. Incandescent light is thus incoherent both spatially and also temporally. In contrast, we show that laser light is an excellent source of (spatially and temporally) coherent light.

Nonlaser light can always be made spatially coherent by using an arrangement similar to Figure 22.3 to create a point source of such light generating a plane wave as discussed earlier. Thus, all across a plane wave light will be in phase because it has traveled the same path length from the point source. However, the longitudinal distance over which the light is coherent is determined by its temporal coherence, the time during which there is a definite phase relation. Light emitted by specific electron transitions from excited states to lower energy states in atoms or molecules (discussed further in Chapter 25) takes place over a finite "lifetime," or coherence time τ_{coh}, of those electronic states.

The longer the coherence time, the greater the distance along the direction of travel over which there is a definite phase variation and so we identify this as the *coherence*

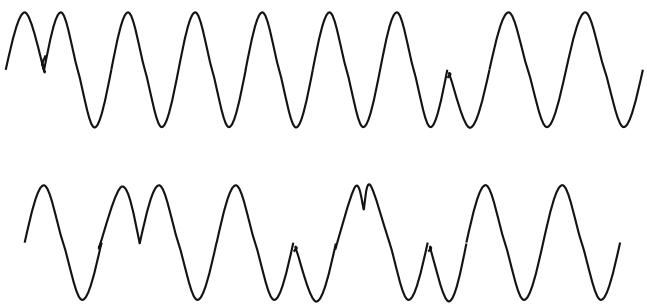


FIGURE 22.12 *Two waves of the same frequency and amplitude, but with different coherence lengths, traveling to the right. The top wave has a longer ℓ_{coh} than the bottom.*

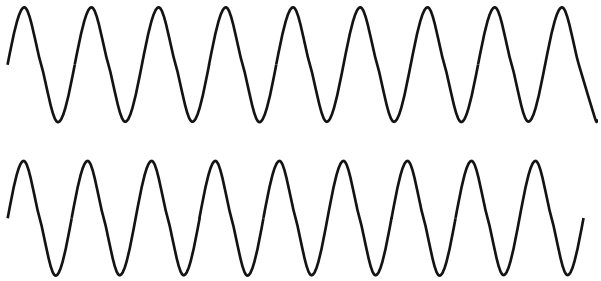


FIGURE 22.13 *Two waves of slightly different frequency, or wavelength, starting out in phase at the left and traveling toward the right will arrive with no definite phase relation.*

length, ℓ_{coh} given by $\ell_{coh} = c\tau_{coh}$ (Figure 22.12). Greater coherence lengths correspond more closely to pure sine waves and a shorter coherence time corresponds to a larger range of frequencies Δf, present with

$$\tau_{coh} \propto \frac{1}{\Delta f}.$$

Incandescent light sources have very broad frequency (color) content and corresponding extremely short coherence lengths. If a colored filter is used to reduce the range of frequencies present then the coherence length of the light can be increased somewhat. The purer the color of the light beam is, the longer its coherence time and length will be. Some lasers generate extremely pure colors of light and have coherence lengths of many kilometers.

Now, to understand the effect of coherence on an interference experiment, first consider the simple case of two waves that have slightly different frequencies, but that start in phase. After traveling some distance, they will lose their common phase because of the frequency difference, as shown in Figure 22.13. In an interference experiment with light, the phase relations between different beams are of utmost importance. Light from any real source, such as the filament of an incandescent light bulb, can be made spatially coherent as we have seen, after focusing down to a point and using the distant plane wave produced, but it will still have a particular coherence length due to temporal coherence. If the coherence length is not longer than the distances involved in the experimental geometry, each beam will itself not have a definite phase over the entire path, even if both start out in phase together, and it will be impossible to observe any interference between different beams. Light from an incandescent bulb has a very short coherence time of about 10^{-10} s, corresponding to a coherence length of a few cm. Unless the optical path length differences are very small, interference experiments with incandescent bulbs are not generally possible.

2.2. SINGLE-SLIT DIFFRACTION

So far we have only very loosely defined diffraction to be the bending of light at a sharp edge or obstacle and have discussed in principle how Huygens' construction can be used to determine the angular spread of the diffracted light. One of the early and most striking demonstrations of diffraction, and the one most responsible for the final acceptance of a wave picture for light in the early 1800s, is the bending of light around a small circular obstacle such as a coin. At that time, Newton's theory of light, treating light as a particle, still dominated the scientific world. Fresnel's early wave theory of light submitted to the French Academy of Sciences in 1818 quickly led Poisson, a nonbeliever of the wave theory and member of the Academy, to deduce a prediction of Fresnel's theory that seemed absurd to him. Poisson claimed that the wave theory should lead to a central bright spot in the shadow of the object, a prediction that

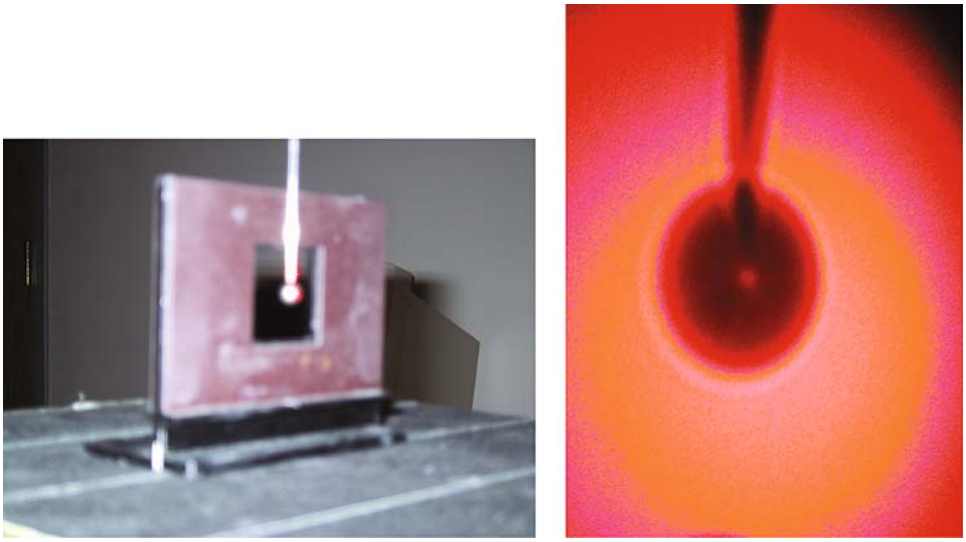


FIGURE 22.14 (left) *A ball bearing magnetically held in the path of a laser beam;* (right) *The Poisson–Arago spot, the bright spot in the center of the geometric shadow of the ball bearing, seen on a distant screen.*

invalidated the theory as far as he was concerned. Arago then performed the experiment and, to everyone's amazement, discovered what is now termed the Poisson–Arago spot (see Figure 22.14). This dramatic event led to the rapid acceptance of the wave theory of light. Rather than analyze this experiment, we consider the mathematically simpler case of the diffraction of light at a narrow slit.

Figure 22.15 shows the experimental arrangement with a single-slit of width a illuminated by a plane wave of monochromatic light and the pattern of light examined on a screen located a distance D from the slit, with $D >> a$. This is an example of *Fraunhofer diffraction,* in which the diffracted light is examined at a large distance from the slit, in the so-called far-field. If the screen were close to the slit, the diffraction pattern would be more complex as well as more mathematically difficult to analyze; this near-field diffraction is known as *Fresnel diffraction.*

The Fraunhofer diffraction pattern from a single-slit consists of a central bright maximum surrounded by a series of secondary maxima of decreasing intensity known as fringes (Figure 22.16). Notice in the figure that the central maximum is wider than the other secondary maxima; its width is inversely related to the slit width. The narrower the slit is, the greater the extent of diffraction and correspondingly the wider the central maximum. We can determine the locations of the maxima and minima, or fringe boundaries, in the diffraction pattern, by using a simple argument based on the

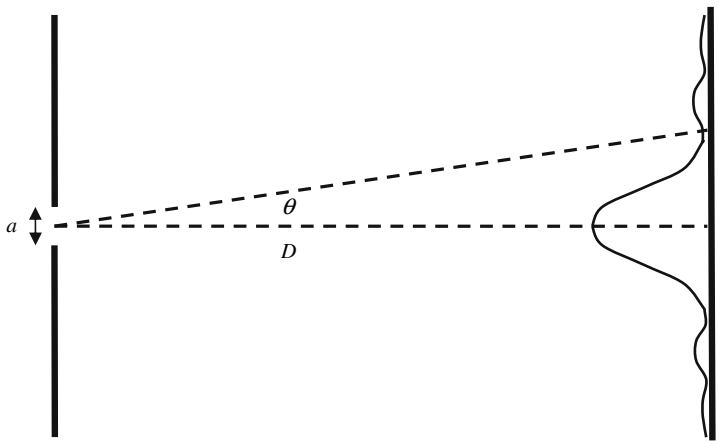


FIGURE 22.15 *Single-slit Fraunhofer diffraction pattern.*

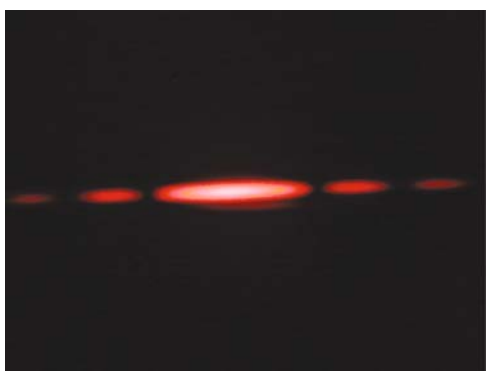


FIGURE 22.16 *Single-slit diffraction using a red He–Ne laser.*

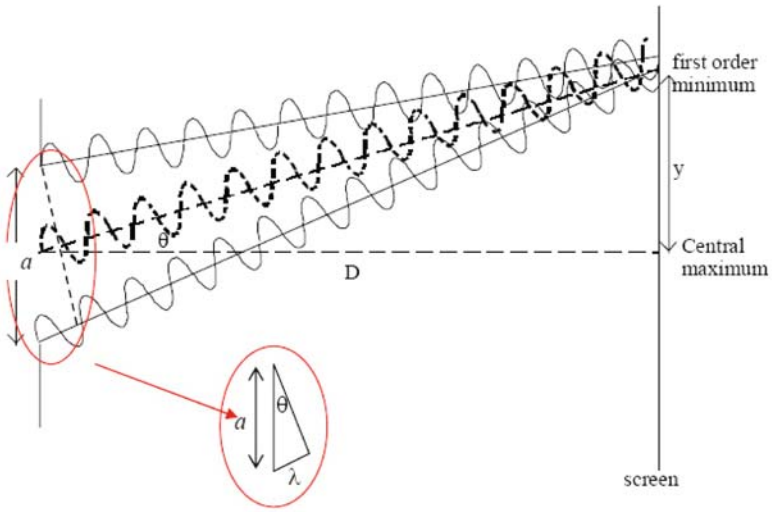


FIGURE 22.17 *Construction to find the condition for the first-order minimum in the single-slit diffraction pattern, Equation (22.10).*

phase relations of the wavelets emitted at the slit when they arrive at the screen. Those rays that emerge parallel to the optic axis remain in phase and produce a bright central maximum. To find the position of the first minimum on either side, consider those rays that are deviated by an angle θ from the optic axis, such that the path difference between a ray from one edge of the slit is one wavelength λ more than that from the opposite edge as shown in Figure 22.17. Because all these rays are parallel, at this angular condition the rays will cancel in pairs as can be seen from the following argument. The ray from the center of the slit will have a path length to the screen of $\lambda/2$ more than the ray from the bottom of the slit and hence these two rays will destructively interfere. Using the same argument repeatedly, we can consider neighboring rays just above each of those in a stepwise fashion, rays that will cancel pairwise because the path difference will remain at $\lambda/2$ all the way up the slit, to conclude that there is no light at this point. At this angle we have, from the insert in the diagram,

$$\sin\theta = \frac{\lambda}{a}, \qquad \text{(first diffraction minimum)}, \tag{22.10}$$

defining the distance along the screen from the optic axis y to the first minimum to be $y = D\tan\theta \approx D\sin\theta = D\lambda/a$.

The next larger angle at which we can have pairwise destructive interference and thus a diffraction minimum is shown in Figure 22.18. The slit is imagined to be divided into four equal portions with a total path difference of 2λ between the top and bottom, so that there will again be pairwise cancellation for rays within each separate half of the slit just as above. For this case we must have that

$$(a/2)\sin\theta = \lambda \quad or \quad a\sin\theta = 2\lambda.$$

Continuing this construction, the general condition for a diffraction minimum becomes

$$a\sin\theta = m\lambda, \qquad \text{(diffraction minima)}, \tag{22.11}$$

where m is a nonzero integer. Notice that this equation for the single-slit dark fringes is very similar to Equation (22.5) for the location of bright interference fringes in the double-slit experiment and the symbol meanings must be kept carefully in mind.

Equation (22.11) predicts that for a given wavelength of light, the width of the diffraction pattern on a screen is inversely related to the slit width; that is, for small angles $\sin\theta \sim \theta \propto 1/a$. This means that the smaller the slit width is, the wider the observed fringe pattern on a distant screen. Conversely, slits that are very wide compared to the wavelength of light only show a faint fringe pattern near the geometric shadow of the slit edges, with no other diffraction effects occurring in this geometrical optics limit. This was the limit that we investigated in the previous two chapters.

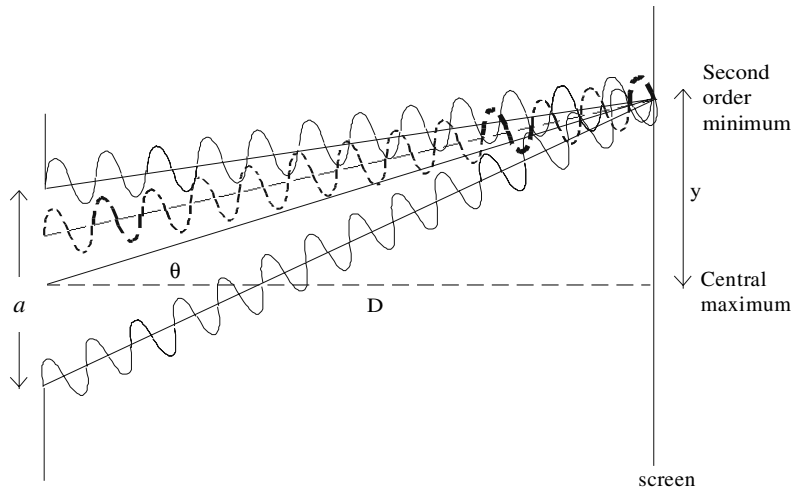


FIGURE 22.18 *Construction to find the second-order minimum location, Equation (22.11) with* m = *2, for the single-slit diffraction pattern.*

The pattern of intensity variation in the diffraction fringes from a single-slit is more difficult to derive but is shown graphically in Figure 22.15. The central bright fringe is twice as wide as the other maxima and is much brighter. The secondary maxima are less than 5% as intense as the central maximum with the intensity rapidly decreasing for higher order maxima.

2.3. DOUBLE-SLIT INTERFERENCE RECONSIDERED

Returning to the double-slit experiment, the intensity profile of the interference fringes will be governed by diffraction from each slit. Instead of the profile shown earlier in Figure 22.11 where diffraction effects were ignored, the actual pattern observed, an example of which is shown in Figure 22.19, is a convolution (productlike combination) of the interference and diffraction effects and depends on the particular slit widths and separation. With the individual slit widths small compared to the slit separation, the bright central diffraction maximum will have many interference minima where no light falls on the screen in contrast to the diffraction pattern of a single-slit shown before in Figure 22.16. Light from each slit interferes with that from the other slit and produces a characteristic fringe pattern with a fringe spacing given in the small angle approximation by $\Delta y = (\lambda/d)D$ and the diffraction minima are spaced by

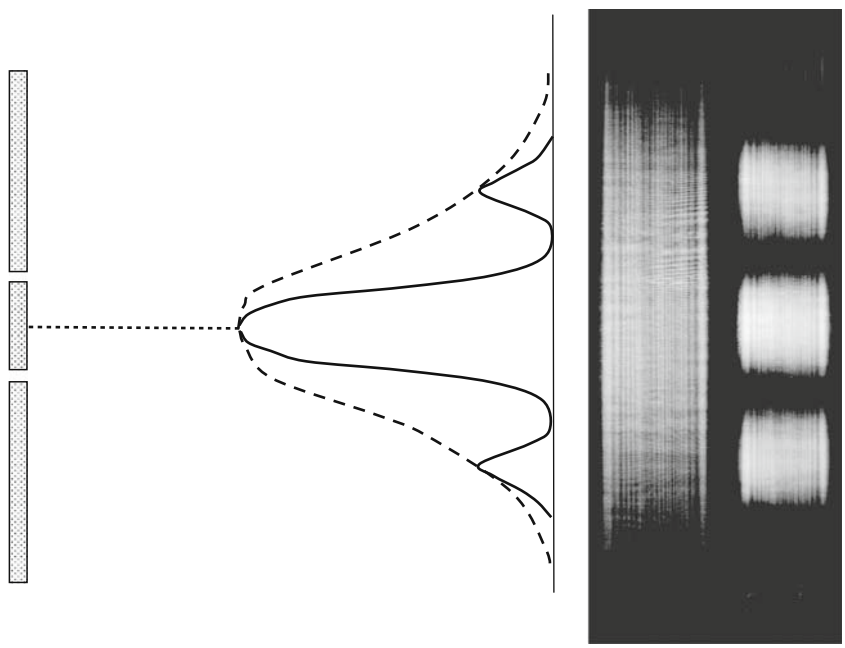


FIGURE 22.19 *Intensity pattern for the double-slit interference experiment shown on the left. The dotted line shows the overall single-slit diffraction pattern obtained if either slit is covered. On the right are the actual patterns observed with a red He–Ne laser.*

$\Delta y = (\lambda/a)D$. Because $d > a$ the fringe spacing is smaller than the widths of the diffraction peaks. Fringe intensity from Figure 22.11 is modulated by the single-slit diffraction patterns from each slit that essentially overlap on the screen.

Example 22.3 Let's return to Example 22.2 with two slits separated by 0.05 mm with a screen 5 m away and examine the problem in more detail, Suppose that the two slits are identical and each has a width of 0.01 mm. What will the pattern on the screen look like when illuminated with light at 460 nm?

Solution: In the previous example we were told that the interference fringes were spaced 4.6 cm apart on the screen. Having learned about diffraction from slits, we now know that the overall intensity on the screen will be modulated by the single-slit diffraction pattern. The central diffraction maximum lies within an angle given by $\sin\theta = \lambda/a$, where a is the given slit width, so $\sin\theta = 460\text{ nm}/0.01\text{ mm} = 460 \times 10^{-9}\text{ m}/1 \times 10^{-5}\text{ m} = 0.046$, corresponding to a distance along the screen of $y \approx D\sin\theta = (5\text{ m})(0.046) = 23$ cm from the optic axis to the first diffraction minimum. Because according to Example 22.2 each interference fringe is spaced 4.6 cm from the next, the fifth fringe from the central one, on either side, actually falls directly on the first diffraction minimum and therefore will have zero intensity and not be seen. Within the central diffraction maximum there then will be four fringes visible on either side of the central fringe on the optic axis, making a total of nine fringes within the central diffraction maximum. If we look further off axis, the next minimum in the diffraction pattern occurs when $\sin\theta = 2\lambda/a = 0.092$, so that it occurs at a distance of 46 cm from the optic axis. We find that the $(46/4.6) = 10$th interference fringe lies at this location and so is also not visible. Therefore within the second diffraction maximum on either side there are four interference fringes. The resulting pattern of fringes is somewhat similar to the photo shown in Figure 22.19, but with different numbers of fringes observed.

2.4. MULTIPLE SLITS AND DIFFRACTION GRATINGS

As the number of slits with the same width and spacing is increased beyond two, the patterns of light and dark on a distant screen at first become more complex, with each slit still creating the same Fraunhofer diffraction pattern as in our single-slit discussions but with the interference pattern within the diffraction peaks having more detail. The angular positions of the bright interference fringes are the same as for the double-slit, Equation (22.5) above, independent of the number of slits, namely

$$d\sin\theta = m\lambda, \tag{22.12}$$

where d is the uniform slit spacing between any adjacent pair of slits and m is the order number. In fact, the same exact reasoning holds in deriving this equation, because if light from two neighboring slits has a path difference of $m\lambda$, then light from any pair of slits, neighboring or not, will still have a path difference that is a multiple of the wavelength of light.

On the other hand, the nature of the minima and the width of the maxima both change with the number of slits. With an increasing number of slits, the secondary maxima dramatically decrease in intensity and the central maximum narrows in width. Figure 22.20 illustrates this sharpening of the central maximum with an increasing number of slits. The larger number of slits makes the condition for constructive interference from all the slits that much more stringent. With only two slits, points on the screen near

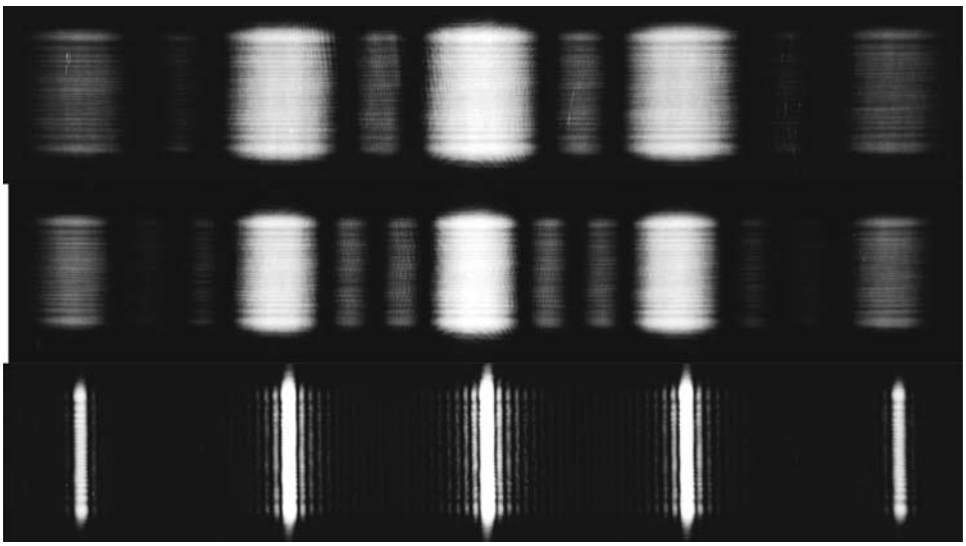


FIGURE 22.20 (top) *Three-slit diffraction showing one minor peak between the major double-slit pattern peaks;* (middle) *4-slit pattern with two minor peaks;* (bottom) *23-slit pattern showing the sharpening of the central maximum.*

the central maximum peak have path differences that are only slightly different from an integer number of wavelengths and so there is only a gradual decrease in intensity away from the peak. With many slits, even if the rays from two neighboring slits have a path difference of only a small fraction of a wavelength, the path difference from the 100th slit away is increased by a factor of 100 and much more destructive interference occurs. This is the reason for the much sharper central maximum with many slits.

Diffraction gratings are devices that have a very large number of very narrow slits, separated by distances comparable to the wavelength of the light. The best gratings for visible light have more than 30,000 lines per inch (or spacings of less than 1 μm apart). There are two fundamental types of gratings for optical work: transmission gratings, of the type we have been discussing, and reflection gratings that have their fine rulings made on a mirrored surface. Diffraction gratings give very sharp interference peaks, so that with monochromatic light, such as that from a laser, there will be a series of small spots, one for each order of Equation (22.12).

The real utility of diffraction gratings is their ability to analyze polychromatic light as "spectrum analyzers," dispersing all the colors present in a particular light source (Figures 22.21 and 22.22). Equation (22.12) indicates that for a given slit spacing, or its inverse, known as the grating constant which is the number of lines per unit distance,

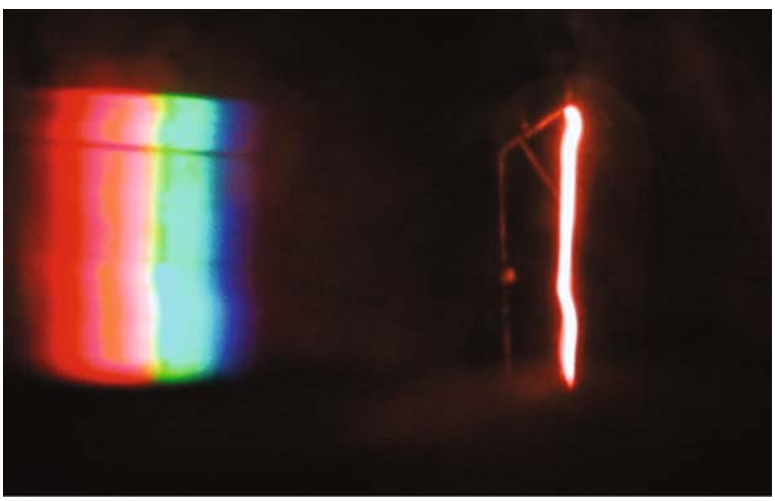


FIGURE 22.21 *Diffraction pattern observed from a grating in front of a white light slit source; note the continuous spectrum of colors observed in the first-order peak to the left of the central peak.*

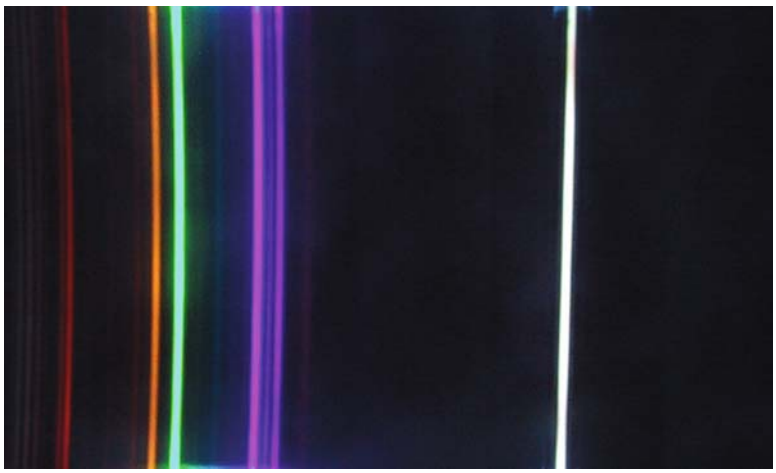


FIGURE 22.22 *Diffraction pattern observed in a reflection grating from a mercury slit lamp; note that only discrete colors are present. We study these spectra—known as line spectra—later in the book.*

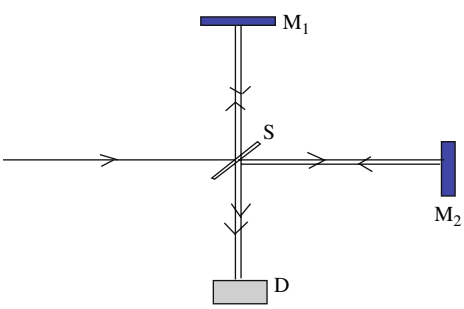
different wavelengths of light will be diffracted at different angles. Gratings can thereby serve as prisms, dispersing light of different colors to produce a *spectrum.* According to Equation (22.12), light of longer wavelengths will be diffracted by larger angles in contrast to a prism which refracts light of shorter wavelengths more because of dispersion. For each order, the grating will produce an entire spectrum, except for the central maximum (zeroth order) at which all colors superimpose. In grating spectroscopy, a source of light is collimated, directed on a grating, and the diffracted light detected. We show the application of spectroscopy in the study of atomic physics in the next chapter.

2.5. INTERFEROMETERS

Optical devices known as interferometers split a light beam into two beams that travel different routes and are then brought together to interfere. One important example, known as a Michelson interferometer, is shown schematically in Figure 22.23. The beamsplitter S divides the incident light into two portions, one reflected and one transmitted at its back surface. These are separately reflected by mirrors M_1 and M_2 and the reflected beams are recombined by the beamsplitter and observed by a detector *D.* The path differences in the two "arms" of the interferometer must be shorter than the coherence length of the light in order to observe interference effects at the detector. Typically one mirror is slowly and precisely moved along its axis and the fringes shift at a rate of 1 fringe per a path difference equal to the wavelength of light.

Interferometers can be used for a variety of purposes including, for example, measuring the wavelength of light or accurately measuring optical distances or changes in optical distances. This can be accomplished simply by counting fringes at the detector and knowing that each fringe corresponds to an optical path difference of one wavelength. Interferometers are often useful to check on the quality of optical components during and after manufacture. They are also useful to determine refractive indices of transparent materials by inserting a sample in one branch of the interferometer, thus increasing the optical path in that branch and measuring its optical distance compared to its physical distance.

FIGURE 22.23 *A Michelson interferometer. Typically one mirror is able to move along its optic axis parallel to the light beam and the interference fringes are observed by the detector.*



3. RESOLUTION

In order to distinguish by eye two objects that are very close together, whether they be microscopic objects or stars in the sky, we can use lenses to magnify the objects. Aside from lens aberrations that can limit the quality of the images as discussed in the last chapter, diffraction imposes a fundamental

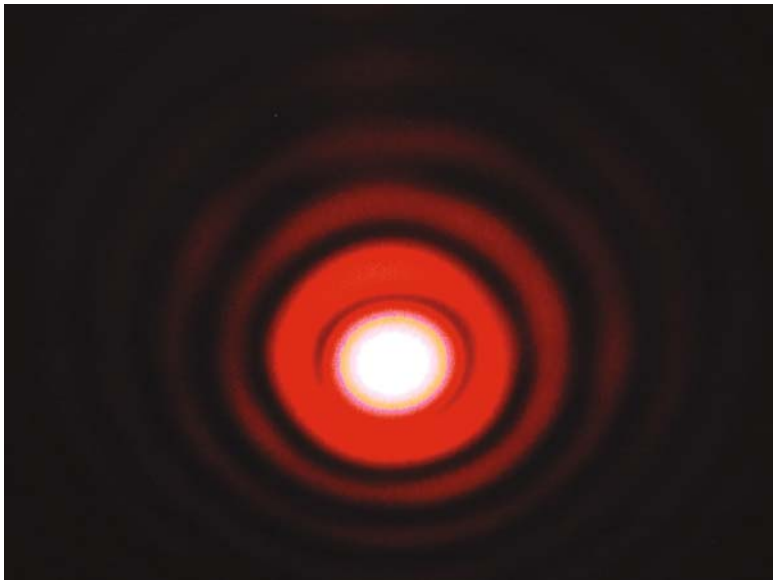


FIGURE 22.24 *Fraunhofer diffraction from a circular aperture; the central spot has saturated the detector and appears white.*

limit on our ability to discern two closely spaced objects. Most optical instruments use circular rather than slit apertures, therefore before discussing such limits on resolution we first briefly consider the diffraction of light by a circular aperture.

The far-field (Fraunhofer) diffraction pattern from a circular aperture consists of a central circular maximum, known as the *Airy disk,* surrounded by a set of circular fringes (Figure 22.24). A similar, but more complex, derivation to that for a slit shows that the angular spread of the Airy disk (first-order minimum location) is given by

$$d \sin \theta = 1.22\lambda, \tag{22.13}$$

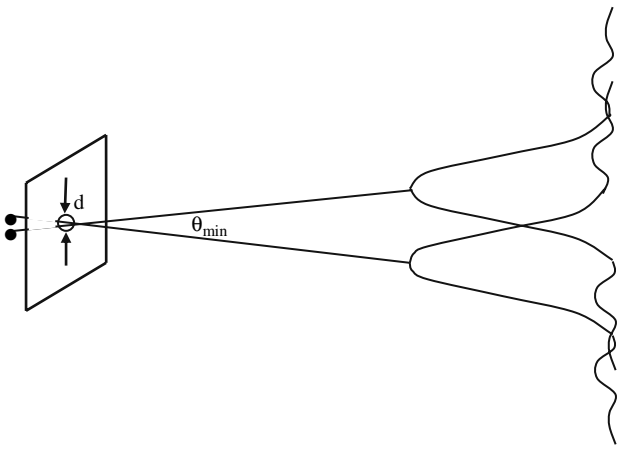
where d is the diameter of the aperture. Effectively, $d/1.22$ is the average width of the equivalent slit representing the circular aperture so that $(d/1.22) \sin \theta = m \lambda$ resembles the single-slit diffraction equation. A photo of the intensity profile of the image of a circular aperture is somewhat deceiving because the secondary maxima are much dimmer ($<5\%$) than the Airy disk.

Two closely spaced objects will each produce a diffraction pattern in the image of an optical system. When the two objects are so close that their Airy disks overlap in the image, it becomes very difficult to distinguish whether there are actually two objects present or just one. The *Rayleigh criterion* is the accepted condition for the resolution of two such objects: *two objects are just resolved when the central maximum of one is superimposed on the first diffraction minimum of the other*. From Equation (22.13), the Rayleigh criterion can be written as

$$\theta_{\text{min}} = \frac{1.22\lambda}{d}, \tag{22.14}$$

where, because the angles are small, θ_{min} represents the minimum angular separation (in radians) of two objects as shown in Figure 22.25 and d is the aperture size. Figure 22.26 shows the diffraction patterns observed when two distant point sources of light get progressively closer. As the angle subtended by the point sources at the aperture gets smaller the images of the two point sources coalesce and blur, so that eventually they cannot be resolved. Thus, in order to increase the resolution of an optical system, the shortest wavelength and the largest aperture possible are desired.

FIGURE 22.25 *Light from two distant point sources of light subtending a small angle θ passes through a circular aperture of diameter d. When the central maximum of one image is at the first diffraction minimum of the other image the two are just resolvable according to Rayleigh's criterion.*



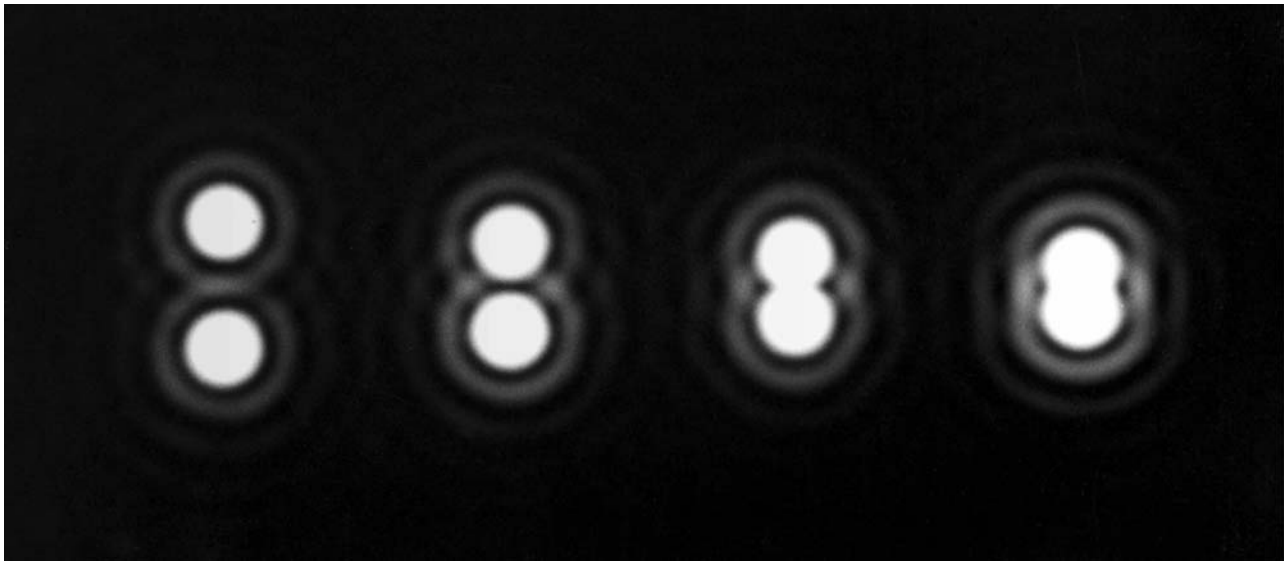


FIGURE 22.26 *Diffraction patterns observed for the situation in Figure 22.25 with progressively closer objects.*

Example 22.4 For the human eye, with a pupil diameter of about 2 mm and using a wavelength of 500 nm, calculate the minimum angle separating two just resolvable points. Then find the actual minimum distance between such just resolvable points as well as the distance between their images on the retina.

Solution: Using Equation (22.14), we find that

$$\theta_{\text{min}} = \frac{(1.22)(500 \times 10^{-9})}{2 \times 10^{-3}} = 3 \times 10^{-4}\,\text{rad},$$

corresponding to about 1 min of arc. The minimum spatial separation occurs when the objects are placed at the near point of the eye (taken as 25 cm). This distance between two just resolved points is then (0.25 m)(3×10^{-4}) = 75 μm, somewhat less than 1/10 mm. So, thinking of the 1 mm divisions on a ruler, our eyes can resolve two objects that are apart by about 1/10 of the smallest mm division on the ruler. The corresponding separation distance between the central maxima of the two images on the surface of the retina (using a lens–retina distance of 2 cm) is (0.02 m) (3×10^{-4}) = 6 μm. Note that cones have an average spacing in the fovea of about 2 μm, so that the detector size is three times smaller than the diffraction limiting image size. To have an effective resolution this high, it appears that at least one nonactivated cone must lie between two other activated cones. Only in the fovea do individual cones have 1:1 connections with nerve cells going to the visual cortex. Our eyes are exquisitely designed to provide the best possible resolution for the physical dimensions of our eyeball. There would be no improvement in the visual resolution of our eyes by having smaller cone cells, because diffraction is the fundamental limit and not the size of the cones.

With a microscope it is more common to discuss resolution in terms of the minimum separation distance of two objects under optimal conditions, known as the *resolving power*, rather than resolving angle. A straightforward (but omitted) derivation shows that the resolving power is given by

$$r_{\text{min}} = \frac{0.61\lambda}{n \sin \alpha} = \frac{0.61\lambda}{NA}, \tag{22.15}$$

where α is the acceptance angle of the light from the objects at the objective lens (Figure 22.27), λ/n is the wavelength of light in the medium between the sample and lens, $0.61 = 1.22/2$, and the product ($n \sin \alpha$) is known as the *numerical aperture* (NA) of the lens. For routine microscopy $n = 1$, whereas for higher magnifications, an oil-immersion objective is often used to increase the resolving power. In this case a drop of immersion oil (typical $n = 1.5$) is placed on the cover slip between the sample and the lens, increasing the resolving power (by decreasing r_{min}) by about 50%. The larger the numerical aperture of a lens, the greater is its resolving power. Numerical apertures of 1.4 are commonly used at high resolutions with optical microscopes. Equation (22.15) then tells us that the very highest resolution obtainable with a light microscope is about $\lambda/4$. Because $\sin \alpha$ is limited to a maximum value of 1, the only way to further improve resolution in a microscope is to decrease the wavelength of the probing radiation. We show how this is done using an electron microscope in the next chapter.

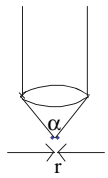


FIGURE 22.27 *The resolving power of a microscope is increased with a larger acceptance angle α using a very short focal length lens.*

CHAPTER SUMMARY

In discussing the different optical wave phenomena in this chapter, what is important is not the physical distance that light travels, but rather the optical path, defined by

$$Optical\ Path = \sum n_i d_i, \qquad (22.2)$$

where n_i and d_i are the index of refraction and distance traveled in the ith segment of the path.

Diffraction is the bending of waves around obstacles or the spreading of waves passing through an aperture. Interference is the superposition of waves in space leading to constructive and destructive interference. One example is interference in thin films of index n and thickness t, where light of wavelength λ reflected normally will experience destructive interference if

$$2nt = m\lambda. \quad \text{(destructive interference).} \qquad (22.3)$$

This phenomenon is exploited in the reflection-interference microscope to study a thin surface layer and in the use of nonreflective glass coatings.

Young's double-slit interference, in which coherent light passes through a pair of slits separated by distance d and is viewed at a distance at an angle θ, leads to constructive interference at angles such that

$$d \sin \theta = m\lambda. \quad \text{(constructive interference).} \qquad (22.5)$$

The order number m is an integer. Each of the slits of width a, in turn, diffracts the light and produces a single-slit diffraction pattern governed by

$$a \sin \theta = m\lambda, \quad \text{(diffraction minima),} \qquad (22.11)$$

where θ is the angle to the first diffraction minimum. The overall pattern observed in a double-slit experiment is the convolution (productlike mix) of both patterns of intensity of light. A common device, the diffraction grating, has numerous closely spaced slits, of separation d, and gives an intensity diffraction pattern of brightness governed by the grating equation

$$d \sin \theta = m\lambda. \qquad (22.12)$$

Resolution is limited by diffraction. Rayleigh's criterion for the threshold of resolution, the minimum angular separation θ_{min} of two barely resolvable objects, viewed through an aperture of size d, is given by

$$\theta_{min} = \frac{1.22\lambda}{d}. \qquad (22.14)$$

In a microscope, this can be shown to give a resolving power, or minimum separation of two just resolved objects, of

$$r_{min} = \frac{0.61\lambda}{n \sin \alpha} = \frac{0.61\lambda}{NA}, \qquad (22.15)$$

where α is the light acceptance angle and NA is the numerical aperture, defined in the equation.

QUESTIONS

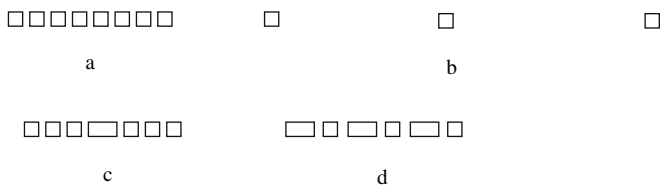
1. Which of the following properties of light do not depend on the material medium in which the light travels: speed, frequency, wavelength, optical path, and index of refraction?
2. A garden hose has an adjustable nozzle which determines whether the water comes out as a tightly focused jet of water or as a wide-angled spray. As the nozzle is adjusted from a jet to a spray is the size of the exit aperture increasing or decreasing?
3. In a concert hall which sound waves bend more on leaving the hall at the open rear doors, high C or low C?
4. Why does a thin film of oil or gasoline on water appear multicolored as in Figure 22.6?
5. Why is the correct thickness of a nonreflective coating on a glass surface equal to $\lambda/4$ where λ is the wavelength of the light in the film? Assume the index of refraction of the coating is less than that of the glass.
6. Show that the two angles labeled θ in Figure 22.10 are in fact equal so that the path difference can be written in terms of the diffraction angle.
7. When listening to a weak radio station on your car radio often the reception will fade in and out while slowly driving along. What is this due to and can you use its source as a way to estimate the wavelength and frequency of the radio waves?
8. Interference is a phenomenon that takes an otherwise uniform energy density and re-distributes the energy into maxima and minima with the same total energy. Discuss this statement.
9. Why does the source of light for Young's double-slit experiment need to be coherent? What pattern would be observed if an ordinary light bulb were used as a light source right behind the slits?
10. Sketch a picture of the central maximum and first peak to either side in the intensity pattern seen on a distant screen in a double slit experiment with slits of width 1/8 of the slit spacing. Check the angle at which the minimum in the diffraction pattern occurs relative to the interference peaks and be sure to draw the correct number of interference peaks within each diffraction peak. What does the pattern look like when one of the two slits is covered?
11. Why is no interference pattern seen when looking at a car's headlights from a distance on a road at night? After all, the light comes from two "point sources" close together.
12. With regard to a diffraction grating, Equation (22.12) is called the grating equation. What is the meaning of d in this equation? What is d for a grating with 10,000 lines per cm?
13. Discuss the meaning of θ_{min} in Equation (22.14). In particular, where is the angle measured from, to where is it measured, what is the meaning of d and does a higher resolution mean a greater or smaller θ?
14. Suppose that the density of cones in the macula was a factor of 10 greater. Would the resolution of the human eye be increased with no other changes? Explain.
15. What is the purpose of immersion oil when used at the objective of a compound microscope? Explain how it works.

MULTIPLE CHOICE QUESTIONS

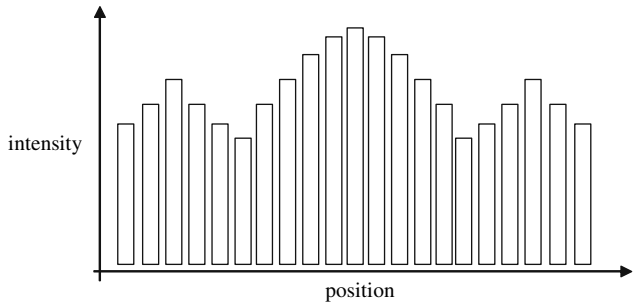
Questions 1 and 2 refer to two parallel light rays, initially in phase and having a 500 nm wavelength, that reach a detector after one of the rays travels through a 10 cm long block of glass with an index of refraction of 1.5 as in Figure 22.2.

1. The optical path difference between the two rays at the detector is (a) 10 cm, (b) 750 nm, (c) 15 cm, (d) 5 cm.
2. The total number of wavelengths shifted between the two rays when they reach the detector is (a) 100,000 (b) 200,000, (c) 300,000, (d) 500,000.
3. The equations for double-slit interference and single-slit diffraction, $m\lambda = d\sin\theta$ and $m\lambda = a\sin\theta$, respectively, are very similar. Which of the following is true? (a) m can take the value zero in the first equation, but not in the second, (b) m can take the value zero in the second equation, but not in the first, (c) m can take exactly the same values in both equations because the two equations describe exactly the same interference pattern, (d) m is the mass of the electron in both equations.

For Questions 4 and 5, consider the four patterns shown to the right. A box corresponds to a bright spot.



4. The pattern most likely to have been produced by a single fairly wide slit is (a) a, (b) b, (c) c, (d) d.
5. The pattern most likely to have been produced by a diffraction grating is (a) a, (b) b, (c) c, (d) d.
6. The figure below shows an approximate plot of intensity as a function of position for a double-slit interference pattern. The vertical bars represent where light is observed. Between the vertical bars are dark spots. The most likely reason the maximum intensities are not all the same is (a) the beam isn't centered properly, (b) the slit openings have a finite size, (c) that's what is predicted for two slits that have infinitesimal openings, (d) there are actually more than two slits.



7. Which of the following best describes the angles at which dark spots will be observed in a double-slit diffraction pattern, if d is the slit separation, λ is the wavelength of the incident radiation, and $m = 0, \pm 1, \pm 2, \pm 3, \ldots$? θ is measured relative to the incident direction and its sine equals (a) $m\lambda/d$, (b) $(2m + 1)\lambda/(2d)$, (c) md/λ, (d) $2md/((2m + 1)\lambda)$.
8. Near-normal 500 nm light is reflected from a thin organic film (with $n = 1.5$) on water. What minimum thickness results in destructive interference? (a) 83 nm, (b) 167 nm, (c) 250 nm, (d) 330 nm.
9. If the first-order double-slit diffraction minimum lies at the same place as the fourth-order interference maximum, how many fringes will be visible in the central diffraction maximum? (a) 3, (b) 5, (c) 6, (d) 7.
10. When two microscope slides are placed together and a laser beam is passed through the two at near normal incidence, if the slides are squeezed together, the pattern of reflected light moves. This is best explained by (a) the gap between the slides is changed by squeezing, (b) the thickness of the slides is changed by squeezing, (c) the index of refraction of the air in the gap between the slides is changed by squeezing, (d) the index of refraction of glass is changed by squeezing.
11. What is the approximate minimum coherence length of a 1 cm diameter 500 nm light beam needed to observe a fringe pattern in a Michelson interferometer with mirror-to-detector distances of 0.2 m? (a) 1 μm, (b) 1 cm, (c) 0.5 m, (d) 1 km.
12. When a car is 500 m ahead of you, you see its tail lights as one long, red light. When the car is 100 m ahead of you, you see that the tail lights are actually several red lights placed close to each other. This is because (a) the pupil of your eye has a finite size, (b) the Doppler effect for light shifts the frequency of the tail lights, (c) light disperses in the lens of your eye, (d) light is made up of photons.
13. When a sheet of paper is 20 cm from your face you can see two small dots of ink. When the paper is a meter from your face the dots appear as a single dot. That is most likely because (a) the pupil of your eye has a finite size, (b) you are farsighted, (c) light disperses in the lens of your eye, (d) light from the dots is polarized.
14. To improve the resolving power of a microscope one can do all of the following except (a) increase the wavelength of light, (b) use immersion oil to index match the glass slide to the glass objective, (c) maximize the acceptance angle, (d) increase the power of the objective lens.
15. The maximum resolution of an optical microscope is about (a) 1 nm, (b) 1 mm, (c) 1 μm, (d) 1 **Å**.

PROBLEMS

1. A thin film of oil with refractive index 1.5 on a water puddle is illuminated from directly overhead by white light. If there is an interference maximum at 600 nm and a minimum at 450 nm with no other minimum in between, what is the film thickness, assumed uniform?
2. A soap film with index of refraction 1.33 is surrounded by air and illuminated by white light. If the film is 265 nm thick, which wavelength in the range 400–750 nm will interfere constructively?
3. If a soap film of refractive index 1.33, surrounded by air, is illuminated by a red (633 nm) and a green (515 nm) light, find the minimum film thickness at which the reflected light will appear red. Repeat to find the thickness at which it will appear green.
4. If a 220 nm thick soap film ($n = 1.33$) is placed on a glass slide ($n = 1.5$) and illuminated with white light, find the wavelengths that interfere constructively in the reflected light. (Note: To do this problem you will need to reanalyze the derivation of Equations (22.3) and (22.4) when there is an additional π phase shift at the second interface.)
5. For a reflection-interference microscope, what is the minimum cell thickness that results in maximum contrast between adhesion sites and the cell region when illuminated with blue light (480 nm)? Take the index of refraction of the cell to be that of water.
6. In a double-slit experiment with red light (633 nm) using slits with 0.12 mm separation, what is the fringe spacing on a screen that is 3.5 m from the slits.
7. If the two slits in a double-slit experiment each have a width of 0.08 mm and a spacing of 0.24 mm, how many interference peaks will lie within the central diffraction maximum using 488 nm light?
8. In a double-slit experiment with 500 nm light, the slits each have a width of 0.1 mm.
   (a) If the interference fringes are 5 mm apart on a screen which is 4 m from the slits, determine the separation of the slits.
   (b) What is the distance from the center of the pattern to the first diffraction minimum on one side of the pattern?
   (c) How many interference fringes will be seen within the central maximum in the diffraction pattern? Draw a sketch of the pattern.

9. In a double-slit experiment with two slits of width 1.0 μm spaced 4 μm apart, suppose a beam of electrons is incident on the slits after being accelerated from rest through a potential difference of 100 V.
(a) We show in Chapter 24 that electrons (and all particles) have a wavelength given by $\lambda = h/p$, where h is Planck's constant and p is the particle's momentum. What is the wavelength of the electrons in this problem?
(b) If the pattern of detected electrons is observed on a fluorescent screen 20 m from the slits, what is the width of the central diffraction maximum?
(c) How many interference fringes will be observed within the central diffraction maximum?

10. In a double-slit experiment, each slit has a width of 0.02 mm and they are spaced 0.14 mm apart. The pattern of light when a coherent 550 nm beam is incident on the slits is observed on a screen 4 m away.
(a) Find the spacing between interference fringes on the screen.
(b) Find the full width of the central diffraction maximum on the screen.
(c) How many fringes are visible within the central diffraction maximum?
(d) What happens to the pattern if the entire apparatus is immersed in water? Find new answers to each of the above parts.
(e) Describe what will happen to the pattern of light if a different polarizer is placed in front of each slit with their transmission axes at right angles to each other (assume the incident light is unpolarized). Will the intensities change? Will the pattern change? Give reasons for your answers.

11. Georges Seurat, a post-impressionist French painter, used a technique of painting with small dots of color placed close together on a canvas. From sufficiently far away, the dots cannot be distinguished and the painting looks normal in appearance.
(a) If the dots on the painting are separated by 1.5 mm and the painting is observed under light of 550 nm wavelength with eyes having a pupil diameter of 2 mm, find the minimum distance you must stand from the painting so that the individual dots cannot be resolved.
(b) Under the conditions of part (a), what would be the distance between the images of dots on the retina of the eye, 2.0 cm behind the lens of the eye.
(c) If a small scaled copy of a Seurat painting is made which is 100-fold smaller in size, the dots would be expected to be spaced only 0.015 mm, too small to be resolved with the naked eye. Using a compound microscope with a 10x eyepiece and 17 cm tube length, what maximum focal length of objective lens would be needed to test the scaled copy for authenticity (meaning that the painting is made from small dots, rather than continuous color).

12. A person's eye has a pupil diameter of 0.2 cm and has a length of about 2.5 cm (from lens to retina). If light of 550 nm is used, find the following.
(a) The minimum distance between resolved images on the retina of the eye, ignoring any lens aberration.
(b) The distance apart that two point sources of light can have at the near point of the eye ($N = 25$ cm) and just be resolved.
(c) What microscope magnification would be needed for the eye to just clearly image these two point sources of light if they are only 500 nm apart?
(d) With a 10x eyepiece and a tube length of 17.0 cm, find the objective focal length needed to achieve the magnification in part c.

13. In a Michelson interferometer when using laser light of 488 nm, how many fringes are scanned at the detector if one of the mirrors is displaced by 2.4 μm.

14. A Michelson interferometer is used to determine the wavelength of a monochromatic light source. If 100 fringes are counted as one of the mirrors is scanned a distance of 31.7 μm, what is the wavelength of the light?

15. Given that the resolving power of the eye corresponds to about 1 minute of arc, how far away can a vehicle be at night where you can still resolve whether it is a car (with two headlights separated by 1.8 m) or a motorcycle?

16. The Hubble space telescope has a resolution of about 0.1 s of arc. If aimed at the moon (3.9×10^5 km away) what is its resolving power? If aimed at Saturn (1.3×10^9 km away)? If aimed at the nearest galaxy (about 2×10^6 light years, where a light year is the distance light travels in a year)?

17. In a double-slit experiment with two slits of width 1.0 μm spaced 4 μm apart, suppose a beam of protons is incident on the slits after being accelerated from rest through a potential difference of 2500 V.
(a) What is the speed of the proton?
(b) What is the wavelength of the proton?
(c) If the pattern of detected protons is observed on a fluorescent screen 20 m from the slits, what is the center-to-center spacing between the constructive interference maxima?
(d) What is the full width of the central diffraction minimum?
(e) How many interference fringes will be observed within the central diffraction minimum?

23 Imaging Using Wave Optics

A number of standard and novel methods to optically image biological and other materials are discussed in this chapter to give the reader a sense of the large variety of tools available. We start with a survey of the arsenal of newer light microscopies available for the study of biological materials, in particular. Another wavelike property of light, its polarization, can be used in several optical polarization methods to study biomolecules. Earlier, we discussed the important imaging technique of MRI in Chapter 18; this chapter concludes with a discussion of two other wave-related imaging techniques: electron microscopy and x-ray diffraction/computed tomography (CT) imaging with x-rays.

1. THE NEW LIGHT MICROSCOPIES

Aside from the resolution needed to form an image of a microscopic object, discussed in the previous chapter for a standard compound microscope, a minimal amount of *contrast* is also needed to clearly detect an image. Contrast can be defined in terms of the visibility of a sample object compared to the background using the percent contrast,

$$\%\ \text{Contrast} = \frac{(I_{\text{bkgd}} - I_{\text{sample}})}{I_{\text{bkgd}}} \times 100, \tag{23.1}$$

where the intensities are average values over those portions of the image. Contrast is determined both by the properties of the object and by those of the microscope. We can distinguish two fundamental types of contrast: amplitude and phase contrast.

Amplitude contrast is due to direct differences in the wave amplitude of the imaged sample and background light due to absorption or scattering from the sample. This is the basis for several types of microscopy including normal or *bright-field microscopy* discussed in the previous chapter. In this technique, the background appears bright white and objects are imaged by their darker or colored appearance due to absorption or scattering. Because most biological materials do not absorb much visible light, usually a colored stain that preferentially sticks to the sample and is washed from the background is used to enhance the contrast. Before defining phase contrast, we take a look at several microscopy methods that use amplitude contrast enhancing schemes.

Very small or thin objects are difficult to see in bright-field microscopy because of the light background and low contrast. If sufficient scattering occurs from an object, it can be better viewed using a variation known as *dark-field microscopy* in which the background light is blocked by a central stop and only the scattered light from the object is imaged. Figure 23.1 shows this microscope arrangement. A hollow cone of light from a special annular aperture is focused on the specimen and the

J. Newman, *Physics of the Life Sciences*, DOI: 10.1007/978-0-387-77259-2_23,

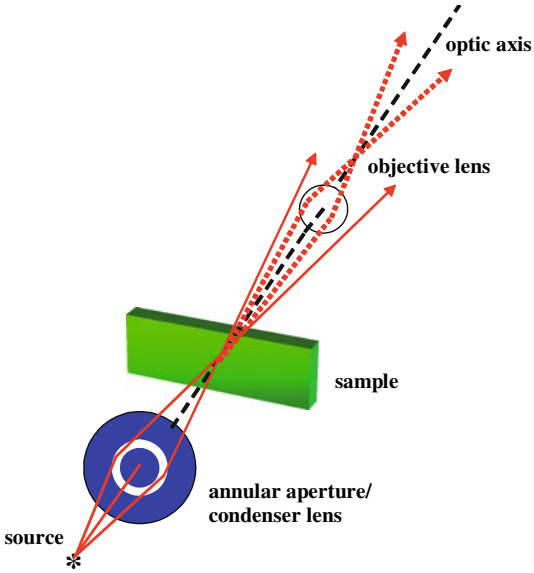


FIGURE 23.1 *Schematic of dark-field microscope optics. An annulus aperture in front of the condenser lens (focusing light on the sample) ensures that none of the unscattered incident light* (smooth red line) *passes through the collection optics. Only light scattered from the sample* (dotted red line) *reaches the image plane (not shown).*

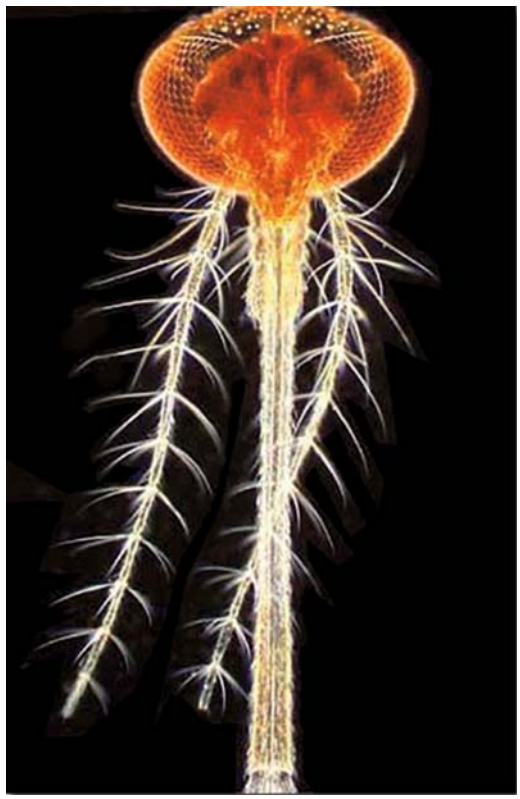
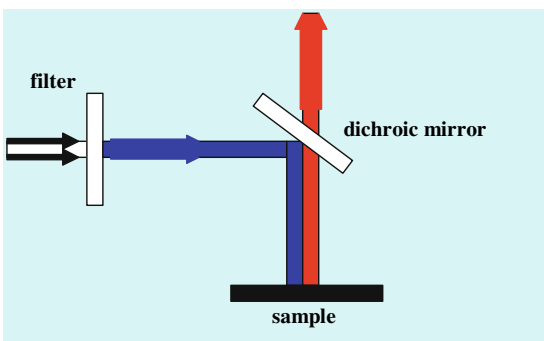


FIGURE 23.2 *A dark-field image of a mosquito head.*

collection optics are arranged so that only the scattered light, and not the directly transmitted cone of light, is collected and focused by the microscope. Figure 23.2 shows an example of a dark-field image.

Fluorescence microscopy is an important variation of amplitude contrast microscopy. Since most samples are not sufficiently fluorescent, usually fluorescent dyes are used to bind to specific sites on the sample and only fluorescent light is imaged in the microscope. To accomplish this, filters must be used to block other wavelengths of light. Unless a laser is used as a light source of the proper excitation wavelength, an excitation filter is used to limit the incident light to the shorter wavelengths capable of exciting the fluorescent dye. The incident light is used in either a dark-field microscope arrangement or in the arrangement shown in Figure 23.3 to direct excitation light onto the sample. Fluorescent light emitted by the sample is then collected and filtered using a barrier filter that passes only longer wavelength fluorescent light, blocking the incident light. In this way there is no background light except for a stray unwanted fluorescent signal from imperfections in the optics. In Figure 23.3, the dichroic mirror is specially coated to reflect only shorter wavelengths but to transmit only longer wavelengths of light, thereby acting both as two filters as well as a beamsplitter. Figure 23.4 shows an image of a multiply labeled fluorescent endothelial cell.

FIGURE 23.3 *Schematic of a fluorescence microscope without the imaging optics shown. The dichroic mirror reflects the shorter wavelength light, but transmits the longer wavelength fluorescent light.*



Recent developments of new multicolor fluorescent dyes for use in microscopy have been partly responsible for a revolution in fluorescent microscopy. Aside from advances in scientists' ability to label specific molecules with a dye, many of the newer dyes have their fluorescence controllable by specific environmental changes. For example, certain dyes can serve as sensors of local pH, with their fluorescence properties depending on pH, whereas others can serve to monitor calcium ions Ca^{2+}, the important messenger and regulating ion in a cell, because their fluorescence is affected by the binding of

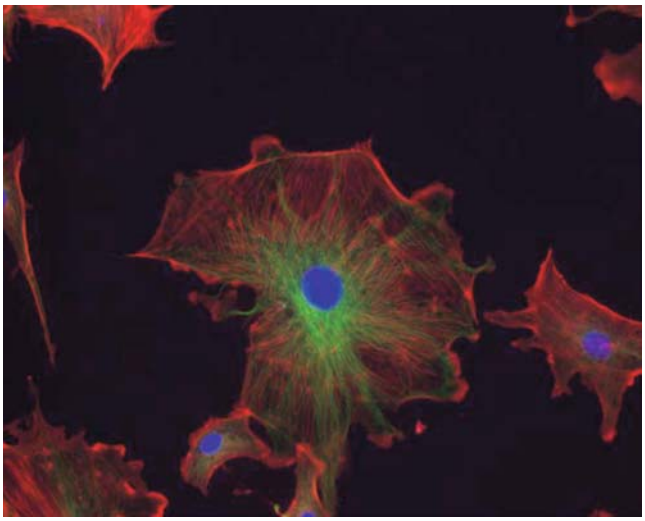


FIGURE 23.4 *Three-color fluorescence image of an endothelial cell showing the tubulin* (green), *nucleus* (blue), *and actin cytoskeleton* (red).

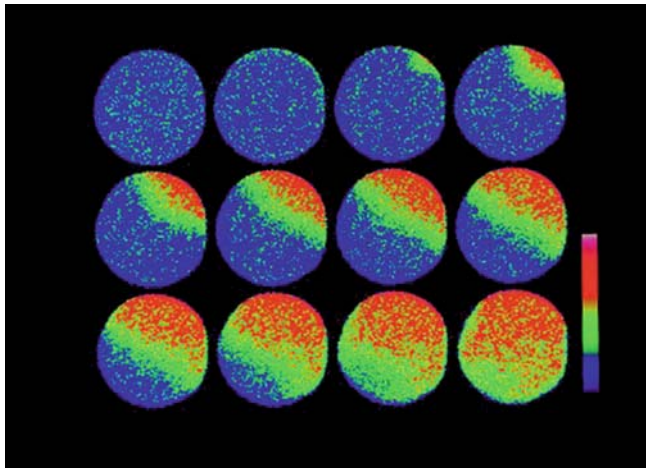


FIGURE 23.5 *A wave of increase in calcium ion concentration sweeps across an egg cell just after fertilization as monitored by the green fluorescence from a Ca-sensitive dye attached to small dextran molecules. The images are taken 5 s apart and show the Ca wave starting around the 1:30 o'clock position and spreading across the cell.*

calcium. An even newer class of fluorescent dyes can be used as optical biosensors to detect conformational changes in macromolecules or binding of ligands (small molecules with specific binding sites) to those molecules. In this way, not only can the locations of specific macromolecules to which the dyes are bound be monitored, but so can their physiological state (Figure 23.5).

As already mentioned, most biological samples for microscopy are essentially completely transparent to visible light, absorbing and scattering very little light, and therefore having very poor contrast (hence, the use of stains and fluorescent dyes). However, all such samples do have somewhat different refractive indices than the surrounding solvent and are therefore called phase objects. These produce a phase shift in the light waves they transmit relative to those through the background, more or less as shown in the last chapter in Figure 22.2. If light is simply allowed to pass through the sample and be imaged, the relative phase shifts will not change the intensity of the light and the objects will be invisible. However, encoded phase information in the light passing through the sample can be used to provide *phase contrast* in several types of microscopies. We discuss two major types: phase contrast and differential-interference-contrast (DIC) microscopy. In both cases the crux of the technique is to separately change the relative phase of the light that interacts with the sample and the undeviated light so that when they are recombined, there will be intensity differences in the images due to interference effects.

Phase contrast microscopy is similar to dark field microscopy in that a hollow cone of light is focused onto the sample but now that light is collected by the objective lens (Figure 23.6). In the absence of a sample, the objective lens produces an image of the annulus used to produce the cone of light at a plane known as the "source image plane." However, light that interacts with the sample will be diffracted from that path (dotted lines in Figure 23.6) with a small phase shift from passing through the more optically dense sample as well. The intensity of this diffracted light will be much less than that of the undeviated light and it will be brought to a focus at a different plane (because

FIGURE 23.6 *Optics of the phase contrast microscope. A cone of light is produced by the annulus and focused on the sample. The undeviated beam is focused by the objective onto a groove in a phase plate located at the source image plane. The groove both attenuates the undeviated beam intensity and shifts its phase with respect to the diffracted light, most of which passes through the rest of the phase plate and is brought to focus at the object image plane. This image is then further magnified by the eyepiece (not shown).*

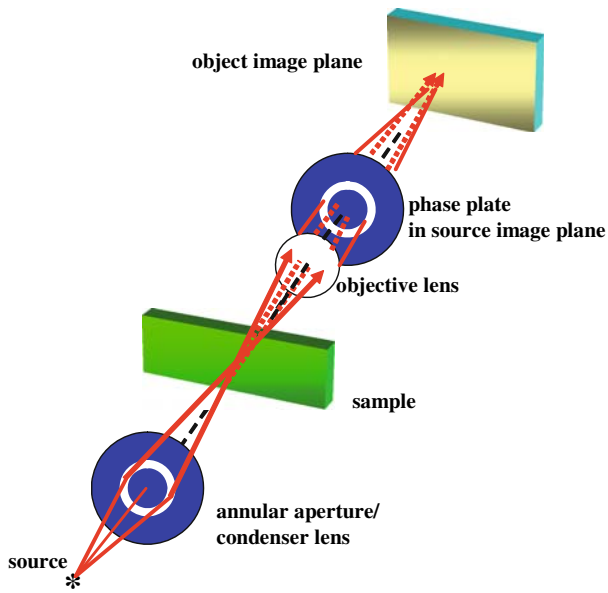


FIGURE 23.7 *Phase contrast micrograph of a paramecium.*

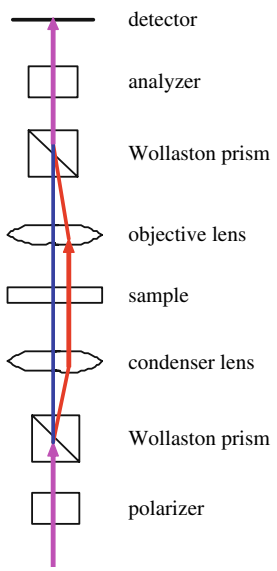


FIGURE 23.8 *Optics of the differential interference contrast (DIC) microscope. The Wollaston prisms are used to create and recombine two beams with slightly offset centers, as well as to introduce a 180° phase shift between the two. If the relative phase of these two is shifted by the sample, then when allowed to interfere after the analyzer, a high-contrast image is formed (not shown). Note magenta = red + blue.*

the object distance is much less than the light source distance from the objective), known as the "object image plane."

In the phase contrast microscope a device known as a phase plate is inserted at the source image plane to improve the contrast. A groove in the phase plate aligned with the image of the annulus is used to shift the phase of the undeviated light relative to the diffracted light. An absorption coating in the groove also decreases the intensity of the undeviated beam, so that it is closer to matching the intensity of the diffracted light in order to provide even better contrast. Phase plates are usually built into objective holders and matched pairs of condenser and objective lenses are used to ensure proper alignment. The total phase difference between undeviated and diffracted portions results in intensity variations in the image that are directly proportional to optical path differences between the sample and background regions. Depending on whether the phase plate gives an additional positive or negative phase shift with respect to the diffracted light, the background can be made dark or bright (Figure 23.7).

In *differential-interference-contrast (DIC) microscopy* there is a complete physical separation of the incident light into two closely spaced beams that probe adjacent portions of the sample. These beams are then used to generate an interference pattern that produces intensity differences in the object image plane. Two special prisms, known as Wollaston prisms, are used both to produce two in-phase beams from one and to recombine them after passing through the sample into one final beam with a 180° phase shift introduced between the two (Figure 23.8). In the absence of a sample and with a uniform background, the two beams completely cancel after recombination due to the 180° phase difference. With a sample present in one beam but not the other, the extra phase differences between the two beams give rise to bright interference light. In this case the image intensities are not proportional to optical path differences, but rather, because of the two spatially separated beams, to the rate of change of optical path transversely (in the direction of the separation of the two beams) across the object. That's the reason for the term "differential interference". Because the rate of change, rather than the absolute optical path difference, is important in DIC microscopy, edge contrast is greater and thinner samples can be better imaged (Figure 23.9).

Wollaston prisms function by spatially separating the two different polarization components of light. They are able to do this because the calcite crystal of which they are made has different refractive indices along two different crystal axes as discussed further in the next section. After the two beams of light travel through the sample this process is reversed in a second matched prism and the two beams recombine after a 180° phase shift introduced by an asymmetric placement of the second prism. At this point, even though the beams are out of phase and overlapping, they cannot interfere with each other because their polarization directions are orthogonal and hence independent. A polarizer oriented at a 45° angle between these directions serves to analyze that portion of each beam and to allow them to subsequently interfere and produce the image. Thus, the Wollaston prisms are serving solely as a beamsplitter and recombiner, whereas the polarization properties of the beams are not used to produce the DIC image. Polarized light can be used in microscopy in the polarization microscope discussed in the next section.

Most current versions of the above microscopic techniques use modern methods of digital recording and computers to further increase the resolution and contrast of images. Developments in detector technology have made use of CCDs (charge-coupled devices) and image intensifiers very commonplace in microscopy. CCD video cameras, based on arrays of discrete light-sensitive detectors, allow digital recording of time-dependent processes in two-dimensional arrays of picture elements, or pixels. These arrays are now relatively inexpensive and are widely used in digital

cameras, whose pictures can then be printed out on ordinary computer printers. Digitally stored video frames from microscopy can be computer-enhanced and manipulated to allow improved resolution, contrast, and quantitative measurements using special software.

Within the last ten years or so many new microscope techniques have been developed that use laser illumination, including confocal microscopy and multiphoton microscopy. *Laser-scanning confocal microscopy* focuses a laser beam to an extremely small spot within the sample and images light only from that spot onto the detector. A pinhole in front of the detector serves to eliminate out-of-focus light from other regions of the sample, only allowing light from the focused spot to be collected. The spot is then scanned over the sample, by moving either the microscope stage or laser beam, in a raster pattern to map out the sample image, having remarkable depth and lifelike appearance (Figure 23.10).

FIGURE 23.9 *DIC image of a deer tick. Note the sharp edges and high contrast.*

Multiphoton microscopy uses a pulsed laser to provide an intense beam of low-energy photons that is scanned across the sample similar to confocal microscopy. When two (for two-photon microscopy) or more (for three- or multiphoton microscopy) of these photons with identical energy are simultaneously absorbed by a fluorescent molecule they can provide the same total energy that a single photon would in the usual fluorescence microscope. The incident photon beam is tuned to the proper wavelength so that two or three or more photons, when combined, give an energy resonant with the fluorescent material, producing subsequent fluorescence emission. Quantum mechanics allows this additive resonance only when the multiple photons are absorbed nearly simultaneously, requiring very high laser intensities. One important advantage of this method is that there is virtually no absorption of these lower-energy photons at any other location in the sample where the beam is not focused and the density of photons is not sufficient to allow multiphoton absorption. Thus instead of using high-energy photons that can damage the sample to produce fluorescence, one can use much lower energy photons and excite the fluorescent molecules through the combined energy of several photons only where the beam is focused. This technique is sensitive enough to image the intrinsic or autofluorescent light from the amino acid tryptophan and other fluorescent macromolecular groups within the sample itself without the addition of fluorescent dyes. High-resolution, high-contrast, three-dimensional images can be obtained using these methods even with samples as thick as 0.5 mm (Figure 23.11).

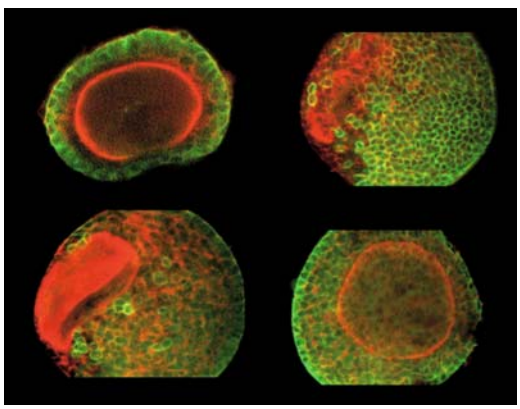


FIGURE 23.10 *Laser-scanning confocal microscopic images of mouse oocytes showing microtubules in red and actin filaments in green.*

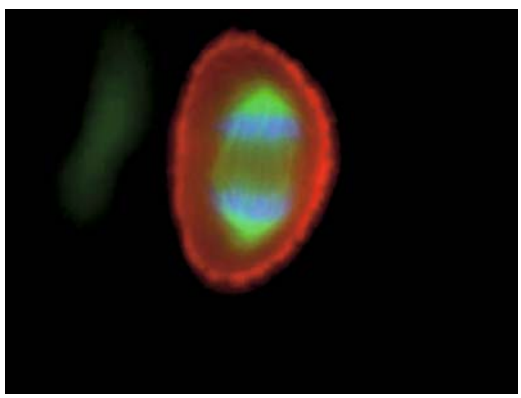


FIGURE 23.11 *Confocal microscopic image of anaphase in a cultured epithelial cell showing chromosomes* (blue), *spindle apparatus* (green), *and actin* (red).

2. OPTICAL ACTIVITY; APPLICATIONS OF LIGHT POLARIZATION

In Chapter 19, we introduced the concept of polarization of a light beam and discussed linearly polarized light as well as the use of Polaroid as a polarizing device to preferentially absorb light with its electric field oriented along one direction. Here, we further discuss the notion of circularly and elliptically polarized light and the use of polarization methods in the study of biomolecules.

Consider two light waves with the same frequency linearly polarized along perpendicular directions as shown in Figure 23.12. If the amplitudes and phases of the two waves are equal, then the superposition of the two waves results in a linearly polarized wave along the vertical direction in (a). With different amplitudes for the two waves, the resultant wave will still be linearly polarized so long as the phases are equal (b). If two waves of equal amplitude are 90° ($\pi/2$ rad or $\lambda/4$) out of phase then when one component is at a zero the other will be at a maximum or minimum. The superposition of those two waves will describe a helical path as the tip of the electric field vector executes circular motion in the transverse wavefront plane itself traveling along at speed c in a vacuum (c). Depending on the relative phases, the circular polarization can be left- or right-handed. Handedness is defined in terms of an observer looking back at the source and the light is right-handed if $\vec{E}$ rotates clockwise.

We can make these ideas quantitative by writing out expressions for the two linearly polarized electric fields (say, along x- and y-axes) as

$$\begin{aligned} E_x &= E_{ox}\cos(\omega t) \\ E_y &= E_{oy}\sin(\omega t) \end{aligned}, \tag{23.2}$$

where we have assumed that E_x leads E_y by 90° (at time 0, E_x is at a maximum and E_y is zero; after 1/4 of a period, E_x is now zero and E_y has increased to a maximum, etc.), and E_{ox} and E_{oy} are the amplitudes of the fields. By using the trigonometry identity $\cos^2\theta + \sin^2\theta = 1$, we find that the components of the vector E satisfy

$$\left(\frac{E_x}{E_{ox}}\right)^2 + \left(\frac{E_y}{E_{oy}}\right)^2 = 1, \tag{23.3}$$

which is the equation of an ellipse. If the two amplitudes are equal (so that $E_{ox} = E_{oy} = E_o$) then Equation (23.3) becomes the equation of a circle ($E_x^2 + E_y^2 = E_o^2$, with radius E_o), the case shown in Figure 23.12c. In the transverse plane the tip of $\vec{E}$ will describe these closed ellipses or circles, but the light wave is actually propagating at the speed of light along the z-direction and the tip of $\vec{E}$ actually describes a helical path in space. The projection of the helix in the x–y plane will be a circle or an ellipse, depending on the amplitudes of the x- and y-components of $\vec{E}$ In a similar way one can show that linearly polarized light can be considered to be the sum of in-phase right- and left-handed circularly polarized light. For example, if the left-handed circularly polarized beam shown in Figure 23.12c is added to its mirror image right-handed beam, the resulting beam has an $\vec{E}$ that is vertically polarized (imagine the summation in the figure: the horizontal components will always cancel with the mirror-image beam). This idea is used below in a discussion of optical activity.

Circularly polarized light can be produced most easily by sending linearly polarized light through a special device known as a quarter-wave plate, or $\lambda/4$ plate. These are made from a *birefringent* (double-refracting) material, one having

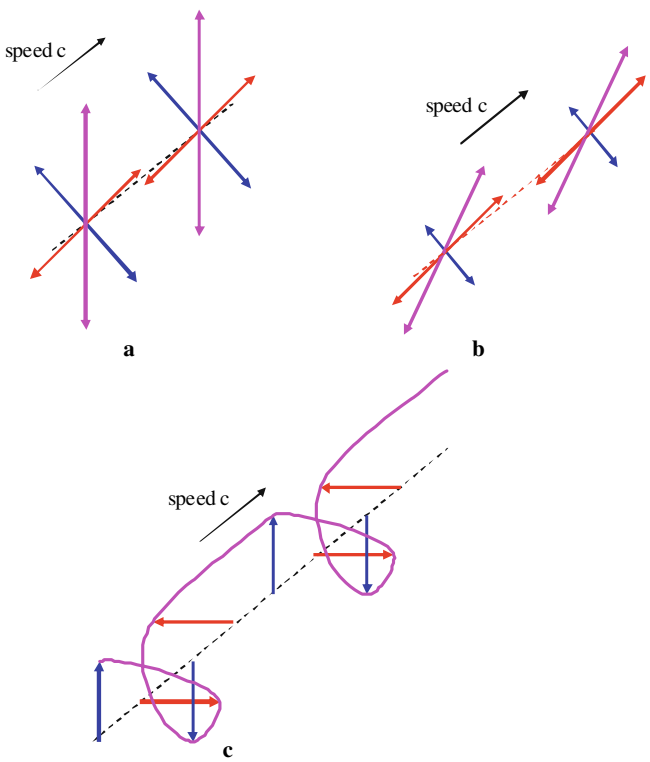


FIGURE 23.12 *Combining two orthogonally polarized waves* (red and blue E field vectors) *at 45° with respect to the vertical. (a) Equal amplitude waves in phase to give a vertically polarized wave* (magenta)*, (b) Unequal amplitude waves still in phase to give a linearly polarized wave at a fixed angle with the vertical* (magenta)*, and (c) Equal amplitude waves 90° out of phase* (red and blue)*, so that the tip of the net E field vector rotates around in a circle as the wave propagates* (magenta)*. In this case a left-handed circularly polarized wave is shown, handedness defined looking backward at the source.*

two crystal axes with different refractive indices, as mentioned in the last section in connection with a Wollaston prism. When linearly polarized light passes through such a material, the different polarization components along either axis travel at different speeds, because $v = c/n_1$ or c/n_2, and will develop phase differences. Furthermore, one beam, called the "ordinary" beam, will be transmitted undeviated, whereas the other, called the "extraordinary" beam, will be refracted and physically separated from the ordinary beam (see Figure 23.8). By adjusting the thickness of the material, a quarter-wave phase difference can be introduced between the two beams, producing fields governed by Equations (23.2). In general this will produce elliptically polarized light but if the wave plate is adjusted to have its axis at 45° to the incident polarization direction then circularly polarized light is produced.

Example 23.1 Suppose that a vertically polarized beam of 500 nm light is incident on a birefringent crystal of mica with a mean index of refraction of 1.552 and which has its crystal axes making a 45° angle with respect to the vertical. If the birefringence of the crystal is $\Delta n = 0.006$, find the minimum thickness of the crystal along the transmission direction of beam so as to produce circularly polarized light.

Solution: If we call the unknown thickness t, then the optical path difference of the two equal components of the vertical polarization along the two crystal axes will be $t\Delta n$ (see Equation (22.2)). In order to produce circularly polarized light, this difference should be set equal to 1/4 wavelength of the light, so that, as in the Figure 23.12c, after leaving the crystal there will be two equal components of electric field that are 90° out of phase, combining to produce a circularly polarized beam. We therefore require $t\Delta n = 1/4$ (500 nm), so that $t = 2.1 \times 10^{-5}$ m $= 0.021$ mm. Mica can be cleaved and polished to produce such quarter-wave plates designed for different wavelengths.

Many biological systems contain components that are anisotropic. These are ordered structures that look different in different directions; for example, the fibrils within a muscle fiber or the crystal-like proteins of the lens of the eye. Polarized light will interact with electrons in such a material in different ways depending on relative orientations and can be used to gain information about such structures. Because of the anisotropy there will be changes in the polarization of transmitted light. Polarization microscopy is yet another way to get images of such anisotropic structures. Linearly polarized light is used as a light source and the imaged light through the objective is passed through a crossed-polarizer. In the absence of any sample, the background light is completely extinguished by the crossed-polarizer. Any resolvable structures that produce some depolarization of the incident light will then produce a bright image (Figure 23.13).

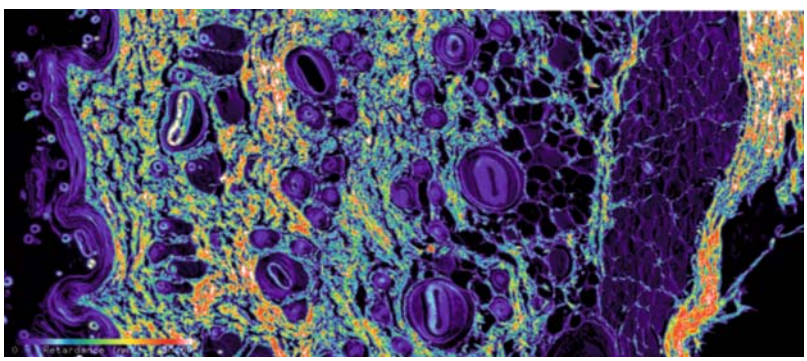


FIGURE 23.13 *Polarization microscope image of rat skin color-coded by the birefringence retardation (see text) which is related to the degree of depolarization of the transmitted light.*

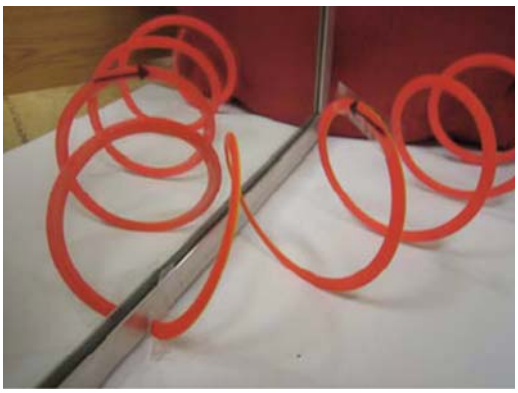


FIGURE 23.14 *Handedness changes in a plane mirror; the left-handed slinky helix (spiraling counterclockwise around the helix axis) changes to right-handed in a plane mirror image (seen on the* left*).*

Most individual biological macromolecules are asymmetric, meaning that they appear different from their mirror image. Most simple molecules are symmetric. Water, carbon dioxide, and many more complex molecules look the same as their mirror images. Biopolymers tend to be formed, at least partially, from helical arrays of molecules, and these will have a handedness. Handedness is a property that changes when viewed in a mirror. As shown in Figure 23.14, a right-handed coiled spring will appear to be a left-handed spring when viewed in a mirror.

On the other hand, a solution of randomly oriented asymmetric molecules will not produce an image in a polarization microscope because the solution as a whole is isotropic. However, asymmetric molecules do have an effect on the polarization properties of light that can be used to gain information about the macromolecules. Asymmetric molecules are said to have *optical activity* and are characterized by different refractive indices for left- and right-handed circularly polarized light. Asymmetric molecules will interact differently with left- and right-handed circularly polarized light because of their handedness.

A simple example may help to clarify this. Imagine a solution of small left-handed helical molecules. Because the electric field vector of the light interacts with the electrons of the helical molecule, left-handed circularly polarized light will allow a stronger interaction with the electrons of a left-handed helical molecule, with the ability to drive them around the helix, and therefore a larger fraction of such light will be absorbed than would be the case for right-handed circularly polarized light. This is somewhat similar to the reason why Polaroid film, with its oriented long polymers, preferentially absorbs light polarized along the polymers: the electric field can then interact more strongly with polymer electrons.

Because linearly polarized light can be considered a sum of left and right circularly polarized light, a solution of optically active molecules probed with linearly polarized light will interact differently with each of these components and affect the polarization of the transmitted light. If the sample absorbs no light, then the light remains linearly polarized, but has its direction of polarization rotated due to different effective optical paths for each polarization. Molecules that rotate the polarization in a left-handed sense are called levarotatory (L) and those that rotate the polarization in a right-handed sense are called dextrorotatory (D). It is a fact that all proteins and most other biological molecules are found only in the L form in nature.

When linearly polarized light is incident on an optically active solution, there can be both phase and amplitude changes associated with the equivalent left- and right-handed circular polarization components making up the incident linear polarization. These can be characterized by two quantities: the *circular birefringence* Δn,

$$\Delta n = (n_L - n_R), \tag{23.4}$$

for the phase changes, where n_L and n_R are the refractive indices for left and right circularly polarized light; and the *circular dichroism* $\Delta\varepsilon$,

$$\Delta\varepsilon = \varepsilon_L - \varepsilon_R, \tag{23.5}$$

for the amplitude changes, where ε_L and ε_R are the absorption coefficients for left and right circularly polarized light. Recall from Chapter 19 (Section 6) that the absorption coefficient is a measure of the intensity of light absorbed in a unit path length and per unit concentration of sample.

Both the circular birefringence and dichroism values depend on the wavelength of light used on a given optically active sample. Spectra showing the wavelength dependence of the birefringence (using the technique known as optical rotary dispersion or ORD experiments) and of the dichroism (using circular dichroism or CD experiments) can be used to characterize biological materials. These techniques are used most to probe the optically active regions of macromolecules, determining their helical content

or following relatively slow kinetic changes that can occur from conformational changes due to environmental factors or to the binding of small ligands. Figure 23.15 shows an example CD spectrum for standards in particular conformations and for a real protein, myoglobin.

3. ELECTRON MICROSCOPY

In our discussion of the resolution possible in a microscope, the resolving power, or closest distance that two distinct objects can lie and still be distinguished under optimal conditions, was given by Equation (22.15) to be no less than $\lambda/4$. For visible light this limits the resolution under the best conditions to about 200 nm. Any further improvement on this limit requires that the wavelength of the probing radiation be decreased. Although ultraviolet microscopes have been developed, the most feasible method for improving resolution is to use electrons in place of light. We show in the next chapter that electrons have an associated wavelength that depends on their momentum (or, in turn, on their energy). Just as with photons, where higher-energy photons have a correspondingly shorter wavelength, we show that higher-energy electrons also have a shorter wavelength. Exactly what it means for an electron or another elementary "particle" to have a wavelength is explored further in the next chapter. For now, we can use the notion of a wave packet introduced in Chapter 19 (Section 5) to picture an electron as having wavelike properties.

Electrons accelerated through a potential difference of 50 kV, typical for an electron microscope (EM), have a wavelength of 0.005 nm, allowing a theoretical improvement in resolution over a light microscope by a factor of 40,000. Unfortunately, other problems limit the practical resolution of the EM, although using a particular variation of electron microscopy has allowed resolutions approaching 0.1 nm at which individual atoms can be directly imaged. The recently developed method of scanning tunneling microscopy (STM), described in the next chapter, allows even higher resolution of surface topography with a resolution of better than 0.1 nm.

The general plan of an EM is shown in Figure 23.16. An electron "gun," or filament and anode combination, is the source of electrons boiled off a tungsten filament heated to very high temperature, similar to a light bulb. The electrons are accelerated through a large potential difference of typically 40–100 kV reducing the wavelength of the electron as it gains kinetic energy. The entire microscope column is evacuated to a fairly high vacuum, reducing energy losses of the electrons from collisions with air molecules. Because electrons can be steered in a magnetic field, a "magnetic condenser lens" is used to focus the electron beam at or near the sample plane down to a spot size of several microns. Samples are supported on copper grids with an array of typically 100 μm × 100 μm square holes coated with a thin uniform layer of a supporting material, such as carbon, that is essentially transparent to electrons. Copper is used

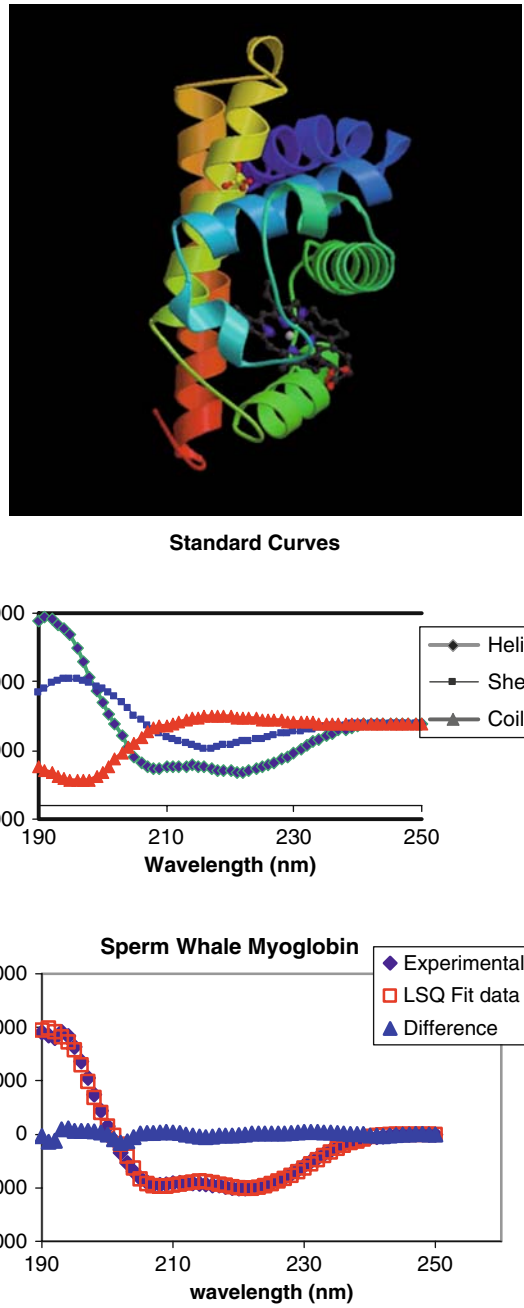


FIGURE 23.15 (top) *Molecular model of the protein sperm whale myoglobin showing helical and coil regions;* (middle) *standard CD curves for pure helix, sheet, or coil;* (bottom) *CD spectrum of the sperm whale myoglobin, showing the best fit to the experimental data as a mix of three different standard components.*

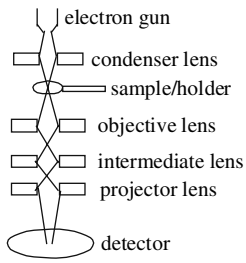


FIGURE 23.16 *Schematic diagram of a transmission electron microscope. The lenses are electromagnets; the entire electron beam path is evacuated. Typical detectors are fluorescent screens, photographic film, or image intensifiers to record digital images.*

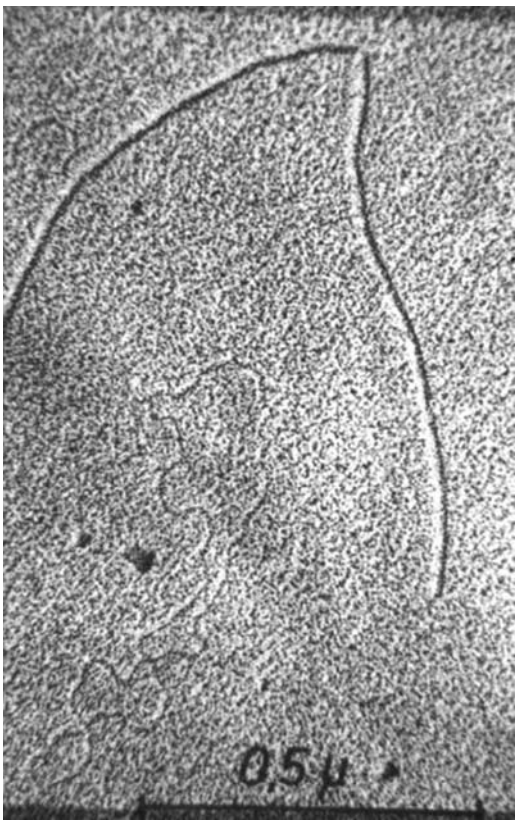


FIGURE 23.17 *TEM of two fd virus particles with one single-stranded DNA from a virus in the lower center.*

because it is a good electrical and thermal conductor, carrying away any heat from the interaction with the beam, and also minimizing the distortion of the focusing magnetic field. The sample is mounted on a movable stage for positioning it in the focused electron beam.

After interacting with the sample, electrons are collected by a (magnetic) objective lens and a magnified image is projected onto a detector by a system of other lenses. Overall magnifications can range from 1000 to over 300,000 times, limited mainly by aberrations in the magnetic lenses. The simplest detector is a fluorescent screen that emits light when struck by the electrons and can be viewed directly by eye or with some further magnification using optical lenses. Other detectors include photographic film or image intensifiers that allow digitization and computer enhancement of images.

Three types of EMs can be distinguished: transmission (TEM), scanning (SEM), and the less common scanning transmission (STEM). Normal TEM, developed in the 1940s, basically creates a greatly enlarged shadow of the sample at the detector. Samples must be very thin for good resolution and thin sections or evaporated deposits of solutions are used. Biological materials are made of smaller atoms (mostly H, C, O, N, P, S) that do not strongly interact with the electron beam and so the contrast is very poor. In order to "see" the sample, some contrast improvement is needed in order to cast a shadow. The usual method is to deposit a heavy metal with high electron scattering power (such as osmium, platinum, gold, or uranium) to coat the structures of interest. This is done in a variety of ways including "shadowing" by direct deposit of heavy metals on the grid, or by negative staining in which heavy metal salt solution fills the region immediately around particles of interest producing a dark background edge around bright images of the transparent objects of interest. Figure 23.17 shows a TEM image of two virus particles with a closed loop of its single-stranded DNA.

SEM uses a tightly focused electron beam (spot size of ~10 nm) directed off-axis at a heavy metal-coated sample as shown schematically in Figure 23.18. The beam is made to scan along the sample in a raster, or TV-like, pattern by a set of scanning coils that steer the electron beam and are coupled to the detectors. Electrons or radiation "reflected" from the sample at each scanned point are collected and used to create an image on a TV screen as the electron beam is scanned across the sample. The spatial pattern of the scanned beam is reproduced in the spatial pattern displayed on the TV screen. A variety of different signals from the electron–sample interaction can be measured using different detectors in the SEM, including backscattered electrons and secondary electrons released from the sample itself, as well as x-rays and emitted light. Although the resolution of this method is much lower (~10 nm at best) than the TEM, the depth of focus is extremely large and the images are very three-dimensional and lifelike (Figure 23.19).

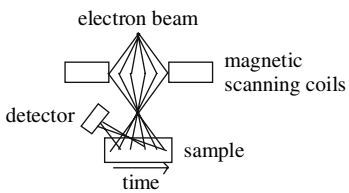


FIGURE 23.18 *Schematic diagram of the final portion of the scanning electron microscope showing the scanned electron beam, in multiple images, steered by magnetic scanning coils and the backscattered electron detector.*

STEM was developed to try to collect not only the "reflected" electrons and radiation as in SEM, but also the transmitted electrons that have interacted with the sample. These transmitted electrons undergo two basic types of interactions, elastic and inelastic, aside from the bulk of the electrons that simply pass through without any interaction at all. Inelastically scattered electrons lose some energy to the sample through excitation of target atoms, whereas elastically scattered electrons, fewer in number, are simply deflected from their path through much larger angles by interaction with the nuclei of target atoms without a change in their energy. The ratio of the intensities of the elastic to inelastic electron scattering is a characteristic of the particular target atom and increases with the number of protons in the nucleus of the atom. STEM scans an even more tightly focused electron beam (~0.5 nm) across the sample simultaneously measuring the elastic and inelastic transmitted electron intensities. Furthermore, the inelastically scattered electrons can be energy-analyzed to determine their energy loss. STEM pictures are at very high resolution (Figure 23.20)

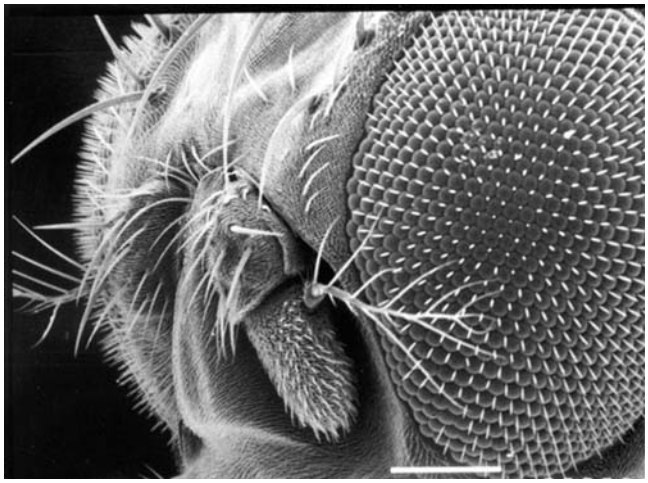


FIGURE 23.19 *SEM of the head of a house fly at 200 X magnification (the bar is 100 μm). The structure on the right is a multifaceted eye.*

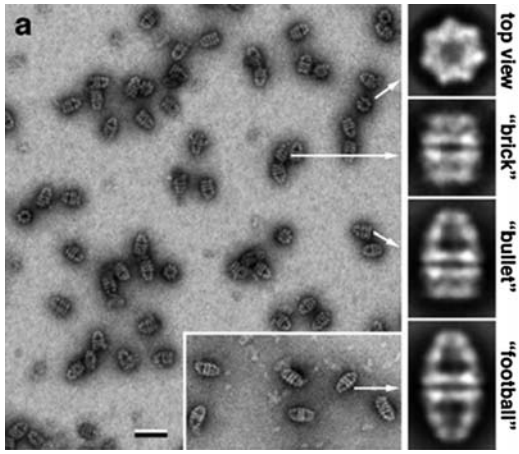


FIGURE 23.20 *STEM images of a particular chaperonin, one of a family of large (~10^6 Da) complexes involved in the folding of proteins, under different solvent conditions showing images used to reconstruct the detailed images of shape shown in the insets on the right. The bar represents 20 nm.*

and can also determine the elemental composition of the sample from point to point. Unfortunately, the fundamental limitation of sample degradation in the electron beam has made STEM less useful in biological imaging than first expected when developed in the 1970s.

4. X-RAYS: DIFFRACTION AND COMPUTED TOMOGRAPHY (CT)

X-ray photons have wavelengths in the range from about 0.01–10 nm, short enough to provide atomic resolution according to the equation for resolving power. Unfortunately, until recently x-rays could not be easily focused and magnified images, such as have been made with light and electron beams, have not yet been produced with x-rays. (In 1996 scientists developed a simple and effective way to focus x-rays; this method is expected to lead to many new applications, particularly in microelectronics.) Even if we had the ability to focus x-rays, their interaction with biological tissue is so weak that there would be virtually no contrast seen in normal thin samples used in microscopy. However, x-rays have two properties that make them extremely useful in both medicine and science. First, because x-rays are a form of electromagnetic radiation, they diffract from objects of comparable dimension to their wavelength, similar to the diffraction of light. Because of their atomic-sized wavelength, x-ray diffraction effects can be used to probe the atomic structure of matter and have been used to determine the structure of many complex biological macromolecules at atomic resolution. Second, because x-ray energies are high, these photons are capable of passing through otherwise opaque materials and x-rays can be used to produce "shadow" pictures of internal structures within thick samples, for example, the human body.

Crystalline materials have a three-dimensional periodic array of their atoms that can diffract x-rays and produce a pattern of detected x-rays containing information about the spatial array of the atoms. In a similar way that a one-dimensional array of slits gives rise to a diffraction pattern with light, the crystalline array of atoms results in a more complicated pattern of diffracted x-rays. In this case the x-rays are scattered, or diffracted, in all directions from the crystalline array of atoms and interference effects result in a detected pattern of x-ray spots.

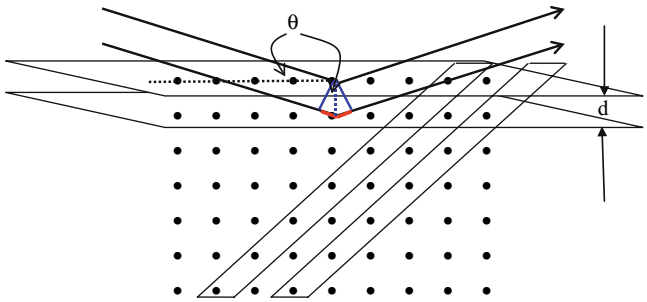


FIGURE 23.21 *Cross-section of a cubic lattice (shown in two dimensions) showing two sets of Bragg planes with different spacings and the diffraction of an x-ray beam from one set with spacing* d. *The extra path difference of the lower beam is shown in red and is equal to* d *sin* θ *for each of the triangles shown, totaling* 2d *sin* θ.

Consider a simple cubic crystal made of identical atoms in a periodic array, or lattice, with separation distance d as shown in cross-section in Figure 23.21. The atoms form planes, known as Bragg planes, and the pattern of diffracted x-rays can be determined by imagining that the x-ray beam reflects from these planes in a process known as Bragg diffraction. This picture greatly simplifies the analysis but gives the correct general result. For the x-ray beams shown in the figure, there will be a path difference for beams reflecting from neighboring planes. From the figure, we see that this path difference will depend on the angle θ between the ray and the Bragg plane and is given by $2d \sin \theta$. (Note that θ is not the usual angle of reflection between the ray and the normal, but is the angle between the ray and the line of atoms in the plane of reflection.) Constructive interference will occur when this path difference is equal to a whole number of wavelengths and the *Bragg equation,*

$$m\lambda = 2\, d \sin \theta, \tag{23.6}$$

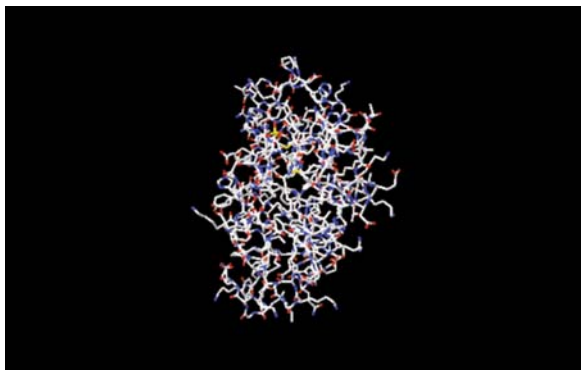
where m is an integer called the order, defines the location of an interference maximum. X-rays incident at an angle given by Equation (23.6), known as a Bragg angle, will produce a diffraction peak, or spot, at some distant detector located at the "reflected" ray. In a noncubic crystal with three different repeat distances along different directions there will be two additional order numbers for the other directions and a generalized Bragg equation. In this case, the "unit cell," or basic repeating structure, dimensions can be found by the location of the Bragg spots.

Example 23.2 In an x-ray diffraction experiment on a cubic crystal with $\lambda = 0.40 \times 10^{-10}$ m, find the crystal plane spacing if the first-order maximum occurs at an angle of 6.4°. At what angle will the third-order maximum be found?

Solution: Using Equation (23.6) with $m = 1$, we have that $d = \lambda/(2 \sin \theta) = 1.79 \times 10^{-10}$ m. The third-order maximum will then be found at the angle given by $\sin \theta = 3\lambda/2d = 0.34$, so that $\theta = 19.6°$.

In the study of macromolecular structure, if a crystal of the macromolecule can be formed, then x-ray diffraction can often be used to determine the three-dimensional arrangement of all its atoms. In such a crystal, the individual scattering centers, or unit cells, may consist of thousands of individual atoms. In addition to the unit cell dimensions affecting the observed diffraction pattern, x-ray scattering from the molecules within the unit cell will affect the pattern due to its "structure factor." In general, if there are N atoms per unit cell, there will be N^2 peaks in the diffraction pattern from the atoms within the unit cell. As N increases for larger molecular crystals, the diffraction patterns become extremely complex and rich with information. From detailed studies of such patterns, together with as much independent information on the macromolecular structure as possible, the detailed three-dimensional atomic arrangement has been determined for many macromolecules. Figure 23.22 shows the three-dimensional structure of myoglobin, a subunit of hemoglobin consisting of 153 amino acids with a total of

FIGURE 23.22 *The structure of myoglobin.*



1260 atoms. To obtain the current resolution of better than 0.2 nm, more than 9600 diffraction spots were measured and analyzed. These pictures of the structure conceal the fact that most macromolecules have extensive flexibility and motion. Because the x-ray pictures are obtained over relatively long times, the resulting 3-D structures represent average positions of the constituent atoms. X-ray diffraction is one of the best methods we have for determining macromolecular structure at an atomic resolution.

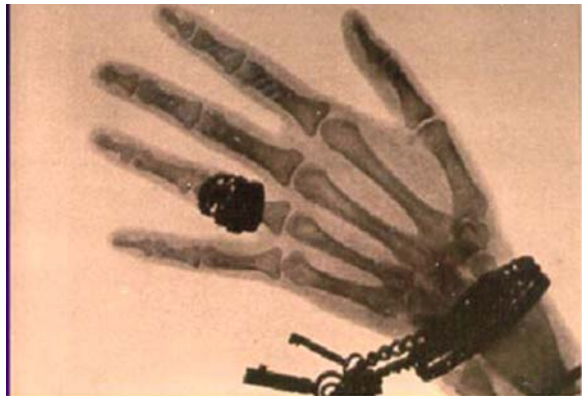


FIGURE 23.23 *The first x-ray picture, obtained in 1896, of the hand of Mrs. Roentgen.*

Not all biological materials can be made to crystallize so that they can be studied by x-ray diffraction. A large class of filamentous macromolecules can, however, be oriented into fibers and studied by x-ray diffraction even though they are not in regular crystalline arrays. Special techniques have been developed for helical proteins and nucleic acids that reveal the symmetries present even when neighboring oriented helices may not be "in register" along the axial direction. Such methods first led to the structural determination of the helical nature of DNA by Watson and Crick and to the basic ideas on how DNA transmits genetic information.

On the much larger dimensional scale of human organs and internal structures, x-rays penetrate though skin and other soft tissue and travel in straight lines without diffraction. In this geometrical optics limit, they can be used to produce shadow images of, for example, bones within the body, based entirely on differences in absorption of x-rays. In fact, when Roentgen first discovered x-rays in 1895, within a week he had obtained the first x-ray picture of a hand (Figure 23.23). The depth of penetration of x-rays depends on the density of the material; denser materials, such as lead, are more effective in absorbing x-rays. Medical technology uses x-rays to obtain pictures of such structures as bone and teeth in x-ray radiography. Softer tissues can be pictured best if a dense material is introduced to increase the contrast. The gastrointestinal tract can be imaged if it is filled with a dense barium solution that casts a shadow in an x-ray picture. Similarly water-soluble organic compounds with iodine are used to give contrast for pictures of the cardiovascular system, the urinary tract or the brain. Mammography can be done without a contrast agent using low-energy x-rays because these give the greatest contrast for soft tissues.

These pictures produce two-dimensional projection images, lacking resolution along the beam direction because the intensity of the x-rays at the detector is determined by an integration or sum through the body along the beam. Thus three-dimensional information is lost on conversion to a two-dimensional picture. Put another way, there is no depth information in an x-ray picture and doctors must infer relative depths of neighboring features in these pictures with much care. Furthermore, it is more difficult to detect small differences in x-ray absorption at neighboring points because there is no resolution along the beam and therefore many minor abnormalities in x-ray radiography are not detectable.

To improve this situation, computed tomography (CT; the Greek word *tomo* means cut or slice) is able to obtain three-dimensional information from a collection of x-ray pictures taken at different orientations. The original CT machines developed in the 1970s used a single x-ray source and detector held in precise register on opposite sides of a patient. These were translated across the sample region, rotated by 1° and scanned across the sample again, and so on, in steps all around the body, so that a sequence of many pictures was obtained in a few minutes that could then be used to reconstruct the depth information in a three-dimensional image. Today, CT machines use a wide fanlike beam and an array of several hundreds to a few thousand x-ray detectors to decrease the time required to a few seconds (Figure 23.24). The newest designed machines have stationary detector arrays with an x-ray beam made to sweep in a circular pattern around the patient with no moving parts. We show in Chapter 25 that x-rays are generated by

FIGURE 23.24 *Modern CT machine used in hospitals and medical imaging facilities.*

transitions of outer electrons to inner empty electron shells after the inner electrons have been ejected by bombardment with high-speed electrons. In modern CT machines a scanning high-energy electron beam that generates the x-rays is an integral part of this design. Projection data can then be obtained in about 50 ms, fast enough to image a beating heart without motion artifacts.

However the various projection data are recorded, a computer will have a record of digitized intensities that needs to be processed to reconstruct the image of a cross-sectional slice in the body. The image consists of a large number of two-dimensional spots, or pixels, each having some grayscale level, a digital value representing the brightness. Grayscale displays are sometimes converted to false color images where the colors represent the brightness level but have nothing to do with the color of the original tissue being imaged. These brightness scales are normally set according to the absorption coefficient of the tissue ε, compared to that of water, ε_w, by the CT number

$$\text{CT number} = 1000\,\frac{\varepsilon - \varepsilon_w}{\varepsilon_w}. \tag{23.7}$$

Table 23.1 shows the CT numbers for different tissue and media for 60 keV x-rays. The absorption coefficients of tissue depend on the x-ray beam energy and corrections usually need to be made for this fact. Note that negative CT numbers indicate that there is less absorption of x-rays than in water.

Table 23.1 *CT Numbers for Various Materials**

Material	*CT Number*
Water	0
Air	−1000
Bone	808
Striated muscle	−48
Fat	−142

* Using 60 keV x-rays.

We briefly try to give the reader a sense of how projection data can be used to determine the CT numbers for an array of pixels in order to generate a cross-sectional

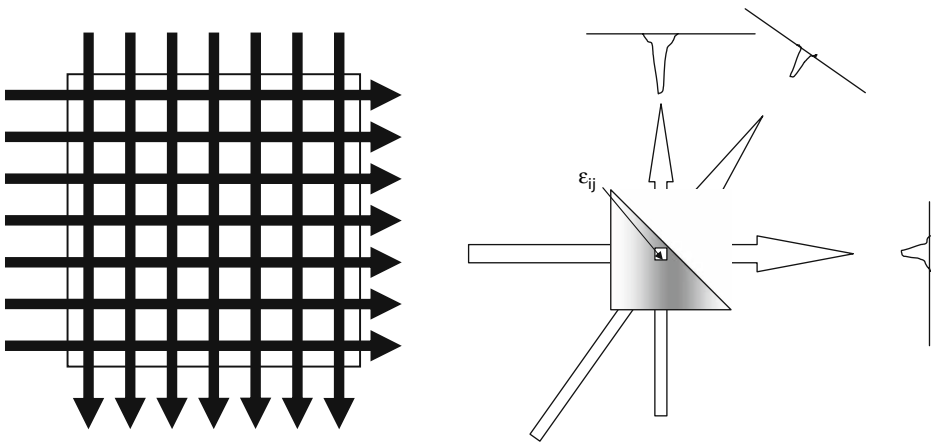


FIGURE 23.25 (left) *An $N \times N$ grid of pixels defined by sets of parallel beams. The transmited (projected) intensities are used to reconstruct the absorption coefficients of each pixel and thus a two-dimensional image based on x-ray absorption.* (right) *A test object with varying shape and absorption coefficient* (symbolized by shading) *probed by several x-ray beams to do a back-projection determining the absorption coefficient ε_{ij} of the overlap region.*

picture of the body based on x-ray contrast. In our context, the absorption coefficient is in the relation

$$I = I_o\, e^{-\varepsilon x} \quad or \quad \log\frac{I_o}{I} = \varepsilon x, \tag{23.8}$$

where I and I_o are the transmitted and incident intensity on a tissue thickness x and the log is to base e. Each x-ray beam can be imagined to have traveled through a distance x in the body and the transmitted intensity detected. We imagine that each of N such neighboring parallel beams (the rows) is divided into N intervals of length x/N (the columns), forming a two-dimensional cross-sectional grid of N rows by N columns, with N typically in the range 256–1024. In the pixel display of this slice, the term εx in Equation (23.8) for the ith row, for example, is given by the sum

$$\varepsilon_i\, x = \sum_{j=1 \text{ to } N} \varepsilon_{ij} \Delta x\,,$$

where we have labeled the ε_{ij} values according to the pixel number (ith row and jth column) and have assumed that the pixel width, $\Delta x = x/N$, is the same in any direction. In the simplest case, imagine that two sets of parallel x-ray beams are used to define a square grid as shown in Figure 23.25 (left) and that the projected (transmitted) intensity is measured for each beam. Using values for the projected intensities, computer algorithms can determine the ε_{ij} for the $N \times N$ pixels, giving a two-dimensional absorption image.

In general, more complex patterns of beams can be used (Figure 23.25 right). Because a set of $N \times N$ pixels is needed to image a given plane, a minimum of N^2 values for ε_{ij} are needed. These can be obtained from at least that many data points for $\log I_o/I$, or εx, obtained by imaging the same region of the body at many, many different orientations. Large numbers of equations must be simultaneously solved on a computer; with $N = 256$, there are at least $N^2 = 65{,}500$ equations to solve. Various computational techniques have been developed to do these calculations rapidly.

With current technology, multiple cross-sectional images can be rapidly obtained and computer techniques allow these to be superposed to produce 3-D images (Figure 23.26). These same tomography methods can be applied to other types of imaging, including ultrasonic (Chapter 11), magnetic resonance (Chapter 18), and to such nuclear decay imaging as positron emission tomography (PET; discussed in Chapter 26). The quality of images from CT and MRI scans are often comparable and the choice of method depends on the type of tissue to be imaged.

FIGURE 23.26 *Three-dimensional rendering of a human heart by CT imaging.*

CHAPTER SUMMARY

Contrast is the other major factor, in addition to resolution discussed in the previous chapter, that determines whether an object can be imaged in a microscope. We can distinguish two types of contrast: amplitude and phase. Microscopes that use amplitude contrast include the standard bright-field compound microscope discussed in the previous chapter, as well as the dark-field and fluorescence microscopes. Phase contrast and DIC (differential interference contrast) microscopes use phase contrast to image objects. Newer microscopies use laser-scanning methods to do point-by-point imaging. These include confocal and multiphoton microscopies.

Optical activity refers to the effect of anisotropic molecules on the circular polarization of light. Such

(*Continued*)

molecules have a different effective index of refraction for left- and right-handed circularly polarized light and are said to have circular birefringence

$$\Delta n = (n_L - n_R). \qquad (23.4)$$

They also absorb left- and right-handed circularly polarized light differently and produce circular dichroism,

$$\Delta\varepsilon = \varepsilon_L - \varepsilon_R\,, \qquad (23.5)$$

where ε is the absorption coefficient. This effect can be measured using the optical technique of circular dichroism CD, and is an important method to determine the percent of helix, beta sheet and random coil composition of macromolecules.

Transmission electron microscopy uses a high-energy beam of electrons to produce a "shadow" image of microscopic objects with a resolution approaching atomic resolution. Scanning electron microscopy (SEM) is a lower-resolution variation that scans a tightly focused electron beam over the sample and detects backscattered, rather than transmitted, electrons. This method gives a greater depth of focus so that the images look three-dimensional. A less used method combines these methods in the high-resolution scanning transmission EM (STEM).

X-rays can be used to study the structure of crystals, even crystals made from complex macromolecules. The basic crystalline array can be determined using the Bragg equation

$$m\lambda = 2\,d\sin\theta, \qquad (23.6)$$

where λ is the wavelength of the x-ray beam, d is the spacing between Bragg planes of the crystal, and θ is the diffraction angle (between the Bragg plane and the incident or exit beam). X-ray beams can be used to produce shadow images through the body because transmission through bone and types of tissue are different. The medical imaging technique known as computed tomography (CT) uses fanlike x-ray beams and multiple detectors to allow images to be reconstructed by computer of cross-sections through the human body at a resolution of about 1 mm.

QUESTIONS

1. Compare image contrast with resolution for a bright-field microscope. How does each enter into producing an image?
2. What is the function of the dichroic mirror in a fluorescent microscope? (See Figure 23.3.)
3. What are the origins of phase and amplitude contrast? Are both always present to some extent?
4. Describe the main differences, in your own words, between phase contrast and differential interference contrast microscopy.
5. What is the function of the Wollaston polarizing prisms in DIC optics? Is the fact that the two beams have different polarizations important in the final image seen?
6. What are the advantages of multiphoton microscopy over single-photon methods?
7. Discuss the superposition of two linear polarized light beams of the same frequency and equal amplitude, one polarized along the x- and one along the y-axis. What is the result if the two are in phase? 90° out of phase? 180° out of phase?
8. Because a plane mirror reverses left and right, but does not reverse up and down, if you hold a coiled right-handed spring and look at its image in a mirror is there an orientation of the spring that results in a right-handed image?
9. Simple molecules produced in chemical reactions, even if they have a handedness, are usually produced in nearly equal quantities of left- and right-handed molecules. Biological molecules, on the other hand, are nearly always found in pure left-handed form. What benefits might be derived from only having one form in living materials?
10. A linearly polarized light beam passes through a birefringent material and two beams emerge. If the beams are each made to pass through one slit of a double-slit experiment, will a standard double-slit interference pattern be produced on a distant screen?
11. What is the difference between circular birefringence and circular dichroism?
12. As the accelerating voltage in an electron microscope is increased, what happens to the theoretical magnification? To the sample degradation? To the magnetic field needed to focus the electron beam?
13. What is the purpose of heavy metal deposition in TEM? How does it affect resolution?
14. Can you argue why the backscattered electrons in SEM allow the images to appear much more three-dimensional than the images transmitted electrons produce in TEM?
15. Fill in the details in the derivation of the Bragg equation, Equation (23.6), using Figure 23.21.
16. Why, when you have a dental x-ray taken, are you covered with a heavy lead-coated gown?
17. Contrast how a CT image is obtained with how you perceive depth with two eyes.

MULTIPLE CHOICE QUESTIONS

1. In dark-field microscopy (a) the sample images darker than the background, (b) an annular aperture is inserted between the sample and the objective lens, (c) the image contrast is usually better than that of bright-field, (d) the samples must be stained to show up.
2. Fluorescent dyes can be used for all but which of the following? (a) Imaging calcium concentration variations, (b) imaging pH variations, (c) localizing specific molecules, (d) high-resolution imaging of molecules.
3. Which of the microscopic techniques usually requires that the sample be stained? (a) Phase contrast, (b) bright field, (c) DIC, (d) polarizing microscopy.
4. In DIC microscopy, the edges of microscopic objects are sharp because (a) that's where the most stain is, (b) that's where there is an extra π phase shift, (c) that's where there is the greatest change in index of refraction, (d) that's where the greatest polarization difference occurs.
5. In three-photon microscopy, to excite a fluor at 450 nm the incident wavelength of light should be (a) 150 nm, (b) 450 nm, (c) 900 nm, (d) 1350 nm.
6. In laser-scanning confocal microscopy all of the following are true except (a) the beam is focused to a very small spot, (b) the beam is moved across the sample, (c) two or more photons are absorbed at the same time, (d) the images appear three-dimensional.
7. A circularly polarized beam of light (a) travels in a spiral around its magnetic field, (b) travels in a spiral around its propagation direction, (c) has an electric field vector whose tip rotates in a closed circle, (d) has an electric field vector whose tip travels in a spiral.
8. Which is not true of a birefringent material? (a) It must be a solid because it has different indices of refraction along two different directions, (b) it can produce two beams of light from one, (c) it can produce circularly polarized light, (d) light can travel through it with two different speeds.
9. Which of the following is not true of an optically active molecule? (a) It produces a circular birefringence signal, (b) it produces a circular dichroism signal, (c) it must be asymmetric, (d) a solution of them can always be imaged in a polarizing microscope.
10. A typical accelerating voltage used in an electron microscope is (a) 100 kV, (b) 1 kV, (c) 10 MV, (d) 100 V.
11. Electron microscope samples must be stained or metal-coated because (a) the atoms are too small to detect otherwise, (b) the samples are not colored otherwise, (c) the samples do not interact with electrons otherwise, (d) the samples would evaporate from the grid otherwise.
12. All of the following are consequences of using high accelerating voltages and small focused spot sizes in scanning electron microscopy except (a) higher resolution, (b) decreased heating of the sample, (c) increased backscattered electrons, (d) more accurate elemental analysis.
13. Which of the following is not true? 60 keV x-rays are absorbed by (a) water more than fat, (b) water more than air, (c) bone more than striated muscle, (d) fat more than striated muscle.
14. The intensity remaining in a beam after traveling 10 cm through a sample with an absorption coefficient of 0.2 cm^{-1} is (a) 1%, (b) 1.4%, (c) 14%, (d) 20%.

PROBLEMS

1. With a compound microscope adjusted poorly, the % contrast for a certain sample is only 5%. If the microscope is adjusted and the sample intensity is reduced by 10% and the background intensity is increased by 20%, what is the new % contrast?
2. In three-photon microscopy, if the peak in the absorption band of a fluorescent molecule to be imaged is at 360 nm, what incident frequency of light should be used?
3. Show that two in-phase linearly polarized beams with the same frequency but along orthogonal axes (x and y) superpose to produce a linearly polarized beam with a polarization direction that depends on the ratio of their amplitudes. What is this polarization angle if $E_{ox} = E_{oy}$? If $E_{ox} = 3E_{oy}$?
4. Show that the tip of the electric field vector produced by the superposition of equal amplitude electric fields given in Equation (23.2) rotates in a circle. Viewed from a location at which the beam is approaching you, does the E vector rotate clockwise or counterclockwise?
5. A birefringent crystal has a birefringence given by $\Delta n = n_1 - n_2 = 0.01$, where n_1 and n_2 are the indices of refraction along its two transverse crystal axes at right angles with each other. Suppose a linearly polarized wave with 550 nm wavelength, is polarized at 45° to the crystal axes. If the crystal has a thickness of 1 cm, what will be the path difference between the two waves polarized along the crystal axes when they exit the crystal? What will be the net phase difference (as a fraction of 2π rad, or modulo 2π rad) of the two waves?
6. Suppose that the spot size in an SEM is 10 nm and that the beam is scanned over a region of 100 μm × 100 μm in a raster pattern, producing a single-scanned image in 10 ms. If the overall region is digitized into a 200 × 200 pixel area,
   (a) What sample area is represented by 1 pixel?
   (b) How long is the beam exposure in each pixel? (This determines resolution time of the detector.)
7. X-rays with a 0.12 nm wavelength produce a first-order diffraction peak at a Bragg angle of 24°. What crystal spacing gave rise to this diffraction?
8. A cubic crystal with identical atoms separated by distance d has sets of Bragg planes separated by distance d. It also has other symmetry planes, as shown,

for example, in Figure 23.21. Using simple trigonometry, draw a two-dimensional square lattice projection of the crystal (as in Figure 23.21) and find two other crystal plane spacings in terms of *d*.

9. If an x-ray beam is incident on a 1.5 cm thick sample and 98% of the beam is transmitted what is the average absorption coefficient of the material in units of m^{-1}?

10. Two samples for an x-ray absorption experiment have the same thickness. With the same incident intensity one has a 95% transmission and the second has an 85% transmission. What is the ratio of their absorption coefficients?

11. Suppose that an x-ray beam is directed on a tissue sample and suppose that 99.3% of the beam is transmitted. If a dummy blank sample of water is used 99.5% of the x-rays are transmitted using exactly the same geometry and beam. What is the CT number of this sample?

24

Special Relativity and Quantum Physics

This and the next chapter bring together a number of fundamental concepts about matter and radiation, some of which we have anticipated in previous discussions when needed. The title of this chapter names the two major theories developed in the 20th century that have most revolutionized physics. Relativity and quantum physics are together sometimes known as modern physics. We begin this chapter with a brief discussion of some aspects of special relativity, a theory developed by Albert Einstein that has brought about major changes in our understanding of the world. Thoroughly tested and consistently found to be correct, special relativity forms a framework on which modern physics rests. The chapter then continues with an overview of the probabilistic view of nature demanded by quantum physics, illustrated by a revisiting of the double-slit experiment. Some of the main features of quantum physics are then discussed, including the Schrödinger equation and the uncertainty principle. The chapter concludes with a discussion of the quantum basis of scanning tunneling microscopy, capable of viewing individual atoms. Our discussion continues in the next chapter with the quantum physics of atoms and molecules and their study by spectroscopy, including the laser which is one of the most important tools in science and medicine today.

1. SPECIAL RELATIVITY: MASS–ENERGY AND DYNAMICS

Special relativity is concerned with our fundamental notions of time, space, mass, energy, and motion at constant velocities. Albert Einstein published the theory of special relativity in 1905 when he was 26 years old. In that same year he also published fundamental papers on Brownian motion and on the photoelectric effect, discussed below, for which he received the Nobel Prize. Twelve years later, in 1917, he published the theory of general relativity, which quantitatively shows the equivalence between accelerated motions and gravity, known as the equivalence principle, replacing the gravitational force with a curvature of space and time. Although Einstein's theory of general relativity has been successfully tested and accepted today, those tests are relatively few in number and its impact on physics is much more limited than that of special relativity. One important everyday application of general relativity is a correction needed for the extremely accurate time keeping required for GPS (global positioning system; Figure 24.1); without general relativity corrections, GPS navigational errors would be about 10 km per day. Special relativity, on the other hand, has been thoroughly tested and is completely ingrained in all areas of modern physics.

Relativity is often thought to be mathematically complex, but it is only general relativity, not discussed here, that involves higher mathematics. Special relativity can be explained without the use of much mathematics and so can be understood by the nonscientist, but it involves ideas that seem contrary to our intuition. We live in a world of extremely slow moving objects compared to the speed of light. Relativity (from now on we omit the word "special" because we limit our discussion to special relativity) deals

J. Newman, *Physics of the Life Sciences*, DOI: 10.1007/978-0-387-77259-2_24,

FIGURE 24.1 *A handheld GPS only works with corrections from general relativity.*

with new phenomena that occur at speeds approaching the speed of light. We have no intuitive basis for understanding such processes since we never experience motion at such speeds. Even a plane traveling at 600 mph travels only at about 1 millionth the speed of light. All of the equations of relativity reduce to equations we have already studied when the speeds of objects are small compared to the speed of light, as we shall see.

Two fundamental postulates form the basis of relativity theory from which all its consequences follow. The first, known as the *principle of relativity*, is that all the laws of physics are the same in all inertial frames of reference. We have already seen an example of this principle in mechanics in the form of Newton's first law. Relative velocities may be different in two different inertial frames, however, accelerations of objects and the description of the forces acting to produce motion will be the same in all inertial reference frames. Einstein's relativity principle extends this notion to cover all the laws of physics, not just those of mechanics. The second postulate concerns the constancy of the speed of light and states that the speed of light in vacuum has the same value c in all inertial reference frames. It is remarkable that these two postulates alone lead to the development of such a powerful theory. We limit our discussion here to those salient features of dynamics that we need later in this book, omitting the fascinating consequences of relativity on our notion of space and time.

Consider a point particle of mass m, moving with a velocity v in the x-direction as seen by an observer. Classically, the momentum of the particle would be defined as $p = mv = m(\Delta x/\Delta t)$, where Δx is the displacement of the particle in a time interval Δt. In place of this, the *relativistic momentum* is defined as

$$p = \frac{mv}{\sqrt{1 - v^2/c^2}} = \gamma mv, \tag{24.1}$$

where the Lorentz factor γ is defined by

$$\gamma = \frac{1}{\sqrt{1 - v^2/c^2}}.$$

Although for a stationary particle $\gamma = 1$, even when the particle moves at 0.1 c, quite a large velocity, the value for γ is only 1.005. Figure 24.2 shows how γ varies with the ratio v/c, confined to lie between 0 and 1; note that γ grows very rapidly as v approaches c. This formula can be directly generalized to three-dimensional motion by treating p and v as vectors.

Note that for small values of v we can neglect the term v^2/c^2 in the denominator of Equation (24.1) so that the expression for momentum reduces to its classical value. As v approaches c, however, the momentum of the particle, being proportional to γ, increases at a much faster rate than the classical linear dependence on v.

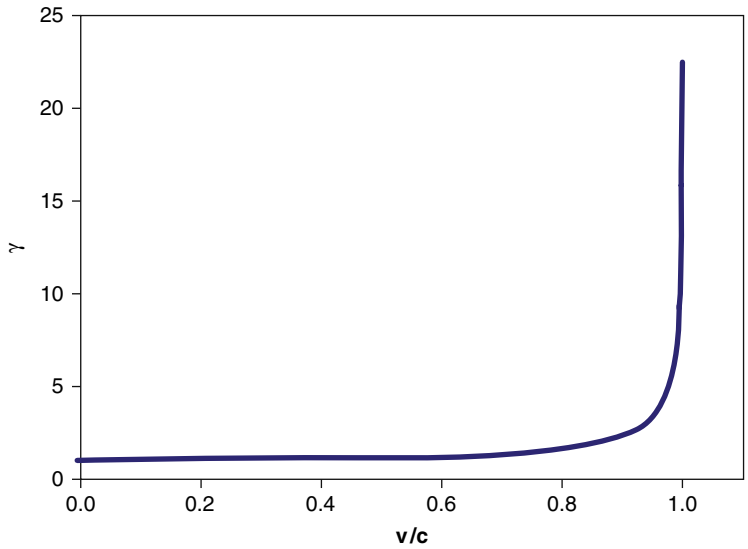


FIGURE 24.2 *The factor γ versus v/c showing its rapid rise as v approaches c.*

Because the momentum increases so rapidly as the particle's velocity approaches c, it requires an ever-increasing force, equal to the rate of change of momentum, to accelerate the particle. If the particle starts from rest, an increase in its velocity by 10% of the speed of light will produce a proportional momentum increase of just about 10%. However, if the particle is already moving at half the speed of light, the change in momentum for a 0.1 c increase in velocity (a 20% increase, from 0.5 c to 0.6 c) will be about 30%, whereas if the particle is already moving at 85% of the speed of light, the corresponding increase in momentum for the same 0.1 c increase (about a 12% increase from 0.85 c to 0.95 c) will be almost 90%. As the velocity of the particle approaches c, its momentum increases very rapidly, and therefore the change in momentum needed to produce the same step increase in its velocity will also dramatically increase. Because an ever-increasing force is needed to increase the particle's momentum, this effect prevents a material particle (one with a nonzero mass) from ever attaining a velocity equal to the speed of light.

Another important variable of dynamics is the kinetic energy, classically given as $\text{KE} = \frac{1}{2}mv^2$. The *relativistic kinetic energy* expression looks quite different and is given by

$$\text{KE} = \frac{mc^2}{\sqrt{1 - v^2/c^2}} - mc^2 = \gamma mc^2 - mc^2. \tag{24.2}$$

This is indeed an energy that depends on motion because if $v = 0$, then $\gamma = 1$ and the expression clearly reduces to KE = 0. Although it is not apparent that for small velocities compared to c this reduces to the classical expression, we can show this by expanding the square root term in Equation (24.2) using the binomial theorem

$$(1 - x^2)^{-\frac{1}{2}} = 1 + \frac{x^2}{2} - \ldots, \tag{24.3}$$

valid for $x << 1$ to find that

$$\text{KE} = mc^2\left(1 + \frac{v^2}{2c^2}\right) - mc^2 = \frac{1}{2}mv^2. \tag{24.4}$$

Therefore, as long as $v/c << 1$, we see that the relativistic kinetic energy reduces to our usual classical physics expression. The relativistic expression for kinetic energy also confirms the idea that it becomes more and more difficult to accelerate a particle of mass m as its speed approaches c because the kinetic energy also grows very rapidly, in proportion to γ.

Now we make a leap in our interpretation of Equation (24.2).

We define the total relativistic energy of the particle to be the first term in Equation (24.2)

$$E = \frac{mc^2}{\sqrt{1 - v^2/c^2}} = \gamma mc^2. \tag{24.5}$$

Then, from Equation (24.2), we can rewrite this as

$$E = \text{KE} + mc^2. \tag{24.6}$$

The total relativistic energy of a particle is therefore made up of its kinetic energy, the first term, and its rest energy, given by mc^2, the energy remaining when $v = 0$ or $\gamma = 1$.

Example 24.1 For an electron traveling at $v = 0.95\,c$, find its momentum, total energy, and kinetic energy and express each of these as multiples of their classical (nonrelativistic) values.

Solution: An electron has a rest mass of 9.1×10^{-31} kg and, at $v = 0.95\,c$, a γ value of $\gamma = \dfrac{1}{\sqrt{1 - (0.95)^2}} = 3.2$.

Therefore its momentum is equal to $p = \gamma mv = (3.2)(9.1 \times 10^{-31})\ (0.95)\ (3 \times 10^8) = 8.2 \times 10^{-22}$ kg-m/s, its total energy is equal to $\gamma mc^2 = 2.6 \times 10^{-13}$ J, and its kinetic energy is $E - mc^2 = 1.8 \times 10^{-13}$ J. Because the classical momentum is just mv, the relativistic momentum is exactly a factor of $\gamma = 3.2$ larger. Classically the total energy and kinetic energy are both equal (because there are no potential energies for an isolated electron) and equal to $\frac{1}{2}mv^2 = 3.7 \times 10^{-14}$ J, and thus the total energy and kinetic energy are actually larger than this by factors of 7.0 and 4.9, respectively.

Einstein's famous formula $E = mc^2$ is really a relation between the rest energy and mass of a particle and shows the equivalence of mass and energy. A particle and its antiparticle, for example, an electron and a positron, each with the same mass, can annihilate on collision converting all of their mass to pure energy in the form of photons (as long as all conservation laws are satisfied). In the inverse reaction, known as pair production, a gamma ray photon with enough energy can create an electron and positron pair. In this case if the photon has an energy greater than the combined rest mass of the electron and positron, then these two particles share the remaining energy in the form of kinetic energy as they fly apart at some appropriate speed.

Nuclear reactions involve small changes in the mass of nuclei with accompanying large changes in energy given by

$$\Delta E = \Delta mc^2. \tag{24.7}$$

For example, a mass change of 1 kg leads to an energy release of about 9×10^{16} J, or enough energy for a U.S. city of $\frac{1}{2}$ million people for a year. Even chemical reactions involve small mass changes of the reacting atoms, although the equivalent energies are much smaller than those of nuclear reactions.

Another way to consider Equations (24.1) and (24.5) for the relativistic momentum and energy of a particle is to consider the term

$$\frac{m}{\sqrt{1 - v^2/c^2}} = \gamma m$$

as a variable, known as the *relativistic mass* (with m known as the rest mass), that depends on the speed of the particle. Viewed in this way, the (relativistic) mass of a particle increases dramatically with speed. This provides an alternative explanation of why it is impossible to surpass the speed of light. The faster a particle travels, the more massive it becomes and the more difficult it becomes to keep it accelerating. Because the relativistic mass grows, unbounded, as the speed approaches c, no finite force can accelerate the material object to the speed of light.

We conclude this section by showing the connection between energy and momentum. Classically, kinetic energy and momentum are related by $\text{KE} = p^2/2m$. With some algebra (see Problem 4), we can show that the relativistic momentum and energy are related by

$$E^2 = p^2c^2 + m^2c^4. \tag{24.8}$$

For a massless particle, such as a photon or neutrino, the rest energy term vanishes and the energy and momentum are proportional to each other

$$E = pc. \quad (\text{if } m = 0 \text{ or } \gamma >> 1). \tag{24.9}$$

This same expression holds for ultrarelativistic massive particles, whose speeds approach c so that $\gamma >> 1$, because the first term on the right in Equation (24.8) dominates and we can neglect the second rest energy term.

The ideas we have developed in this section are used in the remainder of this book in various discussions of modern physics. Relativity also deals with other concepts related to motion at large constant velocities, including fundamental changes in our notion of distance and time. These we leave for the interested reader to find in any one of a large number of popular books that discuss special relativity, including one by Albert Einstein himself.

2. OVERVIEW OF QUANTUM THEORY

We now take a veritable quantum leap and begin considering our current understanding of the atomic world of nature. Earlier in this book we have seen the notion of *wave-particle duality*, that in nature the elementary constituents of matter and radiation can appear to behave as either particles or waves, depending upon the interactions with their environment. For example, photons, the elementary quanta of radiation, can behave as waves (in interference and diffraction), or, in other situations as we soon show, photons can behave as particles. The wave packet picture was introduced in Chapter 19 as a way to visualize this duality, with the wave packet capable of collapsing to be more particlelike or expanding to be more wavelike in space depending on its interactions. Here we discuss this in more general terms and show that the picture also applies to all other elementary "particles" and sometimes even to macroscopic systems. We discuss a series of different experiments that illustrate the wave–particle duality nature of photons and other elementary "particles" such as electrons.

The *photoelectric effect* is a very important process in which light causes the emission of electrons from a metal surface. This phenomenon is the basis for a variety of light-detecting devices that produce electric currents in response to light. Many of the features of the interaction of light with a metal surface could not be explained on the basis of a wave theory of light and these led Albert Einstein to propose a theory of the photoelectric effect in 1905 based on photons.

When light is directed on a metal cathode (negative electrode) within a vacuum tube, as shown in Figure 24.3, an electric current can be generated at the anode (positive electrode) when a potential difference is applied across the electrodes to collect the emitted electrons, even though there is no wire connected between the two electrodes. According to the wave theory of light, the intensity of light should be proportional to the beam energy, and for a

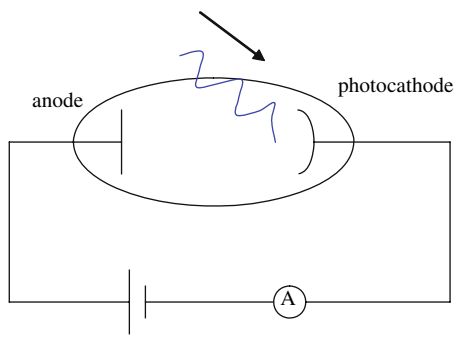


FIGURE 24.3 *The photoelectric effect. Light incident on the photocathode electrode in a vacuum tube causes electrons to be ejected and attracted to the anode (by a positive potential) to make up a current measured by the external ammeter.*

sufficiently intense beam, no matter what the wavelength, one should expect electrons to be ejected from the metal surface after gaining energy from the light. Indeed, for shorter wavelength light, the electric current is proportional to the intensity of the light. However, if the wavelength of the light is long enough, then regardless of the intensity of the beam or the applied voltage supplied by the battery no electrons are generated. Classical wave physics is unable to explain the conditions when such an electric current will appear or will not appear.

Einstein's explanation of the photoelectric effect is based on light consisting of individual photons, each with an energy given by (see Chapter 19)

$$E = hf = \frac{hc}{\lambda}, \tag{24.10}$$

where h is Planck's constant, $h = 6.63 \times 10^{-34}$ J-s, and we have used the fact that $c = f\lambda$. Photons also carry a momentum, according to Equation (24.9), given by

$$p = \frac{E}{c} = \frac{h}{\lambda}. \tag{24.11}$$

Equations (24.10) and (24.11) relate the photon energy and momentum, particlelike properties, to the wavelike properties of wavelength or frequency.

If the wavelength of the light is longer than some threshold value, then the energy of each photon will be too low to provide the minimum energy necessary to eject an electron from the metal surface, an energy known as the *work function* Φ. In this case, no electrons will be ejected.[1] When the photon energy exceeds the work function, a single photon can interact with an atom in the metal surface and eject a single electron. Those electrons that do escape from the "photocathode" surface can be attracted to the anode, by applying a positive potential difference between the electrodes, and make up the detected current. The amount of current is then proportional to the number of photons per second in the beam, this being proportional to the intensity of the beam. Beam intensity is defined as the energy per unit time per cross-sectional area and for a monochromatic beam is determined by the product of the energy of each photon and the number of such photons per second per cross-sectional area.

Now, depending on the wavelength of the incident light, emitted electrons will have more or less kinetic energy. In order to measure the kinetic energy of the emitted electrons, the polarity of the applied voltage can be reversed so that the electrons will be repelled by the anode. When the most energetic electrons are just stopped by this reversed voltage, known as the *stopping potential*, we know that

$$\mathrm{KE}_{\mathrm{max}} = eV_{\mathrm{stop}}, \tag{24.12}$$

and such a measurement can determine the maximum kinetic energy of the electrons emitted in the photoelectric effect. Einstein predicted that this maximum kinetic energy would be given by

$$\mathrm{KE}_{\mathrm{max}} = hf - \Phi, \tag{24.13}$$

so that the excess photon energy above the minimum energy needed to escape from the surface, the work function, equals the maximum kinetic energy. Electrons requiring more energy to escape from the surface will be left with less kinetic energy. Because kinetic energy must be positive, this relation implies that there is a minimum

[1]Strictly speaking, we now know this to be untrue: if the light source is a high-power laser, then there can be such an enormous number of photons that there is a nonnegligible probability that a single electron can absorb two or more subthreshold energy photons simultaneously and gain sufficient energy to escape. This is similar to the basis of multiphoton microscopy discussed at the end of Section 1 of the previous chapter.

frequency of light that is needed for electrons to just escape from the metal surface with essentially no kinetic energy given by

$$f_{\text{min}} = \frac{\Phi}{h}. \tag{24.14}$$

Equation (24.13) also correctly predicts that the maximum kinetic energy of the electrons depends only on the frequency and is independent of the intensity of the light.

Example 24.2 Suppose that red light of $\lambda = 633$ nm or blue light of 488 nm is directed on a photocathode with a work function of 2.25 eV. If 10^{12} photons per second of each color are separately incident on the photocathode, what will be the detected photocurrent in each case assuming 100% efficiency and all the emitted photoelectrons are captured by the anode? If the intensity of each light beam is increased by a factor of 10, what will happen? What is the stopping potential in each case?

Solution: The energy of the red and blue photons are given by hc/λ and are equal to (after converting to eV) 2.0 and 2.5 eV, respectively. Therefore, given the work function of 2.25 eV, red photons have insufficient energy to eject electrons whereas blue photons will each lead to an electron being detected at the anode (given the assumed 100% efficiencies) leading to a photocurrent corresponding to 10^{12} electrons per second or a current of $(10^{12}\ \text{e/s})\ (1.6 \times 10^{-19}\ \text{C/e}) = 1.6 \times 10^{-7}\ \text{A} = 0.16\ \mu\text{A}$. If the intensities are increased by a factor of 10 there will still be no emitted electrons with the red beam because the individual photon energy has not changed, and the photocurrent detected using the blue beam will increase by a factor of 10 to 1.6 μA. The stopping potential for the red beam experiment is zero because no electrons are detected at all whereas for the blue beam experiment, because the electrons are emitted with a maximum kinetic energy of $2.5 - 2.25 = 0.25$ eV, the stopping potential will be 0.25 V. Note carefully the units here.

A second experiment that demonstrates the particlelike nature of photons is the scattering of x-rays, high-energy photons, by the electrons of a material. In the early 1920s Arthur Compton discovered that the wavelength of x-rays gets slightly longer after scattering from a graphite target. He discovered that the process, now known as *Compton scattering*, could be completely explained by assuming that the x-rays carried energy and momentum given by Equations (24.10) and (24.11) and that the scattering simply conserved kinetic energy and momentum. Such an elastic collision is analyzed in a straightforward way using energy and momentum conservation in two dimensions just as it would be for billiard balls on a frictionless table. The resulting shift to longer x-ray wavelengths is due to the electron, initially at rest, gaining some momentum and kinetic energy at the expense of the photon (see Figure 24.4).

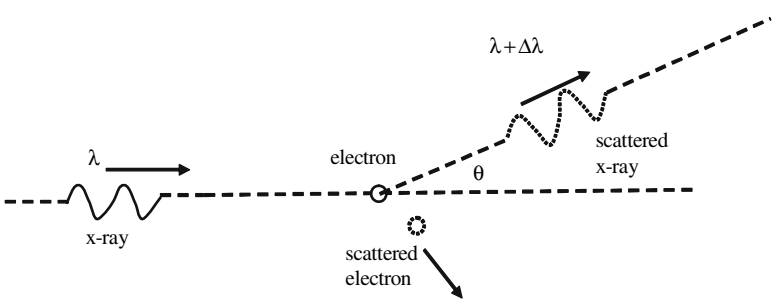


FIGURE 24.4 *Compton scattering of an x-ray photon by an electron. The scattered photon with longer wavelength and the recoil electron are shown dotted.*

A decreased photon momentum or energy has an associated increase in wavelength, known as the Compton wavelength shift, $\Delta\lambda$. Using energy and momentum conservation, Compton derived a formula for the wavelength shift

$$\Delta\lambda = \lambda_c (1 - \cos\theta), \tag{24.15}$$

where θ is the scattering angle and λ_c is the Compton wavelength of the electron, a fundamental constant given by

$$\lambda_c = \frac{h}{mc} = 2.43 \times 10^{-12}\,\text{m},$$

where m is the mass of an electron. Thus, the Compton shift vanishes for forward scattering, where the scattering angle is close to 0° indicating little interaction between the x-ray and electron, and is a maximum for backscattering when θ equals 180° and the x-ray has strongly interacted with the electron. We mention that both Compton scattering and the photoelectric effect are important in the making of a medical x-ray, the first in the x-ray/body interaction and the second in the detection process.

Having just studied two of the important experiments establishing the particlelike nature of photons under certain conditions, let's reconsider the double-slit interference experiment for light discussed earlier in Chapter 22 where we treated light as a wave. Imagine that we reduce the intensity of the light source so low that only one photon at a time arrives at the slits. Figure 24.5 shows the experiment. It is found that individual photons are detected at the screen at localized spots implying that the photon wave packet "collapses" when detected. However, after many such detections, the pattern of the total detected intensity is the same as that observed directly at higher light levels. In other words, even though individual detection events are localized on the screen, no photons ever arrive at positions on the screen that correspond to destructive interference bands whereas many more photons than the average arrive at the positions of constructive interference, according to the path difference equations of Chapter 22. If each individual photon went through one slit or the other, we would not expect to see an interference pattern because, with only one photon at a time, there would be no interference occurring. We must conclude that the *individual photons are going through both slits and interfering with themselves, with their own wave packet*. Given our (brief) discussion of wave packets and the notion of diffraction, it is not impossible to accept this notion. Individual wave packets, representing each photon, must travel through both slits, diffract at each, and recombine according to the rules of interference. When subsequently detected at the detector in the far-field, the wave packets must collapse and interact with the atoms of the detector as a "particle" getting detected at one particular location.

Amazingly, if the same experiment were to be done with electrons (but using different detection equipment), we would observe a similar result. The pattern of detected electrons on a screen far from the double-slits would be that produced by an interference pattern of waves using a wavelength for the electron given by the same expression as

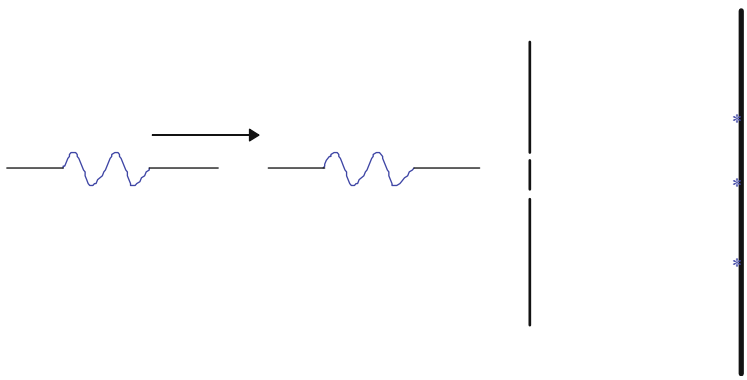


FIGURE 24.5 *The double-slit experiment at very low light levels so that individual photons are detected. A long experiment detecting many photons will build up a multiple exposure that is identical to that detected at higher light levels. The necessary conclusion is that individual photons interfere with themselves in passing through both slits.*

Equation (24.11), $\lambda = h/p$, known as the *de Broglie wavelength* of the electron. The electrons could be detected, for example, by having them strike a fluorescent screen emitting localized flashes of light. The slit separation would need to be made comparable to the de Broglie wavelength of the electron, but by adjustment of the electron's momentum this even can be matched to the same slit size as used for the photon experiment. At high electron beam intensities, an interference pattern would be directly observed on the screen. At very low electron beam intensity, with individual electrons arriving at the double-slit, the same interference pattern would be observed after an extended time exposure, again forcing us to conclude that each electron went through both slits simultaneously and interfered with itself. This seems at first sight to be inconceivable because the electron is known to be a fundamental "particle" that has no internal structure and is not divisible into subpieces. Despite our difficulties in accepting this, the electron does indeed behave as a wave, known as a matter wave. Although proposed much earlier and often used as a conceptual argument, this double-slit experiment with individual electrons was actually performed first in 1961 and has been verified in many ways since.

The first experiment to verify the wave nature of the electron was done by Davisson and Germer in 1927. By studying the diffraction of a beam of electrons from a crystal and observing ring patterns of maxima and minima, these experiments were able to verify the correctness of the de Broglie relation for the wavelength of the electron. Electrons, as well as photons, are said to exhibit wave–particle duality, sometimes behaving as a wave, as in situations showing diffraction and interference effects, and sometimes behaving as a particle, as in the detection process where particle mechanics concepts of momentum and energy "packets" apply.

Our conclusions for electrons also hold for all other elementary particles, each having its own de Broglie wavelength, depending on its momentum. Such wavelike effects of matter are not normally observed for macroscopic matter because the de Broglie wavelengths become extremely tiny. For example, a 1 kg mass traveling at 1 m/s has a de Broglie wavelength of about 10^{-33} m, much too small to produce any observable wave effects. But in the world of elementary particles, the masses are tiny, so that de Broglie wavelengths are large enough to produce dramatic effects. Even nonrelativistic electrons, accelerated through a potential difference of 1 V, have a momentum of $p = \sqrt{2mE} = 5.4 \times 10^{-25}$ kg m/s, and a corresponding de Broglie wavelength of 1.2 nm. This wavelength is large compared to atomic dimensions and such slow moving electrons can therefore be expected to exhibit diffraction and interference effects when interacting with a crystalline array of atoms, just as light does with an array of slits. Figure 24.6 shows an example of an electron diffraction pattern.

In addition to mass and electric charge, each electron carries another intrinsic property called *spin*. Just as mass creates gravity and charge creates the electric force,

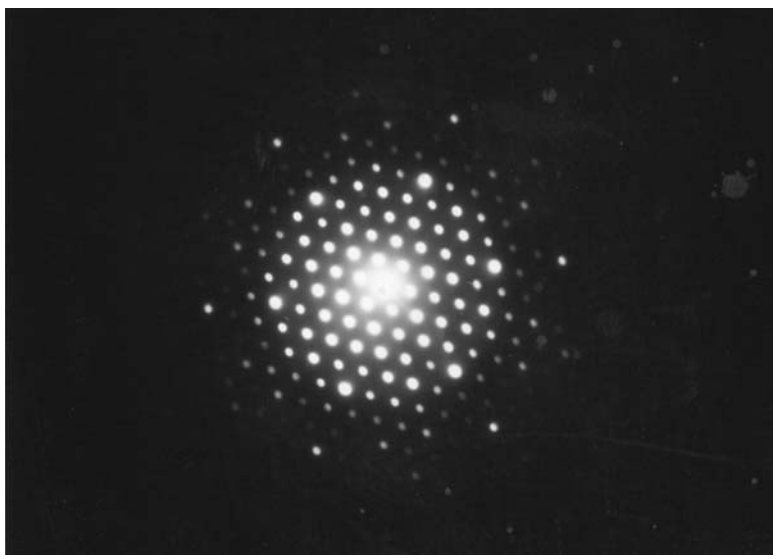


FIGURE 24.6 *Electron diffraction pattern from a thin germanium crystal.*

spin creates an interaction as well, another kind of repulsive force between electrons. Unlike gravity and the electric force, though, we can't write down a specific equation for this interaction. Instead, it is expressed as a rule: *no two electrons occupying the same region of space can be in exactly the same state of motion* (or have the same set of quantum numbers). This rule is called the *Pauli exclusion principle*, and, among other things, it is responsible for the great variety of chemical differences we observe among atoms. We study this further in the next chapter where we show in more detail that this is responsible for the different known types of atoms. Here we need to point out that this principle only applies to particles with half-integral spin.

Some macroscopic systems also exhibit quantum mechanical effects; particularly notable examples are superconductors and superfluids. In some materials at sufficiently low temperature, the conduction electrons pair up so that these "Cooper pairs" have integral spin and are no longer subject to the Pauli exclusion principle. They are all able to occupy the same low energy state and not interact with the material lattice around them. In this case their electrical resistance is, in fact, equal to zero. These materials are called superconductors and a variety of different types of materials have been discovered that become superconductors at sufficiently low temperatures. Superconducting wires are used in large electromagnets to produce very large magnetic fields without heating problems when their temperature is sufficiently low, typically at liquid helium temperatures of about 4 K. For example, these superconducting magnets are used in MRI facilities in hospitals. Such superconductors eliminate I^2R heating and once a current is established in these materials, it persists without the need for a continual energy supply such as a battery or power supply. A major goal of this area of research is to develop materials that are superconducting at ambient, or near ambient, temperatures and that can be fabricated into wires or other types of conductors to avoid the costs of maintaining those extremely low temperatures.

An analogous situation can occur in certain fluids when they are cooled to very low temperatures. For example, when ^{4}He, with paired protons, neutrons, and electrons, is cooled below 2.18 K, it becomes a superfluid with very unusual properties. Superfluids have no viscosity, so that a particle traveling through them moves with no friction. Such superfluids can also flow through microscopic pores and channels that would not be accessible to normal fluids because of surface tension. ^{3}He can also behave as a superfluid at about 1000 times colder temperatures, in a mechanism similar to superconductors, by forming "Cooper pairs" of ^{3}He which behave as integral spin particles, so that they are not subject to the Pauli exclusion principle. Superfluidity is very rare and has only been found in a handful of systems other than helium.

3. WAVE FUNCTIONS; THE SCHRÖDINGER EQUATION

We've seen that photons and other elementary particles such as the electron have both wavelike and particlelike properties that are related to each other. For example, treating light as made of photons, particles of zero rest mass, its energy E and momentum p are connected through the relation $E = pc$. But these quantities are connected with the wavelike properties of frequency and wavelength through Equations (24.10) and (24.11). Furthermore, viewing light as an electromagnetic wave, we've seen that the intensity, or energy per unit area per unit time, is proportional to the square of the electric field. How are these two pictures related to each other?

In our rediscovery of the double-slit experiment for single photons we just saw that the photon wave packet is a representation of the spatial extent of the photon. This implies that the square of the electric field must be a measure of where the photon is located (see below). Knowing that electrons and other elementary particles also exhibit both particlelike and wavelike behavior, scientists were prompted to look for a wave theory of matter. But in that case what is it that is waving; whose square is related to the electron's whereabouts?

Quantum mechanics, developed in the 1920s, introduces a *wave function* Ψ that is dependent on both time and position, and that represents all the possible information

obtainable about an elementary particle or system of particles under study. Note our mix of the words particle and wave function in the same description of the system. An electron, for example, is described completely by its wave function.

The square of the wave function for the electron, $\Psi^2(x, y, z, t)$, multiplied by the volume of a small region in space ΔV located at (x, y, z), represents the probability that the electron will be found within that volume at that position at the specified time

$$\Psi^2(x,y,z,t)\Delta V = \textit{Probability to find electron within } \Delta V \textit{ at } (x,y,z) \textit{ at time } t. \qquad (24.16)$$

According to this definition Ψ^2 represents a probability density, or probability per unit volume. Depending on the dimensionality of a particular problem, we might replace the volume with the surface area or simply the linear distance. For example, in our description of the double-slit experiment with an electron, with x the distance along the screen measured from the central axis, Ψ^2(x, t) Δx would represent the probability of finding an electron within a distance Δx at position x at time t. This probability will have the same spatial variation as the interference patterns with light discussed in the last chapter. Locations of complete destructive interference would have $\Psi^2 = 0$, and interference maxima would correspond to maxima in Ψ^2.

We can make a close analogy between Ψ for matter waves and the electric field E for photons. We know that for photons, the intensity I, proportional to E^2 and representing the photon energy per unit area (or photon flux) per unit time, is also proportional to the number of photons N. If the intensity and therefore number of photons is very small, as in the low-intensity double-slit experiment discussed in the last section, then we can interpret $E^2\Delta x$, evaluated at some point on the screen, as the probability that a photon will be detected within Δx at that point on the screen. Similarly for an electron, for example, $\Psi^2\Delta V$, evaluated at a point represents the probability of finding an electron within the small volume ΔV at that point.

Because we can interpret $\Psi^2\,\Delta V$ as the probability of finding the electron within ΔV, and it is also clear that the electron must be found somewhere within the confines of the system boundary (with certainty, or with a probability of 1), we must have that

$$\sum \Psi^2 \Delta V = 1, \qquad (24.17)$$

where the summation is over all the volume available in the system. This is known as the normalization condition and establishes the scale for quantifying Ψ.

Quantum mechanics provides an equation, the *Schrödinger equation*, which plays the same role as Maxwell's equations play in electromagnetism (see the box below). Schrödinger's equation allows one to compute the space- and time-dependence of the wave function for any quantum system. Only wave functions for simple systems can be analytically determined; those for complicated systems of many bodies must be approximated and calculated using computers.

To give a sense of the nature of wave functions, let's consider the problem of a particle trapped in a box. We consider a one-dimensional problem, with a particle bouncing back and forth between end walls, only experiencing a force at the walls where we imagine the potential energy to rise infinitely steeply as shown in Figure 24.7. The

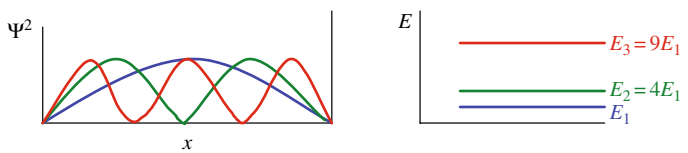


FIGURE 24.7 *A quantum mechanical particle in a one-dimensional box. The first three wave functions are shown, each having a discrete energy shown on the right* (color coded)*; the longer wavelengths correspond to lower energy states as discussed in the text.*

fundamental concept invoked here is that the matter wave Ψ must be a standing wave within the box. Only a standing wave results in nonzero amplitudes and we show that a standing wave also leads to a discrete set of possible energy levels for the particle. A matter wave with energy different from one of those discrete energy levels would, through interference, completely cancel itself on multiple reflection within the box. This is precisely the same idea as was discussed in connection with standing waves on a string or in an air column back in Chapters 10 and 11. We also discuss this further in the next section in connection with the uncertainty principle.

The standing wave expressions for $\Psi(x, t)$ in our one-dimensional box of length L are found by applying the boundary conditions that for all time there are nodes at the ends, $\Psi(0, t) = \Psi(L, t) = 0$. We find that the possible standing wave functions are

$$\Psi_n(x,t) = A_n \sin(n\pi x/L), \tag{24.18}$$

independent of time, where A_n are the amplitudes of the nth harmonic of the wave (see Equation (10.18)) and are chosen according to the normalization requirement of Equation (24.17). Equation (24.18) satisfies the boundary conditions (please check this!) and gives a set of standing waves with wavelengths corresponding to $\lambda_n = 2L/n$, as can be seen by rewriting the argument of the sine function as,

$$\left(\frac{2\pi x}{(2L/n)}\right) = \frac{2\pi x}{\lambda_n}.$$

Using the relation between the de Broglie wavelength and the momentum $p = h/\lambda$, we see that the momentum of the particle is quantized and must satisfy $p_n = nh/2L$, so that the energy of the (nonrelativistic) particle must also be quantized ($E_n = p_n^2/2m$) and given by

$$E_n = n^2 \frac{h^2}{8mL^2}. \tag{24.19}$$

Figure 24.7 shows the first few wave functions and the corresponding energy level diagram for the particle in a one-dimensional box.

The $n = 1$ state is the ground state for this system and has an energy given by

$$E_1 = \frac{h^2}{8mL^2}. \tag{24.20}$$

It is noteworthy that the particle cannot have zero kinetic energy according to our results, but must have at least a minimum energy given by Equation (24.20), known as the *zero-point energy* because the particle will have this same energy even at a temperature of absolute zero.

The particle in a box problem, although perhaps not very realistic, does illustrate some of the basic ideas of quantum mechanics. Other types of potentials, shown in Figure 24.8, can be analyzed in a similar manner to give results applicable to more realistic problems. For example, the "finite square well" problem (curve a in the figure) or better the "Coulomb potential barrier" (curve b) can be used to model particles within the nucleus as we show in the next chapter. In this case, because the wall is not infinitely high, we show that although it is impossible for particles with a small energy to escape from the "well" classically, quantum mechanics predicts some possibility to penetrate the wall and escape. This phenomenon can be used to model radioactive decay of nuclei. Similarly, a particle that meets a "finite barrier potential" (curve c in the figure) with an energy smaller than the barrier should be totally reflected classically, but quantum mechanics predicts

A simplified form of Schrödinger's equation, valid for a particle of mass m and energy E moving along the x-axis in a potential energy PE(x), is a time-independent differential equation for the wave function $\Psi(x)$:

$$\frac{-h^2}{8\pi^2 m}\left(\frac{d^2\Psi(x)}{dx^2}\right) + \text{PE}(x)\Psi(x) = E\Psi(x).$$

The potential energy function represents the total of all the interactions that the particle experiences, with typical model forms for PE being square wells, barriers, Coulomb potentials, harmonic oscillator potentials, or more realistic functions representing molecular potentials (see Figure 24.8).

In a straightforward fashion this equation can be generalized to three-dimensional space and applied directly, for example, to solve for the wave function of the electron in the hydrogen atom using the Coulomb potential due to the proton. The wave functions obtained give the probability density for the electron and give the mean radius of the ground state of the hydrogen atom. In solving the Schrödinger equation, with each wave function there is a corresponding energy of the electron, so that the values of E are discrete and form an energy-level diagram that we have alluded to several times in this text.

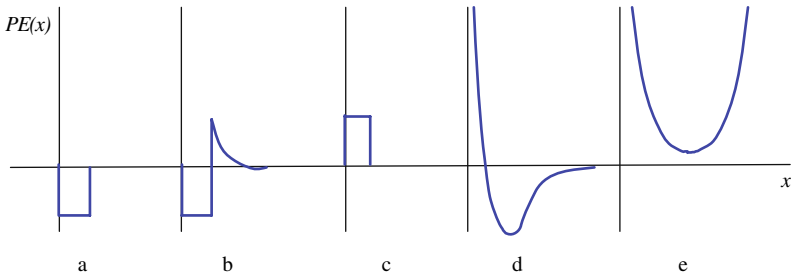


FIGURE 24.8 *Various potential functions commonly used to represent physical situations: (a) finite square well; (b) Coulomb well; (c) finite barrier; (d) molecular potential (Lennard–Jones type); (e) simple harmonic oscillator.*

that there will be some probability that the particle can "tunnel" through the barrier and reach the other side. This phenomenon is important in our discussion of the scanning tunneling microscope in the next section. Finally, various potential energy curves (e.g., curve d) can be used to model the interactions of valence electrons in atoms or molecules, as discussed in the next chapter.

4. UNCERTAINTY PRINCIPLE; SCANNING TUNNELING MICROSCOPE

We have seen that quantum mechanical particles exhibit wave–particle duality, appearing sometimes to have exclusively wavelike and sometimes exclusively particlelike properties. Niels Bohr referred to this as the *principle of complementarity*. Quantum mechanics takes the view that in order to have definite knowledge of a certain parameter describing a particle, such as its position, momentum, or energy, a measurement must be performed. In practice, every such measurement will have an associated uncertainty due to, at the very least, the precision of the measuring instruments and the skill of the measurer. For example, a measurement of a particle's position or velocity may be limited by the precision of the meter stick or of the clock used. No matter how sophisticated the measurement, there will always be limitations on the precision of the measurement.

In the world of elementary quantum mechanical particles there are fundamental intrinsic limitations on the accuracy of measurements due to the interaction of the measuring instrument with the particle. Unlike the usual experimental limits on precision of a measurement, these more fundamental limitations do not depend on the precision of measurement instruments or on the skill of the measurer. If we try to determine both the position and momentum of, say, an electron, then no matter how "gentle" a measurement we make, there is always an uncertainty in precisely how the interaction occurs that is intrinsic in nature. For example, suppose we try to "see" the position of an electron by scattering a photon from it. We know that the photon has a wavelength that will fundamentally limit the resolution with which we can "see" due to diffraction effects. In the scattering process, the photon will also impart some of its energy to the electron. To better locate the position of the electron we might decrease the wavelength of the photon so that during the scattering event we may "see" with greater resolution. In so improving the precision of the electron position measurement, however, the photon's energy and momentum increase and the electron will receive an uncertain fraction of the photon's larger energy leading to a greater uncertainty in the electron's momentum. This is a fundamental problem, not one that can be eliminated by more careful measurement apparatus or skill.

Let's sketch a semiquantitative analysis of the scattering event. The resolution uncertainty is comparable to the wavelength of the photon, so that

$$\Delta x \approx \lambda. \tag{24.21}$$

Because the photon's momentum is given by $p = h/\lambda$, and some indeterminate fraction is imparted to the electron, we also have that

$$\Delta p \approx \frac{h}{\lambda}. \tag{24.22}$$

The product of uncertainty in position, x, and the uncertainty in momentum along the x-direction, p, leads to the Heisenberg uncertainty principle

$$\Delta x \, \Delta p \approx h, \tag{24.23}$$

where it is understood that this expression gives the minimum uncertainty product possible.

The uncertainty principle (see Figure 24.9) tells us that if we know the exact position of a particle, so that Δx is zero, then we can have no knowledge at all of the particle's momentum ($\Delta p \sim \infty$). According to our experience this principle makes no sense at first sight. We can measure the position of, say, a marble with very high precision while it sits quite at rest on a table, so that $p = 0$ very precisely. To see why there is no conflict of this example with the uncertainty principle, we need to examine some numbers. Because h is so very small, 6.6×10^{-34} J-s, the uncertainties that are implied are extremely small for macroscopic objects. If our marble has a mass of 10 g, then dividing h/m, the uncertainty principle leads to the product $\Delta x \, \Delta v \geq 6 \times 10^{-34}\,\text{m}^2/\text{s}$, We can only measure the marble's location to, at very best, the dimension of an atom, 0.5×10^{-10} m, so that the uncertainty in speed of the marble must be at least 10^{-22} m/s. But a velocity of this magnitude corresponds to the marble moving one atomic radius in over 15,000 years! So the uncertainty principle presents no conflict with macroscopic measurements. On the other hand, because of its small mass, to know the position of an electron to within the size of an atom implies an uncertainty in its velocity of over 10^7 m/s!

Position and momentum are said to be conjugate variables since there is an uncertainty relation of the form of Equation (24.23) that links them together. Another important pair of conjugate variables is energy and time, with a similar minimum uncertainty relation

$$\Delta E \, \Delta t \approx h. \tag{24.24}$$

This uncertainty relation has a number of significant consequences. For example, atoms in an excited state have a characteristic lifetime, the average time before emitting a photon and returning to their ground state. This is a statistical process meaning that in a large collection of such excited atoms, the average decay time (the lifetime) is a characteristic of that particular transition, but that for any particular atom undergoing this transition we cannot know the exact transition time. Because of this uncertainty in time, there is a corresponding uncertainty in the energy of the atomic transition, given by Equation (24.24) and hence in the energy of the emitted photon. We can think of this energy uncertainty as arising from a small characteristic energy width of the excited state itself. Narrower, more sharply defined, energy levels have longer lifetimes, whereas broader energy levels have shorter lifetimes.

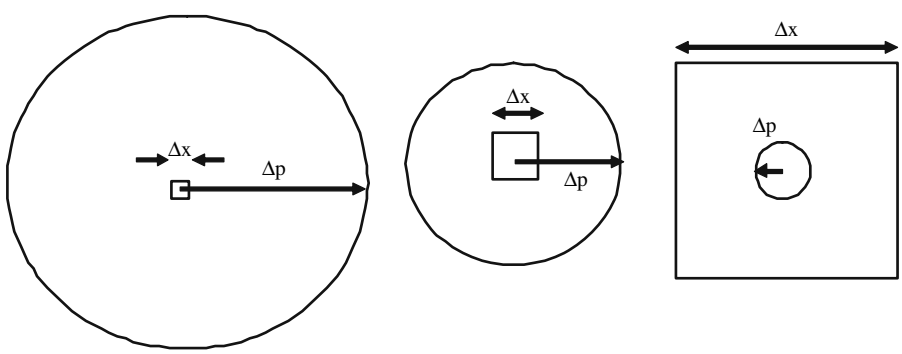


FIGURE 24.9 *The uncertainty in the x-location of a particle is inversely related to the uncertainty in its momentum along the x-direction. The product of these two uncertainties must be at least the order of $h = 6.63 \times 10^{-34}$ J-s.*

Example 24.3 Find the spread in frequencies, or linewidth Δf, when atoms radiate from an excited state with a lifetime of 2×10^{-9} s. Also find the fractional spread in frequencies, $\Delta f/f$, if the emitted photons have a wavelength of 550 nm.

Solution: The lifetime of the transition leads to a spread in energy of the emitted photons. From Equation (24.24) and the fact that, from $E = hf$ we know that $\Delta E = h\Delta f$, we can write that $\Delta f = \Delta E/h \approx (h/\Delta t)/h = 1/\Delta t$. Then because the transition time is a statistical average, its uncertainty is comparable to its value and we have that $\Delta f \approx 1/\Delta t \approx 1/(2 \times 10^{-9}\ \text{s}) = 5 \times 10^{8}$ Hz. The photon frequency is given by $f = c/\lambda = 5.5 \times 10^{14}$ Hz, so the fractional spread in frequencies is then $5 \times 10^{8}/5.5 \times 10^{14} = 9 \times 10^{-7}$. This so-called "intrinsic" linewidth is usually masked by larger spreads in frequency due to thermal motions of the atoms producing random Doppler shifts in frequency.

Another consequence of this uncertainty relation is the possibility of multiphoton spectroscopy, as discussed briefly in Section 1 of Chapter 23 in connection with microscopy. To excite an atom from its ground state to an excited state requires a specific energy photon hf, corresponding to the transition energy. If the photon density is large enough so that the probability for the absorption of two or more photons within a short time Δt is large, then the uncertainty relation allows, for example, N photons, each of energy hf/N, to cause the overall transition even though there are no intermediate energy levels so that no transition to such intermediate energies is possible (Figure 24.10). In other words, as long as the photon absorption occurs within a very short time window, the energy uncertainty that follows from the uncertainty relation is sufficient to allow this process to occur.

Tunneling, mentioned in the last section, is another type of purely quantum mechanical phenomenon that arises from the uncertainty relation. Imagine an electron confined within a one-dimensional box by potential walls, or barriers, such as the one shown in curve c of Figure 24.8, on either side of the box. Classically, if the electron had an energy less than that of the barrier height, it would forever be trapped within the box bouncing back and forth. Quantum mechanics agrees with this as well if the potential barriers are infinitely high and leads to the standing waves studied in the previous section. If the barriers are finite, however, then there is a small probability that the electron can escape or "tunnel" through the barrier wall. Tunneling can be related to the energy–time uncertainty relation. If the time for the electron to pass through the wall is short enough, then the uncertainty in the electron's energy during that time interval may become large enough to allow its energy to exceed the barrier energy. Therefore during that brief time the electron does not violate conservation of energy and the laws of physics will not prevent the electron escaping from the box. The probability that the electron tunnels out of the box is small and depends on the barrier potential height and wall thickness. As bizarre as this appears, it is a real phenomenon and can be used in actual pieces of equipment to study materials on an atomic scale.

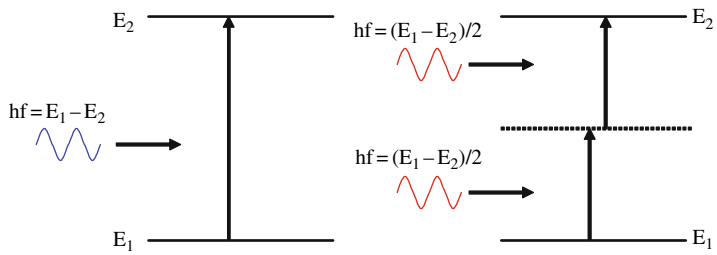


FIGURE 24.10 (left) *Absorption of a single photon causing a transition to an excited state.* (right) *Absorption of two photons each with half the energy needed can occur even though there is no intermediate energy level, as long as the lifetime of the* (virtual) *intermediate state is shorter than the minimum uncertainty dictated by the Heisenberg uncertainty principle.*

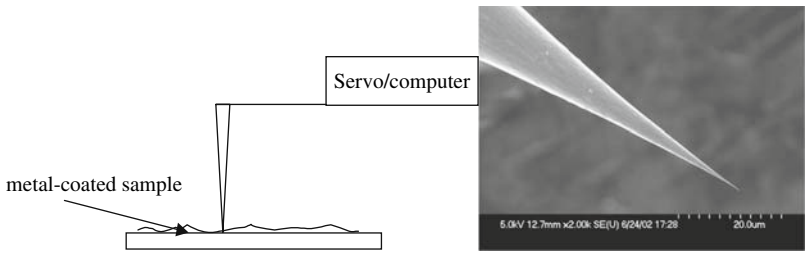


FIGURE 24.11 *Scanning tunneling microscope schematic and EM image of a needle tip used for scanning.*

The *scanning tunneling microscope* uses this phenomenon to image the surface of a microscopic object with unprecedented resolution. A sample is coated with a thin layer of metal to make it electrically conducting. A fine-tipped needle is then placed close to, but not in contact with, the surface and a small potential difference is applied between the needle and the sample surface (Figure 24.11). If the tip-to-sample distance is on the order of 1 nm, then a small electric current can be detected from electrons that have tunneled across the air or vacuum insulating layer. As the needle moves along just above the surface, the gap distance changes and the tunneling current changes as well. Because the tunneling current is so sensitive to the gap (corresponding to the barrier wall thickness), extremely high resolution images of surface sample features is possible. Vertical resolution of better than 10^{-2} nm and lateral resolution about an order of magnitude less is possible, easily allowing individual small atoms at the surface to be visualized. Although, in principle the needles used should have tips with atomic dimensions, it turns out to be fairly straightforward to fabricate such needles because surfaces tend to be fairly rough on atomic dimensions anyway. One commonly used mode of operation has a feedback loop circuit to vary the height of the probe as it is scanned across the sample in order to maintain a constant height above the surface and thereby a constant sample-to-probe current. By scanning the sample, a record of the surface topography is recorded, allowing extremely high resolution of surface features (Figure 24.12).

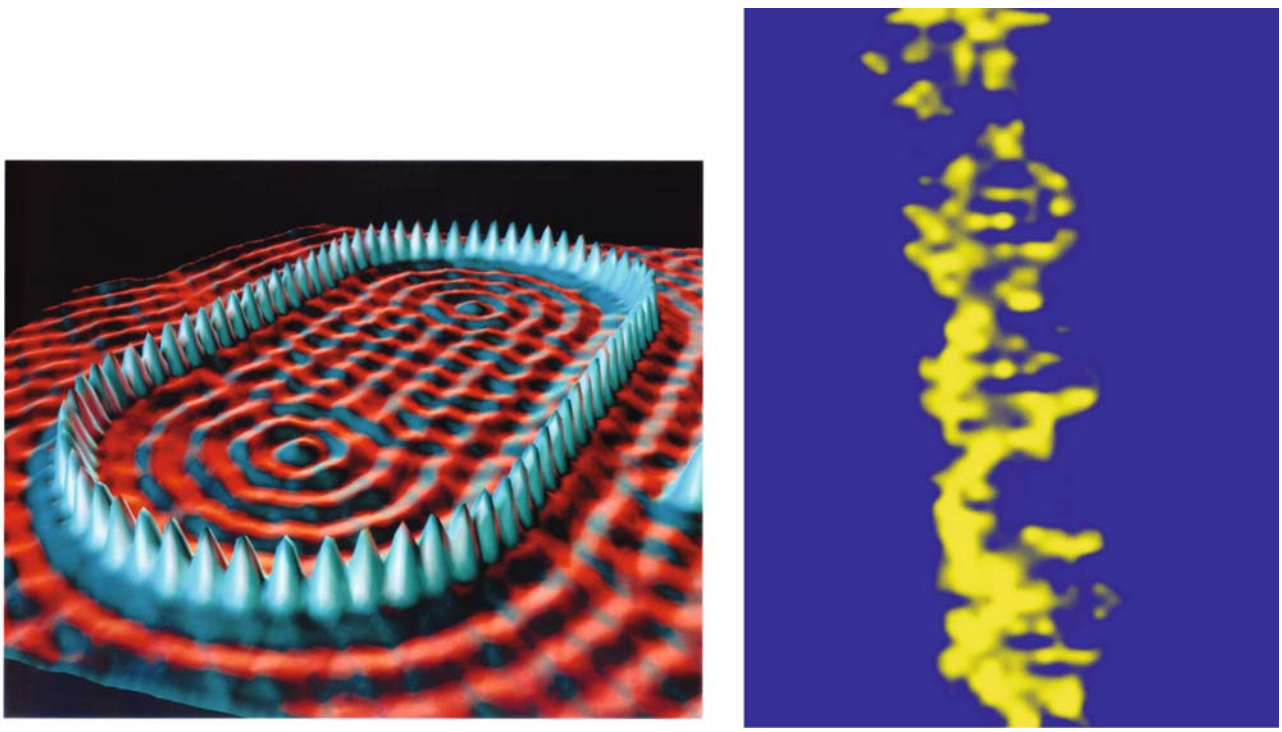


FIGURE 24.12 *False color scanning tunneling microscope images.* (left) *Iron atom array, produced by manipulating the atoms by the STM needle, on a copper atom support film. The wavelike appearance in the background is due to electron matter standing waves that are trapped within the iron atom "corral";* (right) *native DNA image in which the stacked bases are just visible.*

4.1. QUANTUM MECHANICS AND ENERGY LEVELS

In the previous section we saw that a particle trapped in a one-dimensional box has a nonzero minimum energy, the zero-point energy, given by Equation (24.20). This agrees with the uncertainty principle, which also requires that a mass (m) confined to move in a finite region of space (of extent L) must have a smallest speed whose magnitude is approximately given by $v_{QM} \sim h/(mL)$. (Here we use the nonrelativistic expression for $p = mv$.) A confined mass, therefore, must have a minimum kinetic energy, $KE_{QM} = (1/2)m(v_{QM})^2 \sim h^2/(mL^2)$ (remember, "~" means "order of magnitude;" "1/2" has the same order of magnitude as "1"), in qualitative agreement with Equation (24.20). This minimum, irreducible kinetic energy is also called the mass's *ground state kinetic energy*. Let's examine this in some further detail.

Suppose we drop a 1 kg mass 10 cm, calculating that it acquires a KE of $mgh \sim 1$ J. If we substitute into our KE_{QM} expression $m = 1$ kg and $L = 10$ cm (0.1 m), we get a ground state kinetic energy of about 10^{-66} J, using $h \sim 10^{-34}$. Obviously, the 1 J value quoted above for a 1 kg mass falling 10 cm has nothing to do with the ground state motion of the 1 kg mass, a point we return to below. The 1 kg mass consists of about 10^{25} atoms. Each of these is confined to the same 10 cm as the whole body. Thus for one atom with $m \sim 10^{-25}$ kg and $L = 0.1$ m, we have $KE_{QM} \sim 10^{-41}$ J. If we multiply the latter kinetic energy per atom by 10^{25} atoms we might expect to get the kinetic energy of the whole 1 kg body. What we do get is 10^{-16} J. Although neither the 10^{-66} nor the 10^{-16} values are macroscopically measurable, and, therefore, are not of much macroscopic consequence, they differ by a factor of 10^{50}! It would be nice to know which is right.

Resolution of this discrepancy revolves around the notion of *coherent* versus *incoherent* motion as discussed in Chapter 12 (see Figure 12.7). When we use 1 kg for the mass in the calculation we are tacitly assuming that all 10^{25} atoms in the body move together in lock-step fashion, as a coherently synchronized swarm. When we use 10^{-25} kg for the mass we are tacitly assuming that each atom moves independently of the rest. Such unsynchronized motion is incoherent motion. In a solid, where all the atoms are glued together by interatomic forces, the former seems like a reasonable assumption.

But wait! Each atom in a solid is surrounded by neighboring atoms that also confine its motion. Thus each atom shares the macroscopic confinement of the whole body, whereas at the same time each has a microscopic confinement. For an atom in a solid confined by its neighbors, m is about 10^{-25} kg and L is about 10^{-11} m (about 10% of the atom's size). Thus, the ground state speed of the atom (v_{QM}) due to this confinement is on the order of 10^2 m/s and its ground state kinetic energy (KE_{QM}) is about 10^{-21} J (you should verify these values using the equations at the beginning of this discussion above). Clearly, this motion has to be incoherent, because if all of the atoms were moving lock-step together the solid would be careening around at over 100 m/s! Because the motion is incoherent, we can add up the kinetic energies and conclude that the 1 kg solid sitting at rest has about $(10^{-21}$ J/atom$)(10^{25}$ atoms$) = 10^4$ J of ground state kinetic energy due to incoherent, microscopic atom motions (far more than the 1 J you get by dropping all of the atoms coherently a distance of 10 cm).

As each atom jiggles incoherently, it carries its electrons and its nucleus with it. But the electrons are confined by their interaction with the nucleus so they have additional motion internal to the atom's. Use the values $m \sim 10^{-30}$ kg and $L \sim 10^{-10}$ m to find that for each electron $v_{QM} \sim 10^6$ m/s and $KE_{QM} \sim 10^{-18}$ J. As there is an order of 10 e/atom in a typical solid, the incoherent motion of all electrons yields a ground state kinetic energy of about 10^8 J in a 1 kg mass. In addition, the nucleons in each nucleus are confined by their strong nuclear interaction with each other. For them, you should find $m \sim 10^{-27}$ kg and $L \sim 10^{-15}$–10^{-14} m, leading to $v_{QM} \sim 10^7$ to 10^8 m/s (a fair fraction of the speed of light) and $KE_{QM} \sim 10^{-11}$–10^{-13} J for each nucleon. Adding all of this kinetic energy up yields more than 10^{12} J in a 1 kg mass. In other words, in each macroscopic body there is a phenomenally large amount of ground state kinetic energy associated with microscopic, incoherent motion, with the overwhelming majority being associated with motion inside the atomic nuclei.

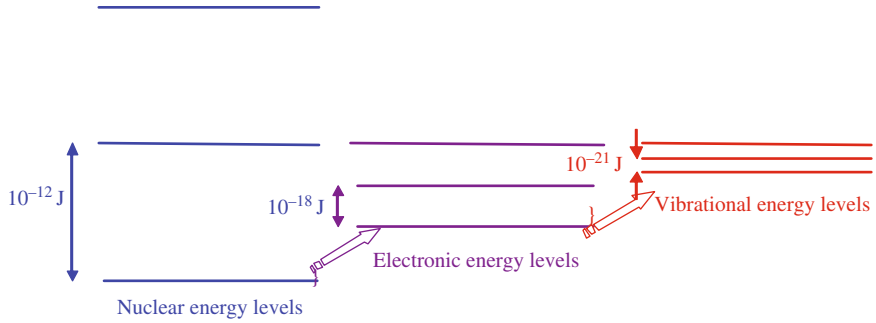


FIGURE 24.13 *Typical allowed quantum states for matter.*

The states of motion allowed by quantum mechanics tend to have different kinetic energies. The energy differences between these allowed states tend to be about the same size as the ground state kinetic energy. Thus the allowed states of nucleon motion tend to differ in energy by about 10^{-12} J. Electronic states tend to differ in energy by about 10^{-18} J, and atomic vibrational states tend to differ in energy by about 10^{-21} J (see Figure 24.13). We return to a discussion of energy levels and their study by spectroscopy in the next chapter.

As we have seen in Chapter 12, the average kinetic energy of an atom or molecule is proportional to the absolute temperature, so that $T \sim \mathrm{KE}_{\mathrm{internal}} \times (10^{23}$ K/J), where $\mathrm{KE}_{\mathrm{internal}}$ is an internal kinetic energy and T is measured in kelvins, K. For atomic vibrations, $\mathrm{KE}_{\mathrm{internal}}$ is about 10^{-21} J, so T for atomic vibrations is of order 10^2 K (e.g., room temperature). For electronic motion in an atom, $\mathrm{KE}_{\mathrm{internal}}$ is about 10^{-18} J, so T for electrons is about 10^5 K. For nucleonic motion in nuclei, $\mathrm{KE}_{\mathrm{internal}}$ is about 10^{-12} J, so T for nucleons is about 10^{11} K. Reciprocally, we can say that if a body has a temperature of a few 100 K it is possible to excite internal atomic vibrations, but not electronic states and, emphatically, not nucleonic states. To excite these requires very high temperatures, indeed. In a body at room temperature, all of the excess energy above the ground state is in atomic vibrations. At room temperature, the body's electrons and nucleons are "frozen" into their respective ground states.

Therefore, here's one of the remarkable secrets of life. In a living cell (whose temperature is roughly 300 K), there's a huge amount of nucleonic internal energy, a much less, but nonetheless significant, amount of electronic internal energy, and, by comparison, an almost negligible amount of atomic vibrational internal energy. Even so, atomic vibrations are the only energy source available for the cell to use, because the other motions are stuck in their ground states. By carefully marshalling and partitioning its puny supply of internal energy, a cell manages to perform all the various tasks of life, including protein replication, locomotion, and cell division.

CHAPTER SUMMARY

The theory of special relativity is based on two fundamental postulates: all the laws of physics are the same in all inertial frames of reference and the speed of light in vacuum has the same value c in all inertial reference frames. This latter postulate seems contrary to our (low-speed) intuition and leads to a large variety of seemingly bizarre, but experimentally confirmed, effects having to do with time and space. Here we focus on the dynamical quantities that we need in the next chapters.

Momentum of a particle of mass m moving at a velocity v is given by

$$p = \frac{mv}{\sqrt{1 - v^2/c^2}} = \gamma mv, \qquad (24.1)$$

where

$$\gamma = \frac{1}{\sqrt{1 - v^2/c^2}}.$$

Similarly, the particle's relativistic energy is given by

$$E = \frac{mc^2}{\sqrt{1 - v^2/c^2}} = \gamma mc^2, \qquad (24.5)$$

which can also be written as the sum of the kinetic energy KE and the rest energy, as

$$E = \text{KE} + mc^2. \qquad (24.6)$$

Small changes in the (rest) mass m, lead to large changes in energy given by

$$\Delta E = \Delta mc^2, \qquad (24.7)$$

and this effect has led, for example, to both atomic bombs and nuclear power plants. Energy and momentum are connected through the equation

$$E^2 = p^2c^2 + m^2c^4. \qquad (24.8)$$

For a massless particle, such as the photon, or in the limit that the velocity approaches c (so that γ becomes very large) this last equation reduces to

$$E = pc. \quad (\text{if } m = 0 \text{ or } \gamma >> 1). \qquad (24.9)$$

Historically, several important experiments revealed the "particlelike" nature of photons and initiated the notion of "wave–particle duality," the idea that all elementary particles exhibit both wave- and particlelike properties depending on their interactions. The photoelectric effect is the production of an electric current proportional to the incident light intensity. But, each incident photon of frequency f needs a minimum threshold energy, the work function Φ, in order to liberate an electron, and because $E = hf$ for a photon, with h = Planck's constant = 6.63×10^{-34} J-s, there will also be a minimum frequency needed. Einstein worked out the explanation for this effect and found that the liberated electrons have a maximum KE given by

$$\text{KE}_{\text{max}} = hf - \Phi, \qquad (24.13)$$

obtained simply from conservation of energy in the individual photon–electron interaction. The Compton effect also treats high-energy x-ray photons as particles colliding with electrons to successfully analyze the scattering results. There is also a further discussion in the chapter of the double-slit interference experiment, but now for single photons, or for electrons, which have a deBroglie wavelength given by

$$\lambda = \frac{h}{p}.$$

This association of a wavelength with a "particlelike" momentum bridges the wave–particle duality notion. Electrons can be seen to exhibit wavelike properties in the phenomenon of electron diffraction, for example.

Quantum mechanics, the theory of the microscopic world, has a central dogma that all the possible information knowable about a system, for example, an electron, can be described by the wave function Ψ, whose square is given by

$$\Psi^2(x, y, z, t)\Delta V = \textit{Probability to find electron within } \Delta V \textit{ at } (x,y,z) \textit{ at time } t. \qquad (24.16)$$

The wave function can be found by solving the Schrödinger equation, which leads to a set of quantum numbers that define the possible energy levels, angular momentum, spin, and so on of the system. A fundamental rule is the Pauli exclusion principle, which states that no two interacting fundamental particles (e.g., electrons) can have the same set of quantum numbers.

Another fundamental principle is the Heisenberg uncertainty principle, which describes a basic limitation in nature on the simultaneous measurement of pairs of conjugate variables

$$\Delta x\, \Delta p \approx h, \qquad (24.23)$$

$$\Delta E\, \Delta t \approx h. \qquad (24.24)$$

These limitations have negligible effect in the macroscopic world, but can produce major effects in the microscopic arena. One notable result is the phenomenon of tunneling and an associated microscopy technique known as scanning tunneling microscopy, which gives atomic resolution images.

A quantum mechanical analysis of the ground state energy contained in matter shows that internal KE from incoherent electron and nucleon motions is huge, much larger than typical translational energies of matter. The different internal energies have corresponding energy levels that can be explored with spectroscopy.

QUESTIONS

1. Relativity requires any particle that travels at the speed of light, such as the photon, to have no rest mass. Why is this necessary?
2. In a photoelectric effect experiment, if a beam of green light produces a photocurrent, will a beam of blue light with the same intensity produce a larger, smaller, or the same photocurrent?
3. If one shade of yellow photons will produce a photocurrent, but another shade does not, will red light produce any photocurrent? Will green light?
4. If when the anode voltage is set to -1.5 V there is just no photocurrent with a particular green wavelength of light, when the wavelength is changed to a blue and the intensity of the blue light is 1/2 that of the green, what happens to the photocurrent? To the maximum kinetic energy of the photoelectrons?
5. Note that the Compton wavelength shift is independent of the actual wavelength of the x-ray photon. How does the percent change in the wavelength of a Compton scattered x-ray photon depend on the wavelength of the photon?
6. How does the Compton wavelength shift for x-ray scattering from protons compare to that from electrons?
7. Discuss how the interference pattern observed in a double-slit experiment with electrons depends on the energy of the electrons.
8. For a particle in a one-dimensional box of length L, where is the particle most likely to be found when in the ground state? In the first excited state?
9. Why is it impossible for an object to be exactly at rest? Discuss this in connection with a car at a stoplight and with an atom in an "atom trap". What is the approximate uncertainty in velocity in each of these cases?
10. Classical physics only allows a photon to be absorbed by a sample if it has an energy equal to the energy difference between the final state and the initial state. If there are no intermediate energy levels between these then no photons with less energy can be absorbed. On the other hand, experimentally it is found that if three photons from a high-intensity laser, each having an energy equal to 1/3 that of that energy difference are absorbed, the sample can reach the final state. Discuss the energy–time uncertainty relation's impact on allowing this process to occur.
11. How do you expect the electron tunneling current to depend on the barrier height? On the barrier thickness? On the electron energy?
12. What is the advantage of false color in representing image data? Think of your nightly weather Doppler radar images.

MULTIPLE CHOICE QUESTIONS

1. As a particle's speed approaches the speed of light, its energy (a) approaches mc^2, where m is the rest mass, (b) approaches its kinetic energy, $1/2\ mv^2$, (c) approaches the product of the particles momentum and the speed of light, pc, (d) approaches γm, where γ is the Lorentz factor.
2. Which of the following is not true about the photoelectric effect? In each case assume that light of a given color is directed onto the emitter plate and a current of electrons is observed to be ejected from the plate. (a) The maximum kinetic energy of the ejected electrons is independent of the intensity of the light. (b) When the intensity of the light is lowered below a finite critical value that depends on the material of the emitter plate, the current abruptly stops. (c) It takes the same very short time to produce a current after turning the light on when the light has intensity I as when it is has intensity $I/2$. (d) The work function of the emitter plate is independent of the color of the light.

Questions 3 and 4 refer to an intensity I of yellow light incident on an ideal 100% efficient metal emitter surface in a phototube producing photoelectrons at a rate of N photons per second.

3. Shining blue light of intensity $I/2$ on the same metal surface in the phototube will (a) not produce any photons, (b) produce $2N$ photons/s, (c) produce $N/2$ photons/s, (d) it is impossible to predict the outcome.
4. Shining red light of intensity $2I$ on the same metal surface in the phototube will (a) produce $2N$ photons/s, (b) produce $N/2$ photons/s, (c) not produce any photons, (d) it is impossible to predict the outcome.
5. The de Broglie wavelength of an electron is associated with what kind of wave? (a) Electric field, (b) magnetic field, (c) probability, (d) sound.
6. The ratio of the Compton shift at forward scattering to that at backward scattering is (a) 2, (b) 1, (c) 0, (d) 1/2.

Questions 7–9 refer to the particle in a box problem, where the particle is confined between 0 and L.

7. A particle in its first excited state is most likely to be found at (a) $L/2$, (b) $L/3$, (c) $L/4$, (d) L.
8. A particle in its second excited state will never be found at (a) $L/4$, (b) $L/3$, (c) $L/2$, (d) it can be found everywhere in the box at some time.
9. When a particle in a box makes a transition from its third excited state to its ground state, the emitted energy equals (a) 9, (b) 5, (c) 8, (d) 2 times its zero-point energy.
10. In quantum mechanics an electron is viewed as being described by a wave function. When confined to a finite region of space, the allowed electron wave functions are standing waves. This explains (a) the results of the photoelectric effect, (b) the results of Compton scattering, (c) why an atom must have a lowest energy state in which its electrons cannot radiate away energy, (d) why the sky is blue.
11. The principle of complementarity refers most closely to (a) the uncertainty principle, (b) wave–particle duality, (c) tunneling, (d) zero-point energy.

12. Heisenberg's uncertainty principle (a) only applies to atomic and subatomic particles, (b) predicts large uncertainties in the velocities of macroscopic objects at rest, (c) states that the product of uncertainties in conjugate variables cannot be zero, (d) explains experimental uncertainties in all measured quantities.
13. Tunneling refers to all but which of the following? (a) An electron escaping from a potential well, (b) an electron traveling in a classically inaccessible region of space for a short time, (c) an electron traveling down a channel between atoms, or a tunnel, in a material, (d) the process used in the STM to image atoms.
14. A scanning tunneling microscope requires all but the following: (a) a fine-tipped needle, (b) a stable, vibration-free sample holder, (c) a vacuum pump to put the sample under vacuum, (d) a stable micromotor to move the needle or sample about.
15. The ground state kinetic energy of a macroscopic body consists mostly of (a) coherent motion of the body as a whole, (b) incoherent motion of the electrons of the atoms, (c) incoherent motions of the nucleons, (d) coherent motions of the atoms of the material.
16. When a block slides across a rough horizontal table surface and stops (a) its coherent center-of-mass energy is transformed into internal kinetic energy, (b) its incoherent atomic energy is transformed into incoherent nucleon energy, (c) its coherent center-of-mass energy is transformed into coherent atomic energy, (d) its coherent center-of-mass energy is transformed into photon energy.
17. In order of increasing energy, the different types of energy of a macroscopic body are due to (a) incoherent nucleon motion, incoherent electron motion, incoherent atomic vibrational motion, coherent center-of-mass motion, (b) coherent center-of-mass motion, incoherent atomic vibrational motion, incoherent electron motion, incoherent nucleon motion, (c) incoherent atomic vibrational motion, incoherent electron motion, incoherent nucleon motion, coherent center-of-mass motion, (d) coherent center-of-mass motion, incoherent nucleon motion, incoherent electron motion, incoherent atomic vibrational motion.

PROBLEMS

1. Compute the momentum and energy of a 1 kg rest mass object traveling at v = 0.8 c, 0.9 c, 0.95 c, 0.99 c, and 0.999 c.
2. Repeat the previous problem for an electron and calculate the energy in MeV.
3. Fill in the steps in the derivation of the classical limit of Equation (24.2).
4. Derive Equation (24.8), the connection between relativistic energy and momentum.
5. Calculate the energy of each of the two photons produced from electron–positron pair annihilation when the electron and positron were nearly at rest. What is their wavelength?
6. A 2.5 MeV photon passes near a stationary atom and produces an electron–positron pair. If all the energy of the photon goes into creating the pair, what is the speed of each when produced?
7. What are the momentum, wavelength, and frequency of a 1.2 MeV photon traveling in space?
8. The work function for cesium is 2.9 eV. Suppose a vacuum tube with a cesium photocathode is configured for the photoelectric effect.
   (a) What is the maximum wavelength photon that will produce a photocurrent?
   (b) If 400 nm photons are used what is the maximum kinetic energy of the emitted electrons?
   (c) If a 1 W beam of 400 nm photons is used, what is the photocurrent that will be detected assuming 100% efficiency (i.e., assuming all emitted photoelectrons are collected by the anode)?
   (d) What maximum work function is needed to allow photoelectron emission using green photons of 500 nm wavelength?
9. Photons of 400 nm wavelength are incident on a photocathode. As the anode potential is made more negative, the photocurrent decreases until it reaches zero when the anode voltage is −0.82 V. Find the work function of the photocathode.
10. A photoelectric experiment is conducted with a sodium surface with work function $\Phi = 2.28$ eV.
    (a) When the surface is illuminated with light with a wavelength of 410 nm, what are the speed and kinetic energy of the emitted electron?
    (b) Is the electron relativistic?
    (c) What is the minimum frequency needed to detect a photocurrent?
    (d) What is the maximum wavelength of light that can be used to detect a photocurrent?
    (e) What are the speed and kinetic energy of the emitted electron if the incident light is 700 nm on the same sodium surface?
11. Suppose that ^{134}Cs, a gamma ray emitter, is used in a Compton effect experiment and the gamma rays are observed to scatter from electrons in an Al target at a 50° angle. ^{134}Cs is radioactive and decays by producing a 1.6 MeV gamma ray, which is just like an x-ray except it has a higher energy. (^{134}Cs also emits β particles in addition to γ rays and has a half-life of about 2.1 years, both of which have nothing to do with the problem.)
    (a) What is the wavelength and momentum of the incident gamma ray?
    (b) Write an expression for the energy of the scattered photon as a function of incident energy photon and the scattering angle Φ.

(c) What is the energy of the scattered γ-ray photon in MeV?
(d) What is the kinetic energy (in MeV) of the recoiling electron?
(e) What is the speed of the recoiling electron as a fraction of c?

12. A 0.012 nm wavelength beam of x-rays is incident on a foil target.
(a) What is the incident x-ray photon energy in MeV?
(b) What is the wavelength and energy of backscattered Compton x-rays?
(c) How much energy is given to the foil target for each backscattered x-ray?

13. Find the relativistic energy (in MeV) of an electron with a de Broglie wavelength of 0.0012 nm.

14. After learning about de Broglie's hypothesis that particles of momentum p have wave characteristics with wavelength $\lambda = h/p$, a 65 kg student has grown concerned about being diffracted when passing through a 90 cm wide doorway.
(a) If the student is traveling at a whopping 0.5 m/s, what is the student's momentum?
(b) What is the de Broglie wavelength of the student?
(c) What would the size of the door need to be in order for there to be noticeable diffraction of the student?

15. Suppose that a 1 mW He–Ne laser ($\lambda = 633$ nm) shines on a screen. How many photons strike the screen each second? (No wonder we are not aware of individual photons!)

16. An electron is trapped in a 10 nm one-dimensional deep potential well. Find the following.
(a) Its ground state energy
(b) The energy of the second excited state above the ground state
(c) The minimum quantum number n corresponding to an energy of at least 100 eV

17. Show that for a particle in a box the difference in energy between consecutive energy levels increases in proportion to the quantum number n.

18. An atomic transition from an excited state to the ground state has a lifetime of 10^{-8} s. What is the uncertainty in the energy of the approximately 550 nm photon emitted? What is the uncertainty in the wavelength of the photon?

19. What is the minimum velocity of an electron in a hydrogen atom, confined within a distance of about 0.1 nm?

20. What is the minimum uncertainty in the velocity of a 2000 kg truck waiting at a red light (or its maximum possible velocity) when its position is measured to an uncertainty of 1.0×10^{-10} m.

21. An electron travels down a channel between two parallel arrays of large atoms along the x-axis separated by 0.12 nm. What is the minimum uncertainty in the y-momentum of the electron?

22. Using the uncertainty principle, derive Equation (24.19) for the zero-point energy of a particle in a box, apart from a small numerical factor.

23. Alpha decay in radioactive nuclei can be thought of as the escape of a helium nucleus from the attractive barrier potential of the larger nucleus. If the nucleus diameter is 5.5 fm, find the maximum velocity of the alpha particle in the nucleus.

25 The Structure of Matter

We continue our study of quantum mechanics with a detailed historical discussion of the simplest atom, hydrogen, and the ad hoc explanation by Bohr and others of its properties. Then we turn to a qualitative discussion of the quantum mechanical theory of atoms and molecules based on allowed quantum numbers. These numbers follow from the Schrödinger equation for the atom or molecule, but we do not show those details here. Some of the spectroscopic implications of our energy level discussions are given for simple systems. Lasers, now found all around us from the supermarket to the CD player, are discussed in the final section of this chapter. We learn how they work, what types of lasers there are, and something about many of their most important applications.

1. THE SIMPLE HYDROGEN ATOM

An early quantum model for simple atoms was developed by Niels Bohr in 1913. Although this theory has been superseded by modern quantum mechanics, discussed further in the next section, it successfully explained the details of the spectra of light emitted by those atoms when they were heated or otherwise given energy. We discuss Bohr's theory because, despite its limitations, it gives some insight into the quantization of fundamental quantities such as energy and angular momentum, as we show.

At the time of Bohr's proposal, atoms were known to consist of a tiny positive nucleus that had just been discovered by Rutherford, surrounded by orbiting electrons. In 1910 Rutherford had established, from observations on the scattering of positive alpha particles, that the positive charges in an atom were localized within a central tiny nucleus. The alpha particles were generated by nuclear decay reactions and were known to have an electric charge of $+2e$; we now know that they are high-energy, doubly ionized positive nuclei of helium atoms. Rutherford found that in directing a beam of these particles on thin foils of metals, although most of them passed straight through, some of them were deflected through large angles, and some even directed backwards. He immediately realized that the only way such large deflections could be produced was if the positive charges of the atoms in the foil were very concentrated so that when the positive alpha particles happen to graze by, there would be a strong electrical repulsion.

In the early 1900s, the stability of atoms was not understandable in terms of the classical physics of Newton and Maxwell. The simplest atom, hydrogen, was pictured to consist of an electron orbiting in a circular path about a single proton under the influence of the electrostatic attractive force. However, according to classical electromagnetism, accelerating charges, such as the electron traveling about the nucleus, should emit radiation at the orbital frequency. Then as the electron would lose this energy in the form of light, it should keep spiraling in towards the nucleus, emitting a continuous spectrum of light as the orbital frequency changes. In reality, of course,

J. Newman, *Physics of the Life Sciences*, DOI: 10.1007/978-0-387-77259-2_25,

atoms are stable and although they do emit light, they emit a discrete spectrum of light with only certain particular "lines" or colors for a type of atom. Classical physics could not explain the simplest atomic or line spectrum.

Bohr proposed a model of the hydrogen atom that was able to predict the wavelengths of its discrete line spectrum with an accuracy of about 0.02%. From several assumed postulates, Bohr was able to derive expressions for the energies and radii of the hydrogen atom in its ground and excited states. He assumed that the orbiting electron in hydrogen can exist in any of a set of discrete orbits, each with a corresponding energy, radius, and angular momentum, known as *stationary states*. In an ad hoc assumption, Bohr proposed that when in one of these stationary states, the electron does not emit any radiation, despite its constant acceleration. Bohr assumed that, aside from the electron not radiating, classical physics otherwise correctly describes the motion of the electron in a stationary state, but that the electron could make abrupt transitions between such states, during which time classical physics is not obeyed. When transitions occur, the energy difference between the stationary states corresponds to the energy of an absorbed or emitted photon. Thus, only during transitions between stationary states does an atom emit radiation. Bohr's final assumption was that the stationary states are characterized by discrete values of the orbital angular momentum. These are given by multiples of a fundamental quantum of angular momentum, $h/2\pi$, so that

$$L_n = n\frac{h}{2\pi} = n\hbar, \tag{25.1}$$

where n is an integer, $n = 1, 2, 3, \ldots$, $\hbar$ is short for $(h/2\pi)$ and is read "h-bar," and L_n is the angular momentum of the nth stationary state.

Bohr's postulate that the angular momentum of the electron in a hydrogen atom should be quantized can be rationalized by invoking the concept of standing de Broglie waves. Just as with waves on a string, sound in an air column, or light, matter waves exhibit interference. If the circumference of the electron orbit matched a nonintegral number of de Broglie waves, then these waves would average away to zero because of destructive interference. Each time the wave traveled around the circular orbit it would arrive at a different phase and this nondefinite phase relation would wash out the matter wave. Only with an integral number of wavelengths fitting around the circumference will constructive interference occur, leading to standing matter waves. If an integer number of de Broglie wavelengths must fit around a circular orbit of the electron then we must have that $n\lambda = 2\pi r$ (Figure 25.1). Substituting h/p for λ we find that requiring standing waves implies that $pr = n\hbar$, but because $pr = mvr = L$, we recover Bohr's angular momentum quantization condition.

From these assumptions we can derive all the information needed to describe the hydrogen atom. As mentioned above, we assume the electron orbits the stationary proton in one of a set of discrete circular orbits of radius r_n with the centripetal force generated by the electrostatic attraction. We can then write $F = ma$ for the electron as

$$\frac{1}{4\pi\varepsilon_o}\frac{e^2}{r_n^2} = m\frac{v_n^2}{r_n}, \tag{25.2}$$

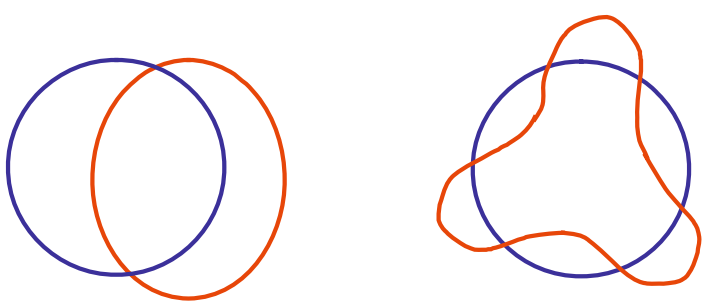


FIGURE 25.1 *Two different states of an orbiting electron:* (left) *ground state with a single wavelength* (red) *fitting around the circular orbit* (blue); (right) *second excited state* (red) *with three wavelengths fitting around the circular orbit* (blue).

where v_n is the orbital speed also satisfying Equation (25.1), rewritten as

$$L_n = mv_n r_n = n\hbar. \tag{25.3}$$

Solving for v_n in Equation (25.3),

$$v_n = \frac{n\hbar}{mr_n}, \tag{25.4}$$

and substituting this expression into Equation (25.2) to eliminate v_n, we find

$$\frac{1}{4\pi\varepsilon_o}\frac{e^2}{r_n^2} = m\left(\frac{n\hbar}{mr_n}\right)^2\frac{1}{r_n}. \tag{25.5}$$

Solving for the radius of the nth orbit we find that

$$r_n = n^2\, r_1\,, \tag{25.6}$$

where the smallest possible orbital radius of the electron, known as the *Bohr radius*, is given by

$$r_1 = \frac{\varepsilon_o\, h^2}{\pi me^2} = 0.53 \times 10^{-10}\, m. \tag{25.7}$$

According to Equation (25.6) the orbital radii grow in proportion to n^2 so that the radii are given by r_1, $4r_1$, $9r_1$, $16r_1$, . . .with the spacing between orbits increasing rapidly.

Using our results thus far, we can next calculate the possible total energy of the electron in the various stationary states. The classical expression for the total energy is given by $E = KE + PE$, or in our case

$$E_n = \frac{1}{2}mv_n^2 - \frac{1}{4\pi\varepsilon_o}\frac{e^2}{r_n}, \tag{25.8}$$

the negative sign arising from the opposite charges of the electron and proton. This can be rewritten, after substituting for $mv_n{}^2$ from Equation (25.2) as,

$$E_n = \frac{1}{2}\left(\frac{e^2}{4\pi\varepsilon_o\, r_n}\right) - \frac{1}{4\pi\varepsilon_o}\frac{e^2}{r_n} = -\frac{e^2}{8\pi\varepsilon_o\, r_n}. \tag{25.9}$$

Then, substituting for r_n from Equations (25.6) and (25.7), we find that

$$E_n = -\frac{me^4}{8\varepsilon_o^2\, h^2}\frac{1}{n^2} = \frac{E_1}{n^2}, \tag{25.10}$$

where

$$E_1 = -\frac{me^4}{8\varepsilon_o^2\, h^2} = -13.6\, eV.$$

The stationary state corresponding to $n = 1$ is known as the *ground state* and is the lowest energy (most negative) and smallest angular momentum state of the hydrogen atom, as well as the orbit with the smallest radius. Note that all of the energies given by Equation (25.10) are negative, with values approaching $E = 0$ as n increases. In general for atomic systems, negative energy values indicate that the electron is bound to the nucleus, and positive energy values mean that the atom is ionized and that the electron is free from the nucleus. All the other states, for $n = 2, 3, \ldots$ are known as *excited states*, with the $n = 2$ state being the first excited state, and so on, and have larger (but still negative) energies given by Equation (25.10). We can picture the energy levels of the hydrogen atom in an *energy level diagram* shown in Figure 25.2.

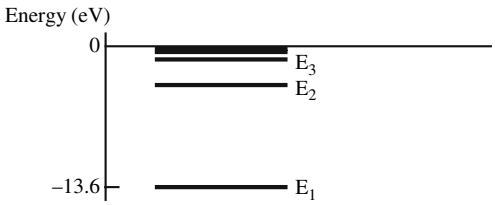


FIGURE 25.2 *Energy level diagram for the hydrogen atom.*

Hydrogen atoms in the ground state cannot radiate energy according to Bohr's postulate, and remain in the ground state unless energy is added, for example, by the absorption of a photon. Our energy level diagram suggests that only certain photon energies can excite hydrogen atoms from the ground state; photons must have energies corresponding to the difference between energies of stationary states

$$E_{\text{photon}} = hf = E_{\text{final}} - E_{\text{initial}}, \tag{25.11}$$

where final and initial correspond to the excited state and the ground state in this case. In general, transitions can occur between all stationary states with a photon either being emitted, if the transition corresponds to a decrease in energy of the atom, or absorbed, if the transition corresponds to an increase in its energy. For the first case, we show the emission spectrum of hydrogen with a set of discrete energy photons being emitted from all possible downward transitions, whereas for the second case one can measure the absorption spectrum of hydrogen with a discrete set of photons being absorbed from all possible upward transitions.

Figure 25.3 shows an energy level diagram with the emission line spectra for hydrogen labeled in different series of lines. Each series corresponds to a set of transitions to a particular final lower energy state n_f from each of the higher energy states n_i. The photon energies can be written, using Equations (25.10) and (25.11), as

$$E_{\text{photon}} = \frac{hc}{\lambda} = E_1\left(\frac{1}{n_f^2} - \frac{1}{n_i^2}\right), \tag{25.12}$$

or rewriting this to solve for $1/\lambda$ after substituting for E_1 we find

$$\frac{1}{\lambda} = R\left(\frac{1}{n_f^2} - \frac{1}{n_i^2}\right). \tag{25.13}$$

The value of R, known as the *Rydberg constant*, is

$$R = \frac{me^4}{8\varepsilon_o^2 h^3 c} = 1.097 \times 10^7\, m^{-1} \tag{25.14}$$

Equation (25.14) gives excellent predictions of the measured line spectra for hydrogen. The Balmer series, for example, consists of those transitions ending at the $n_f = 2$ state. Three of these lines correspond to visible photons as shown in the emission spectrum of hydrogen in Figure 25.3. The success of Bohr's theory was tremendous, although it is clearly an incomplete theory that cannot be used to calculate many measurable quantities. For example, it is impossible to calculate the

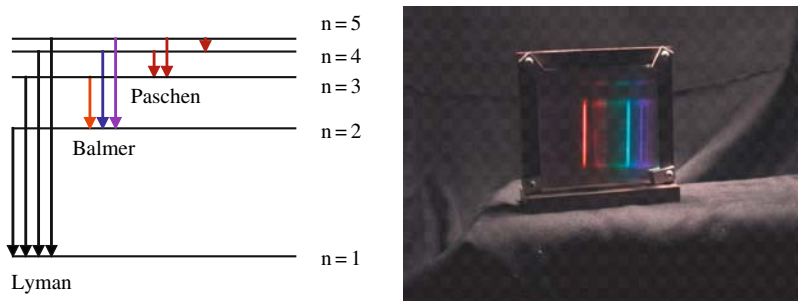


FIGURE 25.3 (left) *Possible energy transitions for hydrogen grouped into families based on the final state. Also shown are the three visible lines of the Balmer series* (red, blue, *and* violet), *the only visible emission spectrum lines for hydrogen. Note that* black *lines are ultraviolet and* dark red *lines are infrared.* (right) *The three visible Balmer lines shown in a reflection grating used to disperse the spectrum.*

lifetimes of excited states. Simple variations on the calculations gave correct results for more complex atoms that had been ionized to have just a single outer (valence) electron. Bohr's theory has been superceded by quantum mechanics, developed in the 1920s and 1930s as briefly discussed in the next section.

2. QUANTUM NUMBERS AND SPIN

Bohr theory, a mix of classical physics and unsupported postulates, leaves much atomic physics unexplained. The energy levels and Bohr radius of the isolated hydrogen atom are correctly predicted and the notion that transitions between stationary states are responsible for line spectra is correct, however, there is no way to calculate how long an atom will remain in an excited state before emitting a photon (equivalently, the transition rates) and no way to understand more complex atoms with additional electrons. Nor does Bohr's theory allow one to understand the interaction of the hydrogen atom with electromagnetic fields. Furthermore, as mentioned above, classical physics predicts that an electron accelerating in a circular orbit should radiate continuously and Bohr's theory simply postulates the stability of stationary states.

Quantum mechanics retains the idea of stationary states for an atom and that line spectra are due to transitions between such states, but the picture of the electron orbiting the nucleus in a classical trajectory, circular or otherwise, is dropped. Bohr's theory for hydrogen atoms centered on the postulate of the quantization of angular momentum in multiples of $\hbar$, where the integer multiple n, also defined the energy level and radius of the circular orbit. Quantum mechanics starts with the Schrödinger equation, mentioned in the last chapter, and the form of the Coulomb potential energy of interaction between the electron and proton in hydrogen to derive the possible energy levels of the isolated hydrogen atom. The result is the same Equation (25.10), with the integer n retained as the *principal quantum number*, but this number now is not related to the angular momentum of the electron. We show that there are three other quantum numbers that are needed to completely specify the possible wave functions for the hydrogen atom.

The simple de Broglie picture of fitting integral numbers of wavelengths in circular orbits does not give the correct angular momentum. The electron's orbital angular momentum is not simply related to n, as Bohr postulated, but to another quantum number ℓ, known as the *orbital quantum number* by

$$L = \sqrt{\ell(\ell + 1)}\,\hbar. \tag{25.15}$$

Here ℓ is an integer that is in the interval from 0 to $(n - 1)$, so that ℓ has the possible values

$$\ell = 0, 1, 2, \ldots, n - 1. \tag{25.16}$$

Thus for a given principal quantum number, n, the orbital angular momentum can have n different possible values. For example if $n = 2$, then ℓ can have two different values, $\ell = 1$ or 0, corresponding to the electron having two possible values of orbital angular momentum, $L = \sqrt{2}\,\hbar$ and $L = 0$, respectively, according to Equation (25.15).

Although for an isolated hydrogen atom the direction of the orbital angular momentum vector can have no significance, as soon as the hydrogen atom is allowed to interact with an external field or another atom, the orientation of this vector becomes important. A third quantum number, the magnetic quantum number m_ℓ gives the projection of the orbital angular momentum along a particular direction in space, usually chosen to be the z-axis, along which the external field lies. The term "magnetic" is used here because the interaction of a hydrogen atom with a magnetic field led to the introduction of this usage. The z-component of orbital angular momentum is given by

$$L_z = m_\ell \hbar, \tag{25.17}$$

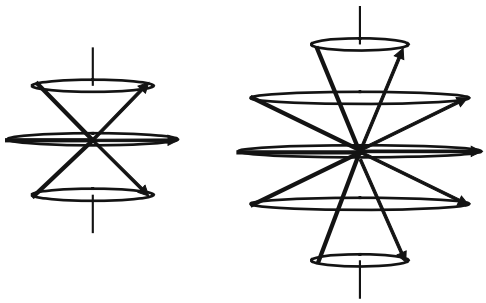


FIGURE 25.4 *Orbital angular momenta for two different values of ℓ. The z-components are given by Equation (25.17), and the magnitude of the vectors is given by Equation (25.15).*

where the magnetic quantum number can take on any integer value between $\pm\ell$; that is,

$$m_\ell = -\ell, -\ell+1, -\ell+2, \ldots, 0, 1, 2, \ldots, \ell. \tag{25.18}$$

For any given value of ℓ, there are $2\ell + 1$ different possible orientations of the orbital angular momentum. Recall that although we are not seeing the details here, these recipes for n, ℓ, and m_ℓ arise directly from solving Schrödinger's equation for the H atom. According to this, not only is the magnitude of the angular momentum quantized, according to Equation (25.15), but so is the orientation of the angular momentum in space according to Equations (25.17) and (25.18). The angular momentum vector is confined to one of several discrete angles with respect to the z-axis. Figure 25.4 shows the orientation of orbital angular momentum for the cases of $\ell = 1$ and 2. Note that the angular momentum vectors lie on cones that have projections on the z-axis given by Equation (25.17). These vectors must have specific projections on the z-axis but can precess (wobble) around the z-axis so long as they lie on one of these cones.

Example 25.1 Compute the possible spatial orientations for the orbital angular momentum of an $\ell = 2$ electron, one with $L = \sqrt{(2)(3)}\hbar = \sqrt{6}\hbar$.

Solution: As shown in Figure 25.4 for the case of an $\ell = 2$ electron, there are five possible spatial orientations of the orbital angular momentum, characterized by m_ℓ values ranging from $+2$ to -2. We can find the angles θ that the orbital angular momentum vector can make with the z-axis by noting that

$$\cos\theta = \frac{L_z}{L} = \frac{m_\ell \hbar}{\sqrt{\ell(\ell+1)}\hbar}.$$

With $\ell = 2$ and the range of m_ℓ values, we find the possible angles to be $\theta = 35.3°$, 65.9°, 90°, 114.1°, and 144.7°. Note that these define five cones (with the one at 90° degenerating into a circle) symmetrically arranged about the horizontal plane as shown in Figure 25.4.

In addition to their orbital angular momentum, electrons also have an *intrinsic spin angular momentum* that is an internal property of the electron. Spin angular momentum S is given by a similar equation to Equation (25.15),

$$S = \sqrt{s(s+1)}\hbar, \tag{25.19}$$

where $s = \frac{1}{2}$ for the electron, so that for the electron S is fixed at

$$S = \sqrt{\left(\frac{1}{2}\right)\left(\frac{3}{2}\right)}\hbar = \frac{\sqrt{3}}{2}\hbar.$$

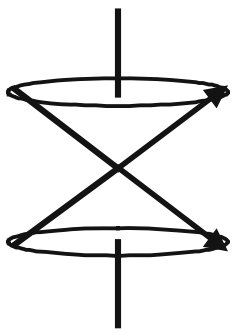


FIGURE 25.5 *Electron spin has two possible orientations, loosely referred to as up and down.*

Just as with orbital angular momentum, spin angular momentum is also spatially quantized in that S can only point in one of two directions (with $s = \frac{1}{2}$, we have that $2s + 1 = 2$ possible orientations, the same rule as for orbital angular momentum) given by

$$S_z = m_s \hbar, \tag{25.20}$$

where the spin quantum number m_s is given by $\pm\frac{1}{2}$ (m_s varies from $+s$ to $-s$, again with the same rule as for orbital angular momentum). Figure 25.5 shows the possible orientations of the electron's spin angular momentum; these orientations are said to

have spins of $\frac{1}{2}$ or $-\frac{1}{2}$, or are said, very loosely, to have spin up or down, respectively, although the spins do not point along the z-axis. The Stern–Gerlach experiment, discussed in Chapter 17, was an important confirmation of the notion of electron spin. Recall that atoms with no orbital angular momentum when steered through an inhomogeneous magnetic field either were deflected upwards or downwards depending on the spin of their outer electron.

We have now introduced the four quantum numbers, n, ℓ, m_ℓ, and m_s, that are needed to fully describe the different possible wave functions (or so-called stationary states) of the hydrogen atom. The ground state of hydrogen is the $n = 1$ state with $\ell = m_\ell = 0$. Figure 25.6 shows the wave function for this state to be spherically symmetric with the greatest density of the electron (position of greatest probability to be found) located at the Bohr radius. The $n = 2$ states can have $\ell = 0$ or 1, with corresponding values of $m_\ell = 0$ if $\ell = 0$, or $m_\ell = 0$ or $\pm$ 1 if $\ell = 1$. States with $\ell = 0$ are spherically symmetric, whereas the others are asymmetric; Figure 25.6 also shows some of these states. Note that all the excited states have regions of zero electron density (probability) located "within" the atom, indicating nodes of the wave function. These nodes are analogous to those of the excited states of a one-dimensional wave function for a particle in a box, seen in the previous chapter. As we show in the next section, the same four quantum numbers can be used to label the wave functions of all types of atoms, although the expressions for the energy levels are different than those for hydrogen.

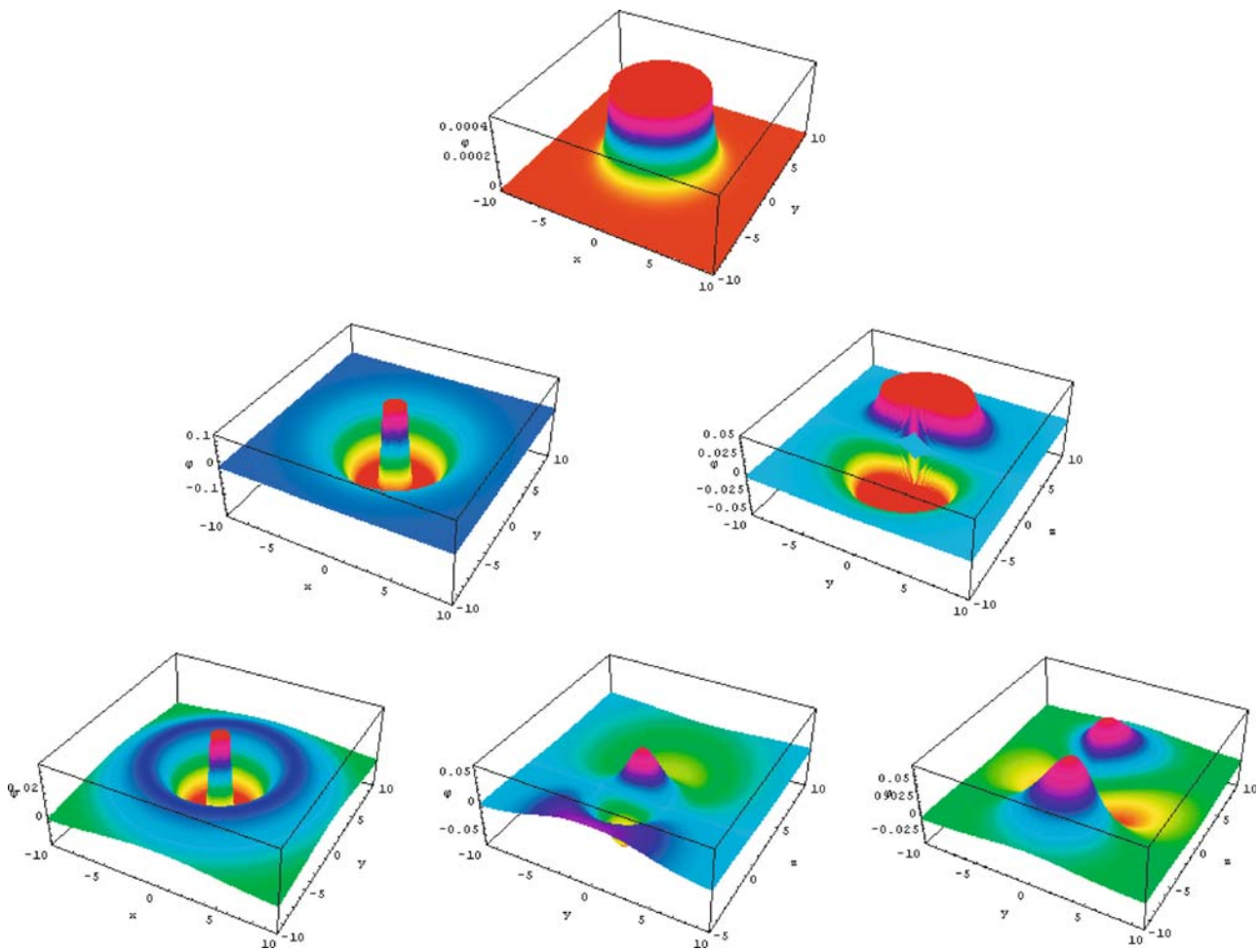


FIGURE 25.6 *Color-coded hydrogen atom wave functions in two dimensions:* (top) *Spherically symmetric n = 1 ground state; (second row,* left to right*), [n = 2, ℓ = 0] and [n = 2 ℓ = 1]; (bottom row,* left to right*), [n = 3, ℓ = 0], [n = 3, ℓ = 1], [n = 3, ℓ = 2], with m_ℓ = 0 in all cases. Note the "internal" node structure of the excited state wave functions. Note also that the ℓ ≠ 0 states are not spherically symmetric, but are more complex with higher ℓ states.*

3. THE PAULI EXCLUSION PRINCIPLE, THE PERIODIC TABLE, AND CHEMISTRY

The four quantum numbers described in the last section for the possible stationary states of the hydrogen atom label a discrete set of such states. Each different combination of quantum numbers represents a different possible state. For an isolated hydrogen atom, because the energy of any state depends only on the principal quantum number n, many different quantum states will have the same energy; these are said to be *degenerate states.* In the presence of an external field or a field due to other nearby atoms, this degeneracy is said to be "split" and the individual states with different values of ℓ, m_ℓ, and m_s may have different energy values, depending on the type of interaction. In general, any quantum mechanical system will have such a set of quantum numbers, perhaps including additional numbers beyond our four that serve to label a discrete number of such states. It is the electron configuration of atoms or molecules that determines all of its physical and chemical properties, with the nuclei playing no direct role in determining those properties. We discuss the nucleus and its structure and interactions in the next chapter in connection with radioactivity.

It is natural to ask whether the electrons in a more complex multielectron atom or molecule, in its ground state, for example, will all occupy the same set of quantum numbers representing the lowest energy state for this atom. For example, do all three electrons of lithium occupy the $n = 1$ state with ℓ and m_ℓ both equal to zero? The answer turns out to be no. New physics is needed to determine the actual quantum numbers of a multielectron atom or more complicated system. The necessary idea is contained in the *Pauli exclusion principle*

At most a single electron can occupy any quantum state.

A quantum state means a state labeled by a complete specific set of quantum numbers. For a multielectron atom, the ground state represents the lowest possible energy of the atom. In the absence of the Pauli exclusion principle, all electrons would occupy the lowest energy level and all would have the same set of quantum numbers, those representing the state of lowest energy. The effect of the Pauli exclusion principle in requiring electrons to have different quantum numbers is to raise the minimum energy of the atom from that hypothetical situation. For each set of quantum numbers n, ℓ, and m_ℓ there are two possible spin quantum numbers m_s, and therefore at most two electrons can have the same set of values for n, ℓ, m_ℓ, as long as one has $m_s = +\frac{1}{2}$, and the other has $m_s = -\frac{1}{2}$, or vice versa.

The elementary particles of nature can be classified according to their intrinsic spin s as either *fermions*, those with half-integral spin such as the electron, proton, and neutron each with spin $\frac{1}{2}$, or as *bosons*, those with integral spin such as the photon with spin 1 or the pion with spin 0. The Pauli exclusion principle holds only for fermions. We show here that when applied to electrons in atoms and molecules, this principle leads to an understanding of their structure and spectra. In the next chapter we show that when applied to protons and neutrons, the Pauli exclusion principle leads to a set of nuclear energy levels with transitions between them producing a "photon line spectra" of high-energy gamma ray photons, very similar to the spectra produced by electron transitions in atomic and molecular systems. Bosons, on the other hand, do not obey the Pauli exclusion principle and are not restricted in their occupancy of any state. Thus, it is possible to have any number of photons simultaneously occupying the same ground state of a system. This gives rise to many interesting phenomena including superfluids and superconductors, discussed at the end of Section 2 of the previous chapter, and the relatively recently discovered phenomenon of Bose–Einstein condensation.

We show that the Pauli exclusion principle can explain the arrangement of all the elements of nature in the Periodic Table of the Elements (Figure 25.7) in terms of the electron configurations associated with the atoms. The elements in any column of the Periodic Table have similar chemical properties. These similarities are due to the arrangement of the outermost or *valence* electrons.

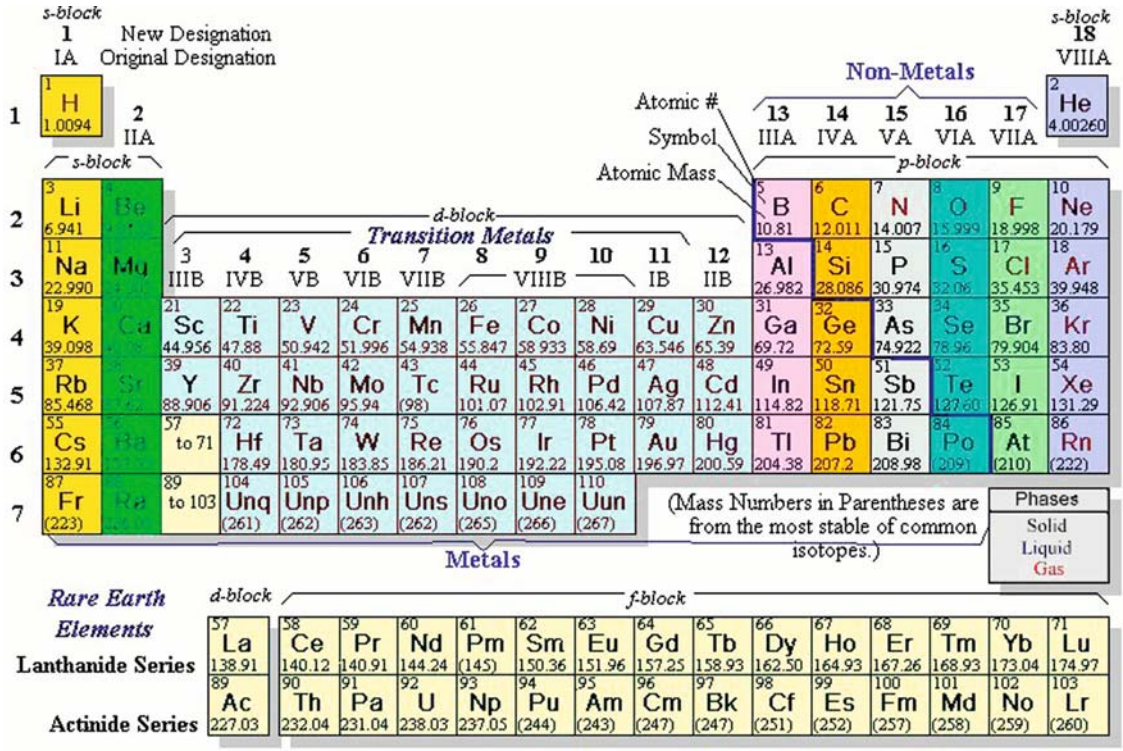


FIGURE 25.7 *The Periodic Table of the Elements.*

Imagine that we are assembling multielectron atoms from their individual parts. If we start with the nucleus for an atom with N protons and the appropriate number of neutrons, and we imagine that we assemble the N electrons to complete the atom in its ground state, then each added electron will occupy the lowest energy level available to it. If $N = 2$ (helium), then the two electrons can occupy the state $n = 1$, $\ell = m_\ell = 0$, completing the $n = 1$ possible states with paired electron spins, $m_s = \pm\frac{1}{2}$. These states are labeled using spectroscopic notation as the $1s^2$ states, where the first 1 indicates the n value, s (for "sharp") indicates that $\ell = 0$, and the superscript indicates the number of electrons in that state. If $N = 3$ (lithium), then the third electron must occupy the lowest energy state in the $n = 2$ level with quantum numbers $\ell = m_\ell = 0$ and $m_s = +\frac{1}{2}$ or $-\frac{1}{2}$, because the $n = 1$ state is filled. Notice that for helium, the two electrons are paired with opposite spins and that they complete the allowed $n = 1$ "shell," known as the K shell. Helium is very stable and unreactive and is therefore chemically inactive and known as an inert, or rare, gas. For lithium, the unpaired $n = 2$ electron is the valence electron and is highly reactive and available to form a chemical bond with an unpaired electron on another atom. In spectroscopic notation the valence electron in ground state lithium is denoted by $2s^1$ and the entire atom is given as $1s^2 2s^1$.

For the $n = 2$ shell of electrons, known as the L shell, there are a total of 8 possible electron states shown in Figure 25.8. As we increase N in our assembly line for atoms, we will fill the L shell when the atom has a total of 10 electrons, 2 in the K shell and 8 in the L shell. A glance at the Periodic Table will show that this 10-electron atom is neon, the second inert gas. In spectroscopic notation neon is given by $1s^2 2s^2 2p^6$, where p (for "principal") stands for $\ell = 1$; of the 8 electrons in the L shell, 2 are s state and 6 are p state electrons. The first two rows of the Periodic Table are arranged according to the

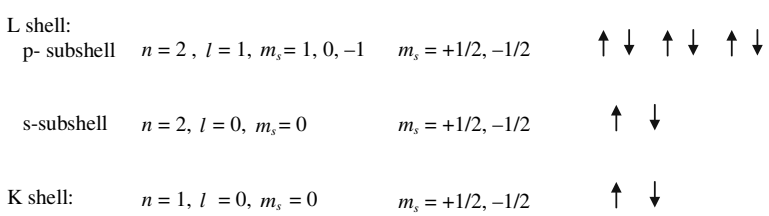


FIGURE 25.8 *Electron configuration for the K and L shells.*

principal quantum numbers $n = 1$ with two elements and the $n = 2$ with 8 elements. Each column of the Periodic Table contains elements with similar chemical characteristics because their outer, or valence, electron configurations are similar. Spectroscopic notation for heavier atoms continues with the $\ell = 2$ state labeled by d (for "diffuse"), $\ell = 3$ by f, with higher ℓ states labeled alphabetically as g, h,. . . . Table 25.1 shows the ground state configurations, in spectroscopic notation, for the first 20 atoms of the Periodic Table.

Table 25.1 *Some Ground State Electron Configurations*

Element (Symbol)	*Valence Electron (n, ℓ)*	*Spectroscopic State*	*Element (Symbol)*	*Valence Electron (n, ℓ)*	*Spectroscopic State*
Hydrogen (H)	(1,0)	$1s^1$	Sodium (Na)	(3,0)	$1s^2 2s^2 2p^6 3s^1$
Helium (He)	(1,0)	$1s^2$	Magnesium (Mg)	(3,0)	$1s^2 2s^2 2p^6 3s^2$
Lithium (Li)	(2,0)	$1s^2 2s^1$	Aluminum (Al)	(3,1)	$1s^2 2s^2 2p^6 3s^2 3p^1$
Beryllium (Be)	(2,0)	$1s^2 2s^2$	Silicon (Si)	(3,1)	$1s^2 2s^2 2p^6 3s^2 3p^2$
Boron (B)	(2,1)	$1s^2 2s^2 2p^1$	Phosphorus (P)	(3,1)	$1s^2 2s^2 2p^6 3s^2 3p^3$
Carbon (C)	(2,1)	$1s^2 2s^2 2p^2$	Sulfur (S)	(3,1)	$1s^2 2s^2 2p^6 3s^2 3p^4$
Nitrogen (N)	(2,1)	$1s^2 2s^2 2p^3$	Chlorine (Cl)	(3,1)	$1s^2 2s^2 2p^6 3s^2 3p^5$
Oxygen (O)	(2,1)	$1s^2 2s^2 2p^4$	Argon (Ar)	(3,1)	$1s^2 2s^2 2p^6 3s^2 3p^6$
Fluorine (F)	(2,1)	$1s^2 2s^2 2p^5$	Potassium (K)	(4,0)	$1s^2 2s^2 2p^6 3s^2 3p^6 4s^1$
Neon (Ne)	(2,1)	$1s^2 2s^2 2p^6$	Calcium (Ca)	(4,0)	$1s^2 2s^2 2p^6 3s^2 3p^6 4s^2$

Now, let's consider the interaction of two electrically neutral atoms. The protons and electrons in one push and pull on the protons and electrons in the second. If the atoms are sufficiently far apart—a few nanometers will typically do—these pushes and pulls tend to cancel out so that there is essentially zero electrical force of one atom on the other. As the two atoms are brought closer, however, the outer electrons in one tend to push away the outer electrons in the second (see Figure 25.9) leaving the nucleus of the second slightly more exposed to the first's electrons. As this distance of separation decreases, the electrons in each atom tend to become somewhat synchronized. The net effect is that the two atoms become (weakly) attracted to each other. On the other hand, if the two atoms are brought close enough that their electrons begin to interpenetrate, the Pauli exclusion principle rears its head. Electrons mixing between the atoms may well find themselves in the same states of motion with spins aligned. Because electrons abhor such a situation, the atoms at that distance of separation experience a very strong repulsion for each other. Somewhere between where the atoms first begin to attract each other and where the repulsion takes over there is a separation at which the atoms feel no net force due to each other. This separation, typically a few tenths of a nanometer, is called the *equilibrium distance* for the atoms. Because repulsion occurs when the two atoms are closer than the equilibrium distance and attraction occurs when they are farther apart, the equilibrium is stable. The two atoms are said to be *bound* to each other, forming a *stable molecule.*

The robustness of the equilibrium between the two atoms depends on how mixed up the atoms' electrons can get when the nuclei are at the equilibrium separation. Three different scenarios are possible. In the first, the electrons on each atom remain clearly "attached" to their respective nuclei. The electrons on one atom don't mix with the electrons on the other. In this case, the bond between the atoms is called a van der Waals bond and the equilibrium is fairly easy to disrupt. Atoms of the noble gases (helium, neon, argon, etc.) interact with other atoms in this way. Molecules formed between such atoms are extremely difficult to make and maintain. Collisions with other atoms easily cause such molecules to fall apart.

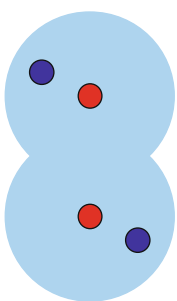


FIGURE 25.9 *Molecular hydrogen.*

In the second scenario, when the two atoms are separated by the equilibrium distance an outer electron from one atom is so strongly attracted to the nucleus of the second that it leaves the first and spends all its time near the second. The atom that loses its electron becomes effectively positively charged and the atom that gains the electron becomes

negatively charged. That is, the two atoms become *ions.* The force of attraction holding the two ions together is said to be an *ionic bond.* Ionic bonds are considerably stronger than van der Waals bonds, but an ionic molecule is still fairly easy to dissociate. For example, the molecule made of one hydrogen atom and one chlorine atom, HCl, is ionically bound. HCl is easily pulled apart when water molecules surround it, producing H^+ and Cl^- ions.

In the third scenario, when the atoms are separated by the equilibrium distance, electrons from each of the two atoms can get scrambled. The two nuclei then share one or more of the atomic electrons, in a kind of game of "electronic catch." Electronic swapping between nuclei results in what is called a *covalent bond.* (Di–)oxygen and (di–)nitrogen are examples of molecules made of two atoms that are covalently bound as is the C–C bond that forms the backbone of most biological molecules. Covalent bonds are generally much stronger than van der Waals or ionic bonds. Although HCl easily dissociates in water at room temperature, O_2 and N_2 do not. Most of the subject matter of chemistry is related to the nature and consequences of molecular bonds. In the next section we consider some of the ways spectroscopy has been used to study the bonding in biomolecular structures.

4. SPECTROSCOPY OF BIOMOLECULES REVISITED

In Chapter 19 we briefly discussed several types of spectroscopic tools in the study of biomolecules. Now that we have a more complete understanding of energy levels in atoms, let's reconsider the types of information that one can learn from spectroscopy, focusing on the study of biomolecules.

In the last section we saw that energy levels in atoms are characterized by a set of quantum numbers defining the valence electron configuration. For atoms, often the energy levels are degenerate in that different angular momentum states of the same principal quantum number have the same energy in the absence of any external interactions. In more complex molecular systems, these degeneracies are not usually present because of the interactions between portions of the molecule and we can, to good approximation, write the total energy of a molecule as

$$E = E_{elec} + E_{vib} + E_{rot}, \tag{25.21}$$

where the three terms are due to the electronic, vibrational, and rotational contributions to the energy, respectively. The total energy is quantized and can be labeled by a set of appropriate quantum numbers with quantized contributions from each term in Equation (25.21). An order of magnitude discussion of the electronic and vibrational energies was given at the end of the last chapter. In this discussion we omit the nuclear contributions which are discussed in the next chapter.

The *electronic energy* term represents the various configurational energies of the molecule's electrons, comparable to the discussion above for atoms, but clearly the energy levels will be richer because there are more valence electrons that interact with each other. *Vibrational energy* arises from relative vibrational motions of the nuclei of the atoms. We saw early in this book (Chapter 4) that any potential energy function appears to be springlike (varying as the square of the distance from equilibrium) for small enough displacements close to equilibrium. Using that, we can picture any chemical bond to be replaced by a spring so that there will be vibrational energies corresponding to small oscillations of the atoms attached by springs. An analysis of this problem using quantum mechanics shows that the vibrational energy levels are given by

$$E_{vib} = (m + \frac{1}{2})hf \quad m = 0,1,2,\ldots, \tag{25.22}$$

where f is the natural frequency of oscillation of the spring, which depends on the details of the chemical bonding interactions. These vibrational energy levels are equally spaced with energy differences about 100 times smaller than electronic energy level differences. Because electronic energy level differences are on the order of eV, corresponding to ultraviolet or visible photons, vibrational energies correspond to infrared photons.

In addition to vibrational energy levels, molecules also have overall rotations with corresponding angular momentum due to this tumbling motion. Because the classical expression for rotational kinetic energy can be written as

$$KE_{rot} = \frac{1}{2} I\omega^2 = \frac{L^2}{2I},$$

where I is the moment of inertia, ω the angular velocity, and L the angular momentum (recall that $L = I\omega$), and using Equation (25.15) for the quantization of L, we can write that the *rotational energy* is

$$E_{rot} = \frac{\ell(\ell + 1)\hbar^2}{2I} \quad \ell = 0, 1, 2, \ldots . \tag{25.23}$$

Rotational energy levels have energy differences about another 100-fold smaller than vibrational and transitions between these levels give rise to far-infrared or microwave photons. Figure 25.10 shows a schematic energy level diagram for a typical small molecule.

Example 25.2 The natural frequency of vibration of the *NO* molecule is 5.63×10^{13} Hz and its moment of inertia is 1.64×10^{-46} kg-m^2. Find which rotational quantum number in the ground vibrational state corresponds to the same energy as the first vibrational excited state.

Solution: The first excited vibrational state has an energy of $E = (3/2)hf = 5.6 \times 10^{-20}$ J $= 0.35$ eV. To find which rotational level has this same energy we can set this energy equal to

$$\frac{\ell(\ell+1)\hbar^2}{2I}$$

and solve for the quantum number ℓ using the given value for I to find that $\ell(\ell + 1) = 41.8$. If $\ell = 6$ these energy levels will match up fairly well.

Spectroscopic techniques that probe the rotational and vibrational energy levels of molecules by examining absorption or emission spectra can be used to determine molecular structure and dynamics. Infrared spectra can give "fingerprints" of molecules because the number and variety of chemical bonds is so large that almost no two biomolecules have the same vibrational spectrum (Figure 25.11). Using a variety of calibration frequencies for particular bonds, obtained from measurements on simple molecules, various peaks in a complex spectrum can be identified. Shifts in the

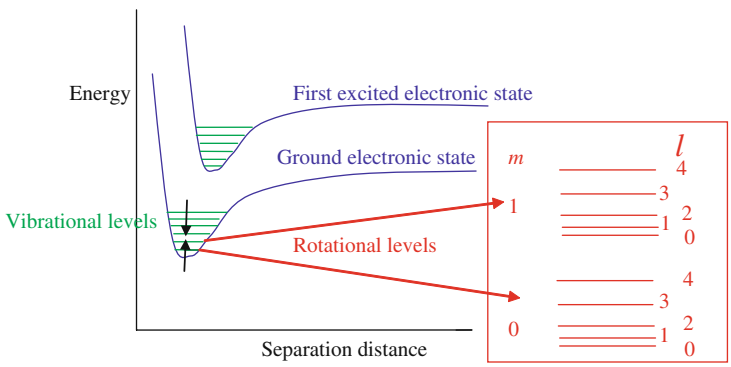


FIGURE 25.10 *Typical molecular energy levels. The curves represent electronic levels* (blue) *with equally spaced vibrational* (m) *levels* (green) *indicated. The detail shows rotational* (ℓ) *energy levels* (red) *within each vibrational level.*

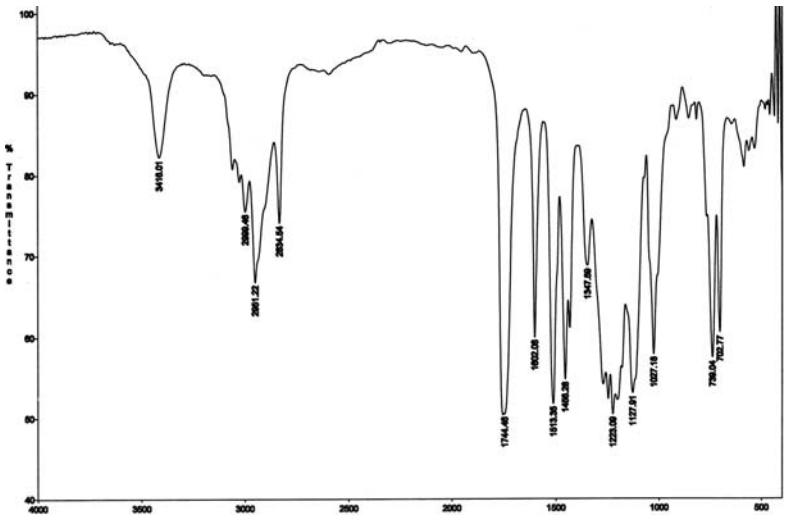


FIGURE 25.11 *IR vibrational spectrum of a small organic molecule with transmittance plotted versus wavenumber. IR spectroscopy uses wavenumber, defined as $2\pi/\lambda$, instead of λ, so that the wavelength range is from about 1.5 mm to 12 mm in this case.*

characteristic frequencies of certain bonds can then be measured as the local environment is changed or as a small ligand molecule binds causing conformational changes. Although rotational spectra have been useful in determining the bond lengths and angles for smaller molecules, they have had limited use for larger macromolecules.

Bulk liquids and solids are really giant molecules made of 10^{20} or more atoms, knit together by small clumps of atoms interacting with other small clumps in macroscopically long chains. Often in solids, atoms are arrayed in orderly repeating patterns called a *crystal.* (See Figure 25.12) The extent of the crystal pattern varies from millions of atoms (a "crystallite"), say, to the entire bulk body (a "single crystal"). In liquids, the regular arrangement of atoms only extends to a few tens of atoms. Liquids are held together by very weak van der Waals type interactions. It is also possible to create a van der Waals solid, but such solids are extraordinarily fragile and only exist at extremely high external pressures and/or low temperatures. Ionic solids, such as common table salt (NaCl), easily dissolve when placed in a surrounding solvent (such as water). Other commonly occurring solids are structurally more robust. They result from some degree of covalent electron swapping between their constituent atoms. Such solids are often distinguished by their electrical properties; in particular, the degree to which electrons can readily travel throughout the solid. A solid in which the covalently shared electrons swap back and forth over short distances between only a few atoms are called *electrical insulators.* Glass and diamond are examples. In solids that are *electrical conductors*, on the other hand, electrons

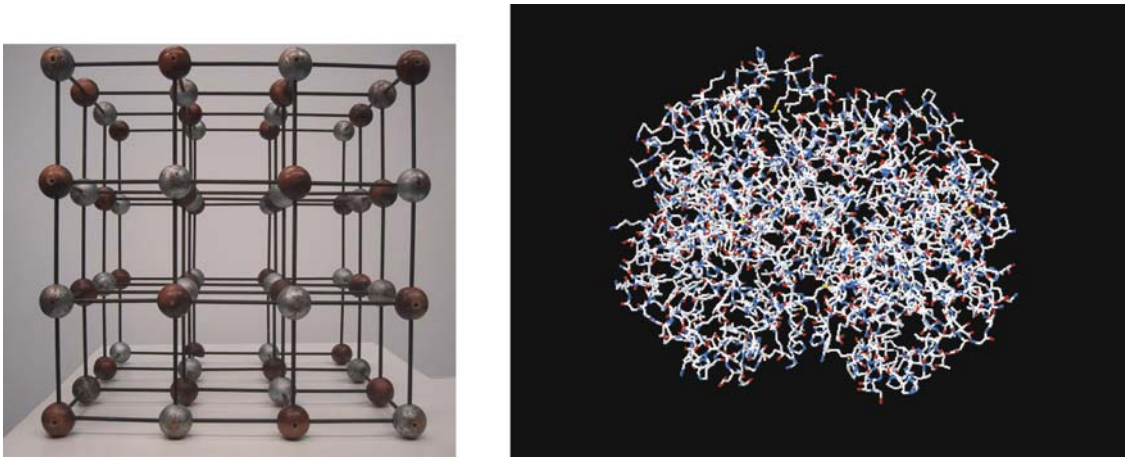


FIGURE 25.12 (left) *Cubic crystal structural model of sodium chloride;* (right) *structure of hemoglobin molecule; in a crystal of hemoglobin, an entire hemoglobin molecule occupies each of the atomic sites of such a sodium chloride crystal in place of each single atom.*

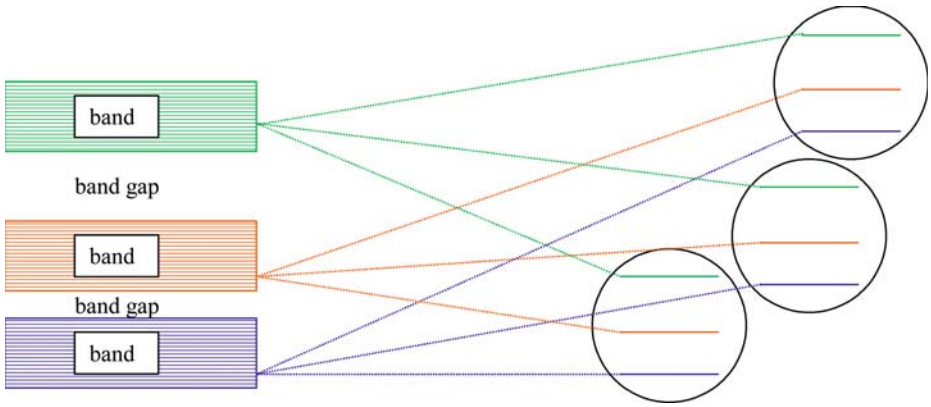


FIGURE 25.13 *Energy level diagram for a macroscopic system of molecules showing their identical discrete energy levels when well-separated schematically on the right and the band structure that forms when they interact (color-coded only for ease of viewing).*

are shared among many atoms over large distances (essentially the whole solid), and are thus relatively easy to move around. The metals aluminum, copper, silver, gold, and so forth are examples of good electrical conductors. Indeed, the particular form of covalency found in electrical conductors is often said to produce a *metallic bond* between atoms.

Let's imagine making a crystalline solid by bringing all N constituent molecules, where N is on the order of 10^{23}, together from infinitely far away where they do not interact. Because the molecules initially do not "see" each other, they will all be in the same ground energy state with the same set of quantum numbers. In this case the Pauli exclusion principle does not apply because the molecules do not interact at all. As they are brought together, once the wave functions overlap in space so that the molecules interact, then the Pauli exclusion principle dictates that no two electrons can have the same set of quantum numbers. Each individual molecular energy level is perturbed and caused to shift, removing the degeneracy, the multiple electrons in the same state, so that in place of a single energy level a huge number of distinct energy levels arise that form a more or less *continuous energy band* of different levels (Figure 25.13). In place of the discrete energy levels in individual atoms when well separated, in solids, where the molecules are closely spaced and interact, there are bands of possible energy levels with unallowed energies in gaps between the bands. If we focus on the outermost electron bands then we can understand some of the fundamental interactions possible in solids, just as the outermost electron orbital of an atom determines its chemical interactions. The electron configurations of solids explain its electrical (and thermal) properties. By examining the outermost two bands, we can distinguish three possible classes of solids: conductors, insulators, and semiconductors (see Figure 25.14).

Insulators, or *dielectrics*, are characterized by a completely filled outermost band, known as the *valence band*, and a large *band gap* of 6 eV or more with no allowed states below the next band, known as the *conduction band* (Figure 25.14). Since there is no way to add small amounts of energy to an insulator because there are no excited states for the electrons to reach (recall that room temperature electron thermal energies, $3/2(kT)$, are on the order of 0.025 eV), electrons are normally trapped in the valence band. In unusual circumstances, large amounts of energy, sufficient to cause dielectric breakdown, can be added to promote electrons to the conduction band. Thus

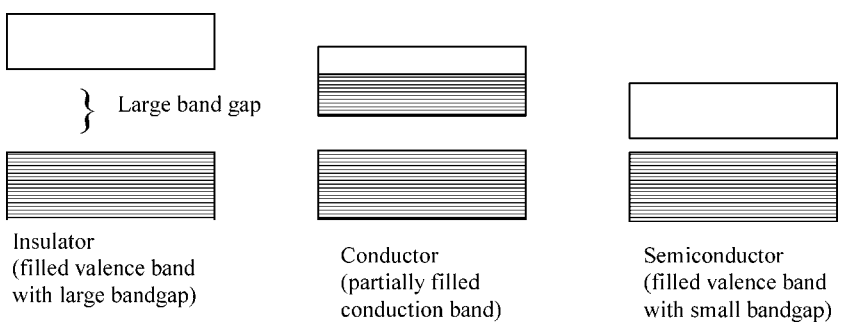


FIGURE 25.14 *The three categories of solids based on their band structure: insulators, conductors, and semiconductors.*

insulators are normally very poor conductors of electricity since the electrons are tightly bound to individual atoms, requiring a large energy to jump the band gap and "detach" from atoms to become conduction, or *free electrons.* For similar reasons these materials are also poor thermal conductors. Dielectric breakdown occurs, for example, during a lightning storm when the air, normally an excellent insulator, becomes conducting due to its ionization by huge electric fields.

Conductors are characterized by a partially filled outer, or conduction band, with lots of nearby available states to which the conduction electrons can be excited. Small additions of energy to the solid are possible with the outermost electrons in this conduction band able to accept such energies and populate near-lying empty energy levels within the band. In conductors, the valence electrons are not firmly attached to specific atoms or sites within the solid, but are termed conduction or free electrons, able to migrate about in the solid under the influence of electric forces. Although the solid as a whole is electrically neutral, each individual atom does not have a permanent complement of electrons. In metals these electrons are said to form a free electron gas because they distribute themselves uniformly within the confines of the metal's boundaries and have some of the characteristics of a gas. Ionic solutions are also an important class of good conductors of electricity, but in this case of a fluid the conducting species is the ion rather than the individual electron.

A third class of materials has an intermediate behavior between a conductor and an insulator, having a band structure with a filled valence band like an insulator but with a much smaller band gap of about 1 eV. These materials are known as *semiconductors*, with silicon and germanium being the most common, and are extremely important materials in our technological lives. Semiconductors are characterized by normally being insulators, but able to become good conductors of electricity by small controlling signals.

Two types of semiconductors can be distinguished: n-type and p-type. The *n-type semiconductors* have free electrons that are in the conduction band, just as in a conductor, although the conductivities are not as high because there are fewer such conduction electrons. In contrast, a *p-type semiconductor* has missing electrons from the valence band and these positive-like "holes" act as the charge carriers. Semiconductors used in electronic devices are universally made starting with an intrinsic semiconductor such as silicon or germanium and adding dopants chosen to enhance the n- or p-type behavior of the material. Figure 25.15 shows two-dimensional molecular pictures of n- and p-type doped semiconductors. To add extra free electrons to an n-type semiconductor, a donor impurity is added that has an extra valence electron compared to the intrinsic atoms. At the donor site, the extra electron is not needed for local bonding and is then contributed to the free electrons. To dope a p-type semiconductor, an acceptor impurity with one fewer valence electron is used. In this case the missing electron acts as an added hole. Even added at a few parts per million, these impurities greatly affect the electrical characteristics of the semiconductor.

Modern semiconductors are fabricated to allow fine control over their electrical characteristics by doping. The basis for most semiconductor chips is a p–n junction formed by butting a p- and n-type semiconductor together. At the junction the excess electrons in the n-type recombine with the holes in the p-type to form a narrow "depletion zone."

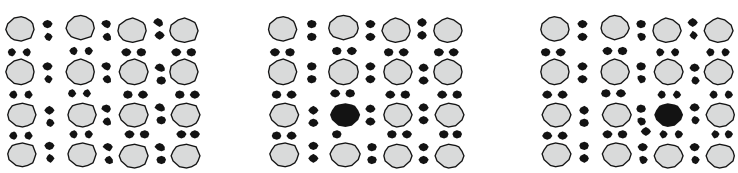


FIGURE 25.15 (left) *A pure intrinsic semiconductor, such as silicon, with each atom having four valence electrons forming a perfect crystal;* (middle) *p-type dopant molecule* (black)*, with only three valence electrons, leaves a "hole" in the crystal;* (right) *n-type dopant molecule* (black)*, with five valence electrons, contributes an extra "free" electron. Both the holes and free electrons migrate about the crystal as charge carriers contributing to its conductivity.*

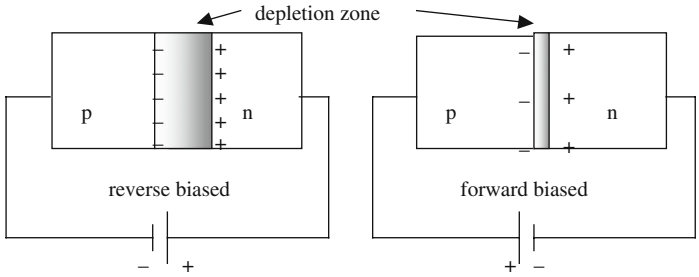


FIGURE 25.16 *A p-n junction diode either reverse or forward biased.*

This recombination reaches an equilibrium because the electrons trapped in the holes build up a layer of charge, as do the excess holes at the other edge of the depletion zone. The layers of charge at the edges of the zone act as a dynamic capacitor, able to rapidly respond to external electric fields. The simplest semiconductor device is the diode. If a voltage is applied to a diode, the relative polarity of the external voltage and diode will either increase the junction voltage (so-called reverse biased) or decrease its voltage (forward biased) as shown in Figure 25.16. In the reverse biased case, no current will flow through the (ideal) diode, whereas in the forward biased case, large amounts of current can freely flow with essentially no voltage across the diode. Thus, a diode can act as a one-way valve, allowing current to flow in only one direction. Other semiconductor devices such as transistors and operational amplifiers (op amps) are controlled devices that also may boost or amplify a signal or act as logic elements in a circuit.

Today, semiconductor "microchips" can be manufactured with specific desired properties and are being ever more miniaturized. Semiconductors are the fundamental basis of modern electronics and are found in nearly every device that plugs into an electrical outlet or runs on batteries.

Throughout this section we have focused on only the valence electrons and have ignored the inner electron core of an atom. At modest energies the inner electrons are locked in and unable to interact and so our neglect of them is justified. However, when sufficient energy is added to an atom to eject an innermost electron, creating a vacant state through, for example, a collision with an energetic external electron, then the outer electrons can make transitions to this lower vacant state, creating a spectrum of high-energy photons in the x-ray region (Figure 25.17). If the ejected electron is an $n = 1$ (K-shell) electron, then transitions from higher n levels will produce these x-ray photons, labeled K_α (from $n = 2$), K_β (from $n = 3$), and so on. Because the innermost electron shell sees the bare unshielded nucleus, the spectrum of x-ray energies is different for each element and is a signature of the atom. These x-rays are known as *characteristic x-rays* and can be used to identify the type of atom present. Specially designed x-ray tubes are used to generate beams of x-rays for x-ray diffraction or CT machines.

5. LASERS AND THEIR APPLICATIONS IN BIOLOGY AND MEDICINE

Lasers have become so widely used that, whether you know it or not, you probably use one quite often (Figure 25.18). There are hundreds of different types of lasers that have been discovered with many of them in commercial production. Nonvisible light

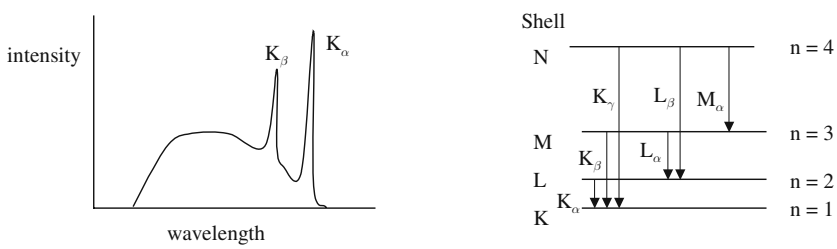


FIGURE 25.17 *Characteristic x-ray spectrum and its interpretation in terms of energy levels transitions.*

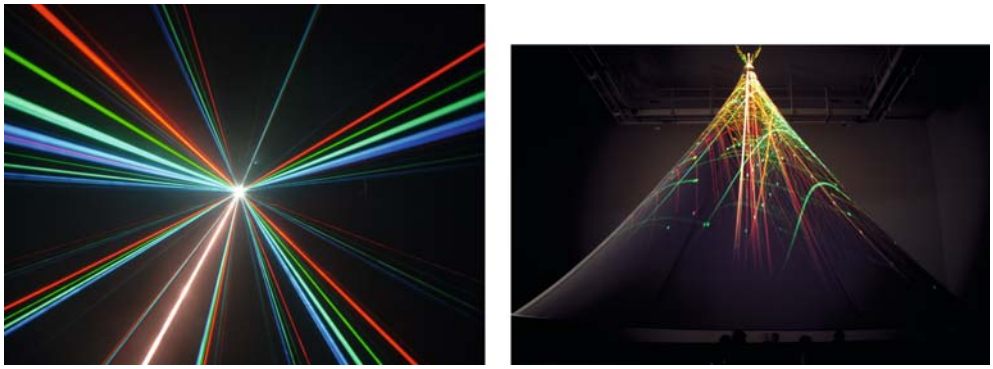


FIGURE 25.18 *Multicolored laser beams.*

lasers are routinely used in remote controls for TVs or stereo systems, in fiber-optic telephone communications, and in CD or DVD players to read the encoded information from the plastic CD/DVDs. Visible lasers put on multicolor dazzling light show displays for entertainment or scan barcodes when you check out of a supermarket. In industry, lasers are used for a large variety of purposes including cutting and processing materials, welding, microfabrication of computer "chips," and even holographic monitoring and testing of precision parts. Medical applications of lasers are ever increasing and include external and internal surgery, eye surgery, and various therapies involving tissue destruction or heating. What are the properties of lasers that allow them to be so useful in such a large array of applications?

Laser light has several quite distinctive properties. Perhaps the most notable is the very narrow frequency or wavelength range of the light. Most types of lasers have extremely pure, or *monochromatic*, light. For example, the common helium–neon (or HeNe) laser produces a red beam of light at a wavelength of 632.8 nm. The frequency of this light is given by $f = c/\lambda = 4.7 \times 10^{14}$ Hz with a bandwidth, or frequency width of the light, of only about 100 MHz. This range of frequencies corresponds to a purity level of about 1 part in 5 million or, said another way, the wavelength of the HeNe laser is 632.8 . . .where the wavelength is known to about 7 digits of precision. Special techniques can be used to stabilize the frequency of the HeNe even further. Different types of lasers emit radiation in the infrared, visible, or ultraviolet regions of the spectrum.

HeNe lasers can produce up to tens of mW (10^{-3} W) of power, representing a huge number of nearly identical photons in the beam. The energy of each 632.8 nm photon is given by $E = hc/\lambda = 3.1 \times 10^{-19}$ J, so that in a 1 mW beam there are $N = (\text{Power})/E = 3 \times 10^{15}$ photons/s, all with the same wavelength. It is impossible to obtain photons with the same degree of purity of color from any other source of light.

Although the several mW power level of the HeNe laser is very low compared to a 100 W incandescent light bulb, for example, the light bulb produces white light with a continuous spectrum of photon energies. Furthermore, light from an incandescent bulb is emitted in all directions, whereas the HeNe laser produces a narrow, roughly 1 mm diameter, beam of light that can be further focused down to a very fine pencil line of light. The range of output powers from different types of lasers is very broad, from about 1 μW = 10^{-6} W to more than 1 TW (terawatt) = 10^{12} W. The power per unit cross-sectional area, known as the power density or intensity, of laser beams can be enormous because they can be focused down to extremely fine diameters. A 1 TW laser beam focused to a spot size of 10 μm has a power density of about 1×10^{22} W/m^2. To give some idea of the magnitude of this number, consider that the average power consumption in the United States is roughly 3×10^{12} W so that a continuous TW laser beam would have roughly the equivalent power to the entire U.S. consumption rate. The flaw in this analogy is that the TW laser is a pulsed laser, with only a very brief duration, so that the total energy in the pulse is many orders of magnitude more modest, although still quite large. In general, lasers are very energy inefficient. They typically require 100–1000 times more input energy than the beam energy they produce, so that overall efficiencies are typically less

than a percent. A few lasers are exceptions to this with overall efficiencies of 20–30%; some of these are used in industry where energy efficiency is particularly important.

Lasers can also be classified according to the nature of their output as either continuous wave (CW) or pulsed. CW lasers emit a steady beam of light, whether visible or not. Pulse durations range from relatively long ms (10^{-3} s) to picosecond pulses (1 ps = 10^{-12} s), with special systems even as fast as femtoseconds (1 fs = 10^{-15} s). These extremely short duration pulses barely contain a single oscillation of the associated electromagnetic wave. Such ultrashort duration pulses have been used for the study of extremely rapid kinetic processes, such as the initial photon absorption steps in chlorophyll for photosynthesis or in rhodopsin for vision within the eye.

Another general and important property of all laser light is that it is *coherent.* The beam of laser light has a plane wavefront, one that is in phase all across the beam diameter. This property of lasers is important in such applications as holography and various types of spectroscopy, although it is unimportant in areas such as surgery or industrial cutting and processing where only the intensity and directional properties are important.

We now take up the basic physics of the laser, using a prototype model to represent a generic laser. Later we briefly discuss a few specific types of lasers. The acronym LASER stands for *Light Amplification by the Stimulated Emission of Radiation*. To understand how a laser functions, we need to first discuss the stimulated emission process.

Recall from our study of thermodynamics that at equilibrium the relative populations of two different energy levels are given by the Boltzmann factor, so that

$$\frac{N_2}{N_1} = e^{-(E_2 - E_1)/k_B T}, \tag{25.24}$$

where the N's and E's represent the populations and energies of two different levels and k_BT is the thermal energy corresponding to the temperature T. At room temperature, with a typical electronic energy level difference ($E_2 - E_1$) ~ 1 eV, we find that N_2/N_1 ~ 4×10^{-18}, so that nearly all the atoms will be in the ground state. To excite atoms, energy must be added by either heating or by collisions with electrons, other atoms, or by the absorption of photons. Figure 25.19 shows a simple two-level atomic system. On the left, the *resonant absorption* of a photon with energy equal to the transition energy is shown. Once an atom is in the upper excited state two processes can occur: *spontaneous emission* of a photon returning the atom to the ground state with no net change (shown in the center), or *stimulated emission* in which another resonant photon induces a transition to the ground state with the emission of a second coherent photon (shown on the right). Einstein first proposed the idea of stimulated emission in 1917; it was experimentally confirmed about 10 years later, but the first proposals for the invention of a laser did not come until 1957, some 40 years later. In that year Gordon Gould, then a graduate student at Columbia University, and, independently, Arthur Schawlow and Charles Townes, then at Bell Laboratories, developed the key concepts that are required for a laser.

In order to get an amplification of the number of coherent photons in our hypothetical two-level system, there must be more atoms in the excited state than in the ground state. Otherwise, because the probability that a resonant photon is absorbed is known to be the same as the probability that a second photon is emitted by stimulated emission, there will be no net increase in the number of photons (see the cartoon in Figure 25.20).

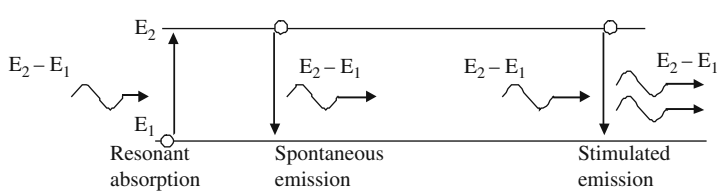


FIGURE 25.19 *The three possible types of atomic transitions between two energy levels.*

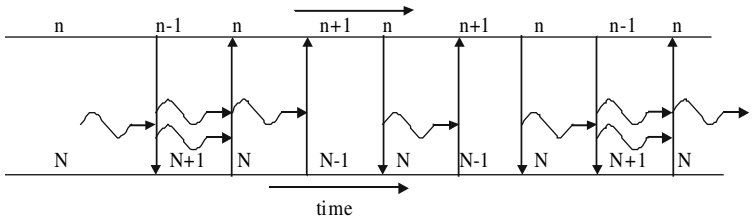


FIGURE 25.20 *Cartoon of energy level populations as a function of time showing that there can be no net amplification of photons because the probability of absorption and stimulated emission are equal.*

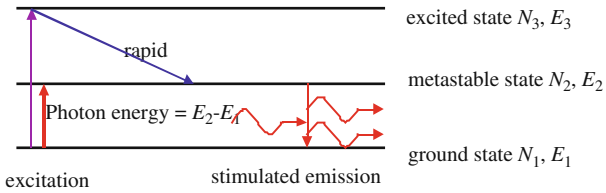


FIGURE 25.21 *A three-level laser. To have a population inversion, so that $N_2 > N_1$, more than half the atoms must be in the metastable state because N_3 is very small.*

Example 25.3 Show that at thermal equilibrium there can never be more atoms in an excited state than in the ground state or than in any lower lying energy excited state.

Solution: According to the Boltzmann distribution, Equation (25.24), which holds at thermal equilibrium the relative populations of any two states are given by Equation (25.24),

$$\frac{N_2}{N_1} = e^{-(E_2-E_1)/k_B T}.$$

If state 2 has a higher energy than state 1 then the exponent must be negative and therefore the ratio of N_2/N_1 must be less than 1, regardless of the energy difference or temperature. Thus, in a huge collection of atoms in thermal equilibrium, the relative populations of energy levels with increasing energies must necessarily decrease; that is, there must be monotonically decreasing numbers of atoms populating a set of increasing energy states. Because of this it is impossible to have any net amplification in the number of photons produced by an atomic system in thermal equilibrium. The solution to producing a net amplification comes from producing a nonequilibrium situation.

What is needed, minimally, to produce a net amplification of photons is a third atomic energy level that has a very long lifetime, known as a *metastable state.* Electrons can be excited from the ground state, ending up in the metastable state for sufficiently long times so as to provide a source of electrons for the stimulated emission of coherent photons. This populated metastable state is then said to produce a *population inversion*, when there are more electrons there (N_2) than in the lower transition state (in this case, the ground state, with N_1; see Figure 25.21). Laser photons are emitted from transitions between the metastable state and the lower energy state. Excitation of ground state electrons, known as "*pumping*," can be by a bright flash of light from a nonlaser source, or even from another laser, in a process known as optical pumping, by direct electron collisions using a high electric current flow through the laser medium, by atom–atom collisions, or by chemical reactions. Different types of lasers use different pumping mechanisms to produce the necessary population inversion.

Once stimulated photons are generated, there needs to be a mechanism for producing the fine pencil-like beam of laser light. This is accomplished by using a mirrored, or resonant, cavity (see Figure 25.22) surrounding the lasing medium, in which the photons reflect back and forth stimulating photon emission primarily along the axis of the cavity. As the photons travel along the cavity axis, their numbers are amplified continually as long as a population inversion is maintained. Those photons that travel in off-axis directions do not contribute to the lasing and simply leave the container holding the lasing medium. The laser beam actually is produced by a leakage of on-axis photons through a front mirror of the cavity designed to be less than 100% reflective. This mirror reflects

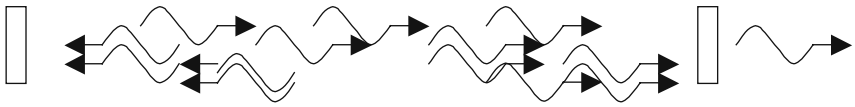


FIGURE 25.22 *The mirrored (resonant) cavity of a laser with its front mirror on the right allowing a laser beam to exit. Your imagination is needed to multiply the numbers of photons shown by a tremendous factor of 10^{15} or so.*

most of the photons, but allows a small (typically $\leq 1\%$) fraction of the photons to be transmitted and exit the laser to constitute the laser beam.

In a CW laser, the population of the metastable state must be continually replenished through continual pumping, whether through optical pumping, collisional pumping, or some other means. In a pulsed laser, after each pulse depopulates the metastable state, the lasing medium must be pumped again to create a population inversion prior to the next pulse.

In summary thus far, the essential ingredients for a laser are a metastable state of the lasing medium, a laser cavity allowing amplification and generating a beam, and a pumping mechanism for populating the metastable state. Most practical lasers involve either three or four particular energy levels of a material, with one metastable state. It is actually easier to produce a population inversion in a four-level laser because the lasing transition is from the metastable state to a lower energy state that is not the ground state (see Figure 25.23). Thus, because the population of that intermediate level is usually relatively small, it is easier to establish and maintain a population inversion ($N_3 > N_2$) than in the three-level laser (where a population inversion requires $N_2 > N_1$, with N_1 large because it is the ground state).

One of the most common lasers is the HeNe, a four-level CW laser with an energy level diagram shown in Figure 25.24. The He atoms are excited through collisions with electrons. It happens that an excited He state has an energy close to that of a metastable state of Ne and, via atomic collisions, energy can be transferred to the Ne electron. A lasing transition occurs in Ne producing photons of 632.8 nm, leading to a bright red beam. Neon also has other lasing transitions in the IR, green, and at several other colors. By using different mirrors, with high reflectivity at a selected wavelength, HeNe lasers of different colors can be manufactured, although the red HeNe is most common. These lasers are relatively small, very rugged, long-lived, trouble and maintenance free, and are relatively inexpensive, so that they are in widespread use.

Lasing materials include gases, liquids, solids, and semiconductor materials. Some of the more commonly used lasers, aside from the HeNe gas laser, include the argon and carbon dioxide gas lasers, the liquid dye lasers that can be tuned to give a laser beam with any color within some range of the particular dye, the neodymium–YAG (yttrium aluminum garnet) and titanium–sapphire solid-state lasers, as well as a host of

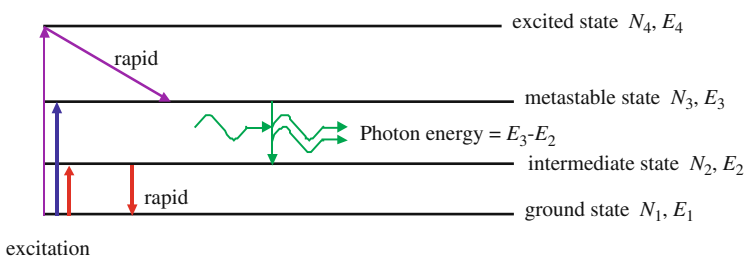


FIGURE 25.23 *A four-level laser. It is easier to reach a population inversion in this case because the metastable state must have a population $N_3 > N_2$ for lasing to occur, and N_2 is relatively small. Thus to reach a population inversion N_3 does not need to be very large and it is correspondingly less costly in energy to achieve. In a three-level laser (see Figure 25.21) the metastable state must have a population greater than the ground state, requiring more input energy.*

miniature semiconductor lasers that can be made smaller than a letter on this page (Figure 25.25).

The various interactions of laser light of different wavelengths with body tissue give rise to myriad medical applications. One fundamental interaction is simply the heating of tissue through the absorption of light. At elevated temperatures proteins will denature or coagulate just as an egg does when cooked. Laser-induced denaturation is known as *photocoagulation* and is of primary importance in laser surgery. Infrared light is very strongly absorbed by water and is therefore particularly strongly absorbed by tissue because water is its major component. Photocoagulation is used in surgery for the destruction of tumors, retinal surgery, and in many internal surgeries using fiber optics to gain access without having to open up the body. A major advantage of laser surgery is the fact that small blood vessels are cauterized, or made to clot through the photocoagulation process, so that there is a large reduction in bleeding. Furthermore, high photon powers in short pulses can deliver large doses of energy to actually vaporize tissue locally, a process known as *photovaporization.* Such high-energy doses raise the local temperature above the boiling point of water for long enough to completely vaporize the tissue, resulting in clean cuts with no bleeding and very limited damage to neighboring tissue. Usually this results in less pain, less swelling (edema), and a more rapid recovery from surgery. Laser surgery is particularly effective in areas of the body that are full of blood vessels and prone to much bleeding, such as the throat, intestines, or uterus. By regulating the intensity of the laser and/or the number of pulses, the depth of the vaporization can be controlled.

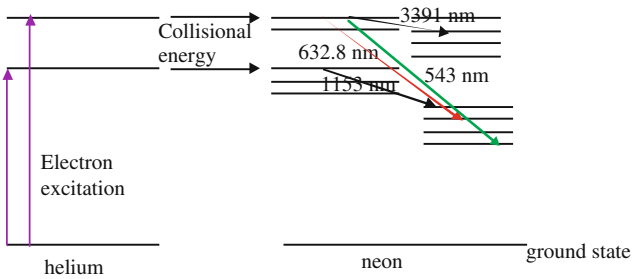


FIGURE 25.24 *Simplified energy level diagram for the HeNe laser. Four of the lasing transitions possible are indicated by the arrows and labeled wavelengths; mirrors that preferentially reflect each of these can be used to selectively produce a particular wavelength HeNe laser.*

Another major advantage of laser surgery is the ability to do microscopic internal (or external) surgery using fiber optics. Figure 25.26 shows the ultrafine surgery possible with lasers. Visible or near-infrared light can be steered using a fine fiber-optics catheter to various internal organs through either blood vessels or the gastrointestinal (GI) tract. These catheters are designed with many fibers, some of which allow imaging of the location of the fiber tip by collecting reflected light (as discussed in Section 3 of Chapter 21), whereas others are used to carry the surgical laser beam, and still others may be designed to suction off waste gases from the vaporization of the tissue. The wavelength of light used, and therefore the type of laser used, will depend on the tissue to be destroyed. Strong absorption lines are used to ensure specific destruction of that type of tissue; for example, blood rich tissue will absorb strongly at 575 nm due to a strong hemoglobin absorption line. Similarly in retinal surgery, the lens of the eye is transparent to visible light so that visible laser light can be used to surgically seal leaky capillaries behind the retina or to reattach a retina by spot-welding it to the back wall of the eye using coagulated blood. The cornea of the eye can be sculpted using a laser to ablate, or photovaporize, material in order to change its curvature and thereby its focusing ability. This type of surgery, known as LASIK, is fast becoming very common to eliminate the need for eyeglasses for certain conditions (Figure 25.27).

We cannot end a discussion of lasers without an introduction to the extremely important application of holography. Holography is photography in three dimensions and more!

FIGURE 25.25 *Diode laser module with actual diode shown as LD on the right. Note the 1 mm bar shown at the top.*

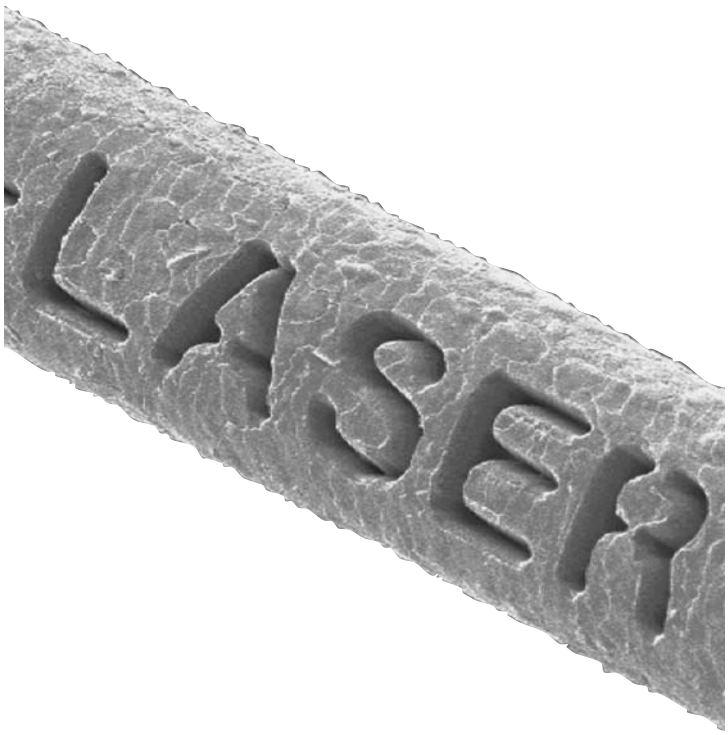


FIGURE 25.26 *A human hair sculpted with a laser beam.*

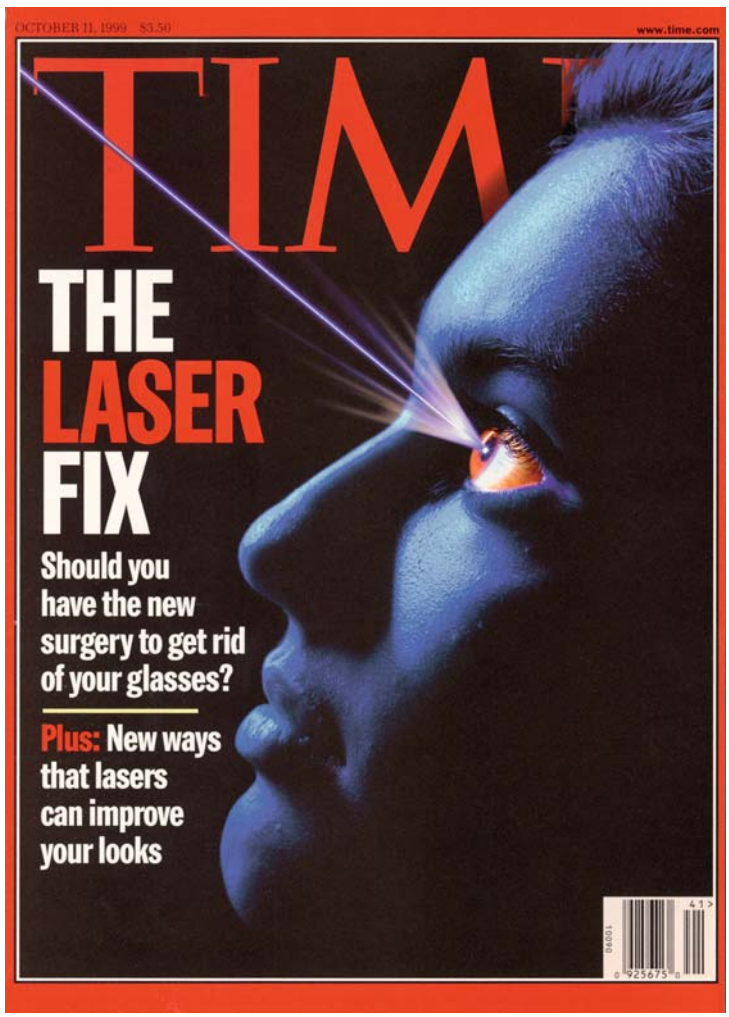


FIGURE 25.27 *Laser corrective eye surgery makes the cover of a national magazine.*

FIGURE 25.28 *A photo of the first hologram, made in 1964.*

It was first proposed and the basic theory developed by Dennis Gabor (1971 Nobel Prize for this work) in 1947 as a tool to improve the resolution of the electron microscope but did not really amount to very much until the development of the laser in the 1960s. Gabor named it from the Greek words for whole (*holos*) and message (*gramma*). In 1964 Emmett Leith and Juris Upatnieks of the University of Michigan made the first hologram (of a train; Figure 25.28), a true three-dimensional image of the original objects.

When photographic film in a standard camera records the intensity of light, all the phase information contained in the original light reaching the film is lost. The film is what is known as a "square-law detector," recording only the intensity of light reaching it, with each grain averaging that intensity over the exposure time. There is no recording of the relative phases of the different rays arriving at the film.

Holography is a method to record both the amplitude and the phase information in the light wave that reaches the film from the viewed object. There are a number of variations of the basic method, but all holography requires laser light (strictly speaking, light that is coherent over the object to be imaged) in order to construct the hologram. Some variations of holography do not require a laser to view the hologram whereas others do. The process of holography has two steps: recording the information on the film (developing the film, just as in normal photography) and then reconstructing the three-dimensional image with either laser light or with white light directed at the proper orientation.

Aside from holographic art, most of us have seen and used holography in our daily lives, probably without knowing it. The scanner used in supermarkets and stores to scan bar code labeling on packages uses holography at its core. Holographic scanners use a spinning CD-like disk to diffract a diode laser beam in a patterned path as the disk spins. The reflected light from the barcode region on the scanned package is detected and analyzed by a photodiode and the code is deciphered and used for pricing and inventory.

The CD-like disk used in a scanner is actually one type of a class of holographic optical elements (HOE). These are specifically designed holograms that can be used as lenses, beamsplitters, spectral filters only passing a narrow range of color, and so on. Many of these functions can be combined in one by multiply exposing the holographic film when being made, so that, for example, one type of HOE can act as a colored filter/lens combination, or a focused beamsplitter. They are lightweight and very thin, but can be made quite large. One example of an application that is becoming

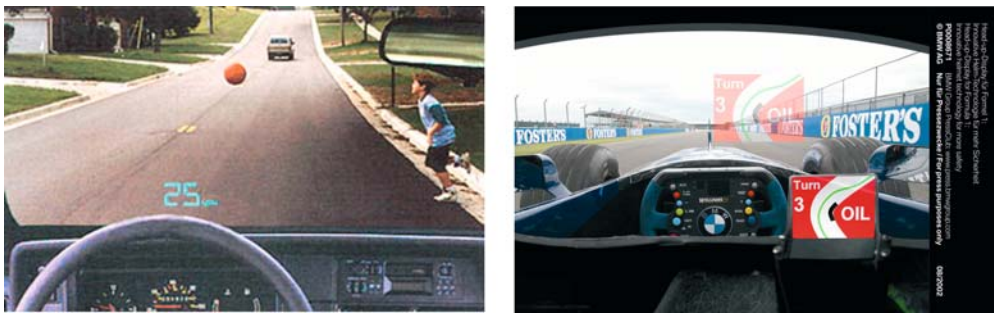


FIGURE 25.29 *Head-up displays on a car windshield are holograms.*

cheaper and more widespread is a head-up display used currently in the windshields of airplanes and high-end cars (Figure 25.29). These displays allow the pilot or driver to keep focused on the outside world, and clearly see a display of controls with a relaxed eye, rather than having to focus on the windshield itself to read the meters. This type of display is finding its way into special glasses for medical surgeons to view other information while focusing on their surgery. They work by projecting images onto an HOE in the windshield or surgeon's face-mask that acts as a narrow band reflector, transparent except for a narrow color range that is reflected.

In the future holograms will become extremely useful as information storage devices. In a high-resolution photo, digital (binary, light, or no light) information can be stored at around 10,000 bits per square mm, about the size of a pinhead. This means that each bit requires an area of about 10 μm $\times$ 10 μm on the film. In a hologram, in principle, one can store digital information with a resolution of about the wavelength of light—or about 0.5 μm—but information can be stored in three dimensions, not just two, by also storing information throughout the thickness of the film. This means that in a 1 mm $\times$ 1 mm $\times$ 1 mm cube one can store about 10^{10} bits of information, equivalent to about one volume of the *Encyclopedia Britannica*, in a 10 cm $\times$ 10 cm (about 4″ $\times$ 4″) hologram. Other advantages include the fact that surface scratches and dust on the hologram do not strongly affect reading out the information and also that the information can be read out in a parallel mode using an array of photodetectors so as to access the information faster.

But how are holograms made? In the simplest scheme for our understanding transmission holography, the first step is the recording of the interference pattern formed between the sample beam, which diffusely reflects from the object and the reference beam which is sent directly to the film (Figure 25.30-top). These two beams must be coherent and are normally obtained from a single laser by splitting its beam. Note that there is no lens used to form an image of the object from the sample beam, but this reflected, or scattered, light is directly mixed with the reference beam. The idea behind this method is to capture the full information present in the light waves of the sample beam as they arrive at the detector using the reference beam as a way to store not only the amplitude information, but also the phase information. When the film is developed (the developed film is called the hologram) and examined directly by eye, it does not look at all like a normal photograph, but is simply a complex interference pattern of light and dark bands in complex shapes and bears no direct resemblance to the original objects. In order to have the fine detail needed to later form clear detailed 3-D images, the film must be very high resolution film that allows interference fringes with spacings of less than one micron, comparable or better than the wavelength of light.

Once the film is developed, the hologram is ready for viewing. For a transmission hologram, reconstruction is done by "playing back" the reference beam with the same orientation to the developed hologram as it had when the hologram was made and viewing the image as shown in Figure 25.30-bottom. What is seen is a virtual image that appears three-dimensional. The interference pattern on the film diffracts the reference beam (known as the reconstruction beam now) to produce a light wave that duplicates the original beam that was scattered from the real object, complete with all amplitude and phase information. Furthermore, the developed hologram acts as a window glass

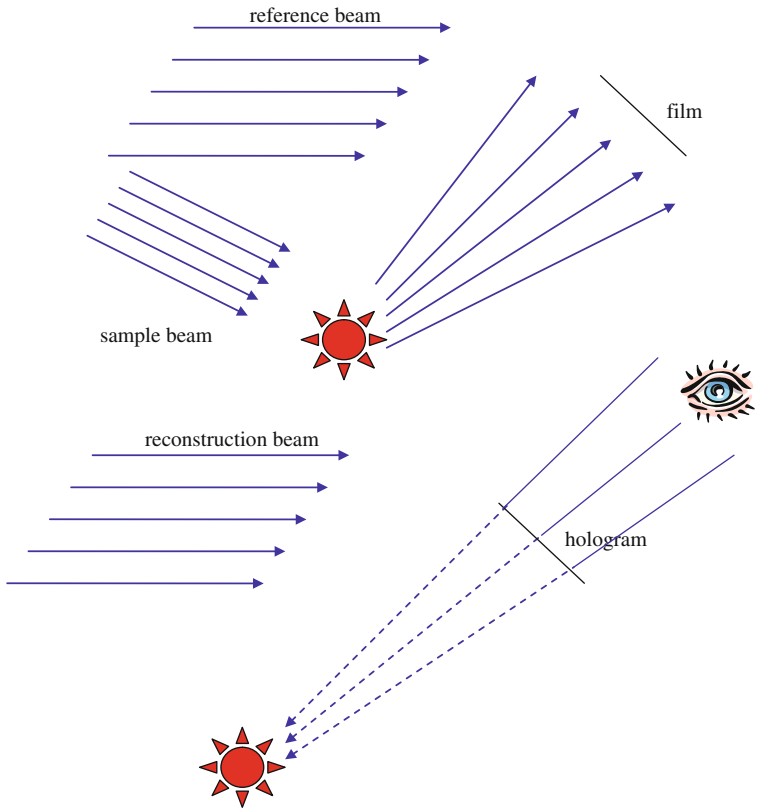


FIGURE 25.30 *(top) Recording of a transmission hologram and (bottom) reconstruction of the three-dimensional virtual image.*

so that if you move your eye around and examine different portions of the diffracted light you will see different views of the virtual image of the object, all in 3-D. So, even if you were to cut up the hologram into pieces, if you looked through one piece you would see a view of the entire object from the perspective of that position on the original intact hologram and not a view of only a portion of the object.

CHAPTER SUMMARY

The Bohr model of the hydrogen atom, or single electron ions, although not based in modern quantum mechanics, was an early successful model that predicted the correct energy levels. It is based on quantizing the orbital angular momentum of the electron according to

$$L_n = mv_n r_n = n\hbar, \qquad (25.3)$$

where n is a positive integer and $\hbar = h/2\pi$. The result gives the radii and energy levels of the different numbered orbitals as

$$r_n = n^2 r_1\,, \quad r_1 = 0.53 \times 10^{-10}\, m \qquad (25.6)$$

$$E_n = \frac{E_1}{n^2} \quad E_1 = -13.6\ \text{eV}. \qquad (25.10)$$

Transitions between different energy levels can occur by the emission or absorption of a photon with an energy corresponding to the difference in energy levels,

$$E_{\text{photon}} = hf = E_{\text{final}} - E_{\text{initial}}. \qquad (25.11)$$

Along with n, the principal quantum number, modern quantum mechanics of atoms introduces other quantum numbers: the orbital quantum number ℓ is an integer in the interval from 0 to $(n - 1)$ that specifies the orbital angular momentum

(*Continued*)

$$L = \sqrt{\ell(\ell + 1)}\,\hbar; \qquad (25.15)$$

the magnetic quantum number m_ℓ, can take on any integer value between $\pm\ \ell$ and specifies the z-component of L

$$L_z = m_\ell \hbar; \qquad (25.17)$$

The spin quantum number for an electron is $s = \frac{1}{2}$ and its z-component is specified by

$$S_z = m_s \hbar \qquad (25.20)$$

where $m_s = \pm\frac{1}{2}$. These quantum numbers, together with the Pauli exclusion principle, stating that at most a single electron can occupy any quantum state, allows the electronic configuration of multielectron atoms to be explained in the Periodic Table of the Elements.

Molecules form from neutral atoms by molecular bonds of which there are three types: covalent (strongest sharing of electrons), ionic (weaker with donor/acceptor groups), and van der Waals (weakest bond). Molecular energies can be divided into three main types (in decreasing energy): electronic, vibrational, and rotational. Electronic levels can be probed by uv-vis spectroscopy, and vibrational and rotational energy levels can be probed by IR spectroscopy. Vibrational energy levels are given by those of a spring of natural frequency f as

$$E_{vib} = \left(m + \frac{1}{2}\right)hf \qquad m = 0,1,2,\ \ldots, \qquad (25.22)$$

and rotational energy levels are given by those of an object with moment of inertia I as

$$E_{rot} = \frac{\ell(\ell + 1)\hbar^2}{2I}. \qquad \ell = 0,\ 1,\ 2,\ldots. \qquad (25.23)$$

Solids can be distinguished in terms of their electrical properties as conductors, insulators, or semiconductors. The discrete energy levels of individual atoms become energy bands in a solid. Conductors have free electrons in the conduction band whereas insulators have large band gaps between the filled valence band and the empty conduction band. Semiconductors have smaller band gaps and can be doped to have either electrons in n-type or holes in p-type semiconductors.

Lasers work by pumping energy into the lasing material so as to build up a population inversion in a metastable state. Stimulated emission then leads to coherent laser light in a resonant cavity that amplifies the beam intensity. Lasers can be CW or pulsed and can produce a pencil-like beam that can be further focused to extremely high intensities of monochromatic light. Applications include the areas of communications, including data storage and retrieval, medicine, as a probe or surgical tool, industry, and science. Holography is a method to produce three-dimensional images using interference of laser light in order to capture all the amplitude and phase information contained in light scattered from some "scene" or object.

QUESTIONS

1. Clarify for yourself what the difference between positive and negative electron energies means.
2. Discuss some of the shortcomings of Bohr's theory, such as the ad hoc nature of its assumptions, the incorrect values for electron angular momentum, and the lack of information or methods to calculate lifetimes of excited states.
3. How many different possible electronic states are there within a subshell with quantum number ℓ?
4. What are degenerate states? How could you "split" this degeneracy to determine how many states there are?
5. In the absence of an external field, what is the degeneracy of the $n = 2$ state in hydrogen?
6. "Electronic configuration" is a notation that consists of (number 1)(letter)$^{(\text{number 2})}$. Fill in the following table to specify the ground state electronic configuration for manganese (Z = 25). You may not need all of the rows, or you may add more below the last row, as needed.

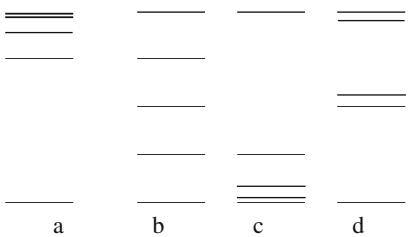
Electrons	*Number 1*	*Letter*	*Number 2*
Lowest energy			
Next lowest energy			
Next lowest energy			
Next lowest energy			
Next lowest energy			
Next lowest energy			
Next lowest energy			
Next lowest energy			
Next lowest energy			
Next lowest energy			
Next lowest energy			
Next lowest energy			
Next lowest energy			

7. Rank order the following contributions to the energy of a molecule in decreasing order: vibrational, rotational, electronic.

8. Why are good electrical conductors also good conductors of heat?
9. Is the number of free electrons in a conductor related to any net charge on the conductor?
10. What is the difference between an energy level and an energy band?
11. Discuss the similarities and differences between a free positive charge and a hole in a semiconductor. What would happen if all the holes in a p-type semiconductor were replaced by actual protons?
12. Suppose that the band gap in an insulator is 6 eV. What happens to a beam of photons with 400 nm wavelength when incident on a thin piece of this material? Will the material be transparent or opaque to this light? What happens if the wavelength is 200 nm?
13. How are x-rays generated from a solid target material?
14. What are the distinctive properties of laser light compared to other types of light such as incandescent or fluorescent?
15. What are the differences between a three-level and a four-level laser and why are four-level lasers so much more widespread?
16. The Boltzmann factor (see Equation (25.24)) predicts that at thermal equilibrium states with higher energy will have a smaller population that lower-energy states. How do you reconcile this with the notion of a population inversion?
17. Discuss some medical applications of the laser. Do some further research on one of these applications and discuss the benefits of using a laser.
18. What is the difference between photocoagulation and photovaporization? Which occurs at lower photon energy?

MULTIPLE CHOICE QUESTIONS

1. Which of the sketches to the right most closely resembles the electronic energy levels in atomic hydrogen? (a) a, (b) b, (c) c, (d) d.



Questions 2–5 concern the hydrogen spectrum viewed using a spectrometer.

2. When hydrogen light is viewed in a spectrometer at a setting of zero degrees (forward direction) it appears pink. This is due to the fact that (a) a grating produces an intensity maximum for all wavelengths of light at zero degrees, (b) a grating produces an intensity maximum at zero degrees only for pink light, (c) light from hydrogen emits a single wavelength of visible light that is pink, (d) pink light goes straight through a grating without interacting with its atoms at all.
3. When you turn the eyepiece of the spectrometer away from the head-on direction the first color of hydrogen light you see is (a) red, (b) yellow, (c) green, (d) blue.
4. The red light emitted by hydrogen results from which of the following n-state transitions? (a) $2 \rightarrow 1$, (b) $3 \rightarrow 1$, (c) $3 \rightarrow 2$, (d) $5 \rightarrow 2$.
5. For radiation emitted from excited hydrogen atoms, there is an empirical formula relating the wavelength of the radiation to some integers n_i and n_f:

$$\frac{1}{\lambda} = R_H\left(\frac{1}{n_f^2} - \frac{1}{n_i^2}\right).$$

The numerical value of the constant R_H in this formula is identical to the numerical value (a) of c, (b) of hc, (c) 13.6 eV, (d) of 13.6 eV/hc.

6. According to Bohr's theory when a hydrogen atom makes a transition from an $n = 5$ to an $n = 2$ state, the angular momentum changes by (a) $5\hbar$, (b) $2\hbar$, (c) $3\hbar$, (d) $3\hbar$.
7. For the same transition as in the previous question the average radial distance of the electron from the nucleus changes by (a) $3r_1$, (b) $25r_1$, (c) $21r_1$, (d) $5r_1$.
8. If a photon is emitted in the transition of multiple choice question 6 just above, its energy will be (a) $0.21 \cdot 13.6$ eV, (b) $21 \cdot 13.6$ eV, (c) $0.04 \cdot 13.6$ eV, (d) $25 \cdot 13.6$ eV.
9. The energy of a violet photon emitted by hydrogen is closest to which of the following values? (a) 0.1 eV, (b) 1 eV, (c) 10 eV, (d) 100 eV.
10. How many electrons could possibly be found in a multielectron atom with orbital quantum numbers $n = 3$, $\ell = 2$, and with spin quantum number "up?" (a) None, that combination is forbidden. (b) 1, (c) 2, (d) 5.
11. The ground state electronic configuration of phosphorous ($Z = 15$) is (a) $1s^2 2s^2 3s^2 4s^2 5s^2 6s^2 7s^2 8s^1$, (b) $1s^2 2s^2 2p^6 3s^2 3p^3$, (c) $1s^2 2s^2 2p^2 3s^2 3p^2 3d^2 4s^2 4p^1$, (d) $1s^{15}$.
12. An apparent anomaly occurs in the order of electron shell filling at $Z = 19$ (potassium). This anomaly is (a) 2d fills before 3s, (b) 3s fills before 4d, (c) 4s fills before 3d, (d) 5s fill before 4d.
13. To excite neon ($Z = 10$) from its ground state to its first excited state (next lowest energy to the ground state) requires transferring one of its electrons from a (a) 1s state to a 2s state, (b) 1s state to a 3s state, (c) 2p state to a 3s state, (d) 2p state to a 2d state.
14. How many electrons could possibly be found with quantum numbers $n = 3$, $\ell = 1$, $m_\ell = -1$ in some multielectron atom? (a) None, because that combination of quantum numbers is not allowed, (b) 1, (c) 2, (d) 3.

15. How many electrons could be found in the $n = 4$, $\ell = 3$ subshell of a multielectron atom? (a) 6, (b) 10, (c) 14, (d) 2.
16. The angle between the z-axis and the spin-up electron spin in the ground state of hydrogen is (a) 54.7°, (b) 30°, (c) 45°, (d) 69.3°.
17. If atoms were made of protons, neutrons, and negative pions instead of electrons, what would be the ground state pionic configuration of the equivalent of lithium with 3 pions? (a) $1s^2 2s^1$, (b) $1s^1 2s^1 2p^1$, (c) $1s^3$, (d) $1s^2 2p^1$.
18. At very low temperature we expect that the rotational and vibrational energies of a material will be (a) both zero, (b) both nonzero, (c) zero rotational and nonzero vibrational, (d) zero vibrational and nonzero rotational.
19. A material with a large ($\geq$6 eV) band gap is a (a) conductor, (b) p-type semiconductor, (c) n-type semiconductor, (d) insulator.
20. Which of the following is not necessary to produce a laser beam? (a) A material with a metastable state, (b) spontaneous emission, (c) a pumping mechanism, (d) a resonant cavity.
21. In a four-level laser, the energy of the emitted photon corresponds to the energy difference between which two states? (a) The excited state and the metastable state, (b) the metastable state and the intermediate state, (c) the intermediate state and the ground state, (d) the metastable state and the ground state.

PROBLEMS

1. Go through the details of the derivation of Equation (25.10), filling in all the details.
2. Repeat the calculation of Equation (25.10) but for a single electron atom with Z protons in the nucleus (a positively charged ion with a single electron) and show that the energy levels are given by $E_n = -(13.6\text{ eV})Z^2/n^2$.
3. Using Bohr theory and conservation of energy, if a hydrogen atom in the $n = 10$ state makes a transition to the $n = 3$ state emitting a single photon, find the energy of the photon.
4. Calculate the possible energy emission spectrum when a collection of hydrogen atoms in the ground state absorbs enough energy to populate the $n = 5$ state; that is, find all possible subsequent emitted wavelengths of light.
5. By relating the constants making up the Rydberg constant to each other through fundamental equations, show that the units work out to be m^{-1} and check the numerical value.
6. Draw a sketch showing the possible spatial orientations for an $\ell = 3$ electron and compute the allowed angles that the orbital angular momentum vector can have with the z-axis.
7. How many electrons can there be in an M shell? List their m_ℓ and m_s values.
8. If the principal quantum number of an electron in a hydrogen atom is 4, what are the possible quantum states? Label them using spectroscopic notation.
9. The short-lived Ω particle has a spin quantum number $s = 3/2$. Find all possible angles that the Ω's spin can make with respect to the z-axis. (Note that the rules for spin angular momentum are similar to those for orbital angular momentum.)
10. Ten identical spin $\frac{1}{2}$ fermions are trapped in the same infinite square well potential. Refer back to the previous chapter and calculate the ratio of the total energy of the ten fermions to the ground state energy. Be sure to apply the Pauli exclusion principle.
11. A diatomic molecule of N_2 has a moment of inertia of 1.67×10^{-46} kg m^2, an effective spring constant of 2300 N/m and an effective mass equal to half the atomic mass of N (2.32×10^{-26} kg).
    (a) Calculate the energies of the first three rotational and the first two vibrational energy levels of N_2.
    (b) Construct a simple energy level diagram showing the six energy levels up through an energy equal to the sum of $E_{\text{vib,second}} + E_{\text{rot,third}}$, labeling the levels with their quantum numbers.
    (c) In making transitions between energy levels there is no restriction on changes in the principal quantum number n. Quantum mechanics tells us, however, that transitions in vibrational and rotational quantum numbers must follow the "selection rule" that $\Delta m = \pm 1$ and $\Delta\ell = \pm 1$. Only those transitions that satisfy both of these are "allowed" to occur. (Actually others do occur, but much less frequently.) Assuming that all of the energy levels in part (a) are populated and that they are the only levels present, how many different allowed emitted photon energies are there? List the initial and final quantum numbers of the transition states.
    (d) Calculate the wavelengths of the emitted photons from the transitions in part (c).
12. A stream of pulses of laser light from a frequency-doubled Nd–YAG laser each has a wavelength of 530 nm, an average power of 10 W, a pulse duration of 10^{-9} s, and the pulses are repeated at 10 Hz.
    (a) What is the total energy delivered by the laser every second?
    (b) How many green photons are emitted every second?
    (c) If the pulses are focused down to a 100 μm diameter spot size, what is the average intensity delivered to the target?
13. Given the first four energy levels of a material to be −13 eV, −11 eV, −8 eV (metastable), and −6 eV, if the material is used as a four-level laser, what is the wavelength of the beam?
14. The National Ignition Facility, an inertial fusion reactor using the world's most powerful laser, is under construction at Lawrence Livermore Laboratory in

California. The facility will use 192 simultaneous laser pulses (3 ns duration) focused onto a small BB-sized fuel pellet from different directions to produce the necessary high density and temperature to cause fusion. On reaching the target the pulses will be in the ultraviolet (350 nm) and each will have a power of about 2.6×10^{12} W for its duration. What is the total energy delivered to the fuel pellet and how many uv photons will there be in the combined pulses?

15. A new tabletop laser (mode-locked Titanium-sapphire) is able to produce 5 fs pulses each with about 20 mJ of energy.

(a) Calculate the pulse power. Compare this to the average U.S. electric grid power consumption of about 0.5 TW (with $1\ \text{TW} = 10^{12}$ W). The power from this tabletop laser rivals that of the NOVA laser, the world's largest (in size) functioning laser.

(b) If the laser light is focused down to a 3 μm diameter spot, calculate the pulse intensity.

(c) To get an idea of the magnitude of this intensity (although it only lasts for the duration of the laser pulse), given that the total power generated by the sun is about 4×10^{26} W, the Earth–sun distance of 150×10^6 km, and the Earth mean radius of 6400 km, find the area onto which the total solar power reaching the Earth would need to be focused to get the same intensity as produced by this laser.

26

Nuclear Physics and Medical Applications

In this concluding chapter, we first summarize our knowledge of the atomic nucleus and discuss the types of nuclear radiation emitted by nuclei and how they can be detected. The rest of the chapter focuses on a variety of applications of nuclear radiation in science and medicine. We start our discussion of applications by introducing the half-life and its use in radioactive dating. Then we introduce some important ideas on dosimetry and the biological effects of radiation, as well as some ideas on nuclear medicine. Two methods (SPECT and PET) are discussed that use nuclear radiation to do imaging of the body, known as radiation tomography. The chapter concludes with the processes of nuclear fission and fusion, two topics that should be understood at a basic level by everyone in this nuclear age.

1. NUCLEAR SIZE, STRUCTURE, AND FORCES

The nucleus is an extremely small dense object in an otherwise nearly empty atom. As we've seen, atomic sizes are about 0.1 nm. The nucleus is typically several fm (10^{-15} m), or about 100,000 times smaller than the atom. To appreciate these relative sizes, imagine that we scale the size of an atom up to the size of a football field (100 yd ~ 100 m). On this scale the nucleus would have a relative size of 100 m/(100,000) = 1 mm, so that it would be like the head of a pin, not in a haystack, but on a football field. This is truly astounding because almost all of the mass of an atom is located inside the nucleus. Matter consists of dense cores in mostly empty space; the head of a pin located within an empty three-dimensional football field of space.

Remember that the nucleus contains the protons and neutrons (together known as *nucleons*) of the atom, representing nearly its entire mass, because protons and neutrons each have more than 1800 times the mass of an electron. Unlike the electron, which appears to be pointlike, having no measurable size, nucleons have a finite size of about 1 fm. Neutral atoms have equal numbers of protons and electrons, with this number known as the *atomic number* and represented by Z; the number of neutrons in a nucleus is known as the *neutron number* and represented by N. The total number of protons and neutrons in a nucleus added together is known as the *mass number* A, where

$$A = Z + N. \tag{26.1}$$

The integer mass number is approximately equal to the atomic mass. Remember that atomic masses are measured in atomic mass units (u), defined as 1/12 the mass of the carbon-12 atom, or 1 u = 1.66×10^{-27} kg.

J. Newman, *Physics of the Life Sciences,* DOI: 10.1007/978-0-387-77259-2_26,

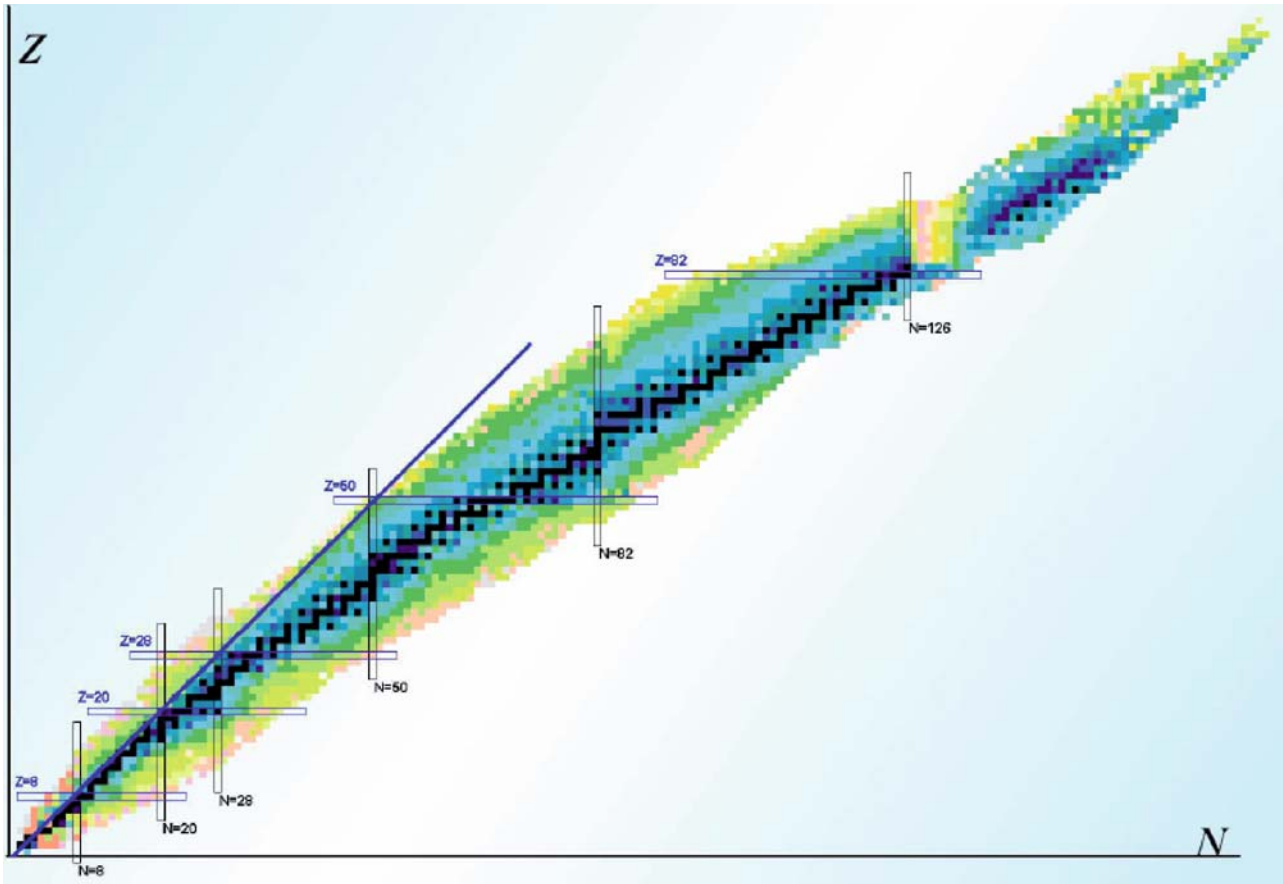


FIGURE 26.1 *Plot of the stable nuclides. The line drawn is for* Z = N *and the vertical and horizontal lines indicate the most stable nuclides (see the discussion of magic numbers below).*

A particular nuclear species is called a *nuclide*, and is represented by the chemical symbol of its neutral atom together with its value of A written as a pre-superscript. For example, ^{13}C represents the nuclide with 6 protons (because all carbon atoms have six protons), and $N\ (= A - Z) = 13 - 6 = 7$ neutrons. Sometimes the Z value will be written explicitly as $^{13}_{6}$C although this is unnecessary because Z is evident from the chemical symbol. Nuclides with the same number of protons but different numbers of neutrons are known as *isotopes*; for example, the two stable isotopes of carbon are ^{12}C and ^{13}C with 6 or 7 neutrons, respectively; other isotopes of carbon are *radioactive*, meaning that they are unstable and "decay" into other nuclides (see below). Figure 26.1 shows a plot of the known stable nuclides.

Scientists have learned about the size and shape of nuclei from high-energy scattering experiments. Electrons are accelerated to energies large enough (>200 MeV) so that their wavelengths become comparable to nuclear dimensions, and are then directed on targets of various nuclei. Recall that $\lambda = h/p$ where $p \approx E/c$ for relativistic electrons (in this case m_0 can be neglected in the expression $E^2 = p^2c^2 + {m_0}^2c^4$), so that

$$\lambda = hc/E = \frac{1.2 \times 10^{-12}\, m}{E\ (\text{in MeV})}.$$

For the energies just mentioned, the electron wavelength is below 6 fm, small enough to probe nuclear dimensions. From such experiments it is known that almost all

nuclei are nearly spherical (although many of the rare-earth element nuclei, those with $Z = 57 - 71$, are ellipsoidal) with somewhat fuzzy boundaries and effective radii R that depend on the mass number A according to

$$R = R_0 A^{1/3}, \tag{26.2}$$

with $R_0 \cong 1.2$ fm.

Because the density of the nucleus is given by the ratio of its mass (proportional to A) to its volume (proportional to R^3, and thus, according to Equation (26.2), also to A), perhaps unexpectedly we see that the density of all nuclei is the same. We can therefore calculate the nuclear density using $A = 1$, to find that $\rho = 1.67 \times 10^{-27}\text{kg}/[(4\pi/3)(1.2 \times 10^{-15}\text{ m})^3] \cong 2 \times 10^{17}$ kg/m^3. This is an extremely high density; note that the density of common materials, and thus of atoms, is only on the order of 10^3 kg/m^3, so that nuclei are 10^{14} times denser than atoms! Both the greater mass of a nucleon compared to the electron and, even more, the tiny size of the nucleus compared to atoms are responsible for this.

Our picture of the nucleus as a dense ball of nucleons that are essentially in contact with one another leads to the striking question of why the nucleus is ever stable. After all, the protons, all with the same positive charge, are extremely close together in the nucleus and their electrical repulsive force is huge. Two protons that are 2 fm apart would experience an electrical repulsive force given by

$$F = \frac{1}{4\pi\varepsilon_0}\frac{e^2}{r^2}, \tag{26.3}$$

where e is the proton charge and r is the 2 fm separation distance. This force is almost equal to 60 N (about 13 lb), a huge force that would instantly rip the nucleus apart if it were the only force acting.

In fact, the nucleus is held together by the strong nuclear force, one of two very short-range nuclear forces (the other, known as the weak nuclear force, is involved in radioactive decay). The strong force between two neighboring protons in a nucleus provides an attractive force roughly 100 times stronger than the electrical repulsion between the two. This attractive force is the same for all protons and neutrons, independent of their electric charge, so that two neighboring neutrons, protons, or a neutron and a proton all feel the same attractive force. However, the strong force rapidly vanishes at distances of even a few fm within the nucleus, and certainly outside the nucleus. A useful simple picture of the nucleus is the *liquid drop model* in which the nucleus is pictured as a tiny drop of liquid. This analogy is appropriate because both the nucleus and a liquid drop have a uniform density, are incompressible, and are held together by large forces: surface tension forces in the case of a liquid, strong forces in the nucleus. This model provides a useful way to look at the process of nuclear fission later in this chapter as analogous to a drop of liquid breaking into two smaller drops.

2. BINDING ENERGY AND NUCLEAR STABILITY

The total energy of the nucleus is the sum of its kinetic and potential energy. Because the potential energy is negative and larger, in magnitude, than the kinetic energy, the total energy of the nucleus is negative, just as we have seen it is for a neutral atom. If the nucleus were disassembled into its constituent protons and neutrons, their total energy would be more than that of the nucleus. This is just the same as the case for atoms where energy is needed to ionize an atom, for example, in hydrogen to separate the electron and proton, so that the energy of the final separated electron and proton have greater energy than that of the ground state atom. This difference is due to the *binding energy* of the atom or nucleus and, in the case

of the nucleus is a considerable amount of energy. For any nucleus of atomic and mass numbers Z and A, the (positive amount of) binding energy is given by

$$\text{Nuclear Binding Energy} = Zm_p c^2 + Nm_n c^2 - mc^2, \tag{26.4}$$

where m_p, m_n, and m are the masses of the proton, neutron, and nucleus, respectively. Because the energy equivalent of 1 atomic mass unit is $(1\text{ u})c^2 = 931.5$ MeV (found from $E = mc^2 = (1.6605 \times 10^{-27}\text{ kg})(2.9979 \times 10^8\text{ m/s})^2(1\text{ eV}/1.6022 \times 10^{-19}\text{ J}) = 9.315 \times 10^8\text{ eV} = 931.5\text{ MeV}$ (with energy conversion to eV)), we see that the nucleons each have an energy equivalent of about 930 MeV, whereas a nucleus of mass number A has an energy equivalent of about $A \times 930$ MeV. A comparable calculation for an atom shows that the atomic binding energy is only on the order of at most tens of eV.

Example 26.1 Calculate the binding energy of ^{2}H, ^{4}He, ^{197}Au, and ^{238}U. Their nuclear masses are, respectively, $m = 2.013552$ u, 4.001503 u, 196.923090 u, and 238.000180 u. Also calculate the binding energy per nucleon for each of these.

Solution: Using Equation (26.4), the Z and N values of each isotope, and the values of $m_p = 1.00727$ u and $m_n = 1.00867$ u, we find for ^{2}H, for example, that the binding energy B is $B = (1 \cdot 1.00727 + 1 \cdot 1.00867 - 2.01355) \cdot 931.5 = 2.226$ MeV. Similarly we find B values for ^{4}He of 28.30 MeV, for ^{197}Au of 1560 MeV, and for ^{238}U of 1802 MeV. On a per nucleon basis, these values are 1.113, 7.075, 7.919, and 7.571 MeV/nucleon.

The nuclear binding energy is about 8 MeV per nucleon for nearly all but the smallest nuclides. This implies that the nuclear binding energy represents about (8 MeV)/(930 MeV) $\cong$ 1% of the total nuclear energy, quite a substantial amount. If each nucleon interacted with all the others in a nucleus we should expect the binding energy per nucleon to grow in proportion to A, since each nucleon would interact with $(A-1)$ others. The binding energy per nucleon remains fairly constant, thus this implies that each nucleon only interacts with its nearest neighbors agreeing with our discussion above of the very short range of the strong nuclear force.

Figure 26.2 shows the binding energy per nucleon of some nuclides as a function of mass number. Note that the larger the binding energy, the more stable the nucleus is. We show that this figure explains the phenomena of both nuclear fission and fusion. Many large nuclei are unstable and will spontaneously fission into two smaller nuclei, each of which has a larger binding energy per nucleon and is more stable. Similarly, under the proper conditions, two protons or other very small nuclei can combine, or fuse, together to form a larger nucleus that is more stable. Both of these reactions liberate substantial amounts of kinetic energy. Fission and fusion are further discussed in the last section of this chapter.

There have been more than 2500 nuclides identified, with only a small number of these (about 280) stable. What determines whether a particular nucleus is stable or unstable? This is a complex issue. Figure 26.1 shows that at small values of N and Z stable nuclides have equal numbers of protons and neutrons, but that as these numbers increase, stable nuclides tend to have significantly more neutrons than protons. We can understand this fact as a consequence of the Pauli exclusion principle and the proton–proton electric repulsion. Recall that the exclusion principle states that interacting identical fermions, those elementary particles with half-integral spin, must have distinct quantum numbers. Protons and neutrons both have spin $^1\!/_2$ and therefore must separately satisfy this principle.

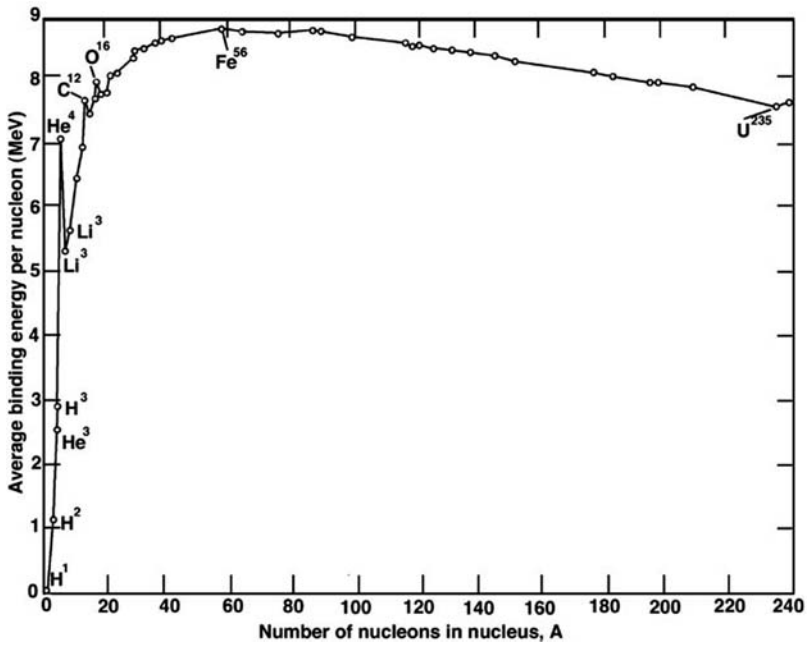


FIGURE 26.2 *The binding energy per nucleon as a function of A. Note the absolute maximum at ^{56}Fe and the minor peak at ^{4}He, as well as the average value of about 8.5 MeV/u.*

For a nucleus, just as for an atom, there are discrete energy levels at which nucleons can reside (discussed further below). If we consider a sequence of increasing Z ground state nuclides, as more protons are found in the nucleus they must occupy higher energy levels because only a spin-up and a spin-down proton can occupy the same otherwise labeled quantum state. An identical situation occurs for neutrons in a sequence of increasing N ground state nuclides. However, because protons and neutrons are different particles, they can occupy the same energy level. This implies that the energy of a nucleus with Z protons and no neutrons will be greater than the energy of a nucleus with $Z/2$ protons and $Z/2$ neutrons, because these can occupy the same set of the lower half of the energy levels (see Figure 26.3). Thus, for a given A, those nuclides with roughly equal Z and N numbers will have the lowest energies based solely on the exclusion principle, and will therefore be more stable.

As more protons are packed into the nucleus, there is a second consideration. The repulsive electrical force between the protons begins to destabilize the nucleus and so there is a tendency for more neutrons than protons to be found in larger stable nuclei. These additional neutrons do not contribute to the electrical potential energy, but they do tend to separate the protons, thus stabilizing the nucleus by reducing the Coulomb interaction energy. This effect explains why the data in Figure 26.1 fall from the $Z = N$ line at larger values. For $Z > 82$ (lead) additional neutrons do not eliminate the destabilization of the nucleus from large numbers of protons and these nuclei are all unstable.

Although nuclear energy levels are very complex and no complete theory of the nucleus yet allows their precise calculation, there are some similarities between atomic and nuclear energy levels. A third factor that affects whether a nucleus is stable has to do with its energy level structure. When we discussed atoms and the Periodic Table of the Elements, we saw that the noble gas elements in the right-hand column all have completed electron shells and are extremely inert and stable. A similar energy level shell structure exists in the nucleus and those nuclides with closed shells are particularly stable. The numbers of nucleons in such closed shells are dubbed the *magic numbers* 2, 8, 20, 28, 50, 82, 126, . . . and apply to both the numbers N and Z (see Figure 26.1). For example, the ^{4}He nuclide (also known as the alpha particle) is extremely stable because it has the magic number 2 for both N and Z. This can be

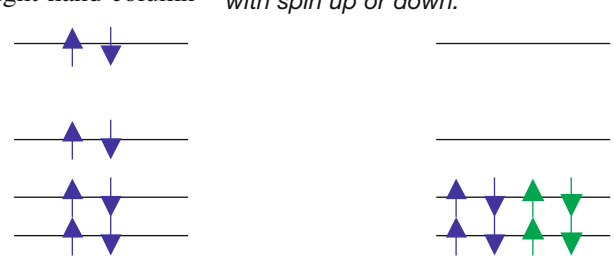


FIGURE 26.3 *Schematic energy level diagram for eight protons* (left) *or four protons and four neutrons* (right)*. The arrows represent protons* (blue) *or neutrons* (green) *with spin up or down.*

seen in Figure 26.2 where ^{4}He has an unusually high binding energy in its region of the curve.

Nuclei that are not completely stable are called radioactive and come apart at some point in time. In the next section we discuss properties of radioactivity, and later sections focus on a variety of applications of nuclear radiation.

3. TYPES OF RADIATION AND THEIR MEASUREMENT

The science of nuclear physics was born in 1896 when Becquerel, while working with photographic plates, accidentally discovered that a mineral (containing uranium) was able to expose the plate while in the dark. Shortly after this, the Curies (Marie and Pierre) isolated two new elements, named polonium and radium, and characterized the radiation they emit. Rutherford and others found that there are three distinct classes of nuclear radiation, based on their penetrating power: one type (named alpha, α, rays) can be stopped by several sheets of paper; a second more penetrating type (beta, β, rays) can be stopped by several mm thickness of aluminum; and a third most penetrating type (gamma, γ, rays) can pass through several cm of lead or through thick concrete walls. It was subsequently discovered that α rays are helium nuclei (^{4}He), that β rays are high-speed electrons, and that γ rays are high-energy photons. These radioactive particles are all emitted from radioactive nuclei.

We might begin by asking why these three types of particles and no others are the products of natural radioactivity. The primary requirement for a nucleus to be radioactive is that it must have more total energy than its products. This requirement can be written as

$$Q = (m_P - \Sigma m_i)c^2 > 0, \tag{26.5}$$

where m_P is the mass of the original nucleus, the so-called *parent*, and the summation is over the masses of all the products, m_i. In α, β, or γ decay, where in addition to the nuclear radiation, one of the products has most of the mass, it is called the *daughter*. The excess energy Q is known as the *decay energy* (or simply the Q of the reaction). Single nucleons are not emitted in nuclear decays because either $Q < 0$ for that reaction (the usual case), or, if the reaction is energetically possible, it occurs so rapidly that the parent nucleus has all decayed away and is not naturally found. An isolated unstable nucleus will decay by the emission of either α, β, or γ radiation. We also study another type of nuclear reaction in which a collision between a nucleus and another nucleon can result in the splitting of the nucleus into two fission products of roughly equal size.

Alpha decay is the spontaneous emission of an alpha particle from a nucleus. Because the alpha particle $^4_2$He consists of two neutrons and two protons, if the parent nucleus has Z protons and N neutrons, the daughter nucleus will have $Z - 2$ protons, $N - 2$ neutrons, and $A - 4$ nucleons in total. In alpha decay the original atom undergoes *transmutation*, becoming another element: ${}^A_Z E \rightarrow {}^{A-4}_{Z-2}E' + {}^4_2$He, where E is the parent and E' the daughter nucleus. The Q for this decay would be given by $Q = (M_E - M_{E'} - M_{\text{He}})c^2$ and, assuming the parent is at rest, Q essentially represents the kinetic energy of the emitted alpha particle (the daughter will necessarily recoil—in order to conserve linear momentum—but will carry off only a small fraction of Q; see Problem 7).

Often the daughter nucleus of an alpha decay itself also undergoes alpha decay. This process defines a radioactive series, whereby each subsequent alpha decay results in a daughter with 4 fewer nucleons, eventually culminating in a stable nucleus. There are four possible series for alpha decay, based on whether the starting nucleus has an A equal to $4n$, $4n + 1$, $4n + 2$, or $4n + 3$. For example, one such

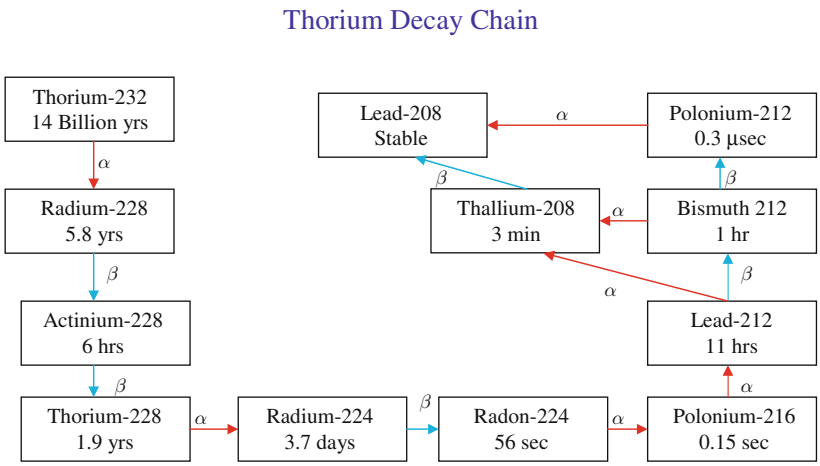


FIGURE 26.4 *The thorium alpha decay series. Note that the only naturally occurring nuclides in this series are ^{232}Th and ^{208}Pb because all of the others have relatively short half-lives. Alpha decays (α) result in a decrease of the mass number A by 4 (and Z by 2 determining the nucleus name) whereas beta decays (β) produce no change in mass number A (but a change in Z of ± 1, changing the nucleus name; see just below).*

series ($4n$) begins with thorium ^{232}Th and ends with the stable lead nuclide ^{208}Pb (see Figure 26.4). Of these four possibilities, only three appear on the Earth, because the longest living element of the $4n + 1$ series has completely decayed away to stable products.

The alpha particle has both charge ($+2e$) and is relatively massive, therefore once it is ejected from a nucleus it will interact with matter rather strongly compared to beta and gamma radiation and has relatively little penetrating power, being stopped by paper. On the other hand, because of its charge and mass, it tends to be the most ionizing type of radiation as it passes through matter. This is discussed further in Section 5.

Beta decay is the spontaneous emission of a high-energy electron (or positron, the antiparticle to the electron with the same mass and charge magnitude, but positive) from a nucleus. The electron is not an orbital electron of the atom, but comes directly from inside the nucleus where it is created just before being ejected. Examples of beta decay are the transmutation of ^{14}C according to ${}^{14}_{6}\mathrm{C} \rightarrow {}^{14}_{7}N + e^- + \nu$ and of ^{19}Ne according to ${}^{19}_{10}Ne \rightarrow {}^{19}_{9}F + e^+ + \nu$, where ν is a *neutrino*, a neutral, nearly massless particle that is very difficult to detect (there are several types of neutrinos; we make no distinctions here). Note carefully that this decay conserves charge as can be seen by adding up the charges or Z numbers on the right, where the electron e^- (positron e^+) is sometimes written as ${}^{0}_{-1}e$ (${}^{0}_{1}e$) to indicate its charge.

Because there are no electrons e^- (or positrons, e^+) to be found in the nucleus, how does this occur? The answer lies in the conversion of one type of nucleon to another within the nucleus. Either a neutron within the nucleus can spontaneously convert to a proton according to

$$n \rightarrow p + e^- + \nu, \tag{26.6}$$

or a proton can convert to a neutron according to

$$p \rightarrow n + e^+ + \nu \tag{26.7}$$

Note that both Equations (26.6) and (26.7) conserve electric charge (in Equation (26.6), both the left and right sides have 0 net charge; in Equation (26.7), both sides have +1 charge) and conserve the number of nucleons. The new nucleon will remain in the daughter nucleus; thus, when a nucleus undergoes beta decay, the mass number A does not change, but the Z will increase (e^- emission) or decrease (e^+ emission) by 1 with N correspondingly decreasing or increasing. The

ejected beta particle (e^- or e^+) and neutrino together acquire essentially the total kinetic energy Q released in the decay ($Q \cong (M_{\text{Parent}} - M_{\text{Daughter}})c^2$, because the beta particle has negligible mass), so that the electron can have any energy between essentially 0 and Q, whereas the neutrino gains the balance of Q in kinetic energy. The beta particle is identical to any electron, but is so named simply to indicate it originates in a nucleus.

When beta decay was first characterized, the variable energy of the emitted beta particle was not understood because the neutrino had not been detected. In addition to an apparent violation of conservation of energy, the laws of conservation of momentum and angular momentum appeared to be violated as well. In 1934 Enrico Fermi worked out a detailed theory of beta decay, proposing not only the existence of the neutrino, but a fourth type of fundamental force in nature known as the weak nuclear force. It was not until 1953 that direct laboratory evidence for the neutrino was obtained, but it had been accepted long before based on scientists' belief in the fundamental conservation laws. It is currently thought that neutrinos are the most ubiquitous of all particles in the universe. In 1998 the first experimental evidence was obtained for a very small, but nonzero, neutrino mass by an international team of 120 scientists working in Japan. These experiments are very difficult and still a bit controversial. If nonzero, even if extremely small, the vast numbers of neutrinos in the universe would contribute substantially toward the total mass of the universe.

Example 26.2 Calculate the Q for the following reactions: (i) the alpha decay of ^{238}U to ^{234}Th; and (ii) the β^- decay of ^{234}Th to ^{234}Pa. Use the following data: the nuclear masses of m(^{238}U) = 238.00018 u, m(^{4}He) = 4.00150 u, m(^{234}Th) = 233.99409 u, and m(^{234}Pa) = 233.99325 u, and m(β) = $(9.11 \times 10^{-31}/1.66 \times 10^{-27}) = 0.00055$ u.

Solution: (i) The alpha decay products of ^{238}U are ^{4}He + ^{234}Th. We calculate the Q for this reaction, to be $Q = [\text{m}(^{238}\text{U}) - \text{m}(^{234}\text{Th}) - \text{m}(^{4}\text{He})]c^2 = (0.00459)(931.5) = 4.28$ MeV.

(ii) In this case the reaction is $^{234}\text{Th} \rightarrow {}^{234}\text{Pa} + \beta^- + \nu$, where the protactinium (Pa) nucleus has one more proton formed in the beta decay. Here we can calculate the Q of the reaction ignoring the neutrino produced. Doing this, we find that $Q = (233.99409 - 233.99325 - 0.00055)(931.5) = 0.270$ MeV. This is the maximum kinetic energy the electron can have because otherwise the neutrino may carry off some energy as well.

Gamma decay, the third type of radioactivity, is the emission of a high-energy photon from a nucleus. The gamma ray is emitted when a nucleus makes a downward transition between two nuclear energy levels, just as a photon is emitted from an atom when it makes a downward transition between atomic energy levels. A major difference is that, because of the much larger energy spacing between nuclear energy levels, a gamma ray has a much higher energy, about a million times more than a photon from an atomic transition. This much larger energy corresponds to a much shorter wavelength for gamma rays, on the order of $10^{-12} - 10^{-15}$ m. Typically, gamma rays are emitted by daughter nuclei that are left in excited states after α or β decays as they relax back to their ground state.

Because gamma rays have no charge, they are the most penetrating of the three types of radiation. Medical imaging techniques that use radioactive isotopes require the emitted radiation to escape from the body in order to be detected. These

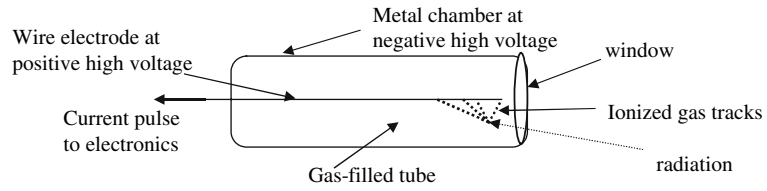


FIGURE 26.5 *Schematic of a Geiger counter used to measure the presence of radiation.*

techniques use gamma emitters because α or β rays have such short penetrating distances that they will not escape from the body. This is discussed further in Section 7 below.

We conclude this section with a discussion of the detection of nuclear radiation. There are several general methods to detect individual radiation particles as well as several methods to visualize the trajectory of these particles. One basic class of detectors is the ion collection detector, consisting of a high Z gas (typically xenon) filled chamber with a thin window through which radiation enters (Figure 26.5). Inside are two electrodes (a negative cathode and positive anode) across which a high voltage is applied. Ionizing radiation that enters the tube interacts with the gas to create ion pairs that travel to the electrodes and make up a current. If the applied voltage is high enough, the current generated is proportional to the amount of ionizing radiation. Such detectors are called *proportional counters*. At even higher applied voltages, a single ionization event will trigger an avalanche of subsequent ionizations of the gas and under these conditions the detector is called a *Geiger–Muller counter* (sometimes a Geiger tube or counter). Geiger counters are excellent for detecting small amounts of radiation because of the large degree of amplification. In general, ionization detectors have limited application in nuclear medicine because they have poor efficiency for gamma rays which are the primary information-containing decay product, as mentioned above.

A second type of radiation monitor is a *scintillation detector*, consisting of a scintillator coupled to a photomultiplier tube (Figure 26.6). The scintillator, or phosphor, is a material (typically NaI crystals, plastics, or a liquid) that emits visible light when excited by radiation. These are dense materials that are very efficiently excited by radiation, including gamma rays, and have relatively fast response times. The number of photons produced is proportional to the energy of the incident radiation and the light produced is then detected by the photomultiplier tube (see the photoelectric effect discussion in Chapter 24) whose output photocurrent can be analyzed to determine the energy of the incident radiation.

Semiconductor detectors that use p–n junctions (see Chapter 25) to detect ionization due to radioactive particles are a third type of detector. Electron–hole pairs created in the p–n junction by radioactive particles constitute an electric current proportional to the radiation energy.

A number of devices allow one to visualize the path of a single charged particle. The simplest is a *photographic emulsion* in which a chemical change along the particle's trajectory can be developed to visualize the path. Two other devices, the *cloud chamber* and *bubble chamber*, make use of either a supercooled gas (that is ready to condense on any ionized particle) or a superheated liquid (that is ready to boil along the path of an ion), respectively, to visualize the trajectory of a high-energy ion. Usually a magnetic field within the chamber causes the charged particles to travel in helical paths. (Do you remember why?) Photographs of the

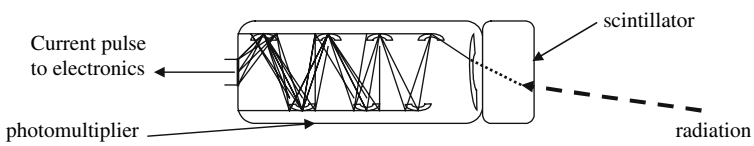


FIGURE 26.6 *A scintillation detector, converting radiation to light in the scintillator, the light then being detected by a photomultiplier and converted to an electric current signal.*

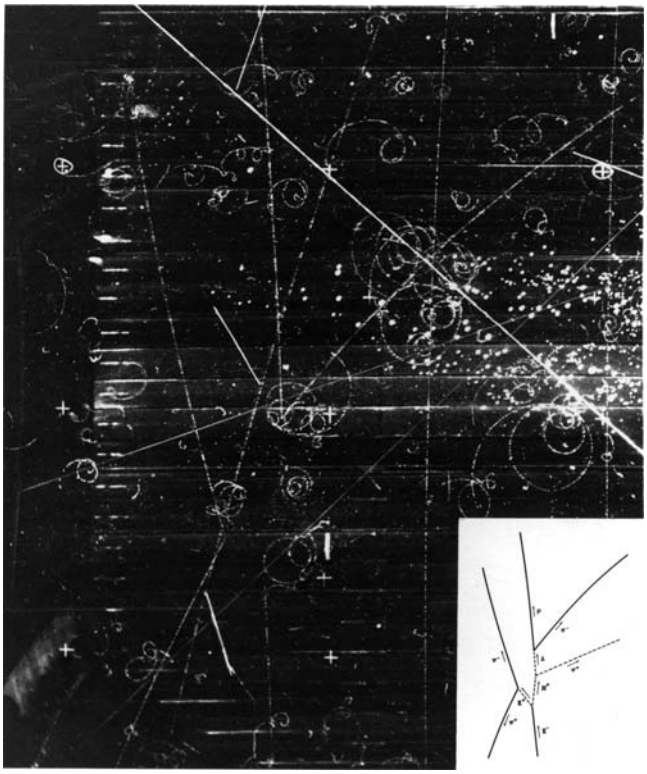


FIGURE 26.7 *Bubble chamber photo showing several interaction sites (vertices where tracks meet) and spirals indicating long-lived charged particles undergoing energy loss.*

charge track (Figure 26.7) can then be used to measure the radii of curvature to deduce the momentum and sign of the charge of the particle.

4. HALF-LIFE AND RADIOACTIVE DATING

In a macroscopic collection of radioactive nuclei, each nucleus decays independently of all the others. In fact because each nucleus is shielded by its atomic electrons, even environmental conditions of pressure, temperature, and the like do not affect radioactivity. It is impossible to predict when any particular nucleus will undergo radioactive decay. The radioactive decay process is a purely random one. We can, however, make statistical predictions about the fraction of nuclei that will decay in a given time interval based on an assumption that the probability for a decay is the same in every equal time interval up until the nucleus actually does decay. Once the parent transmutes to the daughter nuclide, that particular nucleus cannot repeat the process. Only if the daughter is itself radioactive can it decay further, but that process is described by a different probability.

This statistical notion allows us to write that the decrease ΔN in the total number of N nuclei in a sample (ΔN equal to the number of radioactive decays) in a short time interval Δt is proportional to the time interval and to the total number of nuclei in the sample. In symbols we have that

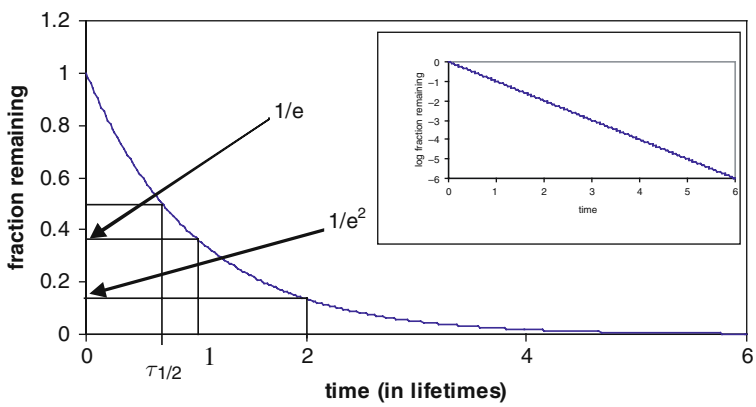
$$\Delta N = -\lambda N \Delta t, \tag{26.8}$$

where the proportionality constant λ is called the *decay constant* whose value depends on the particular radioactive nuclide. Equation (26.8) can be solved for the number of nuclei N at any time t using calculus (see box) to find

$$N(t) = N_0 e^{-\lambda t}, \tag{26.9}$$

where N_0 is the number of nuclei at time $t = 0$. Equation (26.9), plotted in Figure 26.8 normalized to the fraction remaining, is known as the *law of radioactive decay*. The time $\tau = 1/\lambda$ is known as the lifetime of the decay and represents the time for the number of parent nuclei to decay to $N_0/e = N_0/2.718$, as shown in the figure. The

FIGURE 26.8 *Radioactive decay law, normalized to the fraction remaining after some time. The half-life and one and two lifetimes are indicated on the figure. The insert shows that a semilog plot of the natural logarithm of the fraction remaining is plotted versus time, the data decrease linearly.*



number of parent nuclei decreases exponentially with time. In subsequent equal time intervals τ, the number of parent nuclei will continue to decrease by the same ratio of $1/e$ as indicated in the figure, so that after two lifetimes there will be N_0/e^2 nuclei left, after three lifetimes N_0/e^3, and so on.

More commonly the rate of decay is specified by the *half-life*, defined as the time for the number of parent nuclei to decrease by a factor of two, rather than a factor of e (see Figure 26.8). Using Equation (26.9), we can substitute $N(t) = N_0/2$,

$$N_0/2 = N_0 e^{-\lambda t_{1/2}}$$

and then solve for $t_{1/2}$ by taking the logarithm of both sides to find that

$$t_{1/2} = \frac{\log_e 2}{\lambda} = \frac{0.693}{\lambda}. \tag{26.10}$$

After one half-life there are $N_0/2$ nuclei remaining, after two half-lives there are $(N_0/2)/2 = N_0/2^2 = N_0/4$ remaining, after three half lives $(N_0/4)/2 = N_0/2^3 = N_0/8$ remaining, and so on. The half-lives of various radioactive isotopes are listed in Table 26.1. Half-lives in nature vary from vanishingly short (10^{-22} s) to nearly everlasting (10^{21} years).

Table 26.1 *Half-Lives of Some Radioactive Nuclides*

Isotope	*Symbol*	*Half-Life*	*Radioactivity*
Uranium-238	^{238}U	4.5×10^9 years	α, γ
Carbon-14	^{14}C	5730 years	β
Radium-226	^{226}Ra	1600 years	α, γ
Strontium-90	^{90}Sr	29 years	β, γ
Cobolt-60	^{60}Co	5.3 years	β, γ
Iodine-131	^{131}I	8 days	β, γ
Fluorine-18	^{18}F	1.8 h	β
Barium-141	^{141}Ba	18.3 min	β, γ
Krypton-92	^{92}Kr	1.8 s	β, γ
Polonium-214	^{214}Po	164 μs	α, γ

The rate at which radioactive nuclei decay, $\Delta N/\Delta t$, is called the *activity* and is measured in disintegrations/s, or bequerel (Bq), where 1 Bq = 1 disintegration/s. A more common unit of activity is the curie (Ci), with 1 Ci = 3.7×10^{10} Bq. The curie is a rather large unit of activity in nuclear medicine and the mCi and μCi are often used. Activity can be directly measured by detection of the decay products. Because the number of decays in a short time interval is proportional to the number N of parent nuclei (see Equation (26.8)), the activity also decays exponentially with time according to

$$\frac{\Delta N}{\Delta t} = \left(\frac{\Delta N}{\Delta t}\right)_0 e^{-\lambda t}, \tag{26.11}$$

where the subscript again indicates the zero-time value. This should make intuitive sense; if after 10 half-lives there are $1/2^{10}$ fewer radioactive nuclei, then the rate at which decays occur would also be expected to be smaller by the same factor.

Writing the Δs in Equation (26.8) as differentials, we have

$$dN = -\lambda N dt.$$

Dividing by N and integrating both sides from time 0 with N_0 nuclei to some arbitrary time t with $N(t)$ nuclei we have

$$\int_{N_0}^{N(t)} \frac{dN}{N} = -\lambda \int_0^t dt.$$

Remembering that the integral on the left is the natural logarithm of N, we have

$$\log_e(N(t)) - \log_e(N_0) = \log_e\left(\frac{N(t)}{N_0}\right) = -\lambda t.$$

Then, using the definition of the logarithm, we can rewrite this as Equation (26.9). Also note that by differentiating Equation (26.9) we can obtain Equation (26.11) for the activity,

$$\frac{dN}{dt} = -N_0 \lambda e^{-\lambda t} = \left[\frac{dN}{dt}\right]_0 e^{-\lambda t},$$

where we have used the first equation in this box in the second step.

One application of radioactivity is the dating of ancient materials. A commonly used method is ^{14}C *dating* (carbon-14 dating) of the age of once living organisms. All living plants and animals are carbon-based. There are two stable isotopes of carbon with ^{12}C representing close to 99% and ^{13}C about 1%. Carbon-14, a beta emitter with a half-life of 5730 years, is formed in the upper atmosphere by the interaction of cosmic rays with nitrogen in the air. The amount of ^{14}C is very small, roughly 1.3×10^{-12} times as much as ^{12}C, but its net amount has remained stable over many thousands of years due to the balance in its production in the atmosphere and its radioactive decay. All living material incorporates ^{14}C, ultimately by the absorption of CO_2 in the air during photosynthesis in plants. Animals incorporate ^{14}C on eating plants or other animals that have eaten plants earlier in the food chain. However, when an organism dies, no new ^{14}C is further incorporated so that the ratio of ^{14}C to ^{12}C steadily declines with age after death, by a factor of two for every 5730 years. Measurement of ^{14}C activity can thus be used to date the age of the remains of such organisms.

For objects older than about 60,000 years, carbon dating does not work because there is too little ^{14}C activity left to measure accurately. By using other isotopes with much longer half-lives, such as ^{238}U, the geological age of rock formations can be determined in much the same way. A measurement of the parent to daughter ratio can be used to date materials back billions of years. Dating the oldest rocks found, the age of the Earth has been measured to be about 4 billion years. The oldest fossils found date from about 3 billion years ago. Radioactive dating has been critical in a host of geological and evolutionary studies.

Example 26.3 Suppose that you wish to authenticate animal skin remains from one of the earliest known collections of animals, that of Shulgi, a Sumerian ruler of a territory now in Iraq, dating back to 2094 BC. You take a small sample of the skin and chemically analyze it for carbon. From a 10 g sample of carbon, what activity would you expect to measure if the sample is indeed authentic?

Solution: First we need to find the number of carbon nuclei present in the 10 g sample. We do this by assuming that essentially all the carbon is ^{12}C so that there are $(10\text{ g})(1\text{ mol}/12\text{ g})(6.02 \times 10^{23}\text{ nuclei/mol}) = 5.0 \times 10^{23}$ nuclei present. Thus, when alive, the animal would have had about $(1.3 \times 10^{-12})(5 \times 10^{23}) = 6.5 \times 10^{11}$ ^{14}C nuclei present. In that case, from Equations (26.8) and (26.10) and the fact that 5730 $y = 1.81 \times 10^{11}$ s, we know the initial activity was $\lambda N_0 = (\ln 2/\tau_{1/2})N_0 = (0.693/1.81 \times 10^{11})(6.5 \times 10^{11}) = 2.49$ Bq. If the animal skin is indeed authentic, it would be nearly 4100 years old. According to Equation (26.11) then, the expected count rate would be

$$\frac{\Delta N}{\Delta t} = (2.49\text{ Bq})(e^{-4100/5730}) = 1.22\text{ Bq}.$$

Note that this rate would need to be measured very precisely by averaging over long times to ensure a reliable value because the count rate is so low.

Our discussion of radioactive decay in this section thus far has been limited to a single radioactive species decaying away according to Equation (26.9). In a more typical situation, there are several radioactive nuclides that decay successively from one to another in a radioactive series such as the thorium decay chain discussed in the previous section. In this case the parent nuclide will decay according to Equation (26.9), but each of the other nuclides in the series will be produced by the preceding decay and so the populations of these nuclides need to be found from their production rates, shown schematically in Figure 26.9.

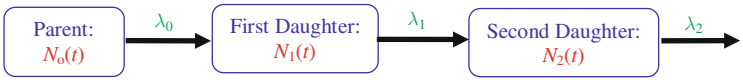


FIGURE 26.9 *Schematic for series of radioactive decays, where the N's are the populations and the λ's are the decay constants.*

Using a similar analysis to that above for Equation (26.8), the change in first daughter population will be given by

$$\Delta N_1(t) = [N_0(t)\lambda_0 - N_1(t)\lambda_1]\Delta t, \tag{26.12}$$

where the first term on the right-hand side is the rate at which the population of first daughters increases from decays of parent nuclides, and the second term on the right is the rate at which first daughters decrease from its own decay at rate λ_1. A similar equation will hold for each subsequent daughter population.

These equations can be solved for a variety of interesting cases, but the most common situation is one in which the parent decay rate is the slowest. Then over very long times, the parent population will decrease exponentially, according to Equation (26.9). But over much shorter times, the parent population N_0 will essentially remain constant and will thus supply the first daughter population at a constant rate. Now because the first daughter decay rate is much faster, its population N_1 will remain constant at a value controlled by the parent supply of first daughters. Under this condition, after a sufficient equilibrium time, all the $N(t)$ will be constant in time, so that the left-hand side of Equation (26.12) becomes equal to zero. Then we find that

$$N_0\lambda_0 = N_1\lambda_1 \tag{26.13}$$

and we can find the first daughter population to be a constant $N_1 = (\lambda_0/\lambda_1)N_0$ in terms of the constant parent population N_0. The same story will follow for the second and all other daughter populations in terms of that of the first, or previous, daughter population.

This analysis explains why it is possible to have naturally occurring very short lifetime alpha emitting nuclei, such as are found in the radioactive series discussed in the previous section. If you look back at Figure 26.4 you will see, for example, that polonium-212 decays to lead-208 by alpha emission with a half-life of 0.3 μs. Why should there be any Po^{212} left in naturally occurring ores mined on the Earth? The answer is that Po^{212} is a daughter in the radioactive series that has thorium-232, with a 14 billion year lifetime, as parent. The series of nuclei produced from Th^{232} continually produce new Po^{212} at essentially a constant rate.

5. DOSIMETRY AND BIOLOGICAL EFFECTS OF RADIATION

The interaction of nuclear radiation with matter leads to ionization; in fact, nuclear radiation (as well as uv and x-ray photons) is sometimes also referred to as ionizing radiation. Because energies of only tens of electron volts are sufficient to ionize atoms, α, β, and γ particles, with energies of MeV, are each able to ionize many thousands of atoms before losing their energy. It is this ionization that makes nuclear radiation dangerous to living organisms. Here we introduce various units to measure exposure, and discuss those doses and the relative biological effects of radiation.

A unit of *exposure*, the *roentgen* (*R*), was first introduced to define the extent of ionization produced by x-rays, but is also used for gamma radiation. Defined as the total number of ion pairs produced in a volume of 1 cm^3 of dry air under standard conditions (0°C and 1 atmosphere of pressure), one roentgen is given by $1\ R = 2.58 \times 10^{-4}$ C/kg air. This is a unit of exposure, giving the ionization level in air, but it does not give any information about absorption of radiation by living tissue or its effects on that tissue.

A measure of the *absorbed dose* of radiation, the absorbed energy per unit mass, is the *gray* (*Gy*), where 1 Gy = 1 J/kg. An older unit, still commonly used today, is the rad, where 1 rad = 0.01 Gy. For a given exposure, the absorbed dose will vary greatly depending on the absorption characteristics of the material and the type of radiation.

Furthermore, the amount of damage produced by a constant absorbed dose will also vary depending on the type of radiation. To account for these, another type of quantity is introduced, the *biological dose equivalent*, measured in *sieverts* (Sv), and given by

$$\text{biological dose equivalent (in Sv)} = \text{absorbed dose (in Gy)} \times \text{RBE}, \qquad (26.14)$$

where *RBE* is a dimensionless weighting factor, named for relative biological effectiveness, that depends on the type of radiation. Values for RBE are given in Table 26.2. These values are obtained by considering the dose of radiation needed to produce the same effect as a dose of 200 KeV x-rays. Their exact values are somewhat fuzzy, because the relative effects of radiation on biological tissue depend on the particular choice of assay. From the table, we see that β and γ rays produce similar effects to these x-rays and nucleons or α particles produce considerably more damage to biological tissue. Another unit commonly used for biological dose equivalent is the *rem*, where 1 rem = 0.01 Sv (note that the rem is commonly used when the absorbed dose is measured in rads, so that the biological dose equivalent (in rem) = absorbed dose (in rad) × RBE).

Table 26.2 *Relative Biological Effectiveness (RBE) of Different Types of Radiation*

Type of Radiation	*RBE*
200 KeV x-rays	1
γ	1
β	1
α	20
Neutrons (fast)	10
Protons	10

Our environment has many sources of natural radioactivity. We are all exposed to radioactivity from a continual shower of cosmic rays on the Earth (varying with altitude and latitude), from certain minerals found in building materials, from naturally found radon gas in the Earth that can enter and accumulate in basements, and even from radioactive elements (notably ^{14}C and ^{40}K) within our bodies. In addition, radiation is produced by many of the manmade devices in our environment, including television and cathode ray tube (CRT) computer monitors (but not liquid crystal display (LCD) monitors), luminous dial watches, as well as common dental and medical x-rays. To evaluate health risks posed by exposure to radiation, scientists have measured typical human biological dose equivalents and the U.S. government has established guidelines for maximum permissible occupational exposure. Table 26.3 shows some typical radiation doses from a variety of sources.

Table 26.3 *Typical Human Radiation Doses*

Source	*Annual Dose (Sv)*
Cosmic rays	4×10^{-4}
Cosmic rays (in high altitude airplane)	7×10^{-6} Sv/h
Radioactive ores (external exposure)	6×10^{-4}
Ingested materials (mainly potassium)	2×10^{-4}
Inhalation of radon	2×10^{-4}
Diagnostic x-rays	7×10^{-4}

These doses should be compared to average annual doses that hospital radiologists receive of about 5×10^{-3} Sv or to the maximum natural exposure to cosmic rays in mountainous areas of Brazil of about 10^{-2} Sv/year. Studies of these populations show no effects of these higher doses on mortality statistics. For comparison purposes single whole-body

radiation doses at higher levels do have significant effects at levels over 0.50 Sv. At levels up to about 2 Sv there is a significant reduction in blood platelet and white cell counts. Above this level there is severe blood damage, nausea, hair loss, hemorrhage, and short-term death in many cases. Whole-body doses between 4 and 5 Sv result in death to about 50% of such a population, and doses over 6 Sv result in nearly universal death. Long-term effects of radiation can be due to short-term high exposure or to accumulated chronic low-level exposure. Federal standards indicate an individual maximum annual exposure of 5×10^{-3} Sv, excluding medical sources. This is increased a factor of 10 for people who work with radiation sources, such as radiation technologists.

It is thought that radiation kills cells by damaging their DNA so that the cells cannot reproduce or by causing sufficient other damage to prevent the cell's normal repair mechanisms from working effectively. In medicine, radiation is often used to destroy cancer cells in a limited area of the body. Of course radiation will also kill healthy cells, particularly those that turn over rapidly, such as blood platelets and white cells or the cells lining the intestinal wall. That's why the typical symptoms of radiation sickness are GI problems due to effects on the intestinal wall, immunological suppression due to white cell kill-off, and general weakness due to red cell and platelet kill-off. By giving radiation over a period of time in repeated smaller doses, it is often possible to minimize damage to normal cells while still killing tumor cells. The chemical changes induced by radiation are caused by the formation of free radicals, enhanced by the presence of oxygen. Therefore the oxygen content of a particular tissue or cancer type will affect the success of the radiation treatment. In the following two sections we discuss nuclear medicine further, focusing on the use of radioisotopes for both therapy and diagnostics.

6. RADIOISOTOPES AND NUCLEAR MEDICINE

The key to understanding the use of radioactive isotopes (radioisotopes) in biological studies and in medicine is the fact that chemistry and radioactivity are completely independent processes. Chemistry is based on valence electron interactions and does not depend at all on nuclear properties. As an example of this, hydrogen and its isotope deuterium (an atom made from a single electron and a nucleus with one neutron in addition to a single proton) have exactly the same chemistry. The only difference in these two is their mass difference of nearly a factor of two. Because of this deuterium is often used in science experiments (in various types of spectroscopy) as an indicator of the location of hydrogen atoms because they bind in the same way chemically. Incorporation of radioactive isotopes in cells or in the body at very low doses does not directly change the normal sequence of chemical events that occurs. This fact allows *radiolabeling* (also known as tagging or tracer studies) to follow a particular type of molecule in its pathway through an organism. In this section we discuss several aspects of nuclear medicine, including the production and types of radioisotopes in use, tracer studies and detection methods in biological research, and various diagnostic tests in medicine using radioisotopes.

In order to safely use radioisotopes in medicine, not only must the dose be well controlled, but the half-life of the isotope must be relatively short so that the radioactivity is quickly reduced, causing no long-term problems. The typical dose used in diagnostic tests is so low ($\sim 10^{-8}$ Sv/h) that there is no danger from radiation. Some commonly used radioisotopes are listed in Table 26.4. Technetium(Tc)-99m is the most common of these and can be combined with many different molecules to act as a *radiopharmaceutical*. It has a half-life of only 6 h so that in order to have sufficient amounts available for hospital studies it must be freshly extracted from molybdenum-99, itself having a 67 h half-life—a useful life span of about a week—and itself usually prepared in-house in a major hospital as discussed just below. The ^{99}Mo is bound to a solid matrix in a chromatography column and as the technetium-99m forms by beta decay it is washed from the column and then can be used directly or as a radiopharmaceutical when labeling another molecule. Technetium-99m does not emit beta particles and its gamma emission is at an energy of 140 keV, a relatively low energy so that many escape the body to be detected. Furthermore, it has a very versatile chemistry and can be

incorporated into a wide range of biomolecules that can be used to target different organs or tissues in the body. These are introduced into the body by injection, ingestion, or inhalation and then imaged, as discussed in the next section.

Table 26.4 *Some Commonly Used Radioisotopes in Medicine*

Radioisotope	*Half-Life*	*Radiation*	*Applications*
Technetium-99m	*6 h	γ	Most widely used
Iodine-123	13 h	γ	SPECT brain imaging
Carbon-11	20 min	e^+	PET
Iodine-131	*8.1 days	β,γ	Thyroid disorders
Phosphorus-32	*14 days	β	Large variety of uses in biology and medicine
Thallium-201	74 h	γ	Heart imaging
Gallium-67	78 h	γ	Tumor imaging
Chromium-51	*28 days	γ	Red blood cell survival

*Produced in nuclear reactors; otherwise produced in an accelerator.

When radiopharmaceuticals are used in human diagnostic studies, there are two important characteristic times to consider. First, there is the physical half-life of the parent radioisotope, $\tau_{1/2}$, as discussed above, that is due solely to nuclear decay. A second time constant is also important in these studies, the biological half-life τ_b equal to the time for the body to wash out half of the pharmaceutical. This latter time constant is not of the same well-defined character as the radioactive half-life, but has considerable variability. These two processes occur simultaneously so that the effective decay rate in the body is given by the sum of the two different rate constants. This should make sense since both paths, physical radioactivity and elimination from the body, act to decrease radioactivity within the body and hence the effective rate constant should be their sum. The rate constant is the reciprocal of the corresponding time constants, therefore the overall effective half-life τ_e is given by a "parallel" combination of time constants (similar to the effective resistance of parallel combinations of resistors),

$$\frac{1}{\tau_e} = \frac{1}{\tau_{1/2}} + \frac{1}{\tau_b}. \tag{26.15}$$

Thus, the effective half-life is shorter than either the physical or biological half-life, just as the effective net resistance is less than either resistance in parallel. The most dangerous of environmental sources of radiation are those that are ingested and have long effective half-lives. An example is strontium-90 that can replace calcium in bones. It has a long biological half-life (45 years) as well as a long physical half-life (29 years), with a corresponding effective half-life of over 17 years.

The fact that radioisotopes used in medicine need to have short half-lives means that they must be constantly replenished for use in hospitals and other medical facilities (they really have a built-in shelf life!). Major hospitals have special supply arrangements or even in-house facilities for their production. Two methods are used to produce radioisotopes: nuclear reactors or accelerators. In nuclear reactors, either neutron beams are used to produce radioisotopes with excessive numbers of neutrons that primarily decay by beta, followed by gamma, emission, or the reactor fission products are isolated and purified. This latter method is the primary source for ^{99}Mo, the parent nucleus for technetium-99m, the most often used radioisotope. Cyclotrons (see Problem 21 in Chapter 17) and linear accelerators with proton beams are used to produce proton-rich radioisotopes. The production sources of the radioisotopes listed in Table 26.4 are indicated.

Medical research often uses radioactive tracers as an in vitro tool. When used in test tube studies, radioisotopes provide a variety of methods in cellular and subcellular work. Some of the earliest uses of tracers were to map out biochemical pathways. Radioactive tracers can be used to determine rates of metabolic processes,

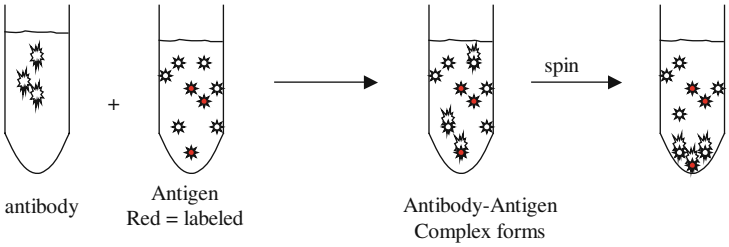


FIGURE 26.10 *Radioimmunological assay to determine the amount of antigen present. Known amounts of radiolabeled antigen and unlabeled antibody are combined and spun to separate the antibody–antigen complex from the antigen. From the ratio of counts in the pellet to that in the supernatant, the amount of antigen originally present can be found.*

predominant pathways for biosynthesis and metabolism reactions, as well as spatial localization information. These are done by various chemical testing methods combined with measuring radioactivity levels at various stages in separations.

Tracers can also be used in amounts too small for chemical testing. For example, a radioimmunological assay can determine the amount of an antigen present even in tiny amounts (~nanograms). In this technique a minute measured amount of radiolabeled antigen is added to the sample along with a measured small amount of antibody, small enough that it is all fully bound with antigen (see Figure 26.10). The antigen will bind to the antibody independent of whether it is labeled. When centrifuged, the antibody–antigen complex can be physically separated from the unbound antigen and the activity of each fraction can be determined. Therefore the ratio of labeled-to-unlabeled antigen bound to the antibody will reflect the same ratio as found in solution. Because the amount of labeled antigen added is known, the amount of antigen in the original sample can simply be computed from that ratio. There are radioimmunological tests for literally hundreds of drugs or proteins found in the blood, urine, and other bodily fluids. These are available in kits that are commonly used in clinical laboratories.

In radioassays, it is important to record as much of the radioactivity as possible. The best detector used in biological research is one in which the sample is directly immersed in the detector itself, in the technique of *liquid scintillation counting* (Figure 26.11). In this method, the sample is dissolved or suspended in a mixture of a special solvent and a fluorescent liquid, together known as a scintillation cocktail. A radioactive particle emitted from the sample will produce a brief flash of light that is then detected by a sensitive photomultiplier tube, whose output electrical current is then a measure of the radioactivity. But more than this, if two different radioisotopes are present in the cocktail they will result in different amplitude current pulses making up the output electric current. A so-called pulse-height analysis of the output current of the photomultiplier tube allows the relative amounts of the two isotopes to be determined.

7. SPECT AND PET: RADIATION TOMOGRAPHY

In this section we discuss two different imaging methods that are based on radioisotopes: single photon emission computer tomography (SPECT) and positron emission tomography (PET). Both of these methods give time-dependent three-dimensional images of the location of radioisotopes.

Earlier imaging methods, using gamma ray cameras, give two-dimensional projections of the locations of radioactive sources within the body. The gamma ray cameras are plane arrays of scintillator/photomultiplier detectors, each with a lead collimating channel to only allow radiation directed toward it to be detected. Lead shielding stops all other radiation so that the detected intensity at each photomultiplier is a measure of the net amount of radioisotope along its axis (see Figure 26.12), giving a projected image of the "object" or location of radioisotopes within the body. These images are relatively poor compared to

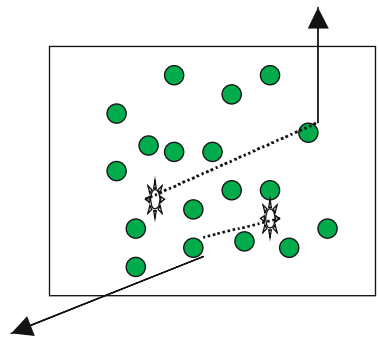


FIGURE 26.11 *Liquid scintillation counting. Radioactive decay particles produce light in a scintillation cocktail; the light is collected and detected by a photomultiplier.*

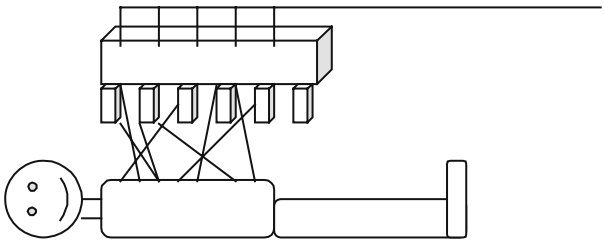


FIGURE 26.12 *A gamma ray camera for obtaining projected images of the location of radioisotopes through the body. The channels at each detector are formed by lead shielding.*

CT or MRI pictures, with resolution limited by multiple scattering of gamma rays as they leave the body and by limited detector resolution to about 1 cm at best. On the other hand, by monitoring the time dependence of the images, information on the metabolism of the radiopharmaceutical can be obtained. Examples of such uses include images of the heart, kidneys, lungs, urinary tract, and so on to determine fluid flow volumes. For imaging, the best radioisotopes are gamma emitters since these will effectively escape the body to be detected.

SPECT uses an imaging system similar to that of CT scans. Either multidetector or rotating gamma ray camera systems are used to capture a series of two-dimensional images, although each image uses a focused collection arrangement to improve resolution and contrast (or ratio of the signal-to-noise of the background radiation). Data are back-projected to reconstruct the three-dimensional image, allowing sequential slices to be imaged with a spatial resolution of about 5 mm at best, compared to the 1 mm resolution of CT scans. Although the resolution is better in CT images, they measure only x-ray absorption through the body, which then must be interpreted in terms of structure of internal organs. SPECT examines images of the distribution of radiopharmaceuticals and the time dependence of the radioactivity signal as well. Because this spatial distribution is determined by the specific binding of the drug to which the radioisotope is attached, clearly these images are directly related to function and not simply to structure.

Most major hospitals have facilities to do SPECT and it is increasingly used since the advent of better detectors and radioisotopes. Some of the organs imaged most often using SPECT include the brain, heart, circulatory system, bones, and tumors, in general. In combination with MRI and CT, this technique offers doctors an excellent tool in making diagnoses.

FIGURE 26.13 *Patient about to have a PET scan, surrounded by a ring of detectors within the housing to look for coincident 180° detections of gamma rays.*

Positron emission tomography (or PET) is an important variation on SPECT that is becoming more common as the associated costs decrease. The radiation source in this case is a positron emitter radioisotope (e.g., fluorine-18 or gallium-68) that is attached to a pharmaceutical and ingested. These positron emitters have short half lives and usually require a hospital to have an accelerator facility to prepare the radioisotopes. An emitted positron is very rapidly annihilated by an electron to form a pair of gamma rays. The energy and momentum of these gamma rays must satisfy the laws of conservation of energy and momentum. If both the electron and positron were at rest, then the total momentum must remain zero (hence the need for two identical gamma rays traveling in exactly opposite directions) and the total energy must equal the total rest energy of the electron and positron. This energy is equivalent to 511 KeV for each gamma ray. Thus the net result of each decay event is the production of a pair of 511 KeV gammas that leave the body in opposite directions. PET detectors 180° apart around the source to be imaged are set to look for the coincident arrival of 511 KeV gamma rays (Figure 26.13). These characteristic events are very clearly due to the positron emission and by projecting the accumulated data from a large number of scans at different angles, and using similar image reconstruction methods to SPECT, high-quality image slices of typically 5 mm resolution can be obtained. Spatial resolution is inherently limited by two facts: the initial kinetic energy and momentum of both the positron and electron is typically small but nonzero so that there is some variability in the 180° angle, and also the positron may travel a short (~1 mm) distance before annihilation. Both of these effects, as well as limits on detector resolution, tend to smear out the images decreasing resolution a bit (see the example image in Figure 26.14).

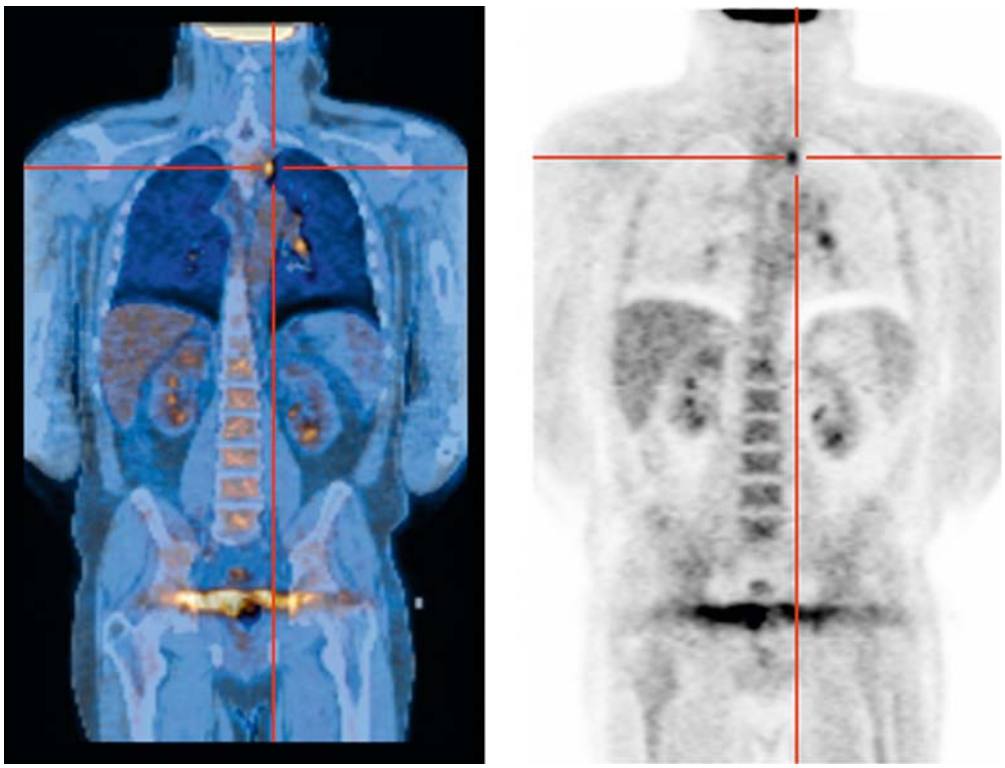


FIGURE 26.14 (left) *CT image of a patient with lung cancer that has spread to the lymph glands as clearly seen in the PET scan* (right) *of the same patient.*

PET scans of the brain, in particular, have revealed physiological correlates to a variety of disorders. Some of the most spectacular images recorded with PET have been brain scans that show brain activity in real-time. By imaging blood flow or glucose or oxygen metabolism and monitoring changes in time as the person is stimulated in various ways (e.g., visually), biochemical events can be directly correlated with brain activity (Figure 26.15). Studies comparing "normal" brains with those of people known to have various psychological disorders have begun to reveal a physiological basis for some of these problems.

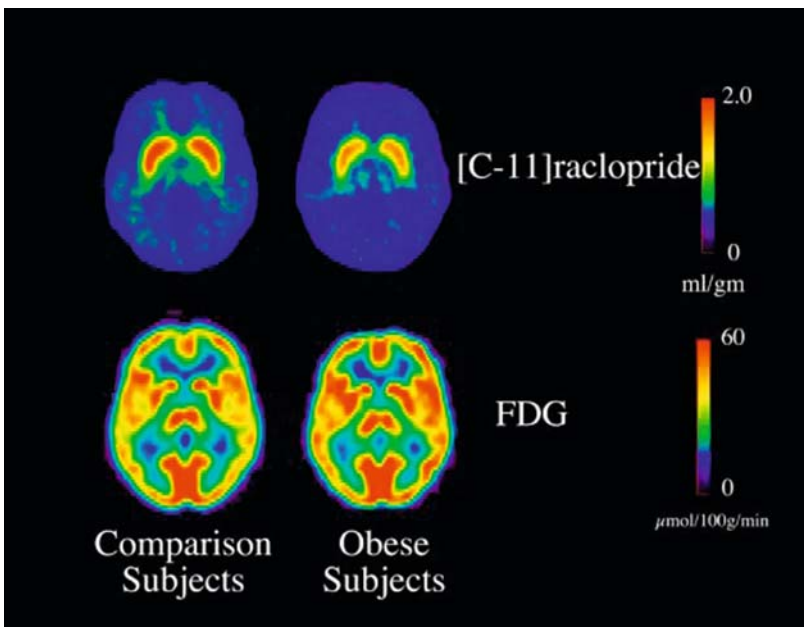


FIGURE 26.15 *PET scan of brain showing the effects of Ritalin (methylphenidate; a drug prescribed for millions of young people with attention deficit hyperactivity disorder) on the number of dopamine transporters available (*red = *more,* blue = *less). These recycle extracellular dopamine, a molecule that has been noted to give pleasure, allowing it to re-enter cells. Thus Ritalin causes an increase in extracellular dopamine levels that apparently correlates well with increased levels of attention and ability to concentrate without distraction.*

8. FISSION AND FUSION

In Section 2 we saw in Figure 26.2 that the binding energy of nuclides with A numbers near iron (56) have more binding energy, and are therefore more stable, than either very low A or very high A nuclides. In most larger nuclei, such as uranium, the long-range Coulomb repulsion of the protons is in a precarious balance with the short-range strong nuclear attractive force between adjacent nucleons. If such a nucleus is perturbed, for example, through a collision with an external nucleon, a new short-lived "excited" nucleus forms. The added energy causes the "liquid drop" nucleus to begin to elongate and once the nucleus becomes sufficiently asymmetric, the Coulomb repulsion of the two portions causes the nucleus to be unstable and decay by dividing into two roughly equal *fission products*. The difference in net binding energy between the higher-energy original nucleus and the total lower energy of the products is given off as kinetic energy of the fission products. This energy is substantial; for example, uranium has a binding energy per nucleon of about 7.6 MeV/nucleon (remember that these binding energies are actually negative, so that a smaller binding energy means a higher energy state), whereas the fission products have values of close to 8.5 MeV/nucleon. The difference of 0.9 MeV/nucleon amounts to about 100 MeV of kinetic energy for each of the two fission products.

Fission was first discovered in 1938 by Hahn and Strassmann, who bombarded uranium with a beam of neutrons and found two fission products, barium and krypton. For each starting nucleus, there are many different pairs of possible fission products, most of them radioactive. One example of a fission reaction for uranium-235 is the reaction

$$ {}^{1}_{0}\mathrm{n} + {}^{235}_{92}\mathrm{U} \rightarrow {}^{141}_{56}\mathrm{Ba} + {}^{92}_{36}\mathrm{Kr} + 3{}^{1}_{0}\mathrm{n}. \qquad (26.16) $$

The fact that there are often additional neutrons emitted, with an average of 2–3 per fission, caused scientists early on to propose that a *chain reaction* of neutron-activated fission could occur. Each fission would lead to two or three neutrons released, some of which would produce further fissions so that there would be a positive feedback and rapid growth in the energy released in fission products. By 1942 Fermi had demonstrated such a chain reaction in the first nuclear reactor.

The first use of nuclear fission was in the form of two atomic bombs dropped over Hiroshima and Nagasaki to end World War II with Japan in 1945. War had united many of Europe's finest scientists with those of the United States in a secret effort to develop the atomic bomb at Los Alamos, New Mexico. Although it is generally agreed that the use of these bombs shortened the war and reduced the total number of deaths, some of the leading scientists who worked on the development of the atomic bomb believed, in retrospect, that it was a mistake and spent much of their subsequent efforts in attempts to bring about nuclear disarmament.

Enrico Fermi's first nuclear reactor had as its main initial function the production of plutonium to be used in two atomic bombs. Today there are about 450 nuclear power reactors used to generate electricity in about 30 countries around the world. Although there are several different designs of these reactors, they all basically use nuclear energy to generate heat, producing steam then used to drive turbines, thereby generating electricity.

There are several key problems to producing controlled nuclear fission in a nuclear reactor. The predominant uranium-238 isotope (representing over 99% of naturally occurring U) is relatively stable against fission, whereas uranium-235 (only about 0.7% abundance) undergoes fission very efficiently when slow neutrons are absorbed. Sometimes uranium ore is processed to enrich the ^{235}U component to a few percent to provide a "richer" fuel. A minimum amount of fuel, the *critical mass*, typically on the order of kg, is needed to have a self-sustaining nuclear reaction. A second problem is that of the two or three neutrons produced in a single fission, only one is needed to sustain a controlled reaction. If more than one neutron from each fission leads to additional fissions, the reaction will "run away," as in a nuclear bomb, whereas if this number is less than 1.0, the reaction will eventually die out. Only by maintaining this number very near to 1.0, by the escape or absorption of excess neutrons in a special device known as a *control rod*, can the reaction be kept at a steady rate. Control rods are made from materials that very effectively absorb

neutrons without undergoing fission. Neutron absorption in ^{235}U leading to fission is most effective for slow thermal neutrons, those that have lost energy often making numerous collisions in a special purpose material known as a *moderator*. Moderators are designed to effectively slow neutrons. Water is commonly used as a moderator in nuclear reactors, with heavy (deuterated) water sometimes used because it absorbs fewer neutrons eliminating the need to enrich the uranium.

Despite huge investments in safety features, there have been two significant accidents at nuclear power plants: one at Three Mile Island, in Pennsylvania in 1979 which was contained, and one at Chernobyl in Ukraine in 1986 where 31 people were initially killed, most from radiation. The Chernobyl accident released about 3–4% of its radioactive material resulting in about 130,000 people receiving significant radiation doses leading to a sharp increase in thyroid cancer among children in that region, with other long-term health effects still unclear. Apart from safety issues of nuclear power plants, there are also literally tons of highly radioactive waste products produced in these plants that need to be safely and securely isolated from our environment for thousands of years. Because of these safety and environmental concerns, alternative sources of electricity other than nuclear fission power are needed. Along with solar, wind, hydroelectric, and other "green" sources of power, a possible long-term solution involves a second type of nuclear reaction.

According to Figure 26.2, two very low mass number nuclides with a small binding energy per nucleon can fuse together to produce a larger nuclide with a much greater binding energy per nucleon, thus releasing a large amount of energy. This process, known as *nuclear fusion*, releases much more energy per nucleon than fission, as can be seen from the steep initial slope in the binding energy per nucleon curve in Figure 26.2. In other words, the magnitude of the energy of the larger fused nucleus is much less than the sum of the energy of the lighter starting nuclei and the difference is liberated in the fusion reaction. For example, in the fusion of deuterium and tritium, two isotopes of hydrogen, an alpha particle and a neutron, form according to

$$^{2}_{1}\mathrm{H} + {}^{3}_{1}\mathrm{H} \rightarrow {}^{4}_{2}\mathrm{He} + {}^{1}_{0}\mathrm{n}. \tag{26.17}$$

Calculating the net difference between the initial and final energies (using the masses of each and the equivalence of mass and energy; see the example just below) gives a net energy release of about 17 MeV for each fusion. Because there are only 5 nucleons involved in this reaction, the energy per nucleon is 3.4 MeV/nucleon, much larger than the 0.9 MeV/nucleon released in fission. On an energy per unit mass basis, fusion is a much more productive process than fission.

Nuclear fusion occurs naturally in stars, including our sun, at extremely high temperatures. These *thermonuclear reactions* in stars are believed to have been responsible for generating all of the larger mass nuclei in the universe starting from hydrogen. We believe that very early in the history of the universe the temperature was too hot for atoms to be stable. As the universe expanded and cooled, hydrogen atoms formed and then condensed locally under gravity to form stars. As stars became more compact due to the force of gravity, the interior temperatures and pressures increased, providing an environment in which nuclear fusion could occur. Stellar fusion first uses hydrogen as a fuel, but as hydrogen is depleted fusion of other light nuclei also occurs. Thus, all the other elements found on Earth and throughout the universe originated in such stellar fusion reactions; *we ourselves are therefore made of stellar material*.

One fusion reaction is the so-called proton–proton cycle:

$$\begin{aligned}
&{}^{1}_{1}\mathrm{H} + {}^{1}_{1}\mathrm{H} \rightarrow {}^{2}_{1}\mathrm{H} + \mathrm{e}^{+} + \text{neutrino} \\
&{}^{1}_{1}\mathrm{H} + {}^{2}_{1}\mathrm{H} \rightarrow {}^{3}_{2}\mathrm{H} + \text{photon} \qquad (26.18) \\
&{}^{3}_{2}\mathrm{H} + {}^{3}_{2}\mathrm{H} \rightarrow {}^{4}_{2}\mathrm{He} + {}^{1}_{1}\mathrm{H} + {}^{1}_{1}\mathrm{H}.
\end{aligned}$$

Net reaction: $4\,{}^{1}_{1}\mathrm{H} \rightarrow {}^{4}_{2}\mathrm{He} + 2\mathrm{e}^{+} + 2\nu + 2\gamma$

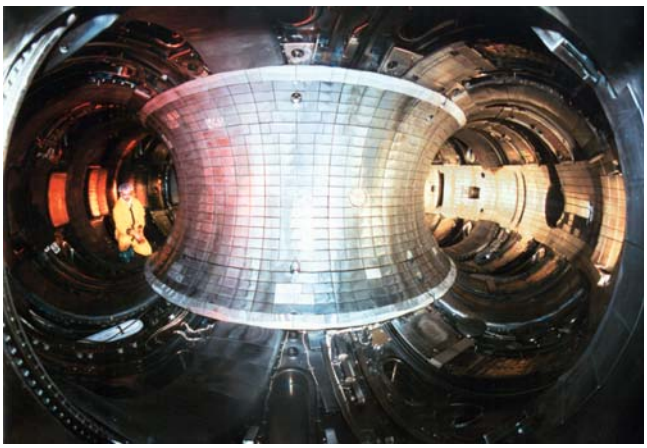


FIGURE 26.16 *Inside the Princeton Tokamak (note the man on the* left *to judge the scale).*

The overall result of these reactions is that four protons have fused to produce one alpha particle plus two each of positrons, neutrinos, and photons, with a net release of 24.7 MeV. The positrons quickly annihilate with electrons to form four additional photons, each with 0.51 MeV, so that the total energy released in the proton–proton cycle is $(24.7 + 4 \times 0.51) = 26.7$ MeV per helium nucleus formed. In order for this reaction sequence to occur, protons must be brought very close together at very high temperature to overcome their mutual electrostatic repulsion and fuse together. Central cores of stars, including our sun, have temperatures and pressures high enough for fusion to occur.

To produce fusion on the Earth, where the pressure is much lower than in the core of a star, even hotter temperatures are required. The first fusion reactions produced were those of hydrogen bombs in which an atomic (fission) bomb was detonated to produce the sufficiently hot temperature necessary to initiate fusion in a deuterium and tritium pellet. Different schemes to produce controlled conditions for nuclear fusion have been tried, each attempting to heat a deuterium–tritium fuel pellet to temperatures of 10^8 K, by either extreme electric currents or particle or laser beams, forming a plasma (ionized gas) confined in space for long enough so that fusion can take place. In one scheme, magnetic confinement, the plasma is trapped by the presence of a very strong magnetic field that exerts magnetic forces on the moving ions traveling around within a toroidal (doughnut) shaped solenoid. Figure 26.16 shows the Princeton Tokamak Fusion Test Reactor for magnetic confinement. A second alternative scheme, inertial confinement, uses many high-powered laser pulses that simultaneously strike a deuterium–tritium fuel pellet from different directions. The beams produce high temperature and pressure so rapidly that the inertia of the fuel does not allow it to escape and fusion occurs. Figure 26.17 shows the target chamber of the NOVA Laser Facility at Lawrence Livermore Laboratory, a facility currently being replaced by an even larger one at the National Ignition Facility (NIF). Short controlled pulses of energy from fusion have been produced by both of these schemes, but much work needs to be done before these become viable commercial sources of energy.

Fusion offers a number of advantages over the current fission nuclear power plants. Fuel for fusion is much more abundant, cheaper, and yields more energy on a per mass basis. The oceans are a vast supply of deuterium fuel. Furthermore, unlike fission, there are no radioactive byproducts, so that there are no long-term storage problems with radioactive waste. There is also the fact that, unlike fission reactions

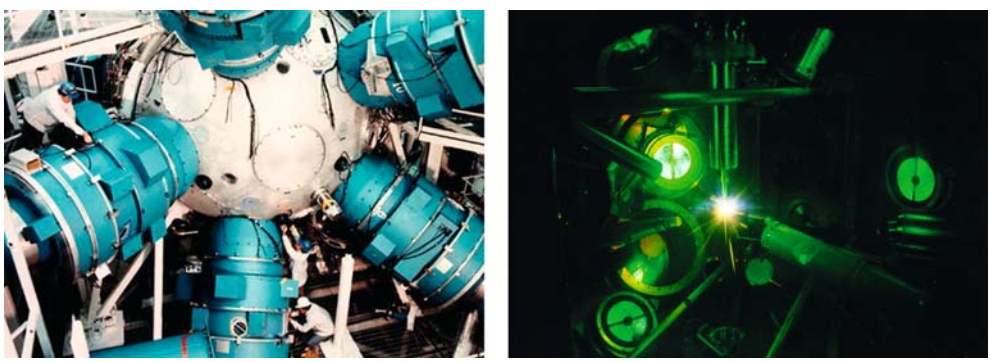


FIGURE 26.17 (left) *The NOVA laser showing some of the arms through which the laser power is focused on the fuel pellet at the center.* (right) *View of the artificial ministar created by inertial confinement fusion in the NOVA.*

in which chain reactions can become uncontrolled if there is a malfunction of control rods producing a melt-down as has happened at Chernobyl and Three Mile Island, failures in fusion reactors would lead to a shut-down of the fusion reactions themselves and no possibility of an out-of-control chain reaction. For these reasons if commercially produced in a reactor, energy from fusion might be the ultimate cure to the world's energy problem.

CHAPTER SUMMARY

The atomic nucleus contains Z protons and N neutrons with a total number of nucleons $A = Z + N$. Nuclei are very small in size, having radii given by

$$R = R_0 A^{1/3}, \tag{26.2}$$

with $R_0 = 1.2$ fm. A nucleus of mass m has a nuclear binding energy given by

$$\text{Nuclear Binding Energy} = Zm_p c^2 + Nm_n c^2 - mc^2, \tag{26.4}$$

and is typically about 8 MeV per nucleon in all but the smallest nuclei.

Three types of nuclear radiation exist known as alpha, beta, and gamma radiation. Alpha radiation is the emission of helium-4 nuclei (2 protons + 2 neutrons) from nuclei through a process of tunneling. Beta emission comes from the production of electrons or positrons within the nucleus because of neutron or proton decay, respectively, and these "beta particles" are emitted from the nucleus along with neutrinos at high energy. Gamma emission comes from transitions from excited nuclear states giving rise to high energy photons. Each of these types of radioactivity are characterized by their Q, or decay energy,

$$Q = (m_P - \Sigma m_i)c^2 > 0, \tag{26.5}$$

where P stands for parent nucleus and the sum is over all the products.

Radioactive decay is governed by an exponential decay of the numbers of radioactive nuclei N,

$$N(t) = N_0 \, e^{-\lambda t}, \tag{26.9}$$

where N_0 is the number of such nuclei at time zero and λ is the decay rate for the process. The half-life of the reaction is the time for 1/2 of the nuclei to decay and is related to the decay rate by

$$t_{1/2} = \frac{\log_e 2}{\lambda} = \frac{0.693}{\lambda}. \tag{26.10}$$

Measures of exposure to radioactivity include: the Roentgen (R) which is simply a measure of the number of decays per unit volume; the Gray (Gy; 1 Gy = 1 J/kg) or rad (1 rad = 0.01 Gy) which are measures of the absorbed dose of radiation; or the biological dose equivalent, measured in sieverts (Sv), or in rem (1 rem = 0.01 Sv),

$$\text{biological dose equivalent (in Sv)} = \text{absorbed dose (in Gy)} \times \text{RBE}, \tag{26.14}$$

where RBE (relative biological effectiveness) is a dimensionless weighting factor describing the effectiveness of different radiation to be absorbed by the body.

Nuclear medicine involves the use of short-lived radioactive tracers (radiolabeling) to follow the path of a particular molecule through the body either by in vivo or in vitro studies. Two imaging methods that use radiotracers are SPECT (single photon emission computer tomography) and PET (positron emission tomography).

Because the binding energy per nucleon for large nuclei is less in magnitude (~0.9 MeV) than for intermediate-sized nuclei, large nuclei can undergo fission releasing the excess energy of ~100 MeV for each of the products, along with several neutrons. Under controlled conditions this energy can be harnessed in nuclear power plants, whereas if left uncontrolled, this serves as the basis for a fission bomb. Fusion is the naturally occurring process in stars whereby hydrogen nuclei, or other small nuclei, are compressed and heated until they fuse to form larger nuclei, releasing large amounts of binding energy (~3.4 MeV per nucleon) in the process. Current research is attempting to produce controlled fusion in the laboratory as a means of generating energy that would be much cleaner than fission nuclear power plants. Very high power lasers are being used to attempt to achieve the very high temperatures and pressures needed to cause fusion.

QUESTIONS

1. Carefully distinguish between Z, A, and N. Which of these change and which remain constant for α, β, and γ decay?
2. Why does Equation (26.2) lead to a picture of the nucleus as a dense-packed ball of nucleons in contact with each other? (Hint: How does the equation predict the nuclear mass will vary with radius?)
3. Compare the notion of nuclear binding energy to the ionization energy of an atom. How is it similar and how is it different?
4. What three factors determine the stability of nuclei?
5. Which of the following nuclei have net spin (indicate whether due to protons or neutrons) in their ground states: ^{1}H, ^{12}C, ^{13}C, ^{15}N, ^{31}P?
6. What are magic numbers and how are they determined?
7. Compare the α, β, and γ decay particles in terms of penetrating power and radiation damage produced.
8. Complete the following nuclear processes by stating what the nucleon X represents: (a)

$$^{60}\text{Co} \rightarrow {}^{60}\text{Ni} + X; (b) X \rightarrow {}^{234}\text{Pa} + \beta^{-};$$
$$(c) X \rightarrow {}^{230}\text{Th} + {}^{4}\text{He}$$

9. Write the complete decay scheme for the β^{-} decay of ^{126}Sn and of ^{60}Co.
10. Does it matter at what time a measurement starts in order to measure the half-life of a radioactive sample? That is, suppose two experimenters take the same sample of a radioactive material with a 3 min half-life and each independently tries to measure the half-life. Does it matter whether they start their measurements at the same time?
11. If a radioactive sample starts out with 10^{20} nuclei, how many will be left after 10 half-lives? After 10 lifetimes?
12. What are the fortuitous circumstances that allow ^{14}C dating of once-living organisms?
13. What is the difference between exposure, absorbed dose, and biological dose equivalent? Which is most important in determining health risks?
14. Which has the longer effective half-life when used as a radiopharmaceutical, an isotope with a 45 h half-life that takes twice as long to wash out of the body as a second isotope with a 75 h half-life and a 7 day biological half-life?
15. In the radioimmunological assay, why is the ratio of the labeled-to-unlabeled antigen bound to the antibody the same as that ratio found in the supernatant?
16. In PET, how is it known that a particular detected gamma came from pair annihilation within the body?
17. In a nuclear fission power plant, what is the purpose of a control rod? A moderator?
18. Which would liberate more energy: assembling 14 protons and 14 neutrons to make one ^{28}Si nucleus or two ^{14}N nuclei?

MULTIPLE CHOICE QUESTIONS

1. The nucleus is about this many times as large as the atom: (a) 10^{-2}, (b) 10^{-3}, (c) 10^{-4}, (d) 10^{-5}.
2. The nuclide of iron $^{56}_{26}$Fe, has (a) 30 neutrons, (b) 56 neutrons, (c) 26 electrons, (d) 56 protons.
3. The Nobelium nucleus, $^{255}_{102}$No, a very short-lived (3 min half-life) manmade nuclide, has an effective radius of about (a) 1.2 fm, (b) 7.6 fm, (c) 5.6 fm, (d) 0.19 fm.
4. Without the strong nuclear force, carbon-based life could not exist. This is primarily because the strong nuclear force (a) permits nuclei to consist of more than just a single proton, (b) keeps gravity from collapsing all matter into a single point, (c) keeps the electric force from attracting all electrons into the nucleus, (d) is responsible for covalent bonds between carbon atoms.
5. Very large nuclei are radioactive because of all but the following (a) the electrical repulsive force destabilizes them, (b) the total binding energy of two fission fragments is smaller than that of the original nucleus, (c) the excess neutrons do not sufficiently shield the repulsive force between protons, (d) it is relatively easy for a proton to escape (tunnel) out of the nucleus.
6. A nucleus with 20 neutrons is unusually stable because (a) it has an unusually low binding energy, (b) it has a closed nuclear shell, (c) it has all spin paired neutrons, (d) it has an unusually high propensity for β decay.
7. The neutrino was first predicted in beta decay because of missing (a) spin, (b) angular momentum, (c) energy, (d) charge.
8. In a scintillation detector, incoming X's are eventually converted into outgoing Y's where X and Y could be (a) gammas and visible photons, (b) electrons and electrons, (c) electrons and visible photons, (d) visible photons and electrons.
9. Starting with 10^{12} radioactive nuclei, after 4 half-lives about (a) 2.5×10^{11}, (b) 1.8×10^{10}, (c) 10^{8}, (d) 6.3×10^{10} nuclei will remain.
10. In radioactive decay, compared to the activity, the total number of radioactive nuclei decays (a) exponentially with the same half-life, (b) logarithmically with the same half-life, (c) exponentially with a greater half-life, (d) exponentially with a smaller half-life.
11. Carbon-14 dating relies on all of the following assumptions except (a) the amount of ^{14}C in the air has remained constant, (b) no new ^{14}C is taken in after death, (c) ^{14}C is the only radioactive form of carbon with a long half-life, (d) ^{14}C is an alpha emitter with a 5730 year half-life.
12. One hundred hours of flying time in a high-altitude jet gives an equivalent radiation dose to a diagnostic x-ray. This large dose is primarily due to the fact that (a) there are fewer people at those altitudes to absorb

the cosmic rays so each person gets a higher dose, (b) the Earth's atmosphere shields us on the ground from most cosmic rays (c) the plane is that much closer to the sun, so the dose is higher, (d) cosmic rays interact with the plane's metal and produce high doses of secondary radiation.

13. Which of the following is not a desired feature of a radiopharmaceutical? (a) A relatively short (hours) half-life, (b) relatively low energy radiation, (c) a long biological half-life, (d) ability to bind to specific target tissue.

14. In liquid scintillation counting which of the following is not true? (a) For each radioactive decay a single electron is detected, (b) a radioactive decay results in a photon of visible light, (c) different radioactive emissions result in different amplitudes of detected current pulses, (d) the solvent that the sample is dissolved or suspended in has a fluorescent component.

15. The detected particles in PET are all but the following. (a) They are an electron and positron, (b) they are measured 180° apart along a line, (c) their energy is equal to the rest mass of the electron, (d) they are produced after ingesting a positron emitting source.

16. Essential features of a nuclear fission power plant include all but the following. (a) Moderators, (b) control rods, (c) deuterium–tritium fuel pellets, (d) steam turbines.

PROBLEMS

1. The largest stable nucleus has a mass number of 209. Find the ratio of the radii, surface areas, and volumes of this largest nucleus to that of a hydrogen nucleus.

2. A neutron star has a diameter of about 20 km and has a density roughly that of the nucleus. What is its mass? How many solar masses is this (solar mass = 2×10^{30} kg)? What is the mass number for the neutron star (i.e., how many nucleons does it contain)?

3. If the sun (mass = 2×10^{30} kg, radius = 7×10^{8} m) collapsed until it had a density equal to that of nuclei, what would be its radius? (Actually, a star cannot collapse to nuclear densities unless its mass exceeds a critical mass, known as the Chandrasekhar mass, of about 1.4 solar masses to overcome the Pauli exclusion repulsion of the electrons.)

4. Using the numbers in Example 26.1 calculate the binding energy of radium 226 ($m = 225.97709$ u), radium 228 ($m = 227.98275$ u), and thorium 232 ($m = 231.98864$ u). Also find their binding energy per nucleon.

5. Calculate Q for the α-decay of $^{232}_{90}$Th using data in the previous problem and in Example 26.1.

6. Calculate Q for the β^- decay of ^{24}Na given the following data: m(^{24}Na) = 23.98492 u, m(^{24}Mg) = 23.97845 u, m(^{24}Ne) = 23.98812 u, m(β^-) = 5.49×10^{-4} u. What is the range of possible energies for the emitted beta particle?

7. Show that in alpha decay from a stationary parent nuclide that conservation of energy and momentum lead to a relation between the Q for the nuclear reaction and the kinetic energy gained by the alpha particle, KE, given by

$$Q = KE\left(1 + \frac{m(^4_2\text{He})}{m(\text{daughter})}\right).$$

Then look back at Example 26.2 and calculate the kinetic energy of the alpha emitted in the decay of ^{238}U.

8. The first successful experiment to detect the neutrino was done in 1953 by Reines who won the 1995 Nobel prize for this work. Neutrinos from the Hanford nuclear reactor were incident on a tank of 200 L of water in which they very infrequently interacted with the water protons to produce a neutron and a positron in the reaction: $\nu + p \rightarrow n + \beta^+$. The positrons subsequently annihilated producing two signature gammas traveling 180° apart and the neutrons were captured by cadmium in the form of 40 kg of $CdCl_2$ salt added to the water to produce several additional gammas. These gammas were detected by three scintillator layers and 110 photomultiplier tubes surrounding the water and Reines and co-workers carried out numerous checks to ensure that these gammas did originate from the above reaction and not from any other source. Although the neutrino flux was as much as 10^{13} /cm^2-s, they detected on average only 0.027 events per hour per phototube in the entire detector. Assuming the cross-sectional dimensions of the water tank to be 2 m on a side and the neutrino flux to be uniform over this area, what fraction of the neutrinos interacted with the water in their detector, assuming 100% collection efficiency of the gammas?

9. Suppose that the phototube of a scintillation detector has a gain of 5×10^5, representing the average number of electrons produced at the anode for each electron emitted at the photocathode. If a 10 μCi gamma emitting radioactive source is detected with 5% efficiency, find the average output current from the phototube.

10. How long does it take for 90% of a ^{60}Co sample originally present to decay?

11. What mass of ^{90}Sr is needed to have an activity of 1 mCi? How long will it take for the activity to decrease to 0.25 mCi?

12. Iodine is selectively accumulated in the thyroid gland where it can build to dangerous levels. When radioactive materials have been released into the atmosphere from either nuclear power plant accidents, such as the major one at Chernobyl, or from nuclear testing, these materials tend to concentrate (through eating of plants by animals) and show up in food products, such as milk, a food which is preferentially eaten by children. Thus, even though the half-life of ^{131}I is relatively

short, this isotope has caused thyroid cancer in many children in affected areas. Some children received up to 1000 rem from ^{131}I release at Chernobyl. By what factor is this above the maximum annual exposure recommended? How long would it have taken for the radiation level to have decreased so that for the same consumption of milk the exposure would have been at 0.1 times the maximum annual recommendation?

13. After the sudden release of radioactivity from the Chernobyl nuclear reactor accident in 1986, the radioactivity of milk in Poland rose to 2000 Bq/L due to iodine-131 present in the grass eaten by dairy cattle. Radioactive iodine, with a half-life 8.0 days, is particularly hazardous because the thyroid gland concentrates iodine.
(a) What is the decay constant that characterizes the decay of ^{131}I if it has a half-life of 8 days?
(b) What is the storage time needed to decrease the ^{131}I content of cheese produced from these cows' milk to 15% of the original level?

14. A bone fragment is found in the desert. If it has a mass of carbon (due to only ^{14}C and ^{12}C) of 200 g, how old is it if it has an activity of 15 decays per second? The ratio of ^{14}C to ^{12}C is 1.3×10^{-12}.

15. To destroy a cancerous tumor, a dose of gamma radiation totaling an energy of 2.12 J is to be delivered in 30.0 days from implanted sealed capsules containing palladium-103. Assuming that this isotope has a half-life of 17.0 days and emits gamma rays of energy 21.0 keV, which are entirely absorbed within the tumor, what is the initial activity of the set of capsules, and what total mass of radioactive palladium should these "seeds" contain?

16. ^{238}U decays by the emission of an alpha particle.
(a) What is the decay sequence?
(b) What is the daughter nucleus?
(c) What is the energy of the alpha particle (its mass is 4.0026 u)?
(d) What is the velocity of the α particle?
(e) Is the alpha particle relativistic?

17. Strontium is chemically similar to calcium and can replace calcium in bones. The radiation from ^{90}Sr can damage the bone marrow where blood cells are produced, and lead to serious health problems. How long would it take for all but 0.01% of a sample of ^{90}Sr to decay?

18. Calculate the activity (in Bq) of one gram of radium-226. (Hint: See Example 26.3; this is the definition of one curie.)

19. An amateur archeologist finds a bone that he believes to be from a dinosaur. He sends a chip of it off to a laboratory for ^{14}C dating. The lab finds that the chip contains 5 g of carbon and has an activity of 0.5 Bq. How old is the bone? Could it be from a dinosaur?

20. An 85 kg person was exposed to a gamma source and received a whole body dose of 0.5 Sv. How much energy was deposited in the person's body? Repeat this calculation if the radiation was from an alpha source.

21. What dose (in Gy) of gammas produces the same biological effects as a 50 rad dose of alpha particles?

22. What fraction of a 1 g sample of ^{90}Sr sitting on a table will remain in 17 years? If the strontium had been ingested and all initially been absorbed into a person's bones, what fraction would remain after 17 years? (The biological half-life of ^{90}Sr is 45 years.)

23. A small amount of phosphorus-32 was accidentally ingested and its activity carefully monitored over time. After 8 days, the activity had halved. The physical half-life of phosphorus-32 is given in Table 26.4. Find its biological half-life.

24. In a radioimmunological assay 10 nM of a ^{125}I labeled antigen and 1 nM of antibody were added to an unlabeled sample of the antigen. The solution was centrifuged and the activity of the supernatant and pellet were measured and found to be in the ratio of 5.4:1. How much antigen was originally present in the solution?

25. Calculate the net energy released in each step of the proton–proton cycle shown in Equation (26.18). Then add up the net release for one step in the cycle being sure to use a balanced net reaction; recall that the mass of the neutrino and photon are zero. Check your result by a direct calculation of the energy released in the net overall reaction. Use the following masses: $m(^1H) = 1.00728$ u; $m(^2H) = 2.01355$ u; $m(e^+) = 5.49 \times 10^{-4}$ u; $m(^3H) = 3.01550$ u; $m(^4He) = 4.00151$ u.

Appendix I

Review of Mathematics

1. SCIENTIFIC NOTATION AND SIGNIFICANT FIGURES

When large or small numbers are written out in decimal form they are often very cumbersome and difficult to read at a glance. For example, writing out the electric charge on the electron in decimal form would mean writing 18 zeroes after the decimal point followed by the digits 1602 in units of coulombs (C): 0.000000000000000001602 C. Scientific notation is the particular form in which numbers are written with the powers of 10 extracted as an exponent factor. The electron's charge is written in the simpler and more quickly grasped form as 1.602×10^{-19} C.

To understand this notation you need to remember that positive powers of 10 are given by:

$$10^0 = 1$$
$$10^1 = 10 \times 1 = 10$$
$$10^2 = 10 \times 10 = 100$$
$$10^3 = 10 \times 10 \times 10 = 1000, \ldots,$$

and negative powers of 10 are given by

$$10^{-1} = \frac{1}{10} = 0.1$$
$$10^{-2} = \frac{1}{10} \times \frac{1}{10} = 0.01$$
$$10^{-3} = \frac{1}{10} \times \frac{1}{10} \times \frac{1}{10} = 0.001, \ldots.$$

In scientific notation numbers are always written in a form with a single nonzero digit to the left of the decimal point according to $D_1 \cdot D_2 D_3 D_4 \times 10^{\pm D_5 D_6}$, where the D_i (i = an integer) are decimal digits. The number of digits written out in the decimal prefactor number (D_1 through D_4 in the above example, and so 4 digits) is called the number of significant figures.

What determines how many significant figures to include in a number? If it is a number that is obtained from a measurement, such as the charge on the electron given above, then the number of significant figures depends on the uncertainty in the measurement. When written in its most precise form, the number should include digits out to the level of uncertainty, so that, for example, if a measurement of the electric charge were done with an uncertainty of $\pm 0.02 \times 10^{-19}$ C, then the number for the charge should include 3 significant digits 1.60×10^{-19} C, because the last digit, 0, has the uncertainty of ± 2 units in its place. Higher-precision measurements include more significant figures.

If the number is derived from other numbers, say by combining other given numbers, each with some number of significant figures, you should be careful about how many significant digits you include in your result. When you enter numbers in your

calculator and compute some result, the calculator display fills with digits, but not all of them are significant. You should retain only as many digits as the least significant number in the calculation. Of course whole number and mathematical values such as π or e have a huge number of significant digits and do not restrict your precision.

Example A.1 Given these values, $x = 2.38$ and $y = 6.45$, compute each of following: $z = x + y$, $A = xy$, and $B = x^{1/2}y^{1/2}$.

Solution: In each case first simply compute the values (probably using a calculator). We find $z = 8.83$, $A = 15.351$, and $B = 3.9180352$. Both x and y have 3 significant digits, therefore we round off our final results to be $z = 8.83$, $A = 15.4$, $B = 3.92$.

2. ALGEBRA

In this section we review basic algebra through the rules for manipulating equations. An equation represents a statement of the equality of both sides. As such, the quantities on both sides must have the same units.

Remember the basic rules for multiplication, division, and addition or subtraction of fractions and working with exponents:

Process	*Rule*	*Example*
Addition or subtraction	$\frac{a}{b} \pm \frac{c}{d} = \frac{ad \pm bc}{bd}$	$\frac{1}{2} - \frac{3}{5} = \frac{(1)(5) - (2)(3)}{10} = -\frac{1}{10}$
Multiplication	$\frac{a}{b} \times \frac{c}{d} = \frac{ac}{bd}$	$\frac{1}{2} \times \frac{3}{5} = \frac{3}{10}$
Division	$\frac{a}{b} \div \frac{c}{d} = \frac{\left(\frac{a}{b}\right)}{\left(\frac{c}{d}\right)} = \left(\frac{a}{b}\right) \times \left(\frac{d}{c}\right) = \frac{ad}{bc}$	$\frac{1}{2} \div \frac{3}{5} = \frac{1}{2} \times \frac{5}{3} = \frac{5}{6}$
Exponents	$\frac{x^m \cdot x^n}{x^p} = x^m \cdot x^n \cdot x^{-p} = x^{m+n-p}$	$\frac{2^3 \cdot 2^2}{2^4} = 2^{3+2-4} = 2^1 = 2$
	$(x^m)^{\pm n} = x^{\pm mn}$	$(2^3)^2 = 2^6 = 64$ and
	$x^{1/n} = \sqrt[n]{x}$	$(2^3)^{1/2} = \sqrt{8}$

Also remember that any operation applied to one side of the equation must be done on the other side as well to preserve the equality. You can add, subtract, multiply, or divide both sides of the equation by equal numbers or quantities (with the obvious exception of dividing by zero). Examples of these operations should help you to recall this.

Example A.2 (a) Given the equation $2x + 7 = 15$, solve for x; (b) Given the equation $P = \frac{1}{2}\rho(v^2 - v_0^2) + \rho gh$, solve for v.

Solutions: (a) We first add -7 to both sides:
$2x + 7 - 7 = 15 - 7$ *or* $2x = 8$ and then divide both sides by 2 to find $\frac{2x}{2} = \frac{8}{2}$ *or* $x = 4$.

(b) We first subtract ρgh from both sides to find $P - \rho gh = \frac{1}{2}\rho(v^2 - v_0^2)$; we then multiply both sides by $2/\rho$ to find

$$\frac{2}{\rho}(P - \rho gh) = \left(\frac{2}{\rho}\right)\frac{1}{2}\rho(v^2 - v_0^2) = (v^2 - v_0^2);$$

next, add v_0^2 to both sides to get

$$\left[\frac{2}{\rho}(P - \rho gh)\right] + v_0^2 = (v^2 - v_0^2) + v_0^2 = v^2;$$

finally, we take the square root of both sides with the result

$$v = \sqrt{\left[\frac{2}{\rho}(P - \rho gh)\right] + v_0^2}.$$

Occasionally you will need to solve a quadratic equation. Remember that if you rewrite the quadratic equation in the form $ax^2 + bx + c = 0$, the solution is given by

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

Often in a physics problem only one of the two roots, the mathematical solutions, will be the correct answer to the physics problem, and some thought as to the consequences of each answer can usually let you make the correct choice.

3. GEOMETRY

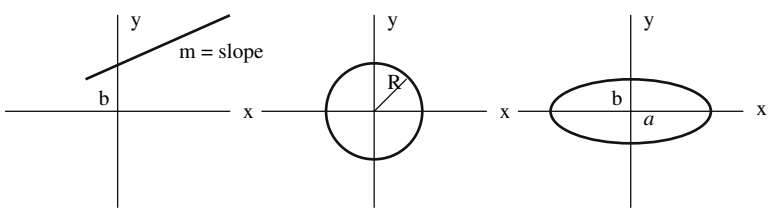
In this section we summarize a number of geometric relationships.

The distance between two points labeled in Cartesian coordinates as (x_1, y_1) and (x_2, y_2) is given by

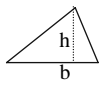
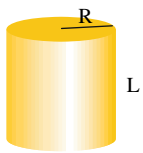
$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}.$$

Some useful equations representing different geometrical shapes include:

$y = mx + b$	straight line of slope m and y – intercept b
$x^2 + y^2 = R^2$	circle, centered at the origin, of radius R
$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$	ellipse, centered at the origin, of semimajor axis a and semiminor axis b (with $a > b$)



Some useful formulae for the areas and volumes of regular solids are given in the following table.

Shape	*Area or Volume*	
Rectangle	Area = $L \times w$	
Circle	Area = πR^2	
Triangle	Area = 1/2 bh	
Sphere	Surface Area = $4\pi R^2$ Volume = $4/3\ \pi\ R^3$	
Cylinder	Surface Area = $2\pi RL$ (plus area of circular end caps) Volume = $\pi R^2 L$	

4. EXPONENTIALS AND LOGARITHMS

Recall that the logarithm of a quantity x, whether it be a number or a variable, with respect to the base b, written $y = \log_b(x)$, means that $x = \text{antilog}_b(y) = b^y$, where y is the exponent to which the base b must be raised to equal x. The two most used bases are base 10, common logarithms, and base e, natural logarithms. Explicitly, these are

$$y = \log_{10}(x) \qquad \text{or} \qquad x = 10^y,$$

and

$$y = \log_e(x) = \ln(x) \qquad \text{or} \qquad x = e^y.$$

Natural logarithms arise directly from calculus, whereas common logarithms are singled out because of the decimal number system we use. We often omit the base when writing logarithms: log implies base 10, and ln is reserved for natural logarithms.

Logarithms have the following general properties:

$\log(xy) = \log(x) + \log(y)$

$\log(x/y) = \log(x) - \log(y)$

$\log(x^n) = n \log(x)$

$\log(1) = 0$

These four expressions are true no matter what the base. Other specific examples for base 10 and e include:

$\log(10) = 1$

$\log(10^n) = n$

$\ln(e) = 1$

$\ln(e^n) = n.$

5. TRIGONOMETRY

Angles can be measured in a variety of units including degrees and radians, those used in this book. Radian measure is defined through the equation

$$s = r\theta,$$

as

$$\theta = \frac{s}{r},$$

where s is the arc length of a circular arc of radius r. Whatever units are used for s and r, θ is dimensionless. In a full circle we have $s = 2\pi r$ and so there are 2π radians corresponding to 360°.

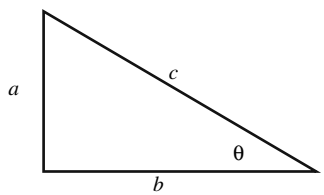
Note that when using a calculator, care must be taken to set the mode in which angles are obtained—degrees or radians—to be sure that your answers, when given as angles are correctly understood. For example, if you enter 1.0 and ask for the inverse tangent you will find 45 if you calculator is set for degrees, but 0.785 if set for radians. Both are correct but you need to know which units for angle the calculator is reporting.

Trigonometry deals with the special properties of right triangles, those with a 90° angle. Because the three angles of a triangle must add to 180° in Euclidean geometry, special relationships arise between the sides and angles in a right triangle. Using the right triangle labeled below, we remind you of the basic definitions of the trigonometric functions (summarized in the famous mnemonic *SOH CAH TOA*):

$$\sin\theta = \frac{\textit{opposite side}}{\textit{hypotenuse}} = \frac{a}{c}$$

$$\cos\theta = \frac{\textit{adjacent side}}{\textit{hypotenuse}} = \frac{b}{c}$$

$$\tan\theta = \frac{\textit{opposite side}}{\textit{adjacent side}} = \frac{\sin\theta}{\cos\theta} = \frac{a}{b}$$



Less common are the trigonometric functions for the inverse of these:

$$\csc\theta = \frac{1}{\sin\theta}; \quad \sec\theta = \frac{1}{\cos\theta}; \quad \cot\theta = \frac{1}{\tan\theta}$$

The other important relationship between the sides of the right triangle is given by the Pythagorean theorem:

$$a^2 + b^2 = c^2.$$

From the above relations a whole host of trigonometric identities can be derived; the most important are listed in the table below.

Trigonometric Identities

$\sin^2\theta + \cos^2\theta = 1$

$\sin 2\theta = 2\sin\theta\cos\theta$

$\cos 2\theta = \cos^2\theta - \sin^2\theta$

$\sin(A \pm B) = \sin A\cos B \pm \cos A\sin B$

$\cos(A \pm B) = \cos A\cos B \mp \sin A\sin B$

$\sin(-\theta) = -\sin\theta$

$\sin^2\frac{\theta}{2} = \frac{1}{2}(1 - \cos\theta)$

$\cos^2\frac{\theta}{2} = \frac{1}{2}(1 + \cos\theta)$

$\sin\theta = \cos(90° - \theta)$

$\cos\theta = \sin(90° - \theta)$

$\cos(-\theta) = \cos\theta$

Appendix II

Table of the Elements

Element	Symbol	Number	Discovery
Actinium	Ac	89	1899
Aluminum	Al	13	1825
Americium	Am	95	1945
Antimony	Sb	51	*
Argon	Ar	18	1894
Arsenic	As	33	*
Astatine	At	85	1940
Barium	Ba	56	1808
Berkelium	Bk	97	1949
Beryllium	Be	4	1798
Bismuth	Bi	83	*
Bohrium	Bh	107	1976
Boron	B	5	1808
Bromine	Br	35	1826
Cadmium	Cd	48	1817
Calcium	Ca	20	1808
Californium	Cf	98	1950
Carbon	C	6	*
Cerium	Ce	58	1803
Cesium	Cs	55	1860
Chlorine	Cl	17	1774
Chromium	Cr	24	1797
Cobalt	Co	27	1737
Copper	Cu	29	*
Curium	Cm	96	1944
Darmstadtium	Ds	110	1994
Dubnium	Db	105	1970
Dysprosium	Dy	66	1886
Einsteinium	Es	99	1952
Erbium	Er	68	1843
Europium	Eu	63	1901
Fermium	Fm	100	1953
Fluorine	F	9	1886
Francium	Fr	87	1939

Gadolinium	Gd	64	1880
Gallium	Ga	31	1875
Germanium	Ge	32	1886
Gold	Au	79	*
Hafnium	Hf	72	1923
Hassium	Hs	108	1984
Helium	He	2	1895
Holmium	Ho	67	1878
Hydrogen	H	1	1766
Indium	In	49	1863
Iodine	I	53	1804
Iridium	Ir	77	1804
Iron	Fe	26	*
Krypton	Kr	36	1898
Lanthanum	La	57	1839
Lawrencium	Lr	103	1961
Lead	Pb	82	*
Lithium	Li	3	1817
Lutetium	Lu	71	1907
Magnesium	Mg	12	1808
Manganese	Mn	25	1774
Meitnerium	Mt	109	1982
Mendelevium	Md	101	1955
Mercury	Hg	80	*
Molybdenum	Mo	42	1778
Neodymium	Nd	60	1925
Neon	Ne	10	1898
Neptunium	Np	93	1940
Nickel	Ni	28	1751
Niobium	Nb	41	1801
Nitrogen	N	7	1772
Nobelium	No	102	1957
Osmium	Os	76	1804
Oxygen	O	8	1774
Palladium	Pd	46	1803
Phosphorus	P	15	1669
Platinum	Pt	78	1735
Plutonium	Pu	94	1940
Polonium	Po	84	1898
Potassium	K	19	1807
Praseodymium	Pr	59	1885
Promethium	Pm	61	1945
Protactinium	Pa	91	1917
Radium	Ra	88	1898
Radon	Rn	86	1898
Rhenium	Re	75	1925
Rhodium	Rh	45	1803
Roentgenium	Rg	111	1994
Rubidium	Rb	37	1861
Ruthenium	Ru	44	1844
Rutherfordium	Rf	104	1969

Samarium	Sm	62	1879
Scandium	Sc	21	1879
Seaborgium	Sg	106	1974
Selenium	Se	34	1817
Silicon	Si	14	1823
Silver	Ag	47	*
Sodium	Na	11	1807
Strontium	Sr	38	1790
Sulfur	S	16	*
Tantalum	Ta	73	1802
Technetium	Tc	43	1937
Tellurium	Te	52	1782
Terbium	Tb	65	1843
Thallium	Tl	81	1861
Thorium	Th	90	1828
Thulium	Tm	69	1879
Tin	Sn	50	*
Titanium	Ti	22	1791
Tungsten	W	74	1783
Uranium	U	92	1789
Vanadium	V	23	1830
Xenon	Xe	54	1898
Ytterbium	Yb	70	1878
Yttrium	Y	39	1794
Zinc	Zn	30	1746
Zirconium	Zr	40	1789

*Known to Ancient Civilization

Color	Group
	1
	2
	3–12
	13
	14
	15
	16
	17
	18
	Series

Appendix III

Answers to Odd-Numbered Multiple Choice and Problems

Chapter 1

MC

1. D
3. D
5. B

P

1. (a) 6.7×10^{-27} kg; (b) 2.7×10^{-26} kg; (c) 2.4×10^{-26} kg; (d) 5.2×10^{-26} kg
3. 2.7×10^{-10} m
5. (a) 1.9×10^{-5} $\mu g/\mu m^3$; (b) 1.9×10^{-8} pg/nm^3
7. (a) 5×10^{-18} kg; (b) 5×10^{-16} kg

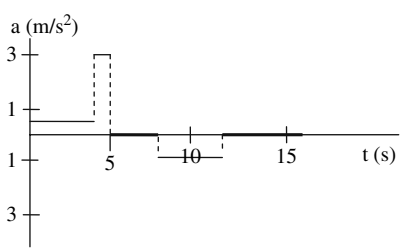
Chapter 2

MC

1. A
3. B
5. D
7. D
9. A
11. A
13. A
15. C
17. D
19. B
21. C
23. B
25. C
27. D
29. C
31. E

P

1.



3. 12.3 s
5. (a) $v = 2.5, 10, 7.5, 2.5$ μm/s; $a = 1.5, -0.5, -1.0$ μm/s^2; (b) $v = 5, 5, 7, 5, 6$ μm/s; $a = 0, 0.5, -0.3$ μm/s^2; (c) $v = 5, 5, 5, 5, -2.5, -2.5, -2.5, -2.5, -2.5$ μm/s; $a = 0, 0, 0, -0.5, 0, 0, 0, 0$ μm/s^2
7. (a) 6.7 m/s, 13.4 m/s, 6.7 m/s, 11.9 m/s = 26.5 mi/h; (b) 33.6 m, 483 m, 40.2 m, 556.8 m; (c) 0 m/s^2
9. (a) $t = 12$ s (b) 360 m $<$ 848.5 m so yes.
11. (a) 49 m/s down; (b) 9.8 m/s^2 up; (c) 19.6 N
13. (a) $x = 300$ m from the police car's starting position; (b) $t = 20$ s from when the police car started moving.
15. 4730 m/s
17. 4.88×10^{16} N
19. 35.3 N/m^2
21. (a) 0.167 m/s^2; (b) 0.143 m/s^2; (c) 0.167 m/s and 0.143 m/s; (d) 60.5 s and 70.5 s
23. (a) 0.67 m/s^2; (b) 13.3 N; (c) 13.3 N; (d) Newton's 3rd law; (e) same
25. (a) 139,000 h, or never, convection will undoubtedly play a role here; (b) 0.45 μm
27. (a) 7.5×10^{-7} m/s; (b) 1.33×10^4 turns/min

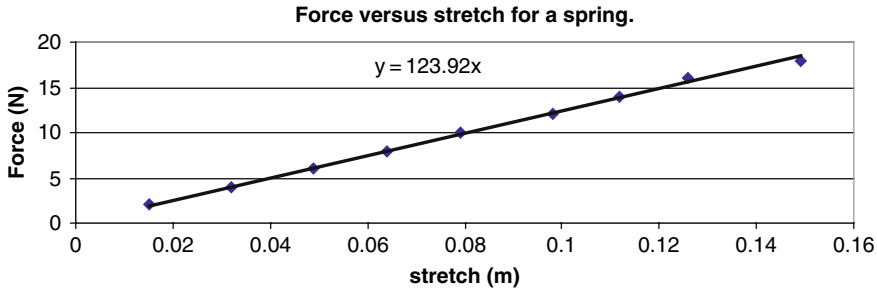
Chapter 3

MC

1. A
3. B
5. C
7. B
9. C
11. C
13. C
15. A
17. D
19. B

P

1. (a) 73 min; (b) 32.7 km/h
3. 25.9 km/h
5. 0.98 m/s^2
7. 32.7 s
9. (a) 75.4 m; (b) 14 m/s
11. (a) 16.1 m/s; (b) 0.83 s; (c) 5.1 m
13. (a) 3.5 s; (b) 9.8 m/s^2 down; 34.3 m/s down
15. No – his speed is 39.5 mph
17. 9 m/s^2
19. (a) 15 m/s; (b) 2 m/s^2 to right; (c) 2.86 N to right, source is 5 kg block
21. (a) 80.5 m; (b) 40.1 m/s
23. 2530 m/s^2 = 258 g's
25. 51.5 m/s
27. (a) 0.58; (b) 1.7; (c) 1.7; (d) 3; (e) 1
29. (a) 7.9 N/m; (b) $x = (0.1\text{ m})\cos(3.14t)$; at $t = 1$ s, $x = 0.1$ m; at $t = 2$ s, $x = -0.1$ m; at $t = 1.5$ s, $x = 0$; at $t = 1.25$ s, $x = 0.07$ m; (c) $v = (0.314\text{ m/s})\sin(3.14t)$; at $t = 1$ s, $v = 0$; at $t = 2$ s, $v = 0$; at $t = 1.5$ s, $v = -0.314$ m/s; at $t = 1.25$ s, $v = -0.22$ m/s
31. (a) 208 m; (b) 3.81×10^4 N; (c) 55.5 X
33. (a) 35.1 N/m; (b) 1.75 m/s^2 at amplitude; (c) 0.15 m/s at equilibrium
35. (a) From the graph below, the slope represents the spring constant and its value is 123.9 N/m.



(b) F = 126.4 N up
37. 0.098 N
39. 2.2×10^{-4} m
41. (a) $\Delta L = 9.6$ μm; (b) $k = 1.31 \times 10^9$ N/m; (c) $T = 6.2$ ms
43. 6200 N

Chapter 4

MC

1. B
3. A
5. D
7. A
9. C
11. B
13. B

P

1. (a) 3.15×10^4 J; (b) ~27
3. (a) 330 J; (b) −300 J; (c) 300 J; (d) 2.4 m/s
5. (a) 22.3 m/s; (b) 1.76 s; (c) 2.79 s; 22.3 m/s
7. 8.01 m/s
9. (a) −127 kJ; (b) 31.8%
11. (a) 1400 J; −1200 J; (b) 1.8 m/s; (c) 0.17 m; 0.18 s; (d) 1480 J
13. 0.24 m/s
15. (a) 0.1 kg; (b) 0.1 J; (c) 1.41 m/s at equilibrium (5 cm below initial unstretched spring position)
17. 1.0×10^4 W
19. (a) assuming to the right is $+x$, $v_J = 6$ m/s to the right, $v_S = 6$ m/s to the right; (b) $a_S = 0.6$ m/s^2 to the left; (c) $F_S = 30$ N to the left; (d) W = −900 J

Chapter 5

MC

1. C
3. A
5. A
7. C
9. B
11. A
13. D
15. C
17. D
19. E

21. C
23. C
25. D
27. C
29. A
31. A
33. C
35. C
37. C
39. B

P

1. From square1 to square2 = (0,1); from square2 to square3 = (2,0); from square3 to square4 = (0,2); from square4 to square5 = (3,0); from square5 to square6 = (−4,0); from square6 to square7 = (0,4); The displacements are the same for both boards regardless of the labeling schemes.
3. (a) 135 N vs 75 N; 125 NW vs 125 NW; 85 S vs 101 NE; (b) 400 and 75 N along with 375 and 125 SW; (c) NYC → BGM and KGN → SYR
5. (a) due to current downstream; (b) 35.7° upstream; (c) 1.5 min
7. (a) $A_1 \| B_1$; $A_2 \| B_2$; $A_1 \perp A_2$; $B_1 \perp B_2$; $A_1 \perp B_2$; $B_1 \perp A_2$; (b) (5,0); (0.5); (5,0); (5,5)
9. (a) A = (3,2); B = (7,7); C = (8,3); (b) 6.4; 4.1; 5.1; (c) (10,9); (15,10); (18, 12); (30,20); (d) (−5,−5)
13. (a) 29.8 m/s at 47.9° below the horizontal; (b) 2.3 s; (c) 45.2 m; (d) 9.8 m/s^2 down
15. (a) ~14 km; (b) directly above the bomb
17. (a) 90.4 m; (b) t = 4.52 s
19. a = 9.0 m/s^2
21. (a) 1.17 s; (b) 9.58 m; (c) 4.79 m horizontally; (d) 10 m/s at −35°; (e) 8.97 m/s at −24.0°
23. 2.15 × 10^6 ft/min^2
25. (a) 0.11 m/s; (b) 0.055 m/s^2; (c) 12 s/rev
27. 329 N; (b) 3.29 × 10^5 J
29. (a) F_{T1} = 11300 N; F_{T2} = 5660 N; (b) v_{max} = 3.24 m/s; (c) ~2 s
31. (a) a = 1.26 m/s^2 = 4.16 ft/s^2; (b) t = 68.9 s; (c) F = 4.1 × 10^6 N
33. (b) 5.23 m/s^2; (c) because the block has no acceleration in part (a); (d) 4.7 kg
35. (a) 4.32 m/s; (b) 4.85 m/s; (c) 0.313 m
37. (a) 17.2 m/s; (b) 25.1 m; (c) 23.5 m
39. (a) 4.15 m/s^2; (b) 4.1 m/s
41. 5.5 cm
43. (a) W_{push} = 60 J; W_{grav} = −29.4 J; W_{frict} = −30.6 J; W_{net} = 0 J
45. (a) 0 N; (b) mg + F sin θ; (c) 117.7 N; (e) yes
47. (a) 0 J; (b) 1.03 J ; (c) 1.03 J
49. (b) 3.92 N; (c) 5.23 m/s^2; (d) 4.67 kg
51. 3.0 × 10^{13} N
53. 26.6°, order does not matter
55. Blocks do not move
57. 5.1 m/s, independent of m
59. (a) 1.02 × 10^3 m/s; (b) 2.72 × 10^{-3} m/s^2 toward Earth
61. (a) 0.03 m/s^2; (b) 0.49 m/s^2 and 1.99 m/s^2
63. 17.1 m/s and 13.7 rpm
65. (b) 5.4 m/s; (c) travels vertically 1.5 m from release point
67. 491 N at 86.2° from the horizontal
69. 5.76 × 10^7 N
71. (a) v = 3.79 m/s; (b) v_b = 17.6 m/s; (c) D = 12 m; (d) W = −58.8 kJ; (e) v = 1.62 m/s

Chapter 6

MC

1. C
3. B
5. A
7. D
9. C

P

1. (a) 0 cm; (b) 8 cm; (c) (3.33, 3.33); (d) (3, 2)
3. (a/2, a/3)
5. (1,1)
7. $(m_1/m_3)x_1$
9. 7.7R
11. (1.5, 1)
13. (a) 1 kgm/s; (b) 200 N; (c) yes, but insignificantly
15. 31.0 m/s at 33.3° below the horizontal in the opposite direction to the 3*m* fragment
17. (a) 21.4 m/s; (b) 11%; (c) 98,800 N; (d) 2.56×10^6 J
19. (a) $v_{f2} = 0.65$ m/s at 38.1° from the initial direction, on the side opposite from the deflected puck; (b) 26.2%
21. (a) $mV = (m + M)V_{\text{after collision}} \rightarrow V = \dfrac{(m + M)}{m}V_{\text{after collision}}$;

    (b) $\dfrac{1}{2}(m + M)V^2_{\text{after collision}} = (m + M)g\Delta h_{\text{cm}} = (m + M)g$

    $(R_{\text{cm}} - R_{\text{cm}}\cos\theta) \rightarrow V_{\text{after collision}} = \sqrt{2gR_{\text{cm}}(1 - \cos\theta)}$;

    (c) $V = \dfrac{(m + M)}{m}\sqrt{2gR_{\text{cm}}(1 - \cos\theta)} = 4.84$ m/s

    $\therefore \vec{V} = 4.84$ m/s in the + x − direction; (d) 83%
23. $v_{\text{block}} = 0.28$ m/s at $\Phi = 23.1°$ from the incident direction, on the side opposite to the deflected bullet
25. (a) $v = \dfrac{v_{ix}}{\sqrt{2}}$; (b) $\Phi = \theta = 45°$

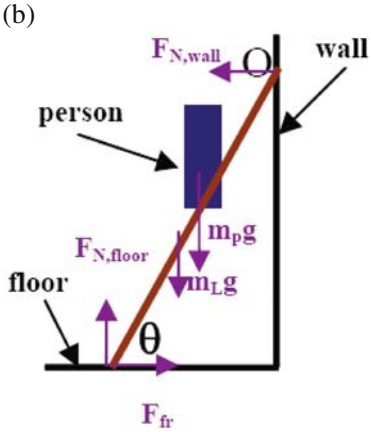
Chapter 7

MC

1. B
3. D
5. D
7. B
9. C
11. D
13. D
15. C
17. D
19. B
21. A
23. Y, Y, N, Y
25. A
27. B
29. B

P

1. (a) 51.7 m/s; (b) 8.1×10^{-6} rad/s
3. 5.48×10^{6} J
5. (a) 4.1×10^{7} rev; (b) 111 rad/s
7. 1.75×10^{-5} m/s
9. 5.0 m/s
11. (a) 89.8 rad/s^2; (b) 22.6 kgm^2/s
13. 5 rad/s; 10 rad/s
15. (a) 0.8 rad/s^2; (b) 57.3 rev (c) tang: 0.2 m/s^2, radial: 144 m/s^2; (d) 7.2 N
17. 1.16 Nm (CCW)
19. (a) 5 rad; (b) $\omega = 5.92$ rad/s; (c) $v = 11.8$ m/s
21. 54°
23. (a) 14 m/s; (b) 0.7 rad/s; (c) 3.5×10^{7} kgm^2/s; (d) 0.088 rad/s; (e) 0.15 m/s^2
25. (a) 1.91 rad/s; (b) $KE_i = 2.53$ J, $KE_f = 6.46$ J
27. (a) 15.7 rad/s; (b) 0.63 Nm, 3.1 N; (c) 3.18 s; (d) 49.3 J
29. (a) 1.67×10^{3} rad/s^2; (b) 5.03 s; (c) 83.8 s; (d) 59,200 rev
31. 31.4 kg
33. 275 N (right end) and 575 N (left end)
35. 0.25 m
37. 498 N; 493 N
39. (a) 24.5 Nm; (b) 19.6 rad/s^2; (c) 6.3 rad/s; (d) 7.83 kgm^2/s
41. (a) $F_w = 8$ N; (b) $F_{min} = F_w / 2$
43. L = 0.20 kgm^2/s
45. (a) $F_w = 600$ N; (b) $F_A = 150$ N, $F_B = 1050$ N
47. (a) $F_T = 113.2$ N; (b) $\alpha = 2.36$ rad /s^2
49. $F = 220$ N
51. (b)



(c) $F_{N,wall} = F_{fr} = \mu_s F_{N,\,floor}$ and $F_{N,\,floor} = m_L g + m_p$; (d) $\theta = 81.1°$

Chapter 8

MC

1. B
3. D
5. A
7. A
9. C
11. D
13. A
15. C
17. A
19. A
21. A

P

1. 0.015 m
3. 5.1 kg/m^3
5. 1.25×10^5 Pa
7. 10.2 m
9. 1.1×10^8 Pa
11. 1.63 m
13. (a) 1.53 cm/s, 9.55 cm/s; (b) 4.4 Pa; (c) 4.1×10^{-8} N
15. 2/3
17. 0.33 m/s
19. (a) v = 0.074 m/s; (b) 1.47×10^{-4} m^3/s; (c) 0.49 m^2; (d) ~16 billion; (e) $KE_{Aorta}/V = 47.25$ J/m^3, $KE_{arteries}/V = 2.84$ J/m^3, $KE_{cap}/V = 4.73 \times 10^{-5}$ J/m^3; (f) 2.5 s
21. v = 9.9 m/s
23. 0.33 mm/s
25. 41 N
27. (a) Area = 4964.7 m^2; (b) waterline area smaller than flight deck by about 75%
29. 5
31. 1.29×10^9 W
33. 110 mph

Chapter 9

MC

1. A
3. D
5. C
7. B
9. B
11. C
13. B
15. C

P

1. $F = 19.6$ mN
3. $\Delta P = 16{,}800$ Pa; 171 cm H_2O
5. $R = 1115$
7. $\eta_{unknown} = 1.071 \times 10^{-3}$ Pa-s
9. (a) $\eta_{spheres} = 1.164 \times 10^{-3}$ Pa-s; (b) 0.001 Pa-s
11. (a) $h = 14.9$ cm (b) $h = 12.9$ cm
13. (a) Pressure higher in smaller bubble. (b) When valve is opened air will rush from the high-pressure bubble into the low-pressure bubble. Thus the pressure decreases and surface tension will tend to collapse the smaller bubble, whereas the larger bubble will grow.
15. $r = 0.026$ mm = 62 μm
17. $\gamma = 0.024$ N/m
19. $L = 0.11$ mm

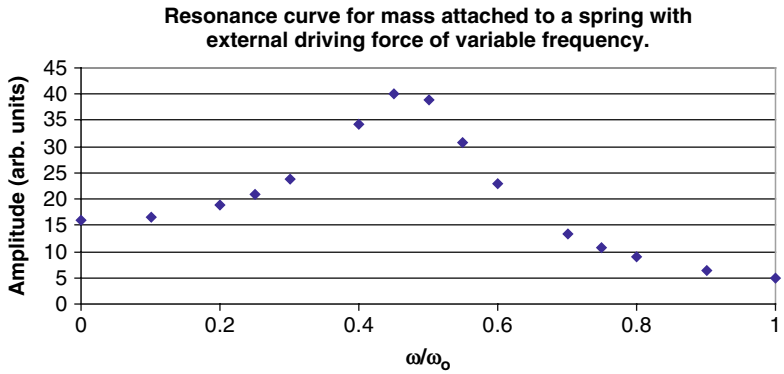
Chapter 10

MC

1. B
3. A
5. D
7. B

9. C
11. A
13. D
15. D
17. B
19. D
21. C
23. A

P

1. $y(t) = 0.1\ e^{-0.069t}\cos(5.7t)$ (in m)
3. (a) 0.503 Hz; (b) 1.99 s; (c) 11.5 s; (d) 5.8; (e) 1%
5. 

7. (a) 16.7 m/s; (b) 4.18 m; (c) $y(x,t) = 0.05\sin(1.5x - 25.1t)$ (in m, s)
9. $y(x,t) = 0.05\sin(25.1x - 628t)$ (in m,s); (b) 25 m/s; (c) 0.016 kg/m
11. (a) 19.8 m/s; (b) 2.83 Hz; (c) 1.75 m
13. 1.31A; λ and f unchanged; $y(x,t) = 1.31A\sin(2\pi x/\lambda - 2\pi ft + \pi/4)$
15. $\lambda = 0.5$ m; $n = 6$
17. 16.5 Hz

Chapter 11

MC

1. C
3. B
5. D
7. C
9. C
11. B
13. B
15. D
17. E
19. C
21. D

P

1. (a) 0.71 s; (b) 0.16 s
3. $v = v_0 + 0.6(T - 20^\circ)$, where $v_0 = 343$ m/s and T in °C
5. 17.2 m and 1.7 cm
7. Open: 8.6 m and 8.6 mm; closed: 4.3 m and 4.3 mm
9. 0.5 J/s
11. 0.0013 N
13. 1715 m

15.

Note	Freq., Hz	String Length, m
C	262	0.67
D	294	0.6
E	330	0.53
F	349	0.5
G	392	0.45
A	440	0.4
B	494	0.35
C	523	0.34

17. Steel wound with copper at 55 mil diameter
19. 191 μs
21. 0.148 mm in water or 0.157 mm in body tissue
23. (a) 1090 N; (b) 4; (c) 0.52 m; (d) by 1.5% or 0.008 m ; (e) closed
25. 58.3 kHz

Chapter 12

MC

1. A
3. B
5. A
7. D
9. A
11. D
13. A
15. C
17. A
19. C
21. A
23. B
25. A
27. A
29. C

P

1. $T_c = -40°C$
3. $\Delta L = 5640\ mm^3$; 0.58%
5. $\Delta l = 2.7 \times 10^{-5}$ m
7. $v_{oxy} = 478$ m/s
9. $P = 46.5$ atm
11. 55.3 mol/L; 3.33×10^{25} H_2O molecules
13. 426 three meter flights of stairs
15. $m_{ice} = 150$ g
17. $P_{athlete} = 330$ W
19. $P_{evaporation} = 970$ W

Chapter 13

MC

1. D
3. D
5. A

P

1. $\Delta S = 27.7 \times 10^3$ J/K
3.

	6 Heads	5 Heads	4 Heads	3 Heads	2 Heads	1 Head	0 Heads		
# Outcomes = 64									
# Ways	1	6	15	20	15	6	1		64
Prob (%) = (#Ways/64)*100	1.5625	9.375	23.4375	31.25	23.4375	9.375	1.5625	Total	100

5. Plotting the given equation, the slope is proportional to ΔH and the intercept is proportional to ΔS.
7.

Macrostate	# Microstates
(2,0,0,0,0,0,1)	3
(0,0,3,0,0,0,0)	3
(1,0,2,0,0,0,0)	6
(1,1,0,0,0,1,0)	6
(1,0,1,0,1,0,0)	6
(0,2,0,0,1,0,0)	6
(0,1,1,1,0,0,0)	6
Total microstates	36

Chapter 14

MC

1. C
3. C
5. C
7. B
9. C
11. A
13. C
15. C
17. A

P

1. $6.25 \times 10^{18} e^-$ and 5.69×10^{-12} kg
3. 0.19 m
5. 2.2×10^6 m/s
7. $Q = 1.8$ μC and $F_T = 0.226$ N
9. $F_{net1} = -1.13\text{N}\hat{i}$; $F_{net2} = 0\text{N}$; $F_{net3} = 1.13\text{N}\hat{i}$
11. $F_{net} = 0$N; F_{net} 1 charge removed $= kQ^2/L^2$ toward the removed charge; L = length of side of hexagon.
13. (a) $F = 3.6 \times 10^{-3}$ N; (b) $q_2 = 8 \times 10^{-7}$C
15. $F = kQq/4R^2$
17. $E = 1.62 \times 10^6$ N/C directed to the midpoint of the side with two -10 μC charges
19. $E_{net} = 2.65 \times 10^4$ N/C directed perpendicular to and away from the plane containing the lines of charge.
21. $E = 1.13 \times 10^8$ N/C
23. For $M_{Na} = 3.89 \times 10^{-26}$ kg $a_{Na} = 4.11 \times 10^{13}$ m/s^2
25. $R = 26.5$ mm

27. (a) $E = 0$; (b) $E = \dfrac{\lambda}{2\pi\varepsilon_0 r}$; (c) $E = 0$

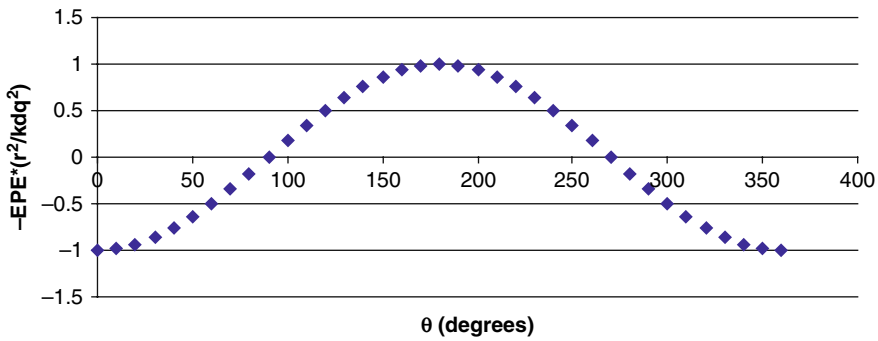
29. $E = \sigma/\varepsilon_0$ toward the negatively charged sheet

Chapter 15

MC

1. C
3. C
5. B
7. D
9. B
11. C
13. D
15. C
17. D
19. B

P

1. PE $= 1.08 \times 10^6$ J
3. (a) $V = 0$; (b) $E = -1.8 \times 10^7$ N/C $\hat{\imath}$; (c) $W = 0$; (d) $V = 1.8 \times 10^6$ V; (e) $E = 0$ N/C; (f) $W = 18$ J
5. (a) 0.36 N/C $\hat{j}$; (b) $V = 1.8$V; (c) -0.18 N/C $\hat{j}$; d) $V = 0$ V
7. $\tau = 6.78 \times 10^5$ Nm
9. 

The equilibrium points are at 0°, 180°, and 360°. 0° and 360° (same physical situation) are stable equilibrium points whereas 180° is an unstable equilibrium point.
11. 6×10^6 V
13. (a) $q = 2 \times 10^{-4}$ C; (b) 4×10^{-4} C
15. (a) $C = 88.5$ pF; (b) $q = 8.85 \times 10^{-10}$ C; (c) $E = 1 \times 10^4$ N/C; (d) $F = 8.9 \times 10^{-6}$ N
17. (a) $C = 3$ μF; (b) $C_{new} = 14.1$ μF; (c) $q = 169$ μC
19. $C/A = 6.4 \times 10^{-2}$ F/m^2
21. 20 mV
23. 4470 V

Chapter 16

MC

1. C
3. C
5. A
7. C
9. C
11. C

13. B
15. A
17. B

P

1. $I = 96.3$ A
3. $I_{avg} = 0.8$ nA
5. $R = 6.44\ \Omega$
7. (a) $I = 10$ A; (b) $R = 10\ \Omega$
9. (a) 5.02×10^5 J; (b) 278.9 W; (c) 23.2 A; (d) $0.52\ \Omega$
11. (a) $R_{eq} = 1730$ W; (b) $I = 6.9$ mA
13. 10% increase
15. $I_{10\,k\Omega} = 0.68$ mA
17. $t_{1/2} = 0.693RC$
19. (a) 0.5 MΩ; (b) $\tau = 100$ s; (c) $\tau = 25$ s
21. (a) 5.5×10^5 V; (b) 0.91 F; (c) 273.3 s; (d) 20,000; (e) 0.52 h; (f) 0.9 million

Chapter 17

MC

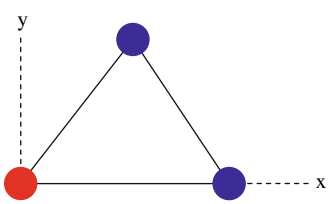
1. A
3. B
5. B
7. C
9. A
11. D
13. D
15. A
17. D
19. A

P

1. $v = 6.25 \times 10^5$ m/s
3. $v = 1.15 \times 10^7$ m/s
5. $B = 6.75 \times 10^{-4}$ T
7. (a) $v = 8.4 \times 10^6$ m/s; (b) $B = 0.64$ mT; (c) $1.1 \times 10^8 s^{-1}$; (d) $f = 1.8 \times 10^7$ Hz and $T = 5.6 \times 10^{-8}$ s
9. $$F_B = qvB = m\frac{v^2}{r} \rightarrow m = \frac{qrB}{v} \quad (1)$$

   and $$qV = \frac{1}{2}mv^2 \rightarrow v^2 = \frac{2qV}{m} \quad (2)$$

   Squaring (1) and *substituting* (2) gives:

   $$m = \frac{q}{2V}(r^2 B^2)$$
11. $\tau_{max} = 1.96 \times 10^{-3}$ Nm; $\tau_{min} = 0$ Nm
13. $\mu = 2.51\ Am^2$
15. (a) $B = 8 \times 10^{-7}$ T North; (b) 4×10^{-7} T East
17. 

Assuming that blue represents currents flowing up out of the page and red represents current flowing down into the page, we have $B = 6.9 \times 10^{-4}$ T at $\phi = -60°$ below the $+x$-axis.

19. (b) $m = 160$ mg
23. (a) $B = 6.9 \times 10^{-4}$ T vertically down; (b) $F = 7.2 \times 10^{-3}$ N to the right in the horizontal plane of the rail gun; (c) $a = 1.45$ m/s^2; (d) $v = 1.7$ m/s; (e) $I = 1500$ A
25. a. $B = \dfrac{\mu_0 I}{2\pi r}$; $r \geq R$ b. $B = \left(\dfrac{\mu_0 I}{2\pi R^2}\right) r$; $r < R$
27. Using the same figure as for Problem 17 above, we have $B = 7.59 \times 10^{-6}$ T at $\phi = -60°$ below the positive x-axis

Chapter 18

MC

1. B
3. A
5. B
7. A
9. D
11. D
13. C
15. D

P

1. $\varepsilon = -1.57 \times 10^{-13}$ V
3. $\varepsilon\,(t) = 30.2$ V sin $(120\pi t)$; $\varepsilon_{max} = 30.2$ V
5. $\varepsilon \sim 1 \times 10^{-14}$ V
7. $v = 750$ m/s ~ 1690 mph ~ Mach 2
9. $\varepsilon = 115$ kV, $I = 10550$ A
11. $E = 2.4$ N/C
13. $\omega = 5.31$ rad/s
15. $n_- = 1.00073\ n_+$
17. For protons $f_{res} = 52$ MHz; for ^{13}C $f_{res} = 13$ MHz

Chapter 19

MC

1. D
3. C
5. D
7. B
9. C
11. C
13. A
15. B
17. A
19. C

P

1. $E = 60$ N/C
3. $I = 29.8$ W/m^2
5. If the force doubles then the intensity must double
7. (a) $S = 0.5\ S_0$; (b) 12.5% of S_0 transmitted

9. As the electric field vector sweeps around at a constant rate, the intensity will be a maximum (equal to I_0) when the electric field vector is parallel to the transmission axis of the polarizer and a minimum (equal to 0) when the electric field vector is perpendicular to the transmission axis. Thus the intensity fluctuates in a periodic way as the electric field sweeps around.

11. $P_{\text{detector}} = 2.24 \times 10^{-4}$ W
13. 4×10^{16} m
15. 5.5×10^{14} Hz
17. (a) Number of photons $= 1.33 \times 10^{20}$; (b) 5×10^{9} W; (c) 50 W
19. $\lambda = (432 \pm 19) \times 10^{-9}$ m
21. (a) $E_{\text{photon}} = 3.74 \times 10^{-19}$ J; $p = 1.25 \times 10^{-27}$ kgm/s; (b) Number of photons $= 1.34 \times 10^{19}$; (c) $F = 16.7$ N
23. $c = 3.30 \times 10^{-4}$ M and 89% of the light is transmitted
25. 22% increase

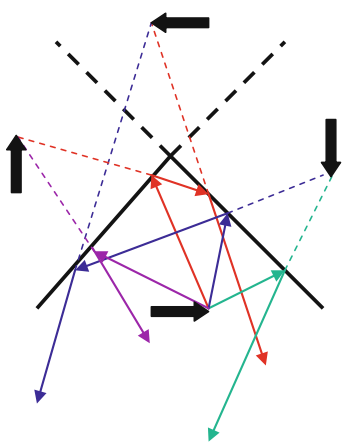
Chapter 20

MC

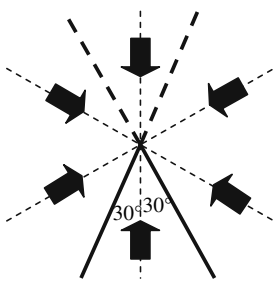
1. C
3. C
5. B
7. C
9. B
11. B
13. B
15. C
17. C

P

1.



3.



For n a small integer, there are $n - 1$ images formed.

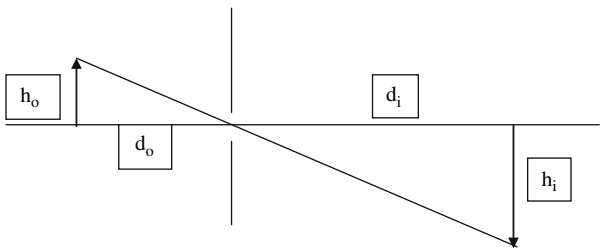
5. (a) $\theta_1 = 0.72°$ $\theta_2 = 0.72°$; (b) $h_{01}/h_{02} = 2$ and $\theta_1/\theta_2, = 2$; (c) From the geometry and because the angles involved are small, $\theta_1'/\theta_2' = h_{02}/h_{01}$ independent of the distances involved.
7. $n_1 \sin \theta_1 = n_2 \sin \theta_2 = n_1 \sin \theta_3$ or $\theta_1 = \theta_3$; $d = 0.39$ cm
9. $d_0 = 0.67$ cm and the image is real and inverted
11. $h_i = 0.25$ m and yes
13. $d_i = -5$ cm, $M = 1.66$ and the image is virtual and erect
15. $r = 9.1$ feet
17. $\theta = 56.2°$
19. (a) $\theta = 44.4°$; (b) $\theta = 8.1°$
21. (a) $\theta_{2blue} = 29.8°$ and $\theta_{2red} = 30.2°$; (b) when the light exits the prism it will bend away from the normal resulting in a rainbow of colors with blue refracted the most and red the least.
23. (a) 96% transmitted; (b) 98% transmitted; (c) 97.6% transmitted
25. 99% lost

Chapter 21

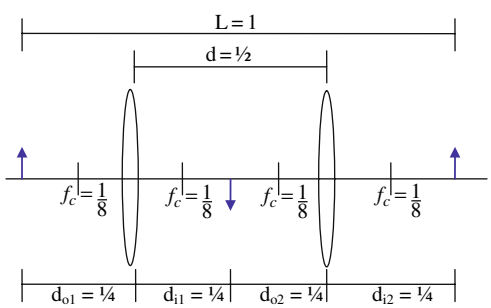
MC

1. B
3. D
5. B
7. A
9. B
11. B
13. B
15. D
17. C

P

1. $P = 4.2$ diopters
3. 

The image is inverted with respect to the original object. The magnification is the ratio of the image to object distances, $h_i/h_0 = d_i/d_0$.
5. (a) $f = 4$ cm; (b) $x = 3.5$ mm
7. (a) $f = 0.33$ m; (b) $v_{screen} = 12.5$ cm/s
9. (a) $d_i = 0.107$ m; (b) error = 7%
11. (a) $f = 0.25$ m and $d_0 = 0.50$ m; (b) The relay system is shown below where all distances are in meters.



13. $M = \left[\dfrac{L - f_e}{f_0}\right]\left[\dfrac{N}{N - f_e}\right] \sim \dfrac{25\text{ cm}\,L}{f_0\,(25\text{ cm} - f_e)}$

Chapter 22

MC

1. D
3. A
5. B
7. B
9. D
11. C
13. A
15. C

P

1. 300 nm
3. (a) $t_{red} = 119$ nm; (b) $t_{green} = 96.8$ nm
5. $t = 180.5$ nm
7. 5 fringes visible
9. (a) $l = 1.22 \times 10^{-10}$ m; (b) width = 4.8 mm; (c) 7 fringes
11. (a) $L = 4.47$ m; (b) $h_i = 6.71$ μm; (c) $f_{obj} = 17$ mm
13. 5 fringes
15. distance = 6190 m
17. (a) $v = 6.9 \times 10^5$ m/s; (b) $\lambda = 5.7 \times 10^{-13}$ m; (c) $\Delta y_m = 2.87$ μm; (d) $\Delta y_m' = 11.4$ μm; (e) 7

Chapter 23

MC

1. B
3. B
5. D
7. D
9. D
11. C
13. D

P

1. $\%_{new} = 28.8\%$
3. (a) $\theta = \tan^{-1}(E_{oy}/E_{ox})$ from x-axis; (b) 45°; (c) 18.4°
5. path difference = 0.01 cm; 0.94 rad
7. $d = 0.148$ nm
9. $\varepsilon = 1.34\text{ m}^{-1}$
11. $CT\,\# = 400$

Chapter 24

MC

1. C
3. C
5. C
7. C
9. A
11. B
13. C

15. C
17. B

P

1. (a) $\gamma = 1.67$; $p = 4 \times 10^8$ kgm/s; $E = 1.5 \times 10^{17}$ J; (b) $\gamma = 2.29$; $p = 6.19 \times 10^8$ kgm/s; $E = 2.07 \times 10^{17}$ J; (c) $\gamma = 3.20$; $p = 9.13 \times 10^8$ kgm/s; $E = 2.88 \times 10^{17}$ J; (d) $\gamma = 7.09$; $p = 2.11 \times 10^9$ kgm/s; $E = 6.38 \times 10^{17}$ J; (e) $\gamma = 22.4$; $p = 6.70 \times 10^9$ kgm/s; $E = 2.0 \times 10^{18}$ J
3. Starting from KE $= \gamma mc^2 - mc^2$ and using the binomial expansion on γ for $v << c$ leads to $(1/2)mv^2$
5. (a) $E = 8.20 \times 10^{-14}$ J $= 0.51$ MeV; (b) $\lambda = 0.0024$ nm
7. (a) $p = 6.4 \times 10^{-22}$ kgm/s; (b) $\lambda = 1.04 \times 10^{-12}$ m; (c) $f = 2.90 \times 10^{20}$ Hz
9. $\Phi = 2.29$ eV
11. (a) $\lambda = 7.7 \times 10^{-13}$ m; $p = 8.53 \times 10^{-22}$ kgm/s; (b) $\dfrac{1}{E'} = \dfrac{1}{E} + \dfrac{(1 - \cos\phi)}{m_e c^2}$; (c) $E' = 0.755$ MeV; (d) 0.845 MeV; (e) $v = 0.926c$
13. $E = 0.104$ MeV
15. #/s $= 3.2 \times 10^{15}$
17. $\Delta E = (2n - 1)\dfrac{h^2}{8mL^2}$.
19. $v_{\text{min}} = 7.28 \times 10^6$ m/s
21. $\Delta p_y = 5.53 \times 10^{-24}$ kgm/s
23. $v_{\text{max}} = 1.79 \times 10^7$ m/s

Chapter 25

MC

1. A
3. A
5. D
7. C
9. B
11. B
13. C
15. C
17. C
19. D
21. B

P

3. $E = 2.20 \times 10^{-19}$ J $= 1.38$ eV
5. Units of R are given as m^{-1} and, substituting in the known values, R $= 1.09 \times 10^7$ m^{-1}
7. For an M shell there can be 18 total electrons. m_s can take the values $\pm 1/2$ and m_ℓ can take the values of $\{0\}$ for $\ell = 0$, $\{-1,0,1\}$ for $\ell = 1$, and $\{-2,-1,0,1,2\}$ for $\ell = 2$
9. The angles (in degrees) are; 39.2; 75; 105; 140.8
11. (a) $E_{\text{rot1}} = 0$; $E_{\text{rot2}} = 6.67 \times 10^{-23}$ J; $E_{\text{rot3}} = 2.00 \times 10^{-22}$ J; $E_{\text{vib1}} = 1.66 \times 10^{-20}$ J; $E_{\text{vib2}} = 4.98 \times 10^{-20}$ J; (b) for (ℓ,m) where $\ell = 0$, 1, 2 and m = 0, 1 we get the following energies for the states:

(l,m)	*E (eV)*
0,1	0.104
0,2	0.311
1,1	0.108
1,2	0.317
2,1	0.496
2,2	0.704

(c), (d) The allowed transitions and wavelengths are given by:

transition	ΔE *(eV)*	λ *(μm)*
(2,2) to (1,1)	0.596	2.09
(2,1) to (1,2)	0.179	6.94
(1,2) to (0,1)	0.213	5.84
(0,2) to (1,1)	0.203	6.12

13. $\lambda = 4.1 \times 10^{-7}$ m = 410 nm
15. (a) Power = 4.0×10^{12} W = 4 TW; (b) Intensity = 5.66×10^{23} W/m^2; (c) Area = 0.322×10^{-6} m^2

Chapter 26

MC

1. D
3. B
5. D
7. C
9. D
11. D
13. C
15. A

P

1. (a) $r_{lA}/r_H = 5.92$; (b) $SA_{lA}/SA_H = 35.1$; (c) $V_{lA}/V_H = 208$
3. $r = 13.4$ km
5. $Q = 4.09$ MeV
7. Let m_α = mass of helium nucleus; m_d = mass of daughter atom; m_p = mass of parent atom. From conservation of momentum we have $0 = -m_d v_d + m_\alpha v_\alpha$. Whereas from conservation of energy we get $m_p c^2 = m_d c^2 + m_\alpha c^2 + 1/2\, m_d v_d^2 + 1/2\, m_\alpha v_\alpha^2$. Solving for v_d from conservation of momentum and substituting into the equation for conservation of energy produces the desired result. For the decay of uranium we find the kinetic energy of the alpha particle to be 4.21 MeV.
9. $I_{\text{photocurrent}} = 1$ nA
11. $m = 7.09$ μg; 58 years
13. (a) 0.0866 days^{-1}; (b) 22 days
15. (a) 4.22×10^8 Bq; (b) $m = 0.153$ μg
17. 383.4 years
19. 2.39×10^{11} s = 7460 years; the time is too short to be a dinosaur bone.
21. 0.5 Gy
23. $\tau_B = 18.7$ days
25. 0.429 MeV + 4.965 MeV + 5.394 MeV + 13.907 MeV = Q = 24.7 MeV

Figure Credits

Figure 1.1 Courtesy © Dennis Kunkel Microscopy, Inc.
Figure 1.2 Courtesy David Goodsell
Figure 1.4 Courtesy Veeco Instruments
Figure 1.6 Courtesy of Alexander Lyubartsev, Stockholm University
Figure 1.7 Courtesy Barbara Danowski, Union College
Figure 2.10 Courtesy National Institute of Standards and Technology
Figure 2.12 Courtesy National Aeronautics and Space Administration
Figure 3.4 Courtesy Matt Amengual
Figure 3.5 Courtesy Mammoth Mountain
Figure 3.6 Right, Courtesy Big Sky Fishing.com
Figure 3.7 Courtesy Stéphane Levet
Figure 3.15 Courtesy Brian Hoffmann, Park University
Figure 3.19 Courtesy Veeco Instruments
Figure 3.20 Left and Right courtesy of J. E. Heuser, Washington University School of Medicine
Figure 4.1 Courtesy Matt Amengual
Figure 4.5 Courtesy Christine Hogan
Figure 4.13 Courtesy National Renewable Energy Laboratory
Figure 5.17 Left, Courtesy www.viewCalgary.com. Center, Courtesy Steven Wolf. Right: Courtesy Mammoth Mountain Ski Area
Figure 5.18 Courtesy IBM Research, Almaden Research Center. Unauthorized use not permitted.
Figure 5.21 Courtesy Christine Hogan
Figure 5.22 Courtesy Ian Fetterley, Victoria, Canada
Figure 5.23 Courtesy Christine Hogan
Figure 5.27 Courtesy Cornell University
Figure 6.2 Courtesy National Aeronautics and Space Administration
Figure 6.4 Left and Right, Courtesy National Oceanic and Atmospheric Administration. Center, Courtesy National Aeronautics and Space Administration
Figure 6.5 Courtesy National Oceanic and Atmospheric Administration
Figure 6.10 Courtesy National Aeronautics and Space Administration
Figure 6.11 Top, Courtesy Christine Hogan. Bottom, Courtesy Florida Today
Figure 7.19 Courtesy Cindy Withington Newman
Figure 7.22 Courtesy, Ueli Aebi, Müller Institute of Microscopy, Basel
Figure 7.25 Left, Courtesy of Tatsuo Ushiki, Niigata University. Center and Right, Courtesy Ueli Aebi, Müller Institute of Microscopy, Basel
Figure 7.28 Left, from A. Vander et al., *Human Physiology*, 2004, reproduced with permission of The McGraw-Hill Companies
Figure 8.3 Courtesy Remy Pujol, CRIC, INSERM and University of Montpellier
Figure 8.6 Courtesy U.S. Geological Survey
Figure 8.12 Courtesy BI Ogungbo, Newcastle General Hospital, UK
Figure 8.13 Courtesy Liz Soilleux, Addenbrooke's Hospital, UK
Figure 8.21 Right, Courtesy Protech Environmental Services

Figure 9.8 Left, Courtesy Z. Shao, University of Virginia Medical School and http://hms.medweb.harvard.edu/HS_Heme/ AtlasTOC.htm. Right, Courtesy Dr. Barbara Safiejko-Mroczka and Dr. Paul B. Bell, Dept Zoology, University of Oklahoma
Figure 9.12 Courtesy of Wesley Norman, Georgetown University, http://my webpages. comcast.net/wnor/homepage.htm
Figure 9.15 Courtesy David Hu, MIT
Figure 10.1 Left, Courtesy Anthony Cook, Griffith Observatory. Right, Courtesy Washington State Department of Transportation
Figure 10.6 Courtesy U.S. Geological Survey
Figure 10.19 Courtesy, Jim Richardson, Scotland 1986
Figure 11.1 Courtesy Peter Belafsky
Figure 11.9 Left, Courtesy Bruce Duncan, Union College. Right, © 1992, 1997 James Boyk, Caltech Music Lab. All rights reserved. Used by permission.
Figure 11.11 Top, Courtesy Photo Researchers Inc. Bottom, Courtesy *Journal of the American Medical Association* and Richard Ehman
Figure 11.12 Courtesy Joe Wolfe, University of New South Wales
Figure 11.15 Courtesy Daniel Russell, Kettering University
Figure 11.16 Courtesy, Remy Pujol et al., CRIC, INSERM and University Montpellier
Figure 11.17 *The Far Side*® by Gary Larson © 1990 FarWorks, Inc. All Rights Reserved. Used with permission.
Figure 11.18 Courtesy Remy Pujol, CRIC, INSERM and University of Montpelier
Figure 11.19 Courtesy Remy Pujol, CRIC, INSERM and University of Montpelier
Figure 11.20 Courtesy Remy Pujol, CRIC, INSERM and University of Montpelier
Figure 11.21 Courtesy Remy Pujol, CRIC, INSERM and University of Montpelier
Figure 11.28 Courtesy Ian Burgess, Glenn Watson, and Mater Imaging
Figure 11.29 Courtesy of GE Medical Systems
Figure 11.30 Courtesy GE Medical Imaging
Figure 12.2 Courtesy National Aeronautics and Space Administration
Figure 12.3 Courtesy Rick Adler www.RESnapshot.com
Figure 12.4 Right, Courtesy Alan Nazerian, MD
Figure 12.6 Courtesy National Oceanic and Atmospheric Administration
Figure 12.12 Center, Courtesy Ken Ackerman. Right, Courtesy Mick Batt, Fairbanks Alaska
Figure 12.14 Courtesy of P.K.J. Kinnunen and the *Biophysical Journal* Vol. 78, No. 5, p. 2459–2469, May 2000
Figure 12.16 Courtesy Matt Amengual
Figure 12.19 Courtesy Lynn McCutchen, Kilgore College
Figure 12.20 Courtesy Fresenius Medical Care
Figure 12.21 Right, Courtesy of A. Anderson, Union College
Figure 12.22 Courtesy Sierra Pacific, www.x20.org
Figure 14.1 Courtesy National Oceanic and Atmospheric Administration
Figure 14.3 Courtesy Lawrence Berkeley National Laboratory
Figure 14.7 Courtesy John Miles
Figure 14.13 Courtesy of U.S. Geological Survey
Figure 14.24 Courtesy J. Kalinowski, Institute for Genome Research, Bielefeld University
Figure 15.2 Courtesy Knott's Berry Farm
Figure 15.10 Courtesy Barry Honig, Columbia University
Figure 15.17 Courtesy Barry Honig, Columbia University
Figure 15.18 Courtesy David Goodyear, Department of Physics and Physical Oceanography, Memorial University of Newfoundland
Figure 15.23 Courtesy Peter Bond, University of Oxford
Figure 15.24 Courtesy Tim Smith
Figure 15.25 Courtesy Eric Chudler, University of Washington

Figure 15.26 Courtesy Vernier Software, Inc. and Qubit Systems, Inc.
Figure 15.27 Courtesy Peter van Hese, Ghent University
Figure 16.17 Courtesy of Thomas Euler, Max Planck Institute for Medical Research, Heidelberg
Figure 16.22 Redrawn from *Cellular Biophysics* by T.F. Weiss with permission from MIT Press
Figure 16.24 From *Cellular Biophysics* by T.F. Weiss with permission from MIT Press
Figure 16.26 From *Molecular Biology of the Cell* by Bruce Alberts, et al., with permission
Figure 17.1 Courtesy Drs. Timothy St. Pierre and Ralph James of the Biophysics program in the School of Physics at the University of Western Australia
Figure 17.6 Courtesy J.-L. Aubagnac, Université Montpellier, France
Figure 17.9 Courtesy National Aeronautics and Space Administration
Figure 17.17 Courtesy CERN
Figure 17.21 Courtesy NewsCast
Figure 18.9 Courtesy CTF Systems Inc.
Figure 18.11 Courtesy CTF Systems, Inc.
Figure 18.15 Courtesy Jim Adrian, Union College
Figure 18.18 Courtesy R. R. Ernst, ETH Zurich and the American Institute of Physics
Figure 18.19 Courtesy Jeff Dunham, Middlebury College
Figure 18.20 Courtesy GE Medical Systems
Figure 18.21 Courtesy Joseph Hornak
Figure 18.22 Courtesy, W. Edelstein and T. Dixon, GE Global Research, Schenectady, NY
Figure 18.23 Courtesy National Library of Medicine
Figure 19.8 Courtesy D. DuBois, Desert Research Institute
Figure 19.11 Left, Courtesy Manfred Schliwa. Right, Courtesy Richard Cole, Wadsworth Center
Figure 19.12 Courtesy, Manfred Schliwa
Figure 19.13 Courtesy Steven M. Block and the *Biophysical Journal* Vol. 77, No. 5, p. 2856–2863, November 1999
Figure 19.14 Courtesy Bar-Ziv et al., *Biophysical Journal*, Vol. 75, No. 1, p. 294–320, July 1998
Figure 19.15 Courtesy Manfred Schliwa
Figure 19.29 Bottom two, Courtesy Jeremy Newman
Figure 19.31 Courtesy, Barbara Danowski, Union College
Figure 20.1 Left, Courtesy National Oceanic and Atmospheric Administration. Right, Courtesy National Aeronautics and Space Administration/Goddard Space Flight Center
Figure 20.20 Courtesy The Education Group (www.physicsdemos.com)
Figure 20.21 Courtesy Pacific Northwest National Laboratory
Figure 20.22 Left, Courtesy LSRO – EPFL
Figure 20.23 Courtesy Hans Bjorknas, Gastrolab, Finland
Figure 21.10 1. Courtesy National Oceanic and Atmospheric Administration. 2. Courtesy C.-K. Shene, Michigan Technological University. 3. Courtesy Scott Kahn. 4. Courtesy American Academy of Opthamology
Figure 21.16 Top, Courtesy Thomas Euler, Max Planck Institute for Medical Research, Heidelberg. Bottom, Courtesy Bio-Rad
Figure 21.17 The images were taken at the University of Rochester and are provided courtesy of Austin Roorda and David Williams
Figure 21.18 Courtesy of Dr. Roger Wagner, Biology Department, University of Delaware
Figure 21.19 Courtesy P. Hargrave, redrawn from *Bioessays* 15, 1, 1993

Figure 22.6 Courtesy W. Brent Daniels and Maarten Rutgers
Figure 22.8 Courtesy Monty Reichert and the *Biophysical Journal* Vol. 78 No. 4, p. 1725–1735, April 2000
Figure 22.9 Courtesy David Elfstrom and istockphoto.com
Figure 22.19 Courtesy Chris Jones, Union College
Figure 22.20 Courtesy Chris Jones, Union College
Figure 22.26 Courtesy Chris Jones, Union College
Figure 23.2 Courtesy *Molecular Expressions*
Figure 23.4 Courtesy, Barbara Danowski, Union College
Figure 23.5 Courtesy *Molecular Probes*
Figure 23.7 Courtesy *Molecular Expressions*
Figure 23.9 Courtesy *Molecular Expressions*
Figure 23.10 Courtesy Confocal Microscope Facility, Department of Anatomy and Cellular Biology, Tufts University School of Medicine
Figure 23.11 Courtesy Imaging Technology Group, Beckman Inst. For Advanced Science and Technology, University of Illinois at Urbana-Champaign
Figure 23.13 Courtesy S. Madihally, OSU
Figure 23.15 Bottom two, Courtesy Bernhard Rupp, Lawrence Livermore National Laboratory
Figure 23.19 Courtesy George Smith, Union College
Figure 23.20 Courtesy Ueli Aebi, Muller Institute for Microscopy, Basel
Figure 23.24 Courtesy GE Medical Systems
Figure 23.26 Courtesy GE Medical Systems
Figure 24.1 Courtesy garmin.de
Figure 24.6 Courtesy MATTER, University of Liverpool
Figure 24.11 Right, Courtesy of Keith O'Doherty, Dept. of Chemistry, University of California at Los Angeles
Figure 24.12 Left, Courtesy of IBM Research, Almaden Research Center. Unauthorized use not permitted. Right, Courtesy of W. Schonert, GSI, Germany
Figure 25.11 Courtesy Jim Adrian, Chemistry Department, Union College
Figure 25.18 Left, Courtesy J&K Laser Productions. Right, Courtesy Jon Huffman, photographer, and Coherent Laser
Figure 25.25 Left, Courtesy Seyffie Maleki, Union College. Right, Courtesy Samual M. Goldwasser, from Sam's Laser FAQ, www.repairfaq.org/sam/lasersam.htm
Figure 25.26 Courtesy *Forschungszentrum Karlsruhe*
Figure 25.27 Courtesy Time Life Pictures/ Getty Images
Figure 25.28 Courtesy Emmett Leith
Figure 25.29 Left, Courtesy Chip Fogg, DuPont. Right, Courtesy BMW North America
Figure 26.7 Courtesy Lawrence Berkeley National Laboratory
Figure 26.13 Courtesy of GE Medical Systems
Figure 26.14 Courtesy GE Medical Systems
Figure 26.15 Courtesy Brookhaven National Laboratory
Figure 26.16 Courtesy Princeton Plasma Physics Laboratory
Figure 26.17 Courtesy University of California/ Lawrence Livermore National Laboratory

Index

B

D

F

N

O

S