Newtonian Physics

Benjamin Crowell

Book 1 in the **Light and Matter** series of introductory physics textbooks
www.lightandmatter.com

Newtonian Physics

The **Light and Matter** series of introductory physics textbooks:

1 Newtonian Physics

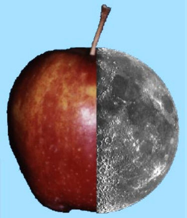
2 Conservation Laws

3 Vibrations and Waves

4 Electricity and Magnetism

5 Optics

6 The Modern Revolution in Physics



Newtonian Physics

Benjamin Crowell

www.lightandmatter.com

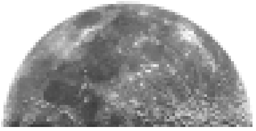
Light and Matter

Fullerton, California
www.lightandmatter.com

Edition 2.1
rev. 2001-04-09

ISBN 0-9704670-1-X

To Paul Herrschaft and Rich Muller.

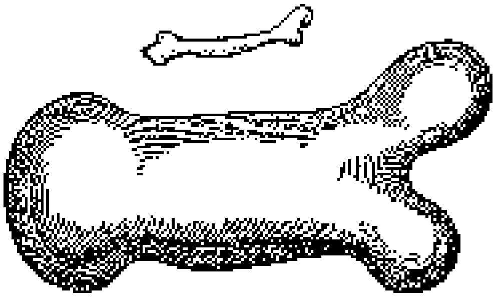


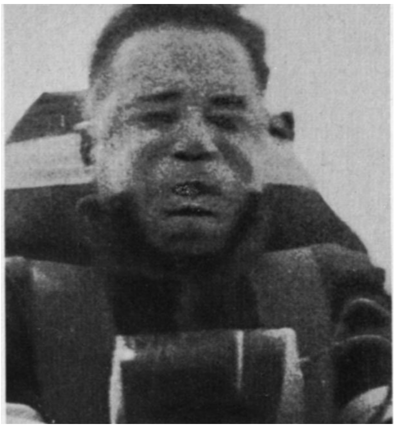
Brief Contents

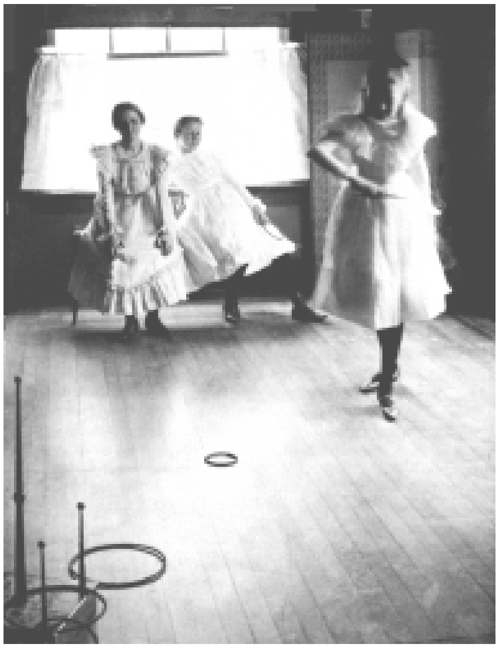
Motion in One Dimension

Motion in Three Dimensions

Contents







Preface

Why a New Physics Textbook?

We assume that our economic system will always scamper to provide us with the products we want. Special orders don't upset us! I want my MTV! The truth is more complicated, especially in our education system, which is paid for by the students but controlled by the professoriate. Witness the perverse success of the bloated science textbook. The newspapers continue to compare our system unfavorably to Japanese and European education, where depth is emphasized over breadth, but we can't seem to create a physics textbook that covers a manageable number of topics for a one-year course and gives honest explanations of everything it touches on.

The publishers try to please everybody by including every imaginable topic in the book, but end up pleasing nobody. There is wide agreement among physics teachers that the traditional one-year introductory textbooks cannot in fact be taught in one year. One cannot surgically remove enough material and still gracefully navigate the rest of one of these kitchen-sink textbooks. What is far worse is that the books are so crammed with topics that nearly all the explanation is cut out in order to keep the page count below 1100. Vital concepts like energy are introduced abruptly with an equation, like a first-date kiss that comes before "hello."

The movement to reform physics texts is steaming ahead, but despite excellent books such as Hewitt's **Conceptual Physics** for non-science majors and Knight's **Physics: A Contemporary Perspective** for students who know calculus, there has been a gap in physics books for life-science majors who haven't learned calculus or are learning it concurrently with physics. This book is meant to fill that gap.

Learning to Hate Physics?

When you read a mystery novel, you know in advance what structure to expect: a crime, some detective work, and finally the unmasking of the evildoer. When Charlie Parker plays a blues, your ear expects to hear certain landmarks of the form regardless of how wild some of his notes are. Surveys of physics students usually show that they have *worse* attitudes about the subject after instruction than before, and their comments often boil down to a complaint that the person who strung the topics together had not learned what Agatha Christie and Charlie Parker knew intuitively about form and structure: students become bored and demoralized because the "march through the topics" lacks a coherent story line. You are reading the first volume of the **Light and Matter** series of introductory physics textbooks, and as implied by its title, the story line of the series is built around light and matter: how they behave, how they are different from each other, and, at the end of the story, how they turn out to be similar in some very bizarre ways. Here is a guide to the structure of the one-year course presented in this series:

1 Newtonian Physics *Matter* moves at constant speed in a straight line unless a force acts on it. (This seems intuitively wrong only because we tend to forget the role of friction forces.) Material objects can exert forces on each other, each changing the other's motion. A more massive object changes its motion more slowly in response to a given force.

2 Conservation Laws Newton's matter-and-forces picture of the universe is fine as far as it goes, but it doesn't apply to *light*, which is a form of pure energy without mass. A more powerful world-view, applying equally well to both light and matter, is provided by the conservation laws, for instance the law of conservation of energy, which states that energy can never be destroyed or created but only changed from one form into another.

3 Vibrations and Waves *Light* is a wave. We learn how waves travel through space, pass through each other, speed up, slow down, and are reflected.

4 Electricity and Magnetism *Matter* is made out of particles such as electrons and protons, which are held together by electrical forces. *Light* is a wave that is made out of patterns of electric and magnetic force.

5 Optics Devices such as eyeglasses and searchlights use *matter* (lenses and mirrors) to manipulate *light*.

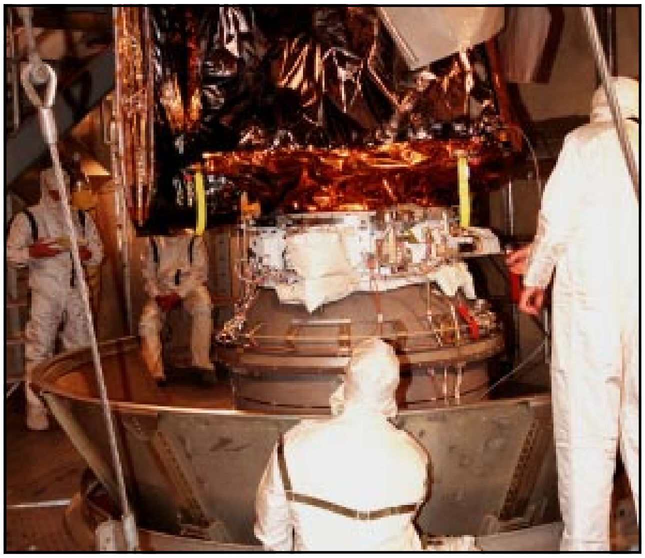
6 The Modern Revolution in Physics Until the twentieth century, physicists thought that *matter* was made out of particles and *light* was purely a wave phenomenon. We now know that both light and matter are made of building blocks that have both particle and wave properties. In the process of understanding this apparent contradiction, we find that the universe is a much stranger place than Newton had ever imagined, and also learn the basis for such devices as lasers and computer chips.

A Note to the Student Taking Calculus Concurrently

Learning calculus and physics concurrently is an excellent idea — it’s not a coincidence that the inventor of calculus, Isaac Newton, also discovered the laws of motion! If you are worried about taking these two demanding courses at the same time, let me reassure you. I think you will find that physics helps you with calculus while calculus deepens and enhances your experience of physics. This book is designed to be used in either an algebra-based physics course or a calculus-based physics course that has calculus as a corequisite. This note is addressed to students in the latter type of course.

It has been said that critics discuss art with each other, but artists talk about brushes. Art needs both a “why” and a “how,” concepts as well as technique. Just as it is easier to enjoy an oil painting than to produce one, it is easier to understand the concepts of calculus than to learn the techniques of calculus. This book will generally teach you the *concepts* of calculus a few weeks before you learn them in your math class, but it does not discuss the *techniques* of calculus at all. There will thus be a delay of a few weeks between the time when a calculus application is first pointed out in this book and the first occurrence of a homework problem that requires the relevant technique. The following outline shows a typical first-semester calculus curriculum side-by-side with the list of topics covered in this book, to give you a rough idea of what calculus your physics instructor might expect you to know at a given point in the semester. The sequence of the calculus topics is the one followed by **Calculus of a Single Variable**, 2nd ed., by Swokowski, Olinick, and Pence.

chapters of this book	topics typically covered at the same point in a calculus course
0-1 introduction	review
2-3 velocity and acceleration	limits
4-5 Newton's laws	the derivative concept
6-8 motion in 3 dimensions	techniques for finding derivatives; derivatives of trigonometric functions
9 circular motion	the chain rule
10 gravity	local maxima and minima
chapters of **Conservation Laws**	
1-3 energy	concavity and the second derivative
4 momentum	
5 angular momentum	the indefinite integral
chapters of **Vibrations and Waves**	
1 vibrations	the definite integral
2-3 waves	the fundamental theorem of calculus



The Mars Climate Orbiter is prepared for its mission. The laws of physics are the same everywhere, even on Mars, so the probe could be designed based on the laws of physics as discovered on earth.

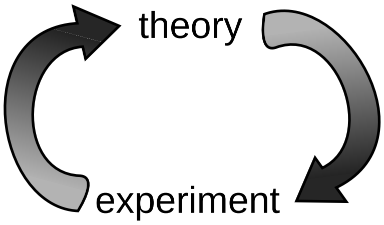
There is unfortunately another reason why this spacecraft is relevant to the topics of this chapter: it was destroyed attempting to enter Mars' atmosphere because engineers at Lockheed Martin forgot to convert data on engine thrusts from pounds into the metric unit of force (newtons) before giving the information to NASA. Conversions are important!

0 Introduction and Review

If you drop your shoe and a coin side by side, they hit the ground at the same time. Why doesn't the shoe get there first, since gravity is pulling harder on it? How does the lens of your eye work, and why do your eye's muscles need to squash its lens into different shapes in order to focus on objects nearby or far away? These are the kinds of questions that physics tries to answer about the behavior of light and matter, the two things that the universe is made of.

0.1 The Scientific Method

Until very recently in history, no progress was made in answering questions like these. Worse than that, the *wrong* answers written by thinkers like the ancient Greek physicist Aristotle were accepted without question for thousands of years. Why is it that scientific knowledge has progressed more since the Renaissance than it had in all the preceding millennia since the beginning of recorded history? Undoubtedly the industrial revolution is part of the answer. Building its centerpiece, the steam engine, required improved techniques for precise construction and measurement. (Early on, it was considered a major advance when English machine shops learned to build pistons and cylinders that fit together with a gap narrower than the thickness of a penny.) But even before the industrial revolution, the pace of discovery had picked up, mainly because of the introduction of the modern scientific method. Although it evolved over time, most scientists today would agree on something like the following list of the basic principles of the scientific method:



(1)*Science is a cycle of theory and experiment.* Scientific theories are created to explain the results of experiments that were created under certain conditions. A successful theory will also make new predictions about new experiments under new conditions. Eventually, though, it always seems to happen that a new experiment comes along, showing that under certain conditions the theory is not a good approximation or is not valid at all. The ball is then back in the theorists' court. If an experiment disagrees with the current theory, the theory has to be changed, not the experiment.

(2)*Theories should both predict and explain.* The requirement of predictive power means that a theory is only meaningful if it predicts something that can be checked against experimental measurements that the theorist did not already have at hand. That is, a theory should be testable. Explanatory value means that many phenomena should be accounted for with few basic principles. If you answer every "why" question with "because that's the way it is," then your theory has no explanatory value. Collecting lots of data without being able to find any basic underlying principles is not science.

(3)*Experiments should be reproducible.* An experiment should be treated with suspicion if it only works for one person, or only in one part of the world. Anyone with the necessary skills and equipment should be able to get the same results from the same experiment. This implies that science transcends national and ethnic boundaries; you can be sure that nobody is doing actual science who claims that their work is "Aryan, not Jewish," "Marxist, not bourgeois," or "Christian, not atheistic." An experiment cannot be reproduced if it is secret, so science is necessarily a public enterprise.

A satirical drawing of an alchemist's laboratory. H. Cock, after a drawing by Peter Brueghel the Elder (16th century).

As an example of the cycle of theory and experiment, a vital step toward modern chemistry was the experimental observation that the chemical elements could not be transformed into each other, e.g. lead could not be turned into gold. This led to the theory that chemical reactions consisted of rearrangements of the elements in different combinations, without any change in the identities of the elements themselves. The theory worked for hundreds of years, and was confirmed experimentally over a wide range of pressures and temperatures and with many combinations of elements. Only in the twentieth century did we learn that one element could be transformed into one another under the conditions of extremely high pressure and temperature existing in a nuclear bomb or inside a star. That observation didn't completely invalidate the original theory of the immutability of the ele-ments, but it showed that it was only an approximation, valid at ordinary temperatures and pressures.

Self-Check

A psychic conducts seances in which the spirits of the dead speak to the participants. He says he has special psychic powers not possessed by other people, which allow him to "channel" the communications with the spirits. What part of the scientific method is being violated here? [Answer below.]

The scientific method as described here is an idealization, and should not be understood as a set procedure for doing science. Scientists have as many weaknesses and character flaws as any other group, and it is very common for scientists to try to discredit other people's experiments when the results run contrary to their own favored point of view. Successful science also has more to do with luck, intuition, and creativity than most people realize, and the restrictions of the scientific method do not stifle individuality and self-expression any more than the fugue and sonata forms

If only he has the special powers, then his results can never be reproduced.

Science is creative.

stifled Bach and Haydn. There is a recent tendency among social scientists to go even further and to deny that the scientific method even exists, claiming that science is no more than an arbitrary social system that determines what ideas to accept based on an in-group's criteria. I think that's going too far. If science is an arbitrary social ritual, it would seem difficult to explain its effectiveness in building such useful items as airplanes, CD players and sewers. If alchemy and astrology were no less scientific in their methods than chemistry and astronomy, what was it that kept them from producing anything useful?

Discussion Questions

Consider whether or not the scientific method is being applied in the following examples. If the scientific method is not being applied, are the people whose actions are being described performing a useful human activity, albeit an unscientific one?

A. Acupuncture is a traditional medical technique of Asian origin in which small needles are inserted in the patient's body to relieve pain. Many doctors trained in the west consider acupuncture unworthy of experimental study because if it had therapeutic effects, such effects could not be explained by their theories of the nervous system. Who is being more scientific, the western or eastern practitioners?

B. Goethe, a famous German poet, is less well known for his theory of color. He published a book on the subject, in which he argued that scientific apparatus for measuring and quantifying color, such as prisms, lenses and colored filters, could not give us full insight into the ultimate meaning of color, for instance the cold feeling evoked by blue and green or the heroic sentiments inspired by red. Was his work scientific?

C. A child asks why things fall down, and an adult answers "because of gravity." The ancient Greek philosopher Aristotle explained that rocks fell because it was their nature to seek out their natural place, in contact with the earth. Are these explanations scientific?

D. Buddhism is partly a psychological explanation of human suffering, and psychology is of course a science. The Buddha could be said to have engaged in a cycle of theory and experiment, since he worked by trial and error, and even late in his life he asked his followers to challenge his ideas. Buddhism could also be considered reproducible, since the Buddha told his followers they could find enlightenment for themselves if they followed a certain course of study and discipline. Is Buddhism a scientific pursuit?

0.2 What Is Physics?

> Given for one instant an intelligence which could comprehend all the forces by which nature is animated and the respective positions of the things which compose it...nothing would be uncertain, and the future as the past would be laid out before its eyes.
>
> Pierre Simon de Laplace

Physics is the study of light and matter.

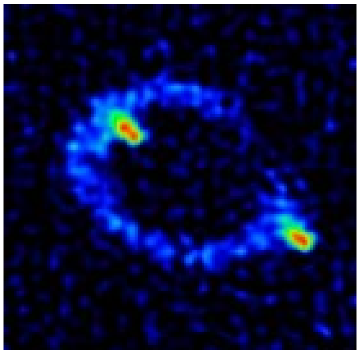
Physics is the use of the scientific method to find out the basic principles governing light and matter, and to discover the implications of those laws. Part of what distinguishes the modern outlook from the ancient mindset is the assumption that there are rules by which the universe functions, and that those laws can be at least partially understood by humans. From the Age of Reason through the nineteenth century, many scientists began to be convinced that the laws of nature not only could be known but, as claimed by Laplace, those laws could in principle be used to predict everything about the universe's future if complete information was available about the present state of all light and matter. In subsequent sections, I'll describe two general types of limitations on prediction using the laws of physics, which were only recognized in the twentieth century.

Weight is what distinguishes light from matter.

Matter can be defined as anything that is affected by gravity, i.e. that has weight or would have weight if it was near the Earth or another star or planet massive enough to produce measurable gravity. Light can be defined as anything that can travel from one place to another through empty space and can influence matter, but has no weight. For example, sunlight can influence your body by heating it or by damaging your DNA and giving you skin cancer. The physicist's definition of light includes a variety of phenomena that are not visible to the eye, including radio waves, microwaves, x-rays, and gamma rays. These are the "colors" of light that do not happen to fall within the narrow violet-to-red range of the rainbow that we can see.

Self-check

At the turn of the 20th century, a strange new phenomenon was discovered in vacuum tubes: mysterious rays of unknown origin and nature. These rays are the same as the ones that shoot from the back of your TV's picture tube and hit the front to make the picture. Physicists in 1895 didn't have the faintest idea what the rays were, so they simply named them "cathode rays," after the name for the electrical contact from which they sprang. A fierce debate raged, complete with nationalistic overtones, over whether the rays were a form of light or of matter. What would they have had to do in order to settle the issue?



This telescope picture shows two images of the same distant object, an exotic, very luminous object called a quasar. This is interpreted as evidence that a massive, dark object, possibly a black hole, happens to be between us and it. Light rays that would otherwise have missed the earth on either side have been bent by the dark object's gravity so that they reach us. The actual direction to the quasar is presumably in the center of the image, but the light along that central line don't get to us because they are absorbed by the dark object. The quasar is known by its catalog number, MG1131+0456, or more informally as Einstein's Ring.

Many physical phenomena are not themselves light or matter, but are properties of light or matter or interactions between light and matter. For instance, motion is a property of all light and some matter, but it is not itself light or matter. The pressure that keeps a bicycle tire blown up is an interaction between the air and the tire. Pressure is not a form of matter in and of itself. It is as much a property of the tire as of the air. Analogously, sisterhood and employment are relationships among people but are not people themselves.

Some things that appear weightless actually do have weight, and so qualify as matter. Air has weight, and is thus a form of matter even though a cubic inch of air weighs less than a grain of sand. A helium balloon has weight, but is kept from falling by the force of the surrounding more dense air, which pushes up on it. Astronauts in orbit around the Earth have weight, and are falling along a curved arc, but they are moving so fast that the curved arc of their fall is broad enough to carry them all the way around the Earth in a circle. They perceive themselves as being weightless because their space capsule is falling along with them, and the floor therefore does not push up on their feet.

Optional Topic

Einstein predicted as a consequence of his theory of relativity that light would after all be affected by gravity, although the effect would be extremely weak under normal conditions. His prediction was borne out by observations of the bending of light rays from stars as they passed close to the sun on their way to the Earth. Einstein also

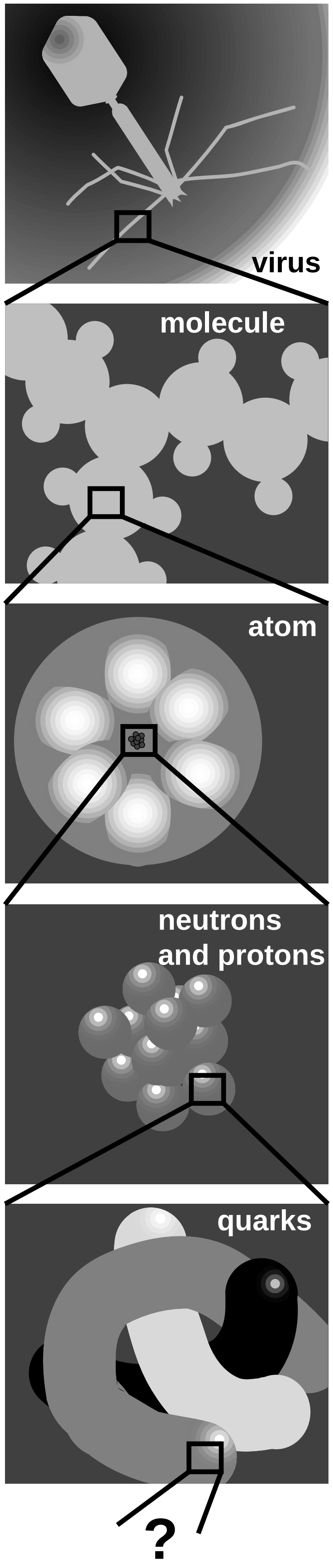
They would have had to weigh the rays, or check for a loss of weight in the object from which they were have emitted. (For technical reasons, this was not a measurement they could actually do, hence the opportunity for disagreement.)

predicted the existence of black holes, stars so massive and compact that their intense gravity would not even allow light to escape. (These days there is strong evidence that black holes exist.)

Einstein’s interpretation was that light doesn’t really have mass, but that energy is affected by gravity just like mass is. The energy in a light beam is equivalent to a certain amount of mass, given by the famous equation $E=mc^2$, where c is the speed of light. Because the speed of light is such a big number, a large amount of energy is equivalent to only a very small amount of mass, so the gravitational force on a light ray can be ignored for most practical purposes.

There is however a more satisfactory and fundamental distinction between light and matter, which should be understandable to you if you have had a chemistry course. In chemistry, one learns that electrons obey the Pauli exclusion principle, which forbids more than one electron from occupying the same orbital if they have the same spin. The Pauli exclusion principle is obeyed by the subatomic particles of which matter is composed, but disobeyed by the particles, called photons, of which a beam of light is made.

Einstein’s theory of relativity is discussed more fully in book 6 of this series.



The boundary between physics and the other sciences is not always clear. For instance, chemists study atoms and molecules, which are what matter is built from, and there are some scientists who would be equally willing to call themselves physical chemists or chemical physicists. It might seem that the distinction between physics and biology would be clearer, since physics seems to deal with inanimate objects. In fact, almost all physicists would agree that the basic laws of physics that apply to molecules in a test tube work equally well for the combination of molecules that constitutes a bacterium. (Some might believe that something more happens in the minds of humans, or even those of cats and dogs.) What differentiates physics from biology is that many of the scientific theories that describe living things, while ultimately resulting from the fundamental laws of physics, cannot be rigorously derived from physical principles.

Isolated systems and reductionism

To avoid having to study everything at once, scientists isolate the things they are trying to study. For instance, a physicist who wants to study the motion of a rotating gyroscope would probably prefer that it be isolated from vibrations and air currents. Even in biology, where field work is indispensable for understanding how living things relate to their entire environment, it is interesting to note the vital historical role played by Darwin’s study of the Galápagos Islands, which were conveniently isolated from the rest of the world. Any part of the universe that is considered apart from the rest can be called a “system.”

Physics has had some of its greatest successes by carrying this process of isolation to extremes, subdividing the universe into smaller and smaller parts. Matter can be divided into atoms, and the behavior of individual atoms can be studied. Atoms can be split apart into their constituent neutrons, protons and electrons. Protons and neutrons appear to be made out of even smaller particles called quarks, and there have even been some claims of experimental evidence that quarks have smaller parts inside them.

This method of splitting things into smaller and smaller parts and studying how those parts influence each other is called reductionism. The hope is that the seemingly complex rules governing the larger units can be better understood in terms of simpler rules governing the smaller units. To appreciate what reductionism has done for science, it is only necessary to examine a 19th-century chemistry textbook. At that time, the existence of atoms was still doubted by some, electrons were not even suspected to exist, and almost nothing was understood of what basic rules governed the way atoms interacted with each other in chemical reactions. Students had to memorize long lists of chemicals and their reactions, and there was no way to understand any of it systematically. Today, the student only needs to remember a small set of rules about how atoms interact, for instance that atoms of one element cannot be converted into another via chemical reactions, or that atoms from the right side of the periodic table tend to form strong bonds with atoms from the left side.

Discussion Questions

A. I've suggested replacing the ordinary dictionary definition of light with a more technical, more precise one that involves weightlessness. It's still possible, though, that the stuff a lightbulb makes, ordinarily called "light," does have some small amount of weight. Suggest an experiment to attempt to measure whether it does.

B. Heat is weightless (i.e. an object becomes no heavier when heated), and can travel across an empty room from the fireplace to your skin, where it influences you by heating you. Should heat therefore be considered a form of light by our definition? Why or why not?

C. Similarly, should sound be considered a form of light?

0.3 How to Learn Physics

> For as knowledges are now delivered, there is a kind of contract of error between the deliverer and the receiver; for he that delivereth knowledge desireth to deliver it in such a form as may be best believed, and not as may be best examined; and he that receiveth knowledge desireth rather present satisfaction than expectant inquiry.
>
> Sir Francis Bacon

Science is not about plugging into formulas.

Many students approach a science course with the idea that they can succeed by memorizing the formulas, so that when a problem is assigned on the homework or an exam, they will be able to plug numbers in to the formula and get a numerical result on their calculator. Wrong! That's not what learning science is about! There is a big difference between memorizing formulas and understanding concepts. To start with, different formulas may apply in different situations. One equation might represent a definition, which is always true. Another might be a very specific equation for the speed of an object sliding down an inclined plane, which would not be true if the object was a rock drifting down to the bottom of the ocean. If you don't work to understand physics on a conceptual level, you won't know which formulas can be used when.

Most students taking college science courses for the first time also have very little experience with interpreting the meaning of an equation. Consider the equation $w=A/h$ relating the width of a rectangle to its height and area. A student who has not developed skill at interpretation might view this as yet another equation to memorize and plug in to when needed. A slightly more savvy student might realize that it is simply the familiar formula $A=wh$ in a different form. When asked whether a rectangle would have a greater or smaller width than another with the same area but a smaller height, the unsophisticated student might be at a loss, not having any numbers to plug in on a calculator. The more experienced student would know how to reason about an equation involving division — if h is smaller, and A stays the same, then w must be bigger. Often, students fail to recognize a sequence of equations as a derivation leading to a final result, so they think all the intermediate steps are equally important formulas that they should memorize.

interpreting an equation

Other Books

PSSC Physics, Haber-Schaim et al., 7th ed., 1986. Kendall/Hunt, Dubuque, Iowa.

A high-school textbook at the algebra-based level. This book distinguishes itself by giving a clear, careful, and honest explanation of every topic, while avoiding unnecessary details.

Physics for Poets, Robert H. March, 4th ed., 1996. McGraw-Hill, New York.

As the name implies, this book's intended audience is liberal arts students who want to understand science in a broader cultural and historical context. Not much math is used, and the page count of this little paperback is about five times less than that of the typical "kitchen sink" textbook, but the intellectual level is actually pretty challenging.

Conceptual Physics, Paul Hewitt. Scott Foresman, Glenview, Ill.

This is the excellent book used for Physics 130 here at Fullerton College. Only simple algebra is used.

When learning any subject at all, it is important to become as actively involved as possible, rather than trying to read through all the information quickly without thinking about it. It is a good idea to read and think about the questions posed at the end of each section of these notes as you encounter them, so that you know you have understood what you were reading.

Many students' difficulties in physics boil down mainly to difficulties with math. Suppose you feel confident that you have enough mathematical preparation to succeed in this course, but you are having trouble with a few specific things. In some areas, the brief review given in this chapter may be sufficient, but in other areas it probably will not. Once you identify the areas of math in which you are having problems, get help in those areas. Don't limp along through the whole course with a vague feeling of dread about something like scientific notation. The problem will not go away if you ignore it. The same applies to essential mathematical skills that you are learning in this course for the first time, such as vector addition.

Sometimes students tell me they keep trying to understand a certain topic in the book, and it just doesn't make sense. The worst thing you can possibly do in that situation is to keep on staring at the same page. Every textbook explains certain things badly — even mine! — so the best thing to do in this situation is to look at a different book. Instead of college textbooks aimed at the same mathematical level as the course you're taking, you may in some cases find that high school books or books at a lower math level give clearer explanations. The three books listed on the left are, in my opinion, the best introductory physics books available, although they would not be appropriate as the primary textbook for a college-level course for science majors.

Finally, when reviewing for an exam, don't simply read back over the text and your lecture notes. Instead, try to use an active method of reviewing, for instance by discussing some of the discussion questions with another student, or doing homework problems you hadn't done the first time.

0.4 Self-Evaluation

The introductory part of a book like this is hard to write, because every student arrives at this starting point with a different preparation. One student may have grown up in another country and so may be completely comfortable with the metric system, but may have had an algebra course in which the instructor passed too quickly over scientific notation. Another student may have already taken calculus, but may have never learned the metric system. The following self-evaluation is a checklist to help you figure out what you need to study to be prepared for the rest of the course.

If you disagree with this statement...	**you should study this section:**
I am familiar with the basic metric units of meters, kilograms, and seconds, and the most common metric prefixes: milli- (m), kilo- (k), and centi- (c).	0.5 Basics of the Metric System
I know about the Newton, a unit of force	0.6 The Newton, the Metric Unit of Force
I am familiar with these less common metric prefixes: mega- (M), micro- (µ), and nano- (n).	0.7 Less Common Metric Prefixes
I am comfortable with scientific notation.	0.8 Scientific Notation
I can confidently do metric conversions.	0.9 Conversions
I understand the purpose and use of significant figures.	0.10 Significant Figures

It wouldn't hurt you to skim the sections you think you already know about, and to do the self-checks in those sections.

0.5 Basics of the Metric System

The metric system

Units were not standardized until fairly recently in history, so when the physicist Isaac Newton gave the result of an experiment with a pendulum, he had to specify not just that the string was 37 $^{7}/_{8}$ inches long but that it was "37 $^{7}/_{8}$ London inches long." The inch as defined in Yorkshire would have been different. Even after the British Empire standardized its units, it was still very inconvenient to do calculations involving money, volume, distance, time, or weight, because of all the odd conversion factors, like 16 ounces in a pound, and 5280 feet in a mile. Through the nineteenth century, schoolchildren squandered most of their mathematical education in preparing to do calculations such as making change when a customer in a shop offered a one-crown note for a book costing two pounds, thirteen shillings and tuppence. The dollar has always been decimal, and British money went decimal decades ago, but the United States is still saddled with the antiquated system of feet, inches, pounds, ounces and so on.

Every country in the world besides the U.S. has adopted a system of units known in English as the "metric system." This system is entirely decimal, thanks to the same eminently logical people who brought about the French Revolution. In deference to France, the system's official name is the Système International, or SI, meaning International System. (The phrase "SI system" is therefore redundant.)

The wonderful thing about the SI is that people who live in countries more modern than ours do not need to memorize how many ounces there are in a pound, how many cups in a pint, how many feet in a mile, etc. The whole system works with a single, consistent set of prefixes (derived from Greek) that modify the basic units. Each prefix stands for a power of ten, and has an abbreviation that can be combined with the symbol for the unit. For instance, the meter is a unit of distance. The prefix kilo- stands for 10^3, so a kilometer, 1 km, is a thousand meters.

The basic units of the metric system are the meter for distance, the second for time, and the gram for mass.

The following are the most common metric prefixes. You should memorize them.

prefix		meaning	example	
kilo-	k	10^3	60 kg	= a person's mass
centi-	c	10^{-2}	28 cm	= height of a piece of paper
milli-	m	10^{-3}	1 ms	= time for one vibration of a guitar string playing the note D

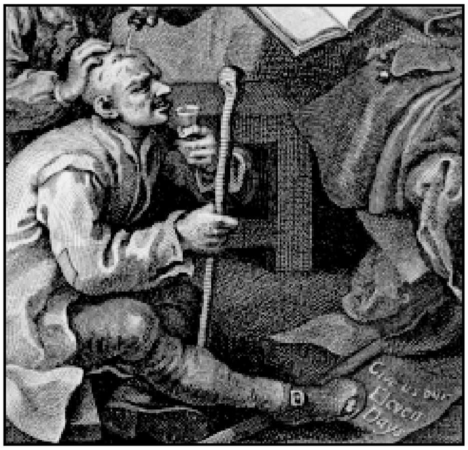
The prefix centi-, meaning 10^{-2}, is only used in the centimeter; a hundredth of a gram would not be written as 1 cg but as 10 mg. The centi- prefix can be easily remembered because a cent is 10^{-2} dollars. The official SI abbreviation for seconds is "s" (not "sec") and grams are "g" (not "gm").

The second

> The sun stood still and the moon halted until the nation had taken vengeance on its enemies...
>
> Joshua 10:12-14

> Absolute, true, and mathematical time, of itself, and from its own nature, flows equably without relation to anything external...
>
> Isaac Newton

When I stated briefly above that the second was a unit of time, it may not have occurred to you that this was not really much of a definition. The two quotes above are meant to demonstrate how much room for confusion exists among people who seem to mean the same thing by a word such as "time." The first quote has been interpreted by some biblical scholars as indicating an ancient belief that the motion of the sun across the sky was not just something that occurred with the passage of time but that the sun actually caused time to pass by its motion, so that freezing it in the sky



Pope Gregory created our modern “Gregorian” calendar, with its system of leap years, to make the length of the calendar year match the length of the cycle of seasons. Not until1752 did Protestant England switched to the new calendar. Some less educated citizens believed that the shortening of the month by eleven days would shorten their lives by the same interval. In this illustration by William Hogarth, the leaflet lying on the ground reads, “Give us our eleven days.”

would have some kind of a supernatural decelerating effect on everyone except the Hebrew soldiers. Many ancient cultures also conceived of time as cyclical, rather than proceeding along a straight line as in 1998, 1999, 2000, 2001,... The second quote, from a relatively modern physicist, may sound a lot more scientific, but most physicists today would consider it useless as a definition of time. Today, the physical sciences are based on *operational definitions*, which means definitions that spell out the actual steps (operations) required to measure something numerically.

Now in an era when our toasters, pens, and coffee pots tell us the time, it is far from obvious to most people what is the fundamental operational definition of time. Until recently, the hour, minute, and second were defined operationally in terms of the time required for the earth to rotate about its axis. Unfortunately, the Earth’s rotation is slowing down slightly, and by 1967 this was becoming an issue in scientific experiments requiring precise time measurements. The second was therefore redefined as the time required for a certain number of vibrations of the light waves emitted by a cesium atoms in a lamp constructed like a familiar neon sign but with the neon replaced by cesium. The new definition not only promises to stay constant indefinitely, but for scientists is a more convenient way of calibrating a clock than having to carry out astronomical measurements.

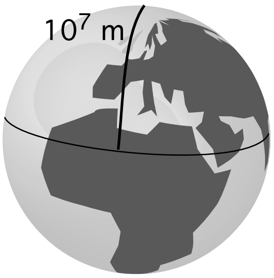
The Time Without Underwear

Unfortunately, the French Revolutionary calendar never caught on. Each of its twelve months was 30 days long, with names like Thermidor (the month of heat) and Germinal (the month of budding). To round out the year to 365 days, a five-day period was added on the end of the calendar, and named the *sans culottides*. In modern French, *sans culottides* means “time without underwear,” but in the 18th century, it was a way to honor the workers and peasants, who wore simple clothing instead of the fancy pants (*culottes*) of the aristocracy.

Self-Check

What is a possible operational definition of how strong a person is?

The meter



The French originally defined the meter as 10^{-7} times the distance from the equator to the north pole, as measured through Paris (of course). Even if the definition was operational, the operation of traveling to the north pole and laying a surveying chain behind you was not one that most working scientists wanted to carry out. Fairly soon, a standard was created in the form of a metal bar with two scratches on it. This definition persisted until 1960, when the meter was redefined as the distance traveled by light in a vacuum over a period of (1/299792458) seconds.

A dictionary might define “strong” as “posessing powerful muscles,” but that’s not an operational definition, because it doesn’t say how to measure strength numerically. One possible operational definition would be the number of pounds a person can bench press.

The kilogram

The third base unit of the SI is the kilogram, a unit of mass. Mass is intended to be a measure of the amount of a substance, but that is not an operational definition. Bathroom scales work by measuring our planet's gravitational attraction for the object being weighed, but using that type of scale to define mass operationally would be undesirable because gravity varies in strength from place to place on the earth.

There's a surprising amount of disagreement among physics textbooks about how mass should be defined, but here's how it's actually handled by the few working physicists who specialize in ultra-high-precision measurements. They maintain a physical object in Paris, which is the standard kilogram, a cylinder made of platinum-iridium alloy. Duplicates are checked against this mother of all kilograms by putting the original and the copy on the two opposite pans of a balance. Although this method of comparison depends on gravity, the problems associated with differences in gravity in different geographical locations are bypassed, because the two objects are being compared in the same place. The duplicates can then be removed from the Parisian kilogram shrine and transported elsewhere in the world.

Combinations of metric units

Just about anything you want to measure can be measured with some combination of meters, kilograms, and seconds. Speed can be measured in m/s, volume in m^3, and density in kg/m^3. Part of what makes the SI great is this basic simplicity. No more funny units like a cord of wood, a bolt of cloth, or a jigger of whiskey. No more liquid and dry measure. Just a simple, consistent set of units. The SI measures put together from meters, kilograms, and seconds make up the mks system. For example, the mks unit of speed is m/s, not km/hr.

Discussion question

Isaac Newton wrote, "...the natural days are truly unequal, though they are commonly considered as equal, and used for a measure of time... It may be that there is no such thing as an equable motion, whereby time may be accurately measured. All motions may be accelerated or retarded..." Newton was right. Even the modern definition of the second in terms of light emitted by cesium atoms is subject to variation. For instance, magnetic fields could cause the cesium atoms to emit light with a slightly different rate of vibration. What makes us think, though, that a pendulum clock is more accurate than a sundial, or that a cesium atom is a more accurate timekeeper than a pendulum clock? That is, how can one test experimentally how the accuracies of different time standards compare?

0.6 The Newton, the Metric Unit of Force

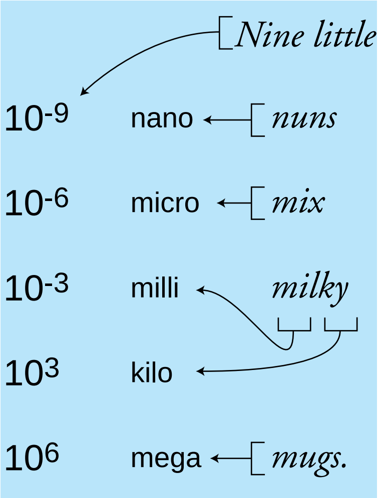
A force is a push or a pull, or more generally anything that can change an object's speed or direction of motion. A force is required to start a car moving, to slow down a baseball player sliding in to home base, or to make an airplane turn. (Forces may fail to change an object's motion if they are canceled by other forces, e.g. the force of gravity pulling you down right now is being canceled by the force of the chair pushing up on you.) The metric unit of force is the Newton, defined as the force which, if applied for one second, will cause a 1-kilogram object starting from rest to reach a speed of 1 m/s. Later chapters will discuss the force concept in more detail. In fact, this entire book is about the relationship between force and motion.

In the previous section, I gave a gravitational definition of mass, but by defining a numerical scale of force, we can also turn around and define a scale of mass without reference to gravity. For instance, if a force of two Newtons is required to accelerate a certain object from rest to 1 m/s in 1 s, then that object must have a mass of 2 kg. From this point of view, mass characterizes an object's resistance to a change in its motion, which we call inertia or inertial mass. Although there is no fundamental reason why an object's resistance to a change in its motion must be related to how strongly gravity affects it, careful and precise experiments have shown that the inertial definition and the gravitational definition of mass are highly consistent for a variety of objects. It therefore doesn't really matter for any practical purpose which definition one adopts.

Discussion Question

Spending a long time in weightlessness is unhealthy. One of the most important negative effects experienced by astronauts is a loss of muscle and bone mass. Since an ordinary scale won't work for an astronaut in orbit, what is a possible way of monitoring this change in mass? (Measuring the astronaut's waist or biceps with a measuring tape is not good enough, because it doesn't tell anything about bone mass, or about the replacement of muscle with fat.)

0.7 Less Common Metric Prefixes



This is a mnemonic to help you remember the most important metric prefixes. The word "little" is to remind you that the list starts with the prefixes used for small quantities and builds upward. The exponent changes by 3 with each step, except that of course we do not need a special prefix for 10^0, which equals one.

The following are three metric prefixes which, while less common than the ones discussed previously, are well worth memorizing.

prefix		meaning	example	
mega-	M	10^6	6.4 Mm	= radius of the earth
micro-	μ	10^{-6}	1 μm	= diameter of a human hair
nano-	n	10^{-9}	0.154 nm	= distance between carbon nuclei in an ethane molecule

Note that the abbreviation for micro is the Greek letter mu, μ — a common mistake is to confuse it with m (milli) or M (mega).

There are other prefixes even less common, used for extremely large and small quantities. For instance, 1 femtometer=10^{-15} m is a convenient unit of distance in nuclear physics, and 1 gigabyte=10^9 bytes is used for computers' hard disks. The international committee that makes decisions about the SI has recently even added some new prefixes that sound like jokes, e.g. 1 yoctogram = 10^{-24} g is about half the mass of a proton. In the immediate future, however, you're unlikely to see prefixes like "yocto-" and "zepto-" used except perhaps in trivia contests at science-fiction conventions or other geekfests.

Self-Check

Suppose you could slow down time so that according to your perception, a beam of light would move across a room at the speed of a slow walk. If you perceived a nanosecond as if it was a second, how would you perceive a microsecond?

0.8 Scientific Notation

Most of the interesting phenomena our universe has to offer are not on the human scale. It would take about 1,000,000,000,000,000,000,000 bacteria to equal the mass of a human body. When the physicist Thomas Young discovered that light was a wave, it was back in the bad old days before scientific notation, and he was obliged to write that the time required for one vibration of the wave was 1/500 of a millionth of a millionth of a second. Scientific notation is a less awkward way to write very large and very small numbers such as these. Here's a quick review.

Scientific notation means writing a number in terms of a product of something from 1 to 10 and something else that is a power of ten. For instance,

$32 = 3.2 \text{ x } 10^1$
$320 = 3.2 \text{ x } 10^2$
$3200 = 3.2 \text{ x } 10^3$ ...

Each number is ten times bigger than the previous one.

Since 10^1 is ten times smaller than 10^2, it makes sense to use the notation 10^0 to stand for one, the number that is in turn ten times smaller than 10^1. Continuing on, we can write 10^{-1} to stand for 0.1, the number ten times smaller than 10^0. Negative exponents are used for small numbers:

$3.2 = 3.2 \text{ x } 10^0$
$0.32 = 3.2 \text{ x } 10^{-1}$
$0.032 = 3.2 \text{ x } 10^{-2}$ ...

A common source of confusion is the notation used on the displays of many calculators. Examples:

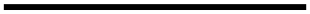
$3.2 \text{ x } 10^6$	(written notation)
3.2E+6	(notation on some calculators)
3.2^6	(notation on some other calculators)

The last example is particularly unfortunate, because 3.2^6 really stands for the number 3.2x3.2x3.2x3.2x3.2x3.2 = 1074, a totally different number from $3.2 \text{ x } 10^6 = 3200000$. The calculator notation should never be used in writing. It's just a way for the manufacturer to save money by making a simpler display.

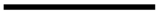
A microsecond is 1000 times longer than a nanosecond, so it would seem like 1000 seconds, or about 20 minutes.

Self-Check

A student learns that 10^4 bacteria, standing in line to register for classes at Paramecium Community College, would form a queue of this size:



The student concludes that 10^2 bacteria would form a line of this length:



Why is the student incorrect?

0.9 Conversions

I suggest you avoid memorizing lots of conversion factors between SI units and U.S. units. Suppose the United Nations sends its black helicopters to invade California (after all who wouldn't rather live here than in New York City?), and institutes water fluoridation and the SI, making the use of inches and pounds into a crime punishable by death. I think you could get by with only two mental conversion factors:

1 inch = 2.54 cm
An object with a weight on Earth of 2.2 lb has a mass of 1 kg.

The first one is the present definition of the inch, so it's exact. The second one is not exact, but is good enough for most purposes. The pound is a unit of gravitational force, while the kg is a unit of mass, which measures how hard it is to accelerate an object, not how hard gravity pulls on it. Therefore it would be incorrect to say that 2.2 lb literally equaled 1 kg, even approximately.

More important than memorizing conversion factors is understanding the right method for doing conversions. Even within the SI, you may need to convert, say, from grams to kilograms. Different people have different ways of thinking about conversions, but the method I'll describe here is systematic and easy to understand. The idea is that if 1 kg and 1000 g represent the same mass, then we can consider a fraction like

$$\frac{10^3\ \text{g}}{1\ \text{kg}}$$

to be a way of expressing the number one. This may bother you. For instance, if you type 1000/1 into your calculator, you will get 1000, not one. Again, different people have different ways of thinking about it, but the justification is that it helps us to do conversions, and it works! Now if we want to convert 0.7 kg to units of grams, we can multiply 0.7 kg by the number one:

$$0.7\ \text{kg} \times \frac{10^3\ \text{g}}{1\ \text{kg}}$$

If you're willing to treat symbols such as "kg" as if they were variables as used in algebra (which they're really not), you can then cancel the kg on top with the kg on the bottom, resulting in

Exponents have to do with multiplication, not addition. The first line should be 100 times longer than the second, not just twice as long.

$$0.7\ \cancel{\text{kg}} \times \frac{10^3\ \text{g}}{1\ \cancel{\text{kg}}} = 700\ \text{g} \quad .$$

To convert grams to kilograms, you would simply flip the fraction upside down.

One advantage of this method is that it can easily be applied to a series of conversions. For instance, to convert one year to units of seconds,

$$1\ \cancel{\text{year}} \times \frac{365\ \cancel{\text{days}}}{1\ \cancel{\text{year}}} \times \frac{24\ \cancel{\text{hours}}}{1\ \cancel{\text{day}}} \times \frac{60\ \cancel{\text{min}}}{1\ \cancel{\text{hour}}} \times \frac{60\ \text{s}}{1\ \cancel{\text{min}}}$$

$$= 3.15 \text{ x } 10^7\ \text{s} \quad .$$

Should that exponent be positive or negative?

A common mistake is to write the conversion fraction incorrectly. For instance the fraction

$$\frac{10^3\ \text{kg}}{1\ \text{g}} \qquad \text{(incorrect)}$$

checking conversions using common sense

does not equal one, because 10^3 kg is the mass of a car, and 1 g is the mass of a raisin. One correct way of setting up the conversion factor would be

$$\frac{10^{-3}\ \text{kg}}{1\ \text{g}} \quad . \qquad \text{(correct)}$$

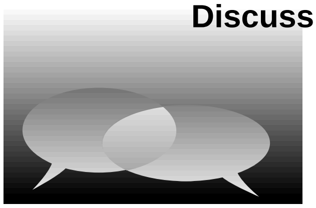
You can usually detect such a mistake if you take the time to check your answer and see if it is reasonable.

checking conversions using the idea of "compensating"

If common sense doesn't rule out either a positive or a negative exponent, here's another way to make sure you get it right. There are big prefixes and small prefixes:

big prefixes:	k	M	
small prefixes:	m	µ	n

(It's not hard to keep straight which are which, since "mega" and "micro" are evocative, and it's easy to remember that a kilometer is bigger than a meter and a millimeter is smaller.) In the example above, we want the top of the fraction to be the same as the bottom. Since k is a big prefix, we need to *compensate* by putting a small number like 10^{-3} in front of it, not a big number like 10^3.



Discussion Question

Each of the following conversions contains an error. In each case, explain what the error is.

(a) $1000 \text{ kg x } \frac{1\ \text{kg}}{1000\ \text{g}} = 1\ \text{g}$ (b) $50 \text{ m x } \frac{1\ \text{cm}}{100\ \text{m}} = 0.5\ \text{cm}$

(c) "Nano" is 10^{-9}, so there are 10^{-9} nm in a meter.

(d) "Micro" is 10^{-6}, so 1 kg is 10^6 µg.

0.10 Significant Figures

An engineer is designing a car engine, and has been told that the diameter of the pistons (which are being designed by someone else) is 5 cm. He knows that 0.02 cm of clearance is required for a piston of this size, so he designs the cylinder to have an inside diameter of 5.04 cm. Luckily, his supervisor catches his mistake before the car goes into production. She explains his error to him, and mentally puts him in the "do not promote" category.

What was his mistake? The person who told him the pistons were 5 cm in diameter was wise to the ways of significant figures, as was his boss, who explained to him that he needed to go back and get a more accurate number for the diameter of the pistons. That person said "5 cm" rather than "5.00 cm" specifically to avoid creating the impression that the number was extremely accurate. In reality, the pistons' diameter was 5.13 cm. They would never have fit in the 5.04-cm cylinders.

Significant figures communicate the accuracy of a number.

The number of digits of accuracy in a number is referred to as the number of significant figures, or "sig figs" for short. As in the example above, sig figs provide a way of showing the accuracy of a number. In most cases, the result of a calculation involving several pieces of data can be no more accurate than the least accurate piece of data. In other words, "garbage in, garbage out." Since the 5 cm diameter of the pistons was not very accurate, the result of the engineer's calculation, 5.04 cm, was really not as accurate as he thought. In general, your result should not have more than the number of sig figs in the least accurate piece of data you started with. The calculation above should have been done as follows:

5 cm	(1 sig fig)
+ 0.04 cm	(1 sig fig)
= 5 cm	(rounded off to 1 sig fig)

The fact that the final result only has one significant figure then alerts you to the fact that the result is not very accurate, and would not be appropriate for use in designing the engine.

Note that the leading zeroes in the number 0.04 do not count as significant figures, because they are only placeholders. On the other hand, a number such as 50 cm is ambiguous — the zero could be intended as a significant figure, or it might just be there as a placeholder. The ambiguity involving trailing zeroes can be avoided by using scientific notation, in which 5 x 10^1 cm would imply one sig fig of accuracy, while 5.0 x 10^1 cm would imply two sig figs.

Dealing correctly with significant figures can save you time! Often, students copy down numbers from their calculators with eight significant figures of precision, then type them back in for a later calculation. That's a waste of time, unless your original data had that kind of incredible precision.

The rules about significant figures are only rules of thumb, and are not a substitute for careful thinking. For instance, $20.00 + $0.05 is $20.05. It need not and should not be rounded off to $20. In general, the sig fig rules work best for multiplication and division, and we also apply them when doing a complicated calculation that involves many types of operations. For simple addition and subtraction, it makes more sense to maintain a fixed number of digits after the decimal point. When in doubt, don't use the sig fig rules at all: just observe the effect on your final result when you change one piece of your initial data by the maximum amount by which you think it could have been inaccurate.

Self-Check

How many significant figures are there in each of the following measurements?

(a) 9.937 m

(b) 4.0 s

(c) 0.0000037 kg

(a) 4; (b) 2; (c) 2

Summary

Selected Vocabulary

matter .... Anything that is affected by gravity.

light.................................. Anything that can travel from one place to another through empty space and can influence matter, but is not affected by gravity.

operational definition .... A definition that states what operations should be carried out to measure the thing being defined.

Système International .... A fancy name for the metric system.

mks system .... The use of metric units based on the meter, kilogram, and second. Example: meters per second is the mks unit of speed, not cm/s or km/hr.

mass .... A numerical measure of how difficult it is to change an object's motion.

significant figures .... Digits that contribute to the accuracy of a measurement.

Notation

m	symbol for mass, or the meter, the metric distance unit
kg	kilogram, the metric unit of mass
s	second, the metric unit of time
M-	the metric prefix mega-, 10^6
k-	the metric prefix kilo-, 10^3
m-	the metric prefix milli-, 10^{-3}
µ-	the metric prefix micro-, 10^{-6}
n-	the metric prefix nano-, 10^{-9}

Summary

Physics is the use of the scientific method to study the behavior of light and matter. The scientific method requires a cycle of theory and experiment, theories with both predictive and explanatory value, and reproducible experiments.

The metric system is a simple, consistent framework for measurement built out of the meter, the kilogram, and the second plus a set of prefixes denoting powers of ten. The most systematic method for doing conversions is shown in the following example:

$$370\text{ ms}\times\frac{10^{-3}\text{ s}}{1\text{ ms}} = 0.37\text{ s}$$

Mass is a measure of the amount of a substance. Mass can be defined gravitationally, by comparing an object to a standard mass on a double-pan balance, or in terms of inertia, by comparing the effect of a force on an object to the effect of the same force on a standard mass. The two definitions are found experimentally to be proportional to each other to a high degree of precision, so we usually refer simply to "mass," without bothering to specify which type.

A force is that which can change the motion of an object. The metric unit of force is the Newton, defined as the force required to accelerate a standard 1-kg mass from rest to a speed of 1 m/s in 1 s.

Scientific notation means, for example, writing 3.2x10^5 rather than 320000.

Writing numbers with the correct number of significant figures correctly communicates how accurate they are. As a rule of thumb, the final result of a calculation is no more accurate than, and should have no more significant figures than, the least accurate piece of data.

Homework Problems

1. Correct use of a calculator: (a✓) Calculate $\frac{74658}{53222 + 97554}$ on a calculator.

[Self-check: The most common mistake results in 97555.40.]

(b) Which would be more like the price of a TV, and which would be more like the price of a house, \$ 3.5x10^5 or \$ 3.5^5?

2. Compute the following things. If they don't make sense because of units, say so.

(a) 3 cm + 5 cm

(b) 1.11 m + 22 cm

(c) 120 miles + 2.0 hours

(d) 120 miles / 2.0 hours

3. Your backyard has brick walls on both ends. You measure a distance of 23.4 m from the inside of one wall to the inside of the other. Each wall is 29.4 cm thick. How far is it from the outside of one wall to the outside of the other? Pay attention to significant figures.

4 ✓. The speed of light is 3.0x10^8 m/s. Convert this to furlongs per fortnight. A furlong is 220 yards, and a fortnight is 14 days. An inch is 2.54 cm.

5 ✓. Express each of the following quantities in micrograms: (a) 10 mg, (b) 10^4 g, (c) 10 kg, (d) 100x10^3 g, (e) 1000 ng.

6 S. Convert 134 mg to units of kg, writing your answer in scientific notation.

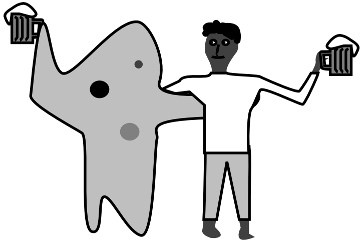
7✓. In the last century, the average age of the onset of puberty for girls has decreased by several years. Urban folklore has it that this is because of hormones fed to beef cattle, but it is more likely to be because modern girls have more body fat on the average and possibly because of estrogen-mimicking chemicals in the environment from the breakdown of pesticides. A hamburger from a hormone-implanted steer has about 0.2 ng of estrogen (about double the amount of natural beef). A serving of peas contains about 300 ng of estrogen. An adult woman produces about 0.5 mg of estrogen per day (note the different unit!). (a) How many hamburgers would a girl have to eat in one day to consume as much estrogen as an adult woman's daily production? (b) How many servings of peas?

S A solution is given in the back of the book. ★ A difficult problem.
✓ A computerized answer check is available. ∫ A problem that requires calculus.

Life would be very different if you were the size of an insect.

1 Scaling and Order-of-Magnitude Estimates

1.1 Introduction



Amoebas this size are seldom encountered.

Why can't an insect be the size of a dog? Some skinny stretched-out cells in your spinal cord are a meter tall — why does nature display no single cells that are not just a meter tall, but a meter wide, and a meter thick as well? Believe it or not, these are questions that can be answered fairly easily without knowing much more about physics than you already do. The only mathematical technique you really need is the humble conversion, applied to area and volume.

Area and volume

Area can be defined by saying that we can copy the shape of interest onto graph paper with 1 cm x 1 cm squares and count the number of squares inside. Fractions of squares can be estimated by eye. We then say the area equals the number of squares, in units of square cm. Although this might seem less "pure" than computing areas using formulae like $A=\pi r^2$ for a circle or $A=wh/2$ for a triangle, those formulae are not useful as definitions of area because they cannot be applied to irregularly shaped areas.

Units of square cm are more commonly written as cm^2 in science. Of course, the unit of measurement symbolized by "cm" is not an algebra symbol standing for a number that can be literally multiplied by itself. But it is advantageous to write the units of area that way and treat the units as if they were algebra symbols. For instance, if you have a rectangle with an area of 6 m^2 and a width of 2 m, then calculating its length as $(6\ \text{m}^2)/(2\ \text{m})=3$ m gives a result that makes sense both numerically and in terms of units. This algebra-style treatment of the units also ensures that our methods of converting units work out correctly. For instance, if we accept the fraction

$$\frac{100\text{ cm}}{1\text{ m}}$$

as a valid way of writing the number one, then one times one equals one, so we should also say that one can be represented by

$$\frac{100\text{ cm}}{1\text{ m}} \times \frac{100\text{ cm}}{1\text{ m}}$$

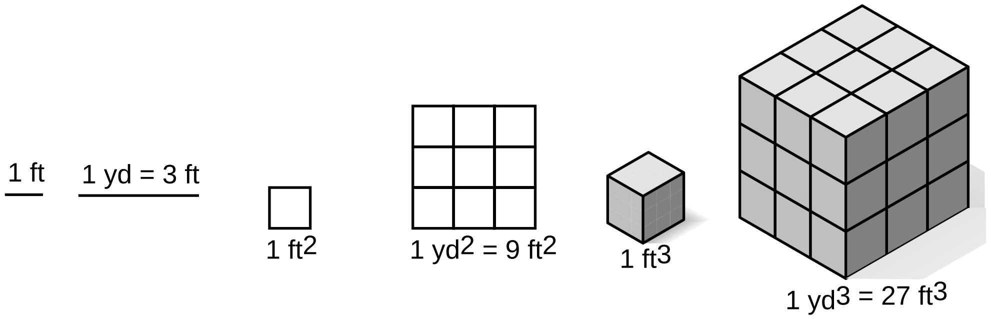
which is the same as

$$\frac{10000\text{ cm}^2}{1\text{ m}^2} \quad .$$

That means the conversion factor from square meters to square centimeters is a factor of 10^4, i.e. a square meter has 10^4 square centimeters in it.

All of the above can be easily applied to volume as well, using one-cubic-centimeter blocks instead of squares on graph paper.

To many people, it seems hard to believe that a square meter equals 10000 square centimeters, or that a cubic meter equals a million cubic centimeters — they think it would make more sense if there were 100 cm^2 in 1 m^2, and 100 cm^3 in 1 m^3, but that would be incorrect. The examples shown in the figure below aim to make the correct answer more believable, using the traditional U.S. units of feet and yards. (One foot is 12 inches, and one yard is three feet.)



Self-Check

Based on the figure, convince yourself that there are 9 ft^2 in a square yard , and 27 ft^3 in a cubic yard, then demonstrate the same thing symbolically (i.e. with the method using fractions that equal one).

Discussion question

A. How many square centimeters are there in a square inch? (1 inch=2.54 cm) First find an approximate answer by making a drawing, then derive the conversion factor more accurately using the symbolic method.

1 yd^2x(3 ft/1 yd)^2=9 ft^2. 1 yd^3x(3 ft/1 yd)^3=27 ft^3.

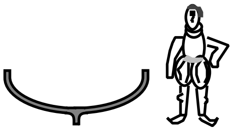
GALILEO GALILEI LINCEO FILOSOFO E MATEMATICO DEL SER.MO GRAN DUCA DI TOSCA

Galileo Galilei (1564-1642) was a Renaissance Italian who brought the scientific method to bear on physics, creating the modern version of the science. Coming from a noble but very poor family, Galileo had to drop out of medical school at the University of Pisa when he ran out of money. Eventually becoming a lecturer in mathematics at the same school, he began a career as a notorious troublemaker by writing a burlesque ridiculing the university's regulations — he was forced to resign, but found a new teaching position at Padua. He invented the pendulum clock, investigated the motion of falling bodies, and discovered the moons of Jupiter. The thrust of his life's work was to discredit Aristotle's physics by confronting it with contradictory experiments, a program which paved the way for Newton's discovery of the relationship between force and motion. In Chapter 3 we'll come to the story of Galileo's ultimate fate at the hands of the Church.

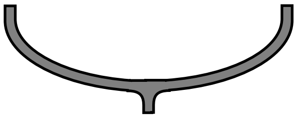
1.2 Scaling of Area and Volume

Great fleas have lesser fleas
Upon their backs to bite 'em.
And lesser fleas have lesser still,
And so ad infinitum.
Jonathan Swift

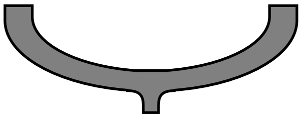
Now how do these conversions of area and volume relate to the questions I posed about sizes of living things? Well, imagine that you are shrunk like Alice in Wonderland to the size of an insect. One way of thinking about the change of scale is that what used to look like a centimeter now looks like perhaps a meter to you, because you're so much smaller. If area and volume scaled according to most people's intuitive, incorrect expectations, with 1 m^2 being the same as 100 cm^2, then there would be no particular reason why nature should behave any differently on your new, reduced scale. But nature does behave differently now that you're small. For instance, you will find that you can walk on water, and jump to many times your own height. The physicist Galileo Galilei had the basic insight that the scaling of area and volume determines how natural phenomena behave differently on different scales. He first reasoned about mechanical structures, but later extended his insights to living things, taking the then-radical point of view that at the fundamental level, a living organism should follow the same laws of nature as a machine. We will follow his lead by first discussing machines and then living things.



The small boat holds up just fine.



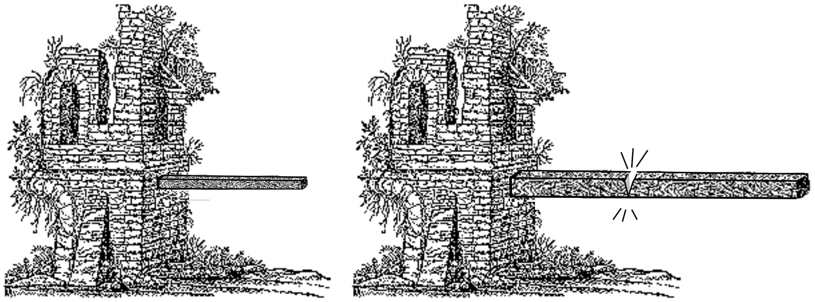
A larger boat built with the same proportions as the small one will collapse under its own weight.



A boat this large needs to have timbers that are thicker compared to its size.

Galileo on the behavior of nature on large and small scales

One of the world's most famous pieces of scientific writing is Galileo's **Dialogues Concerning the Two New Sciences**. Galileo was an entertaining writer who wanted to explain things clearly to laypeople, and he livened up his work by casting it in the form of a dialogue among three people. Salviati is really Galileo's alter ego. Simplicio is the stupid character, and one of the reasons Galileo got in trouble with the Church was that there were rumors that Simplicio represented the Pope. Sagredo is the earnest and intelligent student, with whom the reader is supposed to identify. (The following excerpts are from the 1914 translation by Crew and de Salvio.)



This plank is the longest it can be without collapsing under its own weight. If it was a hundredth of an inch longer, it would collapse.

This plank is made out of the same kind of wood. It is twice as thick, twice as long, and twice as wide. It will collapse under its own weight.

(After Galileo's original drawing.)

> Salviati: ...we asked the reason why [shipbuilders] employed stocks, scaffolding, and bracing of larger dimensions for launching a big vessel than they do for a small one; and [an old man] answered that they did this in order to avoid the danger of the ship parting under its own heavy weight, a danger to which small boats are not subject?
>
> Sagredo: Yes, that is what I mean; and I refer especially to his last assertion which I have always regarded as false...; namely, that in speaking of these and other similar machines one cannot argue from the small to the large, because many devices which succeed on a small scale do not work on a large scale. Now, since mechanics has its foundations in geometry, where mere size [is unimportant], I do not see that the properties of circles, triangles, cylinders, cones and other solid figures will change with their size. If, therefore, a large machine be constructed in such a way that its parts bear to one another the same ratio as in a smaller one, and if the smaller is sufficiently strong for the purpose for which it is designed, I do not see why the larger should not be able to withstand any severe and destructive tests to which it may be subjected.

Salviati contradicts Sagredo:

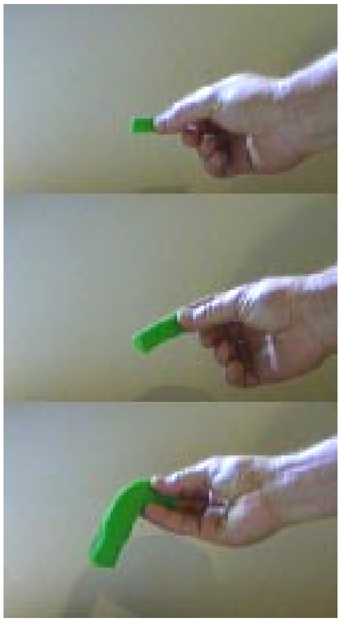
> Salviati: ...Please observe, gentlemen, how facts which at first seem improbable will, even on scant explanation, drop the cloak which has hidden them and stand forth in naked and simple beauty. Who does not know that a horse falling from a height of three or four cubits will break his bones, while a dog falling from the same height or a cat from a height of eight or ten cubits will suffer no injury? Equally harmless would be the fall of a grasshopper from a tower or the fall of an ant from the distance of the moon.

The point Galileo is making here is that small things are sturdier in proportion to their size. There are a lot of objections that could be raised, however. After all, what does it really mean for something to be "strong", to be "strong in proportion to its size," or to be strong "out of proportion to its size?" Galileo hasn't spelled out operational definitions of things like "strength," i.e. definitions that spell out how to measure them numerically.

Also, a cat is shaped differently from a horse — an enlarged photograph of a cat would not be mistaken for a horse, even if the photo-doctoring experts at the National Inquirer made it look like a person was riding on its back. A grasshopper is not even a mammal, and it has an exoskeleton instead of an internal skeleton. The whole argument would be a lot more convincing if we could do some isolation of variables, a scientific term that means to change only one thing at a time, isolating it from the other variables that might have an effect. If size is the variable whose effect we're interested in seeing, then we don't really want to compare things that are different in size but also different in other ways.

Also, Galileo is doing something that would be frowned on in modern science: he is mixing experiments whose results he has actually observed (building boats of different sizes), with experiments that he could not possibly have done (dropping an ant from the height of the moon).

After this entertaining but not scientifically rigorous beginning, Galileo starts to do something worthwhile by modern standards. He simplifies everything by considering the strength of a wooden plank. The variables involved can then be narrowed down to the type of wood, the width, the thickness, and the length. He also gives an operational definition of what it means for the plank to have a certain strength "in proportion to its size," by introducing the concept of a plank that is the longest one that would not snap under its own weight if supported at one end. If you increased its length by the slightest amount, without increasing its width or thickness, it would break. He says that if one plank is the same shape as another but a different size, appearing like a reduced or enlarged photograph of the other, then the planks would be strong "in proportion to their sizes" if both were just barely able to support their own weight.



Galileo discusses planks made of wood, but the concept may be easier to imagine with clay. All three clay rods in the figure were originally the same shape. The medium-sized one was twice the height, twice the length, and twice the width of the small one, and similarly the large one was twice as big as the medium one in all its linear dimensions. The big one has four times the linear dimensions of the small one, 16 times the cross-sectional area when cut perpendicular to the page, and 64 times the volume. That means that the big one has 64 times the weight to support, but only 16 times the strength compared to the smallest one.

He now relates how he has done actual experiments with such planks, and found that, according to this operational definition, they are not strong in proportion to their sizes. The larger one breaks. He makes sure to tell the reader how important the result is, via Sagredo's astonished response:

> SAGREDO: My brain already reels. My mind, like a cloud momentarily illuminated by a lightning flash, is for an instant filled with an unusual light, which now beckons to me and which now suddenly mingles and obscures strange, crude ideas. From what you have said it appears to me impossible to build two similar structures of the same material, but of different sizes and have them proportionately strong.

In other words, this specific experiment, using things like wooden planks that have no intrinsic scientific interest, has very wide implications because it points out a general principle, that nature acts differently on different scales.

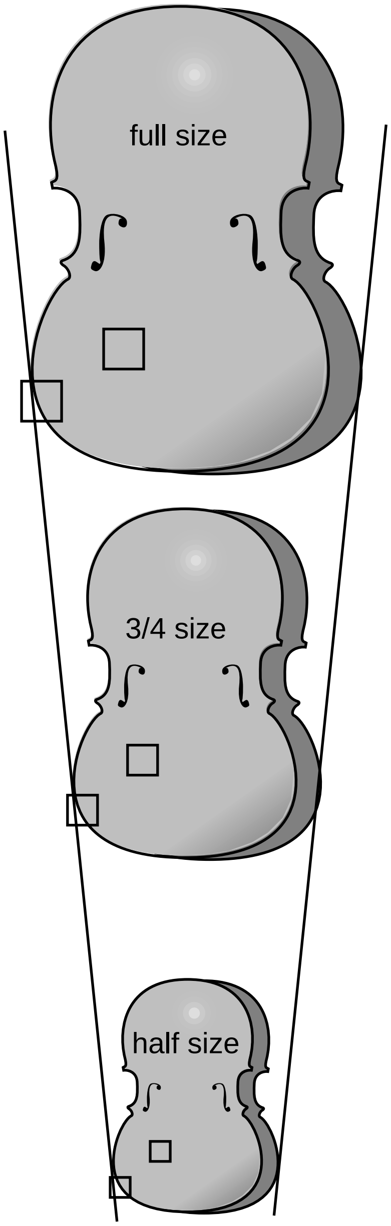
To finish the discussion, Galileo gives an explanation. He says that the strength of a plank (defined as, say, the weight of the heaviest boulder you could put on the end without breaking it) is proportional to its cross-sectional area, that is, the surface area of the fresh wood that would be exposed if you sawed through it in the middle. Its weight, however, is proportional to its volume.

How do the volume and cross-sectional area of the longer plank compare with those of the shorter plank? We have already seen, while discussing conversions of the units of area and volume, that these quantities don't act the way most people naively expect. You might think that the volume and area of the longer plank would both be doubled compared to the shorter plank, so they would increase in proportion to each other, and the longer plank would be equally able to support its weight. You would be wrong, but Galileo knows that this is a common misconception, so he has

Salviati address the point specifically:

> SALVIATI: ...Take, for example, a cube two inches on a side so that each face has an area of four square inches and the total area, i.e., the sum of the six faces, amounts to twenty-four square inches; now imagine this cube to be sawed through three times [with cuts in three perpendicular planes] so as to divide it into eight smaller cubes, each one inch on the side, each face one inch square, and the total surface of each cube six square inches instead of twenty-four in the case of the larger cube. It is evident therefore, that the surface of the little cube is only one-fourth that of the larger, namely, the ratio of six to twenty-four; but the volume of the solid cube itself is only one-eighth; the volume, and hence also the weight, diminishes therefore much more rapidly than the surface... You see, therefore, Simplicio, that I was not mistaken when ... I said that the surface of a small solid is comparatively greater than that of a large one.

The same reasoning applies to the planks. Even though they are not cubes, the large one could be sawed into eight small ones, each with half the length, half the thickness, and half the width. The small plank, therefore, has more surface area in proportion to its weight, and is therefore able to support its own weight while the large one breaks.



Scaling of area and volume for irregularly shaped objects

You probably are not going to believe Galileo's claim that this has deep implications for all of nature unless you can be convinced that the same is true for any shape. Every drawing you've seen so far has been of squares, rectangles, and rectangular solids. Clearly the reasoning about sawing things up into smaller pieces would not prove anything about, say, an egg, which cannot be cut up into eight smaller egg-shaped objects with half the length.

Is it always true that something half the size has one quarter the surface area and one eighth the volume, even if it has an irregular shape? Take the example of a child's violin. Violins are made for small children in lengths that are either half or 3/4 of the normal length, accommodating their small hands. Let's study the surface area of the front panels of the three violins.

Consider the square in the interior of the panel of the full-size violin. In the 3/4-size violin, its height and width are both smaller by a factor of 3/4, so the area of the corresponding, smaller square becomes 3/4x3/4=9/16 of the original area, not 3/4 of the original area. Similarly, the corresponding square on the smallest violin has half the height and half the width of the original one, so its area is 1/4 the original area, not half.

The same reasoning works for parts of the panel near the edge, such as the part that only partially fills in the other square. The entire square scales down the same as a square in the interior, and in each violin the same fraction (about 70%) of the square is full, so the contribution of this part to the total area scales down just the same.

Since any small square region or any small region covering part of a square scales down like a square object, the entire surface area of an irregularly shaped object changes in the same manner as the surface area of a square: scaling it down by 3/4 reduces the area by a factor of 9/16, and so on.

In general, we can see that any time there are two objects with the same shape, but different linear dimensions (i.e. one looks like a reduced photo of the other), the ratio of their areas equals the ratio of the squares of their linear dimensions:

$$\frac{A_1}{A_2}=\left(\frac{L_1}{L_2}\right)^2 \quad .$$

Note that it doesn't matter where we choose to measure the linear size, L, of an object. In the case of the violins, for instance, it could have been measured vertically, horizontally, diagonally, or even from the bottom of the left f-hole to the middle of the right f-hole. We just have to measure it in a consistent way on each violin. Since all the parts are assumed to shrink or expand in the same manner, the ratio L_1/L_2 is independent of the choice of measurement.

It is also important to realize that it is completely unnecessary to have a formula for the area of a violin. It is only possible to derive simple formulas for the areas of certain shapes like circles, rectangles, triangles and so on, but that is no impediment to the type of reasoning we are using.

Sometimes it is inconvenient to write all the equations in terms of ratios, especially when more than two objects are being compared. A more compact way of rewriting the previous equation is

$$A \propto L^2 \quad .$$

The symbol "$\propto$" means "is proportional to." Scientists and engineers often speak about such relationships verbally using the phrases "scales like" or "goes like," for instance "area goes like length squared."

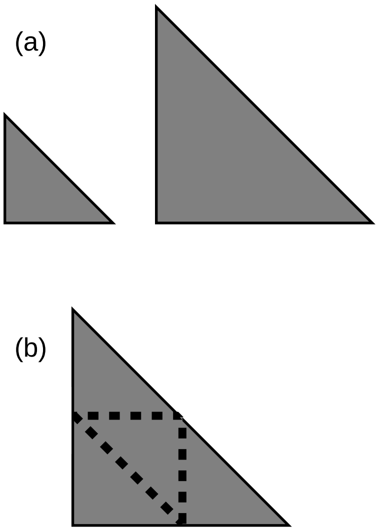
All of the above reasoning works just as well in the case of volume. Volume goes like length cubed:

$$V \propto L^3 \quad .$$

If different objects are made of the same material with the same density, $\rho=m/V$, then their masses, $m=\rho V$, are proportional to L^3, and so are their weights. (The symbol for density is ρ, the lower-case Greek letter "rho".)

An important point is that all of the above reasoning about scaling only applies to objects that are the same shape. For instance, a piece of paper is larger than a pencil, but has a much greater surface-to-volume ratio.

One of the first things I learned as a teacher was that students were not very original about their mistakes. Every group of students tends to come up with the same goofs as the previous class. The following are some examples of correct and incorrect reasoning about proportionality.



The big triangle has four times more area than the little one.

Example: scaling of the area of a triangle

Question: In fig. (a), the larger triangle has sides twice as long. How many times greater is its area?

Correct solution #1: Area scales in proportion to the square of the linear dimensions, so the larger triangle has four times more area ($2^2=4$).

Correct solution #2: You could cut the larger triangle into four of the smaller size, as shown in fig. (b), so its area is four times greater. (This solution is correct, but it would not work for a shape like a circle, which can't be cut up into smaller circles.)

Correct solution #3: The area of a triangle is given by $A = \frac{1}{2}bh$, where b is the base and h is the height. The areas of the triangles are

$$\begin{aligned}
A_1 &= \tfrac{1}{2}b_1h_1 \\
A_2 &= \tfrac{1}{2}b_2h_2 \\
&= \tfrac{1}{2}(2b_1)(2h_1) \\
&= 2b_1h_1 \\
A_2/A_1 &= (2b_1h_1)/(\tfrac{1}{2}b_1h_1) \\
&= 4
\end{aligned}$$

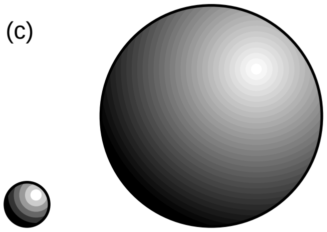
(Although this solution is correct, it is a lot more work than solution #1, and it can only be used in this case because a triangle is a simple geometric shape, and we happen to know a formula for its area.)

Correct solution #4: The area of a triangle is $A = \frac{1}{2}bh$. The comparison of the areas will come out the same as long as the ratios of the linear sizes of the triangles is as specified, so let's just say b_1=1.00 m and b_2=2.00 m. The heights are then also h_1=1.00 m and h_2=2.00 m, giving areas A_1=0.50 m^2 and A_2=2.00 m^2, so A_2/A_1=4.00.

(The solution is correct, but it wouldn't work with a shape for whose area we don't have a formula. Also, the numerical calculation might make the answer of 4.00 appear inexact, whereas solution #1 makes it clear that it is exactly 4.)

Incorrect solution: The area of a triangle is $A = \frac{1}{2}bh$, and if you plug in b=2.00 m and h=2.00 m, you get A=2.00 m^2, so the bigger triangle has 2.00 times more area. (This solution is incorrect because no comparison has been made with the smaller triangle.)

(c)



The big sphere has 125 times more volume than the little one.

Example: scaling of the volume of a sphere

Question: In figure (c), the larger sphere has a radius that is five times greater. How many times greater is its volume?

Correct solution #1: Volume scales like the third power of the linear size, so the larger sphere has a volume that is 125 times greater (5^3=125).

Correct solution #2: The volume of a sphere is $V=\frac{4}{3}\pi r^3$, so

$$\begin{aligned}
V_1 &= \tfrac{4}{3}\pi r_1^3 \\
V_2 &= \tfrac{4}{3}\pi r_2^3 \\
&= \tfrac{4}{3}\pi(5r_1)^3 \\
&= \tfrac{500}{3}\pi r_1^3 \\
V_2/V_1 &= \left(\frac{500}{3}\pi r_1^3\right) / \left(\frac{4}{3}\pi r_1^3\right) \\
&= 125
\end{aligned}$$

Incorrect solution: The volume of a sphere is $V=\frac{4}{3}\pi r^3$, so

$$\begin{aligned}
V_1 &= \tfrac{4}{3}\pi r_1^3 \\
V_2 &= \tfrac{4}{3}\pi r_2^3 \\
&= \tfrac{4}{3}\pi \cdot 5r_1^3 \\
&= \tfrac{20}{3}\pi r_1^3 \\
V_2/V_1 &= (\tfrac{20}{3}\pi r_1^3)/(\tfrac{4}{3}\pi r_1^3) \\
&= 5
\end{aligned}$$

(The solution is incorrect because $(5r_1)^3$ is not the same as $5r_1^3$.)

(d) The 48-point “S” has 1.78 times more area than the 36-point “S.”

Example: scaling of a more complex shape

Question: The first letter “S” in fig. (d) is in a 36-point font, the second in 48-point. How many times more ink is required to make the larger “S”?

Correct solution: The amount of ink depends on the area to be covered with ink, and area is proportional to the square of the linear dimensions, so the amount of ink required for the second “S” is greater by a factor of $(48/36)^2$=1.78.

Incorrect solution: The length of the curve of the second “S” is longer by a factor of 48/36=1.33, so 1.33 times more ink is required.

(The solution is wrong because it assumes incorrectly that the width of the curve is the same in both cases. Actually both the width and the length of the curve are greater by a factor of 48/36, so the area is greater by a factor of $(48/36)^2$=1.78.)

Discussion questions

A. A toy fire engine is 1/30 the size of the real one, but is constructed from the same metal with the same proportions. How many times smaller is its weight? How many times less red paint would be needed to paint it?

B. Galileo spends a lot of time in his dialog discussing what really happens when things break. He discusses everything in terms of Aristotle's now-discredited explanation that things are hard to break, because if something breaks, there has to be a gap between the two halves with nothing in between, at least initially. Nature, according to Aristotle, "abhors a vacuum," i.e. nature doesn't "like" empty space to exist. Of course, air will rush into the gap immediately, but at the very moment of breaking, Aristotle imagined a vacuum in the gap. Is Aristotle's explanation of why it is hard to break things an experimentally testable statement? If so, how could it be tested experimentally?

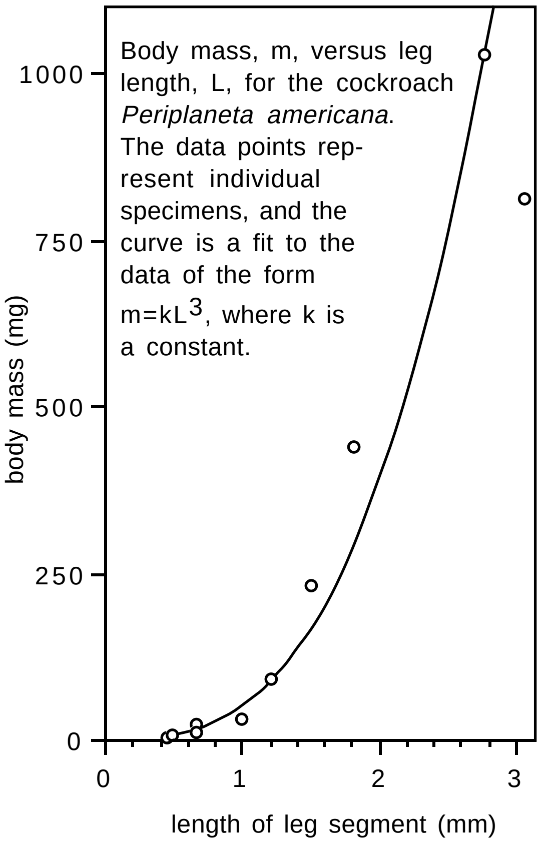
1.3 Scaling Applied to Biology

Organisms of different sizes with the same shape

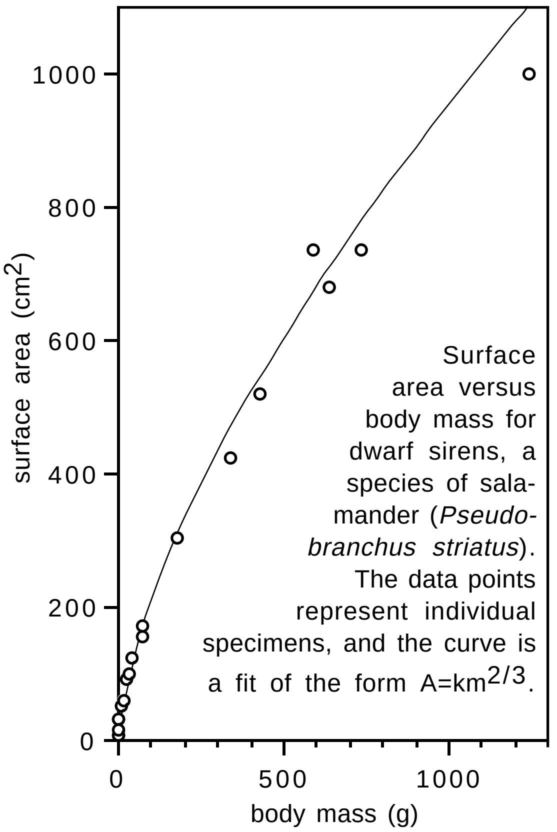
The first of the following graphs shows the approximate validity of the proportionality $m\propto L^3$ for cockroaches (redrawn from McMahon and Bonner). The scatter of the points around the curve indicates that some cockroaches are proportioned slightly differently from others, but in general the data seem well described by $m\propto L^3$. That means that the largest cockroaches the experimenter could raise (is there a 4-H prize?) had roughly the same shape as the smallest ones.

Another relationship that should exist for animals of different sizes shaped in the same way is that between surface area and body mass. If all the animals have the same average density, then body mass should be proportional to the cube of the animal's linear size, $m\propto L^3$, while surface area should vary proportionately to L^2. Therefore, the animals' surface areas should be proportional to $m^{2/3}$. As shown in the second graph, this relationship appears to hold quite well for the dwarf siren, a type of salamander. Notice how the curve bends over, meaning that the surface area does not increase as quickly as body mass, e.g. a salamander with eight times more body mass will have only four times more surface area.

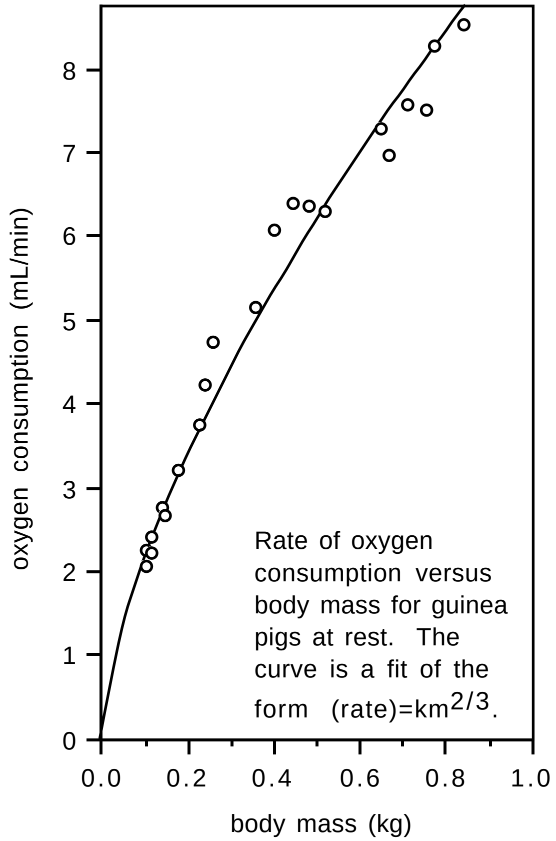
This behavior of the ratio of surface area to mass (or, equivalently, the ratio of surface area to volume) has important consequences for mammals, which must maintain a constant body temperature. It would make sense for the rate of heat loss through the animal's skin to be proportional to its surface area, so we should expect small animals, having large ratios of surface area to volume, to need to produce a great deal of heat in comparison to their size to avoid dying from low body temperature. This expectation is borne out by the data of the third graph, showing the rate of oxygen consumption of guinea pigs as a function of their body mass. Neither an animal's heat production nor its surface area is convenient to measure, but in order to produce heat, the animal must metabolize oxygen, so oxygen consumption is a good indicator of the rate of heat production. Since surface area is proportional to $m^{2/3}$, the proportionality of the rate of oxygen consumption to $m^{2/3}$ is consistent with the idea that the animal needs to produce heat at a rate in proportion to its surface area. Although the smaller animals metabolize less oxygen and produce less heat in absolute terms, the amount of food and oxygen they must consume is greater in proportion to their own mass. The Etruscan pigmy shrew, weighing in at 2 grams as an



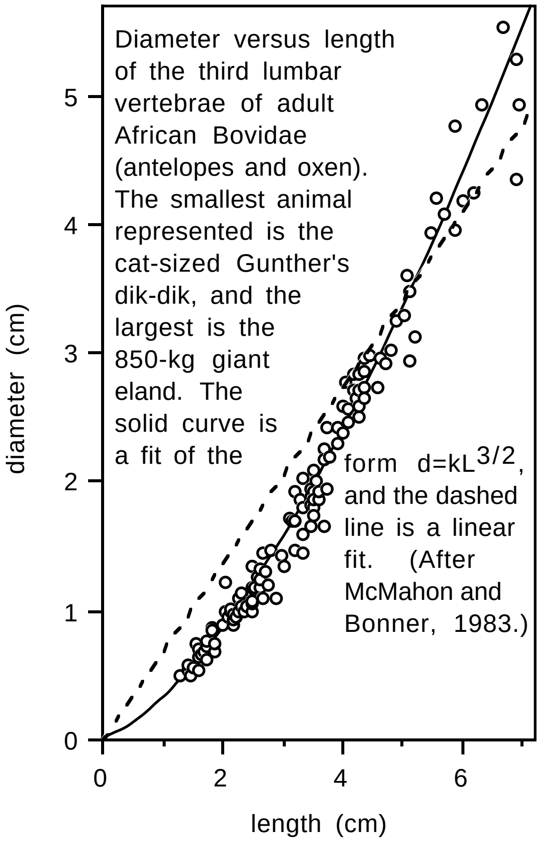
Body mass, m, versus leg length, L, for the cockroach *Periplaneta americana*. The data points represent individual specimens, and the curve is a fit to the data of the form $m=kL^3$, where k is a constant.



Surface area versus body mass for dwarf sirens, a species of salamander (*Pseudobranchus striatus*). The data points represent individual specimens, and the curve is a fit of the form $A=km^{2/3}$.



Rate of oxygen consumption versus body mass for guinea pigs at rest. The curve is a fit of the form (rate)=$km^{2/3}$.

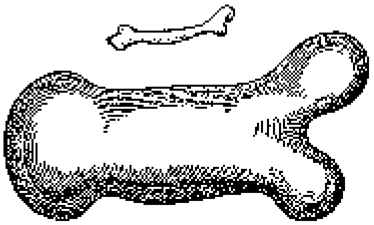


Diameter versus length of the third lumbar vertebrae of adult African Bovidae (antelopes and oxen). The smallest animal represented is the cat-sized Gunther's dik-dik, and the largest is the 850-kg giant eland. The solid curve is a fit of the form $d=kL^{3/2}$, and the dashed line is a linear fit. (After McMahon and Bonner, 1983.)

adult, is at about the lower size limit for mammals. It must eat continually, consuming many times its body weight each day to survive.

Changes in shape to accommodate changes in size

Large mammals, such as elephants, have a small ratio of surface area to volume, and have problems getting rid of their heat fast enough. An elephant cannot simply eat small enough amounts to keep from producing excessive heat, because cells need to have a certain minimum metabolic rate to run their internal machinery. Hence the elephant's large ears, which add to its surface area and help it to cool itself. Previously, we have seen several examples of data within a given species that were consistent with a fixed shape, scaled up and down in the cases of individual specimens. The elephant's ears are an example of a change in shape necessitated by a change in scale.



Galileo's original drawing, showing how larger animals' bones must be greater in diameter compared to their lengths.

Large animals also must be able to support their own weight. Returning to the example of the strengths of planks of different sizes, we can see that if the strength of the plank depends on area while its weight depends on volume, then the ratio of strength to weight goes as follows:

$$\text{strength/weight} \propto A/V \propto 1/L \quad .$$

Thus, the ability of objects to support their own weights decreases inversely in proportion to their linear dimensions. If an object is to be just barely able to support its own weight, then a larger version will have to be proportioned differently, with a different shape.

Since the data on the cockroaches seemed to be consistent with roughly similar shapes within the species, it appears that the ability to support its own weight was not the tightest design constraint that Nature was working under when she designed them. For large animals, structural strength is important. Galileo was the first to quantify this reasoning and to explain why, for instance, a large animal must have bones that are thicker in proportion to their length. Consider a roughly cylindrical bone such as a leg bone or a vertebra. The length of the bone, L, is dictated by the overall linear size of the animal, since the animal's skeleton must reach the animal's whole length. We expect the animal's mass to scale as L^3, so the strength of the bone must also scale as L^3. Strength is proportional to cross-sectional area, as with the wooden planks, so if the diameter of the bone is d, then

$$d^2 \propto L^3$$

or

$$d \propto L^{3/2} \quad .$$

If the shape stayed the same regardless of size, then all linear dimensions, including d and L, would be proportional to one another. If our reasoning holds, then the fact that d is proportional to $L^{3/2}$, not L, implies a change in proportions of the bone. As shown in the graph on the previous page, the vertebrae of African Bovidae follow the rule $d \propto L^{3/2}$ fairly well. The vertebrae of the giant eland are as chunky as a coffee mug, while those of a Gunther's dik-dik are as slender as the cap of a pen.

Discussion questions

A. Single-celled animals must passively absorb nutrients and oxygen from their surroundings, unlike humans who have lungs to pump air in and out and a heart to distribute the oxygenated blood throughout their bodies. Even the cells composing the bodies of multicellular animals must absorb oxygen from a nearby capillary through their surfaces. Based on these facts, explain why cells are always microscopic in size.

B. The reasoning of the previous question would seem to be contradicted by the fact that human nerve cells in the spinal cord can be as much as a meter long, although their widths are still very small. Why is this possible?

1.4 Order-of-Magnitude Estimates

> It is the mark of an instructed mind to rest satisfied with the degree of precision that the nature of the subject permits and not to seek an exactness where only an approximation of the truth is possible.
>
> Aristotle

It is a common misconception that science must be exact. For instance, in the Star Trek TV series, it would often happen that Captain Kirk would ask Mr. Spock, "Spock, we're in a pretty bad situation. What do you think are our chances of getting out of here?" The scientific Mr. Spock would answer with something like, "Captain, I estimate the odds as 237.345 to one." In reality, he could not have estimated the odds with six significant figures of accuracy, but nevertheless one of the hallmarks of a person with a good education in science is the ability to make estimates that are likely to be at least somewhere in the right ballpark. In many such situations, it is often only necessary to get an answer that is off by no more than a factor of ten in either direction. Since things that differ by a factor of ten are said to differ by one order of magnitude, such an estimate is called an order-of-magnitude estimate. The tilde, ~, is used to indicate that things are only of the same order of magnitude, but not exactly equal, as in

$$\text{odds of survival} \sim 100 \text{ to one} .$$

The tilde can also be used in front of an individual number to emphasize that the number is only of the right order of magnitude.

Although making order-of-magnitude estimates seems simple and natural to experienced scientists, it's a mode of reasoning that is completely unfamiliar to most college students. Some of the typical mental steps can be illustrated in the following example.

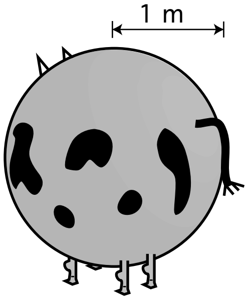
> *Example: Cost of transporting tomatoes*
> **Question**: Roughly what percentage of the price of a tomato comes from the cost of transporting it in a truck?

The following incorrect solution illustrates one of the main ways you can go wrong in order-of-magnitude estimates.

> **Incorrect solution**: Let's say the trucker needs to make a $400 profit on the trip. Taking into account her benefits, the cost of gas, and maintenance and payments on the truck, let's say the total cost is more like $2000. I'd guess about 5000 tomatoes would fit in the back of the truck, so the extra cost per tomato is 40 cents. That means the cost of transporting one tomato is comparable to the cost of the tomato itself. Transportation really adds a lot to the cost of produce, I guess.

The problem is that the human brain is not very good at estimating area or volume, so it turns out the estimate of 5000 tomatoes fitting in the truck is way off. That's why people have a hard time at those contests where you are supposed to estimate the number of jellybeans in a big jar. Another example is that most people think their families use about 10 gallons of water per day, but in reality the average is about 300 gallons per day. When estimating area or volume, you are much better off estimating linear dimensions, and computing volume from the linear dimensions. Here's a better solution:

> **Better solution**: As in the previous solution, say the cost of the trip is $2000. The dimensions of the bin are probably 4 m x 2 m x 1 m, for a volume of 8 m^3. Since the whole thing is just an order-of-magnitude estimate, let's round that off to the nearest power of ten, 10 m^3. The shape of a tomato is complicated, and I don't know any formula for the volume of a tomato shape, but since this is just an estimate, let's pretend that a tomato is a cube, 0.05 m x 0.05 m x 0.05, for a volume of $1.25\text{x}10^{-4}$ m^3. Since this is just a rough estimate, let's round that to 10^{-4} m^3. We can find the total number of tomatoes by dividing the volume of the bin by the volume of one tomato: 10 m^3 / 10^{-4} m^3 = 10^5 tomatoes. The transportation cost per tomato is $2000/$10^5$ tomatoes=$0.02/tomato. That means that transportation really doesn't contribute very much to the cost of a tomato.



Approximating the shape of a tomato as a cube is an example of another general strategy for making order-of-magnitude estimates. A similar situation would occur if you were trying to estimate how many m^2 of leather could be produced from a herd of ten thousand cattle. There is no point in trying to take into account the shape of the cows' bodies. A reasonable plan of attack might be to consider a spherical cow. Probably a cow has roughly the same surface area as a sphere with a radius of about 1 m, which would be $4\pi(1\text{ m})^2$. Using the well-known facts that pi equals three, and four times three equals about ten, we can guess that a cow has a surface area of about 10 m^2, so the herd as a whole might yield 10^5 m^2 of leather.

The following list summarizes the strategies for getting a good order-of-magnitude estimate.

(1) Don't even attempt more than one significant figure of precision.

(2) Don't guess area or volume directly. Guess linear dimensions and get area or volume from them.

(3) When dealing with areas or volumes of objects with complex shapes, idealize them as if they were some simpler shape, a cube or a sphere, for example.

(4) Check your final answer to see if it is reasonable. If you estimate that a herd of ten thousand cattle would yield 0.01 m^2 of leather, then you have probably made a mistake with conversion factors somewhere.

Summary

Notation

$\propto$ .... is proportional to
~ .... on the order of, is on the order of

Summary

Nature behaves differently on large and small scales. Galileo showed that this results fundamentally from the way area and volume scale. Area scales as the second power of length, $A \propto L^2$, while volume scales as length to the third power, $V \propto L^3$.

An order of magnitude estimate is one in which we do not attempt or expect an exact answer. The main reason why the uninitiated have trouble with order-of-magnitude estimates is that the human brain does not intuitively make accurate estimates of area and volume. Estimates of area and volume should be approached by first estimating linear dimensions, which one’s brain has a feel for.

Homework Problems

1 ✓. How many cubic inches are there in a cubic foot? The answer is not 12.

2. Assume a dog's brain is twice is great in diameter as a cat's, but each animal's brain cells are the same size and their brains are the same shape. In addition to being a far better companion and much nicer to come home to, how many times more brain cells does a dog have than a cat? The answer is not 2.

3 ✓. The population density of Los Angeles is about 4000 people/km^2. That of San Francisco is about 6000 people/km^2. How many times farther away is the average person's nearest neighbor in LA than in San Francisco? The answer is not 1.5.

4. A hunting dog's nose has about 10 square inches of active surface. How is this possible, since the dog's nose is only about 1 in x 1 in x 1 in = 1 in^3? After all, 10 is greater than 1, so how can it fit?

5. Estimate the number of blades of grass on a football field.

6. In a computer memory chip, each bit of information (a 0 or a 1) is stored in a single tiny circuit etched onto the surface of a silicon chip. A typical chip stores 64 Mb (megabytes) of data, where a byte is 8 bits. Estimate (a) the area of each circuit, and (b) its linear size.

7. Suppose someone built a gigantic apartment building, measuring 10 km x 10 km at the base. Estimate how tall the building would have to be to have space in it for the entire world's population to live.

8. A hamburger chain advertises that it has sold 10 billion Bongo Burgers. Estimate the total mass of hay required to feed the cows used to make the burgers.

9. Estimate the volume of a human body, in cm^3.

10 S. How many cm^2 is 1 mm^2?

11 S. Compare the light-gathering powers of a 3-cm-diameter telescope and a 30-cm telescope.

S A solution is given in the back of the book. ★ A difficult problem.
✓ A computerized answer check is available. ∫ A problem that requires calculus.

12. S. One step on the Richter scale corresponds to a factor of 100 in terms of the energy absorbed by something on the surface of the Earth, e.g. a house. For instance, a 9.3-magnitude quake would release 100 times more energy than an 8.3. The energy spreads out from the epicenter as a wave, and for the sake of this problem we'll assume we're dealing with seismic waves that spread out in three dimensions, so that we can visualize them as hemispheres spreading out under the surface of the earth. If a certain 7.6-magnitude earthquake and a certain 5.6-magnitude earthquake produce the same amount of vibration where I live, compare the distances from my house to the two epicenters.

Motion in One Dimension

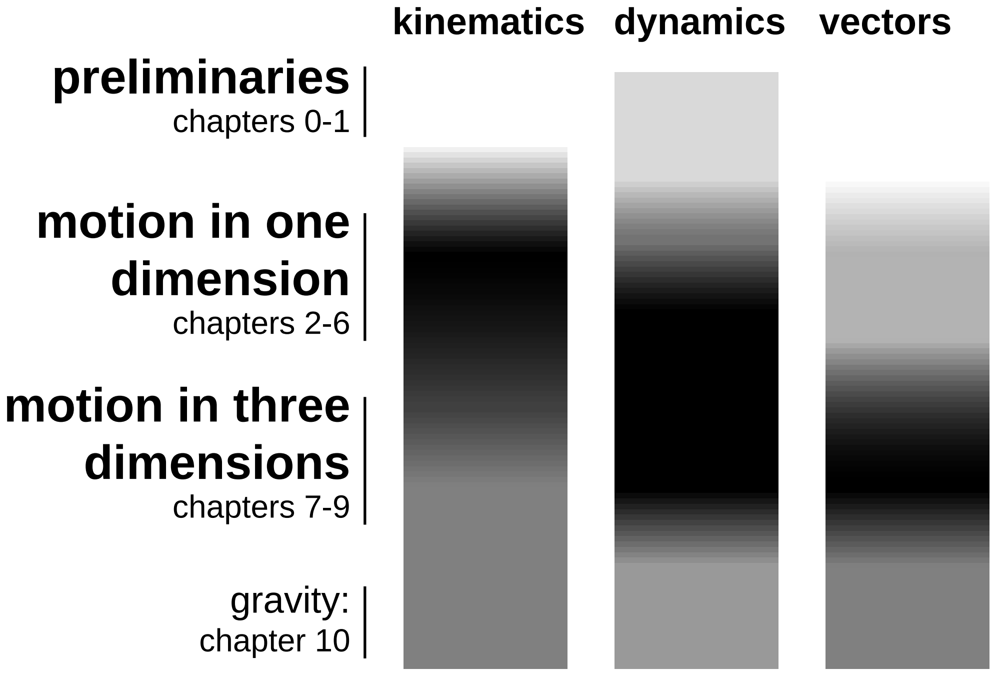
I didn't learn until I was nearly through with college that I could understand a book much better if I mentally outlined it for myself before I actually began reading. It's a technique that warns my brain to get little cerebral file folders ready for the different topics I'm going to learn, and as I'm reading it allows me to say to myself, "Oh, the reason they're talking about this now is because they're preparing for this other thing that comes later," or "I don't need to sweat the details of this idea now, because they're going to explain it in more detail later on."

At this point, you're about to dive in to the main subjects of this book, which are force and motion. The concepts you're going to learn break down into the following three areas:

kinematics — how to describe motion numerically
dynamics — how force affects motion
vectors — a mathematical way of handling the three-dimensional nature of force and motion

Roughly speaking, that's the order in which we'll cover these three areas, but the earlier chapters do contain quite a bit of preparation for the later topics. For instance, even before the present point in the book you've learned about the Newton, a unit of force. The discussion of force properly belongs to dynamics, which we aren't tackling head-on for a few more chapters, but I've found that when I teach kinematics it helps to be able to refer to forces now and then to show why it makes sense to define certain kinematical concepts. And although I don't explicitly introduce vectors until ch. 8, the groundwork is being laid for them in earlier chapters.

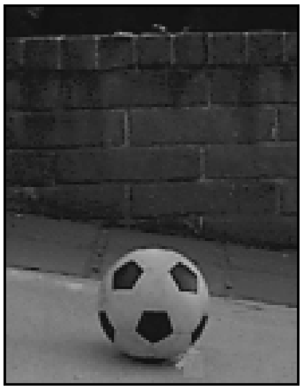
Here's a roadmap to the rest of the book:



2 Velocity and Relative Motion

2.1 Types of Motion

Rotation.



Simultaneous rotation and motion through space.

One person might say that the tipping chair was only rotating in a circle about its point of contact with the floor, but another could describe it as having both rotation and motion through space.

If you had to think consciously in order to move your body, you would be severely disabled. Even walking, which we consider to be no great feat, requires an intricate series of motions that your cerebrum would be utterly incapable of coordinating. The task of putting one foot in front of the other is controlled by the more primitive parts of your brain, the ones that have not changed much since the mammals and reptiles went their separate evolutionary ways. The thinking part of your brain limits itself to general directives such as “walk faster,” or “don’t step on her toes,” rather than micromanaging every contraction and relaxation of the hundred or so muscles of your hips, legs, and feet.

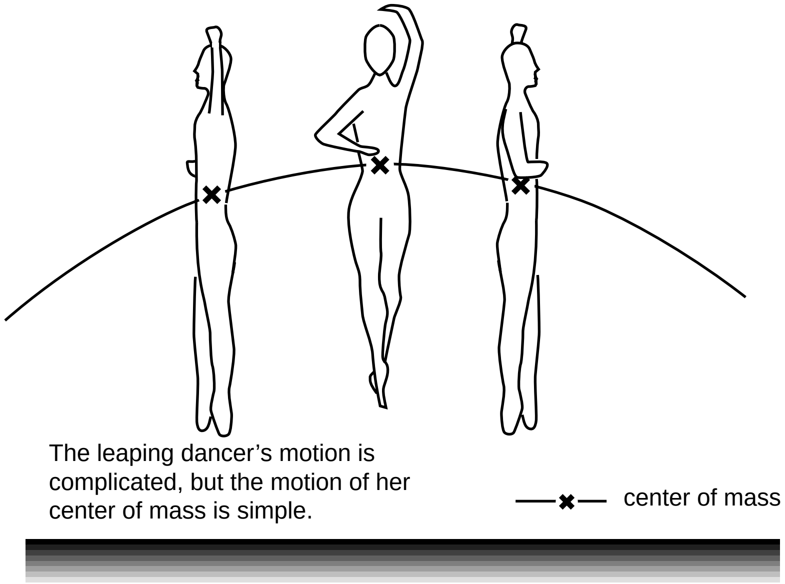
Physics is all about the conscious understanding of motion, but we’re obviously not immediately prepared to understand the most complicated types of motion. Instead, we’ll use the divide-and-conquer technique. We’ll first classify the various types of motion, and then begin our campaign with an attack on the simplest cases. To make it clear what we are and are not ready to consider, we need to examine and define carefully what types of motion can exist.

Rigid-body motion distinguished from motion that changes an object’s shape

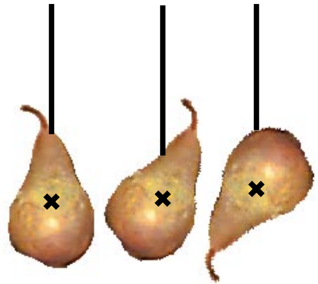
Nobody, with the possible exception of Fred Astaire, can simply glide forward without bending their joints. Walking is thus an example in which there is both a general motion of the whole object and a change in the shape of the object. Another example is the motion of a jiggling water balloon as it flies through the air. We are not presently attempting a mathematical description of the way in which the shape of an object changes. Motion without a change in shape is called rigid-body motion. (The word “body” is often used in physics as a synonym for “object.”)

Center-of-mass motion as opposed to rotation

A ballerina leaps into the air and spins around once before landing. We feel intuitively that her rigid-body motion while her feet are off the ground consists of two kinds of motion going on simultaneously: a rotation and a motion of her body as a whole through space, along an arc. It is not immediately obvious, however, what is the most useful way to define the distinction between rotation and motion through space. Imagine that you attempt to balance a chair and it falls over. One person might say that the only motion was a rotation about the chair’s point of contact with the floor, but another might say that there was both rotation and motion down and to the side.



The leaping dancer's motion is complicated, but the motion of her center of mass is simple.

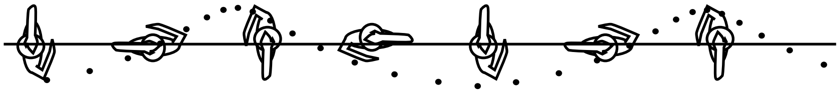


No matter what point you hang the pear from, the string lines up with the pear's center of mass. The center of mass can therefore be defined as the intersection of all the lines made by hanging the pear in this way. Note that the X in the figure should not be interpreted as implying that the center of mass is on the surface — it is actually inside the pear.

It turns out that there is one particularly natural and useful way to make a clear definition, but it requires a brief digression. Every object has a balance point, referred to in physics as the *center of mass*. For a two-dimensional object such as a cardboard cutout, the center of mass is the point at which you could hang the object from a string and make it balance. In the case of the ballerina (who is likely to be three-dimensional unless her diet is particularly severe), it might be a point either inside or outside her body, depending on how she holds her arms. Even if it is not practical to attach a string to the balance point itself, the center of mass can be defined as shown in the figure on the left.

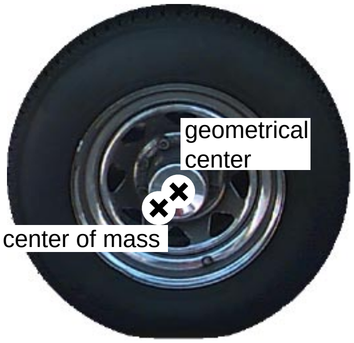
The motion of an object's center of mass is usually much simpler than the motion of any other point on it.

Why is the center of mass concept relevant to the question of classifying rotational motion as opposed to motion through space? As illustrated in the figure above, it turns out that the motion of an object's center of mass is nearly always far simpler than the motion of any other part of the object. The ballerina's body is a large object with a complex shape. We might expect that her motion would be much more complicated that the motion of a small, simply-shaped object, say a marble, thrown up at the same angle as the angle at which she leapt. But it turns out that the motion of the ballerina's center of mass is exactly the same as the motion of the marble. That is, the motion of the center of mass is the same as the motion the ballerina would have if all her mass was concentrated at a point. By restricting our attention to the motion of the center of mass, we can therefore simplify things greatly.



The same leaping dancer, viewed from above. Her center of mass traces a straight line, but a point away from her center of mass, such as her elbow, traces the much more complicated path shown by the dots.

We can now replace the ambiguous idea of "motion as a whole through space" with the more useful and better defined concept of *"center-of-mass motion."* The motion of any rigid body can be cleanly split into rotation and center-of-mass motion. By this definition, the tipping chair does have both rotational and center-of-mass motion. Concentrating on the center of

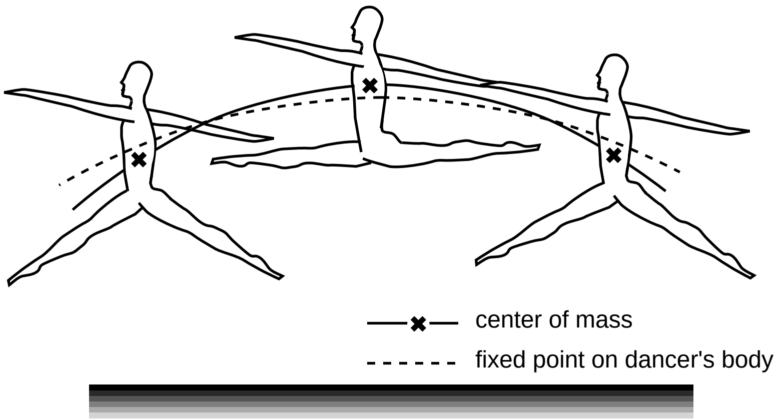


An improperly balanced wheel has a center of mass that is not at its geometric center. When you get a new tire, the mechanic clamps little weights to the rim to balance the wheel.

mass motion allows us to make a simplified *model* of the motion, as if a complicated object like a human body was just a marble or a point-like particle. Science really never deals with reality; it deals with models of reality.

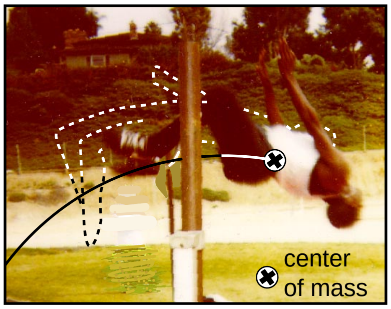
Note that the word "center" in "center of mass" is not meant to imply that the center of mass must lie at the geometrical center of an object. A car wheel that has not been balanced properly has a center of mass that does not coincide with its geometrical center. An object such as the human body does not even have an obvious geometrical center.

It can be helpful to think of the center of mass as the average location of all the mass in the object. With this interpretation, we can see for example that raising your arms above your head raises your center of mass, since the higher position of the arms' mass raises the average.



A fixed point on the dancer's body follows a trajectory that is flatter than what we expect, creating an illusion of flight.

Ballerinas and professional basketball players can create an illusion of flying horizontally through the air because our brains intuitively expect them to have rigid-body motion, but the body does not stay rigid while executing a grand jete or a slam dunk. The legs are low at the beginning and end of the jump, but come up higher at the middle. Regardless of what the limbs do, the center of mass will follow the same arc, but the low position of the legs at the beginning and end means that the torso is higher compared to the center of mass, while in the middle of the jump it is lower compared to the center of mass. Our eye follows the motion of the torso and tries to interpret it as the center-of-mass motion of a rigid body. But since the torso follows a path that is flatter than we expect, this attempted interpretation fails, and we experience an illusion that the person is flying horizontally. Another interesting example from the sports world is the high jump, in which the jumper's curved body passes over the bar, but the center of mass passes under the bar! Here the jumper lowers his legs and upper body at the peak of the jump in order to bring his waist higher compared to the center of mass.



The high-jumper's body passes over the bar, but his center of mass passes under it.

Photo by Dunia Young.

Later in this course, we'll find that there are more fundamental reasons (based on Newton's laws of motion) why the center of mass behaves in such a simple way compared to the other parts of an object. We're also postponing any discussion of numerical methods for finding an object's center of mass. Until later in the course, we will only deal with the motion of objects' centers of mass.

Center-of-mass motion in one dimension

In addition to restricting our study of motion to center-of-mass motion, we will begin by considering only cases in which the center of mass moves along a straight line. This will include cases such as objects falling straight down, or a car that speeds up and slows down but does not turn.

Note that even though we are not explicitly studying the more complex aspects of motion, we can still analyze the center-of-mass motion while ignoring other types of motion that might be occurring simultaneously . For instance, if a cat is falling out of a tree and is initially upside-down, it goes through a series of contortions that bring its feet under it. This is definitely not an example of rigid-body motion, but we can still analyze the motion of the cat's center of mass just as we would for a dropping rock.

eck

Consider a person running, a person pedaling on a bicycle, a person coasting on a bicycle, and a person coasting on ice skates. In which cases is the center-of-mass motion one-dimensional? Which cases are examples of rigid-body motion?

2.2 Describing Distance and Time

Center-of-mass motion in one dimension is particularly easy to deal with because all the information about it can be encapsulated in two variables: x, the position of the center of mass relative to the origin, and t, which measures a point in time. For instance, if someone supplied you with a sufficiently detailed table of x and t values, you would know pretty much all there was to know about the motion of the object's center of mass.

A point in time as opposed to duration

In ordinary speech, we use the word "time" in two different senses, which are to be distinguished in physics. It can be used, as in "a short time" or "our time here on earth," to mean a length or duration of time, or it can be used to indicate a clock reading, as in "I didn't know what time it was," or "now's the time." In symbols, t is ordinarily used to mean a point in time, while Δt signifies an interval or duration in time. The capital Greek letter delta, Δ, means "the change in...," i.e. a duration in time is the change or difference between one clock reading and another. The notation Δt does not signify the product of two numbers, Δ and t, but rather one single number, Δt. If a matinee begins at a point in time t=1 o'clock and ends at t=3 o'clock, the duration of the movie was the change in t,

$$\Delta t = 3 \text{ hours} - 1 \text{ hour} = 2 \text{ hours} \quad .$$

To avoid the use of negative numbers for Δt, we write the clock reading "after" to the left of the minus sign, and the clock reading "before" to the right of the minus sign. A more specific definition of the delta notation is therefore that delta stands for "after minus before."

Even though our definition of the delta notation guarantees that Δt is positive, there is no reason why t can't be negative. If t could not be negative, what would have happened one second before t=0? That doesn't mean

Coasting on a bike and coasting on skates give one-dimensional center-of-mass motion, but running and pedaling require moving body parts up and down, which makes the center of mass move up and down. The only example of rigid-body motion is coasting on skates. (Coasting on a bike is not rigid-body motion, because the wheels twist.)

that time “goes backward” in the sense that adults can shrink into infants and retreat into the womb. It just means that we have to pick a reference point and call it t=0, and then times before that are represented by negative values of t.

Although a point in time can be thought of as a clock reading, it is usually a good idea to avoid doing computations with expressions such as “2:35” that are combinations of hours and minutes. Times can instead be expressed entirely in terms of a single unit, such as hours. Fractions of an hour can be represented by decimals rather than minutes, and similarly if a problem is being worked in terms of minutes, decimals can be used instead of seconds.

Self-Check

Of the following phrases, which refer to points in time, which refer to time intervals, and which refer to time in the abstract rather than as a measurable number?

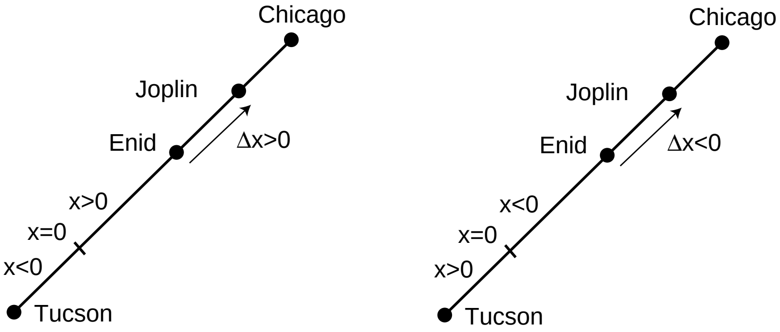
(a) “The time has come.”
(b) “Time waits for no man.”
(c) “The whole time, he had spit on his chin.”

Position as opposed to change in position

As with time, a distinction should be made between a point in space, symbolized as a coordinate x, and a change in position, symbolized as Δx.

As with t, x can be negative. If a train is moving down the tracks, not only do you have the freedom to choose any point along the tracks and call it x=0, but it’s also up to you to decide which side of the x=0 point is positive x and which side is negative x.

Since we’ve defined the delta notation to mean “after minus before,” it is possible that Δx will be negative, unlike Δt which is guaranteed to be positive. Suppose we are describing the motion of a train on tracks linking Tucson and Chicago. As shown in the figure, it is entirely up to you to decide which way is positive.



Two equally valid ways of describing the motion of a train from Tucson to Chicago. In the first example, the train has a positive Δx as it goes from Enid to Joplin. In the second example, the same train going forward in the same direction has a negative Δx.

(a) a point in time; (b) time in the abstract sense; (c) a time interval

Note that in addition to x and Δx, there is a third quantity we could define, which would be like an odometer reading, or actual distance traveled. If you drive 10 miles, make a U-turn, and drive back 10 miles, then your Δx is zero, but your car’s odometer reading has increased by 20 miles. However important the odometer reading is to car owners and used car dealers, it is not very important in physics, and there is not even a standard name or notation for it. The change in position, Δx, is more useful because it is so much easier to calculate: to compute Δx, we only need to know the beginning and ending positions of the object, not all the information about how it got from one position to the other.

Self-Check

A ball hits the floor, bounces to a height of one meter, falls, and hits the floor again. Is the Δx between the two impacts equal to zero, one, or two meters?

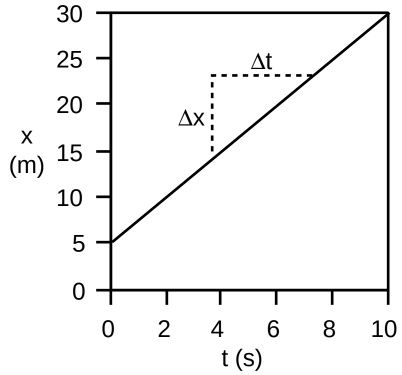
Frames of reference

The example above shows that there are two arbitrary choices you have to make in order to define a position variable, x. You have to decide where to put x=0, and also which direction will be positive. This is referred to as choosing a *coordinate system* or choosing a *frame of reference*. (The two terms are nearly synonymous, but the first focuses more on the actual x variable, while the second is more of a general way of referring to one’s point of view.) As long as you are consistent, any frame is equally valid. You just don’t want to change coordinate systems in the middle of a calculation.

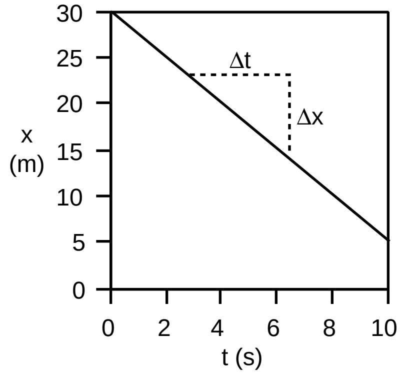
Have you ever been sitting in a train in a station when suddenly you notice that the station is moving backward? Most people would describe the situation by saying that you just failed to notice that the train was moving — it only seemed like the station was moving. But this shows that there is yet a third arbitrary choice that goes into choosing a coordinate system: valid frames of reference can differ from each other by moving relative to one another. It might seem strange that anyone would bother with a coordinate system that was moving relative to the earth, but for instance the frame of reference moving along with a train might be far more convenient for describing things happening inside the train.

Zero, because the “after” and “before” values of x are the same.

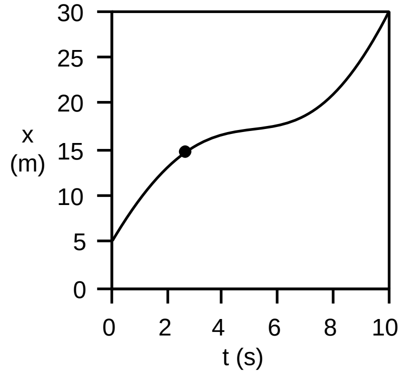
2.3 Graphs of Motion; Velocity.



(a) Motion with constant velocity.



(b) Motion that decreases *x* is represented with negative values of Δ*x* and *v*.



(c) Motion with changing velocity.

Motion with constant velocity

In example (a), an object is moving at constant speed in one direction. We can tell this because every two seconds, its position changes by five meters.

In algebra notation, we'd say that the graph of *x* vs. *t* shows the same change in position, Δ*x*=5.0 m, over each interval of Δ*t*=2.0 s. The object's velocity or speed is obtained by calculating *v*=Δ*x*/Δ*t*=(5.0 m)/(2.0 s)=2.5 m/s. In graphical terms, the velocity can be interpreted as the slope of the line. Since the graph is a straight line, it wouldn't have mattered if we'd taken a longer time interval and calculated *v*=Δ*x*/Δ*t*=(10.0 m)/(4.0 s). The answer would still have been the same, 2.5 m/s.

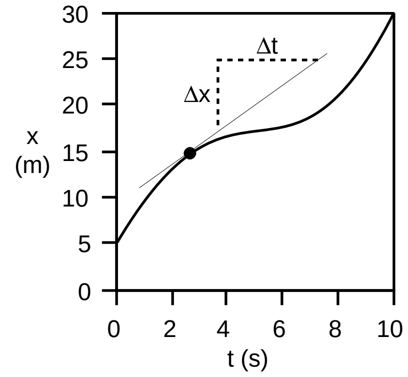
Note that when we divide a number that has units of meters by another number that has units of seconds, we get units of meters per second, which can be written m/s. This is another case where we treat units as if they were algebra symbols, even though they're not.

In example (b), the object is moving in the opposite direction: as time progresses, its x coordinate decreases. Recalling the definition of the Δ notation as "after minus before," we find that Δ*t* is still positive, but Δ*x* must be negative. The slope of the line is therefore negative, and we say that the object has a negative velocity, *v*=Δ*x*/Δ*t*=(-5.0 m)/(2.0 s)=-2.5 m/s. We've already seen that the plus and minus signs of Δ*x* values have the interpretation of telling us which direction the object moved. Since Δ*t* is always positive, dividing by Δ*t* doesn't change the plus or minus sign, and the plus and minus signs of velocities are to be interpreted in the same way. In graphical terms, a positive slope characterizes a line that goes up as we go to the right, and a negative slope tells us that the line went down as we went to the right.

Motion with changing velocity

Now what about a graph like example (c)? This might be a graph of a car's motion as the driver cruises down the freeway, then slows down to look at a car crash by the side of the road, and then speeds up again, disappointed that there is nothing dramatic going on such as flames or babies trapped in their car seats. (Note that we are still talking about one-dimensional motion. Just because the graph is curvy doesn't mean that the car's path is curvy. The graph is not like a map, and the horizontal direction of the graph represents the passing of time, not distance.)

Example (c) is similar to example (a) in that the object moves a total of 25.0 m in a period of 10.0 s, but it is no longer true that it makes the same amount of progress every second. There is no way to characterize the entire graph by a certain velocity or slope, because the velocity is different at every moment. It would be incorrect to say that because the car covered 25.0 m in 10.0 s, its velocity was 2.5 m/s . It moved faster than that at the beginning and end, but slower in the middle. There may have been certain instants at which the car was indeed going 2.5 m/s, but the speedometer swept past that value without "sticking," just as it swung through various other values of speed. (I definitely want my next car to have a speedometer calibrated in m/s and showing both negative and positive values.)



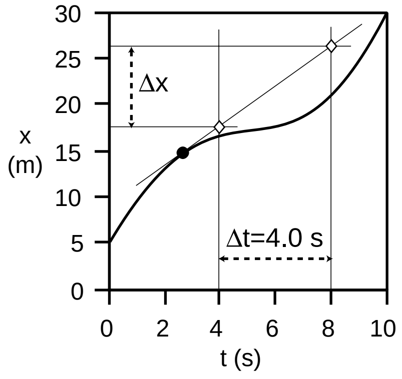
(d) The velocity at any given moment is defined as the slope of the tangent line through the relevant point on the graph.

We assume that our speedometer tells us what is happening to the speed of our car at every instant, but how can we define speed mathematically in a case like this? We can't just define it as the slope of the curvy graph, because a curve doesn't have a single well-defined slope as does a line. A mathematical definition that corresponded to the speedometer reading would have to be one that attached a different velocity value to a single point on the curve, i.e. a single instant in time, rather than to the entire graph. If we wish to define the speed at one instant such as the one marked with a dot, the best way to proceed is illustrated in (d), where we have drawn the line through that point called the tangent line, the line that "hugs the curve." We can then adopt the following definition of velocity:

definition of velocity

The velocity of an object at any given moment is the slope of the tangent line through the relevant point on its *x*-*t* graph.

One interpretation of this definition is that the velocity tells us how many meters the object would have traveled in one second, if it had continued moving at the same speed for at least one second. To some people the graphical nature of this definition seems "inaccurate" or "not mathematical." The equation $v=\Delta x/\Delta t$ by itself, however, is only valid if the velocity is constant, and so cannot serve as a general definition.



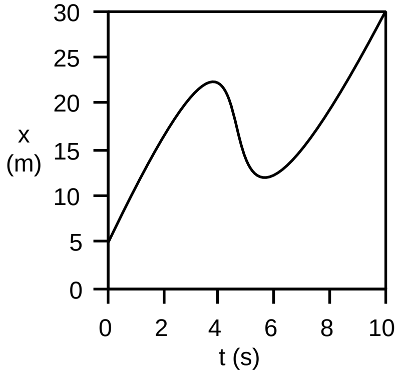
Example: finding the velocity at the point indicated with the dot.

Example

Question: What is the velocity at the point shown with a dot on the graph?

Solution: First we draw the tangent line through that point. To find the slope of the tangent line, we need to pick two points on it. Theoretically, the slope should come out the same regardless of which two points we picked, but in practical terms we'll be able to measure more accurately if we pick two points fairly far apart, such as the two white diamonds. To save work, we pick points that are directly above labeled points on the t axis, so that Δt=4.0 s is easy to read off. One diamond lines up with $x\approx$17.5 m, the other with $x\approx$26.5 m, so Δx=9.0 m. The velocity is $\Delta x/\Delta t$=2.2 m/s.

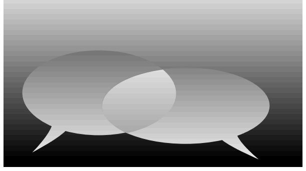
Conventions about graphing



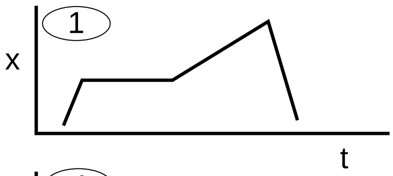
(e) Reversing the direction of motion.

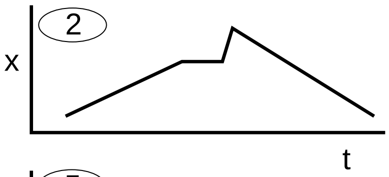
The placement of *t* on the horizontal axis and *x* on the upright axis may seem like an arbitrary convention, or may even have disturbed you, since your algebra teacher always told you that *x* goes on the horizontal axis and *y* goes on the upright axis. There is a reason for doing it this way, however. In example (e), we have an object that reverses its direction of motion twice. It can only be in one place at any given time, but there can be more than one time when it is at a given place. For instance, this object passed through *x*=17 m on three separate occasions, but there is no way it could have been in more than one place at *t*=5.0 s. Resurrecting some terminology you learned in your trigonometry course, we say that *x* is a function of *t*, but *t* is not a function of *x*. In situations such as this, there is a useful convention that the graph should be oriented so that any vertical line passes through the curve at only one point. Putting the *x* axis across the page and *t* upright would have violated this convention. To people who are used to interpreting graphs, a graph that violates this convention is as annoying as fingernails scratching on a chalkboard. We say that this is a graph of "*x* versus *t*." If the axes were the other way around, it would be a graph of "*t* versus *x*." I remember the "versus" terminology by visualizing the labels on the *x* and *t* axes and remembering that when you read, you go from left to right and from top to bottom.

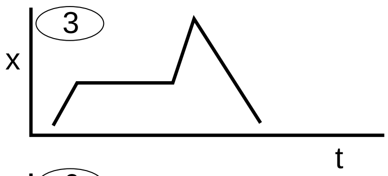
Discussion questions

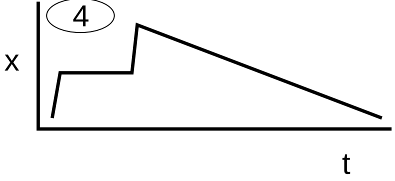


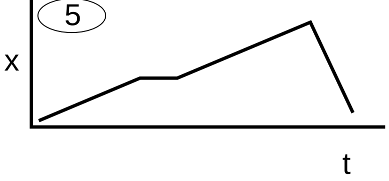
A. An ant walks forward, pauses, then runs quickly ahead. It then suddenly reverses direction and walks slowly back the way it came. Which graph could represent its motion?

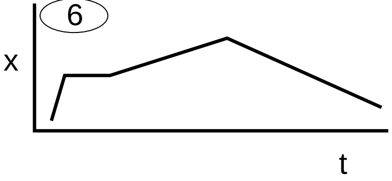




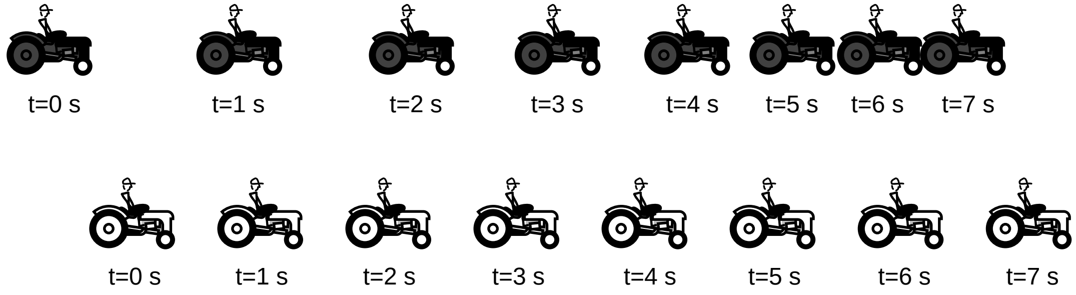








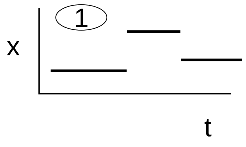
B. The figure shows a sequence of positions for two racing tractors. Compare the tractors' velocities as the race progresses.

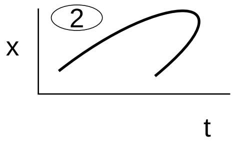


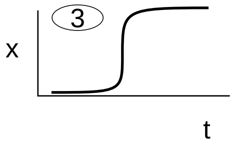
C. If an object had a straight-line motion graph with $\Delta x=0$ and $\Delta t\neq 0$, what would be true about its velocity? What would this look like on a graph? What about $\Delta t=0$ and $\Delta x\neq 0$?

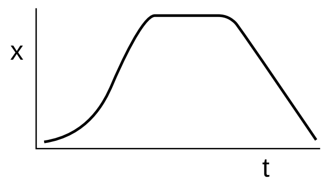
D. If an object has a wavy motion graph like the one in example (e) on the previous page, which are the points at which the object reverses its direction? What is true about the object's velocity at these points?

E. Discuss anything unusual about the following three graphs.









Discussion question G.

F. I have been using the term “velocity” and avoiding the more common English word “speed,” because some introductory physics texts define them to mean different things. They use the word “speed,” and the symbol “s” to mean the absolute value of the velocity, $s=|v|$. Although I have thrown in my lot with the minority of books that don’t emphasize this distinction in technical vocabulary, there are clearly two different concepts here. Can you think of an example of a graph of x vs. t in which the object has constant speed, but not constant velocity?

G. In the graph on the left, describe how the object’s velocity changes.

H. Two physicists duck out of a boring scientific conference to go get beer. On the way to the bar, they witness an accident in which a pedestrian is injured by a hit-and-run driver. A criminal trial results, and they must testify. In her testimony, Dr. Transverz Waive says, “The car was moving along pretty fast, I’d say the velocity was +40 mi/hr. They saw the old lady too late, and even though they slammed on the brakes they still hit her before they stopped. Then they made a U turn and headed off at a velocity of about -20 mi/hr, I’d say.” Dr. Longitud N.L. Vibrasheun says, “He was really going too fast, maybe his velocity was -35 or -40 mi/hr. After he hit Mrs. Hapless, he turned around and left at a velocity of, oh, I’d guess maybe +20 or +25 mi/hr.” Is their testimony contradictory? Explain.

2.4 The Principle of Inertia

Physical effects relate only to a change in velocity

Consider two statements that were at one time made with the utmost seriousness:

> People like Galileo and Copernicus who say the earth is rotating must be crazy. We know the earth can't be moving. Why, if the earth was really turning once every day, then our whole city would have to be moving hundreds of leagues in an hour. That's impossible! Buildings would shake on their foundations. Gale-force winds would knock us over. Trees would fall down. The Mediterranean would come sweeping across the east coasts of Spain and Italy. And furthermore, what force would be making the world turn?
>
> All this talk of passenger trains moving at forty miles an hour is sheer hogwash! At that speed, the air in a passenger compartment would all be forced against the back wall. People in the front of the car would suffocate, and people at the back would die because in such concentrated air, they wouldn't be able to expel a breath.

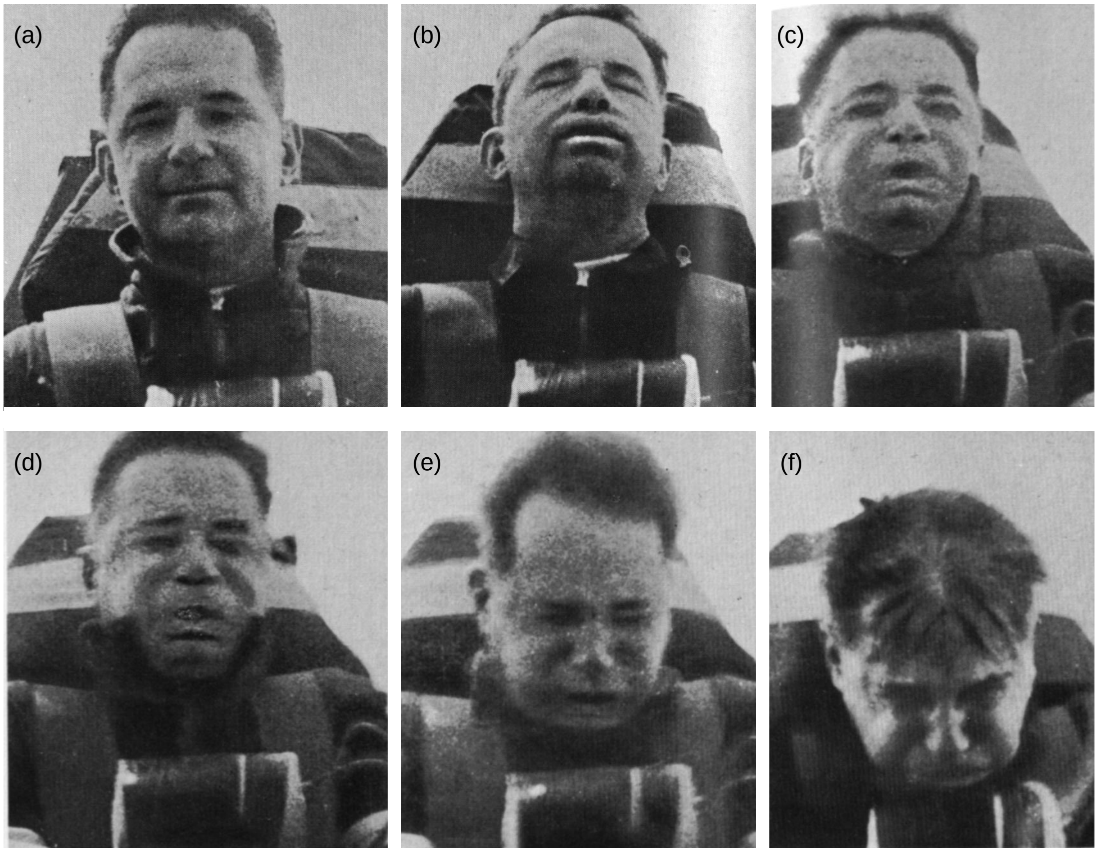
Some of the effects predicted in the first quote are clearly just based on a lack of experience with rapid motion that is smooth and free of vibration. But there is a deeper principle involved. In each case, the speaker is assuming that the mere fact of motion must have dramatic physical effects. More subtly, they also believe that a force is needed to keep an object in motion: the first person thinks a force would be needed to maintain the earth's rotation, and the second apparently thinks of the rear wall as pushing on the air to keep it moving.

Common modern knowledge and experience tell us that these people's predictions must have somehow been based on incorrect reasoning, but it is not immediately obvious where the fundamental flaw lies. It's one of those things a four-year-old could infuriate you by demanding a clear explanation of. One way of getting at the fundamental principle involved is to consider how the modern concept of the universe differs from the popular conception at the time of the Italian Renaissance. To us, the word "earth" implies a planet, one of the nine planets of our solar system, a small ball of rock and dirt that is of no significance to anyone in the universe except for members of our species, who happen to live on it. To Galileo's contemporaries, however, the earth was the biggest, most solid, most important thing in all of creation, not to be compared with the wandering lights in the sky known as planets. To us, the earth is just another object, and when we talk loosely about "how fast" an object such as a car "is going," we really mean the car-object's velocity relative to the earth-object.

Motion is relative

There is nothing special about motion or lack of motion relative to the planet earth.

According to our modern world-view, it really isn't that reasonable to expect that a special force should be required to make the air in the train have a certain velocity relative to our planet. After all, the "moving" air in the "moving" train might just happen to have zero velocity relative to some other planet we don't even know about. Aristotle claimed that things "naturally" wanted to be at rest, lying on the surface of the earth. But experiment after experiment has shown that there is really nothing so



This Air Force doctor volunteered to ride a rocket sled as a medical experiment. The obvious effects on his head and face are not because of the sled's speed but because of its rapid changes in speed: increasing in (b) and (c), and decreasing in (e) and (f).In (d) his speed is greatest, but because his speed is not increasing or decreasing very much at this moment, there is little effect on him.

special about being at rest relative to the earth. For instance, if a mattress falls out of the back of a truck on the freeway, the reason it rapidly comes to rest with respect to the planet is simply because of friction forces exerted by the asphalt, which happens to be attached to the planet.

Galileo's insights are summarized as follows:

The Principle of Inertia
No force is required to maintain motion with constant velocity in a straight line, and absolute motion does not cause any observable physical effects.

There are many examples of situations that seem to disprove the principle of inertia, but these all result from forgetting that friction is a force. For instance, it seems that a force is needed to keep a sailboat in motion. If the wind stops, the sailboat stops too. But the wind's force is not the only force on the boat; there is also a frictional force from the water. If the sailboat is cruising and the wind suddenly disappears, the backward frictional force still exists, and since it is no longer being counteracted by the wind's forward force, the boat stops. To disprove the principle of inertia, we would have to find an example where a moving object slowed down even though no forces whatsoever were acting on it.

Self-Check

What is incorrect about the following supposed counterexamples to the principle of inertia?

(1) When astronauts blast off in a rocket, their huge velocity does cause a physical effect on their bodies — they get pressed back into their seats, the flesh on their faces gets distorted, and they have a hard time lifting their arms.

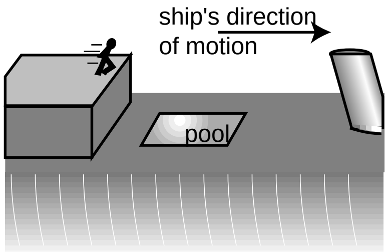
(2) When you're driving in a convertible with the top down, the wind in your face is an observable physical effect of your absolute motion.

Discussion questions

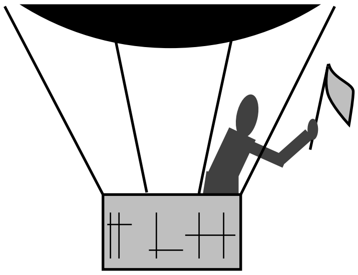
A. A passenger on a cruise ship finds, while the ship is docked, that he can leap off of the upper deck and just barely make it into the pool on the lower deck. If the ship leaves dock and is cruising rapidly, will this adrenaline junkie still be able to make it?

B. You are a passenger in the open basket hanging under a helium balloon. The balloon is being carried along by the wind at a constant velocity. If you are holding a flag in your hand, will the flag wave? If so, which way? [Based on a question from PSSC Physics.]

C. Aristotle stated that all objects naturally wanted to come to rest, with the unspoken implication that "rest" would be interpreted relative to the surface of the earth. Suppose we could transport Aristotle to the moon, put him in a space suit, and kick him out the door of the spaceship and into the lunar landscape. What would he expect his fate to be in this situation? If intelligent creatures inhabited the moon, and one of them independently came up with the equivalent of Aristotelian physics, what would they think about objects coming to rest?

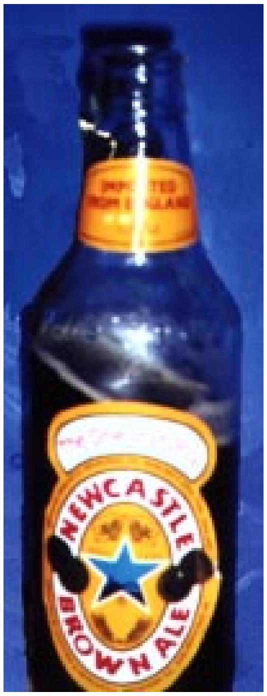


Discussion question A.



Discussion question B.

D. The bottle is sitting on a level table in a train's dining car, but the surface of the beer is tilted. What can you infer about the motion of the train?



Discussion question D.

(1) The effect only occurs during blastoff, when their velocity is changing. Once the rocket engines stop firing, their velocity stops changing, and they no longer feel any effect. (2) It is only an observable effect of your motion relative to the air.

2.5 Addition of Velocities

Addition of velocities to describe relative motion

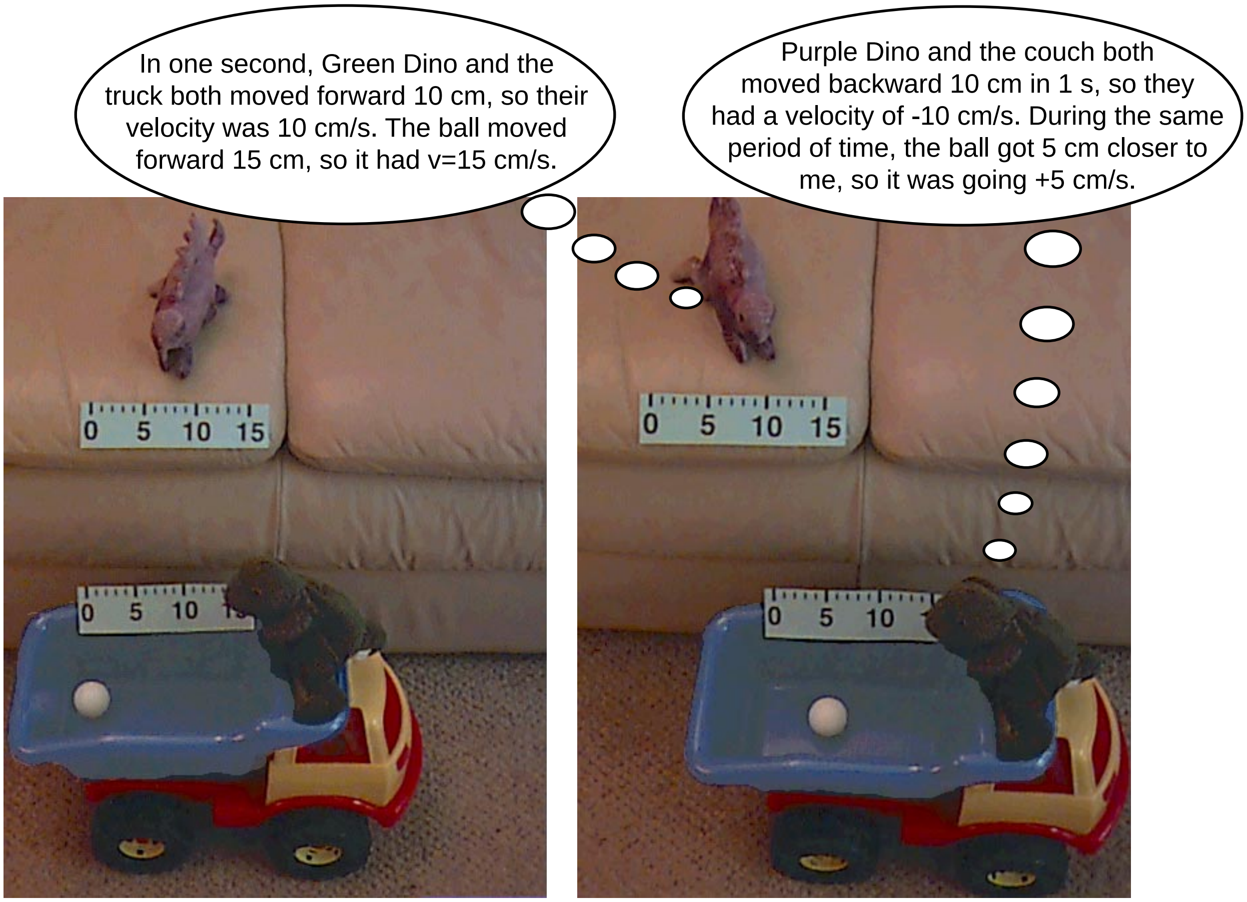
Since absolute motion cannot be unambiguously measured, the only way to describe motion unambiguously is to describe the motion of one object relative to another. Symbolically, we can write v_{PQ} for the velocity of object P relative to object Q.

Relative velocities add together.

Velocities measured with respect to different reference points can be compared by addition. In the figure below, the ball's velocity relative to the couch equals the ball's velocity relative to the truck plus the truck's velocity relative to the couch:

$$\begin{aligned} v_{BC} &= v_{BT} + v_{TC} \\ &= 5\text{ cm/s} + 10\text{ cm/s} \end{aligned}$$

The same equation can be used for any combination of three objects, just by substituting the relevant subscripts for B, T, and C. Just remember to write the equation so that the velocities being added have the same subscript twice in a row. In this example, if you read off the subscripts going from left to right, you get BC...=...BTTC. The fact that the two "inside" subscripts on the right are the same means that the equation has been set up correctly. Notice how subscripts on the left look just like the subscripts on the right, but with the two T's eliminated.


In one second, Green Dino and the truck both moved forward 10 cm, so their velocity was 10 cm/s. The ball moved forward 15 cm, so it had v=15 cm/s.
Purple Dino and the couch both moved backward 10 cm in 1 s, so they had a velocity of -10 cm/s. During the same period of time, the ball got 5 cm closer to me, so it was going +5 cm/s.
0 5 10 15
0 5 10 15
0 5 10 15
0 5 10

These two highly competent physicists disagree on absolute velocities, but they would agree on relative velocities. Purple Dino considers the couch to be at rest, while Green Dino thinks of the truck as being at rest. They agree, however, that the truck's velocity relative to the couch is v_{TC}=10 cm/s, the ball's velocity relative to the truck is v_{BT}=5 cm/s, and the ball's velocity relative to the couch is $v_{BC}=v_{BT}+v_{TC}$=15 cm/s.

Negative velocities in relative motion

My discussion of how to interpret positive and negative signs of velocity may have left you wondering why we should bother. Why not just make velocity positive by definition? The original reason why negative numbers were invented was that bookkeepers decided it would be convenient to use the negative number concept for payments to distinguish them from receipts. It was just plain easier than writing receipts in black and payments in red ink. After adding up your month's positive receipts and negative payments, you either got a positive number, indicating profit, or a negative number, showing a loss. You could then show the that total with a high-tech "+" or "-" sign, instead of looking around for the appropriate bottle of ink.

If you consistently label velocities as positive or negative depending on their directions, then adding velocities will also give signs that consistently relate to direction.

Nowadays we use positive and negative numbers for all kinds of things, but in every case the point is that it makes sense to add and subtract those things according to the rules you learned in grade school, such as "minus a minus makes a plus, why this is true we need not discuss." Adding velocities has the significance of comparing relative motion, and with this interpretation negative and positive velocities can used within a consistent framework. For example, the truck's velocity relative to the couch equals the truck's velocity relative to the ball plus the ball's velocity relative to the couch:

$$\begin{aligned} v_{TC} &= v_{TB} + v_{BC} \\ &= -5\text{ cm/s} + 15\text{ cm/s} \\ &= 10\text{ cm/s} \end{aligned}$$

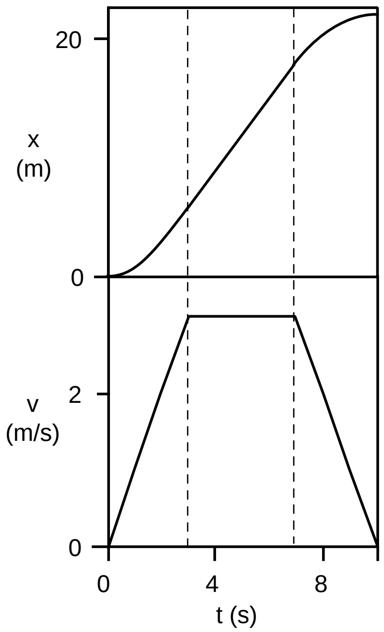
If we didn't have the technology of negative numbers, we would have had to remember a complicated set of rules for adding velocities: (1) if the two objects are both moving forward, you add, (2) if one is moving forward and one is moving backward, you subtract, but (3) if they're both moving backward, you add. What a pain that would have been.

Discussion questions

A. Interpret the general rule $v_{AB} = -v_{BA}$ in words.

B. If we have a specific situation where $v_{AB} + v_{BC} = 0$, what is going on?

2.6 Graphs of Velocity Versus Time



Since changes in velocity play such a prominent role in physics, we need a better way to look at changes in velocity than by laboriously drawing tangent lines on *x*-versus-*t* graphs. A good method is to draw a graph of velocity versus time. The examples on the left show the *x-t* and *v-t* graphs that might be produced by a car starting from a traffic light, speeding up, cruising for a while at constant speed, and finally slowing down for a stop sign. If you have an air freshener hanging from your rear-view mirror, then you will see an effect on the air freshener during the beginning and ending periods when the velocity is changing, but it will not be tilted during the period of constant velocity represented by the flat plateau in the middle of the *v-t* graph.

Students often mix up the things being represented on these two types of graphs. For instance, many students looking at the top graph say that the car is speeding up the whole time, since "the graph is becoming greater." What is getting greater throughout the graph is *x*, not *v*.

Similarly, many students would look at the bottom graph and think it showed the car backing up, because "it's going backwards at the end." But what is decreasing at the end is *v*, not *x*. Having both the *x-t* and *v-t* graphs in front of you like this is often convenient, because one graph may be easier to interpret than the other for a particular purpose. Stacking them like this means that corresponding points on the two graphs' time axes are lined up with each other vertically. However, one thing that is a little counterintuitive about the arrangement is that in a situation like this one involving a car, one is tempted to visualize the landscape stretching along the horizontal axis of one of the graphs. The horizontal axes, however, represent time, not position. The correct way to visualize the landscape is by mentally rotating the horizon 90 degrees counterclockwise and imagining it stretching along the upright axis of the *x-t* graph, which is the only axis that represents different positions in space.

2.7 $\int$ Applications of Calculus

The integral symbol, $\int$, in the heading for this section indicates that it is meant to be read by students in calculus-based physics. Students in an algebra-based physics course should skip these sections. The calculus-related sections in this book are meant to be usable by students who are taking calculus concurrently, so at this early point in the physics course I do not assume you know any calculus yet. This section is therefore not much more than a quick preview of calculus, to help you relate what you're learning in the two courses.

Newton was the first person to figure out the tangent-line definition of velocity for cases where the *x-t* graph is nonlinear. Before Newton, nobody had conceptualized the description of motion in terms of *x-t* and *v-t* graphs. In addition to the graphical techniques discussed in this chapter, Newton also invented a set of symbolic techniques called calculus. If you have an equation for *x* in terms of *t*, calculus allows you, for instance, to find an equation for *v* in terms of *t*. In calculus terms, we say that the function *v*(*t*) is the derivative of the function $x(t)$. In other words, the derivative of a function is a new function that tells how rapidly the original function was changing. We now use neither Newton's name for his technique (he called it "the method of fluxions") nor his notation. The more commonly used notation is due to Newton's German contemporary Leibnitz, whom the English accused of plagiarizing the calculus from Newton. In the Leibnitz notation, we write

$$v = \frac{dx}{dt}$$

to indicate that the function $v(t)$ equals the slope of the tangent line of the graph of $x(t)$ at every time t. The Leibnitz notation is meant to evoke the delta notation, but with a very small time interval. Because the dx and dt are thought of as very small Δx's and Δt's, i.e. very small differences, the part of calculus that has to do with derivatives is called differential calculus.

Differential calculus consists of three things:

- The concept and definition of the derivative, which is covered in this book, but which will be discussed more formally in your math course.
- The Leibnitz notation described above, which you'll need to get more comfortable with in your math course.
- A set of rules for that allows you to find an equation for the derivative of a given function. For instance, if you happened to have a situation where the position of an object was given by the equation $x=2t^7$, you would be able to use those rules to find $dx/dt=14t^6$. This bag of tricks is covered in your math course.

Summary

Selected Vocabulary

center of mass	the balance point of an object
velocity	the rate of change of position; the slope of the tangent line on an *x-t* graph.

Notation

x	a point in space
t	a point in time, a clock reading
Δ	"change in;" the value of a variable afterwards minus its value before
Δx	a distance, or more precisely a change in x, which may be less than the distance traveled; its plus or minus sign indicates direction
Δt	a duration of time
v	velocity
v_{AB}	the velocity of object A relative to object B

Standard Terminology Avoided in This Book

displacement	a name for the symbol Δx.
speed	the absolute value of the velocity, i.e. the velocity stripped of any information about its direction

Summary

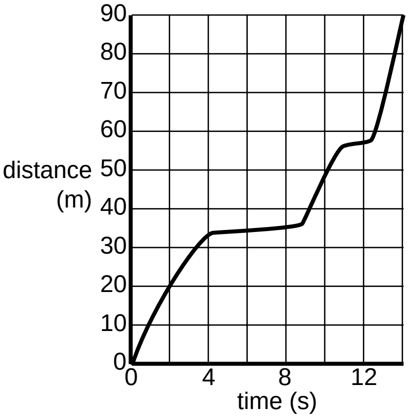
An object's center of mass is the point at which it can be balanced. For the time being, we are studying the mathematical description only of the motion of an object's center of mass in cases restricted to one dimension. The motion of an object's center of mass is usually far simpler than the motion of any of its other parts.

It is important to distinguish location, x, from distance, Δx, and clock reading, t, from time interval Δt. When an object's *x-t* graph is linear, we define its velocity as the slope of the line, Δx/Δt. When the graph is curved, we generalize the definition so that the velocity is the slope of the tangent line at a given point on the graph.

Galileo's principle of inertia states that no force is required to maintain motion with constant velocity in a straight line, and absolute motion does not cause any observable physical effects. Things typically tend to reduce their velocity relative to the surface of our planet only because they are physically rubbing against the planet (or something attached to the planet), not because there is anything special about being at rest with respect to the earth's surface. When it seems, for instance, that a force is required to keep a book sliding across a table, in fact the force is only serving to cancel the contrary force of friction.

Absolute motion is not a well-defined concept, and if two observers are not at rest relative to one another they will disagree about the absolute velocities of objects. They will, however, agree about relative velocities. If object A is in motion relative to object B, and B is in motion relative to C, then A's velocity relative to C is given by $v_{AC}=v_{AB}+v_{BC}$. Positive and negative signs are used to indicate the direction of an object's motion.

Homework Problems



Problem 1.

1 ✓. The graph shows the motion of a car stuck in stop-and-go freeway traffic. (a) If you only knew how far the car had gone during this entire time period, what would you think its velocity was? (b) What is the car's maximum velocity?

2. (a) Let θ be the latitude of a point on the Earth's surface. Derive an algebra equation for the distance, *L*, traveled by that point during one rotation of the Earth about its axis, i.e. over one day, expressed in terms of *L*, θ, and *R*, the radius of the earth. Check: Your equation should give *L*=0 for the North Pole.

(b✓) At what speed is Fullerton, at latitude θ=34°, moving with the rotation of the Earth about its axis? Give your answer in units of mi/h. [See the table in the back of the book for the relevant data.]

3★✓. A person is parachute jumping. During the time between when she leaps out of the plane and when she opens her chute, her altitude is given by the equation

$$y=(10000\ \text{m}) - (50\ \text{m/s})[t+(5.0\ \text{s})e^{-t/5.0\ \text{s}}] \quad .$$

Find her velocity at t=7.0 s. (This can be done on a calculator, without knowing calculus.) Because of air resistance, her velocity does not increase at a steady rate as it would for an object falling in vacuum.

4 S. A light-year is a unit of distance used in astronomy, and defined as the distance light travels in one year. The speed of light is $3.0\text{x}10^8$ m/s. Find how many meters there are in one light-year, expressing your answer in scientific notation.

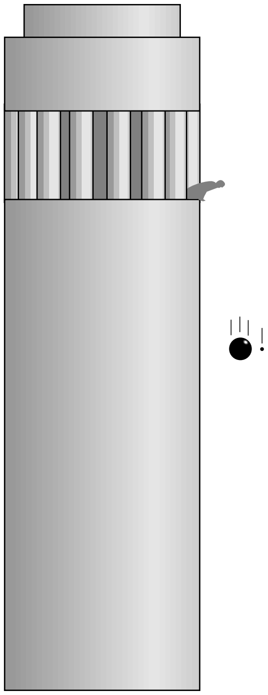
5 S. You're standing in a freight train, and have no way to see out. If you have to lean to stay on your feet, what, if anything, does that tell you about the train's velocity? Its acceleration? Explain.

6 ∫. A honeybee's position as a function of time is given by $x=10t-t^3$, where t is in seconds and x in meters. What is its velocity at t=3.0 s?

S A solution is given in the back of the book. ★ A difficult problem.
✓ A computerized answer check is available. ∫ A problem that requires calculus.

3 Acceleration and Free Fall

3.1 The Motion of Falling Objects



Galileo dropped a cannonball and a musketball simultaneously from a tower, and observed that they hit the ground at nearly the same time. This contradicted Aristotle's long-accepted idea that heavier objects fell faster.

The motion of falling objects is the simplest and most common example of motion with changing velocity. The early pioneers of physics had a correct intuition that the way things drop was a message directly from Nature herself about how the universe worked. Other examples seem less likely to have deep significance. A walking person who speeds up is making a conscious choice. If one stretch of a river flows more rapidly than another, it may be only because the channel is narrower there, which is just an accident of the local geography. But there is something impressively consistent, universal, and inexorable about the way things fall.

Stand up now and simultaneously drop a coin and a bit of paper side by side. The paper takes much longer to hit the ground. That's why Aristotle wrote that heavy objects fell more rapidly. Europeans believed him for two thousand years.

Now repeat the experiment, but make it into a race between the coin and your shoe. My own shoe is about 50 times heavier than the nickel I had handy, but it looks to me like they hit the ground at exactly the same moment. So much for Aristotle! Galileo, who had a flair for the theatrical, did the experiment by dropping a bullet and a heavy cannonball from a tall tower. Aristotle's observations had been incomplete, his interpretation a vast oversimplification.

It is inconceivable that Galileo was the first person to observe a discrepancy with Aristotle's predictions. Galileo was the one who changed the course of history because he was able to assemble the observations into a coherent pattern, and also because he carried out systematic quantitative (numerical) measurements rather than just describing things qualitatively.

Why is it that some objects, like the coin and the shoe, have similar motion, but others, like a feather or a bit of paper, are different? Galileo

Galileo and the Church

Galileo's contradiction of Aristotle had serious consequences. He was interrogated by the Church authorities and convicted of teaching that the earth went around the sun as a matter of fact and not, as he had promised previously, as a mere mathematical hypothesis. He was placed under permanent house arrest, and forbidden to write about or teach his theories. Immediately after being forced to recant his claim that the earth revolved around the sun, the old man is said to have muttered defiantly "and yet it does move."

The story is dramatic, but there are some omissions in the commonly taught heroic version. There was a rumor that the Simplicio character represented the Pope. Also, some of the ideas Galileo advocated had controversial religious overtones. He believed in the existence of atoms, and atomism was thought by some people to contradict the Church's doctrine of transubstantiation, which said that in the Catholic mass, the blessing of the bread and wine literally transformed them into the flesh and blood of Christ. His support for a cosmology in which the earth circled the sun was also disreputable because one of its supporters, Giordano Bruno, had also proposed a bizarre synthesis of Christianity with the ancient Egyptian religion.

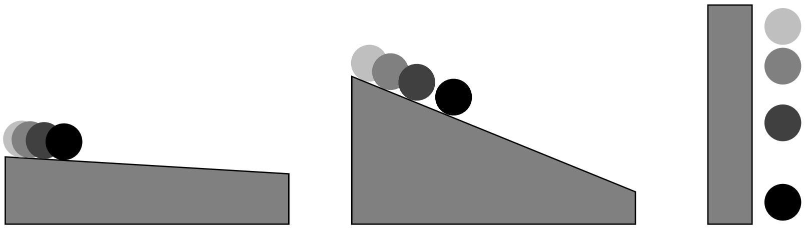
speculated that in addition to the force that always pulls objects down, there was an upward force exerted by the air. Anyone can speculate, but Galileo went beyond speculation and came up with two clever experiments to probe the issue. First, he experimented with objects falling in water, which probed the same issues but made the motion slow enough that he could take time measurements with a primitive pendulum clock. With this technique, he established the following facts:

- All heavy, streamlined objects (for example a steel rod dropped point-down) reach the bottom of the tank in about the same amount of time, only slightly longer than the time they would take to fall the same distance in air.
- Objects that are lighter or less streamlined take a longer time to reach the bottom.

This supported his hypothesis about two contrary forces. He imagined an idealized situation in which the falling object did not have to push its way through any substance at all. Falling in air would be more like this ideal case than falling in water, but even a thin, sparse medium like air would be sufficient to cause obvious effects on feathers and other light objects that were not streamlined. Today, we have vacuum pumps that allow us to suck nearly all the air out of a chamber, and if we drop a feather and a rock side by side in a vacuum, the feather does not lag behind the rock at all.

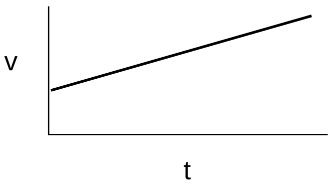
How the speed of a falling object increases with time

Galileo's second stroke of genius was to find a way to make quantitative measurements of how the speed of a falling object increased as it went along. Again it was problematic to make sufficiently accurate time measurements with primitive clocks, and again he found a tricky way to slow things down while preserving the essential physical phenomena: he let a ball roll down a slope instead of dropping it vertically. The steeper the incline, the more rapidly the ball would gain speed. Without a modern video camera, Galileo had invented a way to make a slow-motion version of falling.



Velocity increases more gradually on the gentle slope, but the motion is otherwise the same as the motion of a falling object.

Although Galileo's clocks were only good enough to do accurate experiments at the smaller angles, he was confident after making a systematic study at a variety of small angles that his basic conclusions were generally valid. Stated in modern language, what he found was that the velocity-versus-time graph was a line. In the language of algebra, we know that a line has an equation of the form $y=ax+b$, but our variables are v and t, so it would be $v=at+b$. (The constant b can be interpreted simply as the initial velocity of the object, i.e. its velocity at the time when we started our clock, which we conventionally write as v_o .)



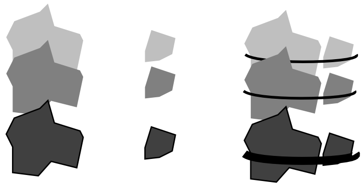
The v-t graph of a falling object is a line.

Self-Check

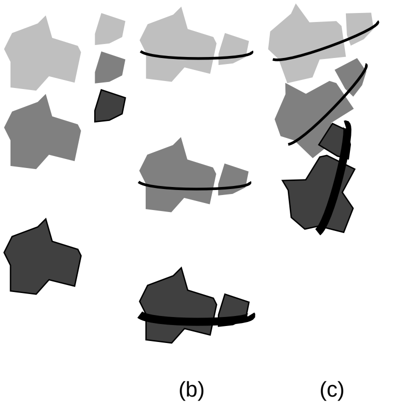
An object is rolling down an incline. After it has been rolling for a short time, it is found to travel 13 cm during a certain one-second interval. During the second after that, if goes 16 cm. How many cm will it travel in the second after that?

A contradiction in Aristotle's reasoning

Galileo's inclined-plane experiment disproved the long-accepted claim by Aristotle that a falling object had a definite "natural falling speed" proportional to its weight. Galileo had found that the speed just kept on increasing, and weight was irrelevant as long as air friction was negligible. Not only did Galileo prove experimentally that Aristotle had been wrong, but he also pointed out a logical contradiction in Aristotle's own reasoning. Simplicio, the stupid character, mouths the accepted Aristotelian wisdom:



(a) Galileo's experiments show that all falling objects have the same motion if air resistance is negligible.



Aristotle said that heavier objects fell faster than lighter ones. If two rocks are tied together, that makes an extra-heavy rock, (b), which should fall faster. But Aristotle's theory would also predict that the light rock would hold back the heavy rock, resulting in a slower fall, (c).

> Simplicio: There can be no doubt but that a particular body ... has a fixed velocity which is determined by nature...
>
> Salviati: If then we take two bodies whose natural speeds are different, it is clear that, [according to Aristotle], on uniting the two, the more rapid one will be partly held back by the slower, and the slower will be somewhat hastened by the swifter. Do you not agree with me in this opinion?
>
> Simplicio: You are unquestionably right.
>
> Salviati: But if this is true, and if a large stone moves with a speed of, say, eight [unspecified units] while a smaller moves with a speed of four, then when they are united, the system will move with a speed less than eight; but the two stones when tied together make a stone larger than that which before moved with a speed of eight. Hence the heavier body moves with less speed than the lighter; an effect which is contrary to your supposition. Thus you see how, from your assumption that the heavier body moves more rapidly than the lighter one, I infer that the heavier body moves more slowly.
>
> [tr. Crew and De Salvio]

What is gravity?

The physicist Richard Feynman liked to tell a story about how when he was a little kid, he asked his father, "Why do things fall?" As an adult, he praised his father for answering, "Nobody knows why things fall. It's a deep mystery, and the smartest people in the world don't know the basic reason for it." Contrast that with the average person's off-the-cuff answer, "Oh, it's because of gravity." Feynman liked his father's answer, because his father realized that simply giving a name to something didn't mean that you understood it. The radical thing about Galileo's and Newton's approach to science was that they concentrated first on describing mathematically what really did happen, rather than spending a lot of time on untestable speculation such as Aristotle's statement that "Things fall because they are trying to reach their natural place in contact with the earth." That doesn't mean that science can never answer the "why" questions. Over the next month or two as you delve deeper into physics, you will learn that there are more fundamental reasons why all falling objects have *v-t* graphs with the same slope, regardless of their mass. Nevertheless, the methods of science always impose limits on how deep our explanation can go.

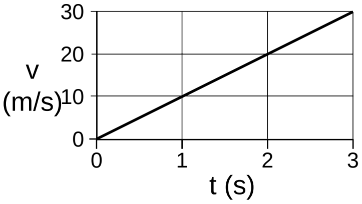
Its speed increases at a steady rate, so in the next second it will travel 19 cm.

3.2 Acceleration

Definition of acceleration for linear v-t graphs

Galileo's experiment with dropping heavy and light objects from a tower showed that all falling objects have the same motion, and his inclined-plane experiments showed that the motion was described by $v=ax+v_o$. The initial velocity v_o depends on whether you drop the object from rest or throw it down, but even if you throw it down, you cannot change the slope, a, of the v-t graph.

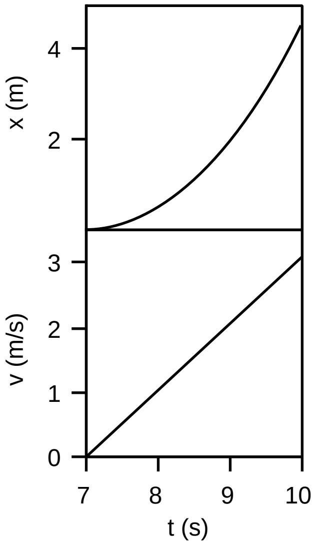
Since these experiments show that all falling objects have linear v-t graphs with the same slope, the slope of such a graph is apparently an important and useful quantity. We use the word *acceleration*, and the symbol a, for the slope of such a graph. In symbols, $a=\Delta v/\Delta t$. The acceleration can be interpreted as the amount of speed gained in every second, and it has units of velocity divided by time, i.e. "meters per second per second," or m/s/s. Continuing to treat units as if they were algebra symbols, we simplify "m/s/s" to read "m/s^2." Acceleration can be a useful quantity for describing other types of motion besides falling, and the word and the symbol "a" can be used in a more general context. We reserve the more specialized symbol "g" for the acceleration of falling objects, which on the surface of our planet equals 9.8 m/s^2. Often when doing approximate calculations or merely illustrative numerical examples it is good enough to use g=10 m/s^2, which is off by only 2%.



Example

Question: A despondent physics student jumps off a bridge, and falls for three seconds before hitting the water. How fast is he going when he hits the water?

Solution: Approximating g as 10 m/s^2, he will gain 10 m/s of speed each second. After one second, his velocity is 10 m/s, after two seconds it is 20 m/s, and on impact, after falling for three seconds, he is moving at 30 m/s.



Example: extracting acceleration from a graph

Question: The x-t and v-t graphs show the motion of a car starting from a stop sign. What is the car's acceleration?

Solution: Acceleration is defined as the slope of the v-t graph. The graph rises by 3 m/s during a time interval of 3 s, so the acceleration is (3 m/s)/(3 s)=1 m/s^2.

Incorrect solution #1: The final velocity is 3 m/s, and acceleration is velocity divided by time, so the acceleration is (3 m/s)/(10 s)=0.3 m/s^2.

✗ The solution is incorrect because you can't find the slope of a graph from one point. This person was just using the point at the right end of the v-t graph to try to find the slope of the curve.

Incorrect solution #2: Velocity is distance divided by time so v=(4.5 m)/(3 s)=1.5 m/s. Acceleration is velocity divided by time, so a=(1.5 m/s)/(3 s)=0.5 m/s^2.

✗ The solution is incorrect because velocity is the slope of the tangent line. In a case like this where the velocity is changing, you can't just pick two points on the x-t graph and use them to find the velocity.

Example: converting g to different units

Question: What is g in units of cm/s^2?

Solution: The answer is going to be how many cm/s of speed a falling object gains in one second. If it gains 9.8 m/s in one second, then it gains 980 cm/s in one second, so g=980 cm/s^2. Alternatively, we can use the method of fractions that equal one:

$$\frac{9.8\ \cancel{\text{m}}}{\text{s}^2} \times \frac{100\ \text{cm}}{1\ \cancel{\text{m}}} = \frac{980\ \text{cm}}{\text{s}^2}$$

Question: What is g in units of miles/hour^2?

Solution:

$$\frac{9.8\ \text{m}}{\text{s}^2} \times \frac{1\ \text{mile}}{1600\ \text{m}} \times \left(\frac{3600\ \text{s}}{1\ \text{hour}}\right)^2 = 7.9 \times 10^4\ \text{mile / hour}^2$$

This large number can be interpreted as the speed, in miles per hour, you would gain by falling for one hour. Note that we had to square the conversion factor of 3600 s/hour in order to cancel out the units of seconds squared in the denominator.

Question: What is g in units of miles/hour/s?

Solution:

$$\frac{9.8\ \text{m}}{\text{s}^2} \times \frac{1\ \text{mile}}{1600\ \text{m}} \times \frac{3600\ \text{s}}{1\ \text{hour}} = 22\ \text{mile/hour/s}$$

This is a figure that Americans will have an intuitive feel for. If your car has a forward acceleration equal to the acceleration of a falling object, then you will gain 22 miles per hour of speed every second. However, using mixed time units of hours and seconds like this is usually inconvenient for problem-solving. It would be like using units of foot-inches for area instead of ft^2 or in^2.

The acceleration of gravity is different in different locations.

Everyone knows that gravity is weaker on the moon, but actually it is not even the same everywhere on Earth, as shown by the sampling of numerical data in the following table.

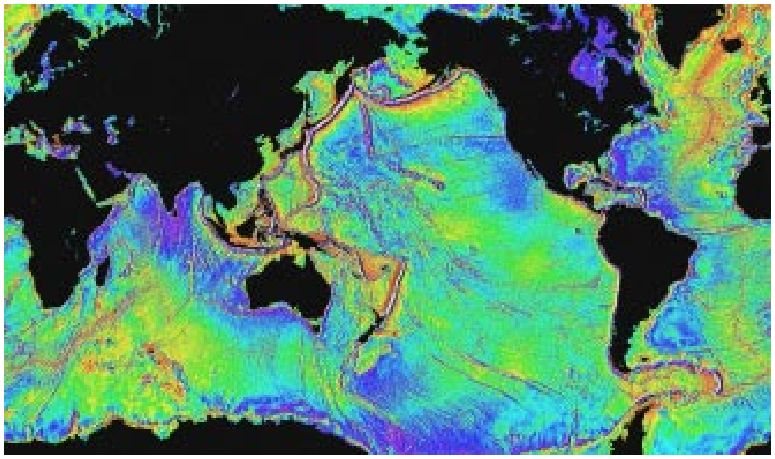
location	latitude	elevation (m)	g (m/s^2)
north pole	90° N	0	9.8322
Reykjavik, Iceland	64° N	0	9.8225
Fullerton, California	34° N	0	9.7957
Guayaquil, Ecuador	2° S	0	9.7806
Mt. Cotopaxi, Ecuador	1° S	5896	9.7624
Mt. Everest	28° N	8848	9.7643

The main variables that relate to the value of *g* on Earth are latitude and elevation. Although you have not yet learned how *g* would be calculated based on any deeper theory of gravity, it is not too hard to guess why *g* depends on elevation. Gravity is an attraction between things that have mass, and the attraction gets weaker with increasing distance. As you ascend from the seaport of Guayaquil to the nearby top of Mt. Cotopaxi, you are distancing yourself from the mass of the planet. The dependence on latitude occurs because we are measuring the acceleration of gravity relative to the earth's surface, but the earth's rotation causes the earth's surface to fall out from under you. (We will discuss both gravity and rotation in more detail later in the course.)

Much more spectacular differences in the strength of gravity can of course be observed away from the Earth's surface:

location	**g (m/s$_2$)**
asteroid Vesta (surface)	0.3
Earth's moon (surface)	1.6
Mars (surface)	3.7
Earth (surface)	9.8
Jupiter (cloud-tops)	26
Sun (visible surface)	270
typical neutron star (surface)	10^{12}
black hole (center)	infinite according to some theories, on the order of 10^{52} according to others

A typical neutron star is not so different in size from a large asteroid, but is orders of magnitude more massive, so the mass of a body definitely correlates with the *g* it creates. On the other hand, a neutron star has about the same mass as our Sun, so why is its *g* billions of times greater? If you had the misfortune of being on the surface of a neutron star, you'd be within a few thousand miles of all its mass, whereas on the surface of the Sun, you'd still be millions of miles from most if its mass.



This false-color map shows variations in the strength of the earth's gravity. Purple areas have the strongest gravity, yellow the weakest. The overall trend toward weaker gravity at the equator and stronger gravity at the poles has been artificially removed to allow the weaker local variations to show up. The map covers only the oceans because of the technique used to make it: satellites look for bulges and depressions in the surface of the ocean. A very slight bulge will occur over an undersea mountain, for instance, because the mountain's gravitational attraction pulls water toward it. The US government originally began collecting data like these for military use, to correct for the deviations in the paths of missiles. The data have recently been released for scientific and commercial use (e.g. searching for sites for off-shore oil wells).

Discussion questions

A. What is wrong with the following definitions of g?

(a) “g is gravity.”

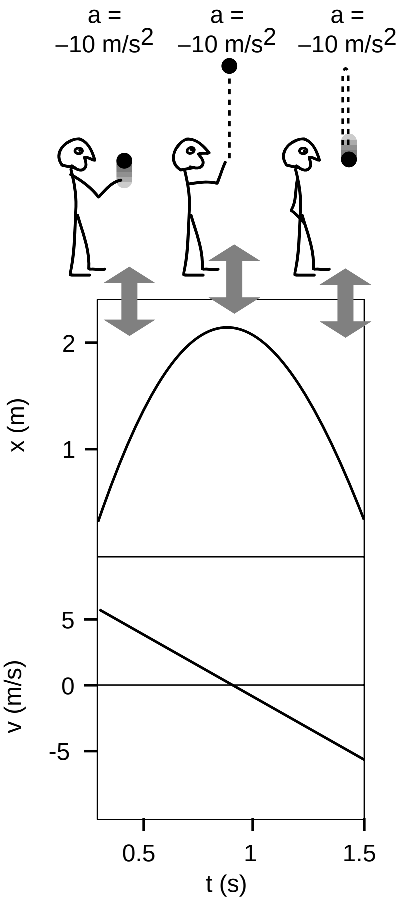
(b) “g is the speed of a falling object.”

(c) “g is how hard gravity pulls on things.”

B. When advertisers specify how much acceleration a car is capable of, they do not give an acceleration as defined in physics. Instead, they usually specify how many seconds are required for the car to go from rest to 60 miles/hour. Suppose we use the notation “a” for the acceleration as defined in physics, and “$a_{\text{car ad}}$” for the quantity used in advertisements for cars. In the US’s non-metric system of units, what would be the units of a and $a_{\text{car ad}}$? How would the use and interpretation of large and small, positive and negative values be different for a as opposed to $a_{\text{car ad}}$?

C. Two people stand on the edge of a cliff. As they lean over the edge, one person throws a rock down, while the other throws one straight up with an exactly opposite initial velocity. Compare the speeds of the rocks on impact at the bottom of the cliff.

3.3 Positive and Negative Acceleration



Gravity always pulls down, but that does not mean it always speeds things up. If you throw a ball straight up, gravity will first slow it down to v=0 and then begin increasing its speed. When I took physics in high school, I got the impression that positive signs of acceleration indicated speeding up, while negative accelerations represented slowing down, i.e. deceleration. Such a definition would be inconvenient, however, because we would then have to say that the same downward tug of gravity could produce either a positive or a negative acceleration. As we will see in the following example, such a definition also would not be the same as the slope of the v-t graph

Let’s study the example of the rising and falling ball. In the example of the person falling from a bridge, I assumed positive velocity values without calling attention to it, which meant I was assuming a coordinate system whose x axis pointed down. In this example, where the ball is reversing direction, it is not possible to avoid negative velocities by a tricky choice of axis, so let’s make the more natural choice of an axis pointing up. The ball’s velocity will initially be a positive number, because it is heading up, in the same direction as the x axis, but on the way back down, it will be a negative number. As shown in the figure, the v-t graph does not do anything special at the top of the ball’s flight, where v equals 0. Its slope is always negative. In the left half of the graph, the negative slope indicates a positive velocity that is getting closer to zero. On the right side, the negative slope is due to a negative velocity that is getting farther from zero, so we say that the ball is speeding up, but its velocity is decreasing!

To summarize, what makes the most sense is to stick with the original definition of acceleration as the slope of the v-t graph, $\Delta v/\Delta t$. By this definition, it just isn’t necessarily true that things speeding up have positive acceleration while things slowing down have negative acceleration. The word “deceleration” is not used much by physicists, and the word “acceleration” is used unblushingly to refer to slowing down as well as speeding up: “There was a red light, and we accelerated to a stop.”

Example

Question: In the example above, what happens if you calculate the acceleration between t=1.0 and 1.5 s?

Answer: Reading from the graph, it looks like the velocity is about -1 m/s at t=1.0 s, and around -6 m/s at t=1.5 s. The acceleration, figured between these two points, is

$$a = \frac{\Delta v}{\Delta t} = \frac{(-6\ \mathrm{m/s}) - (-1\ \mathrm{m/s})}{(1.5\ \mathrm{s}) - (1.0\ \mathrm{s})} = -10\ \mathrm{m/s}^2 \quad .$$

Even though the ball is speeding up, it has a negative acceleration.

Another way of convincing you that this way of handling the plus and minus signs makes sense is to think of a device that measures acceleration. After all, physics is supposed to use operational definitions, ones that relate to the results you get with actual measuring devices. Consider an air freshener hanging from the rear-view mirror of your car. When you speed up, the air freshener swings backward. Suppose we define this as a positive reading. When you slow down, the air freshener swings forward, so we'll call this a negative reading on our accelerometer. But what if you put the car in reverse and start speeding up backwards? Even though you're speeding up, the accelerometer responds in the same way as it did when you were going forward and slowing down. There are four possible cases:

motion of car	accelerometer swings	slope of v-t graph	direction of force acting on car
forward, speeding up	backward	+	forward
forward, slowing down	forward	-	backward
backward, speeding up	forward	-	backward
backward, slowing down	backward	+	forward

Note the consistency of the three right-hand columns — nature is trying to tell us that this is the right system of classification, not the left-hand column.

Because the positive and negative signs of acceleration depend on the choice of a coordinate system, the acceleration of an object under the influence of gravity can be either positive or negative. Rather than having to write things like "g=9.8 m/s^2 or -9.8 m/s^2" every time we want to discuss g's numerical value, we simply define g as the absolute value of the acceleration of objects moving under the influence of gravity. We consistently let g=9.8 m/s^2, but we may have either $a=g$ or $a=-g$, depending on our choice of a coordinate system.

Example

Question: A person kicks a ball, which rolls up a sloping street, comes to a halt, and rolls back down again. The ball has constant acceleration. The ball is initially moving at a velocity of 4.0 m/s, and after 10.0 s it has returned to where it started. At the end, it has sped back up to the same speed it had initially, but in the opposite direction. What was its acceleration?

Solution: By giving a positive number for the initial velocity, the statement of the question implies a coordinate axis that points up the slope of the hill. The "same" speed in the opposite direction should therefore be represented by a negative number, -4.0 m/s. The acceleration is $a=\Delta v/\Delta t=(v_{\text{after}}-\text{v}_{\text{before}})$/10.0 s=[(-4.0 m/s)-(4.0 m/s)]/10.0 s=-0.80 m/s^2. The acceleration was no different during the upward part of the roll than on the downward part of the roll.

Incorrect solution: Acceleration is $\Delta v/\Delta t$, and at the end it's not moving any faster or slower than when it started, so Δv=0 and a=0.

✗ The velocity does change, from a positive number to a negative number.

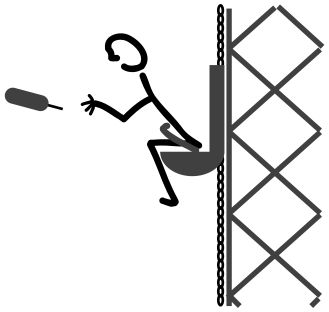
Discussion questions

A. A child repeatedly jumps up and down on a trampoline. Discuss the sign and magnitude of his acceleration, including both the time when he is in the air and the time when his feet are in contact with the trampoline.

B. Sally is on an amusement park ride which begins with her chair being hoisted straight up a tower at a constant speed of 60 miles/hour. Despite stern warnings from her father that he'll take her home the next time she misbehaves, she decides that as a scientific experiment she really needs to release her corndog over the side as she's on the way up. She does not throw it. She simply sticks it out of the car, lets it go, and watches it against the background of the sky, with no trees or buildings as reference points. What does the corndog's motion look like as observed by Sally? Does its speed ever appear to her to be zero? What acceleration does she observe it to have: is it ever positive? negative? zero? What would her enraged father answer if asked for a similar description of its motion as it appears to him, standing on the ground?

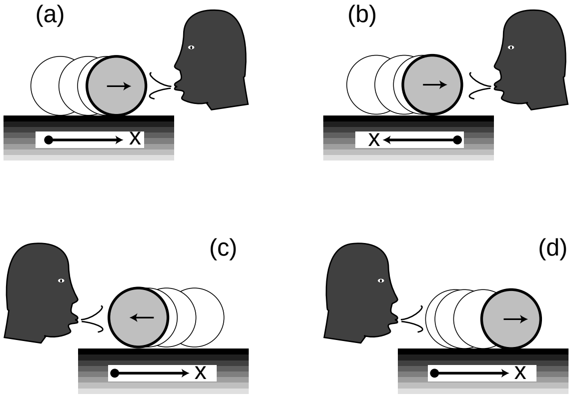
C. Can an object maintain a constant acceleration, but meanwhile reverse the direction of its velocity?

D. Can an object have a velocity that is positive and increasing at the same time that its acceleration is decreasing?



Discussion question B.

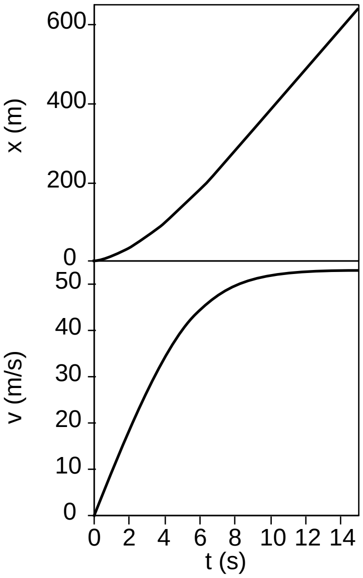
E. The four figures show a refugee from a Picasso painting blowing on a rolling water bottle. In some cases the person's blowing is speeding the bottle up, but in others it is slowing it down. The arrow inside the bottle shows which direction it is going, and a coordinate system is shown at the bottom of each figure. In each case, figure out the plus or minus signs of the velocity and acceleration. It may be helpful to draw a v-t graph in each case.



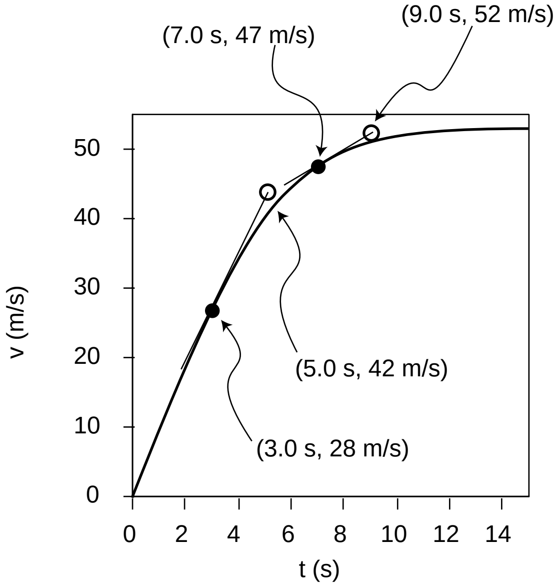
3.4 Varying Acceleration

So far we have only been discussing examples of motion for which the *v-t* graph is linear. If we wish to generalize our definition to v-t graphs that are more complex curves, the best way to proceed is similar to how we defined velocity for curved *x-t* graphs:

definition of acceleration
The acceleration of an object at any instant is the slope of the tangent line passing through its *v*-versus-*t* graph at the relevant point.

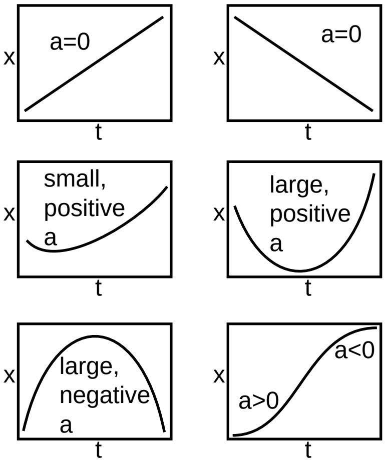


Example: a skydiver
Question: The graphs show the results of a fairly realistic computer simulation of the motion of a skydiver, including the effects of air friction. The x axis has been chosen pointing down, so x is increasing as she falls. Find (a) the skydiver's acceleration at t=3.0 s, and also (b) at t=7.0 s.

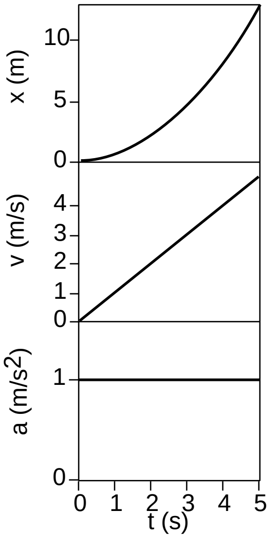


Solution: I've added tangent lines at the two points in question.
(a) To find the slope of the tangent line, I pick two points on the line (not necessarily on the actual curve): (3.0 s, 28 m/s) and (5.0 s, 42 m/s). The slope of the tangent line is (42 m/s-28 m/s)/(5.0 s - 3.0 s)=7.0 m/s^2.
(b) Two points on this tangent line are (7.0 s, 47 m/s) and (9.0 s, 52 m/s). The slope of the tangent line is (52 m/s-47 m/s)/(9.0 s - 7.0 s)=2.5 m/s^2.
Physically, what's happening is that at t=3.0 s, the skydiver is not yet going very fast, so air friction is not yet very strong. She therefore has an acceleration almost as great as g. At t=7.0 s, she is moving almost twice as fast (about 100 miles per hour), and air friction is extremely strong, resulting in a significant departure from the idealized case of no air friction.

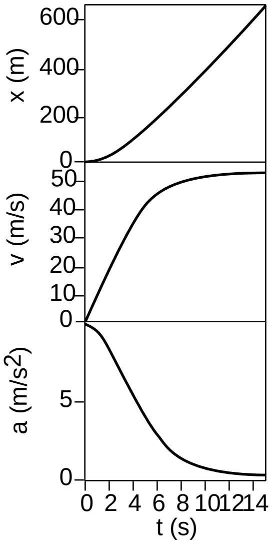
In the above example, the *x-t* graph was not even used in the solution of the problem, since the definition of acceleration refers to the slope of the *v-t* graph. It is possible, however, to interpret an *x-t* graph to find out something about the acceleration. An object with zero acceleration, i.e. constant velocity, has an *x-t* graph that is a straight line. A straight line has no curvature. A change in velocity requires a change in the slope of the x-t graph, which means that it is a curve rather than a line. Thus acceleration relates to the curvature of the *x-t* graph. Figure (c) shows some examples.



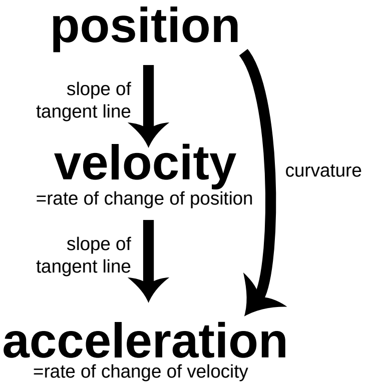
In the skydiver example, the *x-t* graph was more strongly curved at the beginning, and became nearly straight at the end. If the *x-t* graph is nearly straight, then its slope, the velocity, is nearly constant, and the acceleration is therefore small. We can thus interpret the acceleration as representing the curvature of the *x-t* graph. If the "cup" of the curve points up, the acceleration is positive, and if it points down, the acceleration is negative.



(a)



(b)



(c)

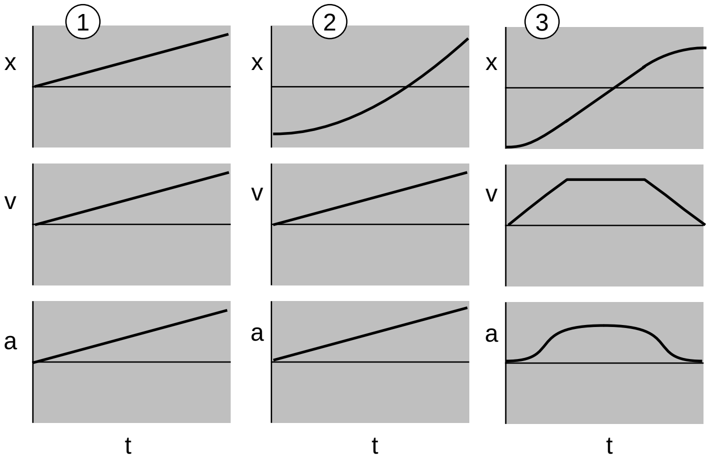
Since the relationship between *a* and *v* is analogous to the relationship between *v* and *x*, we can also make graphs of acceleration as a function of time, as shown in figures (a) and (b) above.

Figure (c) summarizes the relationships among the three types of graphs.

Discussion questions

A. Describe in words how the changes in the *a-t* graph for the skydiver relate to the behavior of the *v-t* graph.

B. Explain how each set of graphs contains inconsistencies.



Discussion question B.

3.5 The Area Under the Velocity-Time Graph

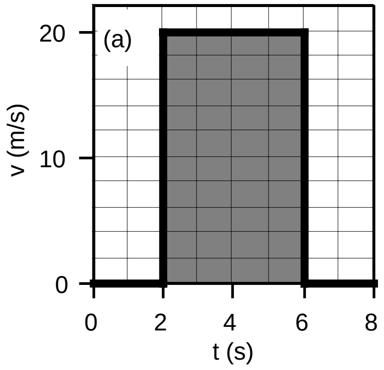
A natural question to ask about falling objects is how fast they fall, but Galileo showed that the question has no answer. The physical law that he discovered connects a cause (the attraction of the planet Earth's mass) to an effect, but the effect is predicted in terms of an acceleration rather than a velocity. In fact, no physical law predicts a definite velocity as a result of a specific phenomenon, because velocity cannot be measured in absolute terms, and only changes in velocity relate directly to physical phenomena.

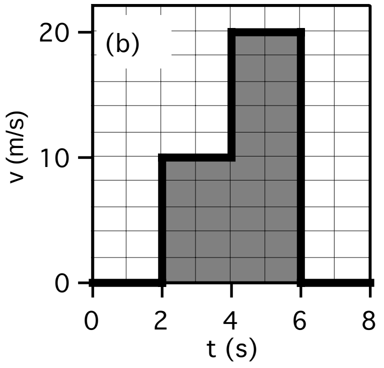
The unfortunate thing about this situation is that the definitions of velocity and acceleration are stated in terms of the tangent-line technique, which lets you go from *x* to *v* to *a*, but not the other way around. Without a technique to go backwards from *a* to *v* to *x*, we cannot say anything quantitative, for instance, about the *x-t* graph of a falling object. Such a technique does exist, and I used it to make the *x-t* graphs in all the examples above.

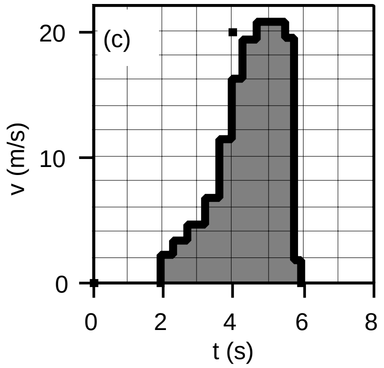
First let's concentrate on how to get *x* information out of a *v-t* graph. In example (a), an object moves at a speed of 20 m/s for a period of 4.0 s. The distance covered is $\Delta x = v\Delta t$=(20 m/s)x(4.0 s)=80 m. Notice that the quantities being multiplied are the width and the height of the shaded rectangle — or, strictly speaking, the time represented by its width and the velocity represented by its height. The distance of Δx=80 m thus corresponds to the area of the shaded part of the graph.

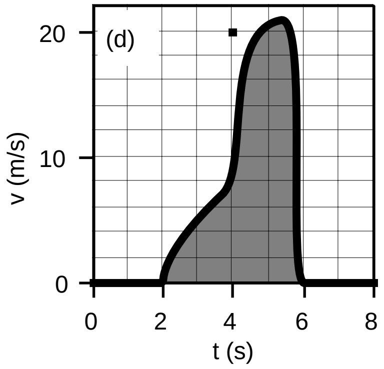
The next step in sophistication is an example like (b), where the object moves at a constant speed of 10 m/s for two seconds, then for two seconds at a different constant speed of 20 m/s. The shaded region can be split into a small rectangle on the left, with an area representing Δx=20 m, and a taller one on the right, corresponding to another 40 m of motion. The total distance is thus 60 m, which corresponds to the total area under the graph.

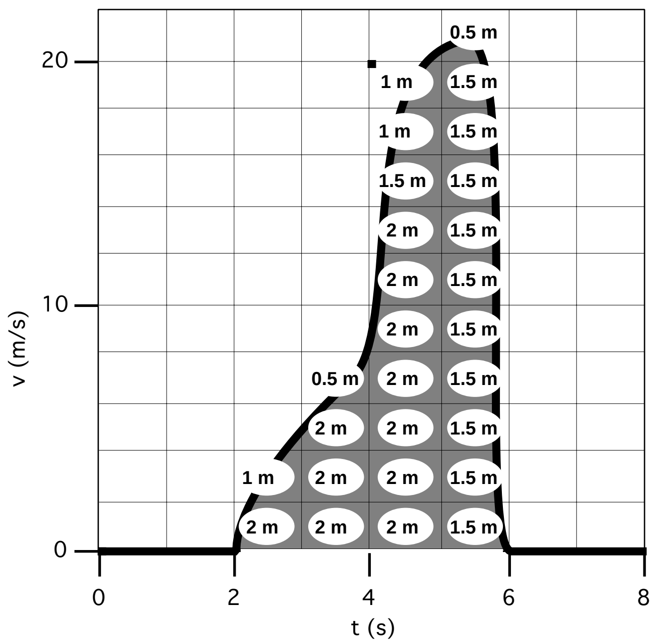
An example like (c) is now just a trivial generalization; there is simply a large number of skinny rectangular areas to add up. But notice that graph (c) is quite a good approximation to the smooth curve (d). Even though we have no formula for the area of a funny shape like (d), we can approximate its area by dividing it up into smaller areas like rectangles, whose area is easier to calculate. If someone hands you a graph like (d) and asks you to find the area under it, the simplest approach is just to count up the little rectangles on the underlying graph paper, making rough estimates of fractional rectangles as you go along.











That's what I've done above. Each rectangle on the graph paper is 1.0 s wide and 2 m/s tall, so it represents 2 m. Adding up all the numbers gives Δx=41 m. If you needed better accuracy, you could use graph paper with smaller rectangles.

It's important to realize that this technique gives you Δx, not x. The *v-t* graph has no information about where the object was when it started.

The following are important points to keep in mind when applying this technique:

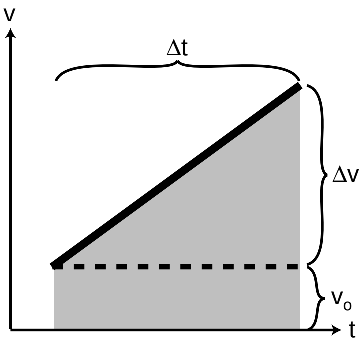
- If the range of v values on your graph does not extend down to zero, then you will get the wrong answer unless you compensate by adding in the area that is not shown.
- As in the example, one rectangle on the graph paper does not necessarily correspond to one meter of distance.
- Negative velocity values represent motion in the opposite direction, so area under the t axis should be subtracted, i.e. counted as "negative area."
- Since the result is a Δx value, it only tells you x_{after}-x_{before}, which may be less than the actual distance traveled. For instance, the object could come back to its original position at the end, which would correspond to Δx=0, even though it had actually moved a nonzero distance.

Finally, note that one can find Δv from an *a-t* graph using an entirely analogous method. Each rectangle on the *a-t* graph represents a certain amount of velocity change.

Discussion question

Roughly what would a pendulum's *v-t* graph look like? What would happen when you applied the area-under-the-curve technique to find the pendulum's Δx for a time period covering many swings?

3.6 Algebraic Results for Constant Acceleration



Although the area-under-the-curve technique can be applied to any graph, no matter how complicated, it may be laborious to carry out, and if fractions of rectangles must be estimated the result will only be approximate. In the special case of motion with constant acceleration, it is possible to find a convenient shortcut which produces exact results. When the acceleration is constant, the *v-t* graph is a straight line, as shown in the figure. The area under the curve can be divided into a triangle plus a rectangle, both of whose areas can be calculated exactly: $A=bh$ for a rectangle and $A=\frac{1}{2}bh$ for a triangle. The height of the rectangle is the initial velocity, v_o, and the height of the triangle is the change in velocity from beginning to end, Δv. The object's Δx is therefore given by the equation $\Delta x = v_o\Delta t + \frac{1}{2}\Delta v\Delta t$. This can be simplified a little by using the definition of acceleration, $a=\Delta v/\Delta t$ to eliminate Δv, giving

$$\Delta x = v_o\Delta t + \frac{1}{2}a\Delta t^2 \quad \text{[motion with constant acceleration]} \quad .$$

Since this is a second-order polynomial in Δt, the graph of Δx versus Δt is a parabola, and the same is true of a graph of *x* versus *t* — the two graphs differ only by shifting along the two axes. Although I have derived the equation using a figure that shows a positive v_o, positive *a*, and so on, it still turns out to be true regardless of what plus and minus signs are involved.

Another useful equation can be derived if one wants to relate the change in velocity to the distance traveled. This is useful, for instance, for finding the distance needed by a car to come to a stop. For simplicity, we start by deriving the equation for the special case of $v_o=0$, in which the final velocity v_f is a synonym for Δv. Since velocity and distance are the variables of interest, not time, we take the equation $\Delta x = \frac{1}{2} a\Delta t^2$ and use $\Delta t=\Delta v/a$ to eliminate Δt. This gives $\Delta x = \frac{1}{2} (\Delta v)^2/a$, which can be rewritten as

$$v_f^2 = 2a\Delta x \quad \text{[motion with constant acceleration, } v_o = 0\text{]} \quad .$$

For the more general case where $v_o \neq 0$, we skip the tedious algebra leading to the more general equation,

$$v_f^2 = v_o^2 + 2a\Delta x \quad \text{[motion with constant acceleration]} \quad .$$

To help get this all organized in your head, first let's categorize the variables as follows:

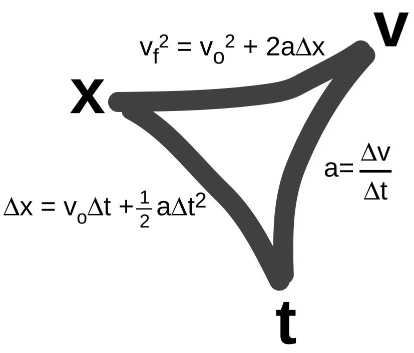
Variables that change during motion with constant acceleration:

x, *v*, *t*

Variable that doesn't change:

a

If you know one of the changing variables and want to find another, there is always an equation that relates those two:


$v_f^2 = v_o^2 + 2a\Delta x$
v
x
$a = \frac{\Delta v}{\Delta t}$
$\Delta x = v_o\Delta t + \frac{1}{2}a\Delta t^2$
t

The symmetry among the three variables is imperfect only because the equation relating *x* and *t* includes the initial velocity.

There are two main difficulties encountered by students in applying these equations:

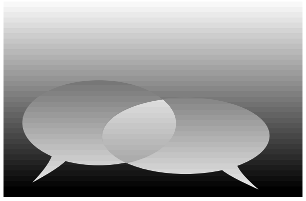
- The equations apply only to motion with constant acceleration. You can't apply them if the acceleration is changing.
- Students are often unsure of which equation to use, or may cause themselves unnecessary work by taking the longer path around the triangle in the chart above. Organize your thoughts by listing the variables you are given, the ones you want to find, and the ones you aren't given and don't care about.

Example
Question: You are trying to pull an old lady out of the way of an oncoming truck. You are able to give her an acceleration of 20 m/s^2. Starting from rest, how much time is required in order to move her 2 m?
Solution: First we organize our thoughts:

Variables given:	Δx, a, v_o
Variables desired:	Δt
Irrelevant variables:	v_f

Consulting the triangular chart above, the equation we need is clearly $\Delta x = v_o\Delta t + \frac{1}{2}a\Delta t^2$, since it has the four variables of interest and omits the irrelevant one. Eliminating the v_o term and solving for Δt gives $\Delta t = \sqrt{\frac{2\,\Delta x}{a}}$ =0.4 s.

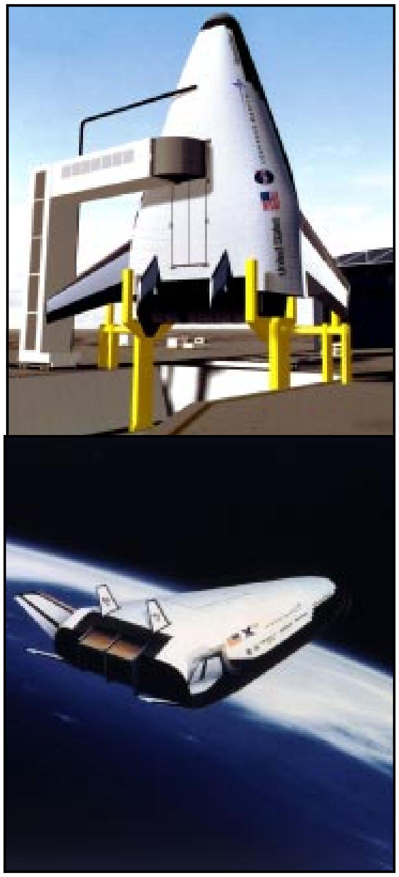
Discussion questions



A Check that the units make sense in the three equations derived in this section.

B. In chapter 1, I gave examples of correct and incorrect reasoning about proportionality, using questions about the scaling of area and volume. Try to translate the incorrect modes of reasoning shown there into mistakes about the following question: If the acceleration of gravity on Mars is 1/3 that on Earth, how many times longer does it take for a rock to drop the same distance on Mars?

3.7* Biological Effects of Weightlessness



Artist's conceptions of the X-33 spaceship, a half-scale uncrewed version of the planned VentureStar vehicle, which was supposed to cut the cost of sending people into space by an order of magnitude. The X-33 program was canceled in March 2001 due to technical failures and budget overruns, so the Space Shuttle will remain the U.S.'s only method of sending people into space for the forseeable future.
Courtesy of NASA.

The usefulness of outer space was brought home to North Americans in 1998 by the unexpected failure of the communications satellite that had been handling almost all of the continent's cellular phone traffic. Compared to the massive economic and scientific payoffs of satellites and space probes, human space travel has little to boast about after four decades. Sending people into orbit has just been too expensive to be an effective scientific or commercial activity. The 1986 Challenger disaster dealt a blow to NASA's confidence, and with the end of the cold war, U.S. prestige as a superpower was no longer a compelling reason to send Americans into space. All that may change soon, with a new generation of much cheaper reusable spaceships. (The space shuttle is not truly reusable. Retrieving the boosters out of the ocean is no cheaper than building new ones, but NASA brings them back and uses them over for public relations, to show how frugal they are.) Space tourism is even beginning to make economic sense! No fewer than three private companies are now willing to take your money for a reservation on a two-to-four minute trip into space, although none of them has a firm date on which to begin service. Within a decade, a space cruise may be the new status symbol among those sufficiently rich and brave.

Space sickness

Well, rich, brave, and possessed of an iron stomach. Travel agents will probably not emphasize the certainty of constant space-sickness. For us animals evolved to function in g=9.8 m/s^2, living in g=0 is extremely unpleasant. The early space program focused obsessively on keeping the astronaut-trainees in perfect physical shape, but it soon became clear that a body like a Greek demigod's was no defense against that horrible feeling that your stomach was falling out from under you and you were never going to catch up. Our inner ear, which normally tells us which way is down, tortures us when down is nowhere to be found. There is contradictory information about whether anyone ever gets over it; the "right stuff" culture creates a strong incentive for astronauts to deny that they are sick.

Effects of long space missions

Worse than nausea are the health-threatening effects of prolonged weightlessness. The Russians are the specialists in long-term missions, in which cosmonauts suffer harm to their blood, muscles, and, most importantly, their bones.

The effects on the muscles and skeleton appear to be similar to those experienced by old people and people confined to bed for a long time. Everyone knows that our muscles get stronger or weaker depending on the amount of exercise we get, but the bones are likewise adaptable. Normally old bone mass is continually being broken down and replaced with new material, but the balance between its loss and replacement is upset when people do not get enough weight-bearing exercise. The main effect is on the bones of the lower body. More research is required to find out whether astronauts' loss of bone mass is due to faster breaking down of bone, slower replacement, or both. It is also not known whether the effect can be suppressed via diet or drugs.

The other set of harmful physiological effects appears to derive from the redistribution of fluids. Normally, the veins and arteries of the legs are tightly constricted to keep gravity from making blood collect there. It is uncomfortable for adults to stand on their heads for very long, because the head's blood vessels are not able to constrict as effectively. Weightless astronauts' blood tends to be expelled by the constricted blood vessels of the lower body, and pools around their hearts, in their thoraxes, and in their heads. The only immediate result is an uncomfortable feeling of bloatedness in the upper body, but in the long term, a harmful chain of events is set in motion. The body's attempts to maintain the correct blood volume are most sensitive to the level of fluid in the head. Since astronauts have extra fluid in their heads, the body thinks that the over-all blood volume has become too great. It responds by decreasing blood volume below normal levels. This increases the concentration of red blood cells, so the body then decides that the blood has become too thick, and reduces the number of blood cells. In missions lasting up to a year or so, this is not as harmful as the musculoskeletal effects, but it is not known whether longer period in space would bring the red blood cell count down to harmful levels.

U.S. and Russian astronauts aboard the International Space Station, October 2000.

The International Space Station, September 2000. The space station will not rotate to provide simulated gravity. The completed station will be much bigger than it is in this picture.

Reproduction in space

For those enthralled by the romance of actual human colonization of space, human reproduction in weightlessness becomes an issue. An already-pregnant Russian cosmonaut did spend some time in orbit in the 1960's, and later gave birth to a normal child on the ground. Recently, one of NASA's public relations concerns about the space shuttle program has been to discourage speculation about space sex, for fear of a potential taxpayers' backlash against the space program as an expensive form of exotic pleasure.

Scientific work has been concentrated on studying plant and animal reproduction in space. Green plants, fungi, insects, fish, and amphibians have all gone through at least one generation in zero-gravity experiments without any serious problems. In many cases, animal embryos conceived in orbit begin by developing abnormally, but later in development they seem to correct themselves. However, chicken embryos fertilized on earth less than 24 hours before going into orbit have failed to survive. Since chickens are the organisms most similar to humans among the species investigated so far, it is not at all certain that humans could reproduce successfully in a zero-gravity space colony.

Simulated gravity

If humans are ever to live and work in space for more than a year or so, the only solution is probably to build spinning space stations to provide the illusion of weight, as discussed in section 9.2. Normal gravity could be simulated, but tourists would probably enjoy g=2 m/s^2 or 5 m/s^2. Space enthusiasts have proposed entire orbiting cities built on the rotating cylinder plan. Although science fiction has focused on human colonization of relatively earthlike bodies such as our moon, Mars, and Jupiter's icy moon Europa, there would probably be no practical way to build large spinning structures on their surfaces. If the biological effects of their 2-3 m/s^2 gravitational accelerations are as harmful as the effect of g=0, then we may be left with the surprising result that interplanetary space is more hospitable to our species than the moons and planets.

More on Apparent Weightlessness
Astronauts in orbit are not really weightless; they are only a few hundred miles up, so they are still affected strongly by the Earth's gravity. Section 10.3 of this book discusses why they experience apparent weightlessness.

More on Simulated Gravity
For more information on simulating gravity by spinning a spacecraft, see section 9.2 of this book.

3.8 ∫ Applications of Calculus

In the Applications of Calculus section at the end of the previous chapter, I discussed how the slope-of-the-tangent-line idea related to the calculus concept of a derivative, and the branch of calculus known as differential calculus. The other main branch of calculus, integral calculus, has to do with the area-under-the-curve concept discussed in section 3.5 of this chapter. Again there is a concept, a notation, and a bag of tricks for doing things symbolically rather than graphically. In calculus, the area under the *v-t* graph between $t=t_1$ and $t=t_2$ is notated like this:

$$\text{area under the curve} = \Delta x = \int_{t_1}^{t_2} v\, dt$$

The expression on the right is called an integral, and the s-shaped symbol, the integral sign, is read as "integral of...."

Integral calculus and differential calculus are closely related. For instance, if you take the derivative of the function $x(t)$, you get the function $v(t)$, and if you integrate the function $v(t)$, you get $x(t)$ back again. In other words, integration and differentiation are inverse operations. This is known as the fundamental theorem of calculus.

On an unrelated topic, there is a special notation for taking the derivative of a function twice. The acceleration, for instance, is the second (i.e. double) derivative of the position, because differentiating x once gives v, and then differentiating v gives a. This is written as

$$a = \frac{d^2x}{dt^2} \quad .$$

The seemingly inconsistent placement of the twos on the top and bottom confuses all beginning calculus students. The motivation for this funny notation is that acceleration has units of m/s^2, and the notation correctly suggests that: the top looks like it has units of meters, the bottom seconds^2. The notation is not meant, however, to suggest that t is really squared.

Summary

Selected Vocabulary

gravity .... A general term for the phenomenon of attraction between things having mass. The attraction between our planet and a human-sized object causes the object to fall.

acceleration .... The rate of change of velocity; the slope of the tangent line on a *v-t* graph.

Notation

a	acceleration
g	the acceleration of objects in free fall

Summary

Galileo showed that when air resistance is negligible all falling bodies have the same motion regardless of mass. Moreover, their *v-t* graphs are straight lines. We therefore define a quantity called acceleration as the slope, $\Delta v/\Delta t$, of an object's *v-t* graph. In cases other than free fall, the *v-t* graph may be curved, in which case the definition is generalized as the slope of a tangent line on the *v-t* graph. The acceleration of objects in free fall varies slightly across the surface of the earth, and greatly on other planets.

Positive and negative signs of acceleration are defined according to whether the *v-t* graph slopes up or down. This definition has the advantage that a force in a given direction always produces the same sign of acceleration.

The area under the *v-t* graph gives Δx, and analogously the area under the *a-t* graph gives Δv.

For motion with constant acceleration, the following three equations hold:

$$\Delta x = v_o \Delta t + \tfrac{1}{2} a \Delta t^2$$

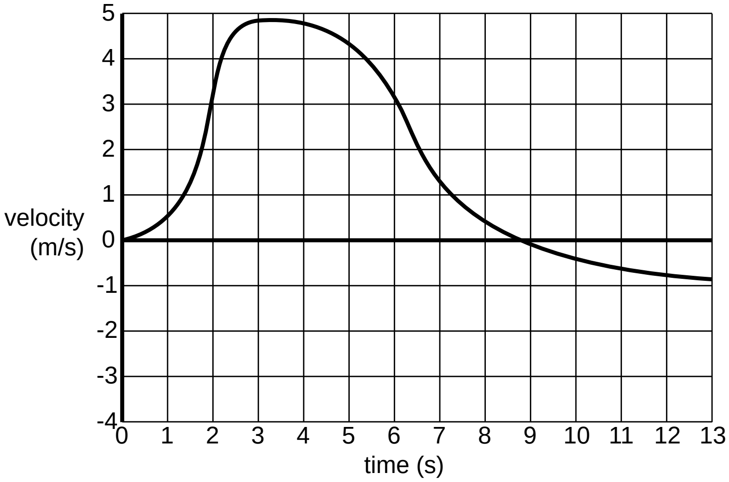
$$v_f^2 = v_o^2 + 2a\Delta x$$

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

They are not valid if the acceleration is changing.

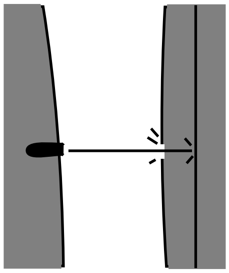
Homework Problems

1 ✓. The graph represents the velocity of a bee along a straight line. At *t*=0, the bee is at the hive. (a) When is the bee farthest from the hive? (b) How far is the bee at its farthest point from the hive? (c) At *t*=13 s, how far is the bee from the hive? [Hint: Try problem 19 first.]



2. A rock is dropped in a pond. Draw plots of its position versus time, velocity versus time, and acceleration versus time. Include its whole motion, from the moment it is dropped to the moment it comes to rest on the bottom of the pond.

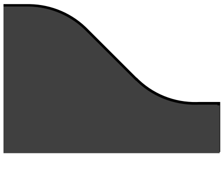
3. In an 18th-century naval battle, a cannon ball is shot horizontally, passes through the side of an enemy ship's hull, flies across the galley, and lodges in a bulkhead. Draw plots of its horizontal position, velocity, and acceleration as functions of time, starting while it is inside the cannon and has not yet been fired, and ending when it comes to rest.



Problem 3.

4. Draw graphs of position, velocity, and acceleration as functions of time for a person bunjee jumping. (In bunjee jumping, a person has a stretchy elastic cord tied to his/her ankles, and jumps off of a high platform. At the bottom of the fall, the cord brings the person up short. Presumably the person bounces up a little.)

5. A ball rolls down the ramp shown in the figure below, consisting of a circular knee, a straight slope, and a circular bottom. For each part of the ramp, tell whether the ball's velocity is increasing, decreasing, or constant, and also whether the ball's acceleration is increasing, decreasing, or constant. Explain your answers. Assume there is no air friction or rolling resistance. Hint: Try problem 20 first. [Based on a problem by Hewitt.]



6. At the end of its arc, the velocity of a pendulum is zero. Is its acceleration also zero at this point? Explain using a *v-t* graph.

7. What is the acceleration of a car that moves at a steady velocity of 100 km/h for 100 seconds? Explain your answer.

8. A physics homework question asks, "If you start from rest and accelerate at 1.54 m/s^2 for 3.29 s, how far do you travel by the end of that time?" A student answers as follows:

$$1.54 \times 3.29 = 5.07 \text{ m}$$

His Aunt Wanda is good with numbers, but has never taken physics. She doesn't know the formula for the distance traveled under constant acceleration over a given amount of time, but she tells her nephew his answer cannot be right. How does she know?

9 ✓. You are looking into a deep well. It is dark, and you cannot see the bottom. You want to find out how deep it is, so you drop a rock in, and you hear a splash 3 seconds later. Approximately how deep is the well?

10★✓. You take a trip in your spaceship to another star. Setting off, you increase your speed at a constant acceleration. Once you get half-way there, you start decelerating, at the same rate, so that by the time you get there, you have slowed down to zero speed. You see the tourist attractions, and then head home by the same method.

(a) Find a formula for the time, *T*, required for the round trip, in terms of *d*, the distance from our sun to the star, and *a*, the magnitude of the acceleration. Note that the acceleration is not constant over the whole trip, but the trip can be broken up into constant-acceleration parts.

(b) The nearest star to the Earth (other than our own sun) is Proxima Centauri, at a distance of d=4x10^{16} m. Suppose you use an acceleration of a=10 m/s^2, just enough to compensate for the lack of true gravity and make you feel comfortable. How long does the round trip take, in years?

(c) Using the same numbers for *d* and *a*, find your maximum speed. Compare this to the speed of light, which is 3.0x10^8 m/s. (Later in this course, you will learn that there are some new things going on in physics when one gets close to the speed of light, and that it is impossible to exceed the speed of light. For now, though, just use the simpler ideas you've learned so far.)

11. You climb half-way up a tree, and drop a rock. Then you climb to the top, and drop another rock. How many times greater is the velocity of the second rock on impact? Explain. (The answer is not two times greater.)

12. Two children stand atop a tall building. One drops a rock over the edge, while simultaneously the second throws a rock downward so that it has an initial speed of 10 m/s. Compare the accelerations of the two objects while in flight.

13 ∫. A person is parachute jumping. During the time between when she leaps out of the plane and when she opens her chute, her altitude is given by an equation of the form

$$y = b - c\left(t + ke^{-t/k}\right) \quad ,$$

where *e* is the base of natural logarithms, and *b*, *c*, and *k* are constants. Because of air resistance, her velocity does not increase at a steady rate as it would for an object falling in vacuum.

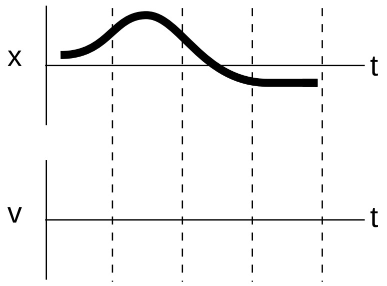
(a) What units would *b*, *c*, and *k* have to have for the equation to make sense?

(b) Find the person's velocity, *v*, as a function of time. [You will need to use the chain rule, and the fact that $d(e^x)/dx=e^x$.]

(c) Use your answer from part (b) to get an interpretation of the constant *c*. [Hint: e^{-x} approaches zero for large values of *x*.]

(d) Find the person's acceleration, *a*, as a function of time.

(e) Use your answer from part (b) to show that if she waits long enough to open her chute, her acceleration will become very small.



Problem 14.

14 S. The top part of the figure shows the position-versus-time graph for an object moving in one dimension. On the bottom part of the figure, sketch the corresponding v-versus-t graph.

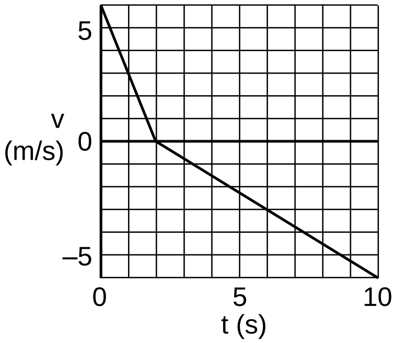
15 S. On New Year's Eve, a stupid person fires a pistol straight up. The bullet leaves the gun at a speed of 100 m/s. How long does it take before the bullet hits the ground?

16 S. If the acceleration of gravity on Mars is 1/3 that on Earth, how many times longer does it take for a rock to drop the same distance on Mars? Ignore air resistance.

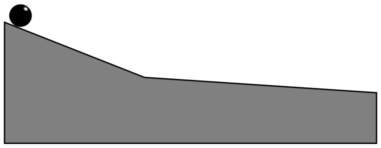
17 S∫. A honeybee's position as a function of time is given by $x=10t-t^3$, where *t* is in seconds and *x* in meters. What is its acceleration at t=3.0 s?

18 S. In July 1999, Popular Mechanics carried out tests to find which car sold by a major auto maker could cover a quarter mile (402 meters) in the shortest time, starting from rest. Because the distance is so short, this type of test is designed mainly to favor the car with the greatest acceleration, not the greatest maximum speed (which is irrelevant to the average person). The winner was the Dodge Viper, with a time of 12.08 s. The car's top (and presumably final) speed was 118.51 miles per hour (52.98 m/s). (a) If a car, starting from rest and moving with *constant* acceleration, covers a quarter mile in this time interval, what is its acceleration? (b) What would be the final speed of a car that covered a quarter mile with the constant acceleration you found in part a? (c) Based on the discrepancy between your answer in part b and the actual final speed of the Viper, what do you conclude about how its acceleration changed over time?

S A solution is given in the back of the book. ★ A difficult problem.
✓ A computerized answer check is available. ∫ A problem that requires calculus.



Problem 19.



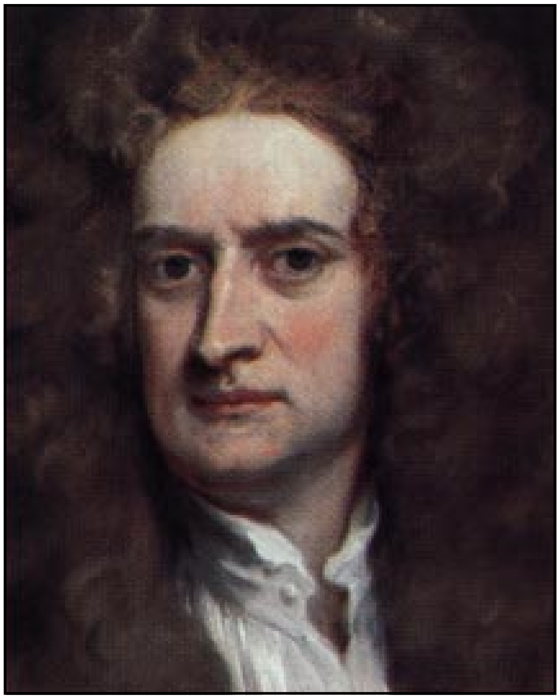
Problem 20.

19 S. The graph represents the motion of a rolling ball that bounces off of a wall. When does the ball return to the location it had at t=0?

20 S. (a) The ball is released at the top of the ramp shown in the figure. Friction is negligible. Use physical reasoning to draw v-t and a-t graphs. Assume that the ball doesn't bounce at the point where the ramp changes slope. (b) Do the same for the case where the ball is rolled up the slope from the right side, but doesn't quite have enough speed to make it over the top.

21 S. You drop a rubber ball, and it repeatedly bounces vertically. Draw graphs of position, velocity, and acceleration as functions of time.

22 S. Starting from rest, a ball rolls down a ramp, traveling a distance L and picking up a final speed v. How much of the distance did the ball have to cover before achieving a speed of v/2? [Based on a problem by Arnold Arons.]



Isaac Newton

Even as great and skeptical a genius as Galileo was unable to make much progress on the causes of motion. It was not until a generation later that Isaac Newton (1642-1727) was able to attack the problem successfully. In many ways, Newton's personality was the opposite of Galileo's. Where Galileo agressively publicized his ideas, Newton had to be coaxed by his friends into publishing a book on his physical discoveries. Where Galileo's writing had been popular and dramatic, Newton originated the stilted, impersonal style that most people think is standard for scientific writing. (Scientific journals today encourage a less ponderous style, and papers are often written in the first person.) Galileo's talent for arousing animosity among the rich and powerful was matched by Newton's skill at making himself a popular visitor at court. Galileo narrowly escaped being burned at the stake, while Newton had the good fortune of being on the winning side of the revolution that replaced King James II with William and Mary of Orange, leading to a lucrative post running the English royal mint.

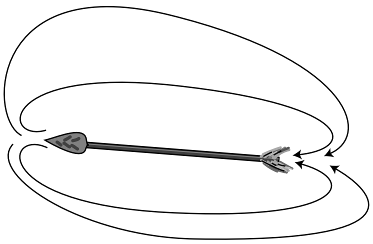
Newton discovered the relationship between force and motion, and revolutionized our view of the universe by showing that the same physical laws applied to all matter, whether living or nonliving, on or off of our planet's surface. His book on force and motion, the **Mathematical Principles of Natural Philosophy**, was uncontradicted by experiment for 200 years, but his other main work, **Optics**, was on the wrong track due to his conviction that light was composed of particles rather than waves. Newton was also an avid alchemist and an astrologer, an embarrassing fact that modern scientists would like to forget.

4 Force and Motion

If I have seen farther than others, it is because I have stood on the shoulders of giants.

Newton, referring to Galileo

4.1 Force



Aristotle said motion had to be caused by a force. To explain why an arrow kept flying after the bowstring was no longer pushing on it, he said the air rushed around behind the arrow and pushed it forward. We know this is wrong, because an arrow shot in a vacuum chamber does not instantly drop to the floor as it leaves the bow. Galileo and Newton realized that a force would only be needed to change the arrow's motion, not to make its motion continue.

We need only explain changes in motion, not motion itself

So far you've studied the measurement of motion in some detail, but not the reasons why a certain object would move in a certain way. This chapter deals with the "why" questions. Aristotle's ideas about the causes of motion were completely wrong, just like all his other ideas about physical science, but it will be instructive to start with them, because they amount to a road map of modern students' incorrect preconceptions.

Aristotle thought he needed to explain both why motion occurs and why motion might change. Newton inherited from Galileo the important counter-Aristotelian idea that motion needs no explanation, that it is only *changes* in motion that require a physical cause.

Aristotle gave three reasons for motion:

- Natural motion, such as falling, came from the tendency of objects to go to their "natural" place, on the ground, and come to rest.
- Voluntary motion was the type of motion exhibited by animals, which moved because they chose to.
- Forced motion occurred when an object was acted on by some other object that made it move.

Motion changes due to an interaction between two objects

In the Aristotelian theory, natural motion and voluntary motion are one-sided phenomena: the object causes its own motion. Forced motion is supposed to be a two-sided phenomenon, because one object imposes its "commands" on another. Where Aristotle conceived of some of the phenomena of motion as one-sided and others as two-sided, Newton realized that a change in motion was always a two-sided relationship of a force acting between two physical objects.

The one-sided "natural motion" description of falling makes a crucial omission. The acceleration of a falling object is not caused by its own "natural" tendencies but by an attractive force between it and the planet Earth. Moon rocks brought back to our planet do not "want" to fly back up to the moon because the moon is their "natural" place. They fall to the floor when you drop them, just like our homegrown rocks. As we'll discuss in more detail later in this course, gravitational forces are simply an attraction that occurs between any two physical objects. Minute gravitational forces can even be measured between human-scale objects in the laboratory.

The idea of natural motion also explains incorrectly why things come to rest. A basketball rolling across a beach slows to a stop because it is interacting with the sand via a frictional force, not because of its own desire to be at rest. If it was on a frictionless surface, it would never slow down. Many of Aristotle's mistakes stemmed from his failure to recognize friction as a force.

"Our eyes receive blue light reflected from this painting because Monet wanted to represent water with the color blue." This is a valid statement at one level of explanation, but physics works at the physical level of explanation, in which blue light gets to your eyes because it is reflected by blue pigments in the paint.

The concept of voluntary motion is equally flawed. You may have been a little uneasy about it from the start, because it assumes a clear distinction between living and nonliving things. Today, however, we are used to having the human body likened to a complex machine. In the modern world-view, the border between the living and the inanimate is a fuzzy no-man's land inhabited by viruses, prions, and silicon chips. Furthermore, Aristotle's statement that you can take a step forward "because you choose to" inappropriately mixes two levels of explanation. At the physical level of explanation, the reason your body steps forward is because of a frictional force acting between your foot and the floor. If the floor was covered with a puddle of oil, no amount of "choosing to" would enable you to take a graceful stride forward.

Forces can all be measured on the same numerical scale

In the Aristotelian-scholastic tradition, the description of motion as natural, voluntary, or forced was only the broadest level of classification, like splitting animals into birds, reptiles, mammals, and amphibians. There might be thousands of types of motion, each of which would follow its own rules. Newton's realization that all changes in motion were caused by two-sided interactions made it seem that the phenomena might have more in common than had been apparent. In the Newtonian description, there is only one cause for a change in motion, which we call force. Forces may be of different types, but they all produce changes in motion according to the same rules. Any acceleration that can be produced by a magnetic force can equally well be produced by an appropriately controlled stream of water. We can speak of two forces as being equal if they produce the same change in motion when applied in the same situation, which means that they pushed or pulled equally hard in the same direction.

The idea of a numerical scale of force and the newton unit were introduced in chapter 0. To recapitulate briefly, a force is when a pair of objects push or pull on each other, and one newton is the force required to accelerate a 1-kg object from rest to a speed of 1 m/s in 1 second.

More than one force on an object

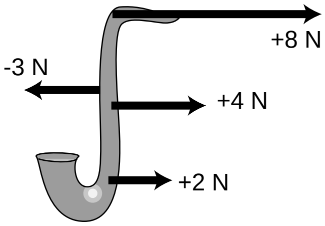
As if we hadn't kicked poor Aristotle around sufficiently, his theory has another important flaw, which is important to discuss because it corresponds to an extremely common student misconception. Aristotle conceived of forced motion as a relationship in which one object was the boss and the other "followed orders." It therefore would only make sense for an object to experience one force at a time, because an object couldn't follow orders from two sources at once. In the Newtonian theory, forces are numbers, not orders, and if more than one force acts on an object at once, the result is found by adding up all the forces. It is unfortunate that the use the English word "force" has become standard, because to many people it suggests that you are "forcing" an object to do something. The force of the earth's gravity cannot "force" a boat to sink, because there are other forces acting on the boat. Adding them up gives a total of zero, so the boat accelerates neither up nor down.

Objects can exert forces on each other at a distance

Aristotle declared that forces could only act between objects that were touching, probably because he wished to avoid the type of occult speculation that attributed physical phenomena to the influence of a distant and invisible pantheon of gods. He was wrong, however, as you can observe when a magnet leaps onto your refrigerator or when the planet earth exerts gravitational forces on objects that are in the air. Some types of forces, such as friction, only operate between objects in contact, and are called *contact forces*. Magnetism, on the other hand, is an example of a *noncontact force*. Although the magnetic force gets stronger when the magnet is closer to your refrigerator, touching is not required.

Weight

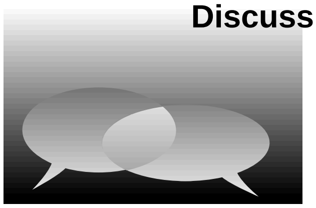
In physics, an object's weight , F_W, is defined as the earth's gravitational force on it. The SI unit of weight is therefore the Newton. People commonly refer to the kilogram as a unit of weight, but the kilogram is a unit of mass, not weight. Note that an object's weight is not a fixed property of that object. Objects weigh more in some places than in others, depending on the local strength of gravity. It is their mass that always stays the same. A baseball pitcher who can throw a 90-mile-per-hour fastball on earth would not be able to throw any faster on the moon, because the ball's inertia would still be the same.



In this example, positive signs have been used consistently for forces to the right, and negative signs for forces to the left. The numerical value of a force carries no information about the place on the saxophone where the force is applied.

Positive and negative signs of force

We'll start by considering only cases of one-dimensional center-of-mass motion in which all the forces are parallel to the direction of motion, i.e. either directly forward or backward. In one dimension, plus and minus signs can be used to indicate directions of forces, as shown in the figure. We can then refer generically to addition of forces, rather than having to speak sometimes of addition and sometimes of subtraction. We add the forces shown in the figure and get 11 N. In general, we should choose a one-dimensional coordinate system with its x axis parallel the direction of motion. Forces that point along the positive x axis are positive, and forces in the opposite direction are negative. Forces that are not directly along the x axis cannot be immediately incorporated into this scheme, but that's OK, because we're avoiding those cases for now.



Discussion questions

In chapter 0, I defined 1 N as the force that would accelerate a 1-kg mass from rest to 1 m/s in 1 s. Anticipating the following section, you might guess that 2 N could be defined as the force that would accelerate the same mass to twice the speed, or twice the mass to the same speed. Is there an easier way to define 2 N based on the definition of 1 N?

4.2 Newton's First Law

We are now prepared to make a more powerful restatement of the principle of inertia.

Newton's First Law
If the total force on an object is zero, its center of mass continues in the same state of motion.

In other words, an object initially at rest is predicted to remain at rest if the total force on it is zero, and an object in motion remains in motion with the same velocity in the same direction. The converse of Newton's first law is also true: if we observe an object moving with constant velocity along a straight line, then the total force on it must be zero.

In a future physics course or in another textbook, you may encounter the term *net force*, which is simply a synonym for *total force*.

What happens if the total force on an object is not zero? It accelerates. Numerical prediction of the resulting acceleration is the topic of Newton's second law, which we'll discuss in the following section.

This is the first of Newton's three laws of motion. It is not important to memorize which of Newton's three laws are numbers one, two, and three. If a future physics teacher asks you something like, "Which of Newton's laws are you thinking of," a perfectly acceptable answer is "The one about constant velocity when there's zero total force." The concepts are more important than any specific formulation of them. Newton wrote in Latin, and I am not aware of any modern textbook that uses a verbatim translation of his statement of the laws of motion. Clear writing was not in vogue in Newton's day, and he formulated his three laws in terms of a concept now called momentum, only later relating it to the concept of force. Nearly all modern texts, including this one, start with force and do momentum later.

Example: an elevator
Question: An elevator has a weight of 5000 N. Compare the forces that the cable must exert to raise it at constant velocity, lower it at constant velocity, and just keep it hanging.
Answer: In all three cases the cable must pull up with a force of exactly 5000 N. Most people think you'd need at least a little more than 5000 N to make it go up, and a little less than 5000 N to let it down, but that's incorrect. Extra force from the cable is

only necessary for speeding the car up when it starts going up or slowing it down when it finishes going down. Decreased force is needed to speed the car up when it gets going down and to slow it down when it finishes going up. But when the elevator is cruising at constant velocity, Newton's first law says that you just need to cancel the force of the earth's gravity.

To many students, the statement in the example that the cable's upward force "cancels" the earth's downward gravitational force implies that there has been a contest, and the cable's force has won, vanquishing the earth's gravitational force and making it disappear. That is incorrect. Both forces continue to exist, but because they add up numerically to zero, the elevator has no center-of-mass acceleration. We know that both forces continue to exist because they both have side-effects other than their effects on the car's center-of-mass motion. The force acting between the cable and the car continues to produce tension in the cable and keep the cable taut. The earth's gravitational force continues to keep the passengers (whom we are considering as part of the elevator-object) stuck to the floor and to produce internal stresses in the walls of the car, which must hold up the floor.

Example: terminal velocity for falling objects
Question: An object like a feather that is not dense or streamlined does not fall with constant acceleration, because air resistance is nonnegligible. In fact, its acceleration tapers off to nearly zero within a fraction of a second, and the feather finishes dropping at constant speed (known as its terminal velocity). Why does this happen?
Answer: Newton's first law tells us that the total force on the feather must have been reduced to nearly zero after a short time. There are two forces acting on the feather: a downward gravitational force from the planet earth, and an upward frictional force from the air. As the feather speeds up, the air friction becomes stronger and stronger, and eventually it cancels out the earth's gravitational force, so the feather just continues with constant velocity without speeding up any more.

The situation for a skydiver is exactly analogous. It's just that the skydiver experiences perhaps a million times more gravitational force than the feather, and it is not until she is falling very fast that the force of air friction becomes as strong as the gravitational force. It takes her several seconds to reach terminal velocity, which is on the order of a hundred miles per hour.

More general combinations of forces

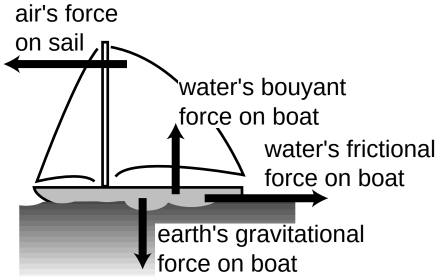
It is too constraining to restrict our attention to cases where all the forces lie along the line of the center of mass's motion. For one thing, we can't analyze any case of horizontal motion, since any object on earth will be subject to a vertical gravitational force! For instance, when you are driving your car down a straight road, there are both horizontal forces and vertical forces. However, the vertical forces have no effect on the center of mass motion, because the road's upward force simply counteracts the earth's downward gravitational force and keeps the car from sinking into the ground.

Later in the book we'll deal with the most general case of many forces acting on an object at any angles, using the mathematical technique of vector addition, but the following slight generalization of Newton's first law allows us to analyze a great many cases of interest:

> Suppose that an object has two sets of forces acting on it, one set along the line of the object's initial motion and another set perpendicular to the first set. If both sets of forces cancel, then the object's center of mass continues in the same state of motion.

Example: a car crash
Question: If you drive your car into a brick wall, what is the mysterious force that slams your face into the steering wheel?
Answer: Your surgeon has taken physics, so she is not going to believe your claim that a mysterious force is to blame. She knows that your face was just following Newton's first law. Immediately after your car hit the wall, the only forces acting on your head were the same canceling-out forces that had existed previously: the earth's downward gravitational force and the upward force from your neck. There were no forward or backward forces on your head, but the car did experience a backward force from the wall, so the car slowed down and your face caught up.

Example: a passenger riding the subway
Question: Describe the forces acting on a person standing in a subway train that is cruising at constant velocity.
Answer: No force is necessary to keep the person moving relative to the ground. He will not be swept to the back of the train if the floor is slippery. There are two vertical forces on him, the earth's downward gravitational force and the floor's upward force, which cancel. There are no horizontal forces on him at all, so of course the total horizontal force is zero.


air's force on sail
water's bouyant force on boat
water's frictional force on boat
earth's gravitational force on boat

Example: forces on a sailboat
Question: If a sailboat is cruising at constant velocity with the wind coming from directly behind it, what must be true about the forces acting on it?
Answer: The forces acting on the boat must be canceling each other out. The boat is not sinking or leaping into the air, so evidently the vertical forces are canceling out. The vertical forces are the downward gravitational force exerted by the planet earth and an upward force from the water.

The air is making a forward force on the sail, and if the boat is not accelerating horizontally then the water's backward frictional force must be canceling it out.

Contrary to Aristotle, more force is not needed in order to maintain a higher speed. Zero total force is always needed to maintain constant velocity. Consider the following made-up numbers:

	boat moving at a low, constant velocity	**boat moving at a high, constant velocity**
forward force of the wind on the sail......	10,000 N	20,000 N
backward force of the water on the hull.......................	-10,000 N	-20,000 N
total force on the boat.....................	0 N	0 N

The faster boat still has zero total force on it. The forward force on it is greater, and the backward force smaller (more negative), but that's irrelevant because Newton's first law has to do with the total force, not the individual forces.

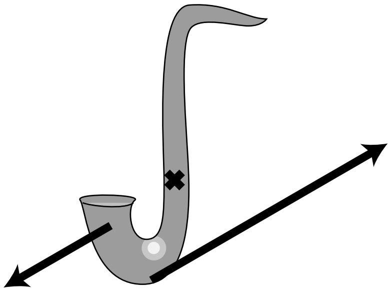
This example is quite analogous to the one about terminal velocity of falling objects, since there is a frictional force that increases with speed. After casting off from the dock and raising the sail, the boat will accelerate briefly, and then reach its terminal velocity, at which the water's frictional force has become as great as the wind's force on the sail.

Discussion questions

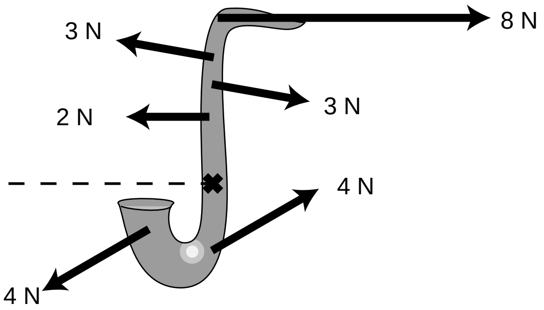
A. Newton said that objects continue moving if no forces are acting on them, but his predecessor Aristotle said that a force was necessary to keep an object moving. Why does Aristotle's theory seem more plausible, even though we now believe it to be wrong? What insight was Aristotle missing about the reason why things seem to slow down naturally?

B. In the first figure, what would have to be true about the saxophone's initial motion if the forces shown were to result in continued one-dimensional motion?

C. The second figure requires an ever further generalization of the preceding discussion. After studying the forces, what does your physical intuition tell you will happen? Can you state in words how to generalize the conditions for one-dimensional motion to include situations like this one?



Discussion question B.



Discussion question C.

4.3 Newton's Second Law

What about cases where the total force on an object is not zero, so that Newton's first law doesn't apply? The object will have an acceleration. The way we've defined positive and negative signs of force and acceleration guarantees that positive forces produce positive accelerations, and likewise for negative values. How much acceleration will it have? It will clearly depend on both the object's mass and on the amount of force.

Experiments with any particular object show that its acceleration is directly proportional to the total force applied to it. This may seem wrong, since we know of many cases where small amounts of force fail to move an object at all, and larger forces get it going. This apparent failure of proportionality actually results from forgetting that there is a frictional force in addition to the force we apply to move the object. The object's acceleration is exactly proportional to the total force on it, not to any individual force on it. In the absence of friction, even a very tiny force can slowly change the velocity of a very massive object.

Experiments also show that the acceleration is inversely proportional to the object's mass, and combining these two proportionalities gives the following way of predicting the acceleration of any object:

Newton's Second Law

$$a = F_{\text{total}}/m \quad ,$$

where

m is an object's mass
F_{total} is the sum of the forces acting on it, and
a is the acceleration of the object's center of mass.

We are presently restricted to the case where the forces of interest are parallel to the direction of motion.

Example: an accelerating bus
Question: A VW bus with a mass of 2000 kg accelerates from 0 to 25 m/s (freeway speed) in 34 s. Assuming the acceleration is constant, what is the total force on the bus?
Solution: We solve Newton's second law for $F_{\text{total}}=ma$, and substitute $\Delta v/\Delta t$ for a, giving

$$\begin{aligned} F_{\text{total}} &= m\Delta v/\Delta t \\ &= (2000\ \text{kg})(25\ \text{m/s} - 0\ \text{m/s})/(34\ \text{s}) \\ &= 1.5\ \text{kN} \quad . \end{aligned}$$

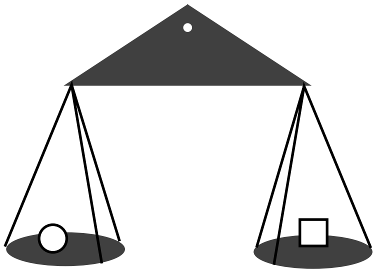
A generalization

As with the first law, the second law can be easily generalized to include a much larger class of interesting situations:

> Suppose an object is being acted on by two sets of forces, one set lying along the object's initial direction of motion and another set acting along a perpendicular line. If the forces perpendicular to the initial direction of motion cancel out, then the object accelerates along its original line of motion according to $a=F_{\text{total}}/m$.

The relationship between mass and weight

Mass is different from weight, but they're related. An apple's mass tells us how hard it is to change its motion. Its weight measures the strength of the gravitational attraction between the apple and the planet earth. The apple's weight is less on the moon, but its mass is the same. Astronauts assembling the International Space Station in zero gravity cannot just pitch massive modules back and forth with their bare hands; the modules are weightless, but not massless.



A simple double-pan balance works by comparing the weight forces exerted by the earth on the contents of the two pans. Since the two pans are at almost the same location on the earth's surface, the value of *g* is essentially the same for each one, and equality of weight therefore also implies equality of mass.

We have already seen the experimental evidence that when weight (the force of the earth's gravity) is the only force acting on an object, its acceleration equals the constant *g*, and *g* depends on where you are on the surface of the earth, but not on the mass of the object. Applying Newton's second law then allows us to calculate the magnitude of the gravitational force on any object in terms of its mass:

$$|F_W| = mg \quad .$$

(The equation only gives the magnitude, i.e. the absolute value, of F_W, because we're defining *g* as a positive number, so it equals the absolute value of a falling object's acceleration.)

Example: calculating terminal velocity

Question: Experiments show that the force of air friction on a falling object such as a skydiver or a feather can be approximated fairly well with the equation $|F_{air}|=c\rho Av^2$, where *c* is a constant, ρ is the density of the air, *A* is the cross-sectional area of the object as seen from below, and *v* is the object's velocity. Predict the object's terminal velocity, i.e. the final velocity it reaches after a long time.

Solution: As the object accelerates, its greater *v* causes the upward force of the air to increase until finally the gravitational force and the force of air friction cancel out, after which the object continues at constant velocity. We choose a coordinate system in which positive is up, so that the gravitational force is negative and the force of air friction is positive. We want to find the velocity at which

$$F_{air} + F_W = 0 \quad \text{, i.e.}$$

$$c\rho Av^2 - mg = 0 \quad .$$

Solving for *v* gives

$$v_{terminal} = \sqrt{\frac{mg}{c\rho A}}$$

Self-Check

It is important to get into the habit of interpreting equations. These two self-check questions may be difficult for you, but eventually you will get used to this kind of reasoning.

(a) Interpret the equation $v_{terminal}=\sqrt{mg/c\rho A}$ in the case of ρ=0.

(b) How would the terminal velocity of a 4-cm steel ball compare to that of a 1-cm ball?

(a) The case of ρ=0 represents an object falling in a vacuum, i.e. there is no density of air. The terminal velocity would be infinite. Physically, we know that an object falling in a vacuum would never stop speeding up, since there would be no force of air friction to cancel the force of gravity. (b) The 4-cm ball would have a mass that was greater by a factor of 4x4x4, but its cross-sectional area would be greater by a factor of 4x4. Its terminal velocity would be greater by a factor of $\sqrt{4^3/4^2}$ =2.

Discussion questions

A. Show that the Newton can be reexpressed in terms of the three basic mks units as the combination kg·m/s^2.

B. What is wrong with the following statements?

1. "g is the force of gravity."
2. "Mass is a measure of how much space something takes up."

C. Criticize the following incorrect statement:

> "If an object is at rest and the total force on it is zero, it stays at rest. There can also be cases where an object is moving and keeps on moving without having any total force on it, but that can only happen when there's no friction, like in outer space."

D. The table on the left gives laser timing data for Ben Johnson's 100 m dash at the 1987 World Championship in Rome. (His world record was later revoked because he tested positive for steroids.) How does the total force on him change over the duration of the race?

x (m)	t (s)
10	1.84
20	2.86
30	3.80
40	4.67
50	5.53
60	6.38
70	7.23
80	8.10
90	8.96
100	9.83

Discussion question D.

4.4 What Force Is Not

Violin teachers have to endure their beginning students' screeching. A frown appears on the woodwind teacher's face as she watches her student take a breath with an expansion of his ribcage but none in his belly. What makes physics teachers cringe is their students' verbal statements about forces. Below I have listed several dicta about what force is not.

Force is not a property of one object.

A great many of students' incorrect descriptions of forces could be cured by keeping in mind that a force is an interaction of two objects, not a property of one object.

> *Incorrect statement*: "That magnet has a lot of force."
> ✗ If the magnet is one millimeter away from a steel ball bearing, they may exert a very strong attraction on each other, but if they were a meter apart, the force would be virtually undetectable. The magnet's strength can be rated using certain electrical units (ampere-meters^2), but not in units of force.

Force is not a measure of an object's motion.

If force is not a property of a single object, then it cannot be used as a measure of the object's motion.

> *Incorrect statement*: "The freight train rumbled down the tracks with awesome force."
> ✗ Force is not a measure of motion. If the freight train collides with a stalled cement truck, then some awesome forces will occur, but if it hits a fly the force will be small.

Force is not energy.

There are two main approaches to understanding the motion of objects, one based on force and one on a different concept, called energy. The SI unit of energy is the Joule, but you are probably more familiar with the calorie, used for measuring food's energy, and the kilowatt-hour, the unit the electric company uses for billing you. Physics students' previous familiarity with calories and kilowatt-hours is matched by their universal unfamiliarity with measuring forces in units of Newtons, but the precise operational definitions of the energy concepts are more complex than those of the force concepts, and textbooks, including this one, almost universally place the force description of physics before the energy description. During the long period after the introduction of force and before the careful definition of energy, students are therefore vulnerable to situations in which, without realizing it, they are imputing the properties of energy to phenomena of force.

> *Incorrect statement:* “How can my chair be making an upward force on my rear end? It has no power!”
>
> ✗ Power is a concept related to energy, e.g. 100-watt lightbulb uses up 100 joules per second of energy. When you sit in a chair, no energy is used up, so forces can exist between you and the chair without any need for a source of power.

Force is not stored or used up.

Because energy can be stored and used up, people think force also can be stored or used up.

> *Incorrect statement:* “If you don’t fill up your tank with gas, you’ll run out of force.”
>
> ✗ Energy is what you’ll run out of, not force.

Forces need not be exerted by living things or machines.

Transforming energy from one form into another usually requires some kind of living or mechanical mechanism. The concept is not applicable to forces, which are an interaction between objects, not a thing to be transferred or transformed.

> *Incorrect statement:* “How can a wooden bench be making an upward force on my rear end? It doesn’t have any springs or anything inside it.”
>
> ✗ No springs or other internal mechanisms are required. If the bench didn’t make any force on you, you would obey Newton’s second law and fall through it. Evidently it does make a force on you!

A force is the direct cause of a change in motion.

I can click a remote control to make my garage door change from being at rest to being in motion. My finger’s force on the button, however, was not the force that acted on the door. When we speak of a force on an object in physics, we are talking about a force that acts directly. Similarly, when you pull a reluctant dog along by its leash, the leash and the dog are making forces on each other, not your hand and the dog. The dog is not even touching your hand.

Self-Check

Which of the following things can be correctly described in terms of force?

(a) A nuclear submarine is charging ahead at full steam.
(b) A nuclear submarine’s propellers spin in the water.
(c) A nuclear submarine needs to refuel its reactor periodically.

Discussion questions

A. Criticize the following incorrect statement: “If you shove a book across a table, friction takes away more and more of its force, until finally it stops.”

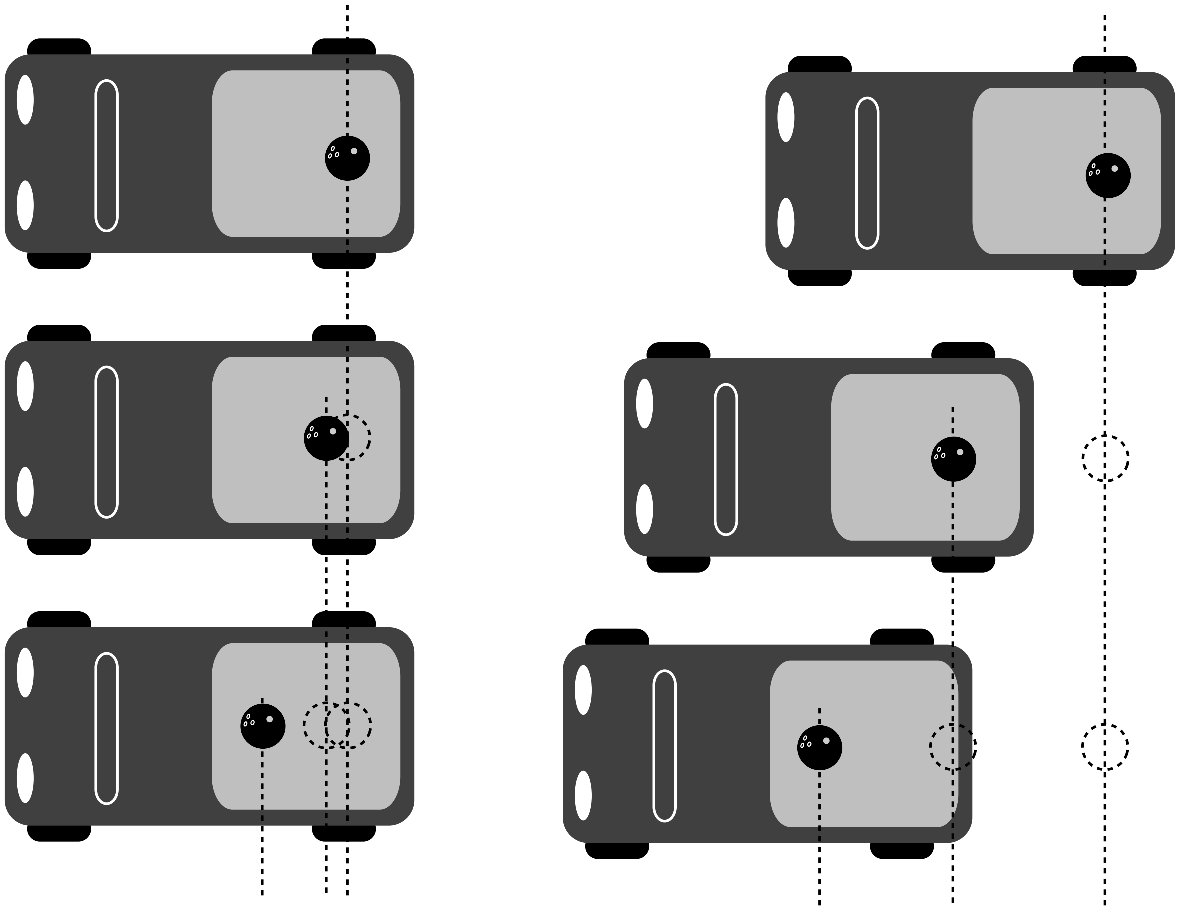
B. You hit a tennis ball against a wall. Explain any and all incorrect ideas in the following description of the physics involved: “The ball gets some force from you when you hit it, and when it hits the wall, it loses part of that force, so it doesn’t bounce back as fast. The muscles in your arm are the only things that a force can come from.”

(a) This is motion, not force. (b) This is a description of how the sub is able to get the water to produce a forward force on it. (c) The sub runs out of energy, not force.

4.5 Inertial and Noninertial Frames of Reference

One day, you're driving down the street in your pickup truck, on your way to deliver a bowling ball. The ball is in the back of the truck, enjoying its little jaunt and taking in the fresh air and sunshine. Then you have to slow down because a stop sign is coming up. As you brake, you glance in your rearview mirror, and see your trusty companion accelerating toward you. Did some mysterious force push it forward? No, it only seems that way because you and the car are slowing down. The ball is faithfully obeying Newton's first law, and as it continues at constant velocity it gets ahead relative to the slowing truck. No forces are acting on it (other than the same canceling-out vertical forces that were always acting on it). The ball only appeared to violate Newton's first law because there was something wrong with your frame of reference, which was based on the truck.

How, then, are we to tell in which frames of reference Newton's laws are valid? It's no good to say that we should avoid moving frames of reference, because there is no such thing as absolute rest or absolute motion. All frames can be considered as being either at rest or in motion. According to



(a) In a frame of reference that moves with the truck, the bowling ball appears to violate Newton's first law by accelerating despite having no horizontal forces on it.

(b) In an inertial frame of reference, which the surface of the earth approximately is, the bowling ball obeys Newton's first law. It moves equal distances in equal time intervals, i.e. maintains constant velocity. In this frame of reference, it is the truck that appears to have a change in velocity, which makes sense, since the road is making a horizontal force on it.

an observer in India, the strip mall that constituted the frame of reference in panel (b) of the figure was moving along with the earth's rotation at hundreds of miles per hour.

The reason why Newton's laws fail in the truck's frame of reference is not because the truck is *moving* but because it is *accelerating*. (Recall that physicists use the word to refer either to speeding up or slowing down.) Newton's laws were working just fine in the moving truck's frame of reference as long as the truck was moving at constant velocity. It was only when its speed changed that there was a problem. How, then, are we to tell which frames are accelerating and which are not? What if you claim that your truck is not accelerating, and the sidewalk, the asphalt, and the Burger King are accelerating? The way to settle such a dispute is to examine the motion of some object, such as the bowling ball, which we know has zero total force on it. Any frame of reference in which the ball appears to obey Newton's first law is then a valid frame of reference, and to an observer in that frame, Mr. Newton assures us that all the other objects in the universe will obey his laws of motion, not just the ball.

Valid frames of reference, in which Newton's laws are obeyed, are called *inertial frames of reference*. Frames of reference that are not inertial are called noninertial frames. In those frames, objects violate the principle of inertia and Newton's first law. While the truck was moving at constant velocity, both it and the sidewalk were valid inertial frames. The truck became an invalid frame of reference when it began changing its velocity.

You usually assume the ground under your feet is a perfectly inertial frame of reference, and we made that assumption above. It isn't perfectly inertial, however. Its motion through space is quite complicated, being composed of a part due to the earth's daily rotation around its own axis, the monthly wobble of the planet caused by the moon's gravity, and the rotation of the earth around the sun. Since the accelerations involved are numerically small, the earth is approximately a valid inertial frame.

Noninertial frames are avoided whenever possible, and we will seldom, if ever, have occasion to use them in this course. Sometimes, however, a noninertial frame can be convenient. Naval gunners, for instance, get all their data from radars, human eyeballs, and other detection systems that are moving along with the earth's surface. Since their guns have ranges of many miles, the small discrepancies between their shells' actual accelerations and the accelerations predicted by Newton's second law can have effects that accumulate and become significant. In order to kill the people they want to kill, they have to add small corrections onto the equation $a=F_{\text{total}}/m$. Doing their calculations in an inertial frame would allow them to use the usual form of Newton's second law, but they would have to convert all their data into a different frame of reference, which would require cumbersome calculations.

Discussion question

If an object has a linear x-t graph in a certain inertial frame, what is the effect on the graph if we change to a coordinate system with a different origin? What is the effect if we keep the same origin but reverse the positive direction of the x axis? How about an inertial frame moving alongside the object? What if we describe the object's motion in a noninertial frame?

Summary

Selected Vocabulary

weight .... the force of gravity on an object, equal to mg

inertial frame .... a frame of reference that is not accelerating, one in which Newton's first law is true

noninertial frame .... an accelerating frame of reference, in which Newton's first law is violated

Terminology Used in Some Other Books

net force .... another way of saying "total force"

Notation

F_W .... the weight force

Summary

Newton's first law of motion states that if all the forces on an object cancel each other out, then the object continues in the same state of motion. This is essentially a more refined version of Galileo's principle of inertia, which did not refer to a numerical scale of force.

Newton's second law of motion allows the prediction of an object's acceleration given its mass and the total force on it, $a=F_{\text{total}}/m$. This is only the one-dimensional version of the law; the full-three dimensional treatment will come in chapter 8, Vectors. Without the vector techniques, we can still say that the situation remains unchanged by including an additional set of vectors that cancel among themselves, even if they are not in the direction of motion.

Newton's laws of motion are only true in frames of reference that are not accelerating, known as inertial frames.

Homework Problems

1. An object is observed to be moving at constant speed along a line. Can you conclude that no forces are acting on it? Explain. [Based on a problem by Serway and Faughn.]

2. A car is normally capable of an acceleration of 3 m/s^2. If it is towing a trailer with half as much mass as the car itself, what acceleration can it achieve? [Based on a problem from PSSC Physics.]

3. (a) Let T be the maximum tension that the elevator's cable can withstand without breaking, i.e. the maximum force it can exert. If the motor is programmed to give the car an acceleration a, what is the maximum mass that the car can have, including passengers, if the cable is not to break? [Numerical check, not for credit: for T=1.0x10^4 N and a=3.0 m/s^2, your equation should give an answer of 780 kg.] (b) Interpret the equation you derived in the special cases of a=0 and of a downward acceleration of magnitude g.

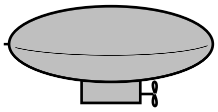
4 . A helicopter of mass m is taking off vertically. The only forces acting on it are the earth's gravitational force and the force, F_{air}, of the air pushing up on the propeller blades. (a) If the helicopter lifts off at t=0, what is its vertical speed at time t? (b✓) Plug numbers into your equation from part a, using m=2300 kg, F_{air}=27000 N, and t=4.0 s.

5★. In the 1964 Olympics in Tokyo, the best men's high jump was 2.18 m. Four years later in Mexico City, the gold medal in the same event was for a jump of 2.24 m. Because of Mexico City's altitude (2400 m), the acceleration of gravity there is lower than that in Tokyo by about 0.01 m/s^2. Suppose a high-jumper has a mass of 72 kg.

(a) Compare his mass and weight in the two locations.

(b✓) Assume that he is able to jump with the same initial vertical velocity in both locations, and that all other conditions are the same except for gravity. How much higher should he be able to jump in Mexico City?

(Actually, the reason for the big change between '64 and '68 was the introduction of the "Fosbury flop.")


Problem 6.

6 ∫. A blimp is initially at rest, hovering, when at t=0 the pilot turns on the motor of the propeller. The motor cannot instantly get the propeller going, but the propeller speeds up steadily. The steadily increasing force between the air and the propeller is given by the equation F=kt, where k is a constant. If the mass of the blimp is m, find its position as a function of time. (Assume that during the period of time you're dealing with, the blimp is not yet moving fast enough to cause a significant backward force due to air resistance.)

7 S. A car is accelerating forward along a straight road. If the force of the road on the car's wheels, pushing it forward, is a constant 3.0 kN, and the car's mass is 1000 kg, then how long will the car take to go from 20 m/s to 50 m/s?

S A solution is given in the back of the book. ★ A difficult problem.
✓ A computerized answer check is available. ∫ A problem that requires calculus.

8. Some garden shears are like a pair of scissors: one sharp blade slices past another. In the “anvil” type, however, a sharp blade presses against a flat one rather than going past it. A gardening book says that for people who are not very physically strong, the anvil type can make it easier to cut tough branches, because it concentrates the force on one side. Evaluate this claim based on Newton’s laws. [Hint: Consider the forces acting on the branch, and the motion of the branch.]

Rockets work by pushing exhaust gases out the back. Newton's third law says that if the rocket exerts a backward force on the gases, the gases must make an equal forward force on the rocket. Rocket engines can function above the atmosphere, unlike propellers and jets, which work by pushing against the surrounding air.

5 Analysis of Forces

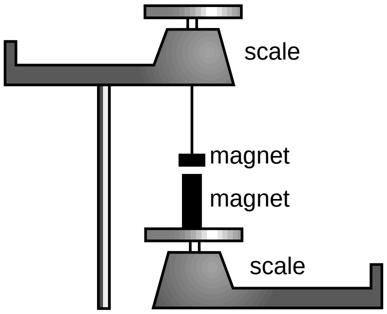
5.1 Newton's Third Law

Newton created the modern concept of force starting from his insight that all the effects that govern motion are interactions between two objects: unlike the Aristotelian theory, Newtonian physics has no phenomena in which an object changes its own motion.

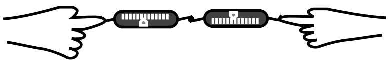
Is one object always the "order-giver" and the other the "order-follower"? As an example, consider a batter hitting a baseball. The bat definitely exerts a large force on the ball, because the ball accelerates drastically. But if you have ever hit a baseball, you also know that the ball makes a force on the bat — often with painful results if your technique is as bad as mine!

How does the ball's force on the bat compare with the bat's force on the ball? The bat's acceleration is not as spectacular as the ball's, but maybe we shouldn't expect it to be, since the bat's mass is much greater. In fact, careful measurements of both objects' masses and accelerations would show that $m_{ball}a_{ball}$ is very nearly equal to $-m_{bat}a_{bat}$, which suggests that the ball's force on the bat is of the same magnitude as the bat's force on the ball, but in the opposite direction.

The figures show two somewhat more practical laboratory experiments for investigating this issue accurately and without too much interference from extraneous forces.



(a) Two magnets exert forces on each other.



(b) Two people's hands exert forces on each other.

In the first experiment, a large magnet and a small magnet are weighed separately, and then one magnet is hung from the pan of the top balance so that it is directly above the other magnet. There is an attraction between the two magnets, causing the reading on the top scale to increase and the reading on the bottom scale to decrease. The large magnet is more "powerful" in the sense that it can pick up a heavier paperclip from the same distance, so many people have a strong expectation that one scale's reading will change by a far different amount than the other. Instead, we find that the two changes are equal in magnitude but opposite in direction, so the upward force of the top magnet on the bottom magnet is of the same magnitude as the downward force of the bottom magnet on the top magnet.

In the second experiment, two people pull on two spring scales. Regardless of who tries to pull harder, the two forces as measured on the spring scales are equal. Interposing the two spring scales is necessary in order to measure the forces, but the outcome is not some artificial result of the scales' interactions with each other. If one person slaps another hard on the hand, the slapper's hand hurts just as much as the slappee's, and it doesn't matter if the recipient of the slap tries to be inactive. (Punching someone in the mouth causes just as much force on the fist as on the lips. It's just that the lips are more delicate. The forces are equal, but not the levels of pain and injury.)

Newton, after observing a series of results such as these, decided that there must be a fundamental law of nature at work:

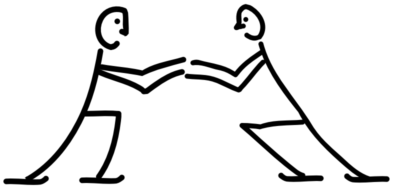
Newton's Third Law

Forces occur in equal and opposite pairs: whenever object A exerts a force on object B, object B must also be exerting a force on object A. The two forces are equal in magnitude and opposite in direction.

In one-dimensional situations, we can use plus and minus signs to indicate the directions of forces, and Newton's third law can be written succinctly as $F_{\text{A on B}} = -F_{\text{B on A}}$.

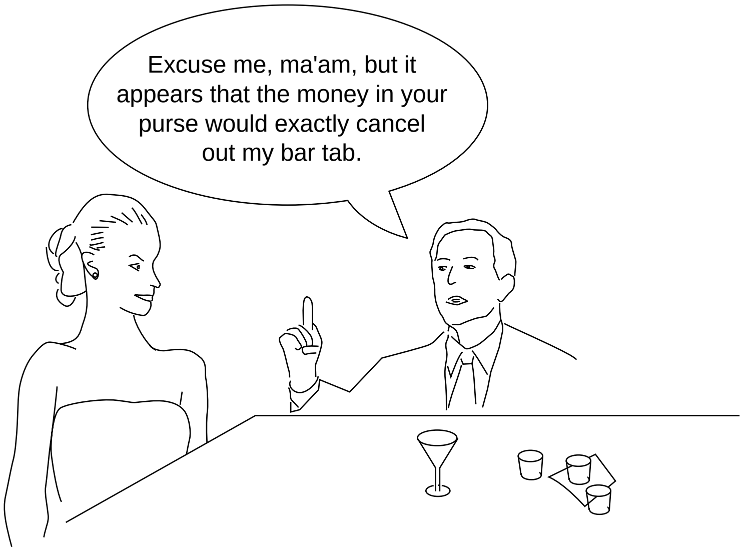
There is no cause and effect relationship between the two forces. There is no "original" force, and neither one is a response to the other. The pair of forces is a relationship, like marriage, not a back-and-forth process like a tennis match. Newton came up with the third law as a generalization about all the types of forces with which he was familiar, such as frictional and gravitational forces. When later physicists discovered a new type force, such as the force that holds atomic nuclei together, they had to check whether it obeyed Newton's third law. So far, no violation of the third law has ever been discovered, whereas the first and second laws were shown to have limitations by Einstein and the pioneers of atomic physics.

The English vocabulary for describing forces is unfortunately rooted in Aristotelianism, and often implies incorrectly that forces are one-way relationships. It is unfortunate that a half-truth such as “the table exerts an upward force on the book” is so easily expressed, while a more complete and correct description ends up sounding awkward or strange: “the table and the book interact via a force,” or “the table and book participate in a force.”



Newton’s third law does not mean that forces always cancel out so that nothing can ever move. If these two figure skaters, initially at rest, push against each other, they will both move.

To students, it often sounds as though Newton’s third law implies nothing could ever change its motion, since the two equal and opposite forces would always cancel. The two forces, however, are always on two different objects, so it doesn’t make sense to add them in the first place — we only add forces that are acting on the same object. If two objects are interacting via a force and no other forces are involved, then *both* objects will accelerate — in opposite directions!


Excuse me, ma'am, but it appears that the money in your purse would exactly cancel out my bar tab.

It doesn’t make sense for the man to talk about the woman’s money canceling out his bar tab, because there is no good reason to combine his debts and her assets. Similarly, it doesn’t make sense to refer to the equal and opposite forces of Newton’s third law as canceling. It only makes sense to add up forces that are acting on the *same* object, whereas two forces related to each other by Newton’s third law are always acting on two *different* objects.

A mnemonic for using Newton’s third law correctly

Mnemonics are tricks for memorizing things. For instance, the musical notes that lie between the lines on the treble clef spell the word FACE, which is easy to remember. Many people use the mnemonic “SOHCAHTOA” to remember the definitions of the sine, cosine, and tangent in trigonometry. I have my own modest offering, POFOSTITO, which I hope will make it into the mnemonics hall of fame. It’s a way to avoid some of the most common problems with applying Newton’s third law correctly:

Pair of
Opposite
Forces

Of the
Same
Type

Involving
Two
Objects

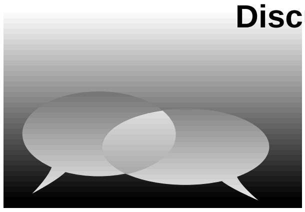
Example
Question: A book is lying on a table. What force is the Newton’s-third-law partner of the earth’s gravitational force on the book?
Answer: Newton’s third law works like “B on A, A on B,” so the partner must be the book’s gravitational force pulling upward on the planet earth. Yes, there is such a force! No, it does not cause the earth to do anything noticeable.
Incorrect answer: The table’s upward force on the book is the Newton’s-third-law partner of the earth’s gravitational force on the book.
✗ This answer violates two out of three of the commandments of POFOSTITO. The forces are not of the same type, because the table’s upward force on the book is not gravitational. Also, three objects are involved instead of two: the book, the table, and the planet earth.

Example
Question: A person is pushing a box up a hill. What force is related by Newton’s third law to the person’s force on the box?
Answer: The box’s force on the person.
Incorrect answer: The person’s force on the box is opposed by friction, and also by gravity.
✗ This answer fails all three parts of the POFOSTITO test, the most obvious of which is that three forces are referred to instead of a pair.

Optional Topic: Newton’s third law and action at a distance
Newton’s third law is completely symmetric in the sense that neither force constitutes a delayed response to the other. Newton’s third law does not even mention time, and the forces are supposed to agree at any given instant. This creates an interesting situation when it comes to noncontact forces. Suppose two people are holding magnets, and when one person waves or wiggles her magnet, the other person feels an effect on his. In this way they can send signals to each other from opposite sides of a wall, and if Newton’s third law is correct, it would seem that the signals are transmitted instantly, with no time lag. The signals are indeed transmitted quite quickly, but experiments with electronically controlled magnets show that the signals do not leap the gap instantly: they travel at the same speed as light, which is an extremely high speed but not an infinite one.

Is this a contradiction to Newton’s third law? Not really. According to current theories, there are no true noncontact forces. Action at a distance does not exist. Although it appears that the wiggling of one magnet affects the other with no need for anything to be in contact with anything, what really happens is that wiggling a magnet unleashes a shower of tiny particles called photons. The magnet shoves the photons out with a kick, and receives a kick in return, in strict obedience to Newton’s third law. The photons fly out in all directions, and the ones that hit the other magnet then interact with it, again obeying Newton’s third law.

Photons are nothing exotic, really. Light is made of photons, but our eyes receive such huge numbers of photons that we do not perceive them individually. The photons you would make by wiggling a magnet with your hand would be of a “color” that you cannot see, far off the red end of the rainbow.



Discussion questions

A. When you fire a gun, the exploding gases push outward in all directions, causing the bullet to accelerate down the barrel. What third-law pairs are involved? [Hint: Remember that the gases themselves are an object.]

B. Tam Anh grabs Sarah by the hand and tries to pull her. She tries to remain standing without moving. A student analyzes the situation as follows. “If Tam Anh’s force on Sarah is greater than her force on him, he can get her to move. Otherwise, she’ll be able to stay where she is.” What’s wrong with this analysis?

C. You hit a tennis ball against a wall. Explain any and all incorrect ideas in the following description of the physics involved: “According to Newton’s third law, there has to be a force opposite to your force on the ball. The opposite force is the ball’s mass, which resists acceleration, and also air resistance.”

5.2 Classification and Behavior of Forces

One of the most basic and important tasks of physics is to classify the forces of nature. I have already referred informally to "types" of forces such as friction, magnetism, gravitational forces, and so on. Classification systems are creations of the human mind, so there is always some degree of arbitrariness in them. For one thing, the level of detail that is appropriate for a classification system depends on what you're trying to find out. Some linguists, the "lumpers," like to emphasize the similarities among languages, and a few extremists have even tried to find signs of similarities between words in languages as different as English and Chinese, lumping the world's languages into only a few large groups. Other linguists, the "splitters," might be more interested in studying the differences in pronunciation between English speakers in New York and Connecticut. The splitters call the lumpers sloppy, but the lumpers say that science isn't worthwhile unless it can find broad, simple patterns within the seemingly complex universe.

Scientific classification systems are also usually compromises between practicality and naturalness. An example is the question of how to classify flowering plants. Most people think that biological classification is about discovering new species, naming them, and classifying them in the class-order-family-genus-species system according to guidelines set long ago. In reality, the whole system is in a constant state of flux and controversy. One very practical way of classifying flowering plants is according to whether their petals are separate or joined into a tube or cone — the criterion is so clear that it can be applied to a plant seen from across the street. But here practicality conflicts with naturalness. For instance, the begonia has separate petals and the pumpkin has joined petals, but they are so similar in so many other ways that they are usually placed within the same order. Some taxonomists have come up with classification criteria that they claim correspond more naturally to the apparent relationships among plants, without having to make special exceptions, but these may be far less practical, requiring for instance the examination of pollen grains under an electron microscope.

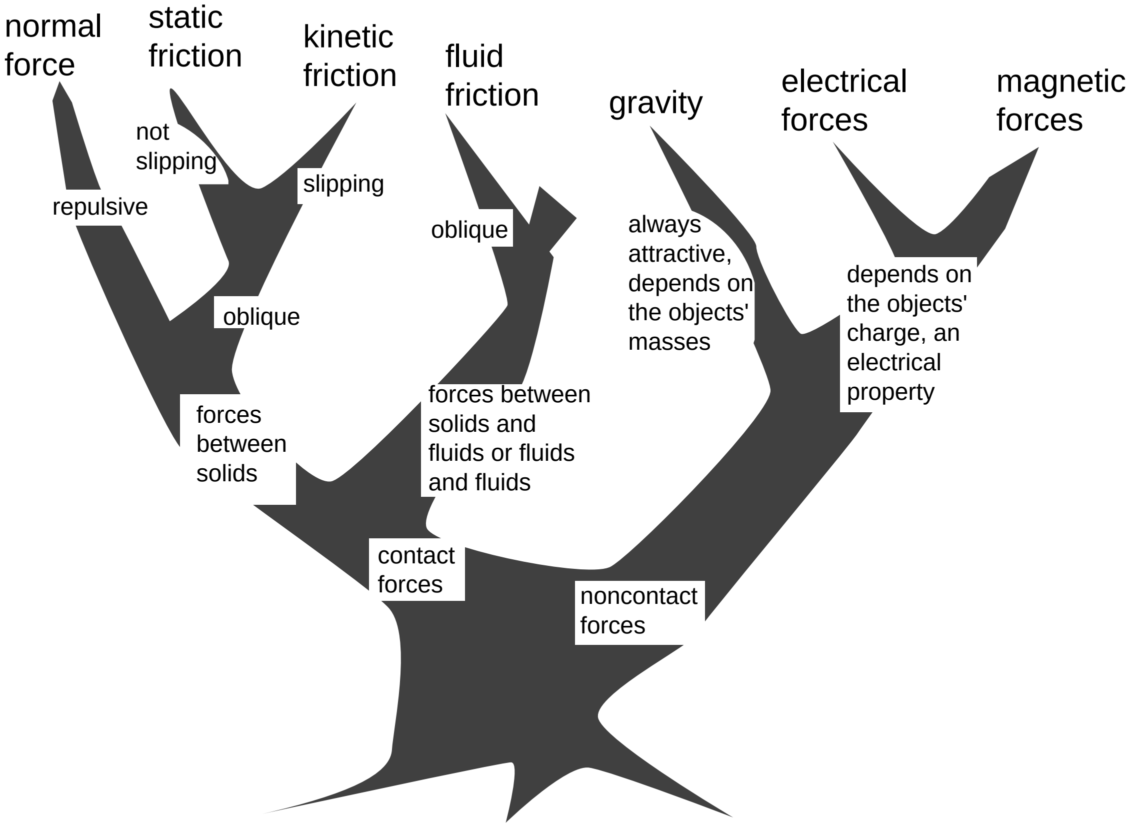
In physics, there are two main systems of classification for forces. At this point in the course, you are going to learn one that is very practical and easy to use, and that splits the forces up into a relatively large number of types: seven very common ones that we'll discuss explicitly in this chapter, plus perhaps ten less important ones such as surface tension, which we will not bother with right now.

Professional physicists, however, are almost all obsessed with finding simple patterns, so recognizing as many as fifteen or twenty types of forces strikes them as distasteful and overly complex. Since about the year 1900, physics has been on an aggressive program to discover ways in which these many seemingly different types of forces arise from a smaller number of fundamental ones. For instance, when you press your hands together, the force that keeps them from passing through each other may seem to have nothing to do with electricity, but at the atomic level, it actually does arise from electrical repulsion between atoms. By about 1950, all the forces of nature had been explained as arising from four fundamental types of forces at the atomic and nuclear level, and the lumping-together process didn't stop there. By the 1960's the length of the list had been reduced to three, and some theorists even believe that they may be able to reduce it to two or one. Although the unification of the forces of nature is one of the most beautiful and important achievements of physics, it makes much more sense to start this course with the more practical and easy system of classification. The unified system of four forces will be one of the highlights of the end of your introductory physics sequence.

The practical classification scheme which concerns us now can be laid out in the form of the tree shown below. The most specific types of forces are shown at the tips of the branches, and it is these types of forces that are referred to in the POFOSTITO mnemonic. For example, electrical and magnetic forces belong to the same general group, but Newton's third law would never relate an electrical force to a magnetic force.

The broadest distinction is that between contact and noncontact forces, which has been discussed in the previous chapter. Among the contact forces, we distinguish between those that involve solids only and those that have to do with fluids, a term used in physics to include both gases and liquids. The terms "repulsive," "attractive," and "oblique" refer to the directions of the forces.

- Repulsive forces are those that tend to push the two participating objects away from each other. More specifically, a repulsive contact force acts perpendicular to the surfaces at which the two objects touch, and a repulsive noncontact force acts along the line between the two objects.
- Attractive forces pull the two objects toward one another, i.e. they act along the same line as repulsive forces, but in the opposite direction.
- Oblique forces are those that act at some other angle.



It should not be necessary to memorize this diagram by rote. It is better to reinforce your memory of this system by calling to mind your commonsense knowledge of certain ordinary phenomena. For instance, we know that the gravitational attraction between us and the planet earth will act even if our feet momentarily leave the ground, and that although magnets have mass and are affected by gravity, most objects that have mass are nonmagnetic.

This diagram is meant to be as simple as possible while including most of the forces we deal with in everyday life. If you were an insect, you would be much more interested in the force of surface tension, which allowed you to walk on water. I have not included the nuclear forces, which are responsible for holding the nuclei of atoms, because they are not evident in everyday life.

You should not be afraid to invent your own names for types of forces that do not fit into the diagram. For instance, the force that holds a piece of tape to the wall has been left off of the tree, and if you were analyzing a situation involving scotch tape, you would be absolutely right to refer to it by some commonsense name such as “sticky force.”

On the other hand, if you are having trouble classifying a certain force, you should also consider whether it is a force at all. For instance, if someone asks you to classify the force that the earth has because of its rotation, you would have great difficulty creating a place for it on the diagram. That’s because it’s a type of motion, not a type of force!

The normal force

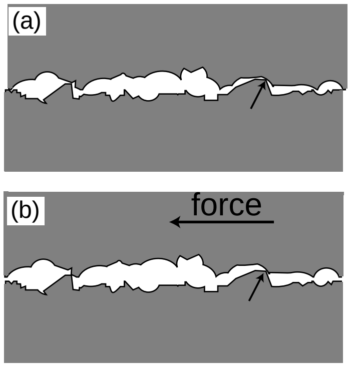
The normal force, F_N, is the force that keeps one solid object from passing through another. “Normal” is simply a fancy word for “perpendicular,” meaning that the force is perpendicular to the surface of contact. Intuitively, it seems the normal force magically adjusts itself to provide whatever force is needed to keep the objects from occupying the same space. If your muscles press your hands together gently, there is a gentle normal force. Press harder, and the normal force gets stronger. How does the normal force know how strong to be? The answer is that the harder you jam your hands together, the more compressed your flesh becomes. Your flesh is acting like a spring: more force is required to compress it more. The same is true when you push on a wall. The wall flexes imperceptibly in proportion to your force on it. If you exerted enough force, would it be possible for two objects to pass through each other? No, typically the result is simply to strain the objects so much that one of them breaks.

Gravitational forces

As we’ll discuss in more detail later in the course, a gravitational force exists between any two things that have mass. In everyday life, the gravitational force between two cars or two people is negligible, so the only noticeable gravitational forces are the ones between the earth and various human-scale objects. We refer to these planet-earth-induced gravitational forces as weight forces, and as we have already seen, their magnitude is given by $|F_W|=mg$.

Static and kinetic friction

If you have pushed a refrigerator across a kitchen floor, you have felt a certain series of sensations. At first, you gradually increased your force on the refrigerator, but it didn't move. Finally, you supplied enough force to unstick the fridge, and there was a sudden jerk as the fridge started moving. Once the fridge is unstuck, you can reduce your force significantly and still keep it moving.



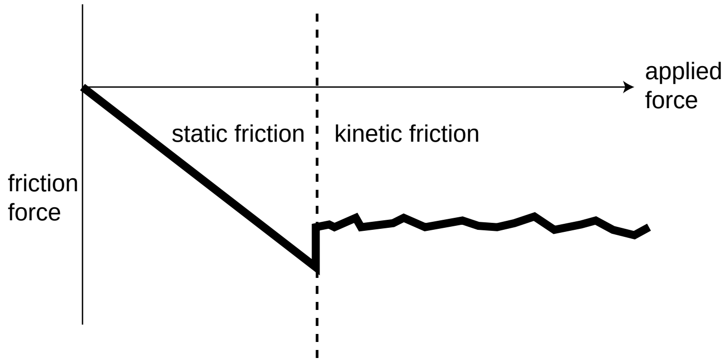
A model that correctly explains many properties of friction. The microscopic bumps and holes in two surfaces dig into each other, causing a frictional force.

While you were gradually increasing your force, the floor's frictional force on the fridge increased in response. The two forces on the fridge canceled, and the fridge didn't accelerate. How did the floor know how to respond with just the right amount of force? The figures on the left show one possible *model* of friction that explains this behavior. (A scientific model is a description that we expect to be incomplete, approximate, or unrealistic in some ways, but that nevertheless succeeds in explaining a variety of phenomena.) Figure (a) shows a microscopic view of the tiny bumps and holes in the surfaces of the floor and the refrigerator. The weight of the fridge presses the two surfaces together, and some of the bumps in one surface will settle as deeply as possible into some of the holes in the other surface. In figure (b), your leftward force on the fridge has caused it to ride up a little higher on the bump in the floor labeled with a small arrow. Still more force is needed to get the fridge over the bump and allow it to start moving. Of course, this is occurring simultaneously at millions of places on the two surfaces.

Once you had gotten the fridge moving at constant speed, you found that you needed to exert less force on it. Since zero total force is needed to make an object move with constant velocity, the floor's rightward frictional force on the fridge has apparently decreased somewhat, making it easier for you to cancel it out. Our model also gives a plausible explanation for this fact: as the surfaces slide past each other, they don't have time to settle down and mesh with one another, so there is less friction.

Even though this model is intuitively appealing and fairly successful, it should not be taken too seriously, and in some situations it is misleading. For instance, fancy racing bikes these days are made with smooth tires that have no tread — contrary to what we'd expect from our model, this does not cause any decrease in friction. Machinists know that two very smooth and clean metal surfaces may stick to each other firmly and be very difficult to slide apart. This cannot be explained in our model, but makes more sense in terms of a model in which friction is described as arising from chemical bonds between the atoms of the two surfaces at their points of contact: very flat surfaces allow more atoms to come in contact.

We choose a coordinate system in which the applied force, i.e. the force trying to move the objects, is positive. The friction force is then negative, since it is in the opposite direction. As you increase the applied force, the force of static friction increases to match it and cancel it out, until the maximum force of static friction is surpassed. The surfaces then begin slipping past each other, and the friction force becomes smaller in absolute value.



Since friction changes its behavior dramatically once the surfaces come unstuck, we define two separate types of frictional forces. *Static friction* is friction that occurs between surfaces that are not slipping over each other. Slipping surfaces experience *kinetic friction*. “Kinetic” means having to do with motion. The forces of static and kinetic friction, notated F_s and F_k, are always parallel to the surface of contact between the two objects.

Self-Check

Does static friction create heat? Kinetic friction?

The maximum possible force of static friction depends on what kinds of surfaces they are, and also on how hard they are being pressed together. The approximate mathematical relationships can be expressed as follows:

$$F_s = -F_{applied}, \text{ when } |F_{applied}| < \mu_s|F_N| \ ,$$

where μ_s is a unitless number, called the coefficient of static friction, which depends on what kinds of surfaces they are. The maximum force that static friction can supply, $\mu_s|F_N|$, represents the boundary between static and kinetic friction. It depends on the normal force, which is numerically equal to whatever force is pressing the two surfaces together. In terms of our model, if the two surfaces are being pressed together more firmly, a greater sideways force will be required in order to make the irregularities in the surfaces ride up and over each other.

Note that just because we use an adjective such as “applied” to refer to a force, that doesn’t mean that there is some special type of force called the “applied force.” The applied force could be any type of force, or it could be the sum of more than one force trying to make an object move.

The force of kinetic friction on each of the two objects is in the direction that resists the slippage of the surfaces. Its magnitude is usually well approximated as

$$|F_k|=\mu_k|F_N|$$

where μ_k is the coefficient of kinetic friction. Kinetic friction is usually more or less independent of velocity.

Only kinetic friction creates heat, as when you rub your hands together. If you move your hands up and down together without sliding them across each other, no heat is produced by the static friction.

Self-Check

Can a frictionless surface exert a normal force? Can a frictional force exist without a normal force?

If you try to accelerate or decelerate your car too quickly, the forces between your wheels and the road become too great, and they begin slipping. This is not good, because kinetic friction is weaker than static friction, resulting in less control. Also, if this occurs while you are turning, the car's handling changes abruptly because the kinetic friction force is in a different direction than the static friction force had been: contrary to the car's direction of motion, rather than contrary to the forces applied to the tire.

Most people respond with disbelief when told of the experimental evidence that both static and kinetic friction are approximately independent of the amount of surface area in contact. Even after doing a hands-on exercise with spring scales to show that it is true, many students are unwilling to believe their own observations, and insist that bigger tires "give more traction." In fact, the main reason why you would not want to put small tires on a big heavy car is that the tires would burst!

Although many people expect that friction would be proportional to surface area, such a proportionality would make predictions contrary to many everyday observations. A dog's feet, for example, have very little surface area in contact with the ground compared to a human's feet, and yet we know that a dog can win a tug-of-war with a child of the same size.

The reason why a smaller surface area does not lead to less friction is that the force between the two surfaces is more concentrated, causing their bumps and holes to dig into each other more deeply.

Fluid friction

Try to drive a nail into a waterfall and you will be confronted with the main difference between solid friction and fluid friction. Fluid friction is purely kinetic; there is no static fluid friction. The nail in the waterfall may tend to get dragged along by the water flowing past it, but it does not stick in the water. The same is true for gases such as air: recall that we are using the word "fluid" to include both gases and liquids.

Unlike solid kinetic friction, the force of fluid friction increases rapidly with velocity. In many cases, the force is approximately proportional to the square of the velocity,

$$F_{\text{fluid friction}} \propto c\rho A v^2 \quad ,$$

where A is the cross-sectional area of the object, ρ is the density of the fluid, and c is a constant of proportionality that depends partly on the type of fluid and partly on how streamlined the object is.

Frictionless ice can certainly make a normal force, since otherwise a hockey puck would sink into the ice. Friction is not possible without a normal force, however: we can see this from the equation, or from common sense, e.g. while sliding down a rope you do not get any friction unless you grip the rope.

Discussion questions

A. A student states that when he tries to push his refrigerator, the reason it won't move is because Newton's third law says there's an equal and opposite frictional force pushing back. After all, the static friction force is equal and opposite to the applied force. How would you convince him he is wrong?

B. Kinetic friction is usually more or less independent of velocity. However, inexperienced drivers tend to produce a jerk at the last moment of deceleration when they stop at a stop light. What does this tell you about the kinetic friction between the brake shoes and the brake drums?

C. Some of the following are correct descriptions of types of forces that could be added on as new branches of the classification tree. Others are not really types of forces, and still others are not force phenomena at all. In each case, decide what's going on, and if appropriate, figure out how you would incorporate them into the tree.

sticky force	makes tape stick to things
opposite force ...	the force that Newton's third law says relates to every force you make
flowing force	the force that water carries with it as it flows out of a hose
surface tension ..	lets insects walk on water
horizontal force .	a force that is horizontal
motor force	the force that a motor makes on the thing it is turning
canceled force ...	a force that is being canceled out by some other force

5.3 Analysis of Forces

Newton's first and second laws deal with the total of all the forces exerted on a specific object, so it is very important to be able to figure out what forces there are. Once you have focused your attention on one object and listed the forces on it, it is also helpful to describe all the corresponding forces that must exist according to Newton's third law. We refer to this as "analyzing the forces" in which the object participates.

Example

A barge is being pulled along a canal by teams of horses on the shores. Analyze all the forces in which the barge participates.

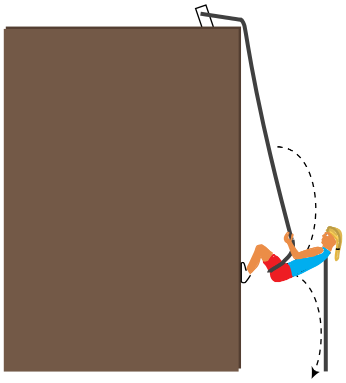
force acting on barge	**force related to it by Newton's third law**
ropes' forward normal forces on barge	barge's backward normal force on ropes
water's backward fluid friction force on barge	barge's forward fluid friction force on water
planet earth's downward gravitational force on barge	barge's upward gravitational force on earth
water's upward "floating" force on barge	barge's downward "floating" force on water

Here I've used the word "floating" force as an example of a sensible invented term for a type of force not classified on the tree in the previous section. A more formal technical term would be "hydrostatic force."

Note how the pairs of forces are all structured as "A's force on B, B's force on A": ropes on barge and barge on ropes; water on barge and barge on water. Because all the forces in the left column are forces acting on the barge, all the forces in the right column are forces being exerted by the barge, which is why each entry in the column begins with "barge."

Often you may be unsure whether you have forgotten one of the forces. Here are three strategies for checking your list:

(1)See what physical result would come from the forces you've found so far. Suppose, for instance, that you'd forgotten the "floating" force on the barge in the example above. Looking at the forces you'd found, you would have found that there was a downward gravitational force on the barge which was not canceled by any upward force. The barge isn't supposed to sink, so you know you need to find a fourth, upward force.



(2) Another technique for finding missing forces is simply to go through the list of all the common types of forces and see if any of them apply.

(3) Make a drawing of the object, and draw a dashed boundary line around it that separates it from its environment. Look for points on the boundary where other objects come in contact with your object. This strategy guarantees that you'll find every contact force that acts on the object, although it won't help you to find non-contact forces.

The following is another example in which we can profit by checking against our physical intuition for what should be happening.

Example

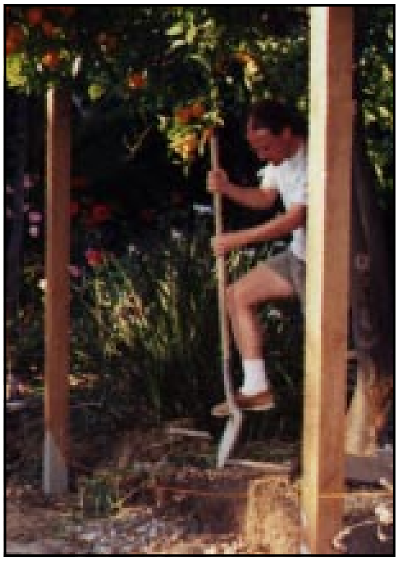
As shown in the figure above, Cindy is rappelling down a cliff. Her downward motion is at constant speed, and she takes little hops off of the cliff, as shown by the dashed line. Analyze the forces in which she participates at a moment when her feet are on the cliff and she is pushing off.

force acting on Cindy	**force related to it by Newton's third law**
planet earth's downward gravitational force on Cindy	Cindy's upward gravitational force on earth
ropes upward frictional force on Cindy (her hand)	Cindy's downward frictional force on the rope
cliff's rightward normal force on Cindy	Cindy's leftward normal force on the cliff

The two vertical forces cancel, which is what they should be doing if she is to go down at a constant rate. The only horizontal force on her is the cliff's force, which is not canceled by any other force, and which therefore will produce an acceleration of Cindy to the right. This makes sense, since she is hopping off. (This solution is a little oversimplified, because the rope is slanting, so it also applies a small leftward force to Cindy. As she flies out to the right, the slant of the rope will increase, pulling her back in more strongly.)

I believe that constructing the type of table described in this section is the best method for beginning students. Most textbooks, however, prescribe a pictorial way of showing all the forces acting on an object. Such a picture is called a free-body diagram. It should not be a big problem if a future physics professor expects you to be able to draw such diagrams, because the conceptual reasoning is the same. You simply draw a picture of the object, with arrows representing the forces that are acting on it. Arrows representing contact forces are drawn from the point of contact, noncontact forces from the center of mass. Free-body diagrams do not show the equal and opposite forces exerted by the object itself.

Discussion questions



Discussion question C.

A. In the example of the barge going down the canal, I referred to a "floating" or "hydrostatic" force that keeps the boat from sinking. If you were adding a new branch on the force-classification tree to represent this force, where would it go?

B. A pool ball is rebounding from the side of the pool table. Analyze the forces in which the ball participates during the short time when it is in contact with the side of the table.

C. The earth's gravitational force on you, i.e. your weight, is always equal to mg, where m is your mass. So why can you get a shovel to go deeper into the ground by jumping onto it? Just because you're jumping, that doesn't mean your mass or weight is any greater, does it?

5.4 Transmission of Forces by Low-Mass Objects

You're walking your dog. The dog wants to go faster than you do, and the leash is taut. Does Newton's third law guarantee that your force on your end of the leash is equal and opposite to the dog's force on its end? If they're not exactly equal, is there any reason why they should be approximately equal?

If there was no leash between you, and you were in direct contact with the dog, then Newton's third law would apply, but Newton's third law cannot relate your force on the leash to the dog's force on the leash, because that would involve three separate objects. Newton's third law only says that your force on the leash is equal and opposite to the leash's force on you,

$$F_{\mathrm{yL}} = -F_{\mathrm{Ly}} \quad ,$$

and that the dog's force on the leash is equal and opposite to its force on the dog

$$F_{\mathrm{dL}} = -F_{\mathrm{Ld}} \quad .$$

Still, we have a strong intuitive expectation that whatever force we make on our end of the leash is transmitted to the dog, and vice-versa. We can analyze the situation by concentrating on the forces that act on the leash, F_{dL} and F_{yL}. According to Newton's second law, these relate to the leash's mass and acceleration:

$$F_{\mathrm{dL}} + F_{\mathrm{yL}} = m_{\mathrm{L}} a_{\mathrm{L}} \quad .$$

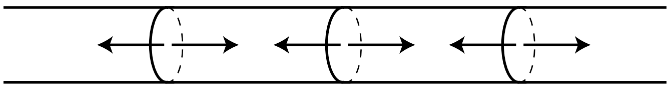
The leash is far less massive then any of the other objects involved, and if m_{L} is very small, then apparently the total force on the leash is also very small, $F_{\mathrm{dL}} + F_{\mathrm{yL}} \approx 0$, and therefore

$$F_{\mathrm{dL}} \approx -F_{\mathrm{yL}} \quad .$$

Thus even though Newton's third law does not apply directly to these two forces, we can approximate the low-mass leash as if it was not intervening between you and the dog. It's at least approximately as if you and the dog were acting directly on each other, in which case Newton's third law would have applied.

In general, low-mass objects can be treated approximately as if they simply transmitted forces from one object to another. This can be true for strings, ropes, and cords, and also for rigid objects such as rods and sticks.

If you look at a piece of string under a magnifying glass as you pull on the ends more and more strongly, you will see the fibers straightening and becoming taut. Different parts of the string are apparently exerting forces on each other. For instance, if we think of the two halves of the string as two objects, then each half is exerting a force on the other half. If we imagine the string as consisting of many small parts, then each segment is transmitting a force to the next segment, and if the string has very little



If we imagine dividing a taut rope up into small segments, then any segment has forces pulling outward on it at each end. If the rope is of negligible mass, then all the forces equal +T or -T, where T, the tension, is a single number.

mass, then all the forces are equal in magnitude. We refer to the magnitude of the forces as the tension in the string, T. Although the tension is measured in units of Newtons, it is not itself a force. There are many forces within the string, some in one direction and some in the other direction, and their magnitudes are only approximately equal. The concept of tension only makes sense as a general, approximate statement of how big all the forces are.

If a rope goes over a pulley or around some other object, then the tension throughout the rope is approximately equal so long as there is not too much friction. A rod or stick can be treated in much the same way as a string, but it is possible to have either compression or tension.

Since tension is not a type of force, the force exerted by a rope on some other object must be of some definite type such as static friction, kinetic friction, or a normal force. If you hold your dog's leash with your hand through the loop, then the force exerted by the leash on your hand is a normal force: it is the force that keeps the leash from occupying the same space as your hand. If you grasp a plain end of a rope, then the force between the rope and your hand is a frictional force.

A more complex example of transmission of forces is the way a car accelerates. Many people would describe the car's engine as making the force that accelerates the car, but the engine is part of the car, so that's impossible: objects can't make forces on themselves. What really happens is that the engine's force is transmitted through the transmission to the axles, then through the tires to the road. By Newton's third law, there will thus be a forward force from the road on the tires, which accelerates the car.

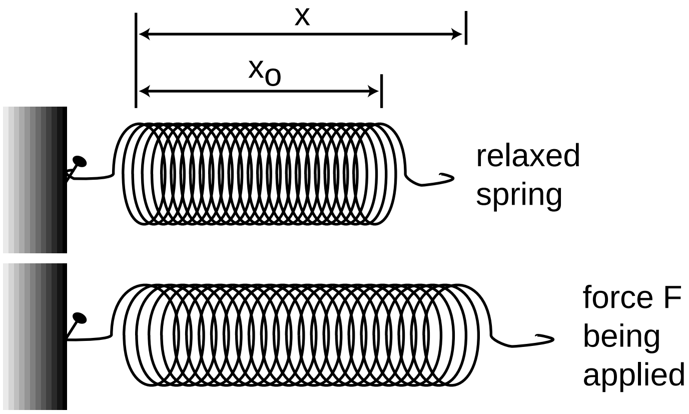
Discussion question

When you step on the gas pedal, is your foot's force being transmitted in the sense of the word used in this section?

5.5 Objects Under Strain

A string lengthens slightly when you stretch it. Similarly, we have already discussed how an apparently rigid object such as a wall is actually flexing when it participates in a normal force. In other cases, the effect is more obvious. A spring or a rubber band visibly elongates when stretched.

Common to all these examples is a change in shape of some kind: lengthening, bending, compressing, etc. The change in shape can be measured by picking some part of the object and measuring its position, x. For concreteness, let's imagine a spring with one end attached to a wall. When no force is exerted, the unfixed end of the spring is at some position x_o. If a force acts at the unfixed end, its position will change to some new value of x. The more force, the greater the departure of x from x_o.



Back in Newton's time, experiments like this were considered cutting-edge research, and his contemporary Hooke is remembered today for doing them and for coming up with a simple mathematical generalization called Hooke's law:

$$F \approx k(x\text{-}x_o) \quad \text{[force required to stretch a spring; valid for small forces only]} \ .$$

Here k is a constant, called the spring constant, that depends on how stiff the object is. If too much force is applied, the spring exhibits more complicated behavior, so the equation is only a good approximation if the force is sufficiently small. Usually when the force is so large that Hooke's law is a bad approximation, the force ends up permanently bending or breaking the spring.

Although Hooke's law may seem like a piece of trivia about springs, it is actually far more important than that, because all solid objects exert Hooke's-law behavior over some range of sufficiently small forces. For example, if you push down on the hood of a car, it dips by an amount that is directly proportional to the force. (But the car's behavior would not be as mathematically simple if you dropped a boulder on the hood!)

Discussion questions

A. A car is connected to its axles through big, stiff springs called shock absorbers, or “shocks.” Although we’ve discussed Hooke’s law above only in the case of stretching a spring, a car’s shocks are continually going through both stretching and compression. In this situation, how would you interpret the positive and negative signs in Hooke’s law?

5.6 Simple Machines: The Pulley

Even the most complex machines, such as cars or pianos, are built out of certain basic units called *simple machines*. The following are some of the main functions of simple machines:

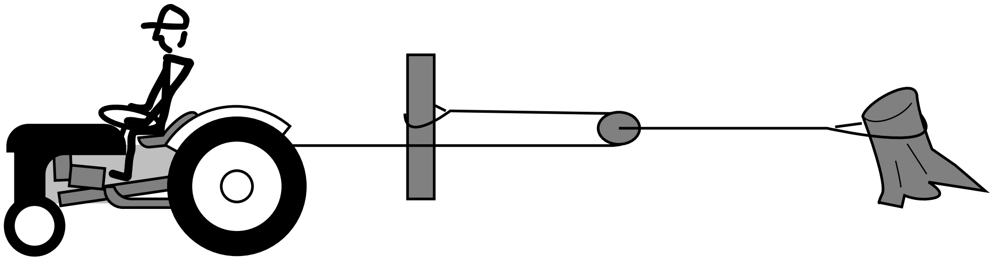
transmitting a force: The chain on a bicycle transmits a force from the crank set to the rear wheel.

changing the direction of a force: If you push down on a seesaw, the other end goes up.

changing the speed and precision of motion: When you make the “come here” motion, your biceps only moves a couple of centimeters where it attaches to your forearm, but your arm moves much farther and more rapidly.

changing the amount of force: A lever or pulley can be used to increase or decrease the amount of force.

You are now prepared to understand one-dimensional simple machines, of which the pulley is the main example.



Example: a pulley
Question: Farmer Bill says this pulley arrangement doubles the force of his tractor. Is he just a dumb hayseed, or does he know what he’s doing?
Solution: To use Newton’s first law, we need to pick an object and consider the sum of the forces on it. Since our goal is to relate the tension in the part of the cable attached to the stump to the tension in the part attached to the tractor, we should pick an object to which both those cables are attached, i.e. the pulley itself. As discussed in section 5.4, the tension in a string or cable remains approximately constant as it passes around a pulley, provided that there is not too much friction. There are therefore two leftward forces acting on the pulley, each equal to the force exerted by the tractor. Since the acceleration of the pulley is essentially zero, the forces on it must be canceling out, so the rightward force of the pulley-stump cable on the pulley must be double the force exerted by the tractor. Yes, Farmer Bill knows what he’s talking about.

Summary

Selected Vocabulary

repulsive	describes a force that tends to push the two participating objects apart
attractive	describes a force that tends to pull the two participating objects together
oblique	describes a force that acts at some other angle, one that is not a direct repulsion or attraction
normal force	the force that keeps two objects from occupying the same space
static friction	a friction force between surfaces that are not slipping past each other
kinetic friction	a friction force between surfaces that are slipping past each other
fluid	a gas or a liquid
fluid friction	a friction force in which at least one of the object is is a fluid
spring constant	the constant of proportionality between force and elongation of a spring or other object under strain

Notation

F_N	a normal force
F_s	a static frictional force
F_k	a kinetic frictional force
μ_s	the coefficient of static friction; the constant of proportionality between the maximum static frictional force and the normal force; depends on what types of surfaces are involved
μ_k	the coefficient of kinetic friction; the constant of proportionality between the kinetic frictional force and the normal force; depends on what types of surfaces are involved
k	the spring constant; the constant of proportionality between the force exerted on an object and the amount by which the object is lengthened or compressed

Summary

Newton’s third law states that forces occur in equal and opposite pairs. If object A exerts a force on object B, then object B must simultaneously be exerting an equal and opposite force on object A. Each instance of Newton’s third law involves exactly two objects, and exactly two forces, which are of the same type.

There are two systems for classifying forces. We are presently using the more practical but less fundamental one. In this system, forces are classified by whether they are repulsive, attractive, or oblique; whether they are contact or noncontact forces; and whether the two objects involved are solids or fluids.

Static friction adjusts itself to match the force that is trying to make the surfaces slide past each other, until the maximum value is reached,

$$|F_s|<\mu_s|F_N| \quad .$$

Once this force is exceeded, the surfaces slip past one another, and kinetic friction applies,

$$|F_k|=\mu_k|F_N| \quad .$$

Both types of frictional force are nearly independent of surface area, and kinetic friction is usually approximately independent of the speed at which the surfaces are slipping.

A good first step in applying Newton’s laws of motion to any physical situation is to pick an object of interest, and then to list all the forces acting on that object. We classify each force by its type, and find its Newton’s-third-law partner, which is exerted by the object on some other object.

When two objects are connected by a third low-mass object, their forces are transmitted to each other nearly unchanged.

Objects under strain always obey Hooke’s law to a good approximation, as long as the force is small. Hooke’s law states that the stretching or compression of the object is proportional to the force exerted on it,

$$F\approx k(x-x_o) \quad .$$

Homework Problems

1. If a big truck and a VW bug collide head-on, which will be acted on by the greater force? Which will have the greater acceleration?

2. The earth is attracted to an object with a force equal and opposite to the force of the earth on the object. If this is true, why is it that when you drop an object, the earth does not have an acceleration equal and opposite to that of the object?

3. When you stand still, there are two forces acting on you, the force of gravity (your weight) and the normal force of the floor pushing up on your feet. Are these forces equal and opposite? Does Newton's third law relate them to each other? Explain.

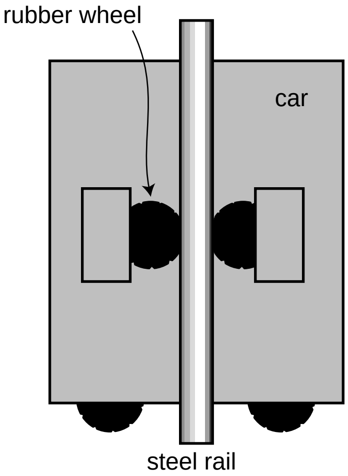
4. A magnet is stuck underneath a car. Analyze the forces in which the magnet participates, using a table in the format shown in section 5.3.

5. Give two examples of objects at rest relative to the earth that are being kept from falling by forces other than the normal force. Do not use objects in outer space. In each case, analyze the forces.

6. A person is rowing a boat, with her legs extended and her feet braced. Analyze the forces in which she participates while she is doing the part of the stroke that propels the boat, with the ends of the oars in the water (not the part where the oars are lifted out of the water).

7. A farmer is in a stall with a cow when the cow decides to press him against the wall, pinning him with his feet off the ground. Analyze the forces in which the farmer participates.

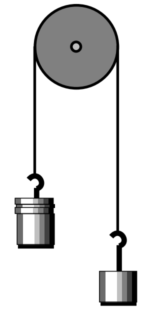
8. A propeller plane is cruising east at constant speed and altitude. Analyze the forces in which the plane participates.



Problem 9.

9 ✓. Today's tallest buildings are really not that much taller than the tallest buildings of the 1940s. The main problem with making an even taller skyscraper is that every elevator needs its own shaft running the whole height of the building. So many elevators are needed to serve the building's thousands of occupants that the elevator shafts start taking up too much of the space within the building. An alternative is to have elevators that can move both horizontally and vertically: with such a design, many elevator cars can share a few shafts, and they don't get in each other's way too much because they can detour around each other. In this design, it becomes impossible to hang the cars from cables, so they would instead have to ride on rails which they grab onto with wheels. Friction would keep them from slipping. The figure shows such a frictional elevator in its vertical travel mode. (The wheels on the bottom are for when it needs to switch to horizontal motion.) (a) If the coefficient of static friction between rubber and steel is μ_s, and the maximum mass of the car plus its passengers is M, how much force must there be pressing each wheel against the rail in order to keep the car from slipping? (Assume the car is not accelerating.) (b) Show that your result has physically reasonable behavior with respect to μ_s. In other words, if there was less friction, would the wheels need to be pressed more firmly or less firmly? Does your equation behave that way?

S A solution is given in the back of the book.	★ A difficult problem.
✓ A computerized answer check is available.	∫ A problem that requires calculus.



Problem 10.

10. Unequal masses *M* and *m* are suspended from a pulley as shown in the figure.

(a) Analyze the forces in which mass *m* participates, using a table the format shown in section 5.3. [The forces in which the other masses participate will of course be similar, but not numerically the same.]

(b) Find the magnitude of the accelerations of the two masses. [Hints: (1) Pick a coordinate system, and use positive and negative signs consistently to indicate the directions of the forces and accelerations. (2) The two accelerations of the two masses have to be equal in magnitude but of opposite signs, since one side eats up rope at the same rate at which the other side pays it out. (3) You need to apply Newton's second law twice, once to each mass, and then solve the two equations for the unknowns: the acceleration, *a*, and the tension in the rope, *T*.]

(c) Many people expect that in the special case of $M=m$, the two masses will naturally settle down to an equilibrium position side by side. Based on your answer from part (b), is this correct?

11. A tugboat of mass *m* pulls a ship of mass *M*, accelerating it. Ignore fluid friction acting on their hulls, although there will of course need to be fluid friction acting on the tug's propellers.

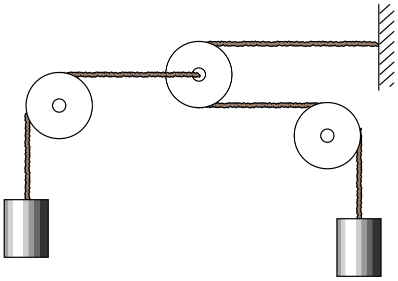
(a) Analyze the forces in which the tugboat participates, using a table in the format shown in section 5.3. Don't worry about vertical forces.

(b) Do the same for the ship.

(c) Assume now that water friction on the two vessels' hulls is negligible. If the force acting on the tug's propeller is *F*, what is the tension, *T*, in the cable connecting the two ships? [Hint: Write down two equations, one for Newton's second law applied to each object. Solve these for the two unknowns *T* and *a*.]

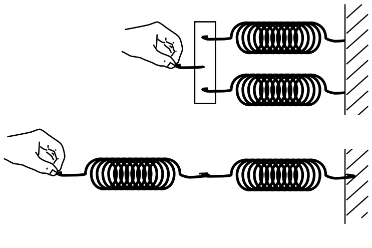
(d) Interpret your answer in the special cases of $M=0$ and $M=\infty$.

12. Explain why it wouldn't make sense to have kinetic friction be stronger than static friction.



Problem 13.

13. In the system shown in the figure, the pulleys on the left and right are fixed, but the pulley in the center can move to the left or right. The two masses are identical. Show that the mass on the left will have an upward acceleration equal to $g/5$.

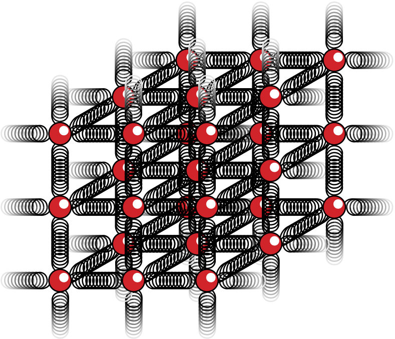


Problem 14.

14 S. The figure shows two different ways of combining a pair of identical springs, each with spring constant *k*. We refer to the top setup as parallel, and the bottom one as a series arrangement. (a) For the parallel arrangement, analyze the forces acting on the connector piece on the left, and then use this analysis to determine the equivalent spring constant of the whole setup. Explain whether the combined spring constant should be interpreted as being stiffer or less stiff. (b) For the series arrangement, analyze the forces acting on each spring and figure out the same things.

15. Generalize the results of problem 14 to the case where the two spring constants are unequal.

16 S. (a) Using the solution of problem 14, which is given in the back of the book, predict how the spring constant of a fiber will depend on its length and cross-sectional area. (b) The constant of proportionality is called the Young's modulus, *E*, and typical values of the Young's modulus are about 10^{10} to 10^{11}. What units would the Young's modulus have in the SI (meter-kilogram-second) system?



Problem 17.

17. This problem depends on the results of problems 14 and 16, whose solutions are in the back of the book. When atoms form chemical bonds, it makes sense to talk about the spring constant of the bond as a measure of how "stiff" it is. Of course, there aren't really little springs — this is just a mechanical model. The purpose of this problem is to estimate the spring constant, *k*, for a single bond in a typical piece of solid matter. Suppose we have a fiber, like a hair or a piece of fishing line, and imagine for simplicity that is made of atoms of a single element stacked in a cubical manner, as shown in the figure, with a center-to-center spacing *b*. A typical value for *b* would be about 10^{-10} m. (a) Find an equation for *k* in terms of *b*, and in terms of the Young's modulus, *E*, defined in problem 16 and its solution. (b) Estimate *k* using the numerical data given in problem 16. (c) Suppose you could grab one of the atoms in a diatomic molecule like H_2 or O_2, and let the other atom hang vertically below it. Does the bond stretch by any appreciable fraction due to gravity?

Motion in Three Dimensions

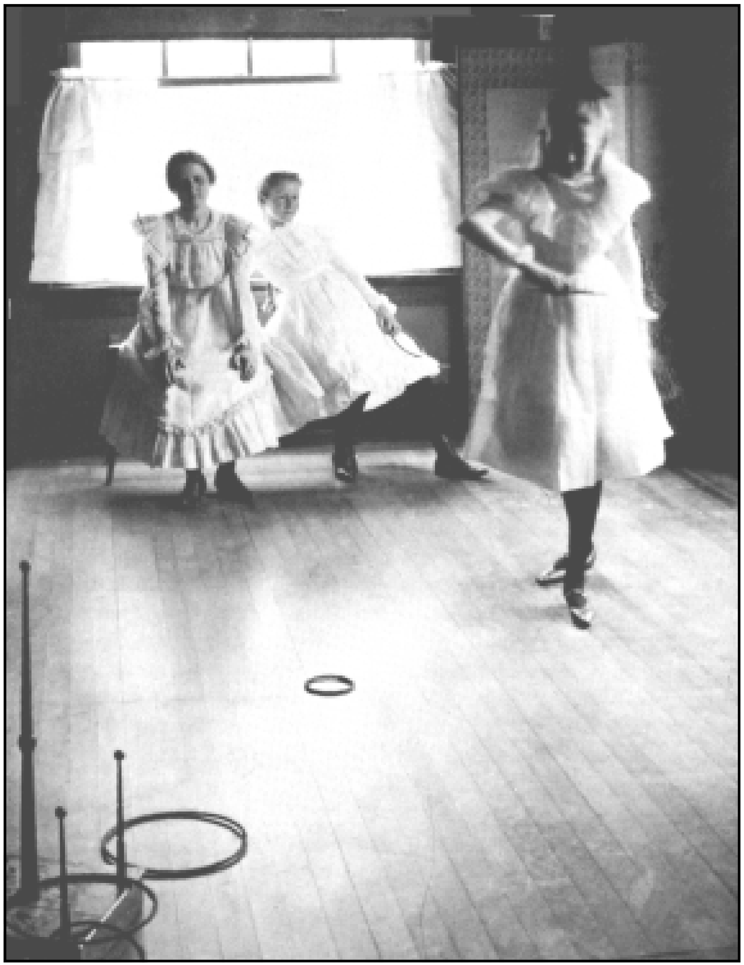
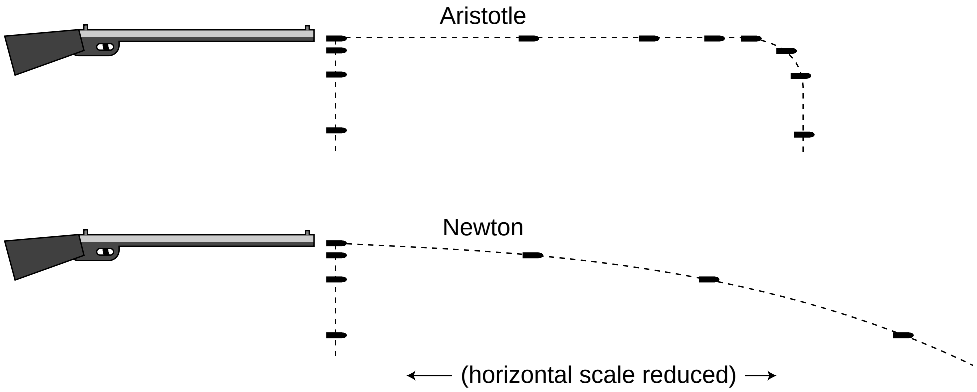


Photo by Clarence White, ca. 1903.

6 Newton's Laws in Three Dimensions

6.1 Forces Have No Perpendicular Effects

Suppose you could shoot a rifle and arrange for a second bullet to be dropped from the same height at the exact moment when the first left the barrel. Which would hit the ground first? Nearly everyone expects that the dropped bullet will reach the dirt first, and Aristotle would have agreed. Aristotle would have described it like this. The shot bullet receives some forced motion from the gun. It travels forward for a split second, slowing



down rapidly because there is no longer any force to make it continue in motion. Once it is done with its forced motion, it changes to natural motion, i.e. falling straight down. While the shot bullet is slowing down, the dropped bullet gets on with the business of falling, so according to Aristotle it will hit the ground first.

Luckily, nature isn't as complicated as Aristotle thought! To convince yourself that Aristotle's ideas were wrong and needlessly complex, stand up now and try this experiment. Take your keys out of your pocket, and begin walking briskly forward. Without speeding up or slowing down, release your keys and let them fall while you continue walking at the same pace.

You have found that your keys hit the ground right next to your feet. Their horizontal motion never slowed down at all, and the whole time they were dropping, they were right next to you. The horizontal motion and the vertical motion happen at the same time, and they are independent of each other. Your experiment proves that the horizontal motion is unaffected by the vertical motion, but it's also true that the vertical motion is not changed in any way by the horizontal motion. The keys take exactly the same amount of time to get to the ground as they would have if you simply dropped them, and the same is true of the bullets: both bullets hit the ground simultaneously.

These have been our first examples of motion in more than one dimension, and they illustrate the most important new idea that is required to understand the three-dimensional generalization of Newtonian physics:

> **Forces have no perpendicular effects.**
> When a force acts on an object, it has no effect on the part of the object's motion that is perpendicular to the force.

In the examples above, the vertical force of gravity had no effect on the horizontal motions of the objects. These were examples of projectile motion, which interested people like Galileo because of its military applications. The principle is more general than that, however. For instance, if a rolling ball is initially heading straight for a wall, but a steady wind begins blowing from the side, the ball does not take any longer to get to the wall. In the case of projectile motion, the force involved is gravity, so we can say more specifically that the vertical acceleration is 9.8 m/s^2, regardless of the horizontal motion.

Relationship to relative motion

These concepts are directly related to the idea that motion is relative. Galileo's opponents argued that the earth could not possibly be rotating as he claimed, because then if you jumped straight up in the air you wouldn't be able to come down in the same place. Their argument was based on their incorrect Aristotelian assumption that once the force of gravity began to act on you and bring you back down, your horizontal motion would stop. In the correct Newtonian theory, the earth's downward gravitational force is acting before, during, and after your jump, but has no effect on your motion in the perpendicular (horizontal) direction.

If Aristotle had been correct, then we would have a handy way to determine absolute motion and absolute rest: jump straight up in the air, and if you land back where you started, the surface from which you jumped must have been in a state of rest. In reality, this test gives the same result as long as the surface under you is an inertial frame. If you try this in a jet plane, you land back on the same spot on the deck from which you started, regardless of whether the plane is flying at 500 miles per hour or parked on the runway. The method would in fact only be good for detecting whether the plane was accelerating.

Discussion Questions

A. The following is an incorrect explanation of a fact about target shooting:

> "Shooting a high-powered rifle with a high muzzle velocity is different from shooting a less powerful gun. With a less powerful gun, you have to aim quite a bit above your target, but with a more powerful one you don't have to aim so high because the bullet doesn't drop as fast."

What is the correct explanation?

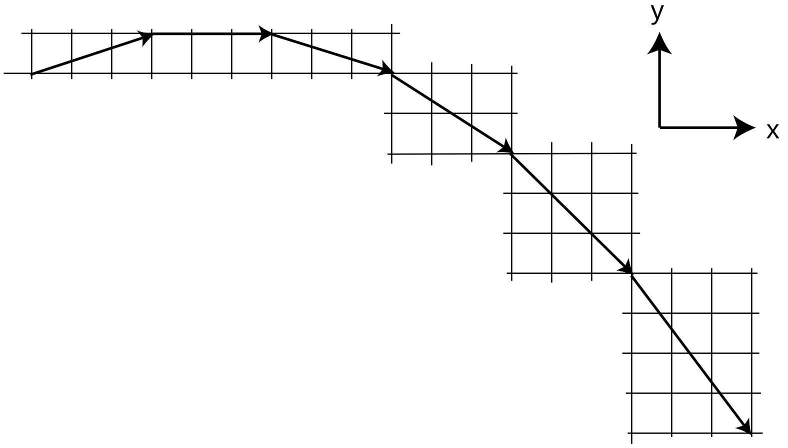
B. You have thrown a rock, and it is flying through the air in an arc. If the earth's gravitational force on it is always straight down, why doesn't it just go straight down once it leaves your hand?

C. Consider the example of the bullet that is dropped at the same moment another bullet is fired from a gun. What would the motion of the two bullets look like to a jet pilot flying alongside in the same direction as the shot bullet and at the same horizontal speed?

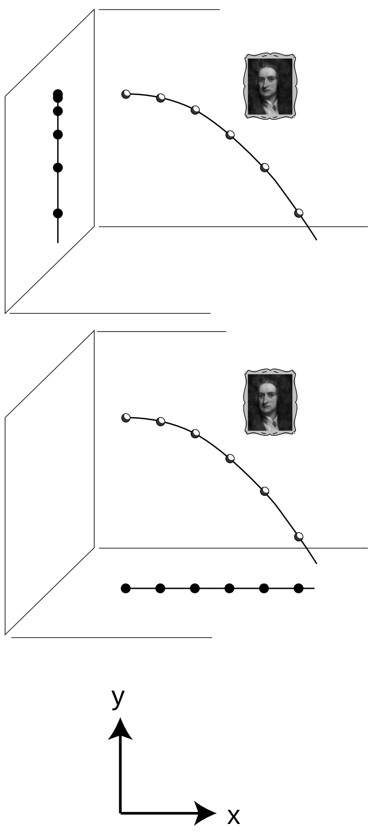
6.2 Coordinates and Components

‘Cause we’re all
Bold as love,
Just ask the axis.
-Jimi Hendrix

How do we convert these ideas into mathematics? The figure below shows a good way of connecting the intuitive ideas to the numbers. In one dimension, we impose a number line with an *x* coordinate on a certain stretch of space. In two dimensions, we imagine a grid of squares which we label with *x* and *y* values.



This object experiences a force that pulls it down toward the bottom of the page. In each equal time interval, it moves three units to the right. At the same time, its vertical motion is making a simple pattern of +1, 0, -1, -2, -3, -4, ... units. Its motion can be described by an *x* coordinate that has zero acceleration and a *y* coordinate with constant acceleration. The arrows labeled *x* and *y* serve to explain that we are defining increasing *x* to the right and increasing *y* as upward.



But of course motion doesn’t really occur in a series of discrete hops like in chess or checkers. The figure on the left shows a way of conceptualizing the smooth variation of the *x* and *y* coordinates. The ball’s shadow on the wall moves along a line, and we describe its position with a single coordinate, *y*, its height above the floor. The wall shadow has a constant acceleration of –9.8 m/s^2. A shadow on the floor, made by a second light source, also moves along a line, and we describe its motion with an *x* coordinate, measured from the wall.

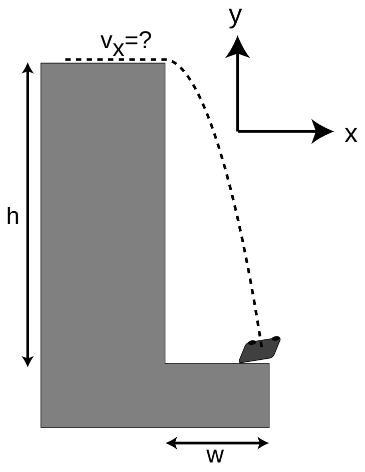
The velocity of the floor shadow is referred to as the *x component* of the velocity, written v_x. Similarly we can notate the acceleration of the floor shadow as a_x. Since v_x is constant, a_x is zero.

Similarly, the velocity of the wall shadow is called v_y, its acceleration a_y. This example has a_y=–9.8 m/s^2.

Because the earth’s gravitational force on the ball is acting along the *y* axis, we say that the force has a negative *y* component, F_y, but F_x=F_z=0.

The general idea is that we imagine two observers, each of whom perceives the entire universe as if it was flattened down to a single line. The *y*-observer, for instance, perceives *y*, v_y, and a_y, and will infer that there is a force, F_y, acting downward on the ball. That is, a *y* component means the aspect of a physical phenomenon, such as velocity, acceleration, or force, that is observable to someone who can only see motion along the *y* axis.

All of this can easily be generalized to three dimensions. In the example above, there could be a *z*-observer who only sees motion toward or away from the back wall of the room.



Example: a car going over a cliff

Question: The police find a car at a distance w=20 m from the base of a cliff of height h=100 m. How fast was the car going when it went over the edge? Solve the problem symbolically first, then plug in the numbers.

Solution: Let's choose y pointing up and x pointing away from the cliff. The car's vertical motion was independent of its horizontal motion, so we know it had a constant vertical acceleration of a=-g=-9.8 m/s^2. The time it spent in the air is therefore related to the vertical distance it fell by the constant-acceleration equation

$$\Delta y = \frac{1}{2}a_y\Delta t^2 \quad ,$$

or

$$-h = \frac{1}{2}(-g)\Delta t^2 \quad .$$

Solving for Δt gives

$$\Delta t = \sqrt{\frac{2h}{g}} \quad .$$

Since the vertical force had no effect on the car's horizontal motion, it had a_x=0, i.e. constant horizontal velocity. We can apply the constant-velocity equation

$$v_x = \frac{\Delta x}{\Delta t} \quad ,$$

i.e.

$$v_x = \frac{w}{\Delta t} \quad .$$

We now substitute for Δt to find

$$v_x = w\,/\sqrt{\frac{2h}{g}} \quad ,$$

which simplifies to

$$v_x = w\sqrt{\frac{g}{2h}} \quad .$$

Plugging in numbers, we find that the car's speed when it went over the edge was 4 m/s, or about 10 mi/hr.

Projectiles move along parabolas

What type of mathematical curve does a projectile follow through space? To find out, we must relate x to y, eliminating t. The reasoning is very similar to that used in the example above. Arbitrarily choosing x=y=t=0 to be at the top of the arc, we conveniently have x=Δx, y=Δy, and t=Δt, so

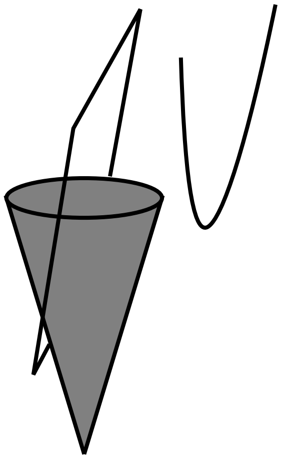
$$y = -\frac{1}{2}a_y t^2$$

$$x = v_x t$$

We solve the second equation for t=x/v_x and eliminate t in the first equation:

$$y = -\frac{1}{2}a_y\left(\frac{x}{v_x}\right)^2 \quad .$$

Since everything in this equation is a constant except for x and y, we conclude that y is proportional to the square of x. As you may or may not recall from a math class, $y \propto x^2$ describes a parabola.



A parabola can be defined as the shape made by cutting a cone parallel to its side. A parabola is also the graph of an equation of the form $y\propto x^2$.

Each water droplet follows a parabola. The faster drops' parabolas are bigger.

Discussion Question

A. At the beginning of this section I represented the motion of a projectile on graph paper, breaking its motion into equal time intervals. Suppose instead that there is no force on the object at all. It obeys Newton's first law and continues without changing its state of motion. What would the corresponding graph-paper diagram look like? If the time interval represented by each arrow was 1 second, how would you relate the graph-paper diagram to the velocity components v_x and v_y?

B. Make up several different coordinate systems oriented in different ways, and describe the a_x and a_y of a falling object in each one.

6.3 Newton's Laws in Three Dimensions

It is now fairly straightforward to extend Newton's laws to three dimensions:

Newton's First Law
If all three components of the total force on an object are zero, then it will continue in the same state of motion.

Newton's Second Law
An object's acceleration components are predicted by the equations

$$a_x = F_{x,total}/m \quad ,$$

$$a_y = F_{y,total}/m \quad \text{, and}$$

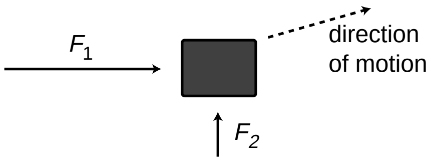
$$a_z = F_{z,total}/m \quad .$$

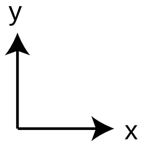
Newton's Third Law
If two objects A and B interact via forces, then the components of their forces on each other are equal and opposite:

$$F_{\text{A on B},x} = -F_{\text{B on A},x} \quad ,$$

$$F_{\text{A on B},y} = -F_{\text{B on A},y} \quad \text{, and}$$

$$F_{\text{A on B},z} = -F_{\text{B on A},z} \quad .$$





Example: forces in perpendicular directions on the same object

Question: An object is initially at rest. Two constant forces begin acting on it, and continue acting on it for a while. As suggested by the two arrows, the forces are perpendicular, and the rightward force is stronger. What happens?

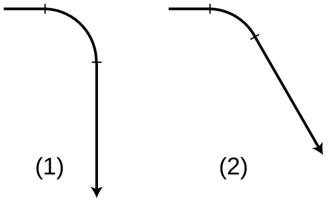
Answer: Aristotle believed, and many students still do, that only one force can "give orders" to an object at one time. They therefore think that the object will begin speeding up and moving in the direction of the stronger force. In fact the object will move along a diagonal. In the example shown in the figure, the object will respond to the large rightward force with a large acceleration component to the right, and the small upward force will give it a small acceleration component upward. The stronger force does not overwhelm the weaker force, or have any effect on the upward motion at all. The force components simply add together:

$$F_{x,total} = F_{1,x} + \overbrace{F_{2,x}}^{=0}$$

$$F_{y,total} = \overbrace{F_{1,y}}^{=0} + F_{2,y}$$

Discussion Question

The figure shows two trajectories, made by splicing together lines and circular arcs, which are unphysical for an object that is only being acted on by gravity. Prove that they are impossible based on Newton's laws.



Summary

Selected Vocabulary

component .... the part of a velocity, acceleration, or force that would be perceptible to an observer who could only see the universe projected along a certain one-dimensional axis

parabola .... the mathematical curve whose graph has *y* proportional to x^2

Notation

x, y, z .... an object's positions along the x, y, and z axes

v_x, v_y, v_z .... the x, y, and z components of an object's velocity; the rates of change of the object's x, y, and z coordinates

a_x, a_y, a_z .... the x, y, and z components of an object's acceleration; the rates of change of v_x, v_y, and v_z

Summary

A force does not produce any effect on the motion of an object in a perpendicular direction. The most important application of this principle is that the horizontal motion of a projectile has zero acceleration, while the vertical motion has an acceleration equal to g. That is, an object's horizontal and vertical motions are independent. The arc of a projectile is a parabola.

Motion in three dimensions is measured using three coordinates, x, y, and z . Each of these coordinates has its own corresponding velocity and acceleration. We say that the velocity and acceleration both have x, y, and z components

Newton's second law is readily extended to three dimensions by rewriting it as three equations predicting the three components of the acceleration,

$$a_x = F_{x,\text{total}}/m \ ,$$
$$a_y = F_{y,\text{total}}/m \ ,$$
$$a_z = F_{z,\text{total}}/m \ ,$$

and likewise for the first and third laws.

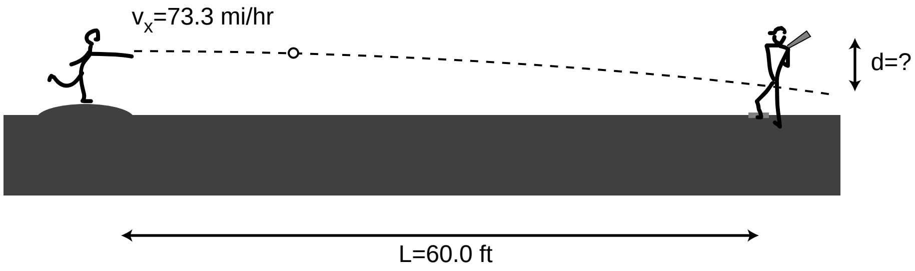
Homework Problems

1. (a) A ball is thrown straight up with velocity v. Find an equation for the height to which it rises.
(b) Generalize your equation for a ball thrown at an angle θ above horizontal, in which case its initial velocity components are $v_x=v\cos\theta$ and $v_y=v\sin\theta$.

2. At the Salinas Lettuce Festival Parade, Miss Lettuce of 1996 drops her bouquet while riding on a float. Compare the shape of its trajectory as seen by her to the shape seen by one of her admirers standing on the sidewalk.

3 . Two daredevils, Wendy and Bill, go over Niagara Falls. Wendy sits in an inner tube, and lets the 30 km/hr velocity of the river throw her out horizontally over the falls. Bill paddles a kayak, adding an extra 10 km/hr to his velocity. They go over the edge of the falls at the same moment, side by side. Ignore air friction. Explain your reasoning.
(a) Who hits the bottom first?
(b) What is the horizontal component of Wendy's velocity on impact?
(c) What is the horizontal component of Bill's velocity on impact?
(d) Who is going faster on impact?

4. A baseball pitcher throws a pitch clocked at v_x=73.3 mi/h. He throws horizontally. By what amount, d, does the ball drop by the time it reaches home plate, L=60.0 ft away? (a) First find a symbolic answer in terms of L, v_x, and g. (b✓) Plug in and find a numerical answer. Express your answer in units of ft. [Note: 1 ft=12 in, 1 mi=5280 ft, and 1 in=2.54 cm]



5 S. A cannon standing on a flat field fires a cannonball with a muzzle velocity v, at an angle θ above horizontal. The cannonball thus initially has velocity components $v_x=v\cos\theta$ and $v_y=v\sin\theta$.

(a) Show that the cannon’s range (horizontal distance to where the cannonball falls) is given by the equation $R=\dfrac{2v^2\sin\theta\cos\theta}{g}$.

(b) Interpret your equation in the cases of θ=0 and θ=90°.

S A solution is given in the back of the book.
✓ A computerized answer check is available.
★ A difficult problem.
∫ A problem that requires calculus.

6 ∫. Assuming the result of the previous problem for the range of a projectile, $R = \frac{2v^2 \sin\theta \cos\theta}{g}$, show that the maximum range is for θ=45°.

Vectors are used in aerial navigation.

7 Vectors

7.1 Vector Notation

The idea of components freed us from the confines of one-dimensional physics, but the component notation can be unwieldy, since every one-dimensional equation has to be written as a set of three separate equations in the three-dimensional case. Newton was stuck with the component notation until the day he died, but eventually someone sufficiently lazy and clever figured out a way of abbreviating three equations as one.

(a)	$\vec{F}_{\text{A on B}} = -\vec{F}_{\text{B on A}}$	stands for	$F_{\text{A on B,x}} = -F_{\text{B on A, x}}$ $F_{\text{A on B,y}} = -F_{\text{B on A, y}}$ $F_{\text{A on B,z}} = -F_{\text{B on A, z}}$
(b)	$\vec{F}_{\text{total}} = \vec{F}_1 + \vec{F}_2 + \ldots$	stands for	$F_{\text{total,x}} = F_{\text{1,x}} + F_{\text{2,x}} + \ldots$ $F_{\text{total,y}} = F_{\text{1,y}} + F_{\text{2,y}} + \ldots$ $F_{\text{total,z}} = F_{\text{1,z}} + F_{\text{2,z}} + \ldots$
(c)	$\vec{a} = \dfrac{\Delta\vec{v}}{\Delta t}$	stands for	$a_x = \Delta v_x \,/\, \Delta t$ $a_y = \Delta v_y \,/\, \Delta t$ $a_z = \Delta v_z \,/\, \Delta t$

Example (a) shows both ways of writing Newton's third law. Which would you rather write?

The idea is that each of the algebra symbols with an arrow written on top, called a *vector*, is actually an abbreviation for three different numbers, the *x*, *y*, and *z* components. The three components are referred to as the components of the vector, e.g. F_x is the *x* component of the vector $\vec{F}$. The notation with an arrow on top is good for handwritten equations, but is unattractive in a printed book, so books use boldface, ***F***, to represent vectors. After this point, I'll use boldface for vectors throughout this book.

A vector has both a direction and an amount. A scalar has only an amount.

In general, the vector notation is useful for any quantity that has both an amount and a direction in space. Even when you are not going to write any actual vector notation, the concept itself is a useful one. We say that force and velocity, for example, are vectors. A quantity that has no direction in space, such as mass or time, is called a *scalar*. The amount of a vector quantity is called its *magnitude*. The notation for the magnitude of a vector ***A*** is |***A***|, like the absolute value sign used with scalars.

Often, as in example (b), we wish to use the vector notation to represent adding up all the *x* components to get a total *x* component, etc. The plus sign is used between two vectors to indicate this type of component-by-component addition. Of course, vectors are really triplets of numbers, not numbers, so this is not the same as the use of the plus sign with individual numbers. But since we don't want to have to invent new words and symbols for this operation on vectors, we use the same old plus sign, and the same old addition-related words like "add," "sum," and "total." Combining vectors this way is called *vector addition*.

Similarly, the minus sign in example (a) was used to indicate negating each of the vector's three components individually. The equals sign is used to mean that all three components of the vector on the left side of an equation are the same as the corresponding components on the right.

Example (c) shows how we abuse the division symbol in a similar manner. When we write the vector $\Delta\boldsymbol{v}$ divided by the scalar Δt, we mean the new vector formed by dividing each one of the velocity components by Δt.

It's not hard to imagine a variety of operations that would combine vectors with vectors or vectors with scalars, but only four of them are required in order to express Newton's laws:

operation	**definition**
vector + **vector**	Add component by component to make a new set of three numbers.
vector - **vector**	Subtract component by component to make a new set of three numbers.
vector · scalar	Multiply each component of the vector by the scalar.
vector / scalar	Divide each component of the vector by the scalar.

As an example of an operation that is not useful for physics, there just aren't any useful physics applications for dividing a vector by another vector component by component. In optional section 7.5, we discuss in more detail the fundamental reasons why some vector operations are useful and others useless.

We can do algebra with vectors, or with a mixture of vectors and scalars in the same equation. Basically all the normal rules of algebra apply, but if you're not sure if a certain step is valid, you should simply translate it into three component-based equations and see if it works.

Example

Question: If we are adding two force vectors, $\boldsymbol{F}+\boldsymbol{G}$, is it valid to assume as in ordinary algebra that $\boldsymbol{F}+\boldsymbol{G}$ is the same as $\boldsymbol{G}+\boldsymbol{F}$?

Answer: To tell if this algebra rule also applies to vectors, we simply translate the vector notation into ordinary algebra notation. In terms of ordinary numbers, the components of the vector $\boldsymbol{F}+\boldsymbol{G}$ would be F_x+G_x, F_y+G_y, and F_z+G_z, which are certainly the same three numbers as G_x+F_x, G_y+F_y, and G_z+F_z. Yes, $\boldsymbol{F}+\boldsymbol{G}$ is the same as $\boldsymbol{G}+\boldsymbol{F}$.

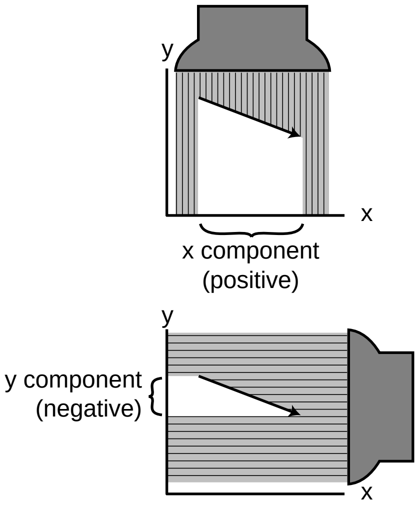
It is useful to define a symbol $\boldsymbol{r}$ for the vector whose components are x, y, and z, and a symbol $\Delta\boldsymbol{r}$ made out of Δx, Δy, and Δz.

Although this may all seem a little formidable, keep in mind that it amounts to nothing more than a way of abbreviating equations! Also, to keep things from getting too confusing the remainder of this chapter focuses mainly on the $\Delta\boldsymbol{r}$ vector, which is relatively easy to visualize.

Self-Check

Translate the equations $v_x=\Delta x/\Delta t$, $v_y=\Delta y/\Delta t$, and $v_z=\Delta z/\Delta t$ for motion with constant velocity into a single equation in vector notation.

$\boldsymbol{v}=\Delta\boldsymbol{r}/\Delta t$



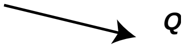
Drawing vectors as arrows

A vector in two dimensions can be easily visualized by drawing an arrow whose length represents its magnitude and whose direction represents its direction. The *x* component of a vector can then be visualized as the length of the shadow it would cast in a beam of light projected onto the *x* axis, and similarly for the y component. Shadows with arrowheads pointing back against the direction of the positive axis correspond to negative components.

In this type of diagram, the negative of a vector is the vector with the same magnitude but in the opposite direction. Multiplying a vector by a scalar is represented by lengthening the arrow by that factor, and similarly for division.

Self-Check

Given vector ***Q*** represented by an arrow below, draw arrows representing the vectors 1.5***Q*** and -***Q***.



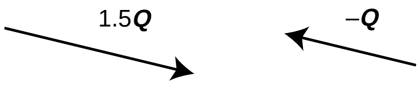
Discussion Questions

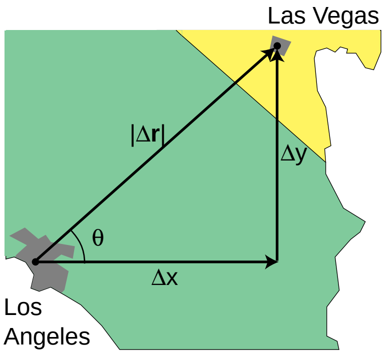
A. Would it make sense to define a zero vector?

B. An object goes from one point in space to another. After it arrives at its destination, how does the magnitude of its Δ***r*** vector compare with the distance it traveled?

7.2 Calculations with Magnitude and Direction

If you ask someone where Las Vegas is compared to Los Angeles, they are unlikely to say that the Δ*x* is 290 km and the Δ*y* is 230 km, in a coordinate system where the positive *x* axis is east and the *y* axis points north. They will probably say instead that it's 370 km to the northeast. If they were being precise, they might specify the direction as 38° counterclockwise from east. In two dimensions, we can always specify a vector's direction like this, using a single angle. A magnitude plus an angle suffice to specify everything about the vector. The following two examples show how we use trigonometry and the Pythagorean theorem to go back and forth between the x-y and magnitude-angle descriptions of vectors.





Example: finding the magnitude and angle from the components

Question: Given that the Δ**r** vector from LA to Las Vegas has Δ*x*=290 km and Δ*y*=230 km, how would we find the magnitude and direction of Δ**r**?

Solution: We find the magnitude of Δ**r** from the Pythagorean theorem:

$$|\Delta\mathbf{r}| = \sqrt{\Delta x^2 + \Delta y^2}$$
$$= 370 \text{ km}$$

We know all three sides of the triangle, so the angle θ can be found using any of the inverse trig functions. For example, we know the opposite and adjacent sides, so

$$\theta = \tan^{-1}\frac{\Delta y}{\Delta x}$$
$$= 38° \quad .$$

Example: finding the components from the magnitude and angle

Question: Given that the straight-line distance from Los Angeles to Las Vegas is 370 km, and that the angle θ in the figure is 38°, how can the *x* and *y* components of the Δ**r** vector be found?

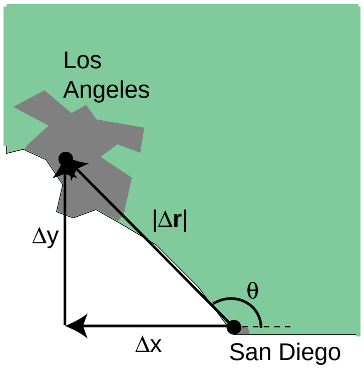
Solution: The sine and cosine of θ relate the given information to the information we wish to find:

$$\cos\theta = \frac{\Delta x}{|\Delta \boldsymbol{r}|}$$

$$\sin\theta = \frac{\Delta y}{|\Delta \boldsymbol{r}|}$$

Solving for the unknowns gives

$$\Delta x = |\Delta \boldsymbol{r}|\cos\theta$$
$$= 290 \text{ km}$$
$$\Delta y = |\Delta \boldsymbol{r}|\sin\theta$$
$$= 230 \text{ km}$$



The following example shows the correct handling of the plus and minus signs, which is usually the main cause of mistakes by students.

Example: negative components

Question: San Diego is 120 km east and 150 km south of Los Angeles. An airplane pilot is setting course from San Diego to Los Angeles. At what angle should she set her course, measured counterclockwise from east, as shown in the figure?

Solution: If we make the traditional choice of coordinate axes, with x pointing to the right and y pointing up on the map, then her Δx is negative, because her final x value is less than her initial x value. Her Δy is positive, so we have

$$\begin{aligned}\Delta x &= -120\ \text{km}\\ \Delta y &= 150\ \text{km}\quad .\end{aligned}$$

If we work by analogy with the previous example, we get

$$\begin{aligned}\theta &= \tan^{-1}\frac{\Delta y}{\Delta x}\\ &= \tan^{-1}\left(-1.25\right)\\ &= -51^{\circ}\quad .\end{aligned}$$

According to the usual way of defining angles in trigonometry, a negative result means an angle that lies clockwise from the x axis, which would have her heading for the Baja California. What went wrong? The answer is that when you ask your calculator to take the arctangent of a number, there are always two valid possibilities differing by 180°. That is, there are two possible angles whose tangents equal -1.25:

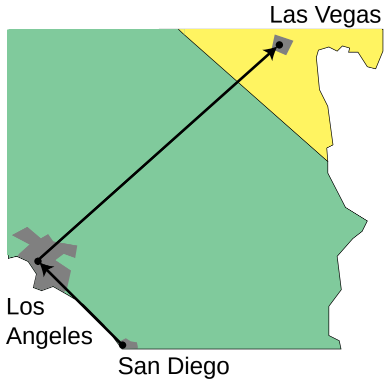
$$\begin{aligned}&\tan 129^{\circ} = -1.25\\ &\tan -51^{\circ} = -1.25\end{aligned}$$

You calculator doesn't know which is the correct one, so it just picks one. In this case, the one it picked was the wrong one, and it was up to you to add 180° to it to find the right answer.

Discussion Question

In the example above, we dealt with *components* that were negative. Does it make sense to talk about positive and negative *vectors*?

7.3 Techniques for Adding Vectors



Addition of vectors given their components

The easiest type of vector addition is when you are in possession of the components, and want to find the components of their sum.

Example

Question: Given the Δx and Δy values from the previous examples, find the Δx and Δy from San Diego to Las Vegas.

Solution:

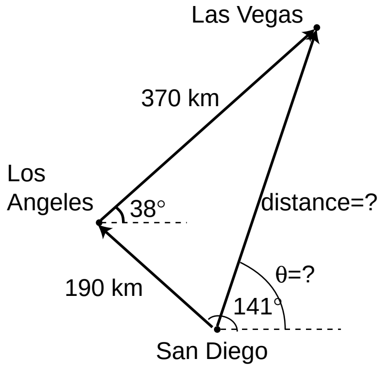
$$\begin{aligned}\Delta x_{\text{total}} &= \Delta x_1 + \Delta x_2\\ &= \text{-120 km} + \text{290 km}\\ &= \text{170 km}\\ \Delta y_{\text{total}} &= \Delta y_1 + \Delta y_2\\ &= \text{150 km} + \text{230 km}\\ &= 380\end{aligned}$$

Addition of vectors given their magnitudes and directions

In this case, you must first translate the magnitudes and directions into components, and the add the components.

Graphical addition of vectors

Often the easiest way to add vectors is by making a scale drawing on a piece of paper. This is known as graphical addition, as opposed to the analytic techniques discussed previously.

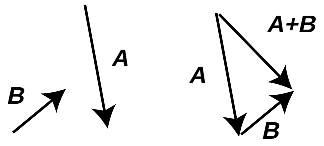


Example

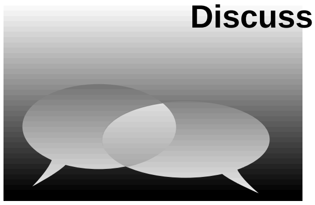
Question: Given the magnitudes and angles of the $\Delta \boldsymbol{r}$ vectors from San Diego to Los Angeles and from Los Angeles to Las Vegas, find the magnitude and angle of the $\Delta \boldsymbol{r}$ vector from San Diego to Las Vegas.

Solution: Using a protractor and a ruler, we make a careful scale drawing, as shown in the figure. A scale of 1 cm→10 km was chosen for this solution. With a ruler, we measure the distance from San Diego to Las Vegas to be 3.8 cm, which corresponds to 380 km. With a protractor, we measure the angle θ to be 71°.

Even when we don't intend to do an actual graphical calculation with a ruler and protractor, it can be convenient to diagram the addition of vectors in this way. With $\Delta \boldsymbol{r}$ vectors, it intuitively makes sense to lay the vectors tip-to-tail and draw the sum vector from the tail of the first vector to the tip of the second vector. We can do the same when adding other vectors such as force vectors.



Vectors can be added graphically by placing them tip to tail, and then drawing a vector from the tail of the first vector to the tip of the second vector.

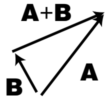


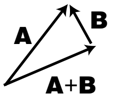
Discussion Questions

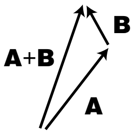
A. If you're doing *graphical* addition of vectors, does it matter which vector you start with and which vector you start from the other vector's tip?

B. If you add a vector with magnitude 1 to a vector of magnitude 2, what magnitudes are possible for the vector sum?

C. Which of these examples of vector addition are correct, and which are incorrect?







7.4* Unit Vector Notation

When we want to specify a vector by its components, it can be cumbersome to have to write the algebra symbol for each component:

$$\Delta x = 290\text{ km},\ \Delta y = 230\text{ km}$$

A more compact notation is to write

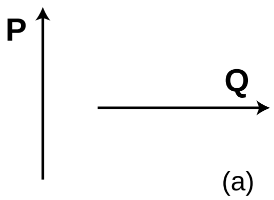
$$\Delta \boldsymbol{r} = (290\text{ km})\hat{\boldsymbol{x}} + (230\text{ km})\hat{\boldsymbol{y}} \quad ,$$

where the vectors $\hat{\boldsymbol{x}}$, $\hat{\boldsymbol{y}}$, and $\hat{\boldsymbol{z}}$, called the unit vectors, are defined as the vectors that have magnitude equal to 1 and directions lying along the *x*, *y*, and *z* axes. In speech, they are referred to as “x-hat” and so on.

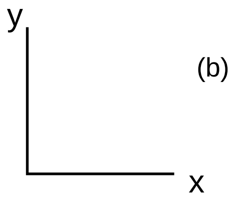
A slightly different, and harder to remember, version of this notation is unfortunately more prevalent. In this version, the unit vectors are called $\hat{\boldsymbol{i}}$, $\hat{\boldsymbol{j}}$, and $\hat{\boldsymbol{k}}$:

$$\Delta r = (290\text{ km})\hat{\boldsymbol{i}} + (230\text{ km})\hat{\boldsymbol{j}} \quad .$$

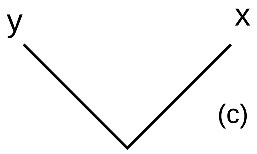
7.5* Rotational Invariance



(a)



(b)



(c)

Let’s take a closer look at why certain vector operations are useful and others are not. Consider the operation of multiplying two vectors component by component to produce a third vector:

$$\begin{aligned} R_x &= P_x Q_x \\ R_y &= P_y Q_y \\ R_z &= P_z Q_z \end{aligned}$$

As a simple example, we choose vectors **P** and **Q** to have length 1, and make them perpendicular to each other, as shown in figure (a). If we compute the result of our new vector operation using the coordinate system shown in (b), we find:

$$\begin{aligned} R_x &= 0 \\ R_y &= 0 \\ R_z &= 0 \end{aligned}$$

The *x* component is zero because P_x=0, the *y* component is zero because Q_y=0, and the *z* component is of course zero because both vectors are in the *x*-*y* plane. However, if we carry out the same operations in coordinate system (c), rotated 45 degrees with respect to the previous one, we find

$$\begin{aligned} R_x &= -1/2 \\ R_y &= 1/2 \\ R_z &= 0 \end{aligned}$$

The operation’s result depends on what coordinate system we use, and since the two versions of ***R*** have different lengths (one being zero and the other nonzero), they don’t just represent the same answer expressed in two different coordinate systems. Such an operation will never be useful in physics, because experiments show physics works the same regardless of which way we orient the laboratory building! The *useful* vector operations, such as addition and scalar multiplication, are rotationally invariant, i.e. come out the same regardless of the orientation of the coordinate system.

Summary

Selected Vocabulary

vector .... a quantity that has both an amount (magnitude) and a direction in space

magnitude .... the "amount" associated with a vector

scalar .... a quantity that has no direction in space, only an amount

Notation

$\boldsymbol{A}$ .... vector with components A_x, A_y, and A_z

$\vec{A}$ .... handwritten notation for a vector

$|A|$ .... the magnitude of vector $\boldsymbol{A}$

$\boldsymbol{r}$ .... the vector whose components are x, y, and z

$\Delta\boldsymbol{r}$..................................... the vector whose components are Δx, Δy, and Δz

$\hat{x}, \hat{y}, \hat{z}$ .... (optional topic) unit vectors; the vectors with magnitude 1 lying along the x, y, and z axes

$\hat{\imath}, \hat{\jmath}, \hat{k}$ .... a harder to remember notation for the unit vectors

Standard Terminology Avoided in This Book

displacement vector .... a name for the symbol $\Delta\boldsymbol{r}$

speed .... the magnitude of the velocity vector, i.e. the velocity stripped of any information about its direction

Summary

A vector is a quantity that has both a magnitude (amount) and a direction in space, as opposed to a scalar, which has no direction. The vector notation amounts simply to an abbreviation for writing the vector's three components.

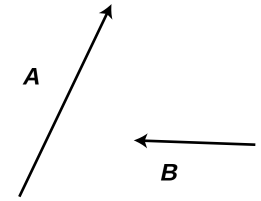
In two dimensions, a vector can be represented either by its two components or by its magnitude and direction. The two ways of describing a vector can be related by trigonometry.

The two main operations on vectors are addition of a vector to a vector, and multiplication of a vector by a scalar.

Vector addition means adding the components of two vectors to form the components of a new vector. In graphical terms, this corresponds to drawing the vectors as two arrows laid tip-to-tail and drawing the sum vector from the tail of the first vector to the tip of the second one. Vector subtraction is performed by negating the vector to be subtracted and then adding.

Multiplying a vector by a scalar means multiplying each of its components by the scalar to create a new vector. Division by a scalar is defined similarly.

Homework Problems



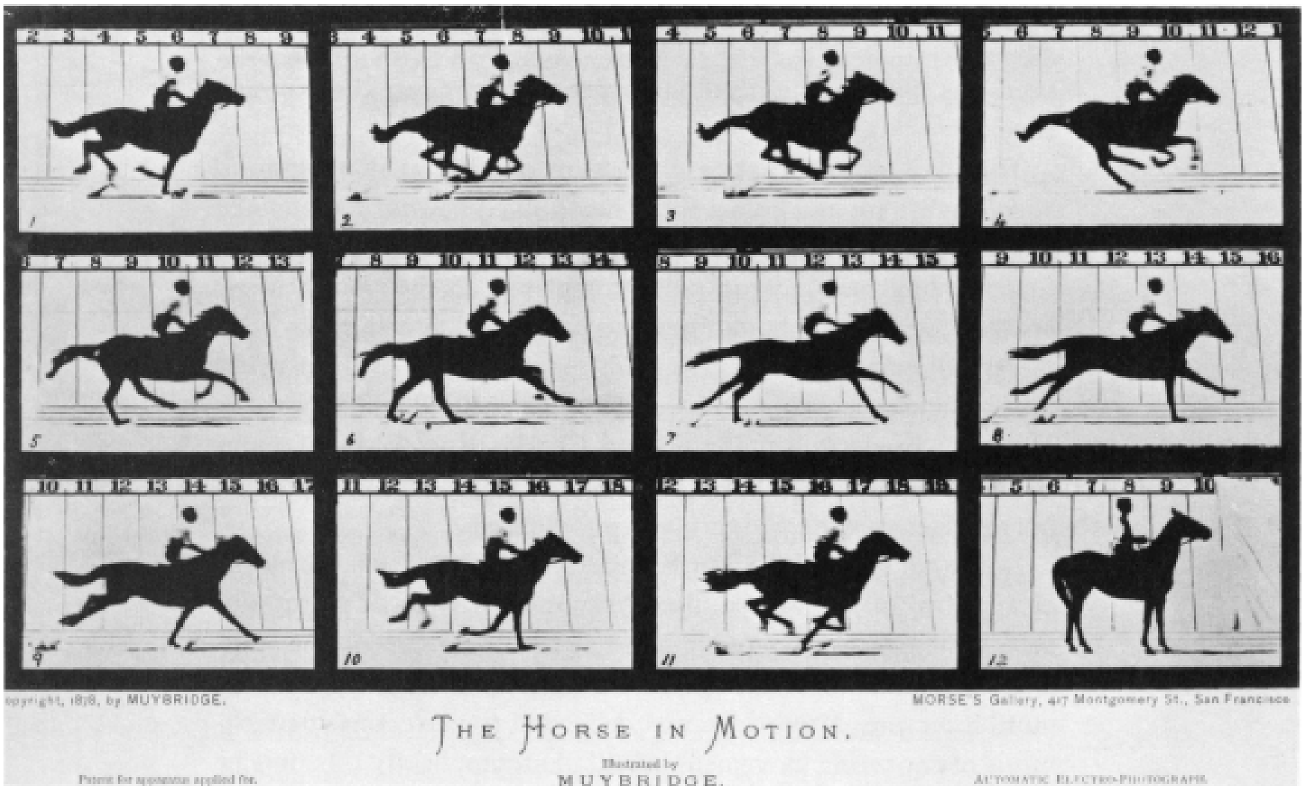
Problem 1.

1. Here are two vectors. Graphically calculate ***A***+***B***, ***A***-***B***, ***B***-***A***, -2***B***, and ***A***-2***B***. No numbers are involved.

2. Phnom Penh is 470 km east and 250 km south of Bangkok. Hanoi is 60 km east and 1030 km north of Phnom Penh. (a) Choose a coordinate system, and translate these data into Δ*x* and Δ*y* values with the proper plus and minus signs. (b✓) Find the components of the Δ***r*** vector pointing from Bangkok to Hanoi.

3 ✓. If you walk 35 km at an angle 25° counterclockwise from east, and then 22 km at 230° counterclockwise from east, find the distance and direction from your starting point to your destination.

S	A solution is given in the back of the book.	★	A difficult problem.
✓	A computerized answer check is available.	∫	A problem that requires calculus.


opyright, 1878, by MUYBRIDGE.
MORSE'S Gallery, 417 Montgomery St., San Francisco
THE HORSE IN MOTION.
Illustrated by
MUYBRIDGE.
Patent for apparatus applied for.
AUTOMATIC ELECTRO-PHOTOGRAPH.

8 Vectors and Motion

In 1872, capitalist and former California governor Leland Stanford asked photographer Eadweard Muybridge if he would work for him on a project to settle a $25,000 bet (a princely sum at that time). Stanford's friends were convinced that a galloping horse always had at least one foot on the ground, but Stanford claimed that there was a moment during each cycle of the motion when all four feet were in the air. The human eye was simply not fast enough to settle the question. In 1878, Muybridge finally succeeded in producing what amounted to a motion picture of the horse, showing conclusively that all four feet did leave the ground at one point. (Muybridge was a colorful figure in San Francisco history, and his acquittal for the murder of his wife's lover was considered the trial of the century in California.)

The losers of the bet had probably been influenced by Aristotelian reasoning, for instance the expectation that a leaping horse would lose horizontal velocity while in the air with no force to push it forward, so that it would be more efficient for the horse to run without leaping. But even for students who have converted wholeheartedly to Newtonianism, the relationship between force and acceleration leads to some conceptual difficulties, the main one being a problem with the true but seemingly absurd statement that an object can have an acceleration vector whose direction is not the same as the direction of motion. The horse, for instance, has nearly constant horizontal velocity, so its a_x is zero. But as anyone can tell you who has ridden a galloping horse, the horse accelerates up and down. The horse's acceleration vector therefore changes back and forth between the up and down directions, but is never in the same direction as the horse's motion. In this chapter, we will examine more carefully the properties of the velocity, acceleration, and force vectors. No new principles are introduced, but an attempt is made to tie things together and show examples of the power of the vector formulation of Newton's laws.

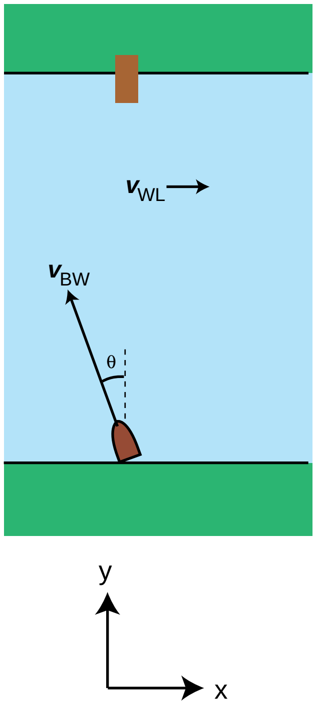
8.1 The Velocity Vector

For motion with constant velocity, the velocity vector is

$$\boldsymbol{v}=\Delta\boldsymbol{r}/\Delta t \qquad \text{[only for constant velocity]} \quad .$$

The $\Delta\boldsymbol{r}$ vector points in the direction of the motion, and dividing it by the scalar Δt only changes its length, not its direction, so the velocity vector points in the same direction as the motion. When the velocity is not constant, i.e. when the x-t, y-t, and z-t graphs are not all linear, we use the slope-of-the-tangent-line approach to define the components v_x, v_y, and v_z, from which we assemble the velocity vector. Even when the velocity vector is not constant, it still points along the direction of motion.

Vector addition is the correct way to generalize the one-dimensional concept of adding velocities in relative motion, as shown in the following example:



Example: velocity vectors in relative motion

Question: You wish to cross a river and arrive at a dock that is directly across from you, but the river's current will tend to carry you downstream. To compensate, you must steer the boat at an angle. Find the angle θ, given the magnitude, $|\boldsymbol{v}_{WL}|$, of the water's velocity relative to the land, and the maximum speed, $|\boldsymbol{v}_{BW}|$, of which the boat is capable relative to the water.

Solution: The boat's velocity relative to the land equals the vector sum of its velocity with respect to the water and the water's velocity with respect to the land,

$$\boldsymbol{v}_{BL} = \boldsymbol{v}_{BW} + \boldsymbol{v}_{WL} \quad .$$

If the boat is to travel straight across the river, i.e. along the y axis, then we need to have $v_{BL,x}$=0. This x component equals the sum of the x components of the other two vectors,

$$v_{BL,x} = v_{BW,x} + v_{WL,x} \quad ,$$

or

$$0 = -|\boldsymbol{v}_{BW}| \sin\theta + |\boldsymbol{v}_{WL}| \quad .$$

Solving for θ, we find

$$\sin\theta = |\boldsymbol{v}_{WL}|/|\boldsymbol{v}_{BW}| \quad ,$$

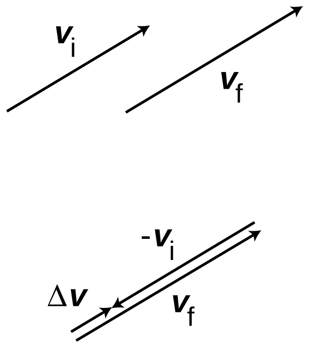
$$\theta = \sin^{-1}\frac{|\boldsymbol{v}_{WL}|}{|\boldsymbol{v}_{BW}|} \quad .$$

Discussion Questions

A. Is it possible for an airplane to maintain a constant velocity vector but not a constant $|\mathbf{v}|$? How about the opposite -- a constant $|\mathbf{v}|$ but not a constant velocity vector? Explain.

B. New York and Rome are at about the same latitude, so the earth's rotation carries them both around nearly the same circle. Do the two cities have the same velocity vector (relative to the center of the earth)? If not, is there any way for two cities to have the same velocity vector?

8.2 The Acceleration Vector



(a) A change in the magnitude of the velocity vector implies an acceleration.

When all three acceleration components are constant, i.e. when the v_x-t, v_y-t, and v_z-t graphs are all linear, we can define the acceleration vector as

$$\mathbf{a}=\Delta\mathbf{v}/\Delta t \qquad \text{[only for constant acceleration]} \quad ,$$

which can be written in terms of initial and final velocities as

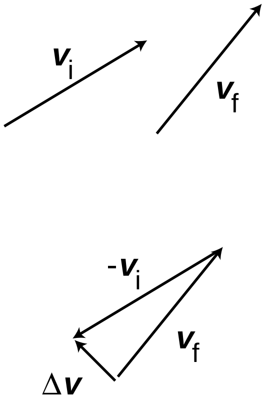
$$\mathbf{a}=(\mathbf{v}_f-\mathbf{v}_i)/\Delta t \qquad \text{[only for constant acceleration]} \quad .$$

If the acceleration is not constant, we define it as the vector made out of the a_x, a_y, and a_z components found by applying the slope-of-the-tangent-line technique to the v_x-t, v_y-t, and v_z-t graphs.

Now there are two ways in which we could have a nonzero acceleration. Either the magnitude or the direction of the velocity vector could change. This can be visualized with arrow diagrams as shown in the figure. Both the magnitude and direction can change simultaneously, as when a car accelerates while turning. Only when the magnitude of the velocity changes while its direction stays constant do we have a $\Delta\mathbf{v}$ vector and an acceleration vector in the same direction as the motion.

Self Check

(1) In figure (a), is the object speeding up or slowing down? (2) What would the diagram look like if $\mathbf{v}_i$ was the same as $\mathbf{v}_f$? (3) Describe how the $\Delta\mathbf{v}$ vector is different depending on whether an object is speeding up or slowing down.



(b) A change in the direction of the velocity vector also produces a nonzero $\Delta\mathbf{v}$ vector, and thus a nonzero acceleration vector, $\Delta\mathbf{v}/\Delta t$.

If this all seems a little strange and abstract to you, you're not alone. It doesn't mean much to most physics students the first time someone tells them that acceleration is a vector, and that the acceleration vector does not have to be in the same direction as the velocity vector. One way to understand those statements better is to imagine an object such as an air freshener or a pair of fuzzy dice hanging from the rear-view mirror of a car. Such a hanging object, called a bob, constitutes an accelerometer. If you watch the bob as you accelerate from a stop light, you'll see it swing backward. The horizontal direction in which the bob tilts is opposite to the direction of the acceleration. If you apply the brakes and the car's acceleration vector points backward, the bob tilts forward.

After accelerating and slowing down a few times, you think you've put your accelerometer through its paces, but then you make a right turn. Surprise! Acceleration is a vector, and needn't point in the same direction as the velocity vector. As you make a right turn, the bob swings outward, to your left. That means the car's acceleration vector is to your right, perpen-

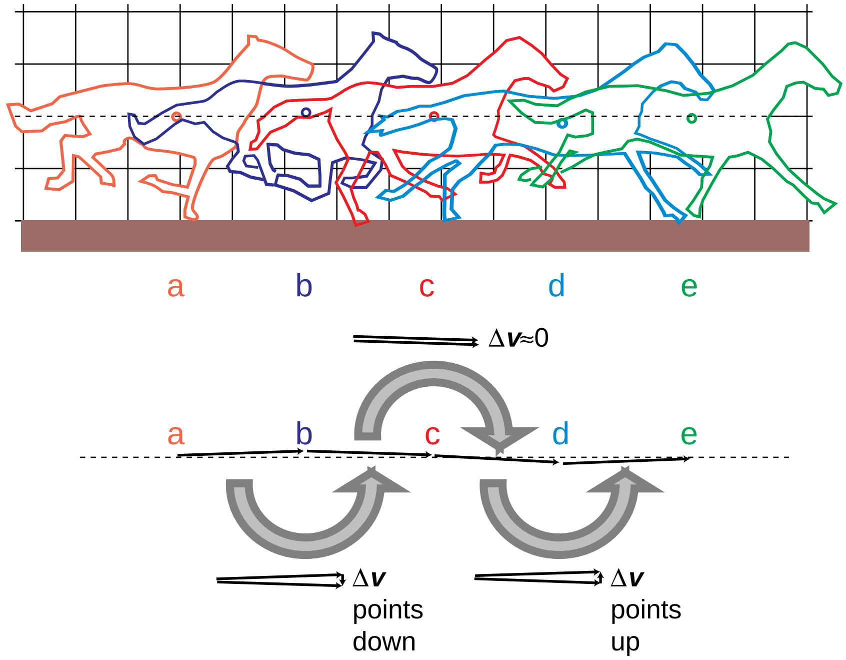
(1) It is speeding up, because the final velocity vector has the greater magnitude. (2) The result would be zero, which would make sense. (3) Speeding up produced a $\Delta\mathbf{v}$ vector in the same direction as the motion. Slowing down would have given a $\Delta\mathbf{v}$ that bointed backward.

dicular to your velocity vector. A useful definition of an acceleration vector should relate in a systematic way to the actual physical effects produced by the acceleration, so a physically reasonable definition of the acceleration vector must allow for cases where it is not in the same direction as the motion.

Self-Check

In projectile motion, what direction does the acceleration vector have?

The following are two examples of force, velocity, and acceleration vectors in complex motion.



The downward force of gravity produces a downward acceleration vector.

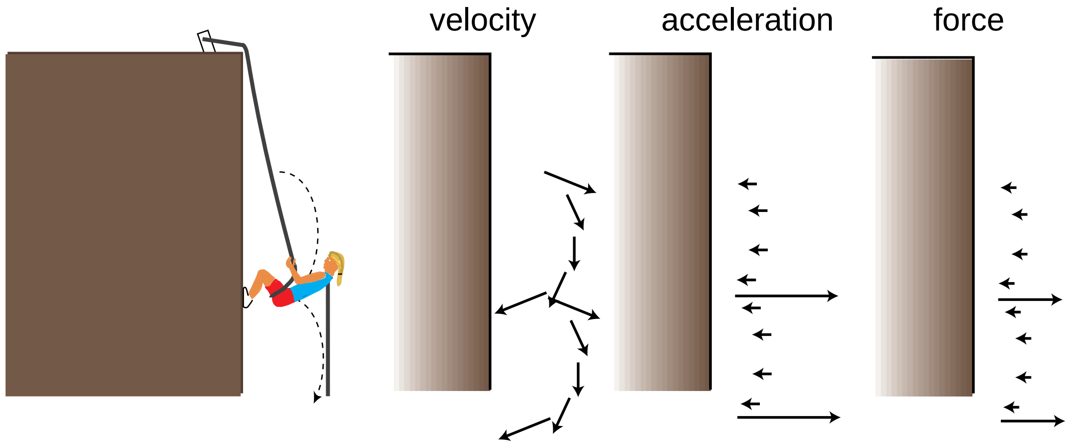
The upward force from the ground is greater than the downward force of gravity. The total force on the horse is upward, giving an upward acceleration.

This figure shows outlines traced from the first, third, fifth, seventh, and ninth frames in Muybridge's series of photographs of the galloping horse. The estimated location of the horse's center of mass is shown with a circle, which bobs above and below the horizontal dashed line.

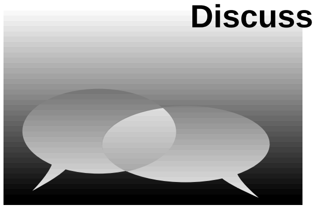
If we don't care about calculating velocities and accelerations in any particular system of units, then we can pretend that the time between frames is one unit. The horse's velocity vector as it moves from one point to the next can then be found simply by drawing an arrow to connect one position of the center of mass to the next. This produces a series of velocity vectors which alternate between pointing above and below horizontal.

The Δ***v*** vector is the vector which we would have to add onto one velocity vector in order to get the next velocity vector in the series. The Δ***v*** vector alternates between pointing down (around the time when the horse is in the air, b) and up (around the time when the horse has two feet on the ground, d).

As we have already seen, the projectile has a_x=0 and a_y=-g, so the acceleration vector is pointing straight down.



In this example, the rappeller’s velocity has long periods of gradual change interspersed with short periods of rapid change. These correspond to periods of small acceleration and force and periods of large acceleration and force.



Discussion Questions

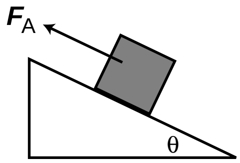
A. When a car accelerates, why does a bob hanging from the rearview mirror swing toward the back of the car? Is it because a force throws it backward? If so, what force? Similarly, describe what happens in the other cases described above.

B. The following is a question commonly asked by students:

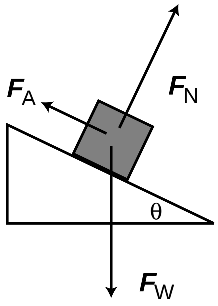
> “Why does the force vector always have to point in the same direction as the acceleration vector? What if you suddenly decide to change your force on an object, so that your force is no longer pointing the same direction that the object is accelerating?”

What misunderstanding is demonstrated by this question? Suppose, for example, a spacecraft is blasting its rear main engines while moving forward, then suddenly begins firing its sideways maneuvering rocket as well. What does the student think Newton’s laws are predicting?

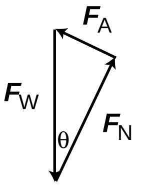
8.3 The Force Vector and Simple Machines



(a) The applied force $\boldsymbol{F}_A$ pushes the block up the frictionless ramp.



(b) Three forces act on the block. Their vector sum is zero.



(c) If the block is to move at constant velocity, Newton's first law says that the three force vectors acting on it must add up to zero. To perform vector addition, we put the vectors tip to tail, and in this case we are adding three vectors, so each one's tail goes against the tip of the previous one. Since they are supposed to add up to zero, the third vector's tip must come back to touch the tail of the first vector. They form a triangle, and since the applied force is perpendicular to the normal force, it is a right triangle.

Force is relatively easy to intuit as a vector. The force vector points in the direction in which it is trying to accelerate the object it is acting on.

Since force vectors are so much easier to visualize than acceleration vectors, it is often helpful to first find the direction of the (total) force vector acting on an object, and then use that information to determine the direction of the acceleration vector. Newton's second law, $\boldsymbol{F}_{\text{total}}=m\boldsymbol{a}$, tells us that the two must be in the same direction.

An important application of force vectors is to analyze the forces acting in two-dimensional mechanical systems, as in the following example.

Example: pushing a block up a ramp

Question: Figure (a) shows a block being pushed up a frictionless ramp at constant speed by an applied force $\boldsymbol{F}_A$. How much force is required, in terms of the block's mass, m, and the angle of the ramp, θ?

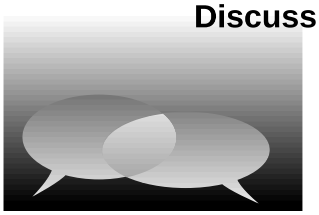
Solution: Figure (b) shows the other two forces acting on the block: a normal force, $\boldsymbol{F}_N$, created by the ramp, and the weight force, $\boldsymbol{F}_W$, created by the earth's gravity. Because the block is being pushed up at constant speed, it has zero acceleration, and the total force on it must be zero. From figure (c), we find

$$\begin{aligned} |\boldsymbol{F}_A| &= |\boldsymbol{F}_W| \sin\theta \\ &= mg \sin\theta \quad . \end{aligned}$$

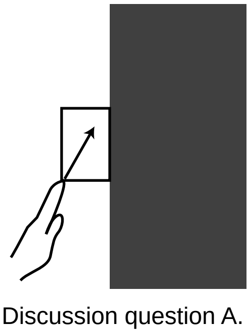
Since the sine is always less than one, the applied force is always less than mg, i.e. pushing the block up the ramp is easier than lifting it straight up. This is presumably the principle on which the pyramids were constructed: the ancient Egyptians would have had a hard time applying the forces of enough slaves to equal the full weight of the huge blocks of stone.

Essentially the same analysis applies to several other simple machines, such as the wedge and the screw.

Discussion Questions



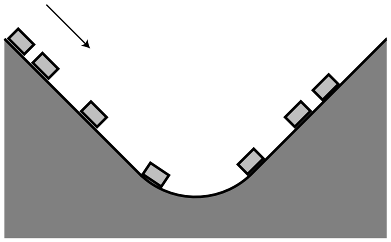
A. The figure shows a block being pressed diagonally upward against a wall, causing it to slide up the wall. Analyze the forces involved, including their directions.



Discussion question A.

B. As discussed in chapter 5, the maximum force of static friction between two surfaces is given by $F_s<\mu_s F_N$, where μ_s is a constant that depends on the types of surfaces, and F_N is the normal force, i.e. the force perpendicular to the surfaces that keeps them from passing through each other. If a block of mass m is resting on an inclined plane, what is the maximum angle of the incline for which static friction can prevent the block from sliding?

C. The figure shows a roller coaster car rolling down and then up under the influence of gravity. Sketch the car's velocity vectors and acceleration vectors. Pick an interesting point in the motion and sketch a set of force vectors acting on the car whose vector sum could have resulted in the right acceleration vector.



Discussion question C.

8.4 ∫ Calculus With Vectors

The definitions of the velocity and acceleration components given in chapter 6 can be translated into calculus notation as

$$\boldsymbol{v} = \frac{dx}{dt}\hat{\boldsymbol{x}} + \frac{dy}{dt}\hat{\boldsymbol{y}} + \frac{dz}{dt}\hat{\boldsymbol{z}}$$

and

$$\boldsymbol{a} = \frac{dv_x}{dt}\hat{\boldsymbol{x}} + \frac{dv_y}{dt}\hat{\boldsymbol{y}} + \frac{dv_z}{dt}\hat{\boldsymbol{z}} \quad .$$

To make the notation less cumbersome, we generalize the concept of the derivative to include derivatives of vectors, so that we can abbreviate the above equations as

$$\boldsymbol{v} = \frac{\mathrm{d}\boldsymbol{r}}{\mathrm{d}t}$$

and

$$\boldsymbol{a} = \frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t} \quad .$$

In words, to take the derivative of a vector, you take the derivatives of its components and make a new vector out of those. This definition means that the derivative of a vector function has the familiar properties

$$\frac{d(cf)}{dt} = c\frac{df}{dt} \qquad [c \text{ is a constant}]$$

and

$$\frac{d(f+g)}{dt} = \frac{df}{dt} + \frac{dg}{dt} \quad . \qquad [c \text{ is a constant}]$$

The integral of a vector is likewise defined as integrating component by component.

Example

Question: Two objects have positions as functions of time given by the equations

$$\mathbf{r}_1 = 3t^2\hat{\mathbf{x}} + t\hat{\mathbf{y}}$$

and

$$\mathbf{r}_2 = 3t^4\hat{\mathbf{x}} + t\hat{\mathbf{y}} \quad .$$

Find both objects' accelerations using calculus. Could either answer have been found without calculus?

Solution: Taking the first derivative of each component, we find

$$\mathbf{v}_1 = 6t\hat{\mathbf{x}} + \hat{\mathbf{y}}$$

$$\mathbf{v}_2 = 12t^3\hat{\mathbf{x}} + \hat{\mathbf{y}} \quad ,$$

and taking the derivatives again gives acceleration,

$$\mathbf{a}_1 = 6\hat{\mathbf{x}}$$

$$\mathbf{a}_2 = 36t^2\hat{\mathbf{x}} \quad .$$

The first object's acceleration could have been found without calculus, simply by comparing the *x* and *y* coordinates with the constant-acceleration equation $\Delta x = v_o\Delta t + \frac{1}{2}a\Delta t^2$. The second equation, however, isn't just a second-order polynomial in *t*, so we really did need calculus to find the corresponding acceleration.

Example: a fire-extinguisher stunt on ice

Question: Prof. Puerile smuggles a fire extinguisher into a skating rink. Climbing out onto the ice without any skates on, he sits down and pushes off from the wall with his feet, acquiring an initial velocity $v_o\hat{\mathbf{y}}$. At *t*=0, he then discharges the fire extinguisher at a 45-degree angle so that it applies a force to him that is backward and to the left, i.e. along the negative y axis and the positive x axis. The fire extinguisher's force is strong at first, but then dies down according to the equation $|\mathbf{F}|=b-ct$, where *b* and *c* are constants. Find the professor's velocity as a function of time.

Solution: Measured counterclockwise from the x axis, the angle of the force vector becomes 315°. Breaking the force down into x and y components, we have

$$\begin{aligned}
F_x &= |\mathbf{F}| \cos 315° \\
&= \tfrac{1}{\sqrt{2}}\ (b-ct) \\
F_y &= |\mathbf{F}| \sin 315° \\
&= \tfrac{1}{\sqrt{2}}\ (-b+ct) \quad .
\end{aligned}$$

In unit vector notation, this is

$$\mathbf{F} = \tfrac{1}{\sqrt{2}}\ (b-ct)\hat{\mathbf{x}} + \tfrac{1}{\sqrt{2}}\ (-b+ct)\hat{\mathbf{y}} \quad .$$

Newton's second law gives

$$\begin{aligned}
\mathbf{a} &= \mathbf{F}/m \\
&= \frac{b-ct}{\sqrt{2}m}\,\hat{\mathbf{x}} + \frac{-b+ct}{\sqrt{2}m}\,\hat{\mathbf{y}} \quad .
\end{aligned}$$

To find the velocity vector as a function of time, we need to integrate the acceleration vector with respect to time,

$$\begin{aligned}\boldsymbol{v} &= \int \boldsymbol{a}\,\mathrm{d}t \\ &= \int \left(\frac{b-ct}{\sqrt{2}m}\hat{\boldsymbol{x}} + \frac{-b+ct}{\sqrt{2}m}\hat{\boldsymbol{y}}\right)\mathrm{d}t \\ &= \frac{1}{\sqrt{2}m}\int \left[\left(b-ct\right)\hat{\boldsymbol{x}} + \left(-b+ct\right)\hat{\boldsymbol{y}}\right]\mathrm{d}t\end{aligned}$$

A vector function can be integrated component by component, so this can be broken down into two integrals,

$$\begin{aligned}\boldsymbol{v} &= \frac{\hat{\boldsymbol{x}}}{\sqrt{2}m}\int \left(b-ct\right)\mathrm{d}t + \frac{\hat{\boldsymbol{y}}}{\sqrt{2}m}\int \left(-b+ct\right)\mathrm{d}t \\ &= \left(\frac{bt-\frac{1}{2}ct^2}{\sqrt{2}m} + \text{const. \#1}\right)\hat{\boldsymbol{x}} \\ &\qquad + \left(\frac{-bt+\frac{1}{2}ct^2}{\sqrt{2}m} + \text{const. \#2}\right)\hat{\boldsymbol{y}}\end{aligned}$$

Here the physical significance of the two constants of integration is that they give the initial velocity. Constant #1 is therefore zero, and constant #2 must equal v_o. The final result is

$$\boldsymbol{v} = \frac{bt-\frac{1}{2}ct^2}{\sqrt{2}m}\hat{\boldsymbol{x}} + \left(\frac{-bt+\frac{1}{2}ct^2}{\sqrt{2}m} + v_o\right)\hat{\boldsymbol{y}} \quad .$$

Summary

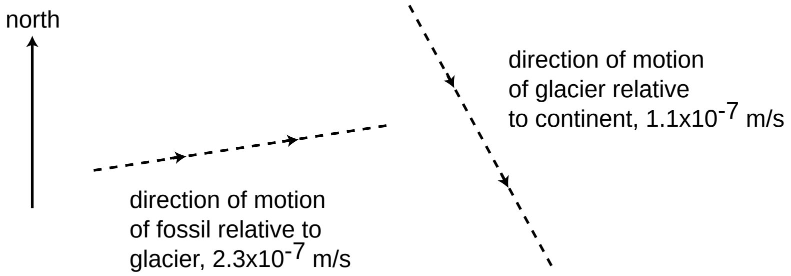
The velocity vector points in the direction of the object’s motion. Relative motion can be described by vector addition of velocities.

The acceleration vector need not point in the same direction as the object’s motion. We use the word “acceleration” to describe any change in an object’s velocity vector, which can be either a change in its magnitude or a change in its direction.

An important application of the vector addition of forces is the use of Newton’s first law to analyze mechanical systems.

Homework Problems

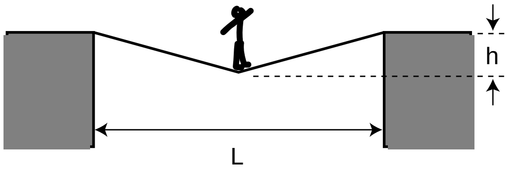
1 ✓. A dinosaur fossil is slowly moving down the slope of a glacier under the influence of wind, rain and gravity. At the same time, the glacier is moving relative to the continent underneath. The dashed lines represent the directions but not the magnitudes of the velocities. Pick a scale, and use graphical addition of vectors to find the magnitude and the direction of the fossil's velocity relative to the continent. You will need a ruler and protractor.

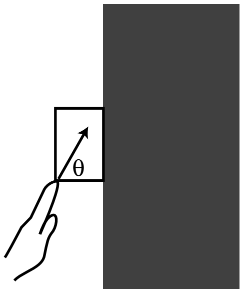

north
direction of motion of glacier relative to continent, 1.1x10^{-7} m/s
direction of motion of fossil relative to glacier, 2.3x10^{-7} m/s

2. Is it possible for a helicopter to have an acceleration due east and a velocity due west? If so, what would be going on? If not, why not?

3 ✓. A bird is initially flying horizontally east at 21.1 m/s, but one second later it has changed direction so that it is flying horizontally and 7° north of east, at the same speed. What are the magnitude and direction of its acceleration vector during that one second time interval? (Assume its acceleration was roughly constant.)

4. A person of mass M stands in the middle of a tightrope, which is fixed at the ends to two buildings separated by a horizontal distance L. The rope sags in the middle, stretching and lengthening the rope slightly. (a) If the tightrope walker wants the rope to sag vertically by no more than a height h, find the minimum tension, T, that the rope must be able to withstand without breaking, in terms of h, g, M, and L. (b) Based on your equation, explain why it is not possible to get h=0, and give a physical interpretation.





Problem 5.

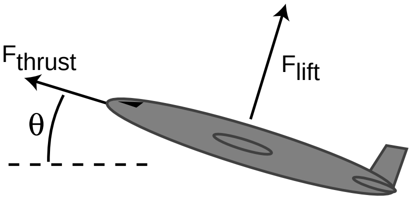
5★. Your hand presses a block of mass m against a wall with a force $\boldsymbol{F}_H$ acting at an angle θ. Find the minimum and maximum possible values of $|\boldsymbol{F}_H|$ that can keep the block stationary, in terms of m, g, θ, and μ_s, the coefficient of static friction between the block and the wall.

S A solution is given in the back of the book.	★ A difficult problem.
✓ A computerized answer check is available.	∫ A problem that requires calculus.

6. A skier of mass m is coasting down a slope inclined at an angle θ compared to horizontal. Assume for simplicity that the treatment of kinetic friction given in chapter 5 is appropriate here, although a soft and wet surface actually behaves a little differently. The coefficient of kinetic friction acting between the skis and the snow is μ_k, and in addition the skier experiences an air friction force of magnitude bv^2, where b is a constant. (a) Find the maximum speed that the skier will attain, in terms of the variables m, θ, μ_k, and b. (b) For angles below a certain minimum angle θ_{min}, the equation gives a result that is not mathematically meaningful. Find an equation for θ_{min}, and give a physical explanation of what is happening for $\theta<\theta_{min}$.

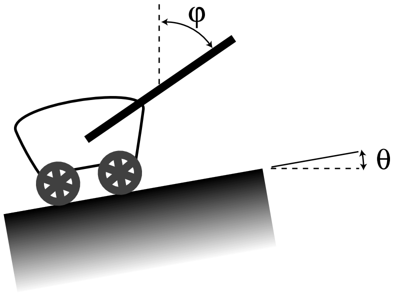
7 ∫. A gun is aimed horizontally to the west, and fired at t=0. The bullet's position vector as a function of time is $\boldsymbol{r} = b\hat{\boldsymbol{x}} + ct\hat{\boldsymbol{y}} + dt\hat{\boldsymbol{z}}$, where b, c, and d are constants. (a) What units would b, c, and d need to have for the equation to make sense? (b) Find the bullet's velocity and acceleration as functions of time. (c) Give physical interpretations of b, c, d, $\hat{\boldsymbol{x}}$, $\hat{\boldsymbol{y}}$, and $\hat{\boldsymbol{z}}$.

8 S. Annie Oakley, riding north on horseback at 30 mi/hr, shoots her rifle, aiming horizontally and to the northeast. The muzzle speed of the rifle is 140 mi/hr. When the bullet hits a defenseless fuzzy animal, what is its speed of impact? Neglect air resistance, and ignore the vertical motion of the bullet.



Problem 9.

9 S. A cargo plane has taken off from a tiny airstrip in the Andes, and is climbing at constant speed, at an angle of θ=17° with respect to horizontal. Its engines supply a thrust of F_{thrust}=200 kN, and the lift from its wings is F_{lift}=654 kN. Assume that air resistance (drag) is negligible, so the only forces acting are thrust, lift, and weight. What is its mass, in kg?



Problem 10.

10 S. A wagon is being pulled at constant speed up a slope θ by a rope that makes an angle φ with the vertical. (a) Assuming negligible friction, show that the tension in the rope is given by the equation

$$F_T = \frac{\sin\theta}{\sin\left(\theta+\varphi\right)} F_W \quad ,$$

where F_W is the weight force acting on the wagon. (b) Interpret this equation in the special cases of φ=0 and φ=180°-θ.

11 S. The angle of repose is the maximum slope on which an object will not slide. On airless, geologically inert bodies like the moon or an asteroid, the only thing that determines whether dust or rubble will stay on a slope is whether the slope is less steep than the angle of repose. (a) Find an equation for the angle of repose, deciding for yourself what are the relevant variables. (b) On an asteroid, where g can be thousands of times lower than on Earth, would rubble be able to lie at a steeper angle of repose?

9 Circular Motion

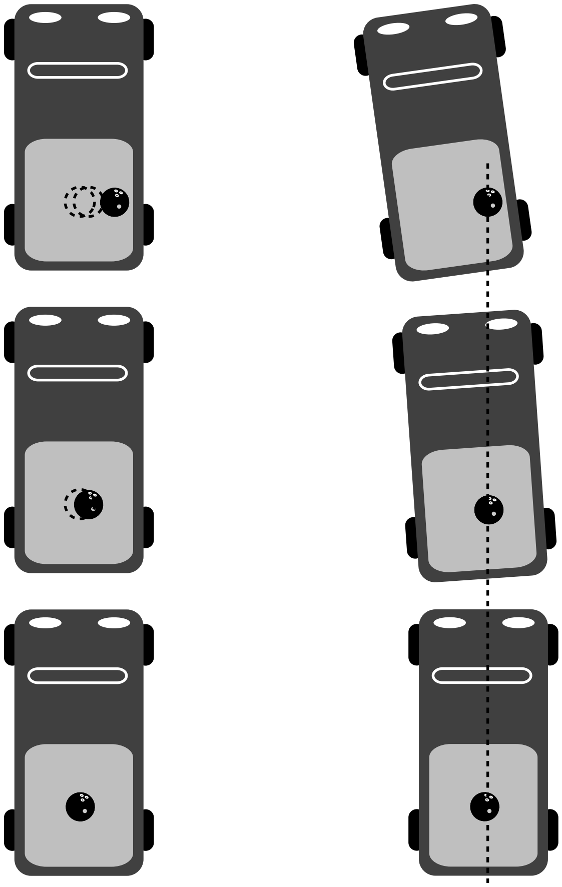
9.1 Conceptual Framework for Circular Motion

I now live fifteen minutes from Disneyland, so my friends and family in my native Northern California think it's a little strange that I've never visited the Magic Kingdom again since a childhood trip to the south. The truth is that for me as a preschooler, Disneyland was not the Happiest Place on Earth. My mother took me on a ride in which little cars shaped like rocket ships circled rapidly around a central pillar. I knew I was going to die. There was a force trying to throw me outward, and the safety features of the ride would surely have been inadequate if I hadn't screamed the whole time to make sure Mom would hold on to me. Afterward, she seemed surprisingly indifferent to the extreme danger we had experienced.

Circular motion does not produce an outward force

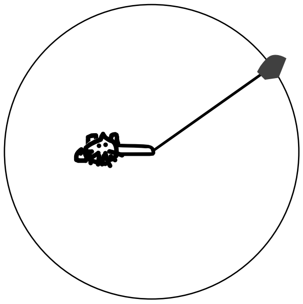
My younger self's understanding of circular motion was partly right and partly wrong. I was wrong in believing that there was a force pulling me outward, away from the center of the circle. The easiest way to understand this is to bring back the parable of the bowling ball in the pickup truck from chapter 4. As the truck makes a left turn, the driver looks in the rearview mirror and thinks that some mysterious force is pulling the ball outward, but the truck is accelerating, so the driver's frame of reference is not an inertial frame. Newton's laws are violated in a noninertial frame, so the ball appears to accelerate without any actual force acting on it. Because we are used to inertial frames, in which accelerations are caused by forces, the ball's acceleration creates a vivid illusion that there must be an outward force.

In an inertial frame everything makes more sense. The ball has no force on it, and goes straight as required by Newton's first law. The truck has a force on it from the asphalt, and responds to it by accelerating (changing the direction of its velocity vector) as Newton's second law says it should.

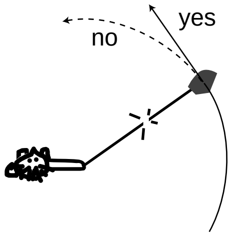


(a) In the turning truck's frame of reference, the ball appears to violate Newton's laws, displaying a sideways acceleration that is not the result of a force-interaction with any other object.

(b) In an inertial frame of reference, such as the frame fixed to the earth's surface, the ball obeys Newton's first law. No forces are acting on it, and it continues moving in a straight line. It is the truck that is participating in an interaction with the asphalt, the truck that accelerates as it should according to Newton's second law.



(a) An overhead view of a person swinging a rock on a rope. A force from the string is required to make the rock's velocity vector keep changing direction.



(b) If the string breaks, the rock will follow Newton's first law and go straight instead of continuing around the circle.

Circular motion does not persist without a force

I was correct about one thing, however. To make me curve around with the car, I really did need some force such as a force from my mother, friction from the seat, or a normal force from the side of the car. (In fact, all three forces were probably adding together.) One of the reasons why Galileo failed to refine the principle of inertia into a quantitative statement like Newton's first law is that he was not sure whether motion without a force would naturally be circular or linear. In fact, the most impressive examples he knew of the persistence of motion were mostly circular: the spinning of a top or the rotation of the earth, for example. Newton realized that in examples such as these, there really were forces at work. Atoms on the surface of the top are prevented from flying off straight by the ordinary force that keeps atoms stuck together in solid matter. The earth is nearly all liquid, but gravitational forces pull all its parts inward.

Uniform and nonuniform circular motion

Circular motion always involves a change in the direction of the velocity vector, but it is also possible for the magnitude of the velocity to change at the same time. Circular motion is referred to as *uniform* if $|\boldsymbol{v}|$ is constant, and *nonuniform* if it is changing.

Your speedometer tells you the magnitude of your car's velocity vector, so when you go around a curve while keeping your speedometer needle steady, you are executing uniform circular motion. If your speedometer reading is changing as you turn, your circular motion is nonuniform. Uniform circular motion is simpler to analyze mathematically, so we will attack it first and then pass to the nonuniform case.

Self-Check

Which of these are examples of uniform circular motion and which are nonuniform?

(a) the clothes in a clothes dryer (assuming they remain against the inside of the drum, even at the top)

(b) a rock on the end of a string being whirled in a vertical circle

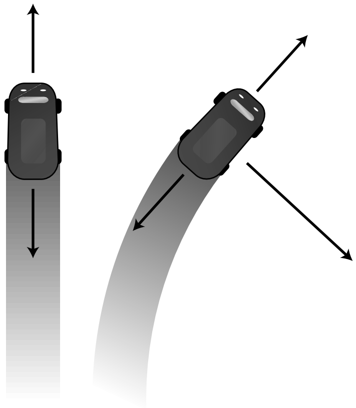
(a) Uniform. They have the same motion as the drum itself, which is rotating as one solid piece. No part of the drum can be rotating at a different speed from any other part. (b) Nonuniform. Gravity speeds it up on the way down and slows it down on the way up.

Only an inward force is required for uniform circular motion.

The figures on the previous page showed the string pulling in straight along a radius of the circle, but many people believe that when they are doing this they must be "leading" the rock a little to keep it moving along. That is, they believe that the force required to produce uniform circular motion is not directly inward but at a slight angle to the radius of the circle. This intuition is incorrect, which you can easily verify for yourself now if you have some string handy. It is only while you are getting the object going that your force needs to be at an angle to the radius. During this initial period of speeding up, the motion is not uniform. Once you settle down into uniform circular motion, you only apply an inward force.

To make the brick go in a circle, I had to exert an inward force on the rope.

If you have not done the experiment for yourself, here is a theoretical argument to convince you of this fact. We have discussed in chapter 6 the principle that forces have no perpendicular effects. To keep the rock from speeding up or slowing down, we only need to make sure that our force is perpendicular to its direction of motion. We are then guaranteed that its forward motion will remain unaffected: our force can have no perpendicular effect, and there is no other force acting on the rock which could slow it down. The rock requires no forward force to maintain its forward motion, any more than a projectile needs a horizontal force to "help it over the top" of its arc.



When a car is going straight at constant speed, the forward and backward forces on it are canceling out, producing a total force of zero. When it moves in a circle at constant speed, there are three forces on it, but the forward and backward forces cancel out, so the vector sum is an inward force.

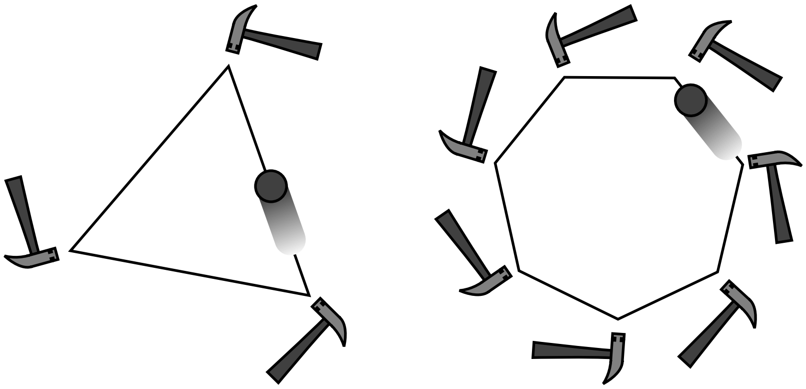
Why, then, does a car driving in circles in a parking lot stop executing uniform circular motion if you take your foot off the gas? The source of confusion here is that Newton's laws predict an object's motion based on the *total* force acting on it. A car driving in circles has three forces on it

(1) an inward force from the asphalt, controlled with the steering wheel;
(2) a forward force from the asphalt, controlled with the gas pedal; and
(3) backward forces from air resistance and rolling resistance.

You need to make sure there is a forward force on the car so that the backward forces will be exactly canceled out, creating a vector sum that points directly inward.

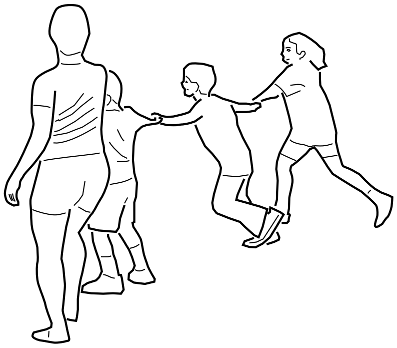
In uniform circular motion, the acceleration vector is inward

Since experiments show that the force vector points directly inward, Newton's second law implies that the acceleration vector points inward as well. This fact can also be proven on purely kinematical grounds, and we will do so in the next section.



A series of three hammer taps makes the rolling ball trace a triangle, seven hammers a heptagon. If the number of hammers was large enough, the ball would essentially be experiencing a steady inward force, and it would go in a circle. In no case is any forward force necessary.

Discussion Questions



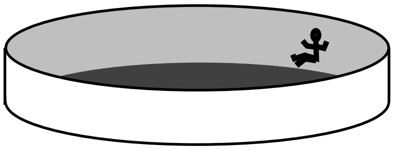
Discussion questions A-D.

A. In the game of crack the whip, a line of people stand holding hands, and then they start sweeping out a circle. One person is at the center, and rotates without changing location. At the opposite end is the person who is running the fastest, in a wide circle. In this game, someone always ends up losing their grip and flying off. Suppose the person on the end loses her grip. What path does she follow as she goes flying off? (Assume she is going so fast that she is really just trying to put one foot in front of the other fast enough to keep from falling; she is not able to get any significant horizontal force between her feet and the ground.)

B. Suppose the person on the outside is still holding on, but feels that she may loose her grip at any moment. What force or forces are acting on her, and in what directions are they? (We are not interested in the vertical forces, which are the earth's gravitational force pulling down, and the ground's normal force pushing up.)

C. Suppose the person on the outside is still holding on, but feels that she may loose her grip at any moment. What is wrong with the following analysis of the situation? "The person whose hand she's holding exerts an inward force on her, and because of Newton's third law, there's an equal and opposite force acting outward. That outward force is the one she feels throwing her outward, and the outward force is what might make her go flying off, if it's strong enough."

D. If the only force felt by the person on the outside is an inward force, why doesn't she go straight in?



Discussion question E.

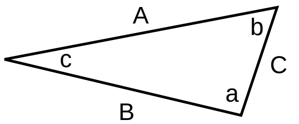
E. In the amusement park ride shown in the figure, the cylinder spins faster and faster until the customer can pick her feet up off the floor without falling. In the old Coney Island version of the ride, the floor actually dropped out like a trap door, showing the ocean below. (There is also a version in which the whole thing tilts up diagonally, but we’re discussing the version that stays flat.) If there is no outward force acting on her, why does she stick to the wall? Analyze all the forces on her.

F. What is an example of circular motion where the inward force is a normal force? What is an example of circular motion where the inward force is friction? What is an example of circular motion where the inward force is the sum of more than one force?

G. Does the acceleration vector always change continuously in circular motion? The velocity vector?

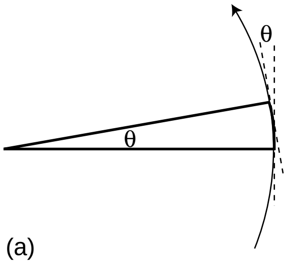
9.2 Uniform Circular Motion

In this section I derive a simple and very useful equation for the magnitude of the acceleration of an object undergoing constant acceleration. The law of sines is involved, so I've recapped it on the left.

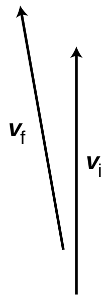


The law of sines:
A/sin a = B/sin b = C/sin c

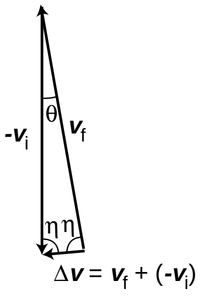
The derivation is brief, but the method requires some explanation and justification. The idea is to calculate a $\Delta\boldsymbol{v}$ vector describing the change in the velocity vector as the object passes through an angle θ. We then calculate the acceleration, $\boldsymbol{a}=\Delta\boldsymbol{v}/\Delta t$. The astute reader will recall, however, that this equation is only valid for motion with constant acceleration. Although the magnitude of the acceleration is constant for uniform circular motion, the acceleration vector changes its direction, so it is not a constant vector, and the equation $\boldsymbol{a}=\Delta\boldsymbol{v}/\Delta t$ does not apply. The justification for using it is that we will then examine its behavior when we make the time interval very short, which means making the angle θ very small. For smaller and smaller time intervals, the $\Delta\boldsymbol{v}/\Delta t$ expression becomes a better and better approximation, so that the final result of the derivation is exact.



(a)



(b)


θ
$-\boldsymbol{v}_i$
$\boldsymbol{v}_f$
η η
$\Delta\boldsymbol{v} = \boldsymbol{v}_f + (-\boldsymbol{v}_i)$

(c)

In figure (a), the object sweeps out an angle θ. Its direction of motion also twists around by an angle θ, from the vertical dashed line to the tilted one. Figure (b) shows the initial and final velocity vectors, which have equal magnitude, but directions differing by θ. In (c), the vectors have been reassembled in the proper orientation for vector subtraction. They form an isosceles triangle with interior angles θ, η, and η. (Eta, η, is my favorite Greek letter.) The law of sines gives

$$\frac{|\Delta\boldsymbol{v}|}{\sin\,\theta}=\frac{|\boldsymbol{v}|}{\sin\eta}\quad.$$

This tells us the magnitude of $\Delta\boldsymbol{v}$, which is one of the two ingredients we need for calculating the magnitude of $\boldsymbol{a}=\Delta\boldsymbol{v}/\Delta t$. The other ingredient is Δt. The time required for the object to move through the angle θ is

$$\Delta t=\frac{\text{length of arc}}{|\boldsymbol{v}|}\quad.$$

Now if we measure our angles in radians we can use the definition of radian measure, which is (angle)=(length of arc)/(radius), giving $\Delta t=\theta r/|\boldsymbol{v}|$. Combining this with the first expression involving $|\Delta\boldsymbol{v}|$ gives

$$\begin{aligned}|\boldsymbol{a}| &= |\Delta\boldsymbol{v}|/\Delta t\\ &= \frac{|\boldsymbol{v}|^2}{r}\cdot\frac{\sin\theta}{\theta}\cdot\frac{1}{\sin\eta}\quad.\end{aligned}$$

When θ becomes very small, the small-angle approximation $\sin\theta\approx\theta$ applies, and also η becomes close to 90°, so $\sin\eta\approx 1$, and we have an equation for $|\boldsymbol{a}|$:

$$|\boldsymbol{a}|=\frac{|\boldsymbol{v}|^2}{r}\qquad\text{[uniform circular motion]}\quad.$$

Example: force required to turn on a bike

Question: A bicyclist is making a turn along an arc of a circle with radius 20 m, at a speed of 5 m/s. If the combined mass of the cyclist plus the bike is 60 kg, how great a static friction force must the road be able to exert on the tires?

Solution: Taking the magnitudes of both sides of Newton's second law gives

$$\begin{aligned} |\mathbf{F}| &= |m\mathbf{a}| \\ &= m|\mathbf{a}| \quad . \end{aligned}$$

Substituting $|\mathbf{a}|=|\mathbf{v}|^2/r$ gives

$$\begin{aligned} |\mathbf{F}| &= m|\mathbf{v}|^2/r \\ &\approx 80 \text{ N} \end{aligned}$$

(rounded off to one sig fig).

Example: Don't hug the center line on a curve!

Question: You're driving on a mountain road with a steep drop on your right. When making a left turn, is it safer to hug the center line or to stay closer to the outside of the road?

Solution: You want whichever choice involves the least acceleration, because that will require the least force and entail the least risk of exceeding the maximum force of static friction. Assuming the curve is an arc of a circle and your speed is constant, your car is performing uniform circular motion, with $|\mathbf{a}|=|\mathbf{v}|^2/r$. The dependence on the square of the speed shows that driving slowly is the main safety measure you can take, but for any given speed you also want to have the largest possible value of r. Even though your instinct is to keep away from that scary precipice, you are actually less likely to skid if you keep toward the outside, because then you are describing a larger circle.

Example: acceleration related to radius and period of rotation

Question: How can the equation for the acceleration in uniform circular motion be rewritten in terms of the radius of the circle and the *period*, T, of the motion, i.e. the time required to go around once?

Solution: The period can be related to the speed as follows:

$$\begin{aligned} |\mathbf{v}| &= \frac{\text{circumference}}{T} \\ &= 2\pi r/T \quad . \end{aligned}$$

Substituting into the equation $|\mathbf{a}|=|\mathbf{v}|^2/r$ gives

$$|\mathbf{a}| = \frac{4\pi^2 r}{T^2} \quad .$$

Example: a clothes dryer

Question: My clothes dryer has a drum with an inside radius of 35 cm, and it spins at 48 revolutions per minute. What is the acceleration of the clothes inside?

Solution: We can solve this by finding the period and plugging in to the result of the previous example. If it makes 48 revolutions in one minute, then the period is 1/48 of a minute, or 1.25 s. To get an acceleration in mks units, we must convert the radius to 0.35 m. Plugging in, the result is 8.8 m/s^2.

Example: more about clothes dryers!

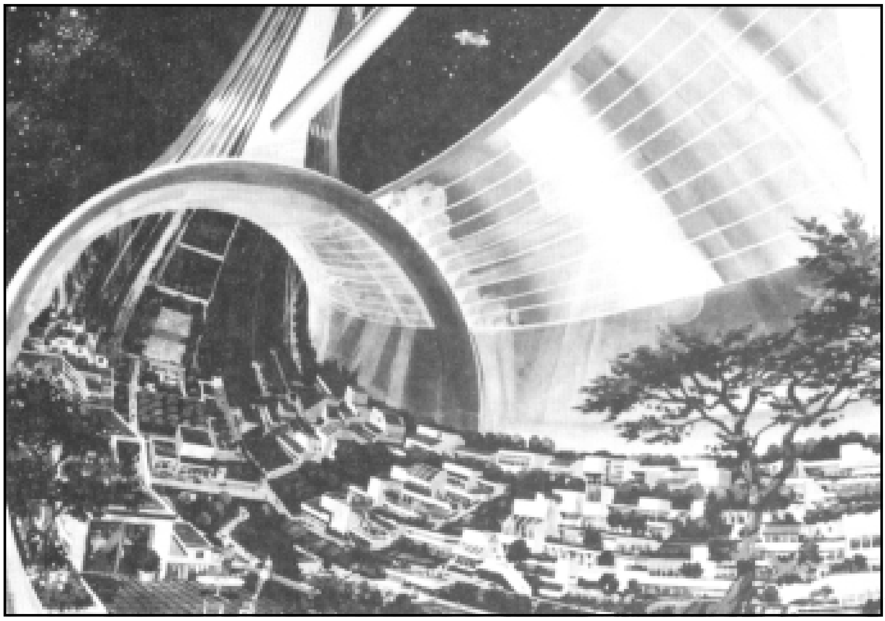
Question: In a discussion question in the previous section, we made the assumption that the clothes remain against the inside of the drum as they go over the top. In light of the previous example, is this a correct assumption?

Solution: No. We know that there must be some minimum speed at which the motor can run that will result in the clothes just barely staying against the inside of the drum as they go over the top. If the clothes dryer ran at just this minimum speed, then there would be no normal force on the clothes at the top: they would be on the verge of losing contact. The only force acting on them at the top would be the force of gravity, which would give them an acceleration of g=9.8 m/s^2. The actual dryer must be running slower than this minimum speed, because it produces an acceleration of only 8.8 m/s^2. My theory is that this is done intentionally, to make the clothes mix and tumble.

Discussion Question

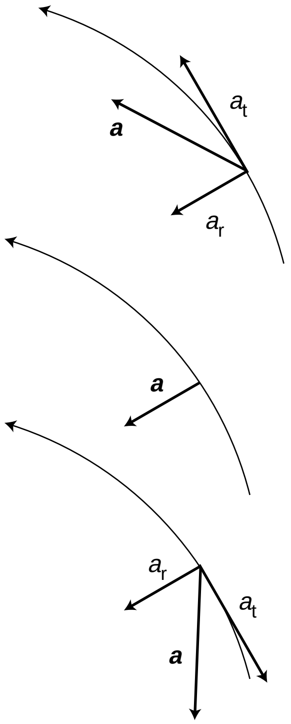
A. A certain amount of force is needed to provide the acceleration of circular motion. What if were are exerting a force perpendicular to the direction of motion in an attempt to make an object trace a circle of radius r, but the force isn't as big as $m|\mathbf{v}|^2/r$?

B. Suppose a rotating space station is built that gives its occupants the illusion of ordinary gravity. What happens when a person in the station lets go of a ball? What happens when she throws a ball straight "up" in the air (i.e. towards the center)?



An artist's conception of a rotating space colony in the form of a giant wheel. A person living in this noninertial frame of reference has an illusion of a force pulling her outward, toward the deck, for the same reason that a person in the pickup truck has the illusion of a force pulling the bowling ball. By adjusting the speed of rotation, the designers can make an acceleration $|v|^2/r$ equal to the usual acceleration of gravity on earth. On earth, your acceleration standing on the ground is zero, and a falling rock heads for your feet with an acceleration of 9.8 m/s^2. A person standing on the deck of the space colony has an *upward* acceleration of 9.8 m/s^2, and when she lets go of a rock, her feet head *up* at the nonaccelerating rock. To her, it seems the same as true gravity.
Art by NASA.

9.3 Nonuniform Circular Motion



An object moving in a circle may speed up (top), keep the magnitude of its velocity vector constant (middle), or slow down (bottom).

What about nonuniform circular motion? Although so far we have been discussing components of vectors along fixed x and y axes, it now becomes convenient to discuss components of the acceleration vector along the radial line (in-out) and the tangential line (along the direction of motion). For nonuniform circular motion, the radial component of the acceleration obeys the same equation as for uniform circular motion,

$$a_r = |\boldsymbol{v}|^2/r \quad ,$$

but the acceleration vector also has a tangential component,

$$a_t = \text{slope of the graph of } |\boldsymbol{v}| \text{ versus t} \quad .$$

The latter quantity has a simple interpretation. If you are going around a curve in your car, and the speedometer needle is moving, the tangential component of the acceleration vector is simply what you would have thought the acceleration was if you saw the speedometer and didn't know you were going around a curve.

Example: Slow down before a turn, not during it.
Question: When you're making a turn in your car and you're afraid you may skid, isn't it a good idea to slow down?
Solution: If the turn is an arc of a circle, and you've already completed part of the turn at constant speed without skidding, then the road and tires are apparently capable of enough static friction to supply an acceleration of $|\boldsymbol{v}|^2/r$. There is no reason why you would skid out now if you haven't already. If you get nervous and brake, however, then you need to have a tangential acceleration component in addition to the radial component you were already able to produce successfully. This would require an acceleration vector with a greater magnitude, which in turn would require a larger force. Static friction might not be able to supply that much force, and you might skid out. As in the previous example on a similar topic, the safe thing to do is to approach the turn at a comfortably low speed.

Summary

Selected Vocabulary

uniform circular motion	circular motion in which the magnitude of the velocity vector remains constant
nonuniform circular motion	circular motion in which the magnitude of the velocity vector changes
radial	parallel to the radius of a circle; the in-out direction
tangential	tangent to the circle, perpendicular to the radial direction

Notation

a_r	radial acceleration; the component of the acceleration vector along the in-out direction
a_t	tangential acceleration; the component of the acceleration vector tangent to the circle

Summary

If an object is to have circular motion, a force must be exerted on it toward the center of the circle. There is no outward force on the object; the illusion of an outward force comes from our experiences in which our point of view was rotating, so that we were viewing things in a noninertial frame.

An object undergoing uniform circular motion has an inward acceleration vector of magnitude

$$|\mathbf{a}| = \frac{|\mathbf{v}|^2}{r} \quad .$$

In nonuniform circular motion, the radial and tangential components of the acceleration vector are

$$a_r = |\mathbf{v}|^2/r$$

$$a_t = \text{slope of the graph of } |\mathbf{v}| \text{ versus } t \quad .$$

Homework Problems

1. When you're done using an electric mixer, you can get most of the batter off of the beaters by lifting them out of the batter with the motor running at a high enough speed. Let's imagine, to make things easier to visualize, that we instead have a piece of tape stuck to one of the beaters. (a) Explain why static friction has no effect on whether or not the tape flies off. (b) Suppose you find that the tape doesn't fly off when the motor is on a low speed, but speeding it up does cause it to fly off. Why would the greater speed change things?

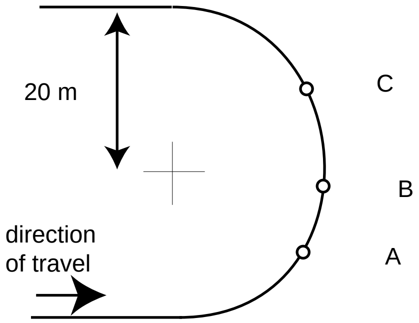
2. Show that the expression $|\boldsymbol{v}|^2/r$ has the units of acceleration.

3 ✓. A plane is flown in a loop-the-loop of radius 1.00 km. The plane starts out flying upside-down, straight and level, then begins curving up along the circular loop, and is right-side up when it reaches the top . (The plane may slow down somewhat on the way up.) How fast must the plane be going at the top if the pilot is to experience no force from the seat or the seatbelt while at the top of the loop?

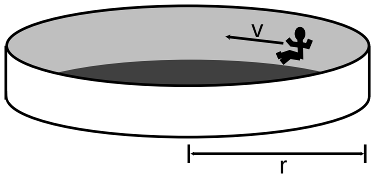
4 ∫. In this problem, you'll derive the equation $|\boldsymbol{a}|=|\boldsymbol{v}|^2/r$ using calculus. Instead of comparing velocities at two points in the particle's motion and then taking a limit where the points are close together, you'll just take derivatives. The particle's position vector is $\boldsymbol{r}=(r\cos\theta)\hat{\boldsymbol{x}} + (r\sin\theta)\hat{\boldsymbol{y}}$, where $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{y}}$ are the unit vectors along the x and y axes. By the definition of radians, the distance traveled since $t=0$ is $r\theta$, so if the particle is traveling at constant speed $v=|\boldsymbol{v}|$, we have $v=r\theta/t$. (a) Eliminate θ to get the particle's position vector as a function of time. (b) Find the particle's acceleration vector. (c) Show that the magnitude of the acceleration vector equals v^2/r.

5 S. Three cyclists in a race are rounding a semicircular curve. At the moment depicted, cyclist A is using her brakes to apply a force of 375 N to her bike. Cyclist B is coasting. Cyclist C is pedaling, resulting in a force of 375 N on her bike. Each cyclist, with her bike, has a mass of 75 kg. At the instant shown, the instantaneous speed of all three cyclists is 10 m/s. On the diagram, draw each cyclist's acceleration vector with its tail on top of her present position, indicating the directions and lengths reasonably accurately. Indicate approximately the consistent scale you are using for all three acceleration vectors. Extreme precision is not necessary as long as the directions are approximately right, and lengths of vectors that should be equal appear roughly equal, etc. Assume all three cyclists are traveling along the road all the time, not wandering across their lane or wiping out and going off the road.

S A solution is given in the back of the book.	★ A difficult problem.
✓ A computerized answer check is available.	∫ A problem that requires calculus.

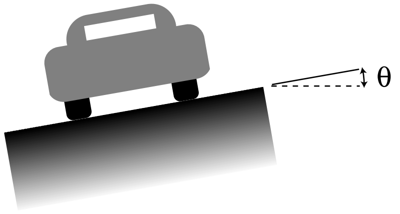


Problem 5.



Problem 6.

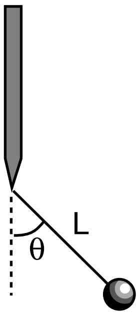
6 S★. The amusement park ride shown in the figure consists of a cylindrical room that rotates about its vertical axis. When the rotation is fast enough, a person against the wall can pick his or her feet up off the floor and remain "stuck" to the wall without falling.
(a) Suppose the rotation results in the person having a speed v. The radius of the cylinder is r, the person's mass is m, the downward acceleration of gravity is g, and the coefficient of static friction between the person and the wall is μ_s. Find an equation for the speed, v, required, in terms of the other variables. (You will find that one of the variables cancels out.)
(b) Now suppose two people are riding the ride. Huy is wearing denim, and Gina is wearing polyester, so Huy's coefficient of static friction is three times greater. The ride starts from rest, and as it begins rotating faster and faster, Gina must wait longer before being able to lift her feet without sliding to the floor. Based on your equation from part a, how many times greater must the speed be before Gina can lift her feet without sliding down?



Problem 7.

7 S. An engineer is designing a curved off-ramp for a freeway. Since the off-ramp is curved, she wants to bank it to make it less likely that motorists going too fast will wipe out. If the radius of the curve is r, how great should the banking angle, θ, be so that for a car going at a speed v, no static friction force whatsoever is required to allow the car to make the curve? State your answer in terms of v, r, and g, and show that the mass of the car is irrelevant.

8 ✓. Lionel brand toy trains come with sections of track in standard lengths and shapes. For circular arcs, the most commonly used sections have diameters of 662 and 1067 mm at the inside of the outer rail. The maximum speed at which a train can take the broader curve without flying off the tracks is 0.95 m/s. At what speed must the train be operated to avoid derailing on the tighter curve?



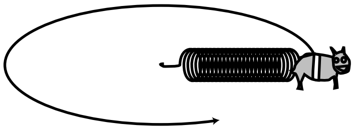
Problem 9.

9. The figure shows a ball on the end of a string of length L attached to a vertical rod which is spun about its vertical axis by a motor. The period (time for one rotation) is P.

(a) Analyze the forces in which the ball participates.

(b) Find how the angle θ depends on P, g, and L. [Hints: (1) Write down Newton's second law for the vertical and horizontal components of force and acceleration. This gives two equations, which can be solved for the two unknowns, θ and the tension in the string. (2) If you introduce variables like v and r, relate them to the variables your solution is supposed to contain, and eliminate them.]

(c) What happens mathematically to your solution if the motor is run very slowly (very large values of P)? Physically, what do you think would actually happen in this case?

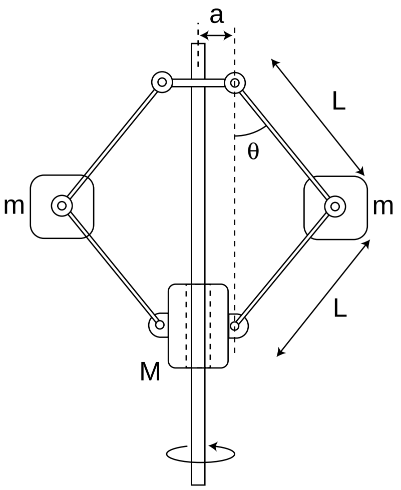


Problem 10.

10. Psychology professor R.O. Dent requests funding for an experiment on compulsive thrill-seeking behavior in hamsters, in which the subject is to be attached to the end of a spring and whirled around in a horizontal circle. The spring has equilibrium length b, and obeys Hooke's law with spring constant k. It is stiff enough to keep from bending significantly under the hamster's weight.

(a) Calculate the length of the spring when it is undergoing steady circular motion in which one rotation takes a time T. Express your result in terms of k, b, and T.

(b) The ethics committee somehow fails to veto the experiment, but the safety committee expresses concern. Why? Does your equation do anything unusual, or even spectacular, for any particular value of T? What do you think is the physical significance of this mathematical behavior?



Problem 11.

11★. The figure shows an old-fashioned device called a flyball governor, used for keeping an engine running at the correct speed. The whole thing rotates about the vertical shaft, and the mass M is free to slide up and down. This mass would have a connection (not shown) to a valve that controlled the engine. If, for instance, the engine ran too fast, the mass would rise, causing the engine to slow back down.

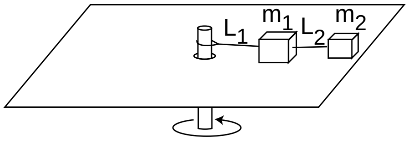
(a) Show that in the special case of a=0, the angle θ is given by

$$\theta = \cos^{-1}\left(\frac{g(m+M)P^2}{4\pi^2 mL}\right) \quad ,$$

where P is the period of rotation (time required for one complete rotation).

(b) There is no closed-form solution for θ in the general case where a is not zero. However, explain how the undesirable low-speed behavior of the a=0 device would be improved by making a nonzero.

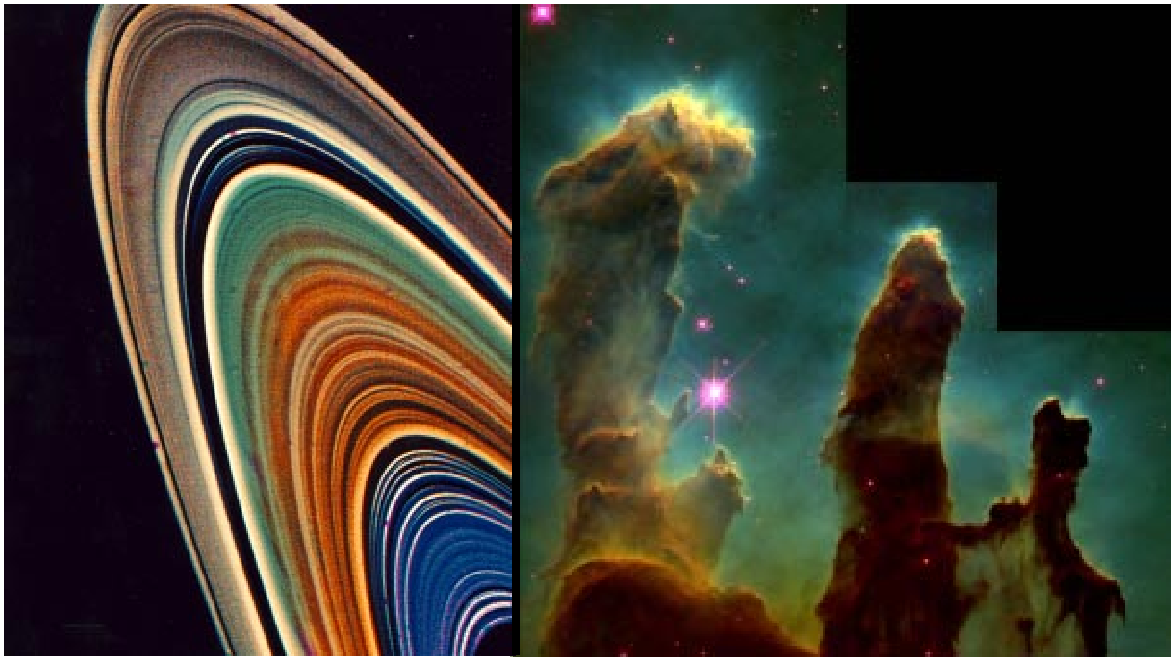
[Based on an example by J.P. den Hartog.]



Problem 12.

12. The figure shows two blocks of masses m_1 and m_2 sliding in circles on a frictionless table. Find the tension in the strings if the period of rotation (time required for one complete rotation) is P.

13. The acceleration of an object in uniform circular motion can be given either by $|\mathbf{a}|=|\mathbf{v}|^2/r$ or, equivalently, by $|\mathbf{a}|=4\pi^2 r/T^2$, where T is the time required for one cycle. (The latter expression comes simply from substituting $|\mathbf{v}|$=distance/time=circumference/T=$2\pi r/T$ into the first expression.) Person A says based on the first equation that the acceleration in circular motion is greater when the circle is smaller. Person B, arguing from the second equation, says that the acceleration is smaller when the circle is smaller. Rewrite the two statements so that they are less misleading, eliminating the supposed paradox. [Based on a problem by Arnold Arons.]



Gravity is the only really important force on the cosmic scale. Left: a false-color image of saturn's rings, composed of innumerable tiny ice particles orbiting in circles under the influence of saturn's gravity. Right: A stellar nursery, the Eagle Nebula. Each pillar of hydrogen gas is about as tall as the diameter of our entire solar system. The hydrogen molecules all attract each other through gravitational forces, resulting in the formation of clumps that contract to form new stars.

10 Gravity

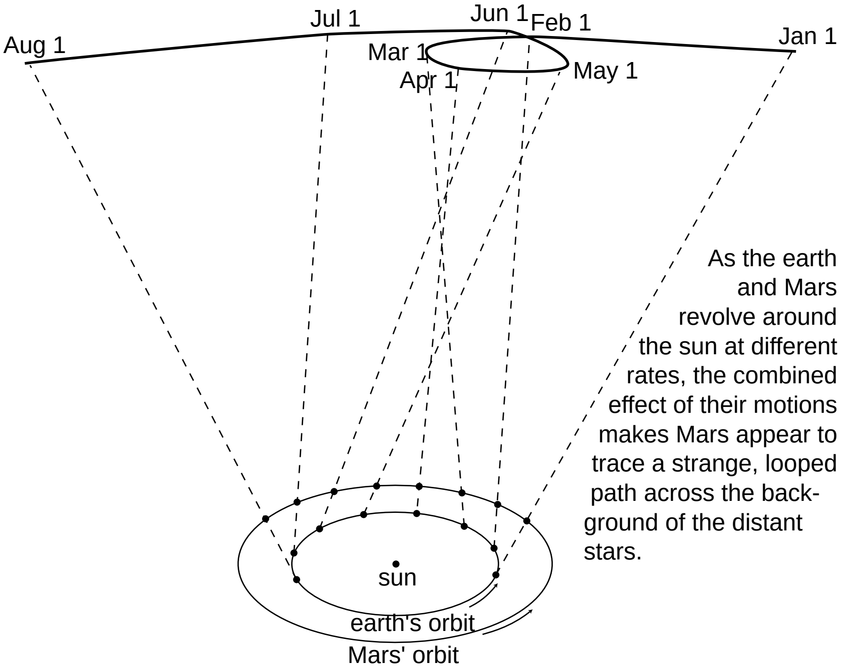
Cruise your radio dial today and try to find any popular song that would have been imaginable without Louis Armstrong. By introducing solo improvisation into jazz, Armstrong took apart the jigsaw puzzle of popular music and fit the pieces back together in a different way. In the same way, Newton reassembled our view of the universe. Consider the titles of some recent physics books written for the general reader: **The God Particle**, **Dreams of a Final Theory**. When the subatomic particle called the neutrino was recently proven for the first time to have mass, specialists in cosmology began discussing seriously what effect this would have on calculations of the ultimate fate of the universe: would the neutrinos' mass cause enough extra gravitational attraction to make the universe eventually stop expanding and fall back together? Without Newton, such attempts at universal understanding would not merely have seemed a little pretentious, they simply would not have occurred to anyone.

This chapter is about Newton's theory of gravity, which he used to explain the motion of the planets as they orbited the sun. Whereas this book has concentrated on Newton's laws of motion, leaving gravity as a dessert, Newton tosses off the laws of motion in the first 20 pages of the **Principia Mathematica** and then spends the next 130 discussing the motion of the planets. Clearly he saw this as the crucial scientific focus of his work. Why? Because in it he showed that the same laws of motion applied to the heavens as to the earth, and that the gravitational force that made an apple fall was the same as the force that kept the earth's motion from carrying it away from the sun. What was radical about Newton was not his laws of motion but his concept of a universal science of physics.

10.1 Kepler’s Laws

Tycho Brahe made his name as an astronomer by showing that the bright new star, today called a supernova, that appeared in the skies in 1572 was far beyond the Earth’s atmosphere. This, along with Galileo’s discovery of sunspots, showed that contrary to Aristotle, the heavens were not perfect and unchanging. Brahe’s fame as an astronomer brought him patronage from King Frederick II, allowing him to carry out his historic high-precision measurements of the planets’ motions. A contradictory character, Brahe enjoyed lecturing other nobles about the evils of dueling, but had lost his own nose in a youthful duel and had it replaced with a prosthesis made of an alloy of gold and silver. Willing to endure scandal in order to marry a peasant, he nevertheless used the feudal powers given to him by the king to impose harsh forced labor on the inhabitants of his parishes. The result of their work, an Italian-style palace with an observatory on top, surely ranks as one of the most luxurious science labs ever built. When the king died and his son reduced Brahe’s privileges, Brahe left in a huff for a new position in Prague, taking his data with him. He died of a ruptured bladder after falling from a wagon on the way home from a party — in those days, it was considered rude to leave the dinner table to relieve oneself.

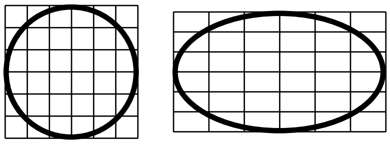
Newton wouldn’t have been able to figure out *why* the planets move the way they do if it hadn’t been for the astronomer Tycho Brahe (1546-1601) and his protege Johannes Kepler (1571-1630), who together came up with the first simple and accurate description of *how* the planets actually do move. The difficulty of their task is suggested by the figure below, which shows how the relatively simple orbital motions of the earth and Mars combine so that as seen from earth Mars appears to be staggering in loops like a drunken sailor.



As the earth and Mars revolve around the sun at different rates, the combined effect of their motions makes Mars appear to trace a strange, looped path across the background of the distant stars.

Brahe, the last of the great naked-eye astronomers, collected extensive data on the motions of the planets over a period of many years, taking the giant step from the previous observations’ accuracy of about 10 seconds of arc (10/60 of a degree) to an unprecedented 1 second. The quality of his work is all the more remarkable considering that his observatory consisted of four giant brass protractors mounted upright in his castle in Denmark. Four different observers would simultaneously measure the position of a planet in order to check for mistakes and reduce random errors.

With Brahe’s death, it fell to his former assistant Kepler to try to make some sense out of the volumes of data. Kepler, in contradiction to his late boss, had formed a prejudice, a correct one as it turned out, in favor of the theory that the earth and planets revolved around the sun, rather than the earth staying fixed and everything rotating about it. Although motion is relative, it is not just a matter of opinion what circles what. The earth’s rotation and revolution about the sun make it a noninertial reference frame, which causes detectable violations of Newton’s laws when one attempts to describe sufficiently precise experiments in the earth-fixed frame. Although such direct experiments were not carried out until the 19th century, what



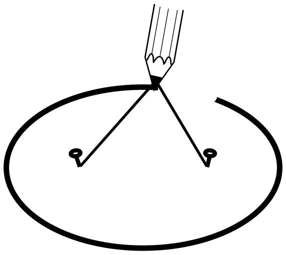
An ellipse is a circle that has been distorted by shrinking and stretching along perpendicular axes.

convinced everyone of the sun-centered system in the 17th century was that Kepler was able to come up with a surprisingly simple set of mathematical and geometrical rules for describing the planets' motion using the sun-centered assumption. After 900 pages of calculations and many false starts and dead-end ideas, Kepler finally synthesized the data into the following three laws:

Kepler's elliptical orbit law: The planets orbit the sun in elliptical orbits with the sun at one focus.

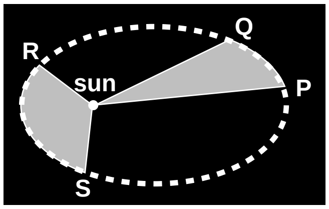
Kepler's equal-area law: The line connecting a planet to the sun sweeps out equal areas in equal amounts of time.

Kepler's law of periods: The time required for a planet to orbit the sun, called its period, is proportional to the long axis of the ellipse raised to the 3/2 power. The constant of proportionality is the same for all the planets.



An ellipse can be constructed by tying a string to two pins and drawing like this with the pencil stretching the string taut. Each pin constitutes one focus of the ellipse.

Although the planets' orbits are ellipses rather than circles, most are very close to being circular. The earth's orbit, for instance, is only flattened by 1.7% relative to a circle. In the special case of a planet in a circular orbit, the two foci (plural of "focus") coincide at the center of the circle, and Kepler's elliptical orbit law thus says that the circle is centered on the sun. The equal-area law implies that a planet in a circular orbit moves around the sun with constant speed. For a circular orbit, the law of periods then amounts to a statement that the time for one orbit is proportional to $r^{3/2}$, where r is the radius. If all the planets were moving in their orbits at the same speed, then the time for one orbit would simply depend on the circumference of the circle, so it would only be proportional to r to the first power. The more drastic dependence on $r^{3/2}$ means that the outer planets must be moving more slowly than the inner planets.



If the time interval taken by the planet to move from P to Q is equal to the time interval from R to S, then according to Kepler's equal-area law, the two shaded areas are equal. The planet is moving faster during interval RS than it did during PQ, which Newton later determined was due to the sun's gravitational force accelerating it. The equal-area law predicts exactly how much it will speed up.

10.2 Newton's Law of Gravity

The sun's force on the planets obeys an inverse square law.

Kepler's laws were a beautifully simple explanation of what the planets did, but they didn't address why they moved as they did. Did the sun exert a force that pulled a planet toward the center of its orbit, or, as suggested by Descartes, were the planets circulating in a whirlpool of some unknown liquid? Kepler, working in the Aristotelian tradition, hypothesized not just an inward force exerted by the sun on the planet, but also a second force in the direction of motion to keep the planet from slowing down. Some speculated that the sun attracted the planets magnetically.

Once Newton had formulated his laws of motion and taught them to some of his friends, they began trying to connect them to Kepler's laws. It was clear now that an inward force would be needed to bend the planets' paths. This force was presumably an attraction between the sun and each planet. (Although the sun does accelerate in response to the attractions of the planets, its mass is so great that the effect had never been detected by the prenewtonian astronomers.) Since the outer planets were moving slowly along more gently curving paths than the inner planets, their accelerations were apparently less. This could be explained if the sun's force was determined by distance, becoming weaker for the farther planets. Physicists were also familiar with the noncontact forces of electricity and magnetism, and knew that they fell off rapidly with distance, so this made sense.

In the approximation of a circular orbit, the magnitude of the sun's force on the planet would have to be

$$F = ma = mv^2/r \ . \qquad (1)$$

Now although this equation has the magnitude, *v*, of the velocity vector in it, what Newton expected was that there would be a more fundamental underlying equation for the force of the sun on a planet, and that that equation would involve the distance, *r*, from the sun to the object, but not the object's speed, *v* — motion doesn't make objects lighter or heavier.

Self-Check

If eq. (1) really was generally applicable, what would happen to an object released at rest in some empty region of the solar system?

Equation (1) was thus a useful piece of information which could be related to the data on the planets simply because the planets happened to be going in nearly circular orbits, but Newton wanted to combine it with other equations and eliminate *v* algebraically in order to find a deeper truth.

To eliminate *v*, Newton used the equation

$$\begin{aligned} v &= \frac{\text{circumference}}{T} \\ &= 2\pi r/T \ . \qquad (2) \end{aligned}$$

Of course this equation would also only be valid for planets in nearly circular orbits. Plugging this into eq. (1) to eliminate *v* gives

$$F = \frac{4\pi^2 mr}{T^2} \ . \qquad (3)$$

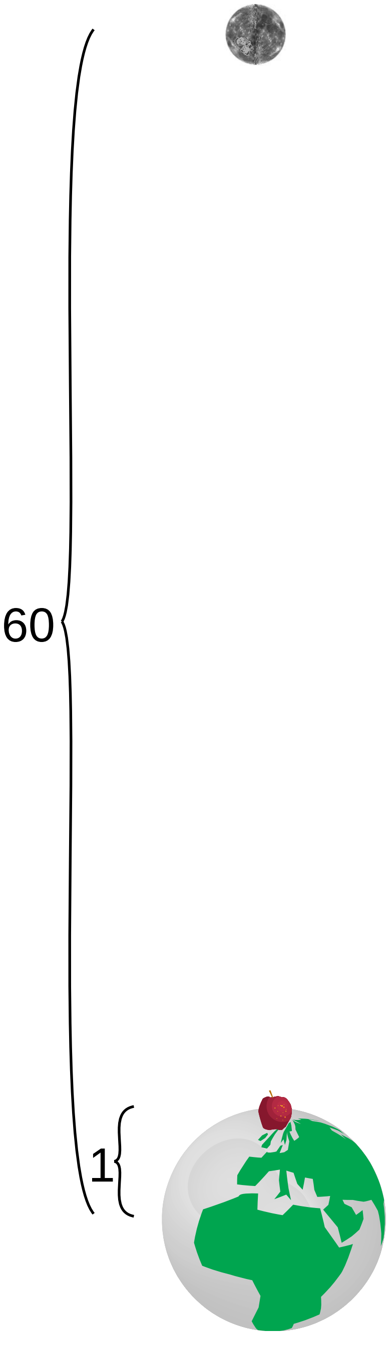
This unfortunately has the side-effect of bringing in the period, *T*, which we expect on similar physical grounds will not occur in the final answer. That's where the circular-orbit case, $T \propto r^{3/2}$, of Kepler's law of periods comes in. Using it to eliminate T gives a result that depends only on the mass of the planet and its distance from the sun:

$$F \propto m/r^2 \ . \qquad \text{[force of the sun on a planet of mass } m \text{ at a distance } r \text{ from the sun; same proportionality constant for all the planets]}$$

(Since Kepler's law of periods is only a proportionality, the final result is a proportionality rather than an equation, and there is this no point in hanging on to the factor of $4\pi^2$.)

It would just stay where it was. Plugging v=0 into eq. (1) would give F=0, so it would not accelerate from rest, and would never fall into the sun. No astronomer had ever observed an object that did that!

As an example, the "twin planets" Uranus and Neptune have nearly the same mass, but Neptune is about twice as far from the sun as Uranus, so the sun's gravitational force on Neptune is about four times smaller.



The forces between heavenly bodies are the same type of force as terrestrial gravity

OK, but what kind of force was it? It probably wasn't magnetic, since magnetic forces have nothing to do with mass. Then came Newton's great insight. Lying under an apple tree and looking up at the moon in the sky, he saw an apple fall. Might not the earth also attract the moon with the same kind of gravitational force? The moon orbits the earth in the same way that the planets orbit the sun, so maybe the earth's force on the falling apple, the earth's force on the moon, and the sun's force on a planet were all the same type of force.

There was an easy way to test this hypothesis numerically. If it was true, then we would expect the gravitational forces exerted by the earth to follow the same $F\propto m/r^2$ rule as the forces exerted by the sun, but with a different constant of proportionality appropriate to the earth's gravitational strength. The issue arises now of how to define the distance, r, between the earth and the apple. An apple in England is closer to some parts of the earth than to others, but suppose we take r to be the distance from the center of the earth to the apple, i.e. the radius of the earth. (The issue of how to measure r did not arise in the analysis of the planets' motions because the sun and planets are so small compared to the distances separating them.) Calling the proportionality constant k, we have

$$F_{\text{earth on apple}} = k\, m_{\text{apple}} / r_{\text{earth}}^2$$

$$F_{\text{earth on moon}} = k\, m_{\text{moon}} / d_{\text{earth-moon}}^2 \;.$$

Newton's second law says $a=F/m$, so

$$a_{\text{apple}} = k / r_{\text{earth}}^2$$

$$a_{\text{moon}} = k / d_{\text{earth-moon}}^2 \;.$$

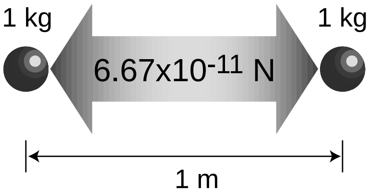
The Greek astronomer Hipparchus had already found 2000 years before that the distance from the earth to the moon was about 60 times the radius of the earth, so if Newton's hypothesis was right, the acceleration of the moon would have to be $60^2=3600$ times less than the acceleration of the falling apple.

Applying $a=v^2/r$ to the acceleration of the moon yielded an acceleration that was indeed 3600 times smaller than 9.8 m/s^2, and Newton was convinced he had unlocked the secret of the mysterious force that kept the moon and planets in their orbits.

Newton's law of gravity

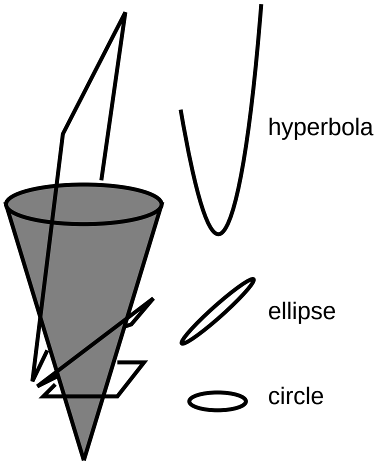
The proportionality $F\propto m/r^2$ for the gravitational force on an object of mass m only has a consistent proportionality constant for various objects if they are being acted on by the gravity of the same object. Clearly the sun's gravitational strength is far greater than the earth's, since the planets all orbit the sun and do not exhibit any very large accelerations caused by the earth (or by one another). What property of the sun gives it its great gravitational strength? Its great volume? Its great mass? Its great temperature? Newton reasoned that if the force was proportional to the mass of the object being acted on, then it would also make sense if the determining factor in the gravitational strength of the object exerting the force was its own mass. Assuming there were no other factors affecting the gravitational force, then the only other thing needed to make quantitative predictions of gravitational forces would be a proportionality constant. Newton called that proportionality constant G, and the complete form of the law of gravity he hypothesized was

$$F = Gm_1m_2/r^2 \quad . \quad \text{[gravitational force between objects of mass } m_1 \text{ and } m_2\text{, separated by a distance } r\text{; } r \text{ is not the radius of anything]}$$



The gravitational attraction between two 1-kg masses separated by a distance of 1 m is 6.67×10^{-11} N. Do not memorize this number!

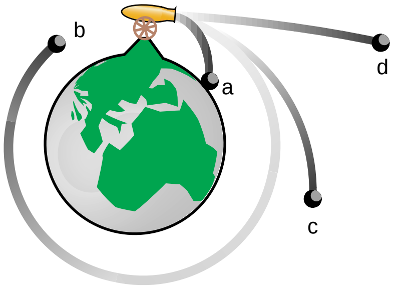
Newton conceived of gravity as an attraction between any two masses in the universe. The constant G tells us the how many newtons the attractive force is for two 1-kg masses separated by a distance of 1 m. The experimental determination of G in ordinary units (as opposed to the special, nonmetric, units used in astronomy) is described in section 10.5. This difficult measurement was not accomplished until long after Newton's death.



The conic sections are the curves made by cutting the surface of an infinite cone with a plane.

The proportionality to $1/r^2$ actually was not entirely unexpected. Proportionalities to $1/r^2$ are found in many other phenomena in which some effect spreads out from a point. For instance, the intensity of the light from a candle is proportional to $1/r^2$, because at a distance r from the candle, the light has to be spread out over the surface of an imaginary sphere of area $4\pi r^2$. The same is true for the intensity of sound from a firecracker, or the intensity of gamma radiation emitted by the Chernobyl reactor. It's important, however, to realize that this is only an analogy. Force does not travel through space as sound or light does, and force is not a substance that can be spread thicker or thinner like butter on toast.

Although several of Newton's contemporaries had speculated that the force of gravity might be proportional to $1/r^2$, none of them, even the ones who had learned Newton's laws of motion, had had any luck proving that the resulting orbits would be ellipses, as Kepler had found empirically. Newton did succeed in proving that elliptical orbits would result from a $1/r^2$ force, but we postpone the proof until the end of the next volume of the textbook because it can be accomplished much more easily using the concepts of energy and angular momentum.



An imaginary cannon able to shoot cannonballs at very high speeds is placed on top of an imaginary, very tall mountain that reaches up above the atmosphere. Depending on the speed at which the ball is fired, it may end up in a tightly curved elliptical orbit, a, a circular orbit, b, a bigger elliptical orbit, c, or a nearly straight hyperbolic orbit, d.

Newton also predicted that orbits in the shape of hyperbolas should be possible, and he was right. Some comets, for instance, orbit the sun in very elongated ellipses, but others pass through the solar system on hyperbolic paths, never to return. Just as the trajectory of a faster baseball pitch is flatter than that of a more slowly thrown ball, so the curvature of a planet's orbit depends on its speed. A spacecraft can be launched at relatively low speed, resulting in a circular orbit about the earth, or it can be launched at a higher speed, giving a more gently curved ellipse that reaches farther from the earth, or it can be launched at a very high speed which puts it in an even less curved hyperbolic orbit. As you go very far out on a hyperbola, it approaches a straight line, i.e. its curvature eventually becomes nearly zero.

Newton also was able to prove that Kepler's second law (sweeping out equal areas in equal time intervals) was a logical consequence of his law of gravity. Newton's version of the proof is moderately complicated, but the proof becomes trivial once you understand the concept of angular momentum, which will be covered later in the course. The proof will therefore be deferred until section 5.7 of book 2.

Self-Check

Which of Kepler's laws would it make sense to apply to hyperbolic orbits?

Discussion Questions

A. How could Newton find the speed of the moon to plug in to $a=v^2/r$?

B. Two projectiles of different mass shot out of guns on the surface of the earth at the same speed and angle will follow the same trajectories, assuming that air friction is negligible. (You can verify this by throwing two objects together from your hand and seeing if they separate or stay side by side.) What corresponding fact would be true for satellites of the earth having different masses?

C. What is wrong with the following statement? "A comet in an elliptical orbit speeds up as it approaches the sun, because the sun's force on it is increasing."

D. Why would it not make sense to expect the earth's gravitational force on a bowling ball to be inversely proportional to the square of the distance between their surfaces rather than their centers?

E. Does the earth accelerate as a result of the moon's gravitational force on it? Suppose two planets were bound to each other gravitationally the way the earth and moon are, but the two planets had equal masses. What would their motion be like?

F. Spacecraft normally operate by firing their engines only for a few minutes at a time, and an interplanetary probe will spend months or years on its way to its destination without thrust. Suppose a spacecraft is in a circular orbit around Mars, and it then briefly fires its engines in reverse, causing a sudden decrease in speed. What will this do to its orbit? What about a forward thrust?

10.3 Apparent Weightlessness

If you ask somebody at the bus stop why astronauts are weightless, you'll probably get one of the following two incorrect answers:

(1) They're weightless because they're so far from the earth.

(2) They're weightless because they're moving so fast.

The first answer is wrong, because the vast majority of astronauts never get more than a thousand miles from the earth's surface. The reduction in gravity caused by their altitude is significant, but not 100%. The second answer is wrong because Newton's law of gravity only depends on distance, not speed.

The correct answer is that astronauts in orbit around the earth are not really weightless at all. Their weightlessness is only apparent. If there was no gravitational force on the spaceship, it would obey Newton's first law and move off on a straight line, rather than orbiting the earth. Likewise, the astronauts inside the spaceship are in orbit just like the spaceship itself, with the earth's gravitational force continually twisting their velocity vectors around. The reason they appear to be weightless is that they are in the same orbit as the spaceship, so although the earth's gravity curves their trajectory

The equal-area law makes equally good sense in the case of a hyperbolic orbit (and observations verify it). The elliptical orbit law had to be generalized by Newton to include hyperbolas. The law of periods doesn't make sense in the case of a hyperbolic orbit, because a hyperbola never closes back on itself, so the motion never repeats.

down toward the deck, the deck drops out from under them at the same rate.

Apparent weightlessness can also be experienced on earth. Any time you jump up in the air, you experience the same kind of apparent weightlessness that the astronauts do. While in the air, you can lift your arms more easily than normal, because gravity does not make them fall any faster than the rest of your body, which is falling out from under them. The Russian air force now takes rich foreign tourists up in a big cargo plane and gives them the feeling of weightlessness for a short period of time while the plane is nose-down and dropping like a rock.

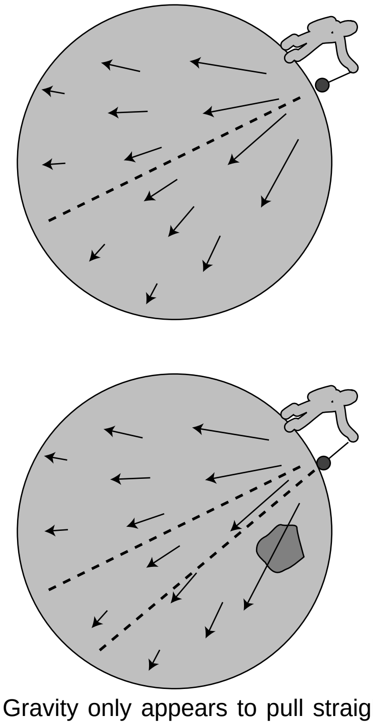
10.4 Vector Addition of Gravitational Forces

Pick a flower on earth and you move the farthest star.

Paul Dirac

When you stand on the ground, which part of the earth is pulling down on you with its gravitational force? Most people are tempted to say that the effect only comes from the part directly under you, since gravity always pulls straight down. Here are three observations that might help to change your mind:

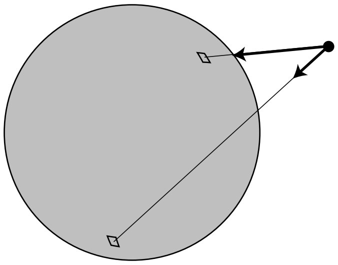
- If you jump up in the air, gravity does not stop affecting you just because you are not touching the earth: gravity is a noncontact force. That means you are not immune from the gravity of distant parts of our planet just because you are not touching them.
- Gravitational effects are not blocked by intervening matter. For instance, in an eclipse of the moon, the earth is lined up directly between the sun and the moon, but only the sun's light is blocked from reaching the moon, not its gravitational force — if the sun's gravitational force on the moon was blocked in this situation, astronomers would be able to tell because the moon's acceleration would change suddenly. A more subtle but more easily observable example is that the tides are caused by the moon's gravity, and tidal effects can occur on the side of the earth facing away from the moon. Thus, far-off parts of the earth are not prevented from attracting you with their gravity just because there is other stuff between you and them.
- Prospectors sometimes search for underground deposits of dense minerals by measuring the direction of the local gravitational forces, i.e. the direction things fall or the direction a plumb bob hangs. For instance, the gravitational forces in the region to the west of such a deposit would point along a line slightly to the east of the earth's center. Just because the total gravitational force on you points down, that doesn't mean that only the parts of the earth directly below you are attracting you. It's just that the sideways components of all the force vectors acting on you come very close to canceling out.



Gravity only appears to pull straight down because the near perfect symmetry of the earth makes the sideways components of the total force on an object cancel almost exactly. If the symmetry is broken, e.g. by a dense mineral deposit, the total force is a little off to the side.

A cubic centimeter of lava in the earth's mantle, a grain of silica inside Mt. Kilimanjaro, and a flea on a cat in Paris are all attracting you with their gravity. What you feel is the vector sum of all the gravitational forces exerted by all the atoms of our planet, and for that matter by all the atoms in the universe.

When Newton tested his theory of gravity by comparing the orbital acceleration of the moon to the acceleration of a falling apple on earth, he assumed he could compute the earth's force on the apple using the distance from the apple to the earth's center. Was he wrong? After all, it isn't just the earth's center attracting the apple, it's the whole earth. A kilogram of dirt a few feet under his backyard in England would have a much greater force on the apple than a kilogram of molten rock deep under Australia, thousands of miles away. There's really no obvious reason why the force should come out right if you just pretend that the earth's whole mass is concentrated at its center. Also, we know that the earth has some parts that are more dense, and some parts that are less dense. The solid crust, on which we live, is considerably less dense that the molten rock underneath on which it floats. By all rights, the computation of the vector sum of all the forces exerted by all the earth's parts should be a horrendous mess.



An object outside a spherical shell of mass will feel gravitational forces from every part of the shell — stronger forces from the closer parts and weaker ones from the parts farther away. The shell theorem states that the vector sum of all the forces is the same as if all the mass had been concentrated at the center of the shell.

Actually, Newton had sound mathematical reasons for treating the earth's mass as if it was concentrated at its center. First, although Newton no doubt suspected the earth's density was nonuniform, he knew that the direction of its total gravitational force was very nearly toward the earth's center. That was strong evidence that the distribution of mass was very symmetric, so that we can think of the earth as being made of many layers, like an onion, with each layer having constant density throughout. (Today there is further evidence for symmetry based on measurements of how the vibrations from earthquakes and nuclear explosions travel through the earth.) Newton then concentrated on the gravitational forces exerted by a single such thin shell, and proved the following mathematical theorem, known as the shell theorem:

> If an object lies outside a thin, uniform shell of mass, then the vector sum of all the gravitational forces exerted by all the parts of the shell is the same as if all the shell's mass was concentrated at its center. If the object lies inside the shell, then all the gravitational forces cancel out exactly.

For terrestrial gravity, each shell acts as though its mass was concentrated at the earth's center, so the final result is the same as if the earth's whole mass was concentrated at its center.

The second part of the shell theorem, about the gravitational forces canceling inside the shell, is a little surprising. Obviously the forces would all cancel out if you were at the exact center of a shell, but why should they still cancel out perfectly if you are inside the shell but off-center? The whole idea might seem academic, since we don't know of any hollow planets in our solar system that astronauts could hope to visit, but actually it's a useful result for understanding gravity within the earth, which is an important issue in geology. It doesn't matter that the earth is not actually hollow.

In a mine shaft at a depth of, say, 2 km, we can use the shell theorem to tell us that the outermost 2 km of the earth has no net gravitational effect, and the gravitational force is the same as what would be produced if the remaining, deeper, parts of the earth were all concentrated at its center.

Discussion Questions

A. If you hold an apple in your hand, does the apple exert a gravitational force on the earth? Is it much weaker than the earth's gravitational force on the apple? Why doesn't the earth seem to accelerate upward when you drop the apple?

B. When astronauts travel from the earth to the moon, how does the gravitational force on them change as they progress?

C. How would the gravity in the first-floor lobby of a massive skyscraper compare with the gravity in an open field outside of the city?

10.5 Weighing the Earth

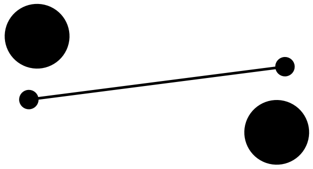
Let's look more closely at the application of Newton's law of gravity to objects on the earth's surface. Since the earth's gravitational force is the same as if its mass was all concentrated at its center, the force on a falling object of mass *m* is given by

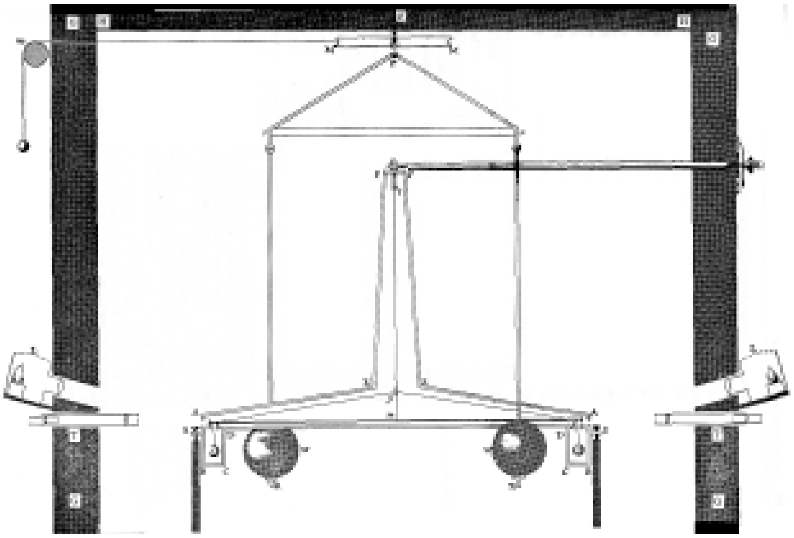
$$F = G M_{\text{earth}} m / r_{\text{earth}}^2 \quad .$$

The object's acceleration equals *F*/*m*, so the object's mass cancels out and we get the same acceleration for all falling objects, as we knew we should:

$$g = G M_{\text{earth}} / r_{\text{earth}}^2 \quad .$$

Newton knew neither the mass of the earth nor a numerical value for the constant *G*. But if someone could measure *G*, then it would be possible for the first time in history to determine the mass of the earth! The only way to measure *G* is to measure the gravitational force between two objects of known mass, but that's an exceedingly difficult task, because the force between any two objects of ordinary size is extremely small. The English physicist Henry Cavendish was the first to succeed, using the apparatus shown in the diagrams. The two larger balls were lead spheres 8 inches in diameter, and each one attracted the small ball near it. The two small balls hung from the ends of a horizontal rod, which itself hung by a thin thread. The frame from which the larger balls hung could be rotated by hand about

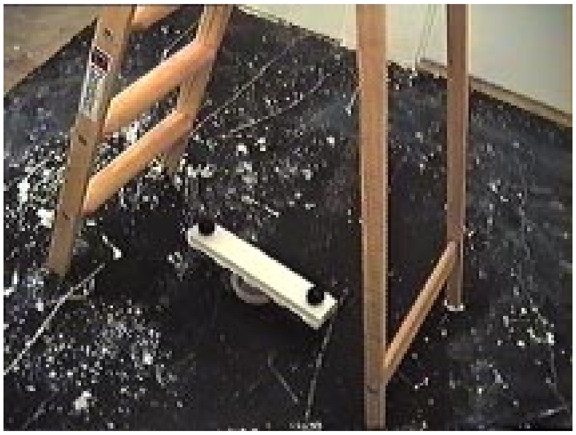


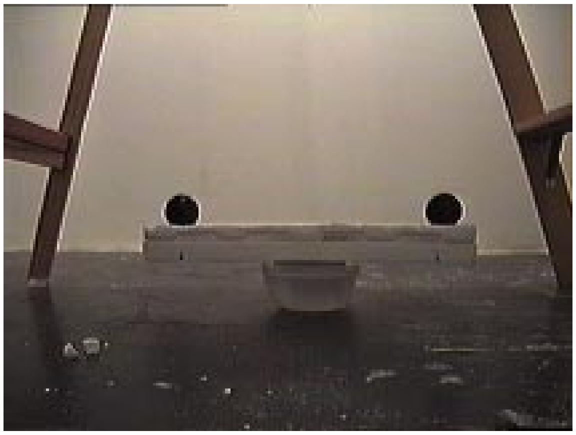


Cavendish's apparatus viewed from the side, and a simplified version viewed from above. The two large balls are fixed in place, but the rod from which the two small balls hang is free to twist under the influence of the gravitational forces.

a vertical axis, so that for instance the large ball on the right would pull its neighboring small ball toward us and while the small ball on the left would be pulled away from us. The thread from which the small balls hung would thus be twisted through a small angle, and by calibrating the twist of the thread with known forces, the actual gravitational force could be determined. Cavendish set up the whole apparatus in a room of his house, nailing all the doors shut to keep air currents from disturbing the delicate apparatus. The results had to be observed through telescopes stuck through holes drilled in the walls. Cavendish's experiment provided the first numerical values for G and for the mass of the earth. The presently accepted value of G is $6.67\text{x}10^{-11}$ $\text{N}\cdot\text{m}^2/\text{kg}^2$.

The following page shows a modern-day Cavendish experiment constructed by one of my students.

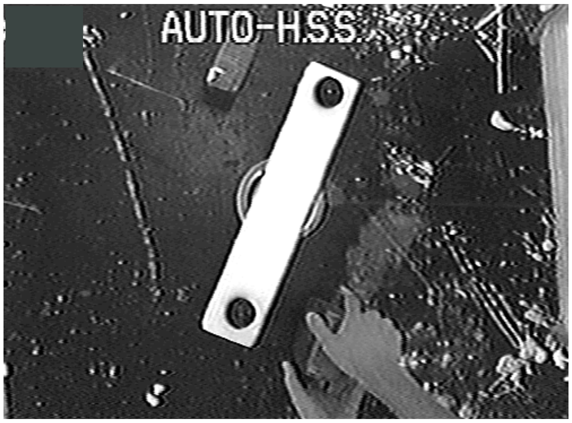


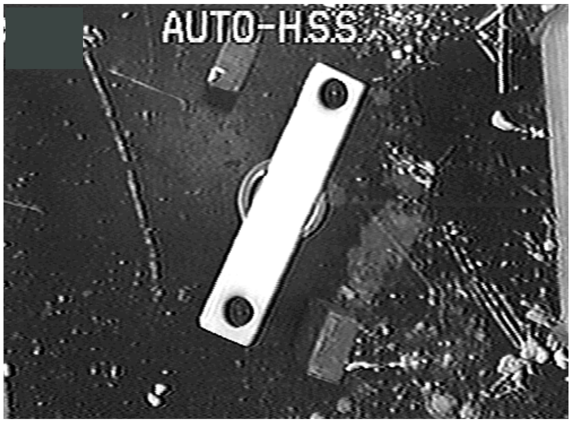


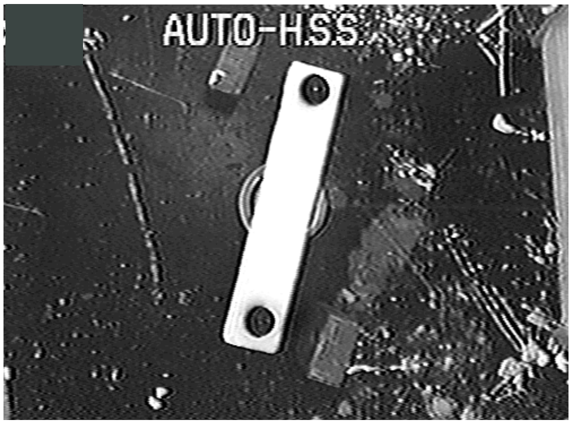
My student Narciso Guzman built this version of the Cavendish experiment in his garage, from a description on the Web at www.fourmilab.to. Two steel balls sit near the ends of a piece of styrofoam, which is suspending from a ladder by fishing line (not visible in this photo). To make vibrations die out more quickly, a small piece of metal from a soda can is attached underneath the styrofoam arm, sticking down into a bowl of water. (The arm is not resting on the bowl.)

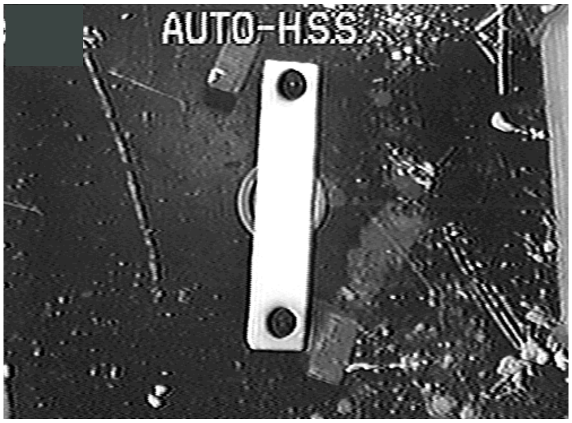
The sequence of four video frames on the right shows the apparatus in action. Initially (top), lead bricks are inserted near the steel balls. They attract the balls, and the arm begins to rotate counterclockwise.

The main difficulties in this experiment are isolating the apparatus from vibrations and air currents. Narciso had to leave the room while the camcorder ran. Also, it is helpful if the apparatus can be far from walls or furniture that would create gravitational forces on it.









Knowing *G* not only allowed the determination of the earth's mass but also those of the sun and the other planets. For instance, by observing the acceleration of one of Jupiter's moons, we can infer the mass of Jupiter. The following table gives the distances of the planets from the sun and the masses of the sun and planets. (Other data are given in the back of the book.)

	average distance from the sun, in units of the earth's average distance from the sun	**mass, in units of the earth's mass**
sun	—	330,000
mercury	0.38	.056
venus	.72	.82
earth	1	1
mars	1.5	.11
jupiter	5.2	320
saturn	9.5	95
uranus	19	14
neptune	30	17
pluto	39	.002

Discussion Questions

A. It would have been difficult for Cavendish to start designing an experiment without at least some idea of the order of magnitude of G. How could he estimate it in advance to within a factor of 10?

B. Fill in the details of how one would determine Jupiter's mass by observing the acceleration of one of its moons. Why is it only necessary to know the acceleration of the moon, not the actual force acting on it? Why don't we need to know the mass of the moon? What about a planet that has no moons, such as Venus — how could its mass be found?

C. The gravitational constant *G* is very difficult to measure accurately, and is the least accurately known of all the fundamental numbers of physics such as the speed of light, the mass of the electron, etc. But that's in the mks system, based on the meter as the unit of length, the kilogram as the unit of mass, and the second as the unit of distance. Astronomers sometimes use a different system of units, in which the unit of distance, called the astronomical unit or a.u., is the radius of the earth's orbit, the unit of mass is the mass of the sun, and the unit of time is the year (i.e. the time required for the earth to orbit the sun). In this system of units, *G* has a precise numerical value simply as a matter of definition. What is it?

10.6* Evidence for Repulsive Gravity

Until recently, physicists thought they understood gravity fairly well. Einstein had modified Newton's theory, but certain characteristrics of gravitational forces were firmly established. For one thing, they were always attractive. If gravity always attracts, then it is logical to ask why the universe doesn't collapse. Newton had answered this question by saying that if the universe was infinite in all directions, then it would have no geometric center toward which it would collapse; the forces on any particular star or planet exerted by distant parts of the universe would tend to cancel out by symmetry. More careful calculations, however, show that Newton's universe would have a tendency to collapse on smaller scales: any part of the universe that happened to be slightly more dense than average would contract further, and this contraction would result in stronger gravitational forces, which would cause even more rapid contraction, and so on.

When Einstein overhauled gravity, the same problem reared its ugly head. Like Newton, Einstein was predisposed to believe in a universe that was static, so he added a special repulsive term to his equations, intended to prevent a collapse. This term was not associated with any attraction of mass for mass, but represented merely an overall tendency for space itself to expand unless restrained by the matter that inhabited it. It turns out that Einstein's solution, like Newton's, is unstable. Furthermore, it was soon discovered observationally that the universe was expanding, and this was interpreted by creating the Big Bang model, in which the universe's current expansion is the aftermath of a fantastically hot explosion. An expanding universe, unlike a static one, was capable of being explained with Einstein's equations, without any repulsion term. The univsrse's expansion would simply slow down over time due to the attractive gravitational forces. After these developments, Einstein said woefully that adding the repulsive term, known as the cosmological constant, had been the greatest blunder of his life.

Book 3, section 3.5 presents some of the evidence for the Big Bang.

This was the state of things until 1999, when evidence began to turn up that the universe's expansion has been speeding up rather than slowing down! The first evidence came from using a telescope as a sort of time machine: light from a distant galaxy may have taken billions of years to reach us, so we are seeing it as it was far in the past. Looking back in time, astronomers saw the universe expanding at speeds that ware lower, rather than higher. At first they were mortified, since this was exactly the opposite of what had been expected. The statistical quality of the data was also not good enough to constute ironclad proof, and there were worries about systematic errors. The case for an accelerating expansion has however been nailed down by high-precision mapping of the dim, sky-wide afterglow of the Big Bang, known as the cosmic microwave background. Some theorists have proposed reviving Einstein's cosmological constant to account for the acceleration, while others believe it is evidence for a mysterious form of matter which exhibits gravitational repulsion. Some recent ideas on this topic can be found in the January 2001 issue of Scientific American, which is available online at

http://www.sciam.com/2001/0101issue/0101currentissue.html .

Summary

Selected Vocabulary

ellipse .... a flattened circle; one of the conic sections

conic section .... a curve formed by the intersection of a plane and an infinite cone

hyperbola .... another conic section; it does not close back on itself

period .... the time required for a planet to complete one orbit; more generally, the time for one repetition of some repeating motion

focus .... one of two special points inside an ellipse: the ellipse consists of all points such that the sum of the distances to the two foci equals a certain number; a hyperbola also has a focus

Notation

G .... the constant of proportionality in Newton's law of gravity; the gravitational force of attraction between two 1-kg spheres at a center-to-center distance of 1 m

Summary

Kepler deduced three empirical laws from data on the motion of the planets:

> **Kepler's elliptical orbit law**: The planets orbit the sun in elliptical orbits with the sun at one focus.
>
> **Kepler's equal-area law**: The line connecting a planet to the sun sweeps out equal areas in equal amounts of time.
>
> **Kepler's law of periods**: The time required for a planet to orbit the sun is proportional to the long axis of the ellipse raised to the 3/2 power. The constant of proportionality is the same for all the planets.

Newton was able to find a more fundamental explanation for these laws. *Newton's law of gravity* states that the magnitude of the attractive force between any two objects in the universe is given by

$$F = Gm_1m_2/r^2 \quad .$$

Weightlessness of objects in orbit around the earth is only apparent. An astronaut inside a spaceship is simply falling along with the spaceship. Since the spaceship is falling out from under the astronaut, it appears as though there was no gravity accelerating the astronaut down toward the deck.

Gravitational forces, like all other forces, add like vectors. A gravitational force such as we ordinarily feel is the vector sum of all the forces exerted by all the parts of the earth. As a consequence of this, Newton proved the *shell theorem* for gravitational forces:

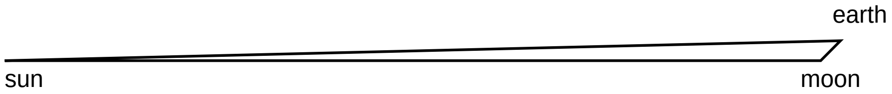
> If an object lies outside a thin, uniform shell of mass, then the vector sum of all the gravitational forces exerted by all the parts of the shell is the same as if all the shell's mass was concentrated at its center. If the object lies inside the shell, then all the gravitational forces cancel out exactly.

Homework Problems

1 ✓. Roy has a mass of 60 kg. Laurie has a mass of 65 kg. They are 1.5 m apart.
(a) What is the magnitude of the gravitational force of the earth on Roy?
(b) What is the magnitude of Roy's gravitational force on the earth?
(c) What is the magnitude of the gravitational force between Roy and Laurie?
(d) What is the magnitude of the gravitational force between Laurie and the sun?

2. During a solar eclipse, the moon, earth and sun all lie on the same line, with the moon between the earth and sun. Define your coordinates so that the earth and moon lie at greater *x* values than the sun. For each force, give the correct sign as well as the magnitude. (a) What force is exerted on the moon by the sun? (b) On the moon by the earth? (c) On the earth by the sun? (d) What total force is exerted on the sun? (e) On the moon? (f) On the earth?

3✓. Suppose that on a certain day there is a crescent moon, and you can tell by the shape of the crescent that the earth, sun and moon form a triangle with a 135° interior angle at the moon's corner. What is the magnitude of the total gravitational force of the earth and the sun on the moon?



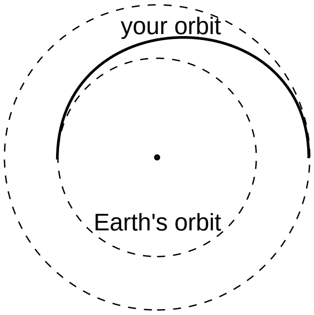
4. How high above the Earth's surface must a rocket be in order to have 1/100 the weight it would have at the surface? Express your answer in units of the radius of the Earth.

5✓. The star Lalande 21185 was found in 1996 to have two planets in roughly circular orbits, with periods of 6 and 30 years. What is the ratio of the two planets' orbital radii?

6. In a Star Trek episode, the Enterprise is in a circular orbit around a planet when something happens to the engines. Spock then tells Kirk that the ship will spiral into the planet's surface unless they can fix the engines. Is this scientifically correct? Why?

S A solution is given in the back of the book.
✓ A computerized answer check is available.
★ A difficult problem.
∫ A problem that requires calculus.

7. (a) Suppose a rotating spherical body such as a planet has a radius r and a uniform density ρ, and the time required for one rotation is T. At the surface of the planet, the apparent acceleration of a falling object is reduced by acceleration of the ground out from under it. Derive an equation for the apparent acceleration of gravity, g, at the equator in terms of r, ρ, T, and G.
(b) Applying your equation from (a), by what fraction is your apparent weight reduced at the equator compared to the poles, due to the Earth's rotation?
(c) Using your equation from (a), derive an equation giving the value of T for which the apparent acceleration of gravity becomes zero, i.e. objects can spontaneously drift off the surface of the planet. Show that T only depends on ρ, and not on r.
(d) Applying your equation from (c), how long would a day have to be in order to reduce the apparent weight of objects at the equator of the Earth to zero? [Answer: 1.4 hours]
(e) Observational astronomers have recently found objects they called pulsars, which emit bursts of radiation at regular intervals of less than a second. If a pulsar is to be interpreted as a rotating sphere beaming out a natural "searchlight" that sweeps past the earth with each rotation, use your equation from (c) to show that its density would have to be much greater than that of ordinary matter.
(f) Theoretical astronomers predicted decades ago that certain stars that used up their sources of energy could collapse, forming a ball of neutrons with the fantastic density of $\sim 10^{17}$ kg/m^3. If this is what pulsars really are, use your equation from (c) to explain why no pulsar has ever been observed that flashes with a period of less than 1 ms or so.



Problem 8.

8. You are considering going on a space voyage to Mars, in which your route would be half an ellipse, tangent to the Earth's orbit at one end and tangent to Mars' orbit at the other. Your spacecraft's engines will only be used at the beginning and end, not during the voyage. How long would the outward leg of your trip last? (Assume the orbits of Earth and Mars are circular.)

9.★ (a) If the earth was of uniform density, would your weight be increased or decreased at the bottom of a mine shaft? Explain. (b) In real life, objects weight slightly more at the bottom of a mine shaft. What does that allow us to infer about the Earth?

10 S. Ceres, the largest asteroid in our solar system, is a spherical body with a mass 6000 times less than the earth's, and a radius which is 13 times smaller. If an astronaut who weighs 400 N on earth is visiting the surface of Ceres, what is her weight?

11 S. Prove, based on Newton's laws of motion and Newton's law of gravity, that all falling objects have the same acceleration if they are dropped at the same location on the earth and if other forces such as friction are unimportant. Do not just say, "g=9.8 m/s^2 -- it's constant." You are supposed to be *proving* that g should be the same number for all objects.

Problem 12.

12 S. The figure shows an image from the Galileo space probe taken during its August 1993 flyby of the asteroid Ida. Astronomers were surprised when Galileo detected a smaller object orbiting Ida. This smaller object, the only known satellite of an asteroid in our solar system, was christened Dactyl, after the mythical creatures who lived on Mount Ida, and who protected the infant Zeus. For scale, Ida is about the size and shape of Orange County, and Dactyl the size of a college campus. Galileo was unfortunately unable to measure the time, T, required for Dactyl to orbit Ida. If it had, astronomers would have been able to make the first accurate determination of the mass and density of an asteroid. Find an equation for the density, ρ, of Ida in terms of Ida's known volume, V, the known radius, r, of Dactyl's orbit, and the lamentably unknown variable T. (This is the same technique that was used successfully for determining the masses and densities of the planets that have moons.)

13 ∫. If a bullet is shot straight up at a high enough velocity, it will never return to the earth. This is known as the escape velocity. We will discuss escape velocity using the concept of energy in the next book of the series, but it can also be gotten at using straightforward calculus. In this problem, you will analyze the motion of an object of mass m whose initial velocity is *exactly* equal to escape velocity. We assume that it is starting from the surface of a spherically symmetric planet of mass M and radius b. The trick is to guess at the general form of the solution, and then determine the solution in more detail. Assume (as is true) that the solution is of the form $r = kt^{p}$, where r is the object's distance from the center of the planet at time t, and k and p are constants. (a) Find the acceleration, and use Newton's second law and Newton's law of gravity to determine k and p. You should find that the result is independent of m. (b) What happens to the velocity as t approaches infinity? (c) Determine escape velocity from the Earth's surface.

14. Astronomers have recently observed stars orbiting at very high speeds around an unknown object near the center of our galaxy. For stars orbiting at distances of about 10^{14} m from the object, the orbital velocities are about 10^{6} m/s. Assuming the orbits are circular, estimate the mass of the object, in units of the mass of the sun, $2\text{x}10^{30}$ kg. If the object was a tightly packed cluster of normal stars, it should be a very bright source of light. Since no visible light is detected coming from it, it is instead believed to be a supermassive black hole.

15 S. Astronomers have detected a solar system consisting of three planets orbiting the star Upsilon Andromedae. The planets have been named b, c, and d. Planet b's average distance from the star is 0.059 A.U., and planet c's average distance is 0.83 A.U., where an astronomical unit or A.U. is defined as the distance from the Earth to the sun. For technical reasons, it is possible to determine the ratios of the planets' masses, but their masses cannot presently be determined in absolute units. Planet c's mass is 3.0 times that of planet b. Compare the star's average gravitational force on planet c with its average force on planet b. [Based on a problem by Arnold Arons.]

16 S. Some communications satellites are in orbits called geosynchronous: the satellite takes one day to orbit the earth from west to east, so that as the earth spins, the satellite remains above the same point on the equator. What is such a satellite's altitude above the surface of the earth?

17 S. As is discussed in more detail in section 5.1 of book 2, tidal interactions with the earth are causing the moon's orbit to grow gradually larger. Laser beams bounced off of a mirror left on the moon by astronauts have allowed a measurement of the moon's rate of recession, which is about 1 cm per year. This means that the gravitational force acting between earth and moon is decreasing. By what fraction does the force decrease with each 27-day orbit? [Hint: If you try to calculate the two forces and subtract, your calculator will probably give a result of zero due to rounding. Instead, reason about the fractional amount by which the quantity $1/r^2$ will change. As a warm-up, you may wish to observe the percentage change in $1/r^2$ that results from changing r from 1 to 1.01. Based on a problem by Arnold Arons.]

18. Suppose that we inhabited a universe in which, instead of Newton's law of gravity, we had $F = k\sqrt{m_1 m_2}\ /\ r^2$, where k is some constant with different units than G. (The force is still attractive.) However, we assume that $a=F/m$ and the rest of Newtonian physics remains true, and we use $a=F/m$ to define our mass scale, so that, e.g., a mass of 2 kg is one which exhibits half the acceleration when the same force is applied to it as to a 1 kg mass. (a) Is this new law of gravity consistent with Newton's third law? (b) Suppose you lived in such a universe, and you dropped two unequal masses side by side. What would happen? (c) Numerically, suppose a 1.0-kg object falls with an acceleration of 10 m/s^2. What would be the acceleration of a rain drop with a mass of 0.1 g? Would you want to go out in the rain? (d) If a falling object broke into two unequal pieces while it fell, what would happen? (e) Invent a law of gravity that results in behavior that is the opposite of what you found in part b. [Based on a problem by Arnold Arons.]

Exercises

Exercise 0A: Models and Idealization

Equipment:

coffee filters
ramps (one per group)
balls of various sizes
sticky tape
vacuum pump and “guinea and feather” apparatus (one)

The motion of falling objects has been recognized since ancient times as an important piece of physics, but the motion is inconveniently fast, so in our everyday experience it can be hard to tell exactly what objects are doing when they fall. In this exercise you will use several techniques to get around this problem and study the motion. Your goal is to construct a scientific *model* of falling. A model means an explanation that makes testable predictions. Often models contain simplifications or idealizations that make them easier to work with, even though they are not strictly realistic.

1. One method of making falling easier to observe is to use objects like feathers that we know from everyday experience will not fall as fast. You will use coffee filters, in stacks of various sizes, to test the following two hypotheses and see which one is true, or whether neither is true:

> Hypothesis 1A: When an object is dropped, it rapidly speeds up to a certain natural falling speed, and then continues to fall at that speed. The falling speed is *proportional* to the object’s weight. (A proportionality is not just a statement that if one thing gets bigger, the other does too. It says that if one becomes three times bigger, the other also gets three times bigger, etc.)
>
> Hypothesis 1B: Different objects fall the same way, regardless of weight.

Test these hypotheses and discuss your results with your instructor.

2. A second way to slow down the action is to let a ball roll down a ramp. The steeper the ramp, the closer to free fall. Based on your experience in part 1, write a hypothesis about what will happen when you race a heavier ball against a lighter ball down the same ramp, starting them both from rest.

Hypothesis:____

Show your hypothesis to your instructor, and then test it.

You have probably found that falling was more complicated than you thought! Is there more than one factor that affects the motion of a falling object? Can you imagine certain idealized situations that are simpler?Try to agree verbally with your group on an informal model of falling that can make predictions about the experiments described in parts 3 and 4.

3. You have three balls: a standard “comparison ball” of medium weight, a light ball, and a heavy ball. Suppose you stand on a chair and (a) drop the light ball side by side with the comparison ball, then (b) drop the heavy ball side by side with the comparison ball, then (c) join the light and heavy balls together with sticky tape and drop them side by side with the comparison ball.

Use your model to make a prediction:____

Test your prediction.

4. Your instructor will pump nearly all the air out of a chamber containing a feather and a heavier object, then let them fall side by side in the chamber.

Use your model to make a prediction:____

Exercise 1A: Scaling Applied to Leaves

Equipment:

leaves of three sizes, having roughly similar proportions of length, width, and thickness
(example: blades of grass, large ficus leaves, and agave leaves)
balance

1. Each group will have one leaf, and should measure its surface area and volume, and determine its surface-to-volume ratio (surface area divided by volume). For consistency, every group should use units of cm^2 and cm^3, and should only find the area of one side of the leaf. The area can be found by tracing the area of the leaf on graph paper and counting squares. The volume can be found by weighing the leaf and assuming that its density is 1 g/cm^3, which is nearly true since leaves are mostly water. Write your results on the board for comparison with the other groups' numbers.

2. Both the surface area and the volume are bigger for bigger leaves, but what about the surface to volume ratios? What implications would this have for the plants' abilities to survive in different environments?

Exercise 2A: Changing Velocity

This exercise involves Michael Johnson's world-record 200-meter sprint in the 1996 Olympics. The table gives the distance he has covered at various times. (The data are made up, except for his total time.) Each group is to find a value of $\Delta x/\Delta t$ between two specified instants, with the members of the group checking each other's answers. We will then compare everyone's results and discuss how this relates to velocity.

	t (s)	x (m)
A	10.200	100.0000
B	10.210	100.0990
C	10.300	100.9912
D	11.200	110.0168
E	19.320	200.0000

group 1:	Find $\Delta x/\Delta t$ using points A and B.
group 2:	Find $\Delta x/\Delta t$ using points A and C.
group 3:	Find $\Delta x/\Delta t$ using points A and D.
group 4:	Find $\Delta x/\Delta t$ using points A and E.

Exercise 3A: Reasoning with Ratios and Powers

Equipment:
ping-pong balls and paddles
two-meter sticks

You have probably bounced a ping pong ball straight up and down in the air. The time between hits is related to the height to which you hit the ball. If you take twice as much time between hits, how many times higher do you think you will have to hit the ball? Write down your hypothesis:_________________________

Your instructor will first beat out a tempo of 240 beats per minute (four beats per second), which you should try to match with the ping-pong ball. Measure the height to which the ball rises:_______________

Now try it at 120 beats per minute:________________

Compare your hypothesis and your results with the rest of the class.

Exercise 4A: Force and Motion

Equipment:

2-meter pieces of butcher paper
wood blocks with hooks
string
masses to put on top of the blocks to increase friction
spring scales (preferably calibrated in Newtons)

Suppose a person pushes a crate, sliding it across the floor at a certain speed, and then repeats the same thing but at a higher speed. This is essentially the situation you will act out in this exercise. What do you think is different about her force on the crate in the two situations? Discuss this with your group and write down your hypothesis:

____

1. First you will measure the amount of friction between the wood block and the butcher paper when the wood and paper surfaces are slipping over each other. The idea is to attach a spring scale to the block and then slide the butcher paper under the block while using the scale to keep the block from moving with it. Depending on the amount of force your spring scale was designed to measure, you may need to put an extra mass on top of the block in order to increase the amount of friction. It is a good idea to use long piece of string to attach the block to the spring scale, since otherwise one tends to pull at an angle instead of directly horizontally.

First measure the amount of friction force when sliding the butcher paper as slowly as possible:____________________

Now measure the amount of friction force at a significantly higher speed, say 1 meter per second. (If you try to go too fast, the motion is jerky, and it is impossible to get an accurate reading.) ____________________

Discuss your results. Why are we justified in assuming that the string's force on the block (i.e. the scale reading) is the same amount as the paper's frictional force on the block?

2. Now try the same thing but with the block moving and the paper standing still. Try two different speeds.

Do your results agree with your original hypothesis? If not, discuss what's going on. How does the block "know" how fast to go?

Exercise 4B: Interactions

Equipment:

- neodymium disc magnets (3/group)
- compass
- triple-arm balance (2/group)
- clamp and 50-cm vertical rod for holding balance up
- string
- tape
- scissors

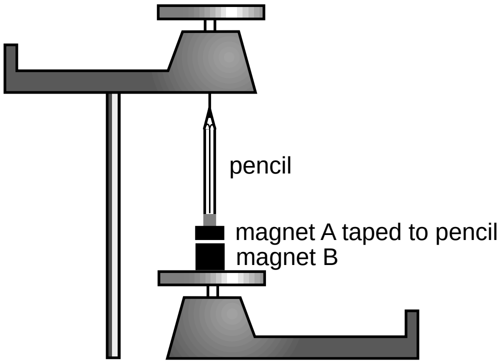
Your goal in this exercise is to compare the forces two magnets exert on each other, i.e. to compare magnet A's force on magnet B to magnet B's force on magnet A. Magnet B will be made out of two of the small disc magnets put together, so it is twice as strong as magnet A.

1. Note that these magnets are extremely strong! Being careful not to pinch your skin, put two disc magnets together to make magnet B.

2. Familiarize yourself with how the magnets behave. In addition to magnets A and B, there are two other magnets that can come into play. The compass needle itself is a magnet, and the planet earth is a magnet. Ordinarily the compass needle twists around under the influence of the earth, but the disc magnets are very strong close up, so if you bring them within a few cm of the compass, the compass is essentially just responding to them. Investigate how different parts of magnets A and B interact with the compass, and label them appropriately. Investigate how magnets A and B can attract or repel one another.

3. You are ready to form a hypothesis about the following situation. Suppose we set up two balances as shown in the figure. The magnets are not touching. The top magnet is hanging from a hook underneath the pan, giving the same result as if it was on top of the pan. Make sure it is hanging under the *center* of the pan. You will want to make sure the magnets are pulling on each other, not pushing each other away, so that the top magnet will stay in one place.

The balances will not show the magnets' true weights, because the magnets are exerting forces on each other. The top balance will read a higher number than it would without any magnetic forces, and the bottom balance will have a lower than normal reading. The difference between each magnet's true weight and the reading on the balance gives a measure of how strongly the magnet is being pushed or pulled by the other magnet.



How do you think the amount of pushing or pulling experienced by the two magnets will compare? In other words, which reading will change more, or will they change by the same amount? Write down a hypothesis:________________________

Before going on to part 4, discuss your hypothesis with your instructor.

4. Now set up the experiment described above with two balances. Since we are interested in the changse in the scale readings caused by the magnetic forces, you will need to take a total of four scale readings: one pair with the balances separated and one pair with the magnets close together as shown in the figure above.

When the balances are together and the magnetic forces are acting, it is not possible to get both balances to reach equilibrium at the same time, because sliding the weights on one balance can cause its magnet to move up or down, tipping the other balance. Therefore, while you take a reading from one balance, you need to immobilize the other in the horizontal position by taping its tip so it points exactly at the zero mark.

You will also probably find that as you slide the weights, the pointer swings suddenly to the opposite side, but you can never get it to be stable in the middle (zero) position. Try bringing the pointer manually to the zero position and then releasing it. If it swings up, you're too low, and if it swings down, you're too high. Search for the dividing line between the too-low region and the too-high region.

If the changes in the scale readings are very small (say a few grams or less), you need to get the magnets closer together. It should be possible to get the scale readings to change by large amounts (up to 10 or 20 g).

Exercise 5A: Friction

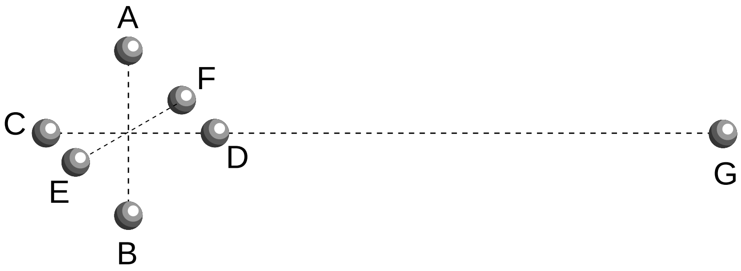
Equipment:

- 2-meter pieces of butcher paper
- wood blocks with hooks
- string
- masses to put on top of the blocks to increase friction
- spring scales (preferably calibrated in Newtons)

1. Using the same equipment as in exercise 4A, test the statement that kinetic friction is approximately independent of velocity.

2. Test the statement that kinetic friction is independent of surface area.

Exrcise 10A: The Shell Theorem



This exercise is an approximate numerical test of the shell theorem. There are seven masses A-G, each being one kilogram. Masses A-E, each one meter from the center, form a shape like two Egyptian pyramids joined at their bases; this is a rough approximation to a six-kilogram spherical shell of mass. Mass G is five meters from the center of the main group. The class will divide into six groups and split up the work required in order to calculate the vector sum of the six gravitational forces exerted on mass G. Depending on the size of the class, more than one group may be assigned to deal with the contribution of the same mass to the total force, and the redundant groups can check each other’s results.

1. Discuss as a class what can be done to simplify the task of calculating the vector sum, and how to organize things so that each group can work in parallel with the others.

2. Each group should write its results on the board in units of piconewtons, retaining six significant figures of precision.

3. The class will determine the vector sum and compare with the result that would be obtained with the shell theorem.

Solutions to Selected Problems

Chapter 0

6. $134\text{ mg}\times\frac{10^{-3}\text{ g}}{1\text{ mg}}\times\frac{10^{-3}\text{ kg}}{1\text{ g}}=1.34\times10^{-4}\text{ kg}$

Chapter 1

10. $1\text{ mm}^2\times\left(\frac{1\text{ cm}}{10\text{ mm}}\right)^2=10^{-2}\text{ cm}^2$

11. The bigger scope has a diameter that's ten times greater. Area scales as the square of the linear dimensions, so its light-gathering power is a hundred times greater (10x10).

12. Since they differ by two steps on the Richter scale, the energy of the bigger quake is 10000 times greater. The wave forms a hemisphere, and the surface area of the hemisphere over which the energy is spread is proportional to the square of its radius. If the amount of vibration was the same, then the surface areas much be in the ratio of 10000:1, which means that the ratio of the radii is 100:1.

Chapter 2

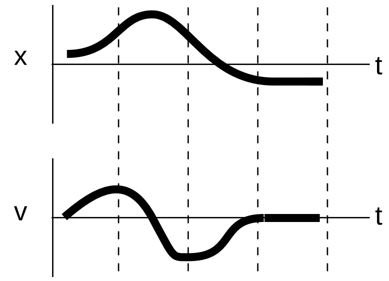
4.

$$\begin{aligned}\text{1 light-year} &= v\Delta t\\ &= \left(3.0\text{x}10^8\text{ m/s}\right)\left(1\text{ year}\right)\\ &\quad\times\left(\frac{365\text{ days}}{1\text{ year}}\right)\left(\frac{24\text{ hours}}{1\text{ day}}\right)\left(\frac{3600\text{ s}}{1\text{ hour}}\right)\\ &= 9.5\text{x}10^{15}\text{ m}\end{aligned}$$

5. Velocity is relative, so having to lean tells you nothing about the train's velocity. Fullerton is moving at a huge speed relative to Beijing, but that doesn't produce any noticeable effect in either city. The fact that you have to lean tells you that the train is accelerating.

Chapter 3

14.



15. Taking g to be 10 m/s, the bullet loses 10 m/s of speed every second, so it will take 10 s to come to a stop, and then another 10 s to come back down, for a total of 20 s.

16. $\Delta x=\frac{1}{2}at^2$, so for a fixed value of Δx, we have $t\propto1/\sqrt{a}$. Decreasing a by a factor of 3 means that t will increase by a factor of $\sqrt{3}$.

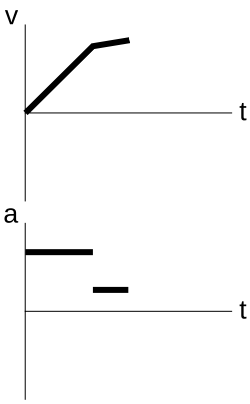
17.

$$\begin{aligned}v &= \frac{dx}{dt}\\ &= 10-3t^2\\ a &= \frac{dv}{dt}\\ &= \text{—}6\text{t}\\ &= \text{—}18\text{ m/s}^2\end{aligned}$$

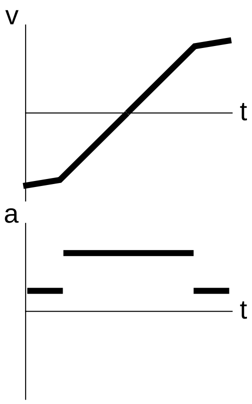
18. (a) Solving $\Delta x=\frac{1}{2}at^2$ for a, we find a=2$\Delta x/t^2$=5.51 m/s^2. (b) v=$\sqrt{2a\Delta x}$ =66.6 m/s. (c) The actual car's final velocity is less than that of the idealized constant-acceleration car. If the real car and the idealized car covered the quarter mile in the same time but the real car was moving more slowly at the end than the idealized one, the real car must have been going faster than the idealized car at the beginning of the race. The real car apparently has a greater acceleration at the beginning, and less acceleration at the end. This make sense, because every car has some maximum speed, which is the speed beyond which it cannot accelerate.

19. Since the lines are at intervals of one m/s and one second, each box represents one meter. From t=0 to t=2 s, the area under the curve represents a positive Δx of 6 m. (The triangle has half the area of the 2x6 rectangle it fits inside.) After t=2 s, the area above the curve represents negative Δx. To get –6 m worth of area, we need to go out to t=6 s, at which point the triangle under the axis has a width of 4 s and a height of 3 m/s, for an area of 6 m (half of 3x4).

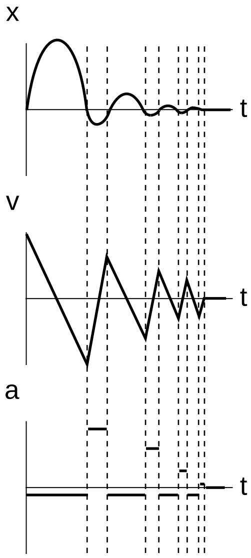
20. (a) We choose a coordinate system with positive pointing to the right. Some people might expect that the ball would slow down once it was on the more gentle ramp. This may be true if there is significant friction, but Galileo's experiments with inclined planes showed that when friction is negligible, a ball rolling on a ramp has constant acceleration, not constant speed. The speed stops increasing as quickly once the ball is on the more gentle slope, but it still keeps on increasing. The a-t graph can be drawn by inspecting the slope of the v-t graph.



(b) The ball will roll back down, so the second half of the motion is the same as in part a. In the first (rising) half of the motion, the velocity is negative, since the motion is in the opposite direction compared to the positive x axis. The acceleration is again found by inspecting the slope of the v-t graph.



21. This is a case where it's probably easiest to draw the acceleration graph first. While the ball is in the air, the only force acting on it is gravity, so it must have the same, constant acceleration during each hop. Choosing a coordinate system where the positive x axis points up, this becomes a negative acceleration (force in the opposite direction compared to the axis). During the short times between hops when the ball is in contact with the ground, it experiences a large acceleration, which turns around its velocity very rapidly. These short positive accelerations probably aren't constant, but it's hard to know how they'd really look. We just idealize them as constant accelerations. Since our acceleration graph consists of constant-acceleration segments, the velocity graph must consist of line segments, and the position graph must consist of parabola.



22. We have $v_f^2=2a\Delta x$, so the distance is proportional to the square of the velocity. To get up to half the speed, the ball needs 1/4 the distance, i.e. L/4.

Chapter 4

7. $a=\frac{\Delta v}{\Delta t}$, and also $a=\frac{F}{m}$, so

$$\begin{aligned}\Delta t &= \frac{\Delta v}{a}\\ &= \frac{m\Delta v}{F}\\ &= \frac{(1000\ \text{kg})(50\ \text{m/s} - 20\ \text{m/s})}{3000\ \text{N}}\\ &= 10\ \text{s}\end{aligned}$$

Chapter 5

14. (a)

top spring’s rightward force on connector
...connector’s leftward force on top spring

bottom spring’s rightward force on connector
...connector’s leftward force on bottom spring

hand’s leftward force on connector
...connector’s rightward force on hand

Looking at the three forces on the connector, we see that the hand’s force must be double the force of either spring. The value of x-x_o is the same for both springs and for the arrangement as a whole, so the spring constant must be 2k. This corresponds to a stiffer spring (more force to produce the same extension).

(b) Forces in which the left spring participates:

hand’s leftward force on left spring
...left spring’s rightward force on hand

right spring’s rightward force on left spring
...left spring’s leftward force on right spring

Forces in which the right spring participates:

left spring’s leftward force on right spring
...right spring’s rightward force on left spring

wall’s rightward force on right spring
...right spring’s leftward force on wall

Since the left spring isn’t accelerating, the total force on it must be zero, so the two forces acting on it must be equal in magnitude. The same applies to the two forces acting on the right spring. The forces between the two springs are connected by Newton’s third law, so all eight of these forces must be equal in magnitude. Since the value of x-x_o for the whole setup is double what it is for either spring individually, the spring constant of the whole setup must be k/2, which corresponds to a less stiff spring.

16. (a) Spring constants in parallel add, so the spring constant has to be proportional to the cross-sectional area. Two springs in series give half the spring constant, three springs in series give 1/3, and so on, so the spring constant has to be inversely proportional to the length. Summarizing, we have $k \propto A/L$.
(b) With the Young’s modulus, we have k=(A/L)E.The spring constant has units of N/m, so the units of E would have to be N/m^2.

Chapter 6

5. (a) The easiest strategy is to find the time spent aloft, and then find the range. The vertical motion and the horizontal motion are independent. The vertical motion has acceleration —g, and the cannonball spends enough time in the air to reverse its vertical velocity component completely, so we have

$$\begin{aligned}\Delta v_y &= v_{yf} - v_{yi}\\ &= -2v\sin\theta \quad .\end{aligned}$$

The time spent aloft is therefore

$$\begin{aligned}\Delta t &= \Delta v_y / a_y\\ &= 2v\sin\theta / g \quad .\end{aligned}$$

During this time, the horizontal distance traveled is

$$\begin{aligned}R &= v_x\Delta t\\ &= 2\ v^2 \sin\theta\cos\theta / g \quad .\end{aligned}$$

(b) The range becomes zero at both θ=0 and at θ=90°. The θ=0 case gives zero range because the ball hits the ground as soon as it leaves the mouth of the cannon. A 90 degree angle gives zero range because the cannonball has no horizontal motion.

Chapter 8

8. We want to find out about the velocity vector $\boldsymbol{v}_{BG}$ of the bullet relative to the ground, so we need to add Annie’s velocity relative to the ground $\boldsymbol{v}_{AG}$ to the bullet’s velocity vector $\boldsymbol{v}_{BA}$ relative to her. Letting the positive x axis be east and y north, we have

$$\begin{aligned}\boldsymbol{v}_{BA,x} &= (140\text{ mi/hr})\cos 45^\circ\\ &= 100\text{ mi/hr}\\ \boldsymbol{v}_{BA,y} &= (140\text{ mi/hr})\sin 45^\circ\\ &= 100\text{ mi/hr}\end{aligned}$$

and

$$\begin{aligned}\boldsymbol{v}_{AG,x} &= 0\\ \boldsymbol{v}_{AG,y} &= 30\text{ mi/hr} \quad .\end{aligned}$$

The bullet’s velocity relative to the ground therefore has components

$$\begin{aligned}\boldsymbol{v}_{BG,x} &= 100\text{ mi/hr} \quad \text{and}\\ \boldsymbol{v}_{BG,y} &= 130\text{ mi/hr} \quad .\end{aligned}$$

Its speed on impact with the animal is the magnitude of this vector

$$\begin{aligned}|\boldsymbol{v}_{BG}| &= \sqrt{(100\text{ mi/hr})^2 + (130\text{ mi/hr})^2}\\ &= 160\text{ mi/hr}\end{aligned}$$

(rounded off to 2 significant figures).

9. Since its velocity vector is constant, it has zero acceleration, and the sum of the force vectors acting on it must be zero. There are three forces acting on the plane: thrust, lift, and gravity. We are given the first two, and if we can find the third we can infer its mass. The sum of the y components of the forces is zero, so

$$\begin{aligned} 0 &= F_{\text{thrust,y}} + F_{\text{lift,y}} + F_{\text{W,y}} \\ &= |\mathbf{F}_{\text{thrust}}| \sin\theta + |\mathbf{F}_{\text{lift}}| \cos\theta — mg \quad . \end{aligned}$$

The mass is

$$\begin{aligned} m &= (|\mathbf{F}_{\text{thrust}}| \sin\theta + |\mathbf{F}_{\text{lift}}| \cos\theta) \,/\, g \\ &= 6.9\text{x}10^4 \text{ kg} \end{aligned}$$

10. (a) Since the wagon has no acceleration, the total forces in both the x and y directions must be zero. There are three forces acting on the wagon: F_T, F_W, and the normal force from the ground, F_N. If we pick a coordinate system with x being horizontal and y vertical, then the angles of these forces measured counterclockwise from the x axis are 90°-φ, 270°, and 90°+θ, respectively. We have

$$\begin{aligned} &F_{\text{x,total}} \\ &\quad = F_T\cos(90°\text{-}\varphi) + F_W\cos(270°) + F_N\cos(90°+\theta) \\ &F_{\text{y,total}} \\ &\quad = F_T\sin(90°\text{-}\varphi) + F_W\sin(270°) + F_N\sin(90°+\theta) \quad , \end{aligned}$$

which simplifies to

$$\begin{aligned} 0 &= F_T \sin\varphi - F_N \sin\theta \\ 0 &= F_T \cos\varphi \ - F_W + F_N \cos\theta \ . \end{aligned}$$

The normal force is a quantity that we are not given and do not with to find, so we should choose it to eliminate. Solving the first equation for $F_N=(\sin\varphi/\sin\theta)F_T$, we eliminate F_N from the second equation,

$$0 = F_T \cos\varphi \ - F_W + F_T \sin\varphi \cos\theta/\sin\theta$$

and solve for F_T, finding

$$F_T = \frac{F_W}{\cos\varphi \ + \sin\varphi \cos\theta \,/\, \sin\theta} \quad .$$

Multiplying both the top and the bottom of the fraction by sin θ, and using the trig identity for sin(θ+φ) gives the desired result,

$$F_T = \frac{\sin\theta}{\sin\left(\theta + \varphi\right)} F_W$$

(b) The case of φ=0, i.e. pulling straight up on the wagon, results in $F_T=F_W$: we simply support the wagon and it glides up the slope like a chair-lift on a ski slope. In the case of φ=180°-θ, F_T becomes infinite. Physically this is because we are pulling directly into the ground, so no amount of force will suffice.

11. (a) If there was no friction, the angle of repose would be zero, so the coefficient of static friction, μ_s, will definitely matter. We also make up symbols θ, m and g for the angle of the slope, the mass of the object, and the acceleration of gravity. The forces form a triangle just like the one in section 8.3, but instead of a force applied by an external object, we have static friction, which is less than $\mu_s F_N$. As in that example, $F_s=mg\sin\theta$, and $F_s<\mu_s F_N$, so

$$mg\sin\theta<\mu_s F_N \quad .$$

From the same triangle, we have $F_N=mg\cos\theta$, so

$$mg\sin\theta < \mu_s mg\cos\theta.$$

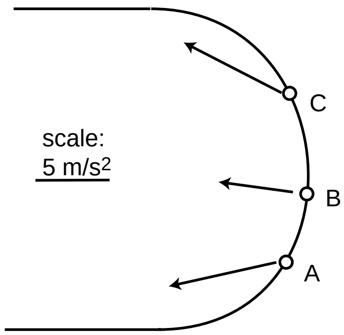
Rearranging,

$$\theta < \tan^{-1}\mu_s \quad .$$

(b) Both m and g canceled out, so the angle of repose would be the same on an asteroid.

Chapter 9

5. Each cyclist has a radial acceleration of v^2/r=5 m/s^2. The tangential accelerations of cyclists A and B are 375 N/75 kg=5 m/s^2.



6. (a) The inward normal force must be sufficient to produce circular motion, so

$$F_N = mv^2 / r \quad .$$

We are searching for the minimum speed, which is the speed at which the static friction force is just barely able to cancel out the downward gravitational force. The maximum force of static friction is

$$|F_s| = \mu_s F_N \quad ,$$

and this cancels the gravitational force, so

$$|F_s| = mg \quad .$$

Solving these three equations for v gives

$$v = \sqrt{\frac{gr}{\mu_s}} \quad .$$

(b) Greater by a factor of $\sqrt{3}$.

7. The inward force must be supplied by the inward component of the normal force,

$$F_N \sin\theta = mv^2/r \quad .$$

The upward component of the normal force must cancel the downward force of gravity,

$$F_N \cos\theta = mg \quad .$$

Eliminating F_N and solving for θ, we find

$$\theta = \tan^{-1}\left(\frac{v^2}{gr}\right) \quad .$$

Chapter 10

10. Newton's law of gravity tells us that her weight will be 6000 times smaller because of the asteroid's smaller mass, but 13^2=169 times greater because of its smaller radius. Putting these two factors together gives a reduction in weight by a factor of 6000/169, so her weight will be (400 N)(169)/(6000)=11 N.

11. Newton's law of gravity says $F=Gm_1m_2/r^2$, and Newton's second law says $F=m_2a$, so $Gm_1m_2/r_2=m_2a$. Since m_2 cancels, a is independent of m_2.

12. Newton's second law gives

$$F = m_D a_D \quad ,$$

where F is Ida's force on Dactyl. Using Newton's universal law of gravity, $F = Gm_I m_D/r^2$,and the equation $a = v^2/r$ for circular motion, we find

$$Gm_I m_D/r^2 = m_D v^2/r \quad .$$

Dactyl's mass cancels out, giving

$$Gm_I/r^2 = v^2/r \quad .$$

Dactyl's velocity equals the circumference of its orbit divided by the time for one orbit: $v=2\pi r/T$. Inserting this in the above equation and solving for m_I, we find

$$m_I = \frac{4\pi^2 r^3}{GT^2} \quad ,$$

so Ida's density is

$$\begin{aligned}\rho &= m_I/\mathrm{V}\\ &= \frac{4\pi^2 r^3}{GVT^2} \quad .\end{aligned}$$

15. Newton's law of gravity depends on the inverse square of the distance, so if the two planets' masses had been equal, then the factor of 0.83/0.059=14 in distance would have caused the force on planet c to be $14^2=2.0\times10^2$ times weaker. However, planet c's mass is 3.0 times greater, so the force on it is only smaller by a factor of $2.0\times10^2/3.0$=65.

16. The reasoning is reminiscent of section 10.2. From Newton's second law we have $F=ma=mv^2/r=m(2\pi r/T)^2/\mathrm{r} = 4\pi^2 mr/T^2$,and Newton's law of gravity gives $F=GMm/r^2$, where M is the mass of the earth.Setting these expressions equal to each other, we have

$$4\pi^2 mr/T^2 = GMm/r^2 \quad ,$$

which gives

$$\begin{aligned}r &= \sqrt[3]{\frac{GMT^2}{4\pi^2}}\\ &= 4.22\times10^4\ \text{km} \quad .\end{aligned}$$

This is the distance from the center of the earth, so to find the altitude, we need to subtract the radius of the earth. The altitude is 3.58×10^4 km.

17. Any fractional change in r results in double that amount of fractional change in $1/r^2$. For example, raising r by 1% causes $1/r^2$ to go down by very nearly 2%. The fractional change in $1/r^2$ is actually

$$2\times\frac{(1/27)\ \text{cm}}{3.84\times10^5\ \text{km}}\times\frac{1\ \text{km}}{10^5\ \text{cm}} = 2\times 10^{-12}$$

Glossary

Acceleration. The rate of change of velocity; the slope of the tangent line on a *v-t* graph.

Attractive. Describes a force that tends to pull the two participating objects together. Cf. repulsive, oblique.

Center of mass. The balance point of an object.

Component. The part of a velocity, acceleration, or force that is along one particular coordinate axis.

Displacement. (avoided in this book) A name for the symbol Δx .

Fluid. A gas or a liquid.

Fluid friction. A friction force in which at least one of the object is is a fluid (i.e. either a gas or a liquid).

Gravity. A general term for the phenomenon of attraction between things having mass. The attraction between our planet and a human-sized object causes the object to fall.

Inertial frame. A frame of reference that is not accelerating, one in which Newton's first law is true

Kinetic friction. A friction force between surfaces that are slipping past each other.

Light. Anything that can travel from one place to another through empty space and can influence matter, but is not affected by gravity.

Magnitude. The "amount" associated with a vector; the vector stripped of any information about its direction.

Mass. A numerical measure of how difficult it is to change an object's motion.

Matter. Anything that is affected by gravity.

Mks system. The use of metric units based on the meter, kilogram, and second. Example: meters per second is the mks unit of speed, not cm/s or km/hr.

Noninertial frame. An accelerating frame of reference, in which Newton's first law is violated

Nonuniform circular motion. Circular motion in which the magnitude of the velocity vector changes

Normal force. The force that keeps two objects from occupying the same space.

Oblique. Describes a force that acts at some other angle, one that is not a direct repulsion or attraction. Cf. attractive, repulsive.

Operational definition. A definition that states what operations should be carried out to measure the thing being defined.

Parabola. The mathematical curve whose graph has *y* proportional to x^2.

Radial. Parallel to the radius of a circle; the in-out direction. Cf. tangential.

Repulsive. Describes a force that tends to push the two participating objects apart. Cf. attractive, oblique.

Scalar. A quantity that has no direction in space, only an amount. Cf. vector.

Significant figures. Digits that contribute to the accuracy of a measurement.

Speed. (avoided in this book) The absolute value of or, in more then one dimension, the magnitude of the velocity, i.e. the velocity stripped of any information about its direction

Spring constant. The constant of proportionality between force and elongation of a spring or other object under strain.

Static friction. A friction force between surfaces that are not slipping past each other.

Système International.. Fancy name for the metric system.

Tangential. Tangent to a curve. In circular motion, used to mean tangent to the circle, perpendicular to the radial direction Cf. radial.

Uniform circular motion. Circular motion in which the magnitude of the velocity vector remains constant

Vector. A quantity that has both an amount (magnitude) and a direction in space. Cf. scalar.

Velocity. The rate of change of position; the slope of the tangent line on an *x-t* graph.

Weight. The force of gravity on an object, equal to *mg*.

Mathematical Review

Algebra

Quadratic equation:

The solutions of $ax^2 + bx + c = 0$

are $x = \dfrac{-b \pm \sqrt{b^2 - 4ac}}{2a}$.

Logarithms and exponentials:

$$\ln (ab) = \ln a + \ln b$$

$$e^{a+b} = e^a e^b$$

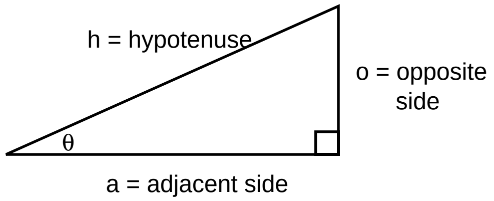
$$\ln e^x = e^{\ln x} = x$$

$$\ln \left(a^b\right) = b \ln a$$

Geometry, area, and volume

area of a triangle of base b and height h	$= \frac{1}{2}bh$
circumference of a circle of radius r	$= 2\pi r$
area of a circle of radius r	$= \pi r^2$
surface area of a sphere of radius r	$= 4\pi r^2$
volume of a sphere of radius r	$= \frac{4}{3}\pi r^3$

Trigonometry with a right triangle



Definitions of the sine, cosine, and tangent:

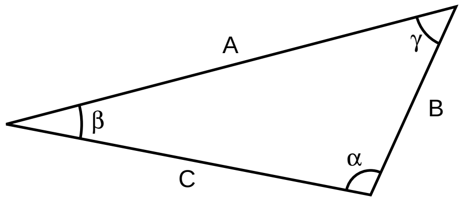
$$\sin\theta = \frac{o}{h}$$

$$\cos\theta = \frac{a}{h}$$

$$\tan\theta = \frac{o}{a}$$

Pythagorean theorem: $h^2 = a^2 + o^2$

Trigonometry with any triangle



Law of Sines:

$$\frac{\sin\alpha}{A} = \frac{\sin\beta}{B} = \frac{\sin\gamma}{C}$$

Law of Cosines:

$$C^2 = A^2 + B^2 - 2AB\cos\gamma$$

Properties of the derivative and integral (for students in calculus-based courses)

Let f and g be functions of x, and let c be a constant.

Linearity of the derivative:

$$\frac{d}{dx}\left(c\,f\right) = c\frac{df}{dx}$$

$$\frac{d}{dx}\left(f + g\right) = \frac{df}{dx} + \frac{dg}{dx}$$

The chain rule:

$$\frac{d}{dx}f(g(x)) = f'(g(x))g'(x)$$

Derivatives of products and quotients:

$$\frac{d}{dx}\left(fg\right) = \frac{df}{dx}g + \frac{dg}{dx}f$$

$$\frac{d}{dx}\left(\frac{f}{g}\right) = \frac{f'}{g} - \frac{fg'}{g^2}$$

Some derivatives:

$$\frac{d}{dx}x^m = mx^{m-1} \quad \text{(except for } m=0\text{)}$$

$$\frac{d}{dx}\sin x = \cos x$$

$$\frac{d}{dx}\cos x = -\sin x$$

$$\frac{d}{dx}e^x = e^x$$

$$\frac{d}{dx}\ln x = \frac{1}{x}$$

The fundamental theorem of calculus:

$$\int \frac{df}{dx}dx = f$$

Linearity of the integral:

$$\int cf(x)dx = c\int f(x)dx$$

$$\int \left[f(x) + g(x)\right]dx = \int f(x)dx + \int g(x)dx$$

Integration by parts:

$$\int f\,dg = fg - \int g\,df$$

Trig Tables

θ	sin θ	cos θ	tan θ
0°	0.000	1.000	0.000
1	0.017	1.000	0.017
2	0.035	0.999	0.035
3	0.052	0.999	0.052
4	0.070	0.998	0.070
5	0.087	0.996	0.087
6	0.105	0.995	0.105
7	0.122	0.993	0.123
8	0.139	0.990	0.141
9	0.156	0.988	0.158
10	0.174	0.985	0.176
11	0.191	0.982	0.194
12	0.208	0.978	0.213
13	0.225	0.974	0.231
14	0.242	0.970	0.249
15	0.259	0.966	0.268
16	0.276	0.961	0.287
17	0.292	0.956	0.306
18	0.309	0.951	0.325
19	0.326	0.946	0.344
20	0.342	0.940	0.364
21	0.358	0.934	0.384
22	0.375	0.927	0.404
23	0.391	0.921	0.424
24	0.407	0.914	0.445
25	0.423	0.906	0.466
26	0.438	0.899	0.488
27	0.454	0.891	0.510
28	0.469	0.883	0.532
29	0.485	0.875	0.554
30°	0.500	0.866	0.577
31	0.515	0.857	0.601
32	0.530	0.848	0.625
33	0.545	0.839	0.649
34	0.559	0.829	0.675
35	0.574	0.819	0.700
36	0.588	0.809	0.727
37	0.602	0.799	0.754
38	0.616	0.788	0.781
39	0.629	0.777	0.810
40	0.643	0.766	0.839
41	0.656	0.755	0.869
42	0.669	0.743	0.900
43	0.682	0.731	0.933
44	0.695	0.719	0.966
45	0.707	0.707	1.000
46	0.719	0.695	1.036
47	0.731	0.682	1.072
48	0.743	0.669	1.111
49	0.755	0.656	1.150
50	0.766	0.643	1.192
51	0.777	0.629	1.235
52	0.788	0.616	1.280
53	0.799	0.602	1.327
54	0.809	0.588	1.376
55	0.819	0.574	1.428
56	0.829	0.559	1.483
57	0.839	0.545	1.540
58	0.848	0.530	1.600
59	0.857	0.515	1.664
60°	0.866	0.500	1.732
61	0.875	0.485	1.804
62	0.883	0.469	1.881
63	0.891	0.454	1.963
64	0.899	0.438	2.050
65	0.906	0.423	2.145
66	0.914	0.407	2.246
67	0.921	0.391	2.356
68	0.927	0.375	2.475
69	0.934	0.358	2.605
70	0.940	0.342	2.747
71	0.946	0.326	2.904
72	0.951	0.309	3.078
73	0.956	0.292	3.271
74	0.961	0.276	3.487
75	0.966	0.259	3.732
76	0.970	0.242	4.011
77	0.974	0.225	4.331
78	0.978	0.208	4.705
79	0.982	0.191	5.145
80	0.985	0.174	5.671
81	0.988	0.156	6.314
82	0.990	0.139	7.115
83	0.993	0.122	8.144
84	0.995	0.105	9.514
85	0.996	0.087	11.430
86	0.998	0.070	14.301
87	0.999	0.052	19.081
88	0.999	0.035	28.636
89	1.000	0.017	57.290
90	1.000	0.000	∞

Index

Photo Credits

All photographs are by Benjamin Crowell, except as noted below.

Cover
Moon: Loewy and Puiseux, 1894.

Chapter 1
Mars Climate Orbiter: NASA/JPL/Caltech. *Red blood cell:* C. Magowan et al.

Chapter 2
High jumper: Dunia Young. *Rocket sled:* U.S. Air Force.

Chapter 3
X-33 art: NASA*. Astronauts and International Space Station:* NASA.
Gravity map: Data from US Navy Geosat and European Space Agency ERS-1 satellites, analyzed by David Sandwell and Walter Smith.

Chapter 4
Isaac Newton: Painting by Sir Godfrey Kneller, National Portrait Gallery, London.

Chapter 5
Space shuttle launch: NASA.

Chapter 6
The Ring Toss: Clarence White, ca. 1903.

Chapter 7
Aerial photo of Mondavi vineyards: NASA.

Chapter 8
Galloping horse: Eadweard Muybridge, 1878.

Useful Data

Metric Prefixes

M-	mega-	10^{6}
k-	kilo-	10^{3}
m-	milli-	10^{-3}
μ- (Greek mu)	micro-	10^{-6}
n-	nano-	10^{-9}

(Centi-, 10^{-2}, is used only in the centimeter.)

Notation and Units

quantity	**unit**	**symbol**
distance	meter, m	x, Δx
time	second, s	t, Δt
mass	kilogram, kg	m
area	m^2 (square meters)	A
volume	m^3 (cubic meters)	V
density	kg/m^3	ρ
force	Newton, 1 N=1 kg·m/s^2	***F***
velocity	m/s	***v***
acceleration	m/s^2	***a***

symbol	**meaning**
∝	is proportional to
≈	is approximately equal to
~	on the order of

The Greek Alphabet

α	Α	alpha	ν	Ν	nu
β	Β	beta	ξ	Ξ	xi
γ	Γ	gamma	ο	Ο	omicron
δ	Δ	delta	π	Π	pi
ε	Ε	epsilon	ρ	Ρ	rho
ζ	Ζ	zeta	σ	Σ	sigma
η	Η	eta	τ	Τ	tau
θ	Θ	theta	υ	Υ	upsilon
ι	Ι	iota	ϕ	Φ	phi
κ	Κ	kappa	χ	Χ	chi
λ	Λ	lambda	ψ	Ψ	psi
μ	Μ	mu	ω	Ω	omega

Conversions

Conversions between SI and other units:

1 inch	=	2.54 cm (exactly)
1 mile	=	1.61 km
1 pound	=	4.45 N
(1 kg)(g)	=	2.2 lb
1 gallon	=	3.78x10^3 cm^3

Conversions between U.S. units:

1 foot	=	12 inches
1 yard	=	3 feet
1 mile	=	5280 ft

Earth, Moon, and Sun

body	**mass (kg)**	**radius (km)**	**radius of orbit (km)**
earth	5.97x10^{24}	6.4x10^3	1.49x10^8
moon	7.35x10^{22}	1.7x10^3	3.84x10^5
sun	1.99x10^{30}	7.0x10^5	

The radii and radii of orbits are average values. The moon orbits the earth and the earth orbits the sun.

Subatomic Particles

particle	**mass (kg)**	**radius (m)**
electron	9.109x10^{-31}	? – less than about 10^{-17}
proton	1.673x10^{-27}	about 1.1x10^{-15}
neutron	1.675x10^{-27}	about 1.1x10^{-15}

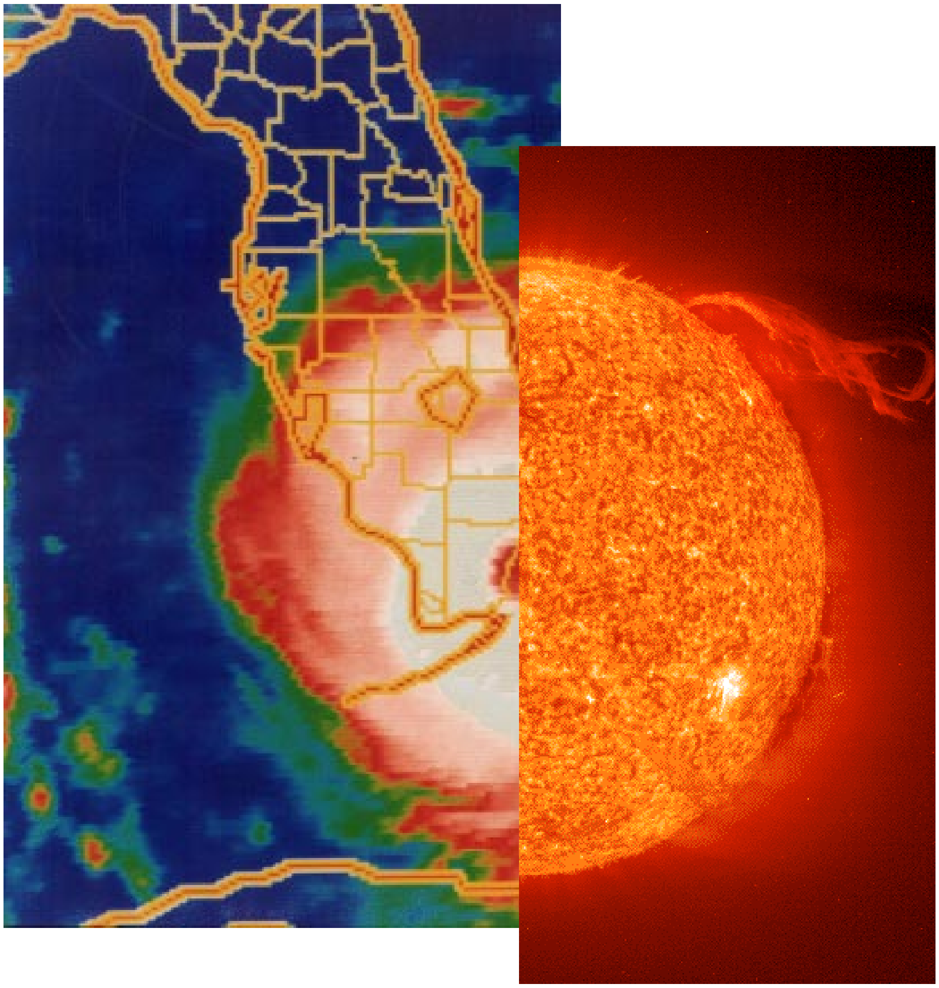
The radii of protons and neutrons can only be given approximately, since they have fuzzy surfaces. For comparison, a typical atom is about 10^{-9} m in radius.

Fundamental Constants

speed of light	c=3.00x10^8 m/s
gravitational constant	G=6.67x10^{-11} N·m^2·kg^{-2}

Conservation Laws

Benjamin Crowell



Book 2 in the **Light and Matter** series of introductory physics textbooks
www.lightandmatter.com

Conservation Laws

The **Light and Matter** series of introductory physics textbooks:

1 Newtonian Physics

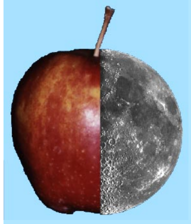
2 Conservation Laws

3 Vibrations and Waves

4 Electricity and Magnetism

5 Optics

6 The Modern Revolution in Physics



Conservation Laws

Benjamin Crowell

www.lightandmatter.com

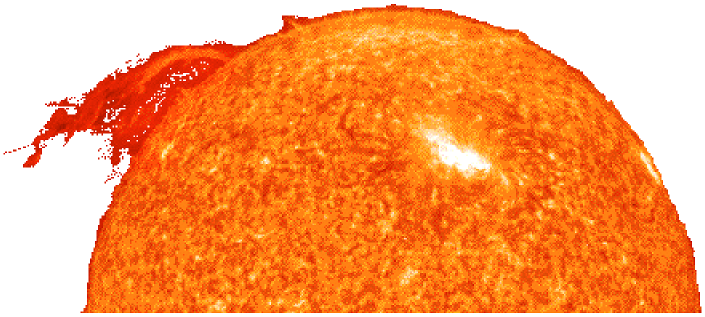
Light and Matter

Fullerton, California
www.lightandmatter.com

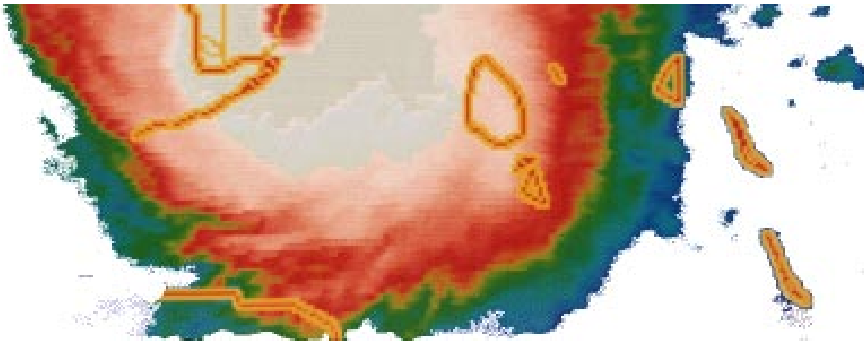
Edition 2.0
rev. 2001-04-17

ISBN 0-9704670-2-8

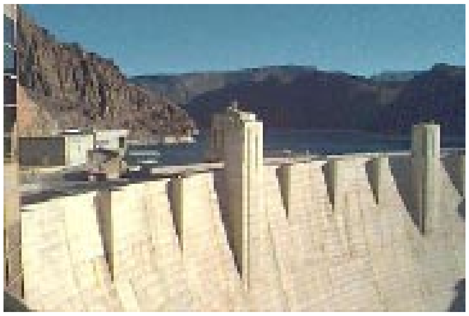
To Uri Haber-Schaim, John Dodge, Robert Gardner, and Edward Shore.



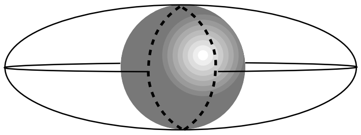
Brief Contents

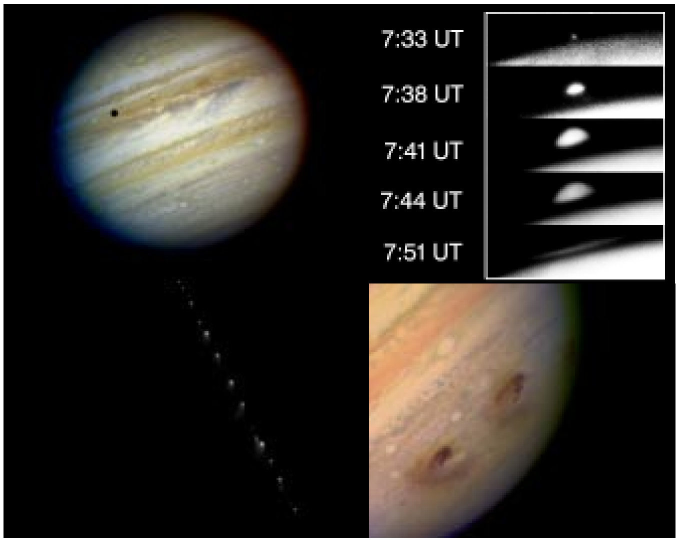


Contents



Supplements 2-6 and 2-7 are available online in electronic format at www.lightandmatter.com.





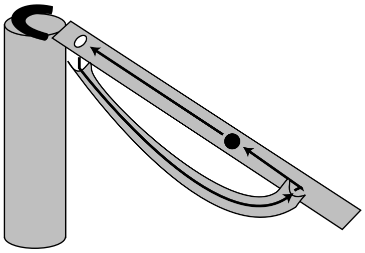
In July of 1994, Comet Shoemaker-Levy struck the planet Jupiter, depositing 7x10 ^{22} joules of energy, and incidentally giving rise to a series of Hollywood movies in which our own planet is threatened by an impact by a comet or asteroid. There is evidence that such an impact caused the extinction of the dinosaurs. Left: Jupiter's gravitational force on the near side of the comet was greater than on the far side, and this difference in force tore up the comet into a string of fragments. Two separate telescope images have been combined to create the illusion of a point of view just behind the comet. (The colored fringes at the edges of Jupiter are artifacts of the imaging system.) Top: A series of images of the plume of superheated gas kicked up by the impact of one of the fragments. The plume is about the size of North America. Bottom: An image after all the impacts were over, showing the damage done.

1 Conservation of Energy

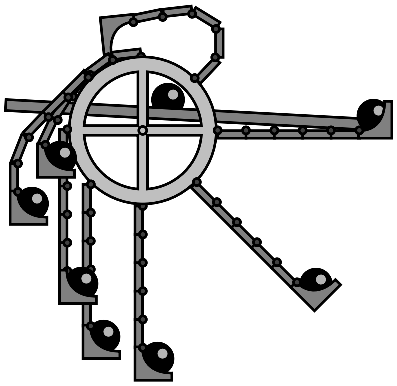
1.1 The Search for a Perpetual Motion Machine

Don't underestimate greed and laziness as forces for progress. Modern chemistry was born from the collision of lust for gold with distaste for the hard work of finding it and digging it up. Failed efforts by generations of alchemists to turn lead into gold led finally to the conclusion that it could not be done: certain substances, the chemical elements, are fundamental, and chemical reactions can neither increase nor decrease the amount of an element such as gold.

Now flash forward to the early industrial age. Greed and laziness have created the factory, the train, and the ocean liner, but in each of these is a boiler room where someone gets sweaty shoveling the coal to fuel the steam



(a) The magnet draws the ball to the top of the ramp, where it falls through the hole and rolls back to the bottom.



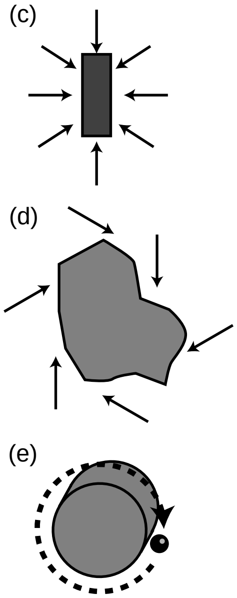
(b) As the wheel spins clockwise, the flexible arms sweep around and bend and unbend. By dropping off its ball on the ramp, the arm is supposed to make itself lighter and easier to lift over the top. Picking its own ball back up again on the right, it helps to pull the right side down.

engine. Generations of inventors have tried to create a machine, called a perpetual motion machine, that would run forever without fuel. Such a machine is not forbidden by Newton's laws of motion, which are built around the concepts of force and inertia. Force is free, and can be multiplied indefinitely with pulleys, gears, or levers. The principle of inertia seems even to encourage the belief that a cleverly constructed machine might not ever run down.

The figures show two of the innumerable perpetual motion machines that have been proposed. The reason these two examples don't work is not much different from the reason all the others have failed. Consider machine (a). Even if we assume that a properly shaped ramp would keep the ball rolling smoothly through each cycle, friction would always be at work. The designer imagined that the machine would repeat the same motion over and over again, so that every time it reached a given point its speed would be exactly the same as the last time. But because of friction, the speed would actually be reduced a little with each cycle, until finally the ball would no longer be able to make it over the top.

Friction has a way of creeping into all moving systems. The rotating earth might seem like a perfect perpetual motion machine, since it is isolated in the vacuum of outer space with nothing to exert frictional forces on it. But in fact our planet's rotation has slowed drastically since it first formed, and the earth continues to slow its rotation, making today just a little longer than yesterday. The very subtle source of friction is the tides. The moon's gravity raises bulges in the earth's oceans, and as the earth rotates the bulges progress around the planet. Where the bulges encounter land, there is friction, which slows the earth's rotation very gradually.

1.2 Energy



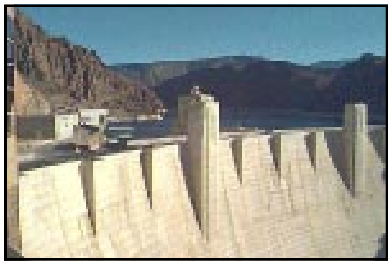
The analysis based on friction is somewhat superficial, however. One could understand friction perfectly well and yet imagine the following situation. Astronauts bring back a piece of magnetic ore from the moon which does not behave like ordinary magnets. A normal bar magnet, (c), attracts a piece of iron essentially directly toward it, and has no left- or right-handedness. The moon rock, however, exerts forces that form a whirlpool pattern around it, (d). NASA goes to a machine shop and has the moon rock put in a lathe and machined down to a smooth cylinder, (e). If we now release a ball bearing on the surface of the cylinder, the magnetic force whips it around and around at ever higher speeds. Of course there is some friction, but there is a net gain in speed with each revolution.

Physicists would lay long odds against the discovery of such a moon rock, not just because it breaks the rules that magnets normally obey but because, like the alchemists, they have discovered a very deep and fundamental principle of nature which forbids certain things from happening. The first alchemist who deserved to be called a chemist was the one who realized one day, "In all these attempts to create gold where there was none before, all I've been doing is shuffling the same atoms back and forth among different test tubes. The only way to increase the amount of gold in my laboratory is to bring some in through the door." It was like having some of your money in a checking account and some in a savings account. Transferring money from one account into the other doesn't change the total amount.

We say that the number of grams of gold is a *conserved* quantity. In this context, the word "conserve" does not have its usual meaning of trying not to waste something. In physics, a conserved quantity is something that you wouldn't be able to get rid of even if you wanted to. Conservation laws in physics always refer to a *closed system*, meaning a region of space with boundaries through which the quantity in question is not passing. In our example, the alchemist's laboratory is a closed system because no gold is coming in or out through the doors.

A similar lightbulb eventually lit up in the heads of the people who had been frustrated trying to build a perpetual motion machine. In perpetual motion machine (b) in the previous section, consider the motion of one of the balls. It performs a cycle of rising and falling. On the way down it gains speed, and coming up it slows back down. Having a greater speed is like having more money in your checking account, and being high up is like having more in your savings account. The device is simply shuffling funds back and forth between the two. Having more balls doesn't change anything fundamentally. Not only that, but friction is always draining off money into a third "bank account:" heat. The reason we rub our hands together when we're cold is that kinetic friction heats things up. The continual buildup in the "heat account" leaves less and less for the "motion account" and "height account," causing the machine eventually to run down.

These insights can be distilled into the following basic principle of physics:



The water behind the Hoover Dam has energy because of its position relative to the planet earth, which is attracting it with a gravitational force. Letting water down to the bottom of the dam converts that energy into energy of motion. When the water reaches the bottom of the dam, it hits turbine blades that drive generators, and its energy of motion is converted into electrical energy.

The Law of Conservation of Energy

It is possible to give a numerical rating, called energy, to the state of a physical system. The total energy is found by adding up contributions coming from characteristics of the system such as motion of objects in it, heating of the objects, and the relative positions of objects that interact via forces. The total energy of a closed system always remains constant. Energy cannot be created or destroyed, but only transferred into or out of a system.

The moon rock story violates conservation of energy because the rock-cylinder and the ball together constitute a closed system. Once the ball has made one revolution around the cylinder, its position relative to the cylinder is exactly the same as before, so the numerical energy rating associated with its position is the same as before. Since the total amount of energy must remain constant, it is impossible for the ball to have a greater speed after one revolution. If it had picked up speed, it would have more energy associated with motion, the same amount of energy associated with position, and a little more energy associated with heating through friction. There cannot be a net increase in energy.

Examples

Dropping a rock: The rock loses energy because of its changing position with respect to the earth. Nearly all that energy is transformed into energy of motion, except for a small amount lost to heat created by air friction.

Sliding in to home base: The runner's energy of motion is nearly all converted into heat via friction with the ground.

Accelerating a car: The gasoline has energy stored in it, which is released as heat by burning it inside the engine. Perhaps 10% of this heat energy is converted into the car's energy of motion. The rest remains in the form of heat, which is carried away by the exhaust.

Cruising in a car: As you cruise at constant speed in your car, all the energy of the burning gas is being converted into heat. The tires and engine get hot, and heat is also dissipated into the air through the radiator and the exhaust.

Stepping on the brakes: All the energy of the car's motion is converted into heat in the brake shoes.

Discussion Question

Hydroelectric power (water flowing over a dam to spin turbines) appears to be completely free. Does this violate conservation of energy? If not, then what is the ultimate source of the electrical energy produced by a hydroelectric plant?

1.3 A Numerical Scale of Energy

Energy comes in a variety of forms, and physicists didn't discover all of them right away. They had to start somewhere, so they picked one form of energy to use as a standard for creating a numerical energy scale. (In fact the history is complicated, and several different energy units were defined before it was realized that there was a single general energy concept that deserved a single consistent unit of measurement.) One practical approach is to define an energy unit based on heating water. The SI unit of energy is the joule, J, (rhymes with "cool"), named after the British physicist James Joule. One Joule is the amount of energy required in order to heat 0.24 g of water by 1°C. The number 0.24 is not worth memorizing.

Note that heat, which is a form of energy, is completely different from temperature, which is not. Twice as much heat energy is required to prepare two cups of coffee as to make one, but two cups of coffee mixed together don't have double the temperature. In other words, the temperature of an object tells us how hot it is, but the heat energy contained in an object also takes into account the object's mass and what it is made of.

Later we will encounter other quantities that are conserved in physics, such as momentum and angular momentum, and the method for defining them will be similar to the one we have used for energy: pick some standard form of it, and then measure other forms by comparison with this standard. The flexible and adaptable nature of this procedure is part of what has made conservation laws such a durable basis for the evolution of physics.

Example: heating a swimming pool

Question: If electricity costs 3.9 cents per MJ (1 MJ = 1 megajoule = 10^6 J), how much does it cost to heat a 26000-gallon swimming pool from 10°C to 18°C?

Solution: Converting gallons to cm^3 gives

$$26000\ \text{gallons} \times \frac{3780\ \text{cm}^3}{1\ \text{gallon}} = 9.8\text{x}10^7\ \text{cm}^3.$$

Water has a density of 1 gram per cubic centimeter, so the mass of the water is $9.8\text{x}10^7$ g. One joule is sufficient to heat 0.24 g by 1°C, so the energy needed to heat the swimming pool is

$$\begin{aligned} 1\ \text{J} \times \frac{9.8\times 10^7\ \text{g}}{0.24\ \text{g}} \times \frac{8°C}{1°C} &= 3.3\text{x}10^9\ \text{J} \\ &= 3.3\text{x}10^3\ \text{MJ}\quad . \end{aligned}$$

The cost of the electricity is $(3.3\text{x}10^3\ \text{MJ})(\$0.039/\text{MJ})=\$130$.

Example: Irish coffee

Question: You make a cup of Irish coffee out of 300 g of coffee at 100°C and 30 g of pure ethyl alcohol at 20°C. One Joule is enough energy to produce a change of 1°C in 0.42 g of ethyl alcohol (i.e. alcohol is easier to heat than water). What temperature is the final mixture?

Solution: Adding up all the energy after mixing has to give the same result as the total before mixing. We let the subscript i stand for the initial situation, before mixing, and f for the final situation, and use subscripts c for the coffee and a for the alcohol. In this notation, we have

$$\begin{aligned} \text{total initial energy} &= \text{total final energy} \\ E_{ci}+E_{ai} &= E_{cf}+E_{af}\quad . \end{aligned}$$

We assume coffee has the same heat-carrying properties as water. Our information about the heat-carrying properties of the two substances is stated in terms of the change in energy required for a certain change in temperature, so we rearrange the equation to express everything in terms of energy differences:

$$E_{af}-E_{ai} = E_{ci}-E_{cf}\quad .$$

Using the given ratios of temperature change to energy change, we have

$$\begin{aligned} E_{ci}-E_{cf} &= (T_{ci}-T_{cf})(m_c)/(0.24\ \text{g}) \\ E_{af}-E_{ai} &= (T_{af}-T_{ai})(m_a)/(0.42\ \text{g}) \end{aligned}$$

Setting these two quantities to be equal, we have

$$(T_{af}-T_{ai})(m_a)/(0.42\ \text{g}) = (T_{ci}-T_{cf})(m_c)/(0.24\ \text{g})\ .$$

In the final mixture the two substances must be at the same temperature, so we can use a single symbol $T_f=T_{cf}=T_{af}$ for the two quantities previously represented by two different symbols,

$$(T_f-T_{ai})(m_a)/(0.42\ \text{g}) = (T_{ci}-T_f)(m_c)/(0.24\ \text{g})\ .$$

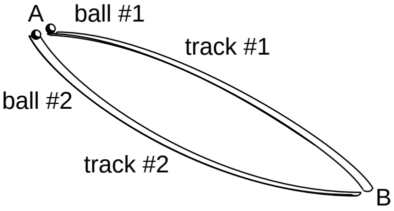
Solving for T_f gives

$$\begin{aligned} T_f &= \frac{T_{ci}\frac{m_c}{.24} + T_{ai}\frac{m_a}{.42}}{\frac{m_c}{.24} + \frac{m_a}{.42}} \\ &= 96°\text{C}. \end{aligned}$$

Once a numerical scale of energy has been established for some form of energy such as heat, it can easily be extended to other types of energy. For instance, the energy stored in one gallon of gasoline can be determined by putting some gasoline and some water in an insulated chamber, igniting the gas, and measuring the rise in the water's temperature. (The fact that the apparatus is known as a "bomb calorimeter" will give you some idea of how dangerous these experiments are if you don't take the right safety precautions.) Here are some examples of other types of energy that can be measured using the same units of joules:

type of energy	**example**
chemical energy released by burning	About 50 MJ are released by burning 1 kg of gasoline.
energy required to break an object	When a person suffers a spiral fracture of the thighbone (a common type in skiing accidents), about 2 J of energy go into breaking the bone.
energy required to melt a solid substance	7 MJ are required to melt 1 kg of tin.
chemical energy released by digesting food	A bowl of Cheerios with milk provides us with about 800 kJ of usable energy.
raising a mass against the force of gravity	Lifting 1.0 kg through a height of 1.0 m requires 9.8 J.
nuclear energy released in fission	1 kg of uranium oxide fuel consumed by a reactor releases 2x10$_{12}$ J of stored nuclear energy.

It is interesting to note the disproportion between the megajoule energies we consume as food and the joule-sized energies we expend in physical activities. If we could perceive the flow of energy around us the way we perceive the flow of water, eating a bowl of cereal would be like swallowing a bathtub's worth of energy, the continual loss of body heat to one's environment would be like an energy-hose left on all day, and lifting a bag of cement would be like flicking it with a few tiny energy-drops. The human body is tremendously inefficient. The calories we "burn" in heavy exercise are almost all dissipated directly as body heat.



Example: You take the high road and I'll take the low road.

Question: The figure shows two ramps which two balls will roll down. Compare their final speeds, when they reach point B. Assume friction is negligible.

Solution: Each ball loses some energy because of its decreasing height above the earth, and conservation of energy says that it must gain an equal amount of energy of motion (minus a little heat created by friction). The balls lose the same amount of height, so their final speeds must be equal.

It's impressive to note the complete impossibility of solving this problem using only Newton's laws. Even if the shape of the track had been given mathematically, it would have been a formidable task to compute the balls' final speed based on vector addition of the normal force and gravitational force at each point along the way.

Forms of Energy Discovered in Recent Times

Einstein showed that mass itself could be converted to and from energy, according to his celebrated equation $E=mc^2$, in which c is the speed of light. We thus speak of mass as simply another form of energy, and it is valid to measure it in units of joules. The mass of a 15-gram pencil corresponds to about 1.3x10^{15} J. The issue is largely academic in the case of the pencil, because very violent processes such as nuclear reactions are required in order to convert any significant fraction of an object's mass into energy. Cosmic rays, however, are continually striking you and your surroundings and converting part of their energy of motion into the mass of newly created particles. A single high-energy cosmic ray can create a "shower" of millions of previously nonexistent particles when it strikes the atmosphere. Einstein's theories are discussed in book 6 of this series.

Even today, when the energy concept is relatively mature and stable, a new form of energy has been proposed based on observations of distant galaxies whose light began its voyage to us billions of years ago. Astronomers have found that the universe's continuing expansion, resulting from the Big Bang, has not been decelerating as rapidly in the last few billion years as would have been expected from gravitational forces. They suggest that a new form of energy may be at work.

How New Forms of Energy Are Discovered

Textbooks often give the impression that a sophisticated physics concept was created by one person who had an inspiration one day, but in reality it is more in the nature of science to rough out an idea and then gradually refine it over many years. The idea of energy was tinkered with from the early 1800s on, and new types of energy kept getting added to the list.

To establish the existence of a new form of energy, a physicist has to

(1) show that it could be converted to and from other forms of energy; and

(2) show that it related to some definite measurable property of the object, for example its temperature, motion, position relative to another object, or being in a solid or liquid state.

For example, energy is released when a piece of iron is soaked in water, so apparently there is some form of energy already stored in the iron. The release of this energy can also be related to a definite measurable property of the chunk of metal: it turns reddish-orange. There has been a chemical change in its physical state, which we call rusting.

Although the list of types of energy kept getting longer and longer, it was clear that many of the types were just variations on a theme. There is an obvious similarity between the energy needed to melt ice and to melt butter, or between the rusting of iron and many other chemical reactions. The topic of the next chapter is how this process of simplification reduced all the types of energy to a very small number (four, according to the way I've chosen to count them).

It might seem that if the principle of conservation of energy ever appeared to be violated, we could fix it up simply by inventing some new type of energy to compensate for the discrepancy. This would be like balancing your checkbook by adding in an imaginary deposit or withdrawal to make your figures agree with the bank's statements. Step (2) above guards against this kind of chicanery. In the 1920s there were experiments that suggested energy was not conserved in radioactive processes. Precise measurements of the energy released in the radioactive decay of a given type of atom showed inconsistent results. One atom might decay and release, say, 1.1x10^{-10} J of energy, which had presumably been stored in some mysterious form in the nucleus. But in a later measurement, an atom of exactly the same type might release 1.2x10^{-10} J. Atoms of the same type are supposed to be identical, so both atoms were thought to have started out with the same energy. If the amount released was random, then apparently the total amount of energy was not the same after the decay as before, i.e. energy was not conserved.

Only later was it found that a previously unknown particle, which is very hard to detect, was being spewed out in the decay. The particle, now called a neutrino, was carrying off some energy, and if this previously unsuspected form of energy was added in, energy was found to be conserved after all. The discovery of the energy discrepancies is seen with hindsight as being step (1) in the establishment of a new form of energy, and the discovery of the neutrino was step (2). But during the decade or so between step (1) and step (2) (the accumulation of evidence was gradual), physicists had the admirable honesty to admit that the cherished principle of conservation of energy might have to be discarded.

Self-Check

How would you carry out the two steps given above in order to establish that some form of energy was stored in a stretched or compressed spring?

1.4 Kinetic Energy

The technical term for the energy associated with motion is *kinetic energy*, from the Greek word for motion. (The root is the same as the word "cinema" for motion picture, and in French the term for kinetic energy is énergie cinématique.) To find how much kinetic energy is possessed by a given moving object, we must convert all its kinetic energy into heat energy, which we have chosen as the standard reference type of energy. We could do this, for example, by firing projectiles into a tank of water and measuring the increase in temperature of the water as a function of the projectile's mass and velocity. Consider the following data from a series of three such experiments:

m (kg)	**v** (m/s)	**energy** (J)
1.00	1.00	0.50
1.00	2.00	2.00
2.00	1.00	1.00

Comparing the first experiment with the second, we see that doubling the

(1) A spring-loaded toy gun can cause a bullet to move, so the spring is capable of storing energy and then converting it into kinetic energy. (2) The amount of energy stored in the spring relates to amount of compression, which can be measured with a ruler.

object's velocity doesn't just double its energy, it quadruples it. If we compare the first and third lines, however, we find that doubling the mass only doubles the energy. This suggests that kinetic energy is proportional to mass and to the square of velocity, $KE \propto mv^2$, and further experiments of this type would indeed establish such a general rule. The proportionality factor equals 0.5 because of the design of the metric system, so the kinetic energy of a moving object is given by

$$KE \quad = \quad \frac{1}{2}mv^2 \quad .$$

The metric system is based on the meter, kilogram, and second, with other units being derived from those. Comparing the units on the left and right sides of the equation shows that the joule can be reexpressed in terms of the basic units as $kg{\cdot}m^2/s^2$.

Students are often mystified by the occurrence of the factor of 1/2, but it is less obscure than it looks. The metric system was designed so that some of the equations relating to energy would come out looking simple, at the expense of some others, which had to have inconvenient conversion factors in front. If we were using the old British Engineering System of units in this course, then we'd have the British Thermal Unit (BTU) as our unit of energy. In that system, the equation you'd learn for kinetic energy would have an inconvenient proportionality constant, $KE=(1.29\text{x}10^{-3})mv^2$, with KE measured in units of BTUs, v measured in feet per second, and so on. At the expense of this inconvenient equation for kinetic energy, the designers of the British Engineering System got a simple rule for calculating the energy required to heat water: one BTU per degree Fahrenheit per gallon. The inventor of kinetic energy, Thomas Young, actually defined it as $KE=mv^2$, which meant that all his other equations had to be different from ours by a factor of two. All these systems of units work just fine as long as they are not combined with one another in an inconsistent way.

Example: energy released by a comet impact

Question: Comet Shoemaker-Levy, which struck the planet Jupiter in 1994, had a mass of roughly $4\text{x}10^{13}$ kg, and was moving at a speed of 60 km/s. Compare the kinetic energy released in the impact to the total energy in the world's nuclear arsenals, which is $2\text{x}10^{19}$ J. Assume for the sake of simplicity that Jupiter was at rest.

Solution: Since we assume Jupiter was at rest, we can imagine that the comet stopped completely on impact, and 100% of its kinetic energy was converted to heat and sound. We first convert the speed to mks units, $v=6\text{x}10^4$ m/s, and then plug in to the equation $KE = \frac{1}{2}mv^2$ to find that the comet's kinetic energy was roughly $7\text{x}10^{22}$ J, or about 3000 times the energy in the world's nuclear arsenals.

Is there any way to derive the equation $KE=\frac{1}{2}mv^2$ mathematically from first principles? No, it is purely empirical. The factor of 1/2 in front is definitely not derivable, since it is different in different systems of units. The proportionality to v^2 is not even quite correct; experiments have shown deviations from the v^2 rule at high speeds, an effect that is related to Einstein's theory of relativity. Only the proportionality to *m* is inevitable. The whole energy concept is based on the idea that we add up energy contributions from all the objects within a system. Based on this philosophy, it is logically necessary that a 2-kg object moving at 1 m/s have the same kinetic energy as two 1-kg objects moving side-by-side at the same speed.

Energy and relative motion

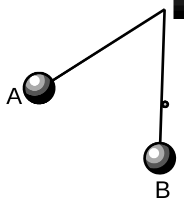
Although I mentioned Einstein's theory of relativity above, it's more relevant right now to consider how conservation of energy relates to the simpler Galilean idea, which we've already studied, that motion is relative. Galileo's Aristotelian enemies (and it is no exaggeration to call them enemies!) would probably have objected to conservation of energy. After all, the Galilean idea that an object in motion will continue in motion indefinitely in the absence of a force is not so different from the idea that an object's kinetic energy stays the same unless there is a mechanism like frictional heating for converting that energy into some other form.

More subtly, however, it's not immediately obvious that what we've learned so far about energy is strictly mathematically consistent with the principle that motion is relative. Suppose we verify that a certain process, say the collision of two pool balls, conserves energy as measured in a certain frame of reference: the sum of the balls' kinetic energies before the collision is equal to their sum after the collision. (In reality we'd need to add in other forms of energy, like heat and sound, that are liberated by the collision, but let's keep it simple.) But what if we were to measure everything in a frame of reference that was in a different state of motion? A particular pool ball might have less kinetic energy in this new frame; for example, if the new frame of reference was moving right along with it, its kinetic energy in that frame would be zero. On the other hand, some other balls might have a greater kinetic energy in the new frame. It's not immediately obvious that the total energy before the collision will still equal the total energy after the collision. After all, the equation for kinetic energy is fairly complicated, since it involves the square of the velocity, so it would be surprising if everything still worked out in the new frame of reference. It *does* still work out. Homework problem 13 in this chapter gives a simple numerical example, and the general proof is taken up in ch. 4, problem 15 (with the solution given in the back of the book).

Discussion Questions

A. Suppose that, like Young or Einstein, you were trying out different equations for kinetic energy to see if they agreed with the experimental data. Based on the meaning of positive and negative signs of velocity, why would you suspect that a proportionality to mv would be less likely than mv^2?

B. The figure shows a pendulum that is released at A and caught by a peg as it passes through the vertical, B. To what height will the bob rise on the right?



Discussion question B.

1.5 Power

A car may have plenty of energy in its gas tank, but still may not be able to increase its kinetic energy rapidly. A Porsche doesn't necessarily have more energy in its gas tank than a Hyundai, it is just able to transfer it more quickly. The rate of transferring energy from one form to another is called *power*. The definition can be written as an equation,

$$P = \frac{\Delta E}{\Delta t} \quad ,$$

where the use of the delta notation in the symbol ΔE has the usual notation: the final amount of energy in a certain form minus the initial amount that was present in that form. Power has units of J/s, which are abbreviated as watts, W (rhymes with "lots").

If the rate of energy transfer is not constant, the power at any instant can be defined as the slope of the tangent line on a graph of E versus t. Likewise ΔE can be extracted from the area under the P-versus-t curve.

Example: converting kilowatt-hours to joules
Question: The electric company bills you for energy in units of kilowatt-hours (kilowatts multiplied by hours) rather than in SI units of joules. How many joules is a kilowatt-hour?
Solution:
1 kilowatt-hour = (1 kW)(1 hour) = (1000 J/s)(3600 s) = 3.6 MJ.

Example: human wattage
Question: A typical person consumes 2000 kcal of food in a day, and converts nearly all of that directly to heat. Compare the person's heat output to the rate of energy consumption of a 100-watt lightbulb.
Solution: Looking up the conversion factor from calories to joules, we find

$$\Delta E = 2000\ \text{kcal} \times \frac{1000\ \text{cal}}{1\ \text{kcal}} \times \frac{4.18\ \text{J}}{1\ \text{cal}} = 8\text{x}10^6\ \text{J}$$

for our daily energy consumption. Converting the time interval likewise into mks,

$$\Delta t = 1\ \text{day} \times \frac{24\ \text{hours}}{1\ \text{day}} \times \frac{60\ \text{min}}{1\ \text{hour}} \times \frac{60\ \text{s}}{1\ \text{min}} = 9\text{x}10^4\ \text{s} \quad .$$

Dividing, we find that our power dissipated as heat is 90 J/s = 90 W, about the same as a lightbulb.

It is easy to confuse the concepts of force, energy, and power, especially since they are synonyms in ordinary speech. The table on the following page may help to clear this up:

	force	energy	power
conceptual definition	A force is an interaction between two objects that causes a push or a pull. A force can be defined as anything that is capable of changing an object's state of motion	Heating an object, making it move faster, or increasing its distance from another object that is attracting it are all examples of things that would require fuel or physical effort. There is a numerical way of measuring all these kinds of things using a single unit of measurement, and we describe them all as forms of energy.	Power is the rate at which energy is transformed from one form to another or transferred from one object to another.
operational definition	A spring scale can be used to measure force.	If we define a unit of energy as the amount required to heat a certain amount of water by a 1°C, then we can measure any other quantity of energy transferring it into heat in water and measuring the temperature increase.	Measure the change in the amount of some form of energy possessed by an object, and divide by the amount of time required for the change to occur.
scalar or vector?	vector – has a direction in space which is the direction in which it pulls or pushes	scalar – has no direction in space	scalar – has no direction in space
unit	newtons (N)	joules (J)	watts (W) = joules/s
Can it run out? Does it cost money?	No. I don't have to pay a monthly bill for the meganewtons of force required to hold up my house.	Yes. We pay money for gasoline, electrical energy, batteries, etc. because they contain energy.	More power means you are paying money at a higher rate. A 100-W lightbulb costs a certain number of cents per hour.
Can it be a property of an object?	No. A force is a relationship between two interacting objects. A home-run baseball doesn't "have" force.	Yes. What a home-run baseball has is kinetic energy, not force.	Not really. A 100-W lightbulb doesn't "have" 100 W. 100 J/s is the rate at which it converts electrical energy into light.

Summary

Selected Vocabulary

energy .... A numerical scale used to measure the heat, motion, or other properties that would require fuel or physical effort to put into an object; a scalar quantity with units of joules (J).

power .... The rate of transferring energy; a scalar quantity with units of watts (W).

kinetic energy .... The energy an object possesses because of its motion.

heat .... The energy that an object has because of its temperature. Heat is different from temperature because an object with twice as much mass requires twice as much heat to increase its temperature by the same amount.

temperature .... What a thermometer measures. Objects left in contact with each other tend to reach the same temperature. Cf. heat. As discussed in more detail in chapter 2, temperature is essentially a measure of the average kinetic energy per molecule.

Notation

E	energy
J	joules, the SI unit of energy
KE	kinetic energy
P	power
W	watts, the SI unit of power; equivalent to J/s

Other Notation and Terminology to be Aware of

Q *or* ΔQ	the amount of heat transferred into or out of an object
K *or* T	alternative symbols for kinetic energy, used in the scientific literature and in most advanced textbooks

thermal energy .... Careful writers make a distinction between heat and thermal energy, but the distinction is often ignored in casual speech, even among physicists. Properly, thermal energy is used to mean the total amount of energy possessed by an object, while heat indicates the amount of thermal energy transferred in or out. The term heat is used in this book to include both meanings.

Summary

Heating an object, making it move faster, or increasing its distance from another object that is attracting it are all examples of things that would require fuel or physical effort. There is a numerical way of measuring all these kinds of things using a single unit of measurement, and we describe them all as forms of *energy*. The SI unit of energy is the Joule. The reason why energy is a useful and important quantity is that it is always conserved. That is, it cannot be created or destroyed but only transferred between objects or changed from one form to another. Conservation of energy is the most important and broadly applicable of all the laws of physics, more fundamental and general even than Newton's laws of motion.

Heating an object requires a certain amount of energy per degree of temperature and per unit mass, which depends on the substance of which the object consists. Heat and temperature are completely different things. Heat is a form of energy, and its SI unit is the joule (J). Temperature is not a measure of energy. Heating twice as much of something requires twice as much heat, but double the amount of a substance does not have double the temperature.

The energy that an object possesses because of its motion is called kinetic energy. Kinetic energy is related to the mass of the object and the magnitude of its velocity vector by the equation

$$KE = \tfrac{1}{2}mv^2 \quad .$$

Power is the rate at which energy is transformed from one form to another or transferred from one object to another,

$$P = \frac{\Delta E}{\Delta t} \quad .$$

The SI unit of power is the watt (W).

Homework Problems

1. Energy is consumed in melting and evaporation. Explain in terms of conservation of energy why sweating cools your body, even though the sweat is at the same temperature as your body.

2. Can the kinetic energy of an object be negative? Explain.

3. Estimate the kinetic energy of an Olympic sprinter.

4✓. You are driving your car, and you hit a brick wall head on, at full speed. The car has a mass of 1500 kg. The kinetic energy released is a measure of how much destruction will be done to the car and to your body. Calculate the energy released if you are traveling at (a) 40 mi/hr, and again (b) if you're going 80 mi/hr. What is counterintuitive about this, and what implication does this have for driving at high speeds?

5✓. A closed system can be a bad thing — for an astronaut sealed inside a space suit, getting rid of body heat can be difficult. Suppose a 60-kg astronaut is performing vigorous physical activity, expending 200 W of power. If none of the heat can escape from her space suit, how long will it take before her body temperature rises by 6°C (11°F), an amount sufficient to kill her? Assume that the amount of heat required to raise her body temperature by 1°C is the same as it would be for an equal mass of water. Express your answer in units of minutes.

6. All stars, including our sun, show variations in their light output to some degree. Some stars vary their brightness by a factor of two or even more, but our sun has remained relatively steady during the hundred years or so that accurate data have been collected. Nevertheless, it is possible that climate variations such as ice ages are related to long-term irregularities in the sun's light output. If the sun was to increase its light output even slightly, it could melt enough ice at the polar icecaps to flood all the world's coastal cities. The total sunlight that falls on the ice caps amounts to about $1\text{x}10^{16}$ watts. Presently, this heat input to the poles is balanced by the loss of heat via winds, ocean currents, and emission of infrared light, so that there is no net melting or freezing of ice at the poles from year to year. Suppose that the sun changes its light output by some small percentage, but there is no change in the rate of heat loss by the polar caps. Estimate the percentage by which the sun's light output would have to increase in order to melt enough ice to raise the level of the oceans by 10 meters over a period of 10 years. (This would be enough to flood New York, London, and many other cities.) Melting 1 kg of ice requires $3\text{x}10^3$ J.

7S. A bullet flies through the air, passes through a paperback book, and then continues to fly through the air beyond the book. When is there a force? When is there energy?

S A solution is given in the back of the book.	★ A difficult problem.
✓ A computerized answer check is available.	∫ A problem that requires calculus.

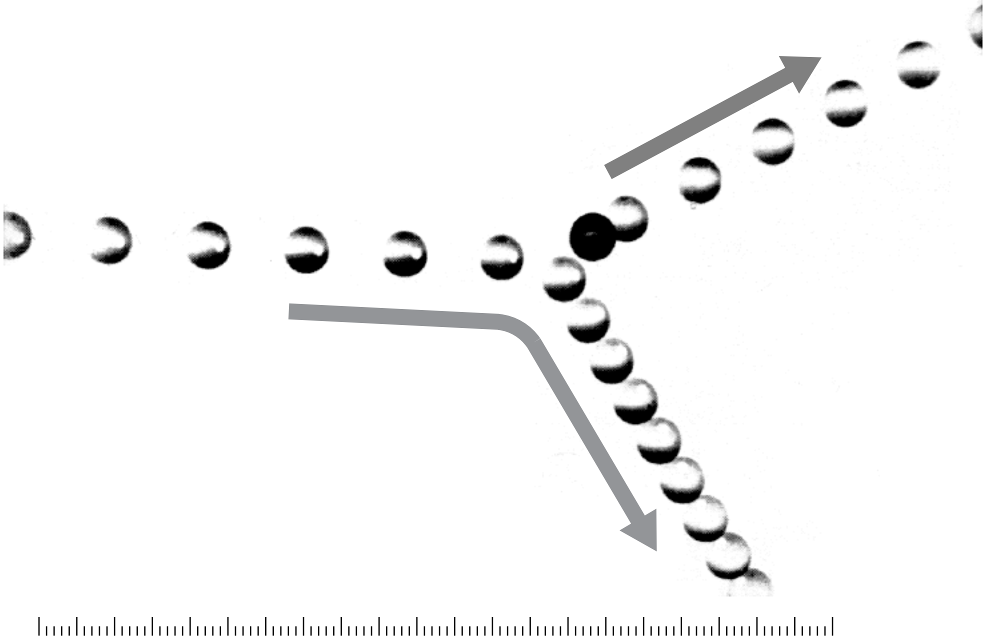
8 S. Experiments show that the power consumed by a boat's engine is approximately proportional to third power of its speed. (We assume that it is moving at constant speed.) (a) When a boat is crusing at constant speed, what type of energy transformation do you think is being performed? (b) If you upgrade to a motor with double the power, by what factor is your boat's crusing speed increased?

9 S. Object A has a kinetic energy of 13.4 J. Object B has a mass that is greater by a factor of 3.77, but is moving more slowly by a factor of 2.34. What is object B's kinetic energy?

10. The moon doesn't really just orbit the Earth. By Newton's third law, the moon's gravitational force on the earth is the same as the earth's force on the moon, and the earth must respond to the moon's force by accelerating. If we consider the earth in moon in isolation and ignore outside forces, then Newton's first law says their common center of mass doesn't accelerate, i.e. the earth wobbles around the center of mass of the earth-moon system once per month, and the moon also orbits around this point. The moon's mass is 81 times smaller than the earth's. Compare the kinetic energies of the earth and moon.

11 S. My 1.25 kW microwave oven takes 126 seconds to bring 250 g of water from room temperature to a boil. What percentage of the power is being wasted? Where might the rest of the energy be going?

12. The multiflash photograph below shows a collision between two pool balls. The ball that was initially at rest shows up as a dark image in its initial position, because its image was exposed several times before it was struck and began moving. By making measurements on the figure, determine whether or not energy appears to have been conserved in the collision. What systematic effects would limit the accuracy of your test? [From an example in PSSC Physics.]



13. This problem is a numerical example of the imaginary experiment discussed at the end of section 1.4 regarding the relationship between energy and relative motion. Let's say that the pool balls both have masses of 1.00 kg. Suppose that in the frame of reference of the pool table, the cue ball moves at a speed of 1.00 m/s toward the eight ball, which is initially at rest. The collision is head-on, and as you can verify for yourself the next time you're playing pool, the result of such a collision is that the incoming ball stops dead and the ball that was struck takes off with the same speed originally possessed by the incoming ball. (This is actually a bit of an idealization. To keep things simple, we're ignoring the spin of the balls, and we assume that no energy is liberated by the collision as heat or sound.) (a) Calculate the total initial kinetic energy and the total final kinetic energy, and verify that they are equal. (b) Now carry out the whole calculation again in the frame of reference that is moving in the same direction that the cue ball was initially moving, but at a speed of 0.50 m/s. In this frame of reference, both balls have nonzero initial and final velocities, which are different from what they were in the table's frame. [See also homework problem 15 in ch. 4.]

Do these forms of energy have anything in common?

2 Simplifying the Energy Zoo

Variety is the spice of life, not of science. The figure shows a few examples from the bewildering array of forms of energy that surrounds us. The physicist's psyche rebels against the prospect of a long laundry list of types of energy, each of which would require its own equations, concepts, notation, and terminology. The point at which we've arrived in the study of energy is analogous to the period in the 1960's when a half a dozen new subatomic particles were being discovered every year in particle accelerators. It was an embarrassment. Physicists began to speak of the "particle zoo," and it seemed that the subatomic world was distressingly complex. The particle zoo was simplified by the realization that most of the new particles being whipped up were simply clusters of a previously unsuspected set of more fundamental particles (which were whimsically dubbed quarks, a made-up word from a line of poetry by James Joyce, "Three quarks for Master Mark.") The energy zoo can also be simplified, and it is the purpose of this chapter to demonstrate the hidden similarities between forms of energy as seemingly different as heat and motion.

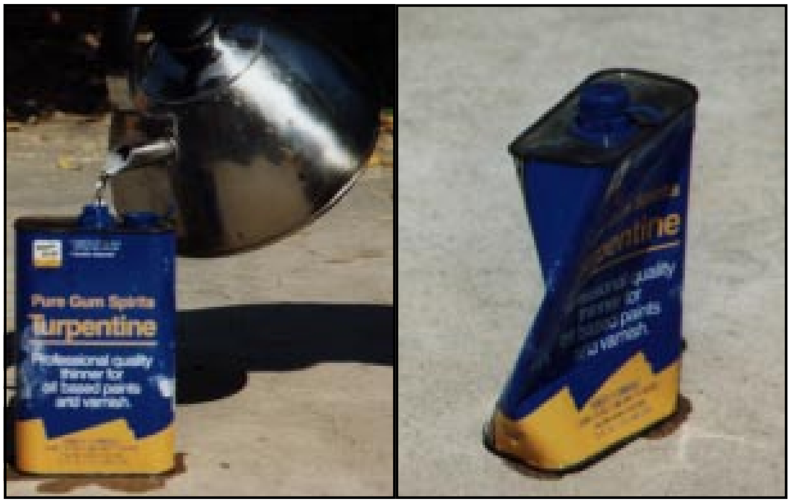
2.1 Heat is Kinetic Energy

What is heat really? Is it an invisible fluid that your bare feet soak up from a hot sidewalk? Can one ever remove all the heat from an object? Is there a maximum to the temperature scale?

The theory of heat as a fluid seemed to explain why colder objects absorbed heat from hotter ones, but once it became clear that heat was a form of energy, it began to seem unlikely that a material substance could transform itself into and out of all those other forms of energy like motion or light. For instance, a compost pile gets hot, and we describe this as a case where, through the action of bacteria, chemical energy stored in the plant cuttings is transformed into heat energy. The heating occurs even if there is no nearby warmer object that could have been leaking “heat fluid” into the pile.

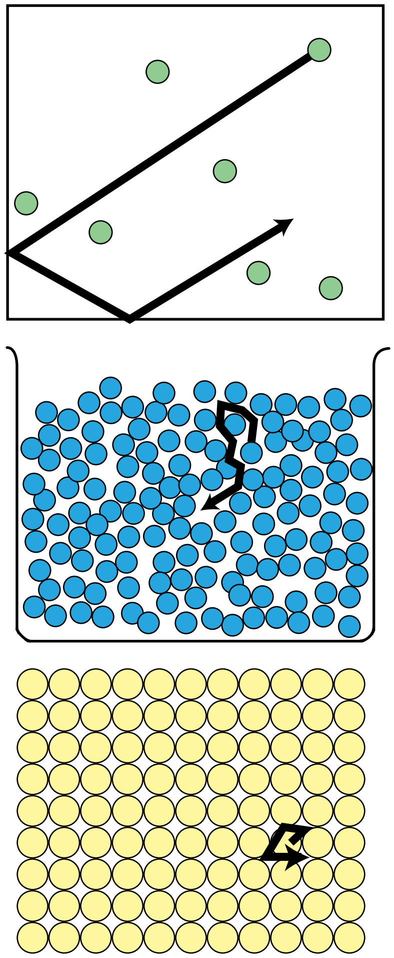
An alternative interpretation of heat was suggested by the theory that matter is made of atoms. Since gases are thousands of times less dense than solids or liquids, the atoms (or clusters of atoms called molecules) in a gas must be far apart. In that case, what is keeping all the air molecules from settling into a thin film on the floor of the room in which you are reading this book? The simplest explanation is that they are moving very rapidly, continually ricocheting off of the floor, walls, and ceiling. Though bizarre, the cloud-of-bullets image of a gas did give a natural explanation for the surprising ability of something as tenuous as a gas to exert huge forces. Your car’s tires can hold it up because you have pumped extra molecules into them. The inside of the tire gets hit by molecules more often than the outside, forcing it to stretch and stiffen.

The outward forces of the air in your car’s tires increase even further when you drive on the freeway for a while, heating up the rubber and the air inside. This type of observation leads naturally to the conclusion that hotter matter differs from colder in that its atoms’ random motion is more


Pure Gum Spirits
Turpentine
Professional quality thinner for oil based paints and varnish.

A vivid demonstration that heat is a form of motion. A small amount of boiling water is poured into the empty can, which rapidly fills up with hot steam. The can is then sealed tightly, and soon crumples. This can be explained as follows. The high temperature of the steam is interpreted as a high average speed of random motions of its molecules. Before the lid was put on the can, the rapidly moving steam molecules pushed their way out of the can, forcing the slower air molecules out of the way. As the steam inside the can thinned out, a stable situation was soon achieved, in which the force from the less dense steam molecules moving at high speed balanced against the force from the more dense but slower air molecules outside. The cap was put on, and after a while the steam inside the can began to cool off. The force from the cooler, thin steam no longer matched the force from the cool, dense air outside, and the imbalance of forces crushed the can.

rapid. In a liquid, the motion could be visualized as people in a milling crowd shoving past each other more quickly. In a solid, where the atoms are packed together, the motion is a random vibration of each atom as it knocks against its neighbors.



Random motion of atoms in a gas, a liquid, and a solid.

We thus achieve a great simplification in the theory of heat. Heat is simply a form of kinetic energy, the total kinetic energy of random motion of all the atoms in an object. With this new understanding, it becomes possible to answer at one stroke the questions posed at the beginning of the section. Yes, it is at least theoretically possible to remove all the heat from an object. The coldest possible temperature, known as absolute zero, is that at which all the atoms have zero velocity, so that their kinetic energies, $KE = \frac{1}{2}mv^2$, are all zero. No, there is no maximum amount of heat that a certain quantity of matter can have, and no maximum to the temperature scale, since arbitrarily large values of v can create arbitrarily large amounts of kinetic energy per atom.

The kinetic theory of heat also provides a simple explanation of the true nature of temperature. Temperature is a measure of the amount of energy per molecule, whereas heat is the total amount of energy possessed by all the molecules in an object.

There is an entire branch of physics, called thermodynamics, that deals with heat and temperature and forms the basis for technologies such as refrigeration. Thermodynamics is discussed in more detail in supplement 2-6, and I have provided here only a brief overview of the thermodynamic concepts that relate directly to energy, glossing over at least one point that would be dealt with more carefully in a thermodynamics course: it is really only true for a gas that all the heat is in the form of kinetic energy. In solids and liquids, the atoms are close enough to each other to exert intense electrical forces on each other, and there is therefore another type of energy involved, the energy associated with the atoms' distances from each other. Strictly speaking, heat energy is defined not as energy associated with random motion of molecules but as any form of energy that can be conducted between objects in contact, without any force.

2.2 Potential Energy: Energy of Distance or Closeness

We have already seen many examples of energy related to the distance between interacting objects. When two objects participate in an attractive noncontact force, energy is required to bring them farther apart. In both of the perpetual motion machines that started off the previous chapter, one of the types of energy involved was the energy associated with the distance between the balls and the earth, which attract each other gravitationally. In the perpetual motion machine with the magnet on the pedestal, there was also energy associated with the distance between the magnet and the iron ball, which were attracting each other.

The opposite happens with repulsive forces: two socks with the same type of static electric charge will repel each other, and cannot be pushed closer together without supplying energy.

In general, the term *potential energy,* with algebra symbol *PE,* is used for the energy associated with the distance between two objects that attract or repel each other via a force that depends on the distance between them. Forces that are not determined by distance do not have potential energy associated with them. For instance, the normal force acts only between objects that have zero distance between them, and depends on other factors besides the fact that the distance is zero. There is no potential energy associated with the normal force.

The skater has converted all his kinetic energy into potential energy on the way up the side of the pool.
Photo by J.D. Rogge, www.sonic.net/~shawn.

The following are some commonplace examples of potential energy:

gravitational potential energy: The skateboarder in the photo has risen from the bottom of the pool, converting kinetic energy into gravitational potential energy. After being at rest for an instant, he will go back down, converting PE back into KE.

magnetic potential energy: When a magnetic compass needle is allowed to rotate, the poles of the compass change their distances from the earth's north and south magnetic poles, converting magnetic potential energy into kinetic energy. (Eventually the kinetic energy is all changed into heat by friction, and the needle settles down in the position that minimizes its potential energy.)

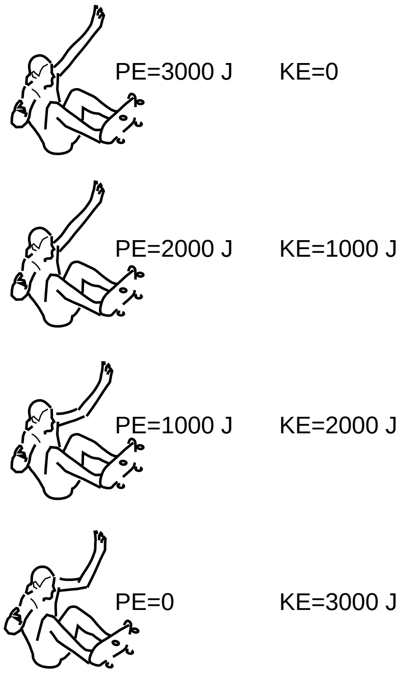
electrical potential energy: Socks coming out of the dryer cling together because of attractive electrical forces. Energy is required in order to separate them.

potential energy of bending or stretching: The force between the two ends of a spring depends on the distance between them, i.e. on the length of the spring. If a car is pressed down on its shock absorbers and then released, the potential energy stored in the spring is transformed into kinetic and gravitational potential energy as the car bounces back up.

I have deliberately avoided introducing the term potential energy up until this point, because it tends to produce unfortunate connotations in the minds of students who have not yet been inoculated with a careful description of the construction of a numerical energy scale. Specifically, there is a tendency to generalize the term inappropriately to apply to any situation where there is the “potential” for something to happen: “I took a break from digging, but I had potential energy because I knew I’d be ready to work hard again in a few minutes.”

An Equation for Gravitational Potential Energy

All the vital points about potential energy can be made by focusing on the example of gravitational potential energy. For simplicity, we treat only vertical motion, and motion close to the surface of the earth, where the gravitational force is nearly constant. (The generalization to the three dimensions and varying forces is more easily accomplished using the concept of work, which is the subject the next chapter.)



As the skater free-falls, his PE is converted into KE. (The numbers would be equally valid as a description of his motion on the way up.)

To find an equation for gravitational PE, we examine the case of free fall, in which energy is transformed between kinetic energy and gravitational PE. Whatever energy is lost in one form is gained in an equal amount in the other form, so using the notation ΔKE to stand for $KE_f - KE_i$ and a similar notation for PE, we have

$$\Delta KE = -\Delta PE_{\text{grav}} \quad . \qquad (1)$$

It will be convenient to refer to the object as falling, so that PE is being changed into KE, but the math applies equally well to an object slowing down on its way up. We know an equation for kinetic energy,

$$KE = \tfrac{1}{2} m v^2 \quad , \qquad (2)$$

so if we can relate v to height, y, we will be able to relate ΔPE to y, which would tell us what we want to know about potential energy. The y component of the velocity can be connected to the height via the constant acceleration equation

$$v_f^2 = v_i^2 + 2a\Delta y \quad , \qquad (3)$$

and Newton’s second law provides the acceleration,

$$a = F/m \quad , \qquad (4)$$

in terms of the gravitational force.

The algebra is simple because both equation (2) and equation (3) have velocity to the second power. Equation (2) can be solved for v^2 to give $v^2 = 2KE/m$, and substituting this into equation (3), we find

$$2KE_f/m = 2KE_i/m + 2a\Delta y \quad .$$

Making use of equations (1) and (4) gives the simple result

$$\Delta PE_{\text{grav}} = -F\Delta y \quad .$$

[change in gravitational PE resulting from a change in height Δy; F is the gravitational force on the object, i.e. its weight; valid only near the surface of the earth, where F is constant]

Example: dropping a rock

Question: If you drop a 1-kg rock from a height of 1 m, how many joules of KE does it have on impact with the ground? (Assume that any energy transformed into heat by air friction is negligible.)

Solution: If we choose the y axis to point up, then F_y is negative, and equals –(1 kg)(g)=-9.8 N. A decrease in y is represented by a negative value of Δy, Δy=–1 m, so the change in potential energy is –(–9.8 N)(–1 m)≈–10 J. (The proof that newtons multiplied by meters give units of joules is left as a homework problem.) Conservation of energy says that the loss of this amount of PE must be accompanied by a corresponding increase in KE of 10 J.

It may be dismaying to note how many minus signs had to be handled correctly even in this relatively simple example: a total of four. Rather than depending on yourself to avoid any mistakes with signs, it is better to check whether the final result make sense physically. If it doesn't, just reverse the sign.

Although the equation for gravitational potential energy was derived by imagining a situation where it was transformed into kinetic energy, the equation can be used in any context, because all the types of energy are freely convertible into each other.

Example: Gravitational PE converted directly into heat

Question: A 50-kg firefighter slides down a 5-m pole at constant velocity. How much heat is produced?

Solution: Since she slides down at constant velocity, there is no change in KE. Heat and gravitational PE are the only forms of energy that change. Ignoring plus and minus signs, the gravitational force on her body equals mg, and the amount of energy transformed is

$$(mg)(5\text{ m}) = 2500\text{ J} \quad .$$

On physical grounds, we know that there must have been an increase (positive change) in the heat energy in her hands and in the flagpole.

Here are some questions and answers about the interpretation of the equation $\Delta PE_{grav} = -F\Delta y$ for gravitational potential energy.

Question: In a nutshell, why is there a minus sign in the equation?
Answer: It is because we increase the PE by moving the object in the *opposite* direction compared to the gravitational force.

Question: Why do we only get an equation for the *change* in potential energy? Don't I really want an equation for the potential energy itself?
Answer: No, you really don't. This relates to a basic fact about potential energy, which is that it is not a well defined quantity in the absolute sense. Only changes in potential energy are unambiguously defined. If you and I both observe a rock falling, and agree that it deposits 10 J of energy in the dirt when it hits, then we will be forced to agree that the 10 J of KE must have come from a loss of 10 joules of PE. But I might claim that it started with 37 J of PE and ended with 27, while you might swear just as truthfully that it had 109 J initially and 99 at the end. It is possible to pick some specific height as a reference level and say that the PE is zero there, but it's easier and safer just to work with changes in PE and avoid absolute PE

altogether.

Question: You referred to potential energy as the energy that *two* objects have because of their distance from each other. If a rock falls, the object is the rock. Where's the other object?
Answer: Newton's third law guarantees that there will always be two objects. The other object is the planet earth.

Question: If the other object is the earth, are we talking about the distance from the rock to the center of the earth or the distance from the rock to the surface of the earth?
Answer: It doesn't matter. All that matters is the change in distance, Δy, not y. Measuring from the earth's center or its surface are just two equally valid choices of a reference point for defining absolute PE.

Question: Which object contains the PE, the rock or the earth?
Answer: We may refer casually to the PE of the rock, but technically the PE is a relationship between the earth and the rock, and we should refer to the earth and the rock together as possessing the PE.

Question: How would this be any different for a force other than gravity?
Answer: It wouldn't. The derivation was derived under the assumption of constant force, but the result would be valid for any other situation where two objects interacted through a constant force. Gravity is unusual, however, in that the gravitational force on an object is so nearly constant under ordinary conditions. The magnetic force between a magnet and a refrigerator, on the other hand, changes drastically with distance. The math is a little more complex for a varying force, but the concepts are the same.

Question: Suppose a pencil is balanced on its tip and then falls over. The pencil is simultaneously changing its height and rotating, so the height change is different for different parts of the object. The bottom of the pencil doesn't lose any height at all. What do you do in this situation?
Answer: The general philosophy of energy is that an object's energy is found by adding up the energy of every little part of it. You could thus add up the changes in potential energy of all the little parts of the pencil to find the total change in potential energy. Luckily there's an easier way! The derivation of the equation for gravitational potential energy used Newton's second law, which deals with the acceleration of the object's center of mass (i.e. its balance point). If you just define Δy as the height change of the center of mass, everything works out. A huge Ferris wheel can be rotated without putting in or taking out any PE, because its center of mass is staying at the same height.

Self-Check

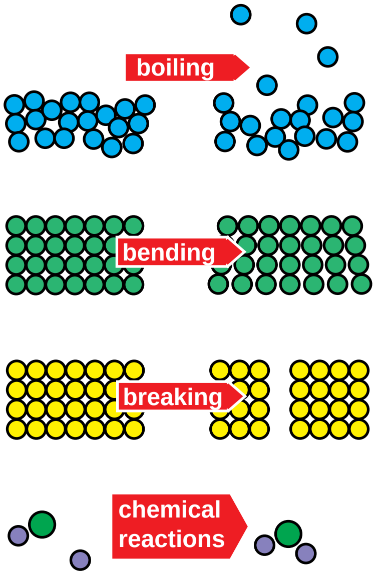
A ball thrown straight up will have the same speed on impact with the ground as a ball thrown straight down at the same speed. How can this be explained using potential energy?

Discussion Question

You throw a steel ball up in the air. How can you prove based on conservation of energy that it has the same speed when it falls back into your hand? What if you threw a feather up? Is energy not conserved in this case?

You throw a feather straight up in the air. When it comes back down to the same height from which you released it, you find that it is going slower. How do you reconcile this with conservation of energy?

2.3 All Energy is Potential or Kinetic

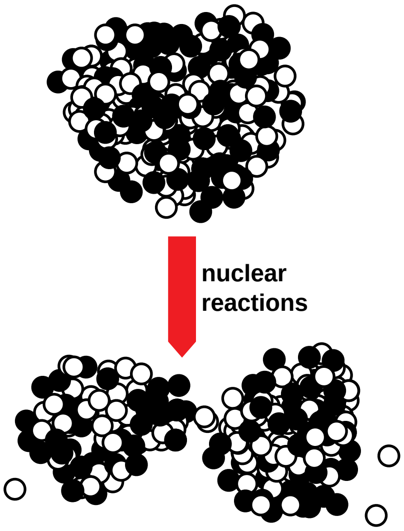


All these energy transformations turn out at the atomic level to be changes in potential energy resulting from changes in the distances between atoms.

In the same way that we found that a change in temperature is really only a change in kinetic energy at the atomic level, we now find that every other form of energy turns out to be a form of potential energy. Boiling, for instance, means knocking some of the atoms (or molecules) out of the liquid and into the space above, where they constitute a gas. There is a net attractive force between essentially any two atoms that are next to each other, which is why matter always prefers to be packed tightly in the solid or liquid state unless we supply enough potential energy to pull it apart into a gas. This explains why water stops getting hotter when it reaches the boiling point: the power being pumped into the water by your stove begins going into potential energy rather than kinetic energy.

As shown in the figure on the left, every stored form of energy that we encounter in everyday life turns out to be a form of potential energy at the atomic level. The forces between atoms are electrical and magnetic in nature, so these are actually electrical and magnetic potential energies.

Both balls start from the same height and end at the same height, so they have the same Δy. This implies that their losses in potential energy are the same, so they must both have gained the same amount of kinetic energy.



This figure looks similar to the previous ones, but the scale is a million times smaller. The little balls are the neutrons and protons that make up the tiny nucleus at the center of the uranium atom. When the nucleus splits (fissions), the potential energy change is partly electrical and partly a change in the potential energy derived from the force that holds atomic nuclei together (known as the strong nuclear force).

Even if we wish to include nuclear reactions in the picture, there still turn out to be only four fundamental types of energy:

kinetic energy (including heat)
gravitational potential energy
electrical and magnetic potential energy
nuclear potential energy

Astute students often ask me how light fits into this picture. This is a very good question, and in fact it could be argued that it is the basic question that led to Einstein's theory of relativity as well as the modern quantum picture of nature. Since these are topics for books 4, 5, and 6 of this series, we will have to be content with half an answer at this point. Essentially we may think of light energy as a form of kinetic energy, but one for which kinetic energy is not given by $\frac{1}{2}mv^2$ but rather by some other equation. (We know that $\frac{1}{2}mv^2$ would not make sense, because light has no mass, and furthermore, high-energy beams of light do not differ in speed from low-energy ones.)

Discussion Question

Referring back to the pictures at the beginning of the chapter, how do all these forms of energy fit into the shortened list of categories given above?

Summary

Selected Vocabulary

potential energy .... the energy having to do with the distance between to objects that interact via a noncontact force

Notation

PE .... potential energy

Alternative Notation to be Aware of

U or V .... symbols used for potential energy in the scientific literature and in most advanced textbooks

Summary

Historically, the energy concept was only invented to include a few phenomena, but it was later generalized more and more to apply to new situations, for example nuclear reactions. This generalizing process resulted in an undesirably long list of types of energy, each of which apparently behaved according to its own rules.

The first step in simplifying the picture came with the realization that heat was a form of random motion on the atomic level, i.e. heat was nothing more than the kinetic energy of atoms.

A second and even greater simplification was achieved with the realization that all the other apparently mysterious forms of energy actually had to do with changing the distances between atoms (or similar processes in nuclei). This type of energy, which relates to the distance between objects that interact via a force, is therefore of great importance. We call it potential energy.

Most of the important ideas about potential energy can be understood by studying the example of gravitational potential energy. The change in an object's gravitational potential energy is given by

$$\Delta PE_{grav} = -F_{grav}\,\Delta y \quad , \qquad \text{[if } F_{grav} \text{ is constant, i.e. the motion is all near the Earth's surface]}$$

The most important thing to understand about potential energy is that there is no unambiguous way to define it in an absolute sense. The only thing that everyone can agree on is how much the potential energy has changed from one moment in time to some later moment in time.

Homework Problems

1. Can the gravitational potential energy of an object ever be negative? Note that the question refers to *PE*, not Δ*PE*, so that you must think about how the choice of a reference level comes into play.

2. A ball is thrown straight up. At what position is its gravitational potential energy at a maximum? At what position is its kinetic energy at a maximum?

3. (a) You release a magnet on a tabletop near a big piece of iron, and the magnet leaps across the table to the iron. Does the magnetic potential energy increase or decrease? Explain. (b) Suppose instead that you have two repelling magnets. You give them an initial push towards each other, so they decelerate while approaching each other. Does the magnetic potential energy increase or decrease? Explain.

4. Let E_b be the energy required to boil one kg of water. (a) Find an equation for the minimum height from which a bucket of water must be dropped if the energy released on impact is to vaporize it. Assume that all the heat goes into the water, not into the dirt it strikes, and ignore the relatively small amount of energy required to heat the water from room temperature to 100°C. [Numerical check, not for credit: Plugging in E_b=2.3 MJ/kg should give a result of 230 km.] (b) Show that the units of your answer in part a come out right based on the units given for E_b.

5 S. A grasshopper with a mass of 110 mg falls from rest from a height of 310 cm. On the way down, it dissipates 1.1 mJ of heat due to air resistance. At what speed, in m/s, does it hit the ground?

6. A person on a bicycle is to coast down a ramp of height *h* and then pass through a circular loop of radius *r*. What is the smallest value of *h* for which the cyclist will complete the loop without falling? (Ignore the kinetic energy of the spinning wheels.)

7★ S. A skateboarder starts at nearly rest at the top of a giant cylinder, and begins rolling down its side. (If she started exactly at rest and exactly at the top, she would never get going!) Show that her board loses contact with the pipe after she has dropped by a height equal to one third the radius of the pipe.

8★. (a) A circular hoop of mass *m* and radius *r* spins like a wheel while its center remains at rest. Its period (time required for one revolution) is *T*. Show that its kinetic energy equals $2\pi^2mr^2/T^2$. (b) If such a hoop rolls with its center moving at velocity *v*, its kinetic energy consists equals $(1/2)mv^2$, plus the amount of kinetic energy found in the first part of this problem. Show that a hoop rolls down an inclined plane with half the acceleration that a frictionless sliding block would have.

9 S. Students are often tempted to think of potential energy and kinetic energy as if they were always related to each other, like yin and yang. To show this is incorrect, give examples of physical situations in which (a) PE is converted to another form of PE, and (b) KE is converted to another form of KE.

S A solution is given in the back of the book.	★ A difficult problem.
✓ A computerized answer check is available.	∫ A problem that requires calculus.

10 ✓. Lord Kelvin, a physicist, told the story of how he encountered James Joule when Joule was on his honeymoon. As he traveled, Joule would stop with his wife at various waterfalls, and measure the difference in temperature between the top of the waterfall and the still water at the bottom. (a) It would surprise most people to learn that the temperature increased. Why should there be any such effect, and why would Joule care? How would this relate to the energy concept, of which he was the principal inventor? (b) How much of a gain in temperature should there be between the top and bottom of a 50-meter waterfall? (c) What assumptions did you have to make in order to calculate your answer to part b? In reality, would the temperature change be more than or less than what you calculated? [Based on a problem by Arnold Arons.]

11 S. Make an order-of-magnitude estimate of the power represented by the loss of gravitational energy of the water going over Niagara Falls. If a hydroelectric plant was built at the bottom of the falls, and could convert 100% of this to electrical power, roughly how many households could be powered?

3 Work: The Transfer of Mechanical Energy

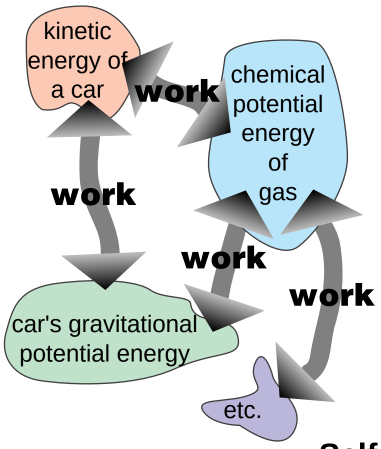
3.1 Work: The Transfer of Mechanical Energy

The concept of work

The mass contained in a closed system is a conserved quantity, but if the system is not closed, we also have ways of measuring the amount of mass that goes in or out. The water company does this with a meter that records your water use.

Likewise, there are many situations in which we would like to know how much energy comes in or out of a system that is not closed. Energy, however, is not a physical substance like water, so energy transfer cannot be measured with the same kind of meter. How can we tell, for instance, how much useful energy a tractor can “put out” on one tank of gas?

The law of conservation of energy guarantees that all the chemical energy in the gasoline will reappear in some form, but not necessarily in a


kinetic energy of a car
work
chemical potential energy of gas
work
work
work
car's gravitational potential energy
etc.

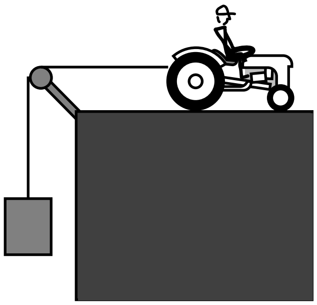
form that is useful for doing farm work. Tractors, like cars, are extremely inefficient, and typically 90% of the energy they consume is converted directly into heat, which is carried away by the exhaust and the air flowing over the radiator. We wish to distinguish the energy that comes out directly as heat from the energy that served to accelerate a trailer or to plow a field, and we define a technical meaning of the ordinary word "work" to express the distinction:

> **definition of work**
> Work is the amount of energy transferred into or out of a system, not counting the energy transferred by heat conduction.

Self-Check

Based on this definition, is work a vector or a scalar? What are its units?

The conduction of heat is to be distinguished from heating by friction. When a hot potato heats up your hands by conduction, the energy transfer occurs without any force, but when friction heats your car's brake shoes, there is a force involved. The transfer of energy with and without a force are measured by completely different methods, so we wish to include heat transfer by frictional heating under the definition of work, but not heat transfer by conduction. The definition of work could thus be restated as the amount of energy transferred by forces.

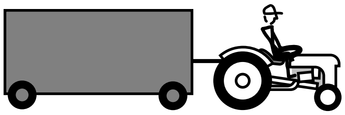


(a) The tractor raises the weight over the pulley, increasing its gravitational potential energy.

Calculating work as force multiplied by distance

The examples in the figures on the left show that there are many different ways in which energy can be transferred. Even so, all these examples have two things in common:

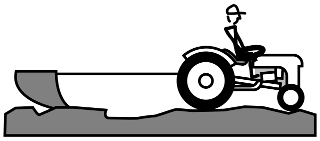
- A force is involved.
- The tractor travels some distance as it does the work.



(b) The tractor accelerates the trailer, increasing its kinetic energy.

In example (a), the increase in the height of the weight, Δy, is the same as the distance the tractor travels, call it d. For simplicity, we discuss the case where the tractor raises the weight at constant speed, so that there is no change in the kinetic energy of the weight, and we assume that there is negligible friction in the pulley, so that the force the tractor applies to the rope is the same as the rope's upward force on the weight. By Newton's first law, these forces are also of the same magnitude as the earth's gravitational force on the weight. The increase in the weight's potential energy is given by $F\Delta y$, so the work done by the tractor on the weight equals Fd, the product of the force and the distance moved:

$$W = Fd \quad .$$



(c) The tractor pulls a plow. Energy is expended in frictional heating of the plow and the dirt, and in breaking dirt clods and lifting dirt up to the sides of the furrow.

In example (b), the tractor's force on the trailer accelerates it, increasing its kinetic energy. If frictional forces on the trailer are negligible, then the increase in the trailer's kinetic energy can be found using the same algebra that was used in the previous chapter to find the potential energy due to gravity. Just as in example (a), we have

$$W = Fd \quad .$$

Work is defined as the transfer of energy, so like energy it is a scalar with units of joules.

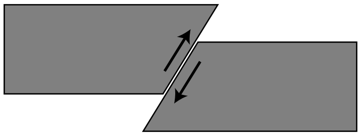
Does this equation always give the right answer? Well, sort of. In example (c), there are two quantities of work you might want to calculate, the work done by the tractor on the plow and the work done by the plow on the dirt. These two quantities can't both equal *Fd*, because the work done by the plow on the dirt is decreased by the heat lost in the plow itself. It turns out that the equation $W=Fd$ gives the work done by the tractor, not the work done by the plow. How are you supposed to know when the equation will work and when it won't? The somewhat complex answer is postponed until section 3.6. Until then, we will restrict ourselves to examples in which $W=Fd$ gives the right answer.

We have also been using examples in which the force is in the same direction as the motion, and the force is constant. (If the force was not constant, we would have to represent it with a function, not a symbol that stands for a number.) To summarize, we have:

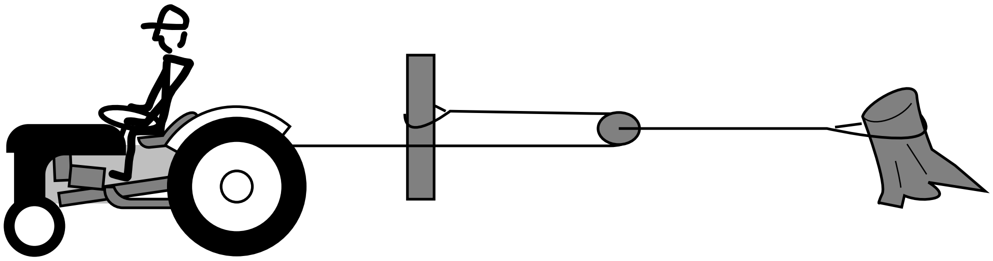
rule for calculating work (simplest version)
The work done by a force can be calculated as

$$W = Fd \quad ,$$

if the force is constant and in the same direction as the motion. Some ambiguities are encountered in cases such as kinetic friction.



Example: mechanical work done in an earthquake
Question: In 1998, geologists discovered evidence for a big prehistoric earthquake in Pasadena, between 10,000 and 15,000 years ago. They found that the two sides of the fault moved 6.7 m relative to one another, and estimated that the force between them was $1.3\text{x}10^{17}$ N. How much energy was released?
Solution: Multiplying the force by the distance gives $9\text{x}10^{17}$ J. For comparison, the Northridge earthquake of 1994, which killed 57 people and did 40 billion dollars of damage, released 22 times less energy.



Machines can increase force, but not work.

The figure above shows a pulley arrangement for doubling the force supplied by the tractor (book 1, section 5.6). The tension in the left-hand rope is equal throughout, assuming negligible friction, so there are two forces pulling the pulley to the left, each equal to the original force exerted by the tractor on the rope. This doubled force is transmitted through the right-hand rope to the stump.

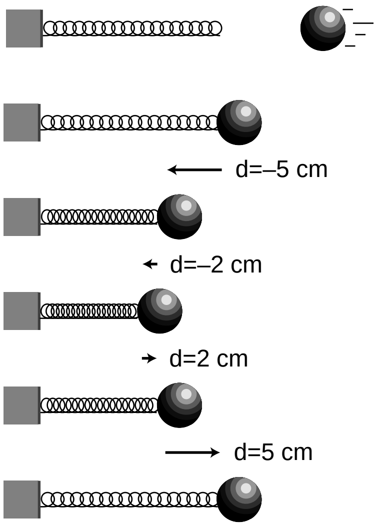
It might seem as though this arrangement would also double the work done by the tractor, but look again. As the tractor moves forward 2 meters, 1 meter of rope comes around the pulley, and the pulley moves 1 m to the left. Although the pulley exerts double the force on the stump, the pulley and stump only move half as far, so the work done on the stump is no greater that it would have been without the pulley.

The same is true for any mechanical arrangement that increases or decreases force, such as the gears on a ten-speed bike. You can't get out more work than you put in, because that would violate conservation of energy. A cyclist simply uses a smaller percentage of her energy in heating her body than does a runner, because there is less friction in her motion.

No work is done without motion.

It strikes most students as nonsensical when they are told that if they stand still and hold a heavy bag of cement, they are doing no work on the bag. Even if it makes sense mathematically that $W=Fd$ gives zero when d is zero, it seems to violate common sense. You would certainly become tired standing there. The solution is simple. Physicists have taken over the common word "work" and given it a new technical meaning, which is the transfer of energy. The energy of the bag of cement is not changing, and that is what the physicist means by saying no work is done on the bag.

There is a transformation of energy, but it is taking place entirely within your own muscles, which are converting chemical energy into heat. Physiologically, a human muscle is not like a tree limb, which can support a weight indefinitely without the expenditure of energy. Each muscle cell's contraction is generated by zillions of little molecular machines, which take turns supporting the tension. When a particular molecule goes on or off duty, it moves, and since it moves while exerting a force, it is doing work. There is work, but it is work done by one molecule in a muscle cell on another.



Positive and negative work

When object A transfers energy to object B, we say that A does positive work on B. B is said to do negative work on A. In other words, a machine like a tractor is defined as doing positive work. This use of the plus and minus signs relates in a logical and consistent way to their use in indicating the directions of force and motion in one dimension. In the example shown on the left, suppose we choose a coordinate system with the x axis pointing to the right. Then the force the spring exerts on the ball is always a positive number. The ball's motion, however, changes directions. The symbol *d* is really just a shorter way of writing the familiar quantity Δx, whose positive and negative signs indicate direction.

While the ball is moving to the left, we use $d<0$ to represent its direction of motion, and the work done by the spring, *Fd*, comes out negative. This indicates that the spring is taking kinetic energy out of the ball, and accepting it in the form of its own potential energy.

As the ball is reaccelerated to the right, it has $d>0$, *Fd* is positive, and the spring does positive work on the ball. Potential energy is transferred out of the spring and deposited in the ball as kinetic energy.

In summary:

rule for calculating work (including cases of negative work)
The work done by a force can be calculated as

$$W = Fd \ ,$$

if the force is constant and along the same line as the motion. The quantity *d* is to be interpreted as a synonym for Δx, i..e. positive and negative signs are used to indicate the direction of motion. Some ambiguities are encountered in cases such as kinetic friction.

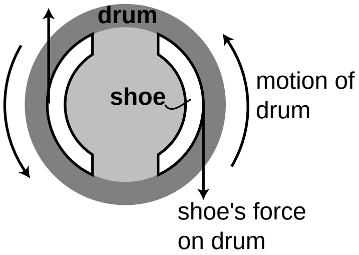
Self-Check

What about the work done by the ball on the spring?

There are many examples where the transfer of energy out of an object cancels out the transfer of energy in. When the tractor pulls the plow with a rope, the rope does negative work on the tractor and positive work on the plow. The total work done by the rope is zero, which makes sense, since it is not changing its energy.

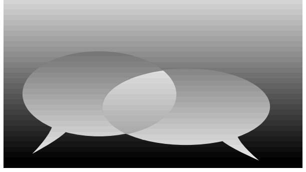
It may seem that when your arms do negative work by lowering a bag of cement, the cement is not really transferring energy into your body. If your body was storing potential energy like a compressed spring, you would be able to raise and lower a weight all day, recycling the same energy. The bag of cement does transfer energy into your body, but your body accepts it as heat, not as potential energy. The tension in the muscles that control the speed of the motion also results in the conversion of chemical energy to heat, for the same physiological reasons discussed previously in the case where you just hold the bag still.

Whenever energy is transferred out of the spring, the same amount has to be transferred into the ball, and vice versa. As the spring compresses, the ball is doing positive work on the spring (giving up its KE and transferring energy into the spring as PE), and as it decompresses the ball is doing negative work (extracting energy).



Because the force is in the opposite direction compared to the motion, the brake shoe does negative work on the drum, i.e. accepts energy from it in the form of heat.

One of the advantages of electric cars over gasoline-powered cars is that it is just as easy to put energy back in a battery as it is to take energy out. When you step on the brakes in a gas car, the brake shoes do negative work on the rest of the car. The kinetic energy of the car is transmitted through the brakes and accepted by the brake shoes in the form of heat. The energy cannot be recovered. Electric cars, however, are designed to use regenerative braking. The brakes don't use friction at all. They are electrical, and when you step on the brake, the negative work done by the brakes means they accept the energy and put it in the battery for later use. This is one of the reasons why an electric car is far better for the environment than a gas car, even if the ultimate source of the electrical energy happens to be the burning of oil in the electric company's plant. The electric car recycles the same energy over and over, and only dissipates heat due to air friction and rolling resistance, not braking. (The electric company's power plant can also be fitted with expensive pollution-reduction equipment that would be prohibitively expensive or bulky for a passenger car.)

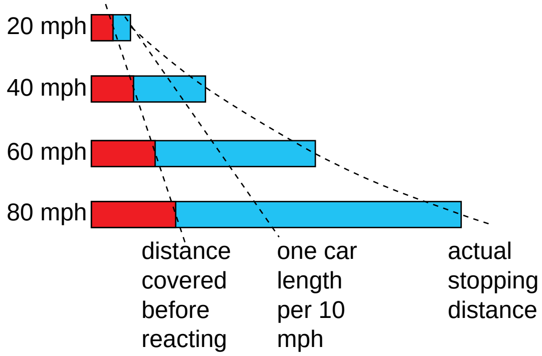


Discussion Questions

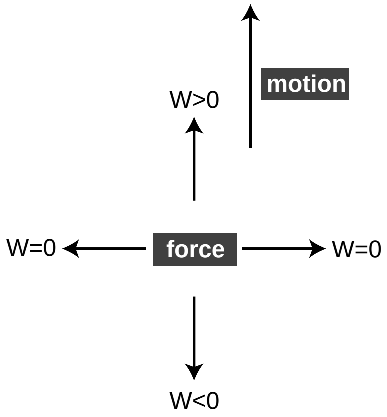
A. Besides the presence of a force, what other things differentiate the processes of frictional heating and heat conduction?

B. Criticize the following incorrect statement: “A force doesn't do any work unless it's causing the object to move.”

C. To stop your car, you must first have time to react, and then it takes some time for the car to slow down. Both of these times contribute to the distance you will travel before you can stop. The figure shows how the average stopping distance increases with speed. Because the stopping distance increases more and more rapidly as you go faster, the rule of one car length per 10 m.p.h. of speed is not conservative enough at high speeds. In terms of work and kinetic energy, what is the reason for the more rapid increase at high speeds?



3.2 Work in Three Dimensions



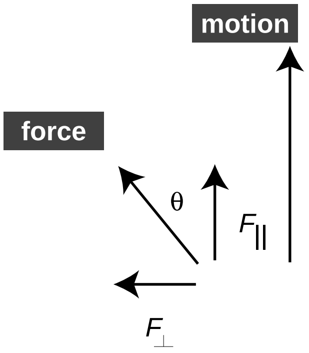
A force perpendicular to the motion does no work.

Suppose work is being done to change an object's kinetic energy. A force in the same direction as its motion will speed it up, and a force in the opposite direction will slow it down. As we have already seen, this is described as doing positive work or doing negative work on the object. All the examples discussed up until now have been of motion in one dimension, but in three dimensions the force can be at any angle θ with respect to the direction of motion.

What if the force is perpendicular to the direction of motion? We have already seen that a force perpendicular to the motion results in circular motion at constant speed. The kinetic energy does not change, and we conclude that no work is done when the force is perpendicular to the motion.

So far we have been reasoning about the case of a single force acting on an object, and changing only its kinetic energy. The result is more generally true, however. For instance, imagine a hockey puck sliding across the ice. The ice makes an upward normal force, but does not transfer energy to or from the puck

Forces at other angles



Suppose the force is at some other angle with respect to the motion, say θ=45°. Such a force could be broken down into two components, one along the direction of the motion and the other perpendicular to it. The force vector equals the vector sum of its two components, and the principle of vector addition of forces thus tells us that the work done by the total force cannot be any different than the sum of the works that would be done by the two forces by themselves. Since the component perpendicular to the motion does no work, the work done by the force must be

$$W = F_{||}\,|\boldsymbol{d}| \quad , \qquad \text{[work done by a constant force]}$$

where the vector $\boldsymbol{d}$ is simply a less cumbersome version of the notation $\Delta\boldsymbol{r}$. This result can be rewritten via trigonometry as

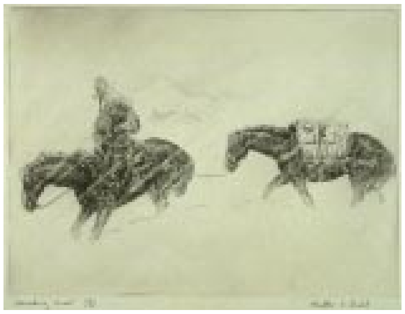
$$W = |\boldsymbol{F}|\,|\boldsymbol{d}|\cos\theta \quad . \qquad \text{[work done by a constant force]}$$

Even though this equation has vectors in it, it depends only on their magnitudes, and the magnitude of a vector is a scalar. Work is therefore still a scalar quantity, which only makes sense if it is defined as the transfer of energy. Ten gallons of gasoline have the ability to do a certain amount of mechanical work, and when you pull in to a full-service gas station you don't have to say "Fill 'er up with 10 gallons of south-going gas."

Students often wonder why this equation involves a cosine rather than a sine, or ask if it would ever be a sine. In vector addition, the treatment of sines and cosines seemed more equal and democratic, so why is the cosine so special now? The answer is that if we are going to describe, say, a velocity vector, we must give both the component *parallel* to the x axis and the component *perpendicular* to the x axis (i.e. the y component). In calculating work, however, the force component perpendicular to the motion is irrelevant — it changes the direction of motion without increasing or decreasing the energy of the object on which it acts. In this context, it is *only* the parallel force component that matters, so only the cosine occurs.

Self-Check

(a) Work is the transfer of energy. According to this definition, is the horse in the picture doing work on the pack? (b) If you calculate work by the method described in this section, is the horse doing work on the pack?



Breaking Trail, by Walter E. Bohl.

Example: pushing a broom

Question: If you exert a force of 21 N on a push broom, at an angle 35 degrees below horizontal, and walk for 5.0 m, how much work do you do? What is the physical significance of this quantity of work?

Solution: Using the second equation above, the work done equals

$$(21\ \text{N})(5.0\ \text{m})(\cos 35°) = 86\ \text{J} \quad .$$

The form of energy being transferred is heat in the floor and the broom’s bristles. This comes from the chemical energy stored in your body. (The majority of the calories you burn are dissipated directly as heat inside your body rather than doing any work on the broom. The 86 J is only the amount of energy transferred through the broom’s handle.)

(a) No. The pack is moving at constant velocity, so its kinetic energy is staying the same. It is only moving horizontally, so its gravitational potential energy is also staying the same. No energy transfer is occurring. (b) No. The horse’s upward force on the pack forms a 90-degree angle with the direction of motion, so cos θ=0, and no work is done.

3.3 Varying Force

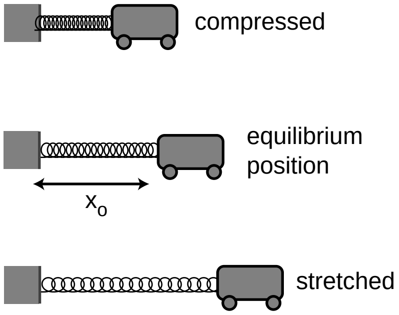
Up until now we have done no actual calculations of work in cases where the force is not constant. The question of how to treat such cases is mathematically analogous to the issue of how to generalize the equation (distance) = (velocity)(time) to cases where the velocity was not constant. There, we found that the correct generalization was to find the area under the graph of velocity versus time. The equivalent thing can be done with work:

> **general rule for calculating work**
> The work done by a force F equals the area under the curve on a graph of $F_{||}$ versus x. (Some ambiguities are encountered in cases such as kinetic friction.)

The examples in this section are ones in which the force is varying, but is always along the same line as the motion, so F is the same as $F_{||}$.

Self-Check

In which of the following examples would it be OK to calculate work using *Fd*, and in which ones would you have to use the area under the *F-x* graph?
(a) A fishing boat cruises with a net dragging behind it.
(b) A magnet leaps onto a refrigerator from a distance.
(c) Earth's gravity does work on an outward-bound space probe.



(a) The spring does work on the cart. (Unlike the ball in section 3.1, the cart is attached to the spring.)

An important and straightforward example is the calculation of the work done by a spring that obeys Hooke's law,

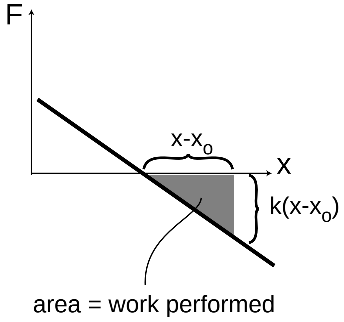
$$F \approx -k(x-x_o) \quad .$$

The minus sign is because this is the force being exerted by the spring, not the force that would have to act on the spring to keep it at this position. That is, if the position of the cart is to the right of equilibrium, the spring pulls back to the left, and vice-versa.

We calculate the work done when the spring is initially at equilibrium and then decelerates the car as the car moves to the right. The work done by the spring on the cart equals the minus area of the shaded triangle, because the triangle hangs below the x axis. The area of a triangle is half its base multiplied by its height, so

$$W = -\tfrac{1}{2}k(x-x_o)^2 \quad .$$

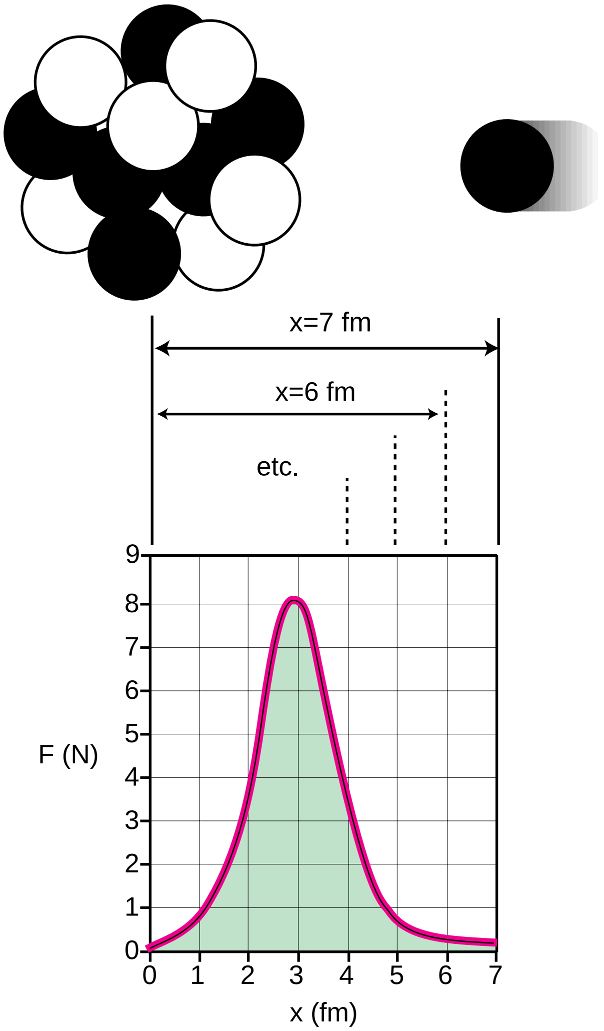
This is the amount of kinetic energy lost by the cart as the spring decelerates it.



(b)The area of the shaded triangle gives the work done by the spring as the cart moves from the equilibrium position to position x.

It was straightforward to calculate the work done by the spring in this case because the graph of F versus x was a straight line, giving a triangular area. But if the curve had not been so geometrically simple, it might not have been possible to find a simple equation for the work done, or an equation might have been derivable only using calculus. Optional section 3.4 gives an important example of such an application of calculus.

(a) If the boat is cruising at constant speed, then the forces are all presumably constant, so *Fd* is correct. (b) The force is changing: weaker at first, and stronger as the magnet approaches the fridge. *Fd* would give the wrong answer. (c) Gravity is getting weaker and weaker and the probe moves away from the earth. *Fd* would give the wrong answer.



Example: energy production in the sun

The sun produces energy through nuclear reactions in which nuclei collide and stick together. The figure depicts one such reaction, in which a single proton (hydrogen nucleus) collides with a carbon nucleus, consisting of six protons and six neutrons. Neutrons and protons attract other neutrons and protons via the strong nuclear force, so as the proton approaches the carbon nucleus it is accelerated. In the language of energy, we say that it loses nuclear potential energy and gains kinetic energy. Together, the seven protons and six neutrons make a nitrogen nucleus. Within the newly put-together nucleus, the neutrons and protons are continually colliding, and the new proton's extra kinetic energy is rapidly shared out among all the neutrons and protons. Soon afterward, the nucleus calms down by releasing some energy in the form of a gamma ray, which helps to heat the sun.

The graph shows the force between the carbon nucleus and the proton as the proton is on its way in, with the distance in units of femtometers (1 fm=10^{-15} m). Amusingly, the force turns out to be a few newtons: on the same order of magnitude as the forces we encounter ordinarily on the human scale. Keep in mind, however, that a force this big exerted on a single subatomic particle such as a proton will produce a truly fantastic acceleration (on the order of 10^{27} m/s^2!).

Why does the force have a peak around x=3 fm, and become smaller once the proton has actually merged with the nucleus? At x=3 fm, the proton is at the edge of the crowd of protons and neutrons. It feels many attractive forces from the left, and none from the right. The forces add up to a large value. However if it later finds itself at the center of the nucleus, x=0, there are forces pulling it from all directions, and these force vectors cancel out.

We can now calculate the energy released in this reaction by using the area under the graph to determine the amount of mechanical work done by the carbon nucleus on the proton. (For simplicity, we assume that the proton came in "aimed" at the center of the nucleus, and we ignore the fact that it has to shove some neutrons and protons out of the way in order to get there.) The area under the curve is about 17 squares, and the work represented by each square is

$$(1\ \text{N})(10^{-15}\ \text{m}) = 10^{-15}\ \text{J} \quad ,$$

so the total energy released is about

$$(10^{-15}\ \text{J/square})(17\ \text{squares}) = 1.7\text{x}10^{-14}\ \text{J} \quad .$$

This may not seem like much, but remember that this is only a reaction between the nuclei of two out of the zillions of atoms in the sun. For comparison, a typical *chemical* reaction between two atoms might transform on the order of 10^{-19} J of electrical potential energy into heat — 100,000 times less energy!

As a final note, you may wonder why reactions such as these only occur in the sun. The reason is that there is a repulsive electrical force between nuclei. When two nuclei are close together, the electrical forces are typically about a million times weaker than the nuclear forces, but the nuclear forces fall off much more quickly with distance than the electrical forces, so the electrical force is the dominant one at longer ranges. The sun is a very hot gas, so the random motion of its atoms is extremely rapid, and a collision between two atoms is sometimes violent enough to overcome this initial electrical repulsion.

3.4 ∫ Applications of Calculus

The student who has studied integral calculus will recognize that the graphical rule given in the previous section can be reexpressed as an integral,

$$W = \int_{x_1}^{x_2} F\,dx \quad .$$

We can then immediately find by the fundamental theorem of calculus that force is the derivative of work with respect to position,

$$F = \frac{\mathrm{d}W}{\mathrm{d}x} \quad .$$

For example, a crane raising a one-ton block on the moon would be transferring potential energy into the block at only one sixth the rate that would be required on Earth, and this corresponds to one sixth the force.

Although the work done by the spring could be calculated without calculus using the area of a triangle, there are many cases where the methods of calculus are needed in order to find an answer in closed form. The most important example is the work done by gravity when the change in height is not small enough to assume a constant force. Newton's law of gravity is

$$F = \frac{GMm}{r^2} \quad ,$$

which can be integrated to give

$$\begin{aligned} \mathrm{W} &= \int_{r_1}^{r_2} \frac{GMm}{r^2} dr \\ &= GMm\left(\frac{1}{r_2} - \frac{1}{r_1}\right) \quad . \end{aligned}$$

3.5 Work and Potential Energy

The techniques for calculating work can also be applied to the calculation of potential energy. If a certain force depends only on the distance between the two participating objects, then the energy released by changing the distance between them is defined as the potential energy, and the amount of potential energy lost equals minus the work done by the force,

$$\Delta PE = -W \quad .$$

The minus sign occurs because positive work indicates that the potential energy is being expended and converted to some other form.

Work and potential energy are not the same thing. We are simply using work as a way of calculating potential energy.

It is sometimes convenient to pick some arbitrary position as a reference position, and derive an equation for once and for all that gives the potential energy relative to this position

$$PE_x = -W_{\text{ref}\to\text{x}} \quad . \text{ [potential energy at a point x]}$$

To find the energy transferred into or out of potential energy, one then subtracts two different values of this equation.

These equations might almost make it look as though work and energy were the same thing, but they are not. First, potential energy measures the energy that a system *has* stored in it, while work measures how much energy is *transferred* in or out. Second, the techniques for calculating work can be used to find the amount of energy transferred in many situations where there is no potential energy involved, as when we calculate the amount of kinetic energy transformed into heat by a car's brake shoes.

Example: a toy gun

Question: A toy gun uses a spring with a spring constant of 10 N/m to shoot a ping-pong ball of mass 5 g. The spring is compressed to 10 cm shorter than its equilibrium length when the gun is loaded. At what speed is the ball released?

Solution: The equilibrium point is the natural choice for a reference point. Using the equation found previously for the work, we have

$$PE_x = \frac{1}{2}k\left(x - x_o\right)^2 \quad .$$

The spring loses contact with the ball at the equilibrium point, so the final potential energy is

$$PE_f = 0 \quad .$$

The initial potential energy is

$$\begin{aligned} PE_i &= (1/2)(10\text{ N/m})(0.10\text{ m})^2 \quad . \\ &= 0.05\text{ J.} \end{aligned}$$

The loss in potential energy of 0.05 J means an increase in kinetic energy of the same amount. The velocity of the ball is found by solving the equation $KE=\frac{1}{2}mv^2$ for v,

$$\begin{aligned} \text{v} &= \sqrt{\frac{2KE}{m}} \\ &= \sqrt{\frac{(2)(0.05\ J)}{0.005\ kg}} \\ &= 4\text{ m/s} \quad . \end{aligned}$$

Example: gravitational potential energy

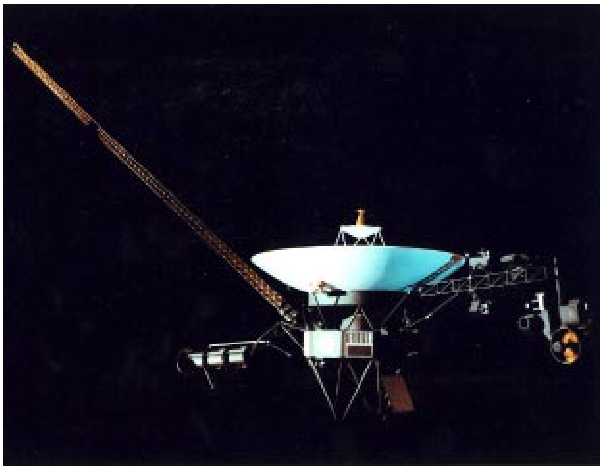
Question: We have already found the equation $\Delta PE = -F\Delta y$ for the gravitational potential energy when the change in height is not enough to cause a significant change in the gravitational force F. What if the change in height is enough so that this assumption is no longer valid? Use the equation $W = GMm\left(\frac{1}{r_2} - \frac{1}{r_1}\right)$ derived in the previous section to find the potential energy, using $r=\infty$ as a reference point.

Solution: The potential energy equals minus the work that would have to be done to bring the object from $r_1 = \infty$ to $r = r_2$, which is

$$PE = -\frac{GMm}{r} \ .$$

This is simpler than the equation for the work, which is an example of why it is advantageous to record an equation for potential energy relative to some reference point, rather than an equation for work.

Although the equations derived in the previous two examples may seem arcane and not particularly useful except for toy designers and rocket scientists, their usefulness is actually greater than it appears. The equation for the potential energy of a spring can be adapted to any other case in which an object is compressed, stretched, twisted, or bent. While you are not likely to use the equation for gravitational potential energy for anything practical, it is directly analogous to an equation that is extremely useful in chemistry, which is the equation for the potential energy of an electron at a distance r from the nucleus of its atom. As discussed in more detail later in the course, the electrical force between the electron and the nucleus is proportional to $1/r^2$, just like the gravitational force between two masses. Since the equation for the force is of the same form, so is the equation for the potential energy.



The twin Voyager space probes were perhaps the greatest scientific successes of the space program. Over a period of decades, they flew by all the planets of the outer solar system, probably accomplishing more of scientific interest than the entire space shuttle program at a tiny fraction of the cost. Both Voyager probes completed their final planetary flybys with speeds greater than the escape velocity at that distance from the sun, and so headed on out of the solar system on hyperbolic orbits, never to return. Radio contact has been lost, and they are now likely to travel interstellar space for billions of years without colliding with anything or being detected by any intelligent species.

Discussion Questions

A. What does the graph of $PE_x = \frac{1}{2}k\left(x - x_o\right)^2$ look like as a function of x? Discuss the physical significance of its features.

B. What does the graph of $PE = -\frac{GMm}{r}$ look like as a function of r? Discuss the physical significance of its features. How would the equation and graph change if some other reference point was chosen rather than $r=\infty$?

C. Starting at a distance r from a planet of mass M, how fast must an object be moving in order to have a hyperbolic orbit, i.e. one that never comes back to the planet? This velocity is called the escape velocity. Interpreting the result, does it matter in what direction the velocity is? Does it matter what mass the object has? Does the object escape because it is moving too fast for gravity to act on it?

D. Does a spring have an “escape velocity”?

E. Calculus-based question: If the form of energy being transferred is potential energy, then the equations $F = \mathrm{d}W / \mathrm{d}x$ and $W = \int F \,\mathrm{d}x$ become $F = \mathrm{d}PE / \mathrm{d}x$ and $PE = \int F \,\mathrm{d}x$. How would you then apply the following calculus concepts: zero derivative at minima and maxima, and the second derivative test for concavity up or down.

3.6* When Does Work Equal Force Times Distance?

In the first section of this chapter I gave an example of a case where the work done by a force did not equal Fd. The purpose of this section is to explain more fully how the quantity Fd can and cannot be used. To simplify things, I write Fd throughout this section, but more generally everything said here would be true for the area under the graph of $F_{||}$ versus d.

The following two theorems allow most of the ambiguity to be cleared up.

the work-kinetic energy theorem

The change in the kinetic energy associated with the motion of an object's center of mass is related to the total force acting on it and to the distance traveled by its center of mass according to the equation $\Delta KE_{\mathrm{cm}} = F_{total} d_{\mathrm{cm}}$.

This can be proven based on Newton’s second law and the equation $KE=\frac{1}{2}mv^2$. Note that despite the traditional name, it does not necessarily tell the amount of work done, since the forces acting on the object could be changing other types of energy besides the KE associated with its center of mass motion.

The second theorem does relate directly to work:

When a contact force acts between two objects and the two surfaces do not slip past each other, the work done equals Fd, where d is the distance traveled by the point of contact.

This one has no generally accepted name, so we refer to it simply as the second theorem.

A great number of physical situations can be analyzed with these two theorems, and often it is advantageous to apply both of them to the same situation.

Example: an ice skater pushing off from a wall

- The work-kinetic energy theorem tells us how to calculate the skater's kinetic energy if we know the amount of force and the distance her center of mass travels while she is pushing off.
- The second theorem tells us that the wall does no work on the skater. This makes sense, since the wall does not have any source of energy.

Example: absorbing an impact without recoiling?

Question: Is it possible to absorb an impact without recoiling? For instance, would a brick wall "give" at all if hit by a ping-pong ball?

Answer: There will always be a recoil. In the example proposed, the wall will surely have some energy transferred to it in the form of heat and vibration. The second theorem tells us that we can only have nonzero work if the distance traveled by the point of contact is nonzero.

Example: dragging a refrigerator at constant velocity

- Newton's first law tells us that the total force on the refrigerator must be zero: your force is canceling the floor's kinetic frictional force. The work-kinetic energy theorem is therefore true but useless. It tells us that there is zero total force on the refrigerator, and that the refrigerator's kinetic energy doesn't change.
- The second theorem tells us that the work you do equals your hand's force on the refrigerator multiplied by the distance traveled. Since we know the floor has no source of energy, the only way for the floor and refrigerator to gain energy is from the work you do. We can thus calculate the total heat dissipated by friction in the refrigerator and the floor.

Note that there is no way to find how much of the heat is dissipated in the floor and how much in the refrigerator.

Example: accelerating a cart

If you push on a cart and accelerate it, there are two forces acting on the cart: your hand's force, and the static frictional force of the ground pushing on the wheels in the opposite direction.

- Applying the second theorem to your force tells us how to calculate the work you do.
- Applying the second theorem to the floor's force tells us that the floor does no work on the cart. There is no motion at the point of contact, because the atoms in the floor are not moving. (The atoms in the surface of the wheel are also momentarily at rest when they touch the floor.) This makes sense, since the floor does not have any source of energy.
- The work-kinetic energy theorem refers to the total force, and because the floor's backward force cancels part of your force, the total force is less than your force. This tells us that only part of your work goes into the kinetic energy associated with the forward motion of the cart's center of mass. The rest goes into rotation of the wheels.

3.7* The Dot Product

Up until now, we have not found any physically useful way to define the multiplication of two vectors. It would be possible, for instance, to multiply two vectors component by component to form a third vector, but there are no physical situations where such a multiplication would be useful.

The equation $W = |\boldsymbol{F}|\,|\boldsymbol{d}| \cos\theta$ is an example of a sort of multiplication of vectors that is useful. The result is a scalar, not a vector, and this is therefore often referred to as the *scalar product* of the vectors $\boldsymbol{F}$ and $\boldsymbol{d}$. There is a standard shorthand notation for this operation,

$$\boldsymbol{A}\cdot\boldsymbol{B} = |\boldsymbol{A}|\,|\boldsymbol{B}| \cos\theta \quad ,$$

[definition of the notation $\boldsymbol{A}\cdot\boldsymbol{B}$;

θ is the angle between vectors $\boldsymbol{A}$ and $\boldsymbol{B}$]

and because of this notation, a more common term for this operation is the *dot product.* In dot product notation, the equation for work is simply

$$W = \boldsymbol{F}\cdot\boldsymbol{d}$$

The dot product has the following geometric interpretation:

$$\begin{aligned}\boldsymbol{A}\cdot\boldsymbol{B} &= |\boldsymbol{A}| \text{ x (component of } \boldsymbol{B} \text{ parallel to } \boldsymbol{A})\\ &= |\boldsymbol{B}| \text{ x (component of } \boldsymbol{A} \text{ parallel to } \boldsymbol{B})\end{aligned}$$

The dot product has some of the properties possessed by ordinary multiplication of numbers,

$$\boldsymbol{A}\cdot\boldsymbol{B} = \boldsymbol{B}\cdot\boldsymbol{A}$$

$$\boldsymbol{A}\cdot(\boldsymbol{B}+\boldsymbol{C}) = \boldsymbol{A}\cdot\boldsymbol{B}+\boldsymbol{A}\cdot\boldsymbol{C}$$

$$(c\boldsymbol{A})\cdot\boldsymbol{B} = c(\boldsymbol{A}\cdot\boldsymbol{B}) \quad ,$$

but it lacks one other: the ability to undo multiplication by dividing.

If you know the components of two vectors, you can easily calculate their dot product as follows:

$$\boldsymbol{A}\cdot\boldsymbol{B} = A_x B_x + A_y B_y + A_z B_z \quad .$$

This can be proven by first analyzing the special case where each vector has only an *x* component, and the similar cases for *y* and *z*. We can then apply the rule $\boldsymbol{A}\cdot(\boldsymbol{B}+\boldsymbol{C}) = \boldsymbol{A}\cdot\boldsymbol{B}+\boldsymbol{A}\cdot\boldsymbol{C}$ to generalize by writing each vector as the sum of its *x*, *y*, and *z* components.

Summary

Selected Vocabulary

work	the amount of energy transferred into or out of a system, excluding energy transferred by heat conduction

Notation

W	work

Summary

Work is a measure of the transfer of mechanical energy, i.e. the transfer of energy by a force rather than by heat conduction. When the force is constant, work can usually be calculated as

$$W = F_{||} |\boldsymbol{d}| \quad , \qquad \text{[only if the force is constant]}$$

where $\boldsymbol{d}$ is simply a less cumbersome notation for $\Delta\boldsymbol{r}$, the vector from the initial position to the final position. Thus,

- A force in the same direction as the motion does positive work, i.e. transfers energy into the object on which it acts.
- A force in the opposite direction compared to the motion does negative work, i.e. transfers energy out of the object on which it acts.
- When there is no motion, no mechanical work is done. The human body burns calories when it exerts a force without moving, but this is an internal energy transfer of energy within the body, and thus does not fall within the scientific definition of work.
- A force perpendicular to the motion does no work.

When the force is not constant, the above equation should be generalized as the area under the graph of $F_{||}$ versus d.

Machines such as pulleys, levers, and gears may increase or decrease a force, but they can never increase or decrease the amount of work done. That would violate conservation of energy unless the machine had some source of stored energy or some way to accept and store up energy.

There are some situations in which the equation $W = F_{||} |\boldsymbol{d}|$ is ambiguous or not true, and these issues are discussed rigorously in section 3.6. However, problems can usually be avoided by analyzing the types of energy being transferred before plunging into the math. In any case there is no substitute for a physical understanding of the processes involved.

The techniques developed for calculating work can also be applied to the calculation of potential energy. We fix some position as a reference position, and calculate the potential energy for some other position, x, as

$$PE_x = -W_{\text{ref}\rightarrow x} \quad .$$

The following two equations for potential energy have broader significance than might be suspected based on the limited situations in which they were derived:

$$PE = \frac{1}{2}\, k(x\text{-}x_o)^2 \quad .$$

[potential energy of a spring having spring constant k, when stretched or compressed from the equilibrium position x_o; analogous equations apply for the twisting, bending, compression, or stretching of any object.]

$$PE = -\frac{GMm}{r}$$

[gravitational potential energy of objects of masses M and m, separated by a distance r; an analogous equation applies to the electrical potential energy of an electron in an atom.]

Homework Problems

1. Two cars with different masses each have the same kinetic energy. (a) If both cars have the same brakes, capable of supplying the same force, how will the stopping distances compare? Explain. (b) Compare the times required for the cars to stop.

2. In each of the following situations, is the work being done positive, negative, or zero? (a) a bull paws the ground; (b) a fishing boat pulls a net through the water behind it; (c) the water resists the motion of the net through it; (d) you stand behind a pickup truck and lower a bale of hay from the truck's bed to the ground. Explain.

3. In the earth's atmosphere, the molecules are constantly moving around. Because temperature is a measure of kinetic energy per molecule, the average kinetic energy of each type of molecule is the same, e.g. the average KE of the O_2 molecules is the same as the average KE of the N_2 molecules. (a) If the mass of an O_2 molecule is eight times greater than that of a He atom, what is the ratio of their average speeds? Which way is the ratio, i.e. which is typically moving faster? (b) Use your result from part a to explain why any helium occurring naturally in the atmosphere has long since escaped into outer space, never to return. (Helium is obtained commercially by extracting it from rocks.)

4. Weiping lifts a rock with a weight of 1.0 N through a height of 1.0 m, and then lowers it back down to the starting point. Bubba pushes a table 1.0 m across the floor at constant speed, requiring a force of 1.0 N, and then pushes it back to where it started. Compare the total work done by Weiping and Bubba.

5. ✓ In one of his more flamboyant moments, Galileo wrote "Who does not know that a horse falling from a height of three or four cubits will break his bones, while a dog falling from the same height or a cat from a height of eight or ten cubits will suffer no injury? Equally harmless would be the fall of a grasshopper from a tower or the fall of an ant from the distance of the moon." Find the speed of an ant that falls to earth from the distance of the moon at the moment when it is about to enter the atmosphere. Assume it is released from a point that is not actually near the moon, so the moon's gravity is negligible.

6. (a) You are moving into a third-floor apartment, and your piano doesn't fit in the elevator. You and your friends put it on the ground outside your window, tie a strong cable to it, and pull it up onto your balcony, 9 m above. The piano has a mass of 650 kg. Define y=0 to be on the ground, and y=+9 m at the balcony. Draw a graph of the force you exert on the piano as a function of y. Find the area under the graph, and determine the work done by you and your friends. Give your answer in Joules.

S A solution is given in the back of the book.	★ A difficult problem.
✓ A computerized answer check is available.	∫ A problem that requires calculus.

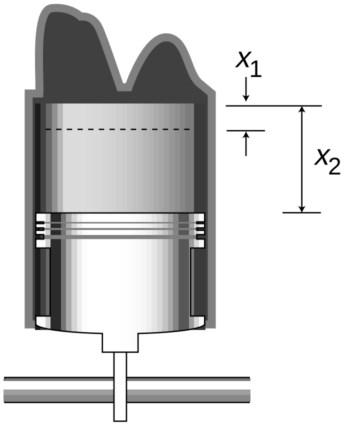
(b) Convert your answer from (a) into units of kcal.

7. (a) The crew of an 18th century warship is raising the anchor. The anchor has a mass of 5000 kg. The water is 30 m deep. The chain to which the anchor is attached has a mass per unit length of 150 kg/m. Before they start raising the anchor, what is the total weight of the anchor plus the portion of the chain hanging out of the ship? (Assume that the buoyancy of the anchor and is negligible.)

(b) After they have raised the anchor by 1 m, what is the weight they are raising?

(c) Define y=0 when the anchor is resting on the bottom, and y=+30 m when it has been raised up to the ship. Draw a graph of the force the crew has to exert to raise the anchor and chain, as a function of y. (Assume that they are raising it slowly, so water resistance is negligible.) It will not be a constant! Now find the area under the graph, and determine the work done by the crew in raising the anchor, in joules.

(d✓) Convert your answer from (c) into units of kcal.



Problem 8: A cylinder from the 1965 Rambler's engine. The piston is shown in its pushed out position. The two bulges at the top are for the valves that let fresh air-gas mixture in.

Based on a figure from Motor Service's Automotive Encyclopedia, Toboldt and Purvis.

8. In the power stroke of a car's gasoline engine, the fuel-air mixture is ignited by the spark plug, explodes, and pushes the piston out. The exploding mixture's force on the piston head is greatest at the beginning of the explosion, and decreases as the mixture expands. It can be approximated by $F = a / x$, where x is the distance from the cylinder to the piston head, and a is a constant with units of N·m. (Actually $a/x^{1.4}$ would be more accurate, but the problem works out more nicely with a/x!)The piston begins its stroke at x=x_1, and ends at x=x_2. The 1965 Rambler had six cylinders, each with a=220 N·m, x_1=1.2 cm, and x_2=10.2 cm.

(a) Draw a neat, accurate graph of F vs x, on graph paper.

(b✓) From the area under the curve, derive the amount of work done in one stroke by one cylinder.

(c✓) Assume the engine is running at 4800 r.p.m., so that during one minute, each of the six cylinders performs 2400 power strokes. (Power strokes only happen every other revolution.) Find the engine's power, in units of horsepower (1 hp=746 W).

(d) The compression ratio of an engine is defined as x_2/x_1. Explain in words why the car's power would be exactly the same if x_1 and x_2 were, say, halved or tripled, maintaining the same compression ratio of 8.5. Explain why this would *not* quite be true with the more realistic force equation F=$a/x^{1.4}$.

9. ∫ The magnitude of the force between two magnets separated by a distance r can be approximated as kr^{-3} for large values of r. The constant k depends on the strengths of the magnets and the relative orientations of their north and south poles. Two magnets are released on a slippery surface at an initial distance r_i, and begin sliding towards each other. What will be the total kinetic energy of the two magnets when they reach a final distance r_f? (Ignore friction.)

10. ∫ A car starts from rest at t=0, and starts speeding up with constant acceleration. (a) Find the car's kinetic energy in terms of its mass, m, acceleration, a, and the time, t. (b) Your answer in the previous part also equals the amount of work, W, done from t=0 until time t. Take the derivative of the previous expression to find the power expended by the car at time t. (c) Suppose two cars with the same mass both start from rest at the same time, but one has twice as much acceleration as the other. At any moment, how many times more power is being dissipated by the more quickly accelerating car? (The answer is not 2.)

11.★ ∫ A space probe of mass m is dropped into a previously unexplored spherical cloud of gas and dust, and accelerates toward the center of the cloud under the influence of the cloud's gravity. Measurements of its velocity allow its potential energy, U, to be determined as a function of the distance r from the cloud's center. The mass in the cloud is distributed in a spherically symmetric way, so its density, $\rho(r)$, depends only on r and not on the angular coordinates. Show that by finding $U(r)$, one can infer $\rho(r)$ as follows:

$$\rho(r) = \frac{1}{4\pi G m r^2}\frac{d}{dr}\left(r^2\frac{dU}{dr}\right) \quad .$$

12. ∫ A rail gun is a device like a train on a track, with the train propelled by a powerful electrical pulse. Very high speeds have been demonstrated in test models, and rail guns have been proposed as an alternative to rockets for sending into outer space any object that would be strong enough to survive the extreme accelerations. Suppose that the rail gun capsule is launched straight up, and that the force of air friction acting on it is given by $F=be^{-cx}$, where x is the altitude, b and c are constants, and e is the base of natural logarithms. The exponential decay occurs because the atmosphere gets thinner with increasing altitude. (In reality, the force would probably drop off even faster than an exponential, because the capsule would be slowing down somewhat.) Find the amount of kinetic energy lost by the capsule due to air friction between when it is launched and when it is completely beyond the atmosphere. (Gravity is negligible, since the air friction force is much greater than the gravitational force.)

13 . A certain binary star system consists of two stars with masses m_1 and m_2, separated by a distance b. A comet, originally nearly at rest in deep space, drops into the system and at a certain point in time arrives at the midpoint between the two stars. For that moment in time, find its velocity, v, symbolically in terms of b, m_1, m_2, and fundamental constants. [Numerical check: For m_1=1.5x10^{30} kg, m_2=3.0x10^{30} kg, and b=2.0x10^{11} m you should find v=7.7x10^4 m/s.]

14 ∫. An airplane flies in the positive direction along the *x* axis, through crosswinds that exert a force $\boldsymbol{F} = \left(a + bx\right)\hat{\boldsymbol{x}} + \left(c + dx\right)\hat{\boldsymbol{y}}$. Find the work done by the wind on the plane, and by the plane on the wind, in traveling from the origin to position *x*.

15. ∫ In 1935, Yukawa proposed an early theory of the force that held the neutrons and protons together in the nucleus. His equation for the potential energy of two such particles, at a center-to-center distance *r*, was $PE(r)=g\,r^{-1}\,e^{-r/a}$, where *g* parametrizes the strength of the interaction, *e* is the base of natural logarithms, and *a* is about 10^{-15} m. Find the force between two nucleons that would be consistent with this equation for the potential energy.

16. Prove that the dot product defined in section 3.7 is rotationally invariant in the sense of book 1, section 7.5.

17. Fill in the details of the proof of $\boldsymbol{A} \cdot \boldsymbol{B}= A_xB_x + A_yB_y + A_zB_z$ in section 3.7.

18 S. Does it make sense to say that work is conserved?

19. (a) Suppose work is done in one-dimensional motion. What happens to the work if you reverse the direction of the positive coordinate axis? Base your answer directly on the definition of work. (b) Now answer the question based on the *W*=*Fd* rule.

4 Conservation of Momentum

In many subfields of physics these days, it is possible to read an entire issue of a journal without ever encountering an equation involving force or a reference to Newton's laws of motion. In the last hundred and fifty years, an entirely different framework has been developed for physics, based on conservation laws.

The new approach is not just preferred because it is in fashion. It applies inside an atom or near a black hole, where Newton's laws do not. Even in everyday situations the new approach can be superior. We have already seen how perpetual motion machines could be designed that were too complex to be easily debunked by Newton's laws. The beauty of conservation laws is that they tell us something must remain the same, regardless of the complexity of the process.

So far we have discussed only two conservation laws, the laws of conservation of mass and energy. Is there any reason to believe that further conservation laws are needed in order to replace Newton's laws as a complete description of nature? Yes. Conservation of mass and energy do not relate in any way to the three dimensions of space, because both are scalars. Conservation of energy, for instance, does not prevent the planet earth from abruptly making a 90-degree turn and heading straight into the sun, because kinetic energy does not depend on direction. In this chapter, we develop a new conserved quantity, called momentum, which is a vector.

4.1 Momentum

A conserved quantity of motion

Your first encounter with conservation of momentum may have come as a small child unjustly confined to a shopping cart. You spot something interesting to play with, like the display case of imported wine down at the end of the aisle, and decide to push the cart over there. But being imprisoned by Dad in the cart was not the only injustice that day. There was a far greater conspiracy to thwart your young id, one that originated in the laws of nature. Pushing forward did nudge the cart forward, but it pushed you backward. If the wheels of the cart were well lubricated, it wouldn't matter how you jerked, yanked, or kicked off from the back of the cart. You could not cause any overall forward motion of the entire system consisting of the cart with you inside.

In the Newtonian framework, we describe this as arising from Newton's third law. The cart made a force on you that was equal and opposite to your force on it. In the framework of conservation laws, we cannot attribute your frustration to conservation of energy. It would have been perfectly possible for you to transform some of the internal chemical energy stored in your body to kinetic energy of the cart and your body.

The following characteristics of the situation suggest that there may be a new conservation law involved:

> **A closed system is involved**. All conservation laws deal with closed systems. You and the cart are a closed system, since the well-oiled wheels prevent the floor from making any forward force on you.
>
> **Something remains unchanged**. The overall velocity of the system started out being zero, and you cannot change it. This vague reference to "overall velocity" can be made more precise: it is the velocity of the system's center of mass that cannot be changed.
>
> **Something can be transferred back and forth without changing the total amount**: If we define forward as positive and backward as negative, then one part of the system can gain positive motion if another part acquires negative motion. If we don't want to worry about positive and negative signs, we can imagine that the whole cart was initially gliding forward on its well-oiled wheels. By kicking off from the back of the cart, you could increase your own velocity, but this inevitably causes the cart to slow down.

It thus appears that there is some numerical measure of an object's quantity of motion that is conserved when you add up all the objects within a system.

Momentum

Although velocity has been referred to, it is not the total velocity of a closed system that remains constant. If it was, then firing a gun would cause the gun to recoil at the same velocity as the bullet! The gun does recoil, but at a much lower velocity than the bullet. Newton's third law tells us

$$F_{\text{gun on bullet}} = -F_{\text{bullet on gun}} \quad ,$$

and assuming a constant force for simplicity, Newton's second law allows us to change this to

$$m_{\text{bullet}}\frac{\Delta v_{\text{bullet}}}{\Delta t} = -m_{\text{gun}}\frac{\Delta v_{\text{gun}}}{\Delta t} \quad .$$

Thus if the gun has 100 times more mass than the bullet, it will recoil at a velocity that is 100 times smaller and in the opposite direction, represented by the opposite sign. The quantity *mv* is therefore apparently a useful measure of motion, and we give it a name, *momentum*, and a symbol, *p*. (As far as I know, the letter "p" was just chosen at random, since "m" was already being used for mass.) The situations discussed so far have been one-dimensional, but in three-dimensional situations it is treated as a vector.

> **definition of momentum for material objects**
> The momentum of a material object, i.e. a piece of matter, is defined as
> $$\mathbf{p} = m\mathbf{v} \quad ,$$
> the product of the object's mass and its velocity vector.

The units of momentum are kg·m/s, and there is unfortunately no abbreviation for this clumsy combination of units.

The reasoning leading up to the definition of momentum was all based on the search for a conservation law, and the only reason why we bother to define such a quantity is that experiments show it is conserved:

> **the law of conservation of momentum**
> In any closed system, the vector sum of all the momenta remains constant,
> $$\mathbf{p}_{1i} + \mathbf{p}_{2i} + \ldots = \mathbf{p}_{1f} + \mathbf{p}_{2fi} + \ldots \quad ,$$
> where *i* labels the initial and *f* the final momenta. (A closed system is one on which no external forces act.)

This chapter first addresses the one-dimensional case, in which the direction of the momentum can be taken into account by using plus and minus signs. We then pass to three dimensions, necessitating the use of vector addition.

A subtle point about conservation laws is that they all refer to "closed systems," but "closed" means different things in different cases. When discussing conservation of mass, "closed" means a system that doesn't have matter moving in or out of it. With energy, we mean that there is no work or heat transfer occurring across the boundary of the system. For momentum conservation, "closed" means there are no external *forces* reaching into the system.

Example: a cannon

Question: A cannon of mass 1000 kg fires a 10-kg shell at a velocity of 200 m/s. At what speed does the cannon recoil?

Solution: The law of conservation of momentum tells us that

$$\boldsymbol{p}_{\text{cannon,i}} + \boldsymbol{p}_{\text{shell,i}} = \boldsymbol{p}_{\text{cannon,f}} + \boldsymbol{p}_{\text{shell,f}} \quad .$$

Choosing a coordinate system in which the cannon points in the positive direction, the given information is

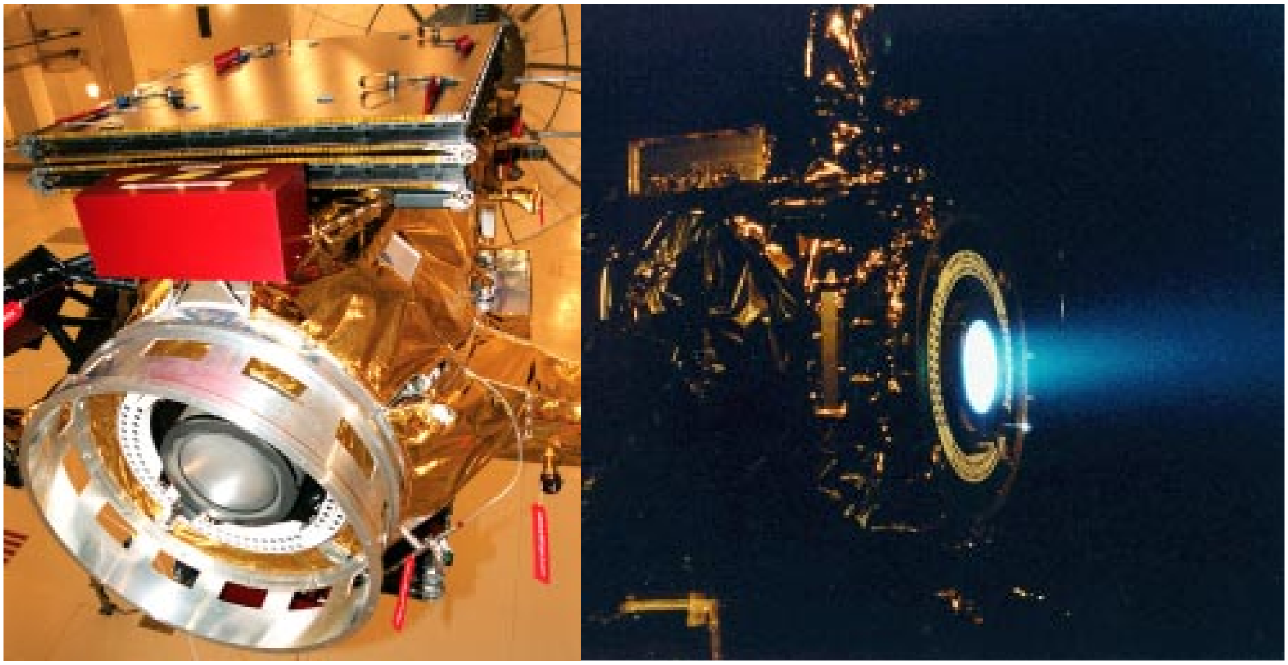
$$\boldsymbol{p}_{\text{cannon,i}} = 0$$
$$\boldsymbol{p}_{\text{shell,i}} = 0$$
$$\boldsymbol{p}_{\text{shell,f}} = 2000\ \text{kg}\cdot\text{m/s} \quad .$$

We must have $\boldsymbol{p}_{\text{cannon,f}}$=–2000 kg·m/s, so the recoil velocity of the cannon is 2 m/s.

Example: ion drive for propelling spacecraft

Question: The experimental solar-powered ion drive of the Deep Space 1 space probe expels its xenon gas exhaust at a speed of 30,000 m/s, ten times faster than the exhaust velocity for a typical chemical-fuel rocket engine. Roughly how many times greater is the maximum speed this spacecraft can reach, compared with a chemical-fueled probe with the same mass of fuel (“reaction mass”) available for pushing out the back as exhaust?

Solution: Momentum equals mass multiplied by velocity. Both spacecraft are assumed to have the same amount of reaction mass, and the ion drive’s exhaust has a velocity ten times greater, so the momentum of its exhaust is ten times greater. Before the engine starts firing, neither the probe nor the exhaust has any momentum, so the total momentum of the system is zero. By conservation of momentum, the total momentum must also be zero after all the exhaust has been expelled. If we define the positive direction as the direction the spacecraft is going, then the negative momentum of the exhaust is canceled by the positive momentum of the spacecraft. The ion drive allows a final



The ion drive engine of the NASA Deep Space 1 probe, shown under construction (left) and being tested in a vacuum chamber (right) prior to its October 1998 launch. Intended mainly as a test vehicle for new technologies, the craft nevertheless is scheduled to carry out a scientific program that includes a 1999 flyby of a near-earth asteroid and a rendezvous with a comet in 2004.

speed that is ten times greater. (This simplified analysis ignores the fact that the reaction mass expelled later in the burn is not moving backward as fast, because of the forward speed of the already-moving spacecraft.)

Generalization of the momentum concept

As with all the conservation laws, the law of conservation of momentum has evolved over time. In the 1800s it was found that a beam of light striking an object would give it some momentum, even though light has no mass, and would therefore have no momentum according to the above definition. Rather than discarding the principle of conservation of momentum, the physicists of the time decided to see if the definition of momentum could be extended to include momentum carried by light. The process is analogous to the process outlined in chapter 1 for identifying new forms of energy. The first step was the discovery that light could impart momentum to matter, and the second step was to show that the momentum possessed by light could be related in a definite way to observable properties of the light. They found that conservation of momentum could be successfully generalized by attributing to a beam of light a momentum vector in the direction of the light's motion and having a magnitude proportional to the amount of energy the light possessed. The momentum of light is negligible under ordinary circumstances, e.g. a flashlight left on for an hour would only absorb about 10^{-5} kg·m/s of momentum as it recoiled.

The reason for bringing this up is not so that you can plug numbers into a formulas in these exotic situations. The point is that the conservation

Momentum is not always equal to *mv*. Halley's comet, shown here, has a very elongated elliptical orbit, like those of many other comets. About once per century, its orbit brings it close to the sun. The comet's head, or nucleus, is composed of dirty ice, so the energy deposited by the intense sunlight boils off water vapor. The bottom photo shows a view of the water boiling off of the nucleus from the European Giotto space probe, which passed within 596 km of the comet's head on March 13, 1986. The sunlight does not just carry energy, however. It also carries momentum. Once the steam boils off, the momentum of the sunlight impacting on it pushes it away from the sun, forming a tail as shown in in the top image, taken through a ground-based telescope. By analogy with matter, for which momentum equals *mv*, you would expect that massless light would have zero momentum, but the equation $p=mv$ is not the correct one for light, and light does have momentum. (Some comets also have a second tail, which is propelled by electrical forces rather than by the momentum of sunlight.)

Modern Changes in the Momentum Concept

Einstein played a role in two major changes in the momentum concept in the 1900s.

First Einstein showed that the equation ***mv*** would not work for a system containing objects moving at very high speeds relative to one another. He came up with a new equation, to which ***mv*** is only the low-velocity approximation.

The second change, and a far stranger one, was the realization that at the atomic level, motion is inescapably random. The electron in a hydrogen atom doesn't really orbit the nucleus, it forms a vague cloud around it. It might seem that this would prove nonconservation of momentum, but in fact the random wanderings of the proton are exactly coordinated with those of the electron so that the total momentum stays exactly constant. In an atom of lead, there are 82 electrons plus the nucleus, all changing their momenta randomly from moment to moment, but all coordinating mysteriously with each other to keep the vector sum constant. In the 1930s, Einstein pointed out that the theories of the atom then being developed would require this kind of spooky coordination, and used this as an argument that there was something physically unreasonable in the new ideas. Experiments, however, have shown that the spooky effects do happen, and Einstein's objections are remembered today only as a historical curiousity.

laws have proven so sturdy exactly because they can easily be amended to fit new circumstances. Newton's laws are no longer at the center of the stage of physics because they did not have the same adaptability. More generally, the moral of this story is the provisional nature of scientific truth.

It should also be noted that conservation of momentum is not a consequence of Newton's laws, as is often asserted in textbooks. Newton's laws do not apply to light, and therefore could not possibly be used to prove anything about a concept as general as the conservation of momentum in its modern form.

Momentum compared to kinetic energy

Momentum and kinetic energy are both measures of the quantity of motion, and a sideshow in the Newton-Leibniz controversy over who invented calculus was an argument over whether mv (i.e. momentum) or mv^2 (i.e. kinetic energy without the 1/2 in front) was the "true" measure of motion. The modern student can certainly be excused for wondering why we need both quantities, when their complementary nature was not evident to the greatest minds of the 1700s. The following table highlights their differences.

Kinetic energy...	**Momentum...**
is a scalar.	is a vector.
is not changed by a force perpendicular to the motion, which changes only the direction of the velocity vector.	is changed by any force, since a change in either the magnitude or direction of the velocity vector will result in a change in the momentum vector.
is always positive, and cannot cancel out.	cancels with momentum in the opposite direction.
can be traded for forms of energy that do not involve motion. KE is not a conserved quantity by itself.	is always conserved in a closed system.
is quadrupled if the velocity is doubled.	is doubled if the velocity is doubled.

Here are some examples that show the different behaviors of the two quantities.

Example: a spinning top
A spinning top has zero total momentum, because for every moving point, there is another point on the opposite side that cancels its momentum. It does, however, have kinetic energy.

Example: momentum and kinetic energy in firing a rifle

The rifle and bullet have zero momentum and zero kinetic energy to start with. When the trigger is pulled, the bullet gains some momentum in the forward direction, but this is canceled by the rifle's backward momentum, so the total momentum is still zero. The kinetic energies of the gun and bullet are both positive scalars, however, and do not cancel. The total kinetic energy is allowed to increase, because both objects' kinetic energies are destined to be dissipated as heat — the gun's "backward" kinetic energy does not refrigerate the shooter's shoulder!

Example: the wobbly earth

As the moon completes half a circle around the earth, its motion reverses direction. This does not involve any change in kinetic energy, and the earth's gravitational force does not do any work on the moon. The reversed velocity vector does, however, imply a reversed momentum vector, so conservation of momentum in the closed earth-moon system tells us that the earth must also change its momentum. In fact, the earth wobbles in a little "orbit" about a point below its surface on the line connecting it and the moon. The two bodies' momentum vectors always point in opposite directions and cancel each other out.

Example: the earth and moon get a divorce

Why can't the moon suddenly decide to fly off one way and the earth the other way? It is not forbidden by conservation of momentum, because the moon's newly acquired momentum in one direction could be canceled out by the change in the momentum of the earth, supposing the earth headed the opposite direction at the appropriate, slower speed. The catastrophe is forbidden by conservation of energy, because both their kinetic energies would have increased greatly.

Example: momentum and kinetic energy of a glacier

A cubic-kilometer glacier would have a mass of about 10^{12} kg. If it moves at a speed of 10^{-5} m/s, then its momentum is 10^{7} kg$\cdot$m/s. This is the kind of heroic-scale result we expect, perhaps the equivalent of the space shuttle taking off, or all the cars in LA driving in the same direction at freeway speed. Its kinetic energy, however, is only 50 J, the equivalent of the calories contained in a poppy seed or the energy in a drop of gasoline too small to be seen without a microscope. The surprisingly small kinetic energy is because kinetic energy is proportional to the square of the velocity, and the square of a small number is an even smaller number.

Discussion Questions

A. If a swarm of ants has a total momentum of zero, what can we conclude? What if their total kinetic energy is zero?

B. If all the air molecules in the room settled down in a thin film on the floor, would that violate conservation of momentum as well as conservation of energy?

C. A refrigerator has coils in back that get hot, and heat is molecular motion. These moving molecules have both energy and momentum. Why doesn't the refrigerator need to be tied to the wall to keep it from recoiling from the momentum it loses out the back?

4.2 Collisions in One Dimension

This Hubble Space Telescope photo shows a small galaxy (yellow blob in the lower right) that has collided with a larger galaxy (spiral near the center), producing a wave of star formation (blue track) due to the shock waves passing through the galaxies' clouds of gas. This is considered a collision in the physics sense, even though it is statistically certain that no star in either galaxy ever struck a star in the other. (This is because the stars are very small compared to the distances between them.)

Physicists employ the term “collision” in a broader sense than ordinary usage, applying it to any situation where objects interact for a certain period of time. A bat hitting a baseball, a radioactively emitted particle damaging DNA, and a gun and a bullet going their separate ways are all examples of collisions in this sense. Physical contact is not even required. A comet swinging past the sun on a hyperbolic orbit is considered to undergo a collision, even though it never touches the sun. All that matters is that the comet and the sun exerted gravitational forces on each other.

The reason for broadening the term “collision” in this way is that all of these situations can be attacked mathematically using the same conservation laws in similar ways. In the first example, conservation of momentum is all that is required.

Example: getting rear-ended

Question: Ms. Chang is rear-ended at a stop light by Mr. Nelson, and sues to make him pay her medical bills. He testifies that he was only going 35 miles per hour when he hit Ms. Chang. She thinks he was going much faster than that. The cars skidded together after the impact, and measurements of the length of the skid marks and the coefficient of friction show that their joint velocity immediately after the impact was 19 miles per hour. Mr. Nelson’s Nissan weighs 3100 pounds, and Ms. Chang ’s Cadillac weighs 5200 pounds. Is Mr. Nelson telling the truth?

Solution: Since the cars skidded together, we can write down the equation for conservation of momentum using only two velocities, v for Mr. Nelson’s velocity before the crash, and v' for their joint velocity afterward:

$$m_N v = m_N v' + m_C v' \quad .$$

Solving for the unknown, v, we find

$$v = \left(1 + \frac{m_C}{m_N}\right) v' \quad .$$

Although we are given the weights in pounds, a unit of force, the ratio of the masses is the same as the ratio of the weights, and we find v=51 miles per hour. He is lying.

The above example was simple because both cars had the same velocity afterward. In many one-dimensional collisions, however, the two objects do not stick. If we wish to predict the result of such a collision, conservation of momentum does not suffice, because both velocities after the collision are unknown, so we have one equation in two unknowns.

Conservation of energy can provide a second equation, but its application is not as straightforward, because kinetic energy is only the particular form of energy that has to do with motion. In many collisions, part of the kinetic energy that was present before the collision is used to create heat or sound, or to break the objects or permanently bending them. Cars, in fact, are carefully designed to crumple in a collision. Crumpling the car uses up energy, and that’s good because the goal is to get rid of all that kinetic energy in a relatively safe and controlled way. At the opposite extreme, a superball is “super” because it emerges from a collision with almost all its original kinetic energy, having only stored it briefly as potential energy while it was being squashed by the impact.

Collisions of the superball type, in which almost no kinetic energy is converted to other forms of energy, can thus be analyzed more thoroughly, because they have $KE_f=KE_i$, as opposed to the less useful inequality $KE_f<KE_i$ for a case like a tennis ball bouncing on grass.

Example: pool balls colliding head-on

Question: Two pool balls collide head-on, so that the collision is restricted to one dimension. Pool balls are constructed so as to lose as little kinetic energy as possible in a collision, so under the assumption that no kinetic energy is converted to any other form of energy, what can we predict about the results of such a collision?

Solution: Pool balls have identical masses, so we use the same symbol m for both. Conservation of energy and no loss of kinetic energy give us the two equations

$$mv_{1i} + mv_{2i} = mv_{1f} + mv_{2f}$$

$$\frac{1}{2}mv_{1i}^2 + \frac{1}{2}mv_{2i}^2 = \frac{1}{2}mv_{1f}^2 + \frac{1}{2}mv_{2f}^2$$

The masses and the factors of 1/2 can be divided out, and we eliminate the cumbersome subscripts by replacing the symbols v_{1i},... with the symbols A, B, C, and D.

$$A+B = C+D$$

$$A^2+B^2 = C^2+D^2 \quad .$$

A little experimentation with numbers shows that given values of A and B, it is impossible to find C and D that satisfy these equations unless C and D equal A and B, or C and D are the same as A and B but swapped around. An algebraic proof is given in the box on the left. In the special case where ball 2 is initially at rest, this tells us that ball 1 is stopped dead by the collision, and ball 2 heads off at the velocity originally possessed by ball 1. This behavior will be familiar to players of pool.

Algebraic Proof of the Result in the Example

The equation $A+B= C+D$ says that the change in one ball's velocity is equal and opposite to the change in the other's. We invent a symbol $x=C-A$ for the change in ball 1's velocity. The second equation can then be rewritten as

$$A^2+B^2 = (A+x)^2+(B-x)^2 \quad .$$

Squaring out the quantities in parentheses gives

$$A^2+B^2$$

$$= A^2+2Ax+x^2+B^2-2Bx+x^2 \quad ,$$

which simplifies to

$$0 = Ax-Bx+x^2 \quad .$$

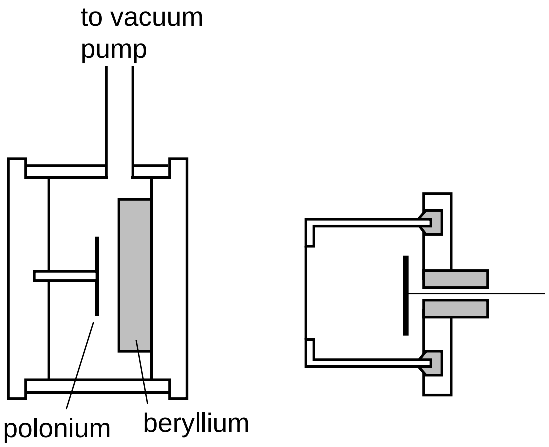
The equation has the trivial solution $x=0$, i.e. neither ball's velocity is changed, but this is physically impossible because the balls cannot travel through each other like ghosts. Assuming $x\neq0$, we can divide by x and solve for $x=B-A$. This means that ball 1 has gained an amount of velocity exactly sufficient to make it equal to ball 2's initial velocity, and vice-versa. The balls must have swapped velocities.

Often, as in the example above, the details of the algebra are the least interesting part of the problem, and considerable physical insight can be gained simply by counting the number of unknowns and comparing to the number of equations. Suppose a beginner at pool notices a case where her cue ball hits an initially stationary ball and stops dead. “Wow, what a good trick,” she thinks. “I bet I could never do that again in a million years.” But she tries again, and finds that she can't help doing it even if she doesn't want to. Luckily she has just learned about collisions in her physics course. Once she has written down the equations for conservation of energy and no loss of kinetic energy, she really doesn't have to complete the algebra. She knows that she has two equations in two unknowns, so there must be a well-defined solution. Once she has seen the result of one such collision, she knows that the same thing must happen every time. The same thing would happen with colliding marbles or croquet balls. It doesn't matter if the masses or velocities are different, because that just multiplies both equations by some constant factor.

The discovery of the neutron

This was the type of reasoning employed by James Chadwick in his 1932 discovery of the neutron. At the time, the atom was imagined to be made out of two types of fundamental particles, protons and electrons. The protons were far more massive, and clustered together in the atom's core, or nucleus. Attractive electrical forces caused the electrons to orbit the nucleus in circles, in much the same way that gravitational forces kept the planets from cruising out of the solar system. Experiments showed that the helium nucleus, for instance, exerted exactly twice as much electrical force on an electron as a nucleus of hydrogen, the smallest atom, and this was explained by saying that helium had two protons to hydrogen's one. The trouble was that according to this model, helium would have two electrons and two protons, giving it precisely twice the mass of a hydrogen atom with one of each. In fact, helium has about four times the mass of hydrogen.

Chadwick suspected that the helium nucleus possessed two additional particles of a new type, which did not participate in electrical forces at all, i.e. were electrically neutral. If these particles had very nearly the same mass as protons, then the four-to-one mass ratio of helium and hydrogen could be explained. In 1930, a new type of radiation was discovered that seemed to fit this description. It was electrically neutral, and seemed to be coming from the nuclei of light elements that had been exposed to other types of radiation. At this time, however, reports of new types of particles were a dime a dozen, and most of them turned out to be either clusters made of previously known particles or else previously known particles with higher energies. Many physicists believed that the "new" particle that had attracted Chadwick's interest was really a previously known particle called a gamma ray, which was electrically neutral. Since gamma rays have no mass, Chadwick decided to try to determine the new particle's mass and see if it was nonzero and approximately equal to the mass of a proton.



Chadwick's subatomic pool table. A disk of the naturally occurring metal polonium provides a source of radiation capable of kicking neutrons out of the beryllium nuclei. The type of radiation emitted by the polonium is easily absorbed by a few mm of air, so the air has to be pumped out of the left-hand chamber. The neutrons, Chadwick's mystery particles, penetrate matter far more readily, and fly out through the wall and into the chamber on the right, which is filled with nitrogen or hydrogen gas. When a neutron collides with a nitrogen or hydrogen nucleus, it kicks it out of its atom at high speed, and this recoiling nucleus then rips apart thousands of other atoms of the gas. The result is an electrical pulse that can be detected in the wire on the right. Physicists had already calibrated this type of apparatus so that they could translate the strength of the electrical pulse into the velocity of the recoiling nucleus. The whole apparatus shown in the figure would fit in the palm of your hand, in dramatic contrast to today's giant particle accelerators.

Unfortunately a subatomic particle is not something you can just put on a scale and weigh. Chadwick came up with an ingenious solution. The masses of the nuclei of the various chemical elements were already known, and techniques had already been developed for measuring the speed of a rapidly moving nucleus. He therefore set out to bombard samples of selected elements with the mysterious new particles. When a direct, head-on collision occurred between a mystery particle and the nucleus of one of the target atoms, the nucleus would be knocked out of the atom, and he would measure its velocity.

Suppose, for instance, that we bombard a sample of hydrogen atoms with the mystery particles. Since the participants in the collision are fundamental particles, there is no way for kinetic energy to be converted into heat or any other form of energy, and Chadwick thus had two equations in three unknowns:

equation #1: conservation of momentum
equation #2: no loss of kinetic energy
unknown #1: mass of the mystery particle
unknown #2: initial velocity of the mystery particle
unknown #3: final velocity of the mystery particle

The number of unknowns is greater than the number of equations, so there is no unique solution. But by creating collisions with nuclei of another element, nitrogen, he gained two more equations at the expense of only one more unknown:

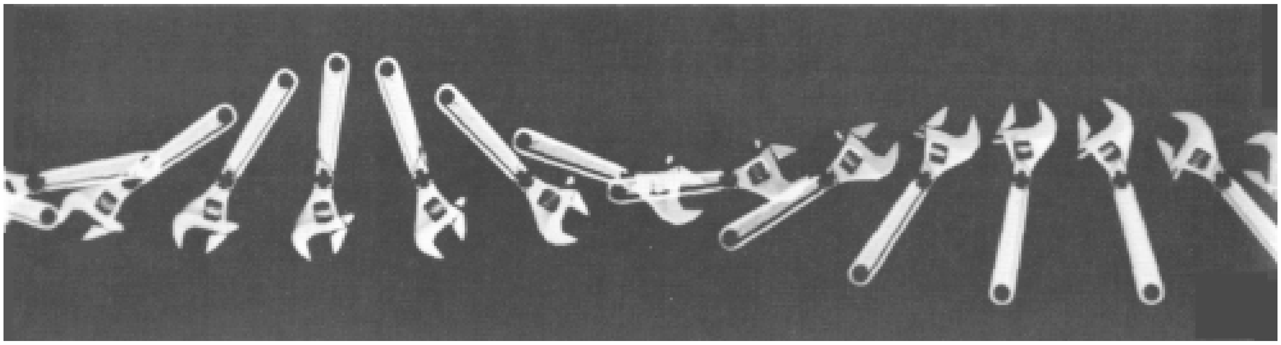
equation #3: conservation of momentum in the new collision
equation #4: no loss of kinetic energy in the new collision
unknown #4: final velocity of the mystery particle in the new collision

He was thus able to solve for all the unknowns, including the mass of the mystery particle, which was indeed within 1% of the mass of a proton. He named the new particle the neutron, since it is electrically neutral.

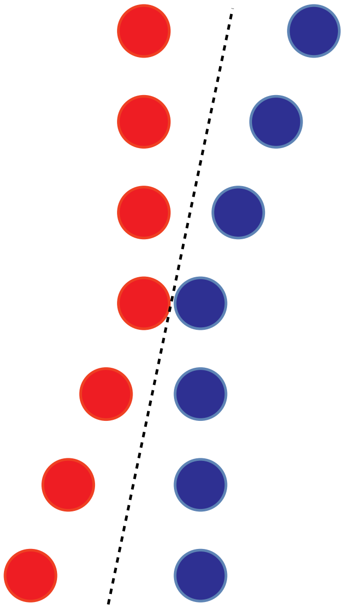
Discussion Question

Good pool players learn to make the cue ball spin, which can cause it not to stop dead in a head-on collision with a stationary ball. If this does not violate the laws of physics, what hidden assumption was there in the example above?

4.3* Relationship of Momentum to the Center of Mass



In this multiple-flash photograph, we see the wrench from above as it flies through the air, rotating as it goes. Its center of mass, marked with the black cross, travels along a straight line, unlike the other points on the wrench, which execute loops.



Two hockey pucks collide. Their mutual center of mass traces the straight path shown by the dashed line.

We have already discussed the idea of the center of mass in the first book of this series, but using the concept of momentum we can now find a mathematical method for defining the center of mass, explain why the motion of an object's center of mass usually exhibits simpler motion than any other point, and gain a very simple and powerful way of understanding collisions.

The first step is to realize that the center of mass concept can be applied to systems containing more than one object. Even something like a wrench, which we think of as one object, is really made of many atoms. The center of mass is particularly easy to visualize in the case shown on the left, where two identical hockey pucks collide. It is clear on grounds of symmetry that their center of mass must be at the midpoint between them. After all, we previously defined the center of mass as the balance point, and if the two hockey pucks were joined with a very lightweight rod whose own mass was negligible, they would obviously balance at the midpoint. It doesn't matter that the hockey pucks are two separate objects. It is still true that the motion of their center of mass is exceptionally simple, just like that of the wrench's center of mass.

The x coordinate of the hockey pucks' center of mass is thus given by $x_{cm}=(x_1+x_2)/2$, i.e. the arithmetic average of their x coordinates. Why is its motion so simple? It has to do with conservation of momentum. Since the hockey pucks are not being acted on by any net external force, they constitute a closed system, and their total momentum is conserved. Their total momentum is

$$\begin{aligned} mv_1+mv_2 &= m(v_1+v_2) \\ &= m\left(\frac{\Delta x_1}{\Delta t}+\frac{\Delta x_2}{\Delta t}\right) \\ &= \frac{m}{\Delta t}\Delta\left(x_1+x_2\right) \\ &= m\frac{2\Delta x_{cm}}{\Delta t} \\ &= m_{total}v_{cm,x} \end{aligned}$$

In other words, the total momentum of the system is the same as if all its mass was concentrated at the center of mass point. Since the total momentum is conserved, the *x* component of the center of mass's velocity vector cannot change. The same is also true for the other components, so the center of mass must move along a straight line at constant speed.

The above relationship between the total momentum and the motion of the center of mass applies to any system, even if it is not closed.

total momentum related to center of mass motion

The total momentum of any system is related to its total mass and the velocity of its center of mass by the equation

$$\boldsymbol{p}_{\text{total}} = m_{\text{total}}\boldsymbol{v}_{\text{cm}} \quad .$$

What about a system containing objects with unequal masses, or containing more than two objects? The reasoning above can be generalized to a weighted average

$$x_{\text{cm}} = \frac{m_1 x_1 + m_2 x_2 + \ldots}{m_1 + m_2 + \ldots} \quad ,$$

with similar equations for the *y* and *z* coordinates.

Momentum in different frames of reference

Absolute motion is supposed to be undetectable, i.e. the laws of physics are supposed to be equally valid in all inertial frames of reference. If we first calculate some momenta in one frame of reference and find that momentum is conserved, and then rework the whole problem in some other frame of reference that is moving with respect to the first, the numerical values of the momenta will all be different. Even so, momentum will still be conserved. All that matters is that we work a single problem in one consistent frame of reference.

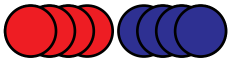
One way of proving this is to apply the equation $\boldsymbol{p}_{\text{total}}=m_{\text{total}}\boldsymbol{v}_{\text{cm}}$. If the velocity of frame B relative to frame A is $\boldsymbol{v}_{\text{BA}}$, then the only effect of changing frames of reference is to change $\boldsymbol{v}_{\text{cm}}$ from its original value to $\boldsymbol{v}_{\text{cm}}+\boldsymbol{v}_{\text{BA}}$. This adds a constant onto the momentum vector, which has no effect on conservation of momentum.

The center of mass frame of reference

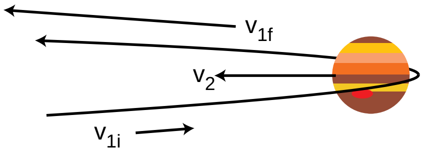
A particularly useful frame of reference in many cases is the frame that moves along with the center of mass, called the center of mass (c.m.) frame. In this frame, the total momentum is zero. The following examples show how the center of mass frame can be a powerful tool for simplifying our understanding of collisions.

Example: a collision of pool balls viewed in the c.m. frame

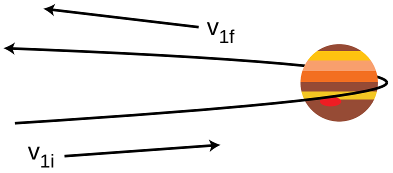
If you move your head so that your eye is always above the point halfway in between the two pool balls, you are viewing things in the center of mass frame. In this frame, the balls come toward the center of mass at equal speeds. By symmetry, they must therefore recoil at equal speeds along the lines on which they entered. Since the balls have essentially swapped paths in the center of mass frame, the same must also be true in any other frame. This is the same result that required laborious algebra to prove previously without the concept of the center of mass frame.



Moving our head so that we are always looking down from right above the center of mass, we observe the collision of the two hockey pucks in the center of mass frame.



(a) The slingshot effect viewed in the sun's frame of reference. Jupiter is moving to the left, and the collision is head-on.



(b) The slingshot viewed in the frame of the center of mass of the Jupiter-spacecraft system.

Example: the slingshot effect

It is a counterintuitive fact that a spacecraft can pick up speed by swinging around a planet, if arrives in the opposite direction compared to the planet's motion. Although there is no physical contact, we treat the encounter as a one-dimensional collision, and analyze it in the center of mass frame. Since Jupiter is so much more massive than the spacecraft, the center of mass is essentially fixed at Jupiter's center, and Jupiter has zero velocity in the center of mass frame, as shown in figure (b). The c.m. frame is moving to the left compared to the sun-fixed frame used in (a), so the spacecraft's initial velocity is greater in this frame.

Things are simpler in the center of mass frame, because it is more symmetric. In the sun-fixed frame, the incoming leg of the encounter is rapid, because the two bodies are rushing toward each other, while their separation on the outbound leg is more gradual, because Jupiter is trying to catch up. In the c.m. frame, Jupiter is sitting still, and there is perfect symmetry between the incoming and outgoing legs, so by symmetry we have $v_{1f}=-v_{1i}$. Going back to the sun-fixed frame, the spacecraft's final velocity is increased by the frames' motion relative to each other. In the sun-fixed frame, the spacecraft's velocity has increased greatly.

The result can also be understood in terms of work and energy. In Jupiter's frame, Jupiter is not doing any work on the spacecraft as it rounds the back of the planet, because the motion is perpendicular to the force. But in the sun's frame, the spacecraft's velocity vector at the same moment has a large component to the left, so Jupiter is doing work on it.

Discussion Question

A. Make up a numerical example of two unequal masses moving in one dimension at constant velocity, and verify the equation $p_{total}=m_{total}v_{cm}$ over a time interval of one second.

B. A more massive tennis racquet or baseball bat makes the ball fly off faster. Explain why this is true, using the center of mass frame. For simplicity, assume that the racquet or bat is simply sitting still before the collision, and that the hitter's hands do not make any force large enough to have a significant effect over the short duration of the impact.

4.4 Momentum Transfer

The rate of change of momentum

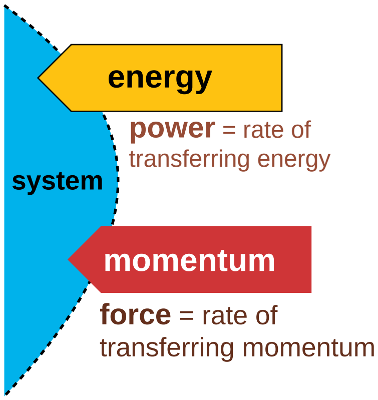
As with conservation of energy, we need a way to measure and calculate the transfer of momentum into or out of a system when the system is not closed. In the case of energy, the answer was rather complicated, and entirely different techniques had to be used for measuring the transfer of mechanical energy (work) and the transfer of heat by conduction. For momentum, the situation is far simpler.

In the simplest case, the system consists of a single object acted on by a constant external force. Since it is only the object's velocity that can change, not its mass, the momentum transferred is

$$\Delta \boldsymbol{p} = m\Delta \boldsymbol{v}$$

which with the help of $\boldsymbol{a}=\boldsymbol{F}/m$ and the constant-acceleration equation $\boldsymbol{a}=\Delta\boldsymbol{v}/\Delta t$ becomes

$$\begin{aligned}\Delta \boldsymbol{p} &= m\boldsymbol{a}\Delta t \\ &= \boldsymbol{F}\Delta t \quad .\end{aligned}$$


energy
power = rate of transferring energy
system
momentum
force = rate of transferring momentum

Thus the rate of transfer of momentum, i.e. the number of kg·m/s absorbed per second, is simply the external force,

$$\boldsymbol{F} = \Delta\boldsymbol{p}/\Delta t \quad .$$

[relationship between the force on an object and the rate of change of its momentum; valid only if the force is constant]

This equation is really just a restatement of Newton's second law, and in fact Newton originally stated it this way. As shown in the diagram, the relationship between force and momentum is directly analogous to that between power and energy.

The situation is not materially altered for a system composed of many objects. There may be forces between the objects, but the internal forces cannot change the system's momentum — if they did, then removing the external forces would result in a closed system that was able to change its own momentum, violating conservation of energy. The equation above becomes

$$\boldsymbol{F}_{\text{total}} = \Delta\boldsymbol{p}_{\text{total}}/\Delta t \quad .$$

[relationship between the total external force on a system and the rate of change of its total momentum; valid only if the force is constant]

Example: walking into a lamppost
Question: Starting from rest, you begin walking, bringing your momentum up to 100 kg·m/s. You walk straight into a lamppost. Why is the momentum change of -100 kg·m/s so much more painful than the change of +100 kg·m/s when you started walking?
Solution: The situation is one-dimensional, so we can dispense with the vector notation. It probably takes you about 1 s to speed up initially, so the ground's force on you is $F=\Delta p/\Delta t\approx$100 N. Your impact with the lamppost, however, is over in the blink of an eye, say 1/10 s or less. Dividing by this much smaller Δt gives a much larger force, perhaps thousands of newtons. (The negative sign simply indicates that the force is in the opposite direction.)

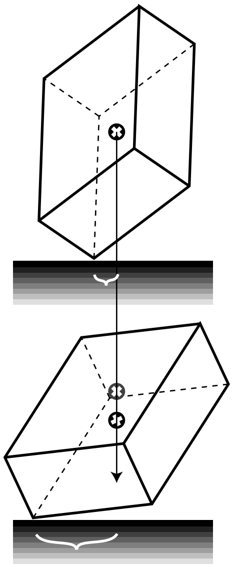
This is also the principle of airbags in cars. The time required for the airbag to decelerate your head is fairly long, the time required for your face to travel 20 or 30 cm. Without an airbag, your face would have been hitting the dashboard, and the time interval would have been the much shorter time taken by your skull to move a couple of centimeters while your face compressed. Note that either way, the same amount of mechanical work has to be done on your head: enough to eliminate all its kinetic energy.

Example: ion drive for spacecraft

Question: The ion drive of the Deep Space 1 spacecraft, pictured earlier in the chapter, produces a thrust of 90 mN (millinewtons). It carries about 80 kg of reaction mass, which it ejects at a speed of 30000 m/s. For how long can the engine continue supplying this amount of thrust before running out of reaction mass to shove out the back?

Solution: Solving the equation $\mathbf{F}=\Delta\mathbf{p}/\Delta t$ for the unknown Δt, and treating force and momentum as scalars since the problem is one-dimensional, we find

$$\begin{aligned}\Delta t &= \frac{\Delta p}{F}\\ &= \frac{m_{\text{exhaust}}\Delta v_{\text{exhaust}}}{F}\\ &= \frac{(80\ \text{kg})(30000\ \text{m/s})}{0.090\ \text{N}}\\ &= 2.7\text{x}10^7\ \text{s}\\ &= 300\ \text{days}\end{aligned}$$



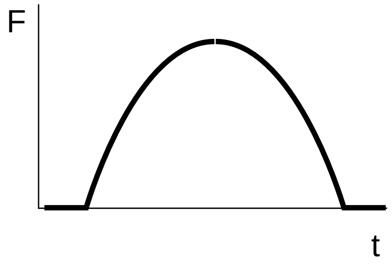
Example: a toppling box

If you place a box on a frictionless surface, it will fall over with a very complicated motion that is hard to predict in detail. We know, however, that its center of mass moves in the same direction as its momentum vector points. There are two forces, a normal force and a gravitational force, both of which are vertical. (The gravitational force is actually many gravitational forces acting on all the atoms in the box.) The total force must be vertical, so the momentum vector must be purely vertical too, and the center of mass travels vertically. This is true even if the box bounces and tumbles. [Based on an example by Kleppner and Kolenkow.]

The area under the force-time graph

Few real collisions involve a constant force. For example, when a tennis ball hits a racquet, the strings stretch and the ball flattens dramatically. They are both acting like springs that obey Hooke's law, which says that the force is proportional to the amount of stretching or flattening. The force is therefore small at first, ramps up to a maximum when the ball is about to reverse directions, and ramps back down again as the ball is on its way back out. The equation $F=\Delta p/\Delta t$, derived under the assumption of constant acceleration, does not apply here, and the force does not even have a single well-defined numerical value that could be plugged in to the equation.

As with similar-looking equations such as $v=\Delta p/\Delta t$, the equation $F=\Delta p/\Delta t$ is correctly generalized by saying that the force is the slope of the p-t graph.



Conversely, if we wish to find Δp from a graph such as the one shown on the left, one approach would be to divide the force by the mass of the ball, rescaling the F axis to create a graph of acceleration versus time. The area under the acceleration-versus-time graph gives the change in velocity, which can then be multiplied by the mass to find the change in momentum. An unnecessary complication was introduced, however, because we began by dividing by the mass and ended by multiplying by it. It would have made just as much sense to find the area under the original *F-t* graph, which would have given us the momentum change directly.

Discussion Question

Many collisions, like the collision of a bat with a baseball, appear to be instantaneous. Most people also would not imagine the bat and ball as bending or being compressed during the collision. Consider the following possibilities:

(1) The collision is instantaneous.

(2) The collision takes a finite amount of time, during which the ball and bat retain their shapes and remain in contact.

(3) The collision takes a finite amount of time, during which the ball and bat are bending or being compressed.

How can two of these be ruled out based on energy or momentum considerations?

4.5 Momentum in Three Dimensions

In this section we discuss how the concepts applied previously to one-dimensional situations can be used as well in three dimensions. Often vector addition is all that is needed to solve a problem:

Example: an explosion

Question: Astronomers observe the planet Mars as the Martians fight a nuclear war. The Martian bombs are so powerful that they rip the planet into three separate pieces of liquified rock, all having the same mass. If one fragment flies off with velocity components v_{1x}=0, v_{1y}=1.0x10^4 km/hr, and the second with v_{2x}=1.0x10^4 km/hr, v_{2y}=0, what is the magnitude of the third one's velocity?

Solution: We work the problem in the center of mass frame, in which the planet initially had zero momentum. After the explosion, the vector sum of the momenta must still be zero. Vector addition can be done by adding components, so

$$mv_{1x} + mv_{2x} + mv_{3x} = 0 \quad \text{, and}$$
$$mv_{1y} + mv_{2y} + mv_{3y} = 0 \quad \text{,}$$

where we have used the same symbol m for all the terms, because the fragments all have the same mass. The masses can be eliminated by dividing each equation by m, and we find

$$v_{3x} = -1.0\text{x}10^4 \text{ km/hr}$$
$$v_{3y} = -1.0\text{x}10^4 \text{ km/hr}$$

which gives a magnitude of

$$\begin{aligned} |\mathbf{v}_3| &= \sqrt{v_{3x}^2 + v_{3y}^2} \\ &= 1.4\text{x}10^4 \text{ km/hr} \end{aligned}$$

The center of mass

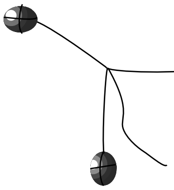
In three dimensions, we have the vector equations

$$\boldsymbol{F}_{\text{total}} = \Delta \boldsymbol{p}_{\text{total}}/\Delta t$$

and

$$\boldsymbol{p}_{\text{total}} = \mathrm{m}_{\text{total}}\,\boldsymbol{v}_{\text{cm}} \quad .$$

The following is an example of their use.



Example: the bola

The bola, similar to the North American lasso, is used by South American gauchos to catch small animals by tangling up their legs in the three leather thongs. The motion of the whirling bola through the air is extremely complicated, and would be a challenge to analyze mathematically. The motion of its center of mass, however, is much simpler. The only forces on it are gravitational, so

$$\boldsymbol{F}_{\text{total}} = m_{\text{total}}\boldsymbol{g} \quad .$$

Using the equation $\boldsymbol{F}_{\text{total}} = \Delta\boldsymbol{p}_{\text{total}}/\Delta t$, we find

$$\Delta\boldsymbol{p}_{\text{total}}/\Delta t = m_{\text{total}}\boldsymbol{g} \quad ,$$

and since the mass is constant, the equation $\boldsymbol{p}_{\text{total}} = m_{\text{total}}\,\boldsymbol{v}_{\text{cm}}$ allows us to change this to

$$m_{\text{total}}\Delta\boldsymbol{v}_{\text{cm}}/\Delta t = m_{\text{total}}\boldsymbol{g} \quad .$$

The mass cancels, and $\Delta\boldsymbol{v}_{\text{cm}}/\Delta t$ is simply the acceleration of the center of mass, so

$$\boldsymbol{a}_{\text{cm}} = \boldsymbol{g} \quad .$$

In other words, the motion of the system is the same as if all its mass was concentrated at and moving with the center of mass. The bola has a constant downward acceleration equal to $\boldsymbol{g}$, and flies along the same parabola as any other projectile thrown with the same initial center of mass velocity. Throwing a bola with the correct rotation is presumably a difficult skill, but making it hit its target is no harder than it is with a ball or a single rock.

[Based on an example by Kleppner & Kolenkow.]

Counting equations and unknowns

Counting equations and unknowns is just as useful as in one dimension, but every object's momentum vector has three components, so an unknown momentum vector counts as three unknowns. Conservation of momentum is a single vector equation, but it says that all three components of the total momentum vector stay constant, so we count it as three equations. Of course if the motion happens to be confined to two dimensions, then we need only count vectors as having two components.

Example: a two-car crash with sticking

Suppose two cars collide, stick together, and skid off together. If we know the cars' initial momentum vectors, we can count equations and unknowns as follows:

unknown #1: x component of cars' final, total momentum
unknown #2: y component of cars' final, total momentum
equation #1: conservation of the total p_x
equation #2: conservation of the total p_y

Since the number of equations equals the number of unknowns, there must be one unique solution for their total momentum vector after the crash. In other words, the speed and direction at which their common center of mass moves off together is unaffected by factors such as whether the cars collide center-to-center or catch each other a little off-center.

Example: shooting pool

Two pool balls collide, and as before we assume there is no decrease in the total kinetic energy, i.e. no energy converted from KE into other forms. As in the previous example, we assume we are given the initial velocities and want to find the final velocities. The equations and unknowns are:

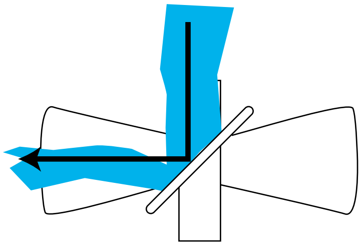
unknown #1: x component of ball #1's final momentum
unknown #2: y component of ball #1's final momentum
unknown #3: x component of ball #2's final momentum
unknown #4: y component of ball #2's final momentum
equation #1: conservation of the total p_x
equation #2: conservation of the total p_y
equation #3: no decrease in total KE

Note that we do not count the balls' final kinetic energies as unknowns, because knowing the momentum vector, one can always find the velocity and thus the kinetic energy. The number of equations is less than the number of unknowns, so no unique result is guaranteed. This is what makes pool an interesting game. By aiming the cue ball to one side of the target ball you can have some control over the balls' speeds and directions of motion after the collision.

It is not possible, however, to choose any combination of final speeds and directions. For instance, a certain shot may give the correct direction of motion for the target ball, making it go into a pocket, but may also have the undesired side-effect of making the cue ball go in a pocket.

Calculations with the momentum vector

The following example illustrates how a force is required to change the direction of the momentum vector, just as one would be required to change its magnitude.



Example: a turbine

Question: In a hydroelectric plant, water flowing over a dam drives a turbine, which runs a generator to make electric power. The figure shows a simplified physical model of the water hitting the turbine, in which it is assumed that the stream of water comes in at a 45° angle with respect to the turbine blade, and bounces off at a 90° angle at nearly the same speed. The water flows at a rate R, in units of kg/s, and the speed of the water is v. What are the magnitude and direction of the water's force on the turbine?

Solution: In a time interval Δt, the mass of water that strikes the blade is $R\Delta t$, and the magnitude of its initial momentum is $mv=mR\Delta t$. The water's final momentum vector is of the same magnitude, but in the perpendicular direction. By Newton's third law, the water's force on the blade is equal and opposite to the blade's force on the water. Since the force is constant, we can use the equation

$$\mathbf{F}_{\text{blade on water}} = \Delta \mathbf{p}_{\text{water}}/\Delta t \quad .$$

Choosing the x axis to be to the right and the y axis to be up, this can be broken down into components as

$$\begin{aligned} F_{\text{blade on water,x}} &= \Delta p_{\text{water,x}}/\Delta t \\ &= (-mR\Delta t\text{-}0)/\Delta t \\ &= -mR \end{aligned}$$

and

$$\begin{aligned} F_{\text{blade on water,y}} &= \Delta p_{\text{water,y}}/\Delta t \\ &= (0-(-mR\Delta t))/\Delta t \\ &= mR \quad . \end{aligned}$$

The water's force on the blade thus has

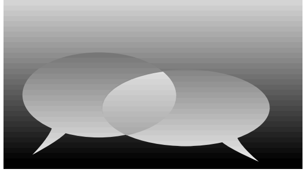
$$\begin{aligned} F_{\text{water on blade,x}} &= mR \\ F_{\text{water on blade,y}} &= -mR \quad . \end{aligned}$$

In situations like this, it is always a good idea to check that the result makes sense physically. The x component of the water's force on the blade is positive, which is correct since we know the blade will be pushed to the right. The y component is negative, which also makes sense because the water must push the blade down. The magnitude of the water's force on the blade is

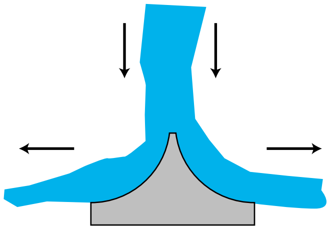
$$|\mathbf{F}_{\text{water on blade}}| = \sqrt{2}\; mR$$

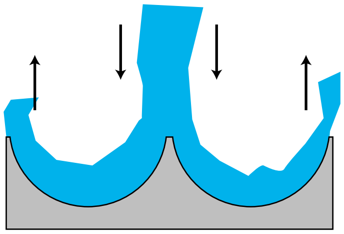
and its direction is at a 45-degree angle down and to the right.

Discussion Questions



The figures show a jet of water striking two different objects. How does the total downward force compare in the two cases? How could this fact be used to create a better waterwheel? (Such a waterwheel is known as a Pelton wheel.)





4.6∫ Applications of Calculus

By now you will have learned to recognized the circumlocutions I use in the sections without calculus in order to introduce calculus-like concepts without using the notation, terminology, or techniques of calculus. It will therefore come as no surprise to you that the rate of change of momentum can be represented with a derivative,

$$\mathbf{F}_{\text{total}} = \frac{\mathrm{d}\mathbf{p}_{\text{total}}}{\mathrm{d}t} \quad .$$

And of course the business about the area under the *F-t* curve is really an integral, $\Delta p_{\text{total}} = \int F_{\text{total}}\mathrm{d}t$, which can be made into an integral of a vector in the more general three-dimensional case:

$$\Delta\mathbf{p}_{\text{total}} = \int \mathbf{F}_{\text{total}}\mathrm{d}t \quad .$$

In the case of a material object that is neither losing nor picking up mass, these are just trivially rearranged versions of familiar equations, e.g. $F = m\frac{\mathrm{d}v}{\mathrm{d}t}$ rewritten as $F = \frac{\mathrm{d}(mv)}{\mathrm{d}t}$. The following is a less trivial example, where *F*=*ma* alone would not have been very easy to work with.

Example: rain falling into a moving cart

Question: If 1 kg/s of rain falls vertically into a 10-kg cart that is rolling without friction at an initial speed of 1.0 m/s, what is the effect on the speed of the cart when the rain first starts falling?

Solution: The rain and the cart make horizontal forces on each other, but there is no external horizontal force on the rain-plus-cart system, so the horizontal motion obeys

$$F = \frac{\mathrm{d}(mv)}{\mathrm{d}t} = 0$$

We use the product rule to find

$$0 = \frac{\mathrm{d}m}{\mathrm{d}t}v + \mathrm{m}\frac{\mathrm{d}v}{\mathrm{d}t} \quad .$$

We are trying to find how *v* changes, so we solve for d*v*/d*t*,

$$\begin{aligned}\frac{\mathrm{d}v}{\mathrm{d}t} &= -\frac{v}{m}\frac{\mathrm{d}m}{\mathrm{d}t} \\ &= -\left(\frac{1 \text{ m/s}}{10 \text{ kg}}\right)(1 \text{ kg/s}) \\ &= -0.1 \text{ m/s}^2 \quad .\end{aligned}$$

(This is only at the moment when the rain starts to fall.)

Finally we note that there are cases where *F*=*ma* is not just less convenient than *F*=d*p*/d*t* but in fact *F*=*ma* is wrong and *F*=d*p*/d*t* is right. A good example is the formation of a comet's tail by sunlight. We cannot use *F*=*ma* to describe this process, since we are dealing with a collision of light with matter, whereas Newton's laws only apply to matter. The equation *F*=d*p*/d*t*, on the other hand, allows us to find the force experienced by an atom of gas in the comet's tail if we know the rate at which the momentum vectors of light rays are being turned around by reflection from the molecule.

Summary

Selected Vocabulary

momentum .... a measure of motion, equal to $m\mathbf{v}$ for material objects
collision .... an interaction between moving objects that lasts for a certain time
center of mass .... the balance point or average position of the mass in a system

Notation

$\mathbf{p}$ .... the momentum vector
cm .... center of mass, as in x_{cm}, a_{cm}, etc.

Standard Terminology and Notation Not Used in This Book

impulse .... the amount of momentum transferred, $\Delta\mathbf{p}$
$\mathbf{I}$, $\mathbf{J}$ .... impulse
elastic collision .... one in which no KE is converted into other forms of energy
inelastic collision .... one in which some KE is converted to other forms of energy

Summary

If two objects interact via a force, Newton's third law guarantees that any change in one's velocity vector will be accompanied by a change in the other's which is in the opposite direction. Intuitively, this means that if the two objects are not acted on by any external force, they cannot cooperate to change their overall state of motion. This can be made quantitative by saying that the quantity $m_1\mathbf{v}_1+m_2\mathbf{v}_2$ must remain constant as long as the only forces are the internal ones between the two objects. This is a conservation law, called the conservation of momentum, and like the conservation of energy, it has evolved over time to include more and more phenomena unknown at the time the concept was invented. The momentum of a material object is

$$\mathbf{p} = m\mathbf{v} \quad ,$$

but this is more like a standard for comparison of momenta rather than a definition. For instance, light has momentum, but has no mass, and the above equation is not the right equation for light. The law of conservation of momentum says that the total momentum of any closed system, i.e. the vector sum of the momentum vectors of all the things in the system, is a constant.

An important application of the momentum concept is to collisions, i.e. interactions between moving objects that last for a certain amount of time while the objects are in contact or near each other. Conservation of momentum tells us that certain outcomes of a collision are impossible, and in some cases may even be sufficient to predict the motion after the collision. In other cases, conservation of momentum does not provide enough equations to find all the unknowns. In some collisions, such as the collision of a superball with the floor, very little kinetic energy is converted into other forms of energy, and this provides one more equation, which may suffice to predict the outcome.

The total momentum of a system can be related to its total mass and the velocity of its center of mass by the equation

$$\mathbf{p}_{total} = m_{total}\mathbf{v}_{cm} \quad .$$

The center of mass, introduced on an intuitive basis in book 1 as the "balance point" of an object, can be generalized to any system containing any number of objects, and is defined mathematically as the weighted average of the positions of all the parts of all the objects,

$$x_{cm} = \frac{m_1x_1 + m_2x_2 + ...}{m_1 + m_2 + ...}$$

with similar equations for the y and z coordinates.

The frame of reference moving with the center of mass of a closed system is always a valid inertial frame, and many problems can be greatly simplified by working them in the inertial frame. For example, any collision between two objects appears in the c.m. frame as a head-on one-dimensional collision.

When a system is not closed, the rate at which momentum is transferred in or out is simply the total force being exerted externally on the system. If the force is constant,

$$\mathbf{F}_{total} = \Delta\mathbf{p}_{total}/\Delta t \quad .$$

When the force is not constant, the force equals the slope of the tangent line on a graph of p versus t, and the change in momentum equals the area under the F-t graph.

Homework Problems

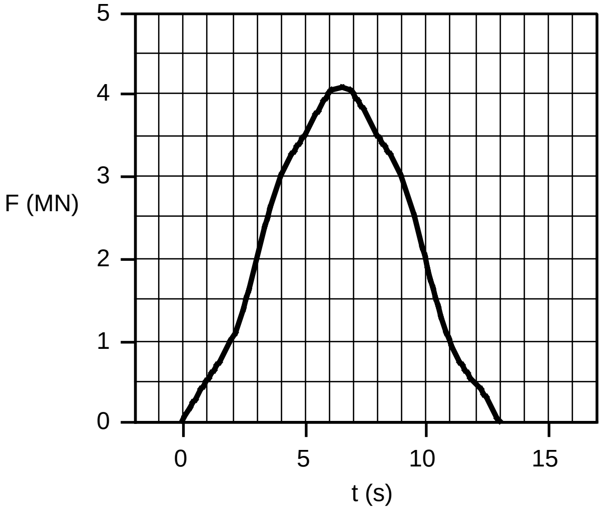
1. Derive a formula expressing the kinetic energy of an object in terms of its momentum and mass.

2. Two people in a rowboat wish to move around without causing the boat to move. What should be true about their total momentum? Explain.

3✓. A learjet traveling due east at 300 mi/hr collides with a jumbo jet which was heading southwest at 150 mi/hr. The jumbo jet's mass is five times greater than that of the learjet. When they collide, the learjet sticks into the fuselage of the jumbo jet, and they fall to earth together. Their engines stop functioning immediately after the collision. On a map, what will be the direction from the location of the collision to the place where the wreckage hits the ground? (Give an angle.)

4. A bullet leaves the barrel of a gun with a kinetic energy of 90 J. The gun barrel is 50 cm long. The gun has a mass of 4 kg, the bullet 10 g.
(a✓) Find the bullet's final velocity.
(b✓) Find the bullet's final momentum.
(c) Find the momentum of the recoiling gun.
(d✓) Find the kinetic energy of the recoiling gun, and explain why the recoiling gun does not kill the shooter.

5✓. The graph below shows the force, in meganewtons, exerted by a rocket engine on the rocket as a function of time. If the rocket's mass if 4000 kg, at what speed is the rocket moving when the engine stops firing? Assume it goes straight up, and neglect the force of gravity, which is much less than a meganewton.



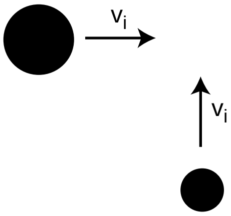
6. Cosmic rays are particles from outer space, mostly protons and atomic nuclei, that are continually bombarding the earth. Most of them, although they are moving extremely fast, have no discernible effect even if they hit your body, because their masses are so small. Their energies vary, however,

S A solution is given in the back of the book. ★ A difficult problem.
✓ A computerized answer check is available. ∫ A problem that requires calculus.

and a very small minority of them have extremely large energies. In some cases the energy is as much as several Joules, which is comparable to the KE of a well thrown rock! If you are in a plane at a high altitude and are so incredibly unlucky as to be hit by one of these rare ultra-high-energy cosmic rays, what would you notice, the momentum imparted to your body, the energy dissipated in your body as heat, or both? Base your conclusions on numerical estimates, not just random speculation. (At these high speeds, one should really take into account the deviations from Newtonian physics described by Einstein's special theory of relativity. Don't worry about that, though.)

7. Show that for a body made up of many *equal* masses, the equation for the center of mass becomes a simple average of all the positions of the masses.

8 S. The figure shows a view from above of a collision about to happen between two air hockey pucks sliding without friction. They have the same speed, v_i, before the collision, but the big puck is 2.3 times more massive than the small one. Their sides have sticky stuff on them, so when they collide, they will stick together. At what angle will they emerge from the collision? In addition to giving a numerical answer, please indicate by drawing on the figure how your angle is defined.



Problem 8.

9 ∫. A flexible rope of mass m and length L slides without friction over the edge of a table. Let x be the length of the rope that is hanging over the edge at a given moment in time.

(a) Show that x satisfies the equation of motion $\frac{d^2x}{dt^2} = \frac{g}{L}x$. [Hint: Use F=dp/dt, which allows you to ignore internal forces in the rope.]

(b) Give a physical explanation for the fact that a larger value of x on the right-hand side of the equation leads to a greater value of the acceleration on the left side.

(c) When we take the second derivative of the function $x(t)$ we are supposed to get essentially the same function back again, except for a constant out in front. The function e^x has the property that it is unchanged by differentiation, so it is reasonable to look for solutions to this problem that are of the form $x=be^{ct}$, where b and c are constants. Show that this does indeed provide a solution for two specific values of c (and for any value of b).

(d) Show that the sum of any two solutions to the equation of motion is also a solution.

(e) Find the solution for the case where the rope starts at rest at t=0 with some nonzero value of x.

10. A very massive object with velocity v collides head-on with an object at rest whose mass is very small. No kinetic energy is converted into other forms. Prove that the low-mass object recoils with velocity $2v$.

11S. When the contents of a refrigerator cool down, the changed molecular speeds imply changes in both momentum and energy. Why, then, does a fridge transfer *power* through its radiator coils, but not *force*?

12. A 10-kg bowling ball moving at 2.0 m/s hits a 1.0-kg bowling pin, which is initially at rest. The other pins are all gone already, and the collision is head-on, so that the motion is one-dimensional. Assume that negligible amounts of heat and sound are produced. Find the velocity of the pin immediately after the collision.

13 ∫★. A rocket ejects exhaust with an exhaust velocity u. The rate at which the exhaust mass is used (mass per unit time) is b. We assume that the rocket accelerates in a straight line starting from rest, and that no external forces act on it. Let the rocket's initial mass (fuel plus the body and payload) be m_i, and m_f be its final mass, after all the fuel is used up. (a) Find the rocket's final velocity, v, in terms of u, m_i, and m_f. (b) A typical exhaust velocity for chemical rocket engines is 4000 m/s. Estimate the initial mass of a rocket that could accelerate a one-ton payload to 10% of the speed of light, and show that this design won't work. (For the sake of the estimate, ignore the mass of the fuel tanks.)

14 S. A firework shoots up into the air, and just before it explodes it has a certain momentum and kinetic energy. What can you say about the momenta and kinetic energies of the pieces immediately after the explosion? [Based on a problem from PSSC Physics.]

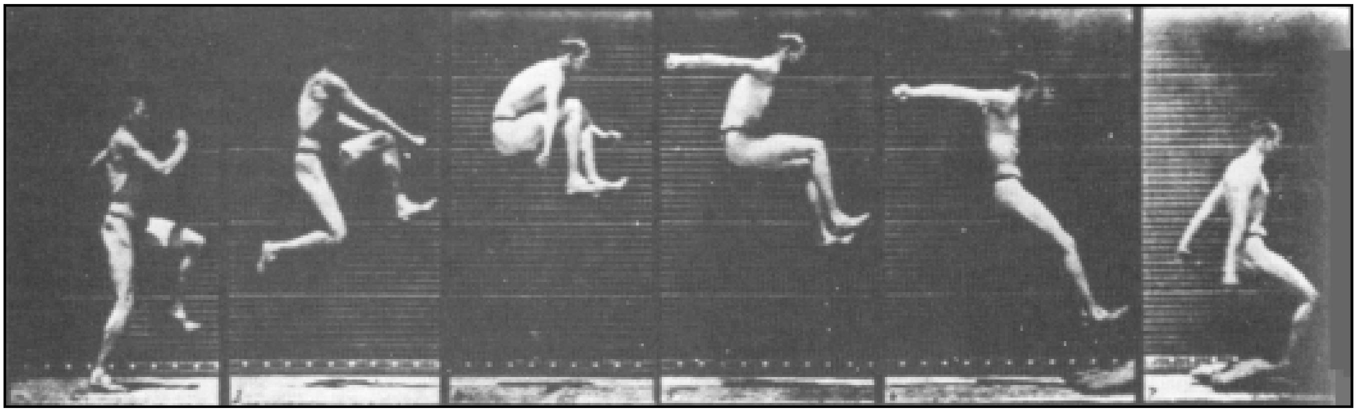
15 ★S. Suppose a system consisting of pointlike particles has a total kinetic energy K_{cm} measured in the center-of-mass frame of reference. Since they are pointlike, they cannot have any energy due to internal motion. (a) Prove that in a different frame of reference, moving with velocity $\boldsymbol{u}$ relative to the center-of-mass frame, the total kinetic energy equals $K_{cm}+M|\boldsymbol{u}|^2/2$, where M is the total mass. [Hint: You can save yourself a lot of writing if you express the total kinetic energy using the dot product.] (b) Use this to prove that if energy is conserved in one frame of reference, then it is conserved in every frame of reference. The total energy equals the total kinetic energy plus the sum of the potential energies due to the particles' interactions with each other, which we assume depends only on the distance between particles. [For a simpler numerical example, see problem 13 in ch. 1.]

A tornado touches down in Spring Hill, Kansas, May 20, 1957.

5 Conservation of Angular Momentum

"Sure, and maybe the sun won't come up tomorrow." Of course, the sun only appears to go up and down because the earth spins, so the cliche should really refer to the unlikelihood of the earth's stopping its rotation abruptly during the night. Why can't it stop? It wouldn't violate conservation of momentum, because the earth's rotation doesn't add anything to its momentum. While California spins in one direction, some equally massive part of India goes the opposite way, canceling its momentum. A halt to Earth's rotation would entail a drop in kinetic energy, but that energy could simply by converted into some other form, such as heat.

Other examples along these lines are not hard to find. A hydrogen atom spins at the same rate for billions of years. A high-diver who is rotating when he comes off the board does not need to make any physical effort to continue rotating, and indeed would be unable to stop rotating before he hit the water.



These observations have the hallmarks of a conservation law:

A closed system is involved. Nothing is making an effort to twist the earth, the hydrogen atom, or the high-diver. They are isolated from rotation-changing influences, i.e. they are closed systems.

Something remains unchanged. There appears to be a numerical quantity for measuring rotational motion such that the total amount of that quantity remains constant in a closed system.

Something can be transferred back and forth without changing the total amount: In the photo of the old-fashioned high jump above, the jumper wants to get his feet out in front of him so he can keep from doing a “face plant” when he lands. Bringing his feet forward would involve a certain quantity of counterclockwise rotation, but he didn’t start out with any rotation when he left the ground. Suppose we consider counterclockwise as positive and clockwise as negative. The only way his legs can acquire some positive rotation is if some other part of his body picks up an equal amount of negative rotation. This is why he swings his arms up behind him, clockwise.

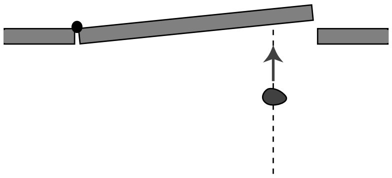
What numerical measure of rotational motion is conserved? Car engines and old-fashioned LP records have speeds of rotation measured in rotations per minute (r.p.m.), but the number of rotations per minute (or per second) is not a conserved quantity. A twirling figure skater, for instance, can pull her arms in to increase her r.p.m.’s. The first section of this chapter deals with the numerical definition of the quantity of rotation that results in a valid conservation law.

5.1 Conservation of Angular Momentum

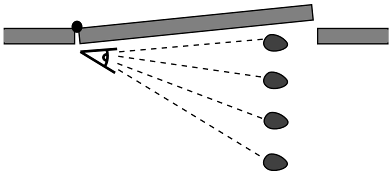
When most people think of rotation, they think of a solid object like a wheel rotating in a circle around a fixed point. Examples of this type of rotation, called rigid rotation or rigid-body rotation, include a spinning top, a seated child’s swinging leg, and a helicopter’s spinning propeller. Rotation, however, is a much more general phenomenon, and includes noncircular examples such as a comet in an elliptical orbit around the sun, or a cyclone, in which the core completes a circle more quickly than the outer parts.

If there is a numerical measure of rotational motion that is a conserved quantity, then it must include nonrigid cases like these, since nonrigid rotation can be traded back and forth with rigid rotation. For instance, there is a trick for finding out if an egg is raw or hardboiled. If you spin a hardboiled egg and then stop it briefly with your finger, it stops dead. But if you do the same with a raw egg, it springs back into rotation because the soft interior was still swirling around within the momentarily motionless shell. The pattern of flow of the liquid part is presumably very complex and nonuniform due to the asymmetric shape of the egg and the different consistencies of the yolk and the white, but there is apparently some way to describe the liquid's total amount of rotation with a single number, of which some percentage is given back to the shell when you release it.

The best strategy is to devise a way of defining the amount of rotation of a single small part of a system. The amount of rotation of a system such as a cyclone will then be defined as the total of all the contributions from its many small parts.



An overhead view of a piece of putty being thrown at a door. Even though the putty is neither spinning nor traveling along a curve, we must define it has having some kind of "rotation" because it is able to make the door rotate.



As seen by someone standing at the axis, the putty changes its angular position. We therefore define it as having angular momentum.

The quest for a conserved quantity of rotation even requires us to broaden the rotation concept to include cases where the motion doesn't repeat or even curve around. If you throw a piece of putty at a door, the door will recoil and start rotating. The putty was traveling straight, not in a circle, but if there is to be a general conservation law that can cover this situation, it appears that we must describe the putty as having had some "rotation," which it then gave up to the door. The best way of thinking about it is to attribute rotation to any moving object or part of an object that changes its angle in relation to the axis of rotation. In the putty-and-door example, the hinge of the door is the natural point to think of as an axis, and the putty changes its angle as seen by someone standing at the hinge. For this reason, the conserved quantity we are investigating is called *angular momentum*. The symbol for angular momentum can't be "*a*" or "*m*," since those are used for acceleration and mass, so the symbol L is arbitrarily chosen instead.

Imagine a 1-kg blob of putty, thrown at the door at a speed of 1 m/s, which hits the door at a distance of 1 m from the hinge. We define this blob to have 1 unit of angular momentum. When it hits the door, it will give up most of its own angular momentum to the door, which will recoil and start rotating.

Experiments show, not surprisingly, that a 2-kg blob thrown in the same way makes the door rotate twice as fast, so the angular momentum of the putty blob must be proportional to mass,

$$L \propto m \quad .$$

Similarly, experiments show that doubling the velocity of the blob will have a similar doubling effect on the result, so its angular momentum must be proportional to its velocity as well,

$$L \propto mv \quad .$$

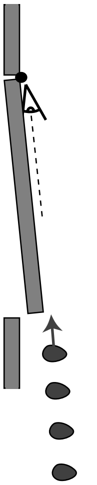
You have undoubtedly had the experience of approaching a closed door with one of those bar-shaped handles on it and pushing on the wrong side, the side close to the hinges. You feel like an idiot, because you have so little leverage that you can hardly budge the door. The same would be true with the putty blob. Experiments would show that the amount of rotation the blob can give to the door is proportional to the distance, *r*, from the axis of rotation, so angular momentum must be proportional to *r* as well,

$$L \propto mvr \quad .$$

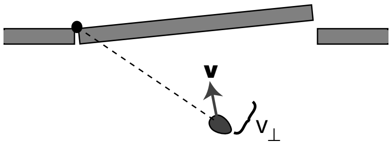
We are almost done, but there is one missing ingredient. We know on grounds of symmetry that a putty ball thrown directly inward toward the hinge will have no angular momentum to give to the door. After all, there would not even be any way to decide whether the ball's rotation was clockwise or counterclockwise in this situation. It is therefore only the component of the blob's velocity vector perpendicular to the door that should be counted in its angular momentum,

$$L = m\, v_\perp\, r \quad .$$

More generally, $v_\perp$ should be thought of as the component of the object's velocity vector that is perpendicular to the line joining the object to the axis of rotation.



A putty blob thrown directly at the axis has no angular motion, and therefore no angular momentum. It will not cause the door to rotate.



Only the component of the velocity vector perpendicular to the line connecting the object to the axis should be counted into the definition of angular momentum.

We find that this equation agrees with the definition of the original putty blob as having one unit of angular momentum, and we can now see that the units of angular momentum are (kg·m/s) · m, i.e. $\text{kg}\cdot\text{m}^2/\text{s}$. This gives us a way of calculating the angular momentum of any material object or any system consisting of material objects:

angular momentum of a material object

The angular momentum of a moving particle is

$$L = m\, v_\perp\, r \quad ,$$

where m is its mass, $v_\perp$ is the component of its velocity perpendicular to the line joining it to the axis of rotation, and r is its distance from the axis of rotation. Positive and negative signs of angular momentum are used to describe opposite directions of rotation.

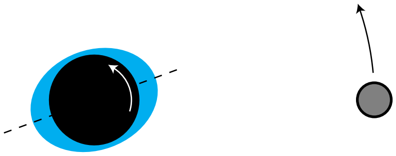
The angular momentum of a finite-sized object or a system of many objects is found by dividing it up into many small parts, applying the equation to each part, and adding to find the total amount of angular momentum.

Note that *r* is not necessarily the radius of a circle. (As implied by the qualifiers, matter isn't the only thing that can have angular momentum. Light can also have angular momentum, and the above equation would not apply to light.)

Conservation of angular momentum has been verified over and over again by experiment, and is now believed to be one of the three most fundamental principles of physics, along with conservation of energy and momentum.

A figure skater pulls in her arms so that she can execute a spin more rapidly.

Example: a figure skater pulls her arms in
When a figure skater is twirling, there is very little friction between her and the ice, so she is essentially a closed system, and her angular momentum is conserved. If she pulls her arms in, she is decreasing r for all the atoms in her arms. It would violate conservation of angular momentum if she then continued rotating at the same speed, i.e. taking the same amount of time for each revolution, her arms' contributions to her angular momentum would have decreased, and no other part of her would have increased its angular momentum. This is impossible because it would violate conservation of angular momentum. If her total angular momentum is to remain constant, the decrease in r for her arms must be compensated for by an overall increase in her rate of rotation. That is, by pulling her arms in, she substantially reduces the time for each rotation.



A view of the earth-moon system from above the north pole. All distances have been highly distorted for legibility. The earth's rotation is counterclockwise from this point of view (arrow). The moon's gravity creates a bulge on the side near it, because its gravitational pull is stronger there, and an "anti-bulge" on the far side, since its gravity there is weaker. For simplicity, let's focus on the tidal bulge closer to the moon. Its frictional force is trying to slow down the earth's rotation, so its force on the earth's solid crust is toward the bottom of the figure. By Newton's third law, the crust must thus make a force on the bulge which is toward the top of the figure. This causes the bulge to be pulled forward at a slight angle, and the bulge's gravity therefore pulls the moon forward, accelerating its orbital motion about the earth and flinging it outward.

Example: Earth's slowing rotation and the receding moon
As noted in chapter 1, the earth's rotation is actually slowing down very gradually, with the kinetic energy being dissipated as heat by friction between the land and the tidal bulges raised in the seas by the earth's gravity. Does this mean that angular momentum is not really perfectly conserved? No, it just means that the earth is not quite a closed system by itself. If we consider the earth and moon as a system, then the angular momentum lost by the earth must be gained by the moon somehow. In fact very precise measurements of the distance between the earth and the moon have been carried out by bouncing laser beams off of a mirror left there by astronauts, and these measurements show that the moon is receding from the earth at a rate of 2 millimeters per year! The moon's greater value of r means that it has a greater angular momentum, and the increase turns out to be exactly the amount lost by the earth. In the days of the dinosaurs, the days were significantly shorter, and the moon was closer and appeared bigger in the sky.
But what force is causing the moon to speed up, drawing it out into a larger orbit? It is the gravitational forces of the earth's tidal bulges. The effect is described qualitatively in the caption of the figure. The result would obviously be extremely difficult to calculate directly, and this is one of those situations where a conservation law allows us to make precise quantitative statements about the outcome of a process when the calculation of the process itself would be prohibitively complex.

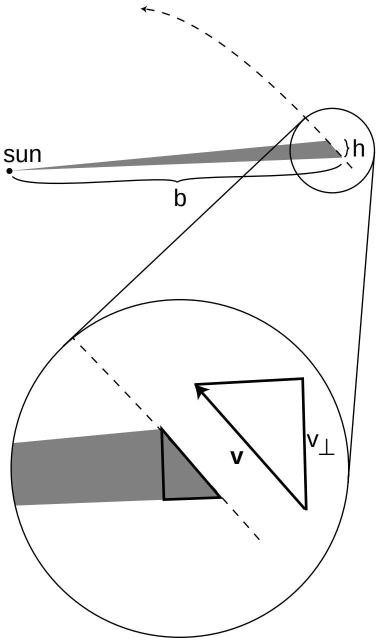
Restriction to rotation in a plane

Is angular momentum a vector or a scalar? It does have a direction in space, but it's a direction of rotation, not a straight-line direction like the directions of vectors such as velocity or force. It turns out that there is a way of defining angular momentum as a vector, but in this book the examples will be confined to a single plane of rotation, i.e. effectively two-dimensional situations. In this special case, we can choose to visualize the plane of rotation from one side or the other, and to define clockwise and counterclockwise rotation as having opposite signs of angular momentum.

Discussion Question

Conservation of plain old momentum, ***p***, can be thought of as the greatly expanded and modified descendant of Galileo's original principle of inertia, that no force is required to keep an object in motion. The principle of inertia is counterintuitive, and there are many situations in which it appears superficially that a force *is* needed to maintain motion, as maintained by Aristotle. Think of a situation in which conservation of angular momentum, *L*, also seems to be violated, making it seem incorrectly that something external must act on a closed system to keep its angular momentum from "running down."

5.2 Angular Momentum in Planetary Motion



We now discuss the application of conservation of angular momentum to planetary motion, both because of its intrinsic importance and because it is a good way to develop a visual intuition for angular momentum.

Kepler's law of equal areas states that the area swept out by a planet in a certain length of time is always the same. Angular momentum had not been invented in Kepler's time, and he did not even know the most basic physical facts about the forces at work. He thought of this law as an entirely empirical and unexpectedly simple way of summarizing his data, a rule that succeeded in describing and predicting how the planets sped up and slowed down in their elliptical paths. It is now fairly simple, however, to show that the equal area law amounts to a statement that the planet's angular momentum stays constant.

There is no simple geometrical rule for the area of a pie wedge cut out of an ellipse, but if we consider a very short time interval, as shown in the figure, the shaded shape swept out by the planet is very nearly a triangle. We do know how to compute the area of a triangle. It is one half the product of the base and the height:

$$\text{area} = \tfrac{1}{2}bh \quad .$$

We wish to relate this to angular momentum, which contains the variables r and $v_\perp$. If we consider the sun to be the axis of rotation, then the variable r is identical to the base of the triangle, r=b. Referring to the magnified portion of the figure, $v_\perp$ can be related to h, because the two right triangles are similar:

$$\frac{h}{\text{distance traveled}} = \frac{v_\perp}{|\boldsymbol{v}|}$$

The area can thus be rewritten as

$$\text{area} = \tfrac{1}{2}r\,\frac{v_\perp(\text{distance traveled})}{|\boldsymbol{v}|} \quad .$$

The distance traveled equals $|\boldsymbol{v}|\Delta t$, so this simplifies to

$$\text{area} = \tfrac{1}{2}r\,v_\perp\,\Delta t \quad .$$

We have found the following relationship between angular momentum and the rate at which area is swept out:

$$L = 2m\frac{\text{area}}{\Delta t} \quad .$$

The factor of 2 in front is simply a matter of convention, since any conserved quantity would be an equally valid conserved quantity if you multiplied it by a constant. The factor of *m* was not relevant to Kepler, who did not know the planets' masses, and who was only describing the motion of one planet at a time.

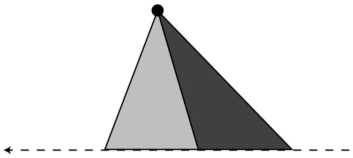
We thus find that Kepler's equal-area law is equivalent to a statement that the planet's angular momentum remains constant. But wait, why should it remain constant? — the planet is not a closed system, since it is being acted on by the sun's gravitational force. There are two valid answers. The first is that it is actually the total angular momentum of the sun plus the planet that is conserved. The sun, however, is millions of times more massive than the typical planet, so it accelerates very little in response to the planet's gravitational force. It is thus a good approximation to say that the sun doesn't move at all, so that no angular momentum is transferred between it and the planet.

The second answer is that to change the planet's angular momentum requires not just a force but a force applied in a certain way. In section 5.4 we discuss the transfer of angular momentum by a force, but the basic idea here is that a force directly in toward the axis does not change the angular momentum.

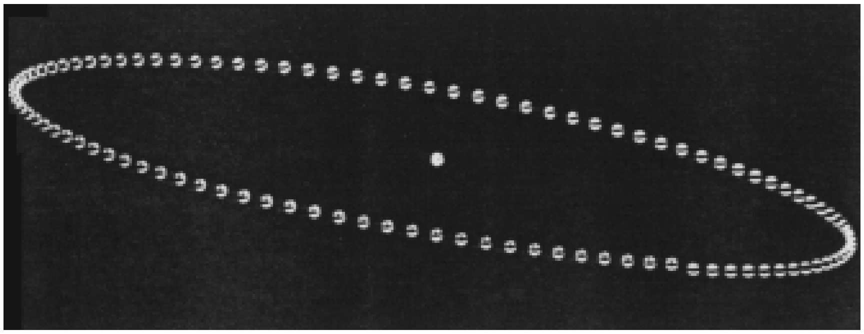
Discussion Questions

A. Suppose an object is simply traveling in a straight line at constant speed. If we pick some point not on the line and call it the axis of rotation, is area swept out by the object at a constant rate?

B. The figure below is a strobe photo of a pendulum bob, taken from underneath the pendulum looking straight up. The black string can't be seen in the photograph. The bob was given a slight sideways push when it was released, so it did not swing in a plane. The bright spot marks the center, i.e. the position the bob would have if it hung straight down at us. Does the bob's angular momentum appear to remain constant if we consider the center to be the axis of rotation?



Discussion question A.



Discussion question B.

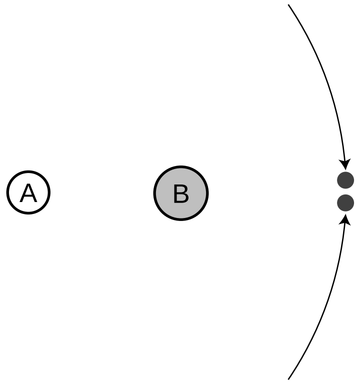
5.3 Two Theorems About Angular Momentum

With plain old momentum, ***p***, we had the freedom to work in any inertial frame of reference we liked. The same object could have different values of momentum in two different frames, if the frames were not at rest with respect to each other. Conservation of momentum, however, would be true in either frame. As long as we employed a single frame consistently throughout a calculation, everything would work.

The same is true for angular momentum, and in addition there is an ambiguity that arises from the definition of an axis of rotation. For a wheel, the natural choice of an axis of rotation is obviously the axle, but what about an egg rotating on its side? The egg has an asymmetric shape, and thus no clearly defined geometric center. A similar issue arises for a cyclone, which does not even have a sharply defined shape, or for a complicated machine with many gears. The following theorem, the first of two presented in this section without proof, explains how to deal with this issue. Although I have put descriptive titles above both theorems, they have no generally accepted names.

choice of axis theorem

It is entirely arbitrary what point one defines as the axis for purposes of calculating angular momentum. If a closed system's angular momentum is conserved when calculated with one choice of axis, then it will also be conserved for any other choice of axis. Likewise, any inertial frame of reference may be used.



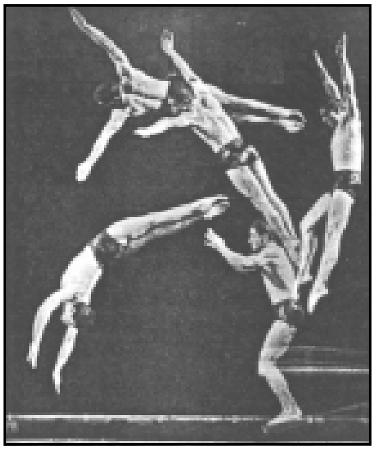
Example: colliding asteroids described with different axes

Observers on planets A and B both see the two asteroids colliding. The asteroids are of equal mass and their impact speeds are the same. Astronomers on each planet decide to define their own planet as the axis of rotation. Planet A is twice as far from the collision as planet B. The asteroids collide and stick. For simplicity, assume planets A and B are both at rest.

With planet A as the axis, the two asteroids have the same amount of angular momentum, but one has positive angular momentum and the other has negative. Before the collision, the total angular momentum is therefore zero. After the collision, the two asteroids will have stopped moving, and again the total angular momentum is zero. The total angular momentum both before and after the collision is zero, so angular momentum is conserved if you choose planet A as the axis.

The only difference with planet B as axis is that *r* is smaller by a factor of two, so all the angular momenta are halved. Even though the angular momenta are different than the ones calculated by planet A, angular momentum is still conserved.

The earth spins on its own axis once a day, but simultaneously travels in its circular one-year orbit around the sun, so any given part of it traces out a complicated loopy path. It would seem difficult to calculate the earth's angular momentum, but it turns out that there is an intuitively appealing shortcut: we can simply add up the angular momentum due to its spin plus that arising from its center of mass's circular motion around the sun. This is a special case of the following general theorem:



Everyone has a strong tendency to think of the diver as rotating about his own center of mass. However, he is flying in an arc, and he also has angular momentum because of this motion.

spin theorem

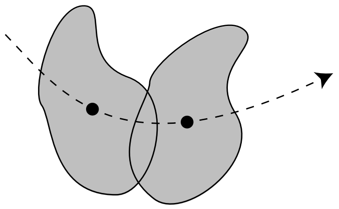
An object's angular momentem with respect to some outside axis A can be found by adding up two parts:

(1) The first part is the object's angular momentum found by using its own center of mass as the axis, i.e. the angular momentum the object has because it is spinning.

(2) The other part equals the angular momentum that the object would have with respect to the axis A if it had all its mass concentrated at and moving with its center of mass.

Example: a system with its center of mass at rest

In the special case of an object whose center of mass is at rest, the spin theorem implies that the object's angular momentum is the same regardless of what axis we choose. (This is an even stronger statement than the choice of axis theorem, which only guarantees that angular momentum is conserved for any given choice of axis, without specifying that it is the same for all such choices.)



This rigid object has angular momentum both because it is spinning about its center of mass and because it is moving through space.

Example: angular momentum of a rigid object

Question: A motorcycle wheel has almost all its mass concentrated at the outside. If the wheel has mass m and radius r, and the time required for one revolution is T, what is the spin part of its angular momentum?

Solution: This is an example of the commonly encountered special case of rigid motion, as opposed to the rotation of a system like a hurricane in which the different parts take different amounts of time to go around. We don't really have to go through a laborious process of adding up contributions from all the many parts of a wheel, because they are all at about the same distance from the axis, and are all moving around the axis at about the same speed. The velocity is all perpendicular to the spokes,

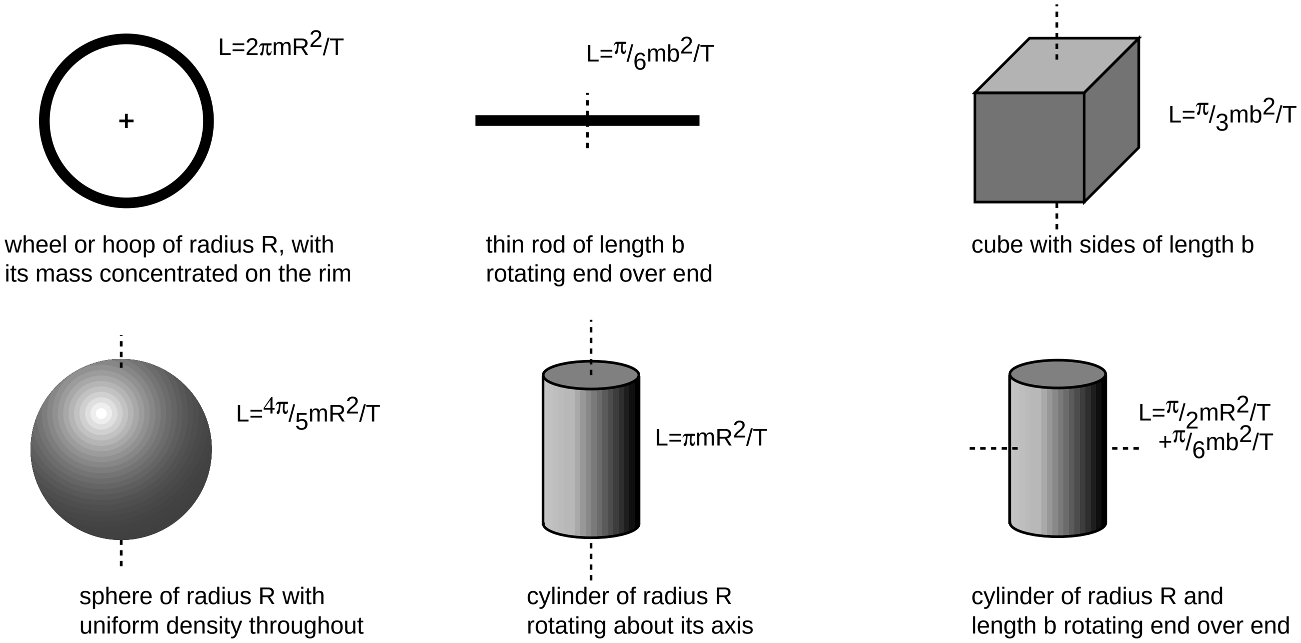
$$\begin{aligned} v_\perp &= v \\ &= \text{(circumference)}/T \\ &= 2\pi\text{r/T} \end{aligned}$$

and the angular momentum of the wheel about its center is

$$\begin{aligned} L &= mv_\perp r \\ &= m(2\pi r/T)\text{r} \\ &= 2\pi mr^2/T\ . \end{aligned}$$

Note that although the factors of 2π in this expression is peculiar to a wheel with its mass concentrated on the rim, the proportionality to m/T would have been the same for any other rigidly rotating object. Although an object with a noncircular shape does not have a radius, it is also true in general that angular momentum is proportional to the square of the object's size for fixed values of m and T. For instance doubling an object's size doubles both the $v_\perp$ and r factors in the contribution of each of its parts to the total angular momentum, resulting in an overall factor of four increase.

The figure shows some examples of angular momenta of various shapes rotating about their centers of mass. The equations for their angular momenta were derived using calculus, using methods discussed in supplement 2-7. Do not memorize these equations!



wheel or hoop of radius R, with its mass concentrated on the rim

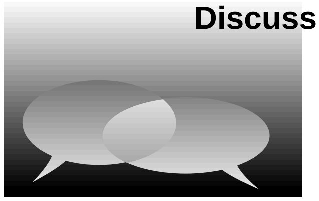
thin rod of length b rotating end over end

cube with sides of length b

sphere of radius R with uniform density throughout

cylinder of radius R rotating about its axis

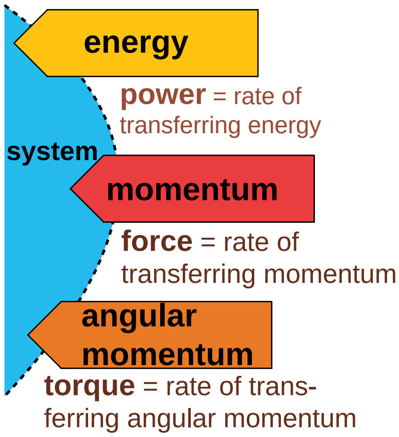
cylinder of radius R and length b rotating end over end



Discussion Question

In the example of the colliding asteroids, suppose planet A was moving toward the top of the page, at the same speed as the bottom asteroid. How would planet A's astronomers describe the angular momenta of the asteroids? Would angular momentum still be conserved?

5.4 Torque: the Rate of Transfer of Angular Momentum


energy
power = rate of transferring energy
system
momentum
force = rate of transferring momentum
angular momentum
torque = rate of transferring angular momentum

Force can be interpreted as the rate of transfer of momentum. The equivalent in the case of angular momentum is called *torque* (rhymes with "fork"). Where force tells us how hard we are pushing or pulling on something, torque indicates how hard we are twisting on it. Torque is represented by the Greek letter tau, τ, and the rate of change of an object's angular momentum equals the total torque acting on it,

$$\tau_{\text{total}} = \Delta L/\Delta t \quad .$$

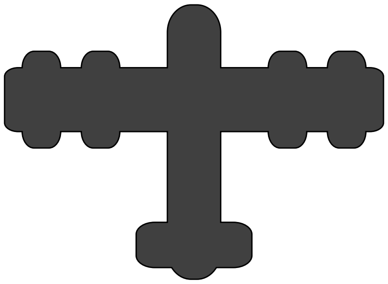
(If the angular momentum does not change at a constant rate, the total torque equals the slope of the tangent line on a graph of L versus t.)

As with force and momentum, it often happens that angular momentum recedes into the background and we focus our interest on the torques. The torque-focused point of view is exemplified by the fact that many scientifically untrained but mechanically apt people know all about torque, but none of them have heard of angular momentum. Car enthusiasts eagerly compare engines' torques, and there is a tool called a torque wrench which allows one to apply a desired amount of torque to a screw and avoid overtightening it.

Torque distinguished from force

Of course a force is necessary in order to create a torque — you can't twist a screw without pushing on the wrench — but force and torque are two different things. One distinction between them is direction. We use positive and negative signs to represent forces in the two possible directions along a line. The direction of a torque, however, is clockwise or counterclockwise, not a linear direction.

The other difference between torque and force is a matter of leverage. A given force applied at a door's knob will change the door's angular momentum twice as rapidly as the same force applied halfway between the knob and the hinge. The same amount of force produces different amounts of torque in these two cases.



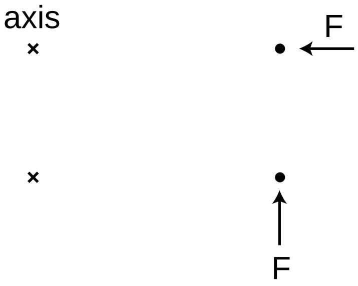
The plane's four engines produce zero total torque but not zero total force.

It is possible to have a zero total torque with a nonzero total force. An airplane with four jet engines would be designed so that their forces are balanced on the left and right. Their forces are all in the same direction, but the clockwise torques of two of the engines are canceled by the counterclockwise torques of the other two, giving zero total torque.

Conversely we can have zero total force and nonzero total torque. A merry-go-round's engine needs to supply a nonzero torque on it to bring it up to speed, but there is zero total force on it. If there was not zero total force on it, its center of mass would accelerate!

Relationship between force and torque

How do we calculate the amount of torque produced by a given force? Since it depends on leverage, we should expect it to depend on the distance between the axis and the point of application of the force. We'll derive an equation relating torque to force for a particular very simple situation, and state without proof that the equation actually applies to all situations.



The simple physical situation we use to derive an equation for torque. A force that points directly in at or out away from the axis produces neither clockwise nor counterclockwise angular momentum. A force in the perpendicular direction does transfer angular momentum.

Consider a pointlike object which is initially at rest at a distance r from the axis we have chosen for defining angular momentum. We first observe that a force directly inward or outward, along the line connecting the axis to the object, does not impart any angular momentum to the object.

A force perpendicular to the line connecting the axis and the object does, however, make the object pick up angular momentum. Newton's second law gives

$$a = F/m \quad ,$$

and assuming for simplicity that the force is constant, the constant acceleration equation $a=\Delta v/\Delta t$ allows us to find the velocity the object acquires after a time Δt,

$$\Delta v = F\Delta t/m \quad .$$

We are trying to relate force to a change in angular momentum, so we multiply both sides of the equation by mr to give

$$m\Delta vr = F\Delta tr$$

$$\Delta L = F\Delta tr \quad .$$

Dividing by Δt gives the torque:

$$\Delta L/\Delta t = Fr$$

$$\tau = Fr \quad .$$

If a force acts at an angle other than 0 or 90° with respect to the line joining the object and the axis, it would be only the component of the force perpendicular to the line that would produce a torque,

$$\tau = F_\perp r \quad .$$

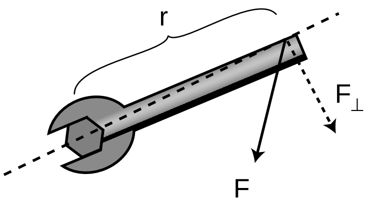
Although this result was proven under a simplified set of circumstances, it is more generally valid.

relationship between force and torque

The rate at which a force transfers angular momentum to an object, i.e. the torque produced by the force, is given by

$$|\tau| = r\,|F_\perp| \quad ,$$

where r is the distance from the axis to the point of application of the force, and $F_\perp$ is the component of the force that is perpendicular to the line joining the axis to the point of application.



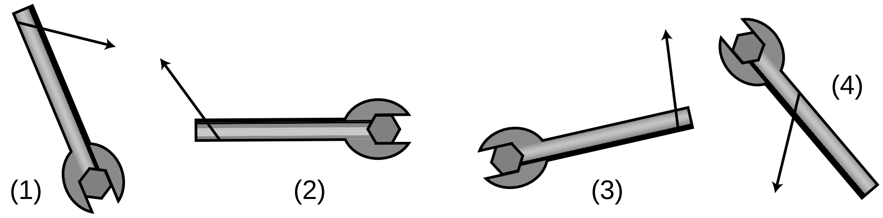
The geometric relationships refered to in the relationship between force and torque.

The equation is stated with absolute value signs because the positive and negative signs of force and torque indicate different things, so there is no useful relationship between them. The sign of the torque must be found by physical inspection of the case at hand.

From the equation, we see that the units of torque can be written as newtons multiplied by meters. Metric torque wrenches are calibrated in N·m, but American ones use foot-pounds, which is also a unit of distance multiplied by a unit of force. We know from our study of mechanical work that newtons multiplied by meters equal joules, but torque is a completely different quantity from work, and nobody writes torques with units of joules, even though it would be technically correct.

Self-Check

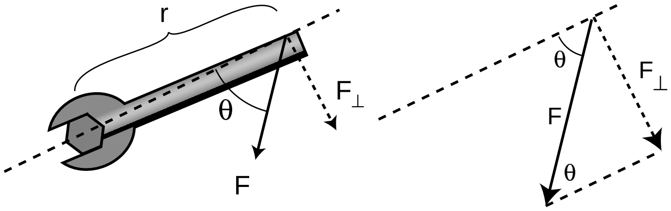
Compare the magnitudes and signs of the four torques shown in the figure. [Answer on next page.]



Example: how torque depends on the direction of the force

Question: How can the torque applied to the wrench in the figure be expressed in terms of r, $|\mathbf{F}|$, and the angle θ?

Solution: The force vector and its $F_\perp$ component form the hypotenuse and one leg of a right triangle,

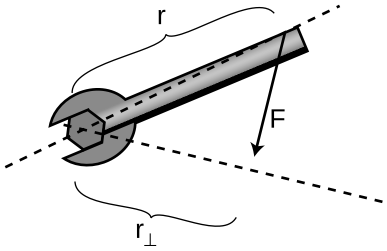


and the interior angle opposite to $F_\perp$ equals θ. The absolute value of $F_\perp$ can thus be expressed as

$$F_\perp = |\mathbf{F}| \sin\theta \quad ,$$

leading to

$$|\tau| = r|\mathbf{F}| \sin\theta \quad .$$



Sometimes torque can be more neatly visualized in terms of the quantity $r_\perp$ shown in the figure on the left, which gives us a third way of expressing the relationship between torque and force:

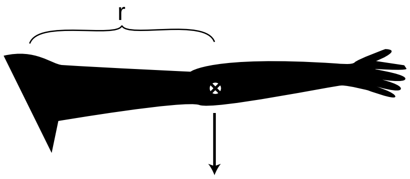
$$|\tau| = r_\perp |F| \quad .$$

Of course you would not want to go and memorize all three equations for torque. Starting from any one of them you could easily derive the other two using trigonometry. Familiarizing yourself with them can however clue you in to easier avenues of attack on certain problems.

The torque due to gravity

Up until now we've been thinking in terms of a force that acts at a single point on an object, such as the force of your hand on the wrench. This is of course an approximation, and for an extremely realistic calculation of your hand's torque on the wrench you might need to add up the torques exerted by each square millimeter where your skin touches the wrench. This is seldom necessary. But in the case of a gravitational force, there is never any single point at which the force is applied. Our planet is exerting a separate tug on every brick in the Leaning Tower of Pisa, and the total gravitational torque on the tower is the sum of the torques contributed by all the little forces. Luckily there is a trick that allows us to avoid such a massive calculation. It turns out that for purposes of computing the total gravitational torque on an object, you can get the right answer by just pretending that the whole gravitational force acts at the object's center of mass.

[Answer to self-check on previous page.] 1, 2, and 4 all have the same sign, because they are trying to twist the wrench clockwise. The sign of 3 is opposite to the signs of 1, 2, and 4. The magnitude of 3 is the greatest, since it has a large r and the force is nearly all perpendicular to the wrench. Torques 1 and 2 are the same because they have the same values of r and $F_\perp$. Torque 4 is the smallest, due to its small r.



Example: gravitational torque on an outstretched arm

Question: Your arm has a mass of 3.0 kg, and its center of mass is 30 cm from your shoulder. What is the gravitational torque on your arm when it is stretched out horizontally to one side, taking the shoulder to be the axis?

Solution: The total gravitational force acting on your arm is

$$|\boldsymbol{F}| = (3.0\text{ kg})(9.8\text{ m/s}^2) = 29\text{ N} \quad .$$

For the purpose of calculating the gravitational torque, we can treat the force as if it acted at the arm's center of mass. The force is straight down, which is perpendicular to the line connecting the shoulder to the center of mass, so

$$F_\perp = |\boldsymbol{F}| = 29\text{ N} \quad .$$

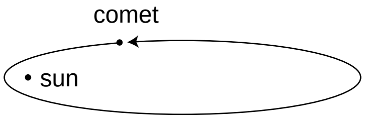
Continuing to pretend that the force acts at the center of the arm, *r* equals 30 cm = 0.30 m, so the torque is

$$\tau = rF_\perp = 9\text{ N}\cdot\text{m} \quad .$$

Discussion Questions

A. This series of discussion questions deals with past students' incorrect reasoning about the following problem.

Suppose a comet is at the point in its orbit shown in the figure. The only force on the comet is the sun's gravitational force.



Throughout the question, define all torques and angular momenta using the sun as the axis.

(1) Is the sun producing a nonzero torque on the comet? Explain.

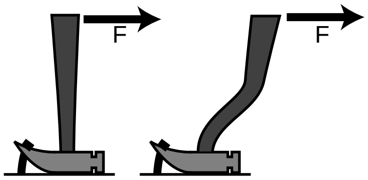
(2) Is the comet's angular momentum increasing, decreasing, or staying the same? Explain.

Explain what is wrong with the following answers. In some cases, the answer is correct, but the reasoning leading up to it is wrong.

(a) Incorrect answer to part (1): "Yes, because the sun is exerting a force on the comet, and the comet is a certain distance from the sun."

(b) Incorrect answer to part (1): "No, because the torques cancel out."

(c) Incorrect answer to part (2): "Increasing, because the comet is speeding up."

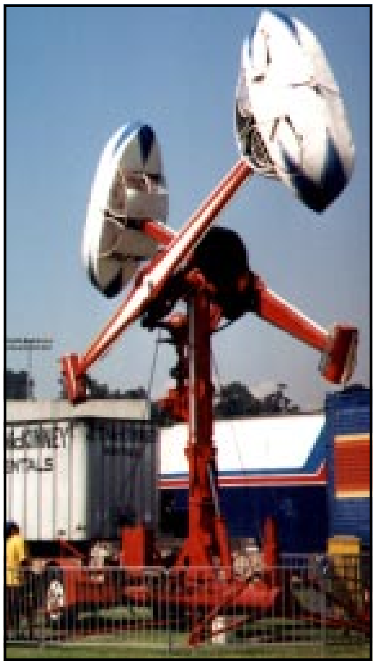


Discussion question B.

B. Which claw hammer would make it easier to get the nail out of the wood if the same force was applied in the same direction?

C. You whirl a rock over your head on the end of a string, and gradually pull in the string, eventually cutting the radius in half. What happens to the rock's angular momentum? What changes occur in its speed, the time required for one revolution, and its acceleration? Why might the string break?

D. A helicopter has, in addition to the huge fan blades on top, a smaller propeller mounted on the tail that rotates in a vertical plane. Why?



Discussion question E.

E. The photo shows an amusement park ride whose two cars rotate in opposite directions. Why is this a good design?

5.5 Statics

Equilibrium

There are many cases where a system is not closed but maintains constant angular momentum. When a merry-go-round is running at constant angular momentum, the engine's torque is being canceled by the torque due to friction.

The windmills are not closed systems, but angular momentum is being transferred out of them at the same rate it is transferred in, resulting in constant angular momentum. To get an idea of the huge scale of the modern windmill farm, note the sizes of the trucks and trailers.

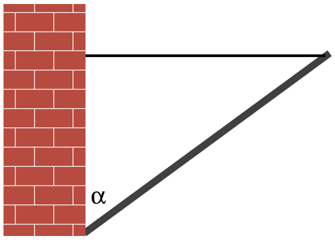
When an object has constant momentum and constant angular momentum, we say that it is in *equilibrium*. This is a scientific redefinition of the common English word, since in ordinary speech nobody would describe a car spinning out on an icy road as being in equilibrium.

Very commonly, however, we are interested in cases where an object is not only in equilibrium but also at rest, and this corresponds more closely to the usual meaning of the word. Trees and bridges have been designed by evolution and engineers to stay at rest, and to do so they must have not just zero total force acting on them but zero total torque. It is not enough that they don't fall down, they also must not tip over. *Statics* is the branch of physics concerned with problems such as these.

Solving statics problems is now simply a matter of applying and combining some things you already know:

- You know the behaviors of the various types of forces, for example that a frictional force is always parallel to the surface of contact.
- You know about vector addition of forces. It is the vector sum of the forces that must equal zero to produce equilibrium.
- You know about torque. The total torque acting on an object must be zero if it is to be in equilibrium.
- You know that the choice of axis is arbitrary, so you can make a choice of axis that makes the problem easy to solve.

In general, this type of problem could involve four equations in four unknowns: three equations that say the force components add up to zero, and one equation that says the total torque is zero. Most cases you'll encounter will not be this complicated. In the example below, only the equation for zero total torque is required in order to get an answer.



Example: a flagpole

Question: A 10-kg flagpole is being held up by a lightweight horizontal cable, and is propped against the foot of a wall as shown in the figure. If the cable is only capable of supporting a tension of 70 N, how great can the angle α be without breaking the cable?

Solution: All three objects in the figure are supposed to be in equilibrium: the pole, the cable, and the wall. Whichever of the three objects we pick to investigate, all the forces and torques on it have to cancel out. It is not particularly helpful to analyze the forces and torques on the wall, since it has forces on it from the ground that are not given and that we don't want to find. We could study the forces and torques on the cable, but that doesn't let us use the given information about the pole. The object we need to analyze is the pole.

The pole has three forces on it, each of which may also result in a torque: (1) the gravitational force, (2) the cable's force, and (3) the wall's force.

We are free to define an axis of rotation at any point we wish, and it is helpful to define it to lie at the bottom end of the pole, since by that definition the wall's force on the pole is applied at r=0 and thus makes no torque on the pole. This is good, because we don't know what the wall's force on the pole is, and we are not trying to find it.

With this choice of axis, there are two nonzero torques on the pole, a counterclockwise torque from the cable and a clockwise torque from gravity. Choosing to represent counterclockwise torques as positive numbers, and using the equation $|\tau| = r|\boldsymbol{F}| \sin\theta$, we have

$$r_{cable}\,|\boldsymbol{F}_{cable}|\sin\theta_{cable} - r_{grav}|\boldsymbol{F}_{grav}|\sin\theta_{grav} = 0 \quad .$$

A little geometry gives θ_{cable}=90°–α and θ_{grav}=α, so

$$r_{cable}\,|\boldsymbol{F}_{cable}|\sin(90°-\alpha) - r_{grav}|\boldsymbol{F}_{grav}|\sin\alpha = 0 \quad .$$

The gravitational force can be considered as acting at the pole's center of mass, i.e. at its geometrical center, so r_{cable} is twice r_{grav}, and we can simplify the equation to read

$$2\,|\boldsymbol{F}_{cable}|\sin(90°-\alpha) - |\boldsymbol{F}_{grav}|\sin\alpha = 0 \quad .$$

These are all quantities we were given, except for α, which is the angle we want to find. To solve for α we need to use the trig identity sin(90°–x) = cos x,

$$2\,|\boldsymbol{F}_{cable}|\cos\alpha - |\boldsymbol{F}_{grav}|\sin\alpha = 0 \quad ,$$

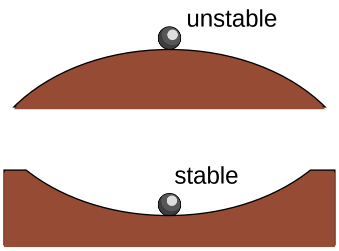
which allows us to find

$$\tan\alpha = 2\frac{|F_{cable}|}{|F_{grav}|} \quad ,$$

$$\alpha = \tan^{-1}\left(2\frac{|F_{cable}|}{|F_{grav}|}\right)$$

$$= \tan^{-1}\left(2\times\frac{70\text{ N}}{98\text{ N}}\right)$$

$$= 55° \quad .$$



Stable and unstable equilibria

A pencil balanced upright on its tip could theoretically be in equilibrium, but even if it was initially perfectly balanced, it would topple in response to the first air current or vibration from a passing truck. The pencil can be put in equilibrium, but not in stable equilibrium. The things around us that we really do see staying still are all in stable equilibrium.

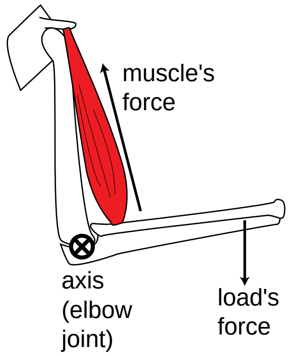
Why is one equilibrium stable and another unstable? Try pushing your own nose to the left or the right. If you push it a millimeter to the left, it responds with a gentle force to the right. If you push it a centimeter to the left, its force on your finger becomes much stronger. The defining characteristic of a stable equilibrium is that the farther the object is moved away from equilibrium, the stronger the force is that tries to bring it back.

The opposite is true for an unstable equilibrium. In the top figure, the ball resting on the round hill theoretically has zero total force on it when it is exactly at the top. But in reality the total force will not be exactly zero, and the ball will begin to move off to one side. Once it has moved, the net force on the ball is greater than it was, and it accelerates more rapidly. In an unstable equilibrium, the farther the object gets from equilibrium, the stronger the force that pushes it farther from equilibrium.

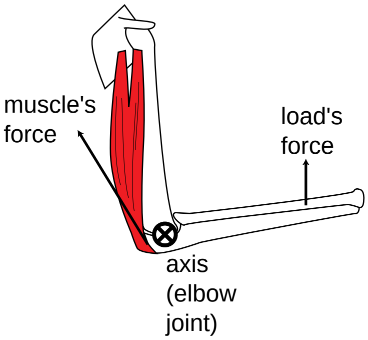
Note that we are using the term “stable” in a weaker sense than in ordinary speech. A domino standing upright is stable in the sense we are using, since it will not spontaneously fall over in response to a sneeze from across the room or the vibration from a passing truck. We would only call it unstable in the technical sense if it could be toppled by *any* force, no matter how small. In everyday usage, of course, it would be considered unstable, since the force required to topple it is so small.

Pooh’s equilibrium is unstable.
(c) 1926 E.H. Shepard

5.6 Simple Machines: The Lever



(a) The biceps muscle flexes the arm.



(b) The triceps muscle extends the arm.

Although we have discussed some simple machines such as the pulley, without the concept of torque we were not yet ready to address the lever, which is the machine nature used in designing living things, almost to the exclusion of all others. (We can speculate what life on our planet might have been like if living things had evolved wheels, gears, pulleys, and screws.) The figures show two examples of levers within your arm. Different muscles are used to flex and extend the arm, because muscles work only by contraction.

Analyzing example (a) physically, there are two forces that do work. When we lift a load with our biceps muscle, the muscle does positive work, because it brings the bone in the forearm in the direction it is moving. The load's force on the arm does negative work, because the arm moves in the direction opposite to the load's force. This makes sense, because we expect our arm to do positive work on the load, so the load must do an equal amount of negative work on the arm. (If the biceps was lowering a load, the signs of the works would be reversed. Any muscle is capable of doing either positive or negative work.)

There is also a third force on the forearm: the force of the upper arm's bone exerted on the forearm at the elbow joint (not shown with an arrow in the figure). This force does no work, because the elbow joint is not moving.

Because the elbow joint is motionless, it is natural to define our torques using the joint as the axis. The situation now becomes quite simple, because the upper arm bone's force exerted at the elbow neither does work nor creates a torque. We can ignore it completely. In any lever there is such a point, called the fulcrum.

If we restrict ourselves to the case in which the forearm rotates with constant angular momentum, then we know that the total torque on the forearm is zero,

$$\tau_{muscle} + \tau_{load} = 0 \quad .$$

If we choose to represent counterclockwise torques as positive, then the muscle's torque is positive, and the load's is negative. In terms of their absolute values,

$$|\tau_{muscle}| = |\tau_{load}| \quad .$$

Assuming for simplicity that both forces act at angles of 90° with respect to the lines connecting the axis to the points at which they act, the absolute values of the torques are

$$r_{muscle}F_{muscle} = r_{load}F_{arm} \quad ,$$

where r_{muscle}, the distance from the elbow joint to the biceps' point of insertion on the forearm, is only a few cm, while r_{load} might be 30 cm or so. The force exerted by the muscle must therefore be about ten times the force exerted by the load. We thus see that this lever is a force reducer. In general, a lever may be used either to increase or to reduce a force.

Why did our arms evolve so as to reduce force? In general, your body is built for compactness and maximum speed of motion rather than maximum force. This is the main anatomical difference between us and the

Neanderthals (their brains covered the same range of sizes as those of modern humans), and it seems to have worked for us.

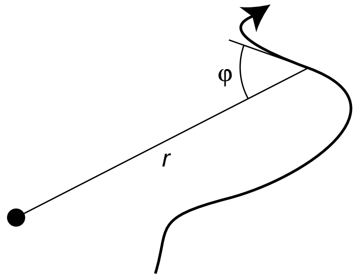
As with all machines, the lever is incapable of changing the amount of mechanical work we can do. A lever that increases force will always reduce motion, and vice versa, leaving the amount of work unchanged.

It is worth noting how simple and yet how powerful this analysis was. It was simple because we were well prepared with the concepts of torque and mechanical work. In anatomy textbooks, whose readers are assumed not to know physics, there is usually a long and complicated discussion of the different types of levers. For example, the biceps lever, (a), would be classified as a class III lever, since it has the fulcrum and load on the ends and the muscle's force acting in the middle. The triceps, (b), is called a class I lever, because the load and muscle's force are on the ends and the fulcrum is in the middle. How tiresome! With a firm grasp of the concept of torque, we realize that all such examples can be analyzed in much the same way. Physics is at its best when it lets us understand many apparently complicated phenomena in terms of a few simple yet powerful concepts.

5.7* Proof of Kepler's Elliptical Orbit Law

Kepler determined purely empirically that the planets' orbits were ellipses, without understanding the underlying reason in terms of physical law. Newton's proof of this fact based on his laws of motion and law of gravity was considered his crowning achievement both by him and by his contemporaries, because it showed that the same physical laws could be used to analyze both the heavens and the earth. Newton's proof was very lengthy, but by applying the more recent concepts of conservation of energy and angular momentum we can carry out the proof quite simply and succinctly, and without calculus.

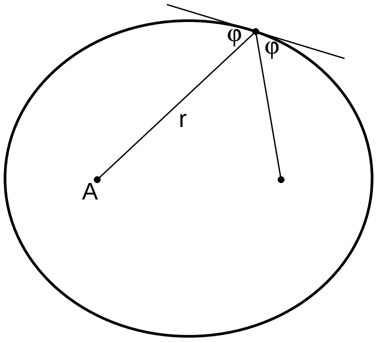
The basic idea of the proof is that we want to describe the shape of the planet's orbit with an equation, and then show that this equation is exactly the one that represents an ellipse. Newton's original proof had to be very complicated because it was based directly on his laws of motion, which include time as a variable. To make any statement about the shape of the orbit, he had to eliminate time from his equations, leaving only space variables. But conservation laws tell us that certain things don't change over time, so they have already had time eliminated from them.



There are many ways of representing a curve by an equation, of which the most familiar is $y=ax+b$ for a line in two dimensions. It would be perfectly possible to describe a planet's orbit using an x-y equation like this, but remember that we are applying conservation of angular momentum, and the space variables that occur in the equation for angular momentum are the distance from the axis, r, and the angle between the velocity vector and the $\boldsymbol{r}$ vector, which we will call φ. The planet will have $\varphi=90°$ when it is moving perpendicular to the $\boldsymbol{r}$ vector, i.e. at the moments when it is at its smallest or greatest distances from the sun. When φ is less than 90° the planet is approaching the sun, and when it is greater than 90° it is receding from it. Describing a curve with an r-φ equation is like telling a driver in a parking lot a certain rule for what direction to steer based on the distance from a certain streetlight in the middle of the lot.

The proof is broken into the three parts for easier digestion. The first part is a simple and intuitively reasonable geometrical fact about ellipses, whose proof we relegate to the caption of a figure; you will not be missing much if you merely absorb the result without reading the proof.

(1) If we use one of the two foci of an ellipse as an axis for defining the variables r and φ, then the angle between the tangent line and the line drawn to the other focus is the same as φ, i.e. the two angles labeled φ in the figure are in fact equal.



Proof that the two angles labeled φ are in fact equal: The definition of an ellipse is that the sum of the distances from the two foci stays constant. If we move a small distance ℓ along the ellipse, then one distance shrinks by an amount $\ell\cos\varphi_1$, while the other grows by $\ell\cos\varphi_2$. These are equal, so $\varphi_1=\varphi_2$.

The other two parts form the meat of our proof. We state the results first and then prove them.

(2) A planet, moving under the influence of the sun's gravity with less then the energy required to escape, obeys an equation of the form

$$\sin \varphi = \frac{1}{\sqrt{-pr^2 + qr}} \quad ,$$

where p and q are constants that depend on the planet's energy and angular momentum and p is greater than zero.

(3) A curve is an ellipse if and only if its r-φ equation is of the form

$$\sin \varphi = \frac{1}{\sqrt{-pr^2 + qr}} \quad ,$$

where p and q are constants that depend on the size and shape of the ellipse and p is greater than zero.

Proof of part (2)

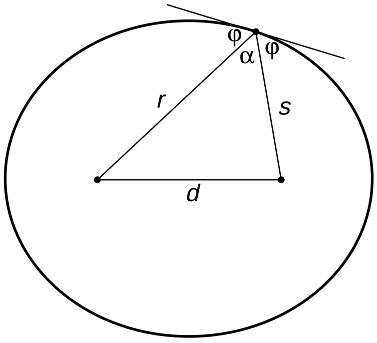
The component of the planet's velocity vector that is perpendicular to the $\boldsymbol{r}$ vector is $v_\perp = v \sin \varphi$, so conservation of angular momentum tells us that $L = mrv \sin \varphi$ is a constant. Since the planet's mass is a constant, this is the same as the condition

$$rv \sin \varphi = \text{constant} \quad .$$

Conservation of energy gives

$$\tfrac{1}{2}mv^2 - G\frac{Mm}{r} = \text{constant} \quad .$$

We solve the first equation for v and plug into the second equation to eliminate v. Straightforward algebra then leads to the equation claimed above, with the constant p being positive because of our assumption that the planet's energy is insufficient to escape from the sun, i.e. its total energy is negative.



Proof of part (3)

We define the quantities α, d, and s as shown in the figure. The law of cosines gives

$$d^2 = r^2 + s^2 - 2rs \cos \alpha \quad .$$

Using α=180°–2φ and the trigonometric identities cos (180°–x)=–cos x and cos $2x$ = 1–2 $\sin^2 x$, we can rewrite this as

$$d^2 = r^2 + s^2 + 2rs\left(1 - \sin^2\varphi\right) \quad .$$

Straightforward algebra transforms this into

$$\sin \varphi = \sqrt{\frac{d^2 + \left(r+s\right)^2}{2rs}} \quad .$$

Since r+s is constant, the top of the fraction is constant, and the denominator can be rewritten as $2rs$=$2r$(constant-r), which is equivalent to the desired form.

Summary

Selected Vocabulary

angular momentum .... a measure of rotational motion; a conserved quantity for a closed system

axis .... An arbitrarily chosen point used in the definition of angular momentum. Any object whose direction changes relative to the axis is considered to have angular momentum. No matter what axis is chosen, the angular momentum of a closed system is conserved.

torque .... the rate of change of angular momentum; a numerical measure of a force's ability to twist on an object

equilibrium .... a state in which an object's momentum and angular momentum are constant

stable equilibrium .... one in which a force always acts to bring the object back to a certain point

unstable equilibrium .... one in which any deviation of the object from its equilibrium position results in a force pushing it even farther away

Notation

$\boldsymbol{L}$ .... angular momentum

τ .... torque

T .... the time required for a rigidly rotating body to complete one rotation

Standard Terminology and Notation Not Used in This Book

period .... a name for the variable T defined above

moment of inertia, I .... the proportionality constant in the equation $L = 2\pi I / T$

Summary

Angular momentum is a measure of rotational motion which is conserved for a closed system. This book only discusses angular momentum for rotation of material objects in two dimensions. Not all rotation is rigid like that of a wheel or a spinning top. An example of nonrigid rotation is a cyclone, in which the inner parts take less time to complete a revolution than the outer parts. In order to define a measure of rotational motion general enough to include nonrigid rotation, we define the angular momentum of a system by dividing it up into small parts, and adding up all the angular momenta of the small parts, which we think of as tiny particles. We arbitrarily choose some point in space, the *axis*, and we say that anything that changes its direction relative to that point possesses angular momentum. The angular momentum of a single particle is

$$L = mv_\perp r \quad ,$$

where $v_\perp$ is the component of its velocity perpendicular to the line joining it to the axis, and r is its distance from the axis. Positive and negative signs of angular momentum are used to indicate clockwise and counterclockwise rotation.

The *choice of axis theorem* states that any axis may be used for defining angular momentum. If a system's angular momentum is constant for one choice of axis, then it is also constant for any other choice of axis.

The *spin theorem* states that an object's angular momentum with respect to some outside axis A can be found by adding up two parts:

(1) The first part is the object's angular momentum found by using its own center of mass as the axis, i.e. the angular momentum the object has because it is spinning.

(2) The other part equals the angular momentum that the object would have with respect to the axis A if it had all its mass concentrated at and moving with its center of mass.

Torque is the rate of change of angular momentum. The torque a force can produce is a measure of its ability to twist on an object. The relationship between force and torque is

$$|\tau| \quad = \quad r|F_\perp| \quad ,$$

where r is the distance from the axis to the point where the force is applied, and $F_\perp$ is the component of the force perpendicular to the line connecting the axis to the point of application. Statics problems can be solved by setting the total force and total torque on an object equal to zero and solving for the unknowns.

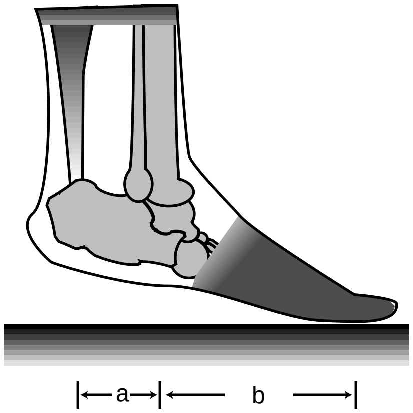
Homework Problems

1✓. You are trying to loosen a stuck bolt on your RV using a big wrench that is 50 cm long. If you hang from the wrench, and your mass is 55 kg, what is the maximum torque you can exert on the bolt?

2✓. A physical therapist wants her patient to rehabilitate his injured elbow by laying his arm flat on a table, and then lifting a 2.1 kg mass by bending his elbow. In this situation, the weight is 33 cm from his elbow. He calls her back, complaining that it hurts him to grasp the weight. He asks if he can strap a bigger weight onto his arm, only 17 cm from his elbow. How much mass should she tell him to use so that he will be exerting the same torque? (He is raising his forearm itself, as well as the weight.)

3. An object thrown straight up in the air is momentarily at rest when it reaches the top of its motion. Does that mean that it is in equilibrium at that point? Explain.

4. An object is observed to have constant angular momentum. Can you conclude that no torques are acting on it? Explain. [Based on a problem by Serway and Faughn.]



Problem 5.

5. A person of weight *W* stands on the ball of one foot. Find the tension in the calf muscle and the force exerted by the shinbones on the bones of the foot, in terms of *W*, *a*, and *b*. For simplicity, assume that all the forces are at 90-degree angles to the foot, i.e. neglect the angle between the foot and the floor.

6. Two objects have the same momentum vector. Can you conclude that their angular momenta are the same? Explain. [Based on a problem by Serway and Faughn.]

In about five billion years, our own sun will become a white dwarf like the small dot at the center of this cloud of cast-off gas.

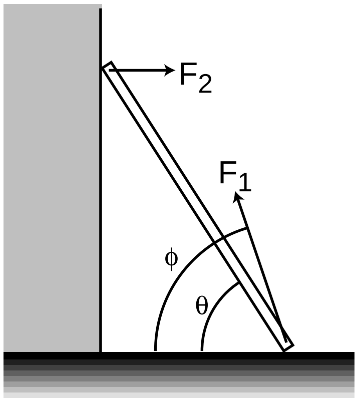
7. The sun turns on its axis once every 26.0 days. Its mass is 2.0×10^{30} kg and its radius is 7.0×10^{8} m. Assume it is a rigid sphere of uniform density.

(a✓) What is the sun's angular momentum?

In a few billion years, astrophysicists predict that the sun will use up all its sources of nuclear energy, and will collapse into a ball of exotic, dense matter known as a white dwarf. Assume that its radius becomes 5.8×10^{6} m (similar to the size of the Earth.) Assume it does not lose any mass between now and then. (Don't be fooled by the photo, which makes it look like nearly all of the star was thrown off by the explosion. The visually prominent gas cloud is actually thinner than the best laboratory vacuum every produced on earth. Certainly a little bit of mass is actually lost, but it is not at all unreasonable to make an approximation of zero loss of mass as we are doing.)

(b) What will its angular momentum be?
(c✓) How long will it take to turn once on its axis?

S	A solution is given in the back of the book.	★	A difficult problem.
✓	A computerized answer check is available.	∫	A problem that requires calculus.



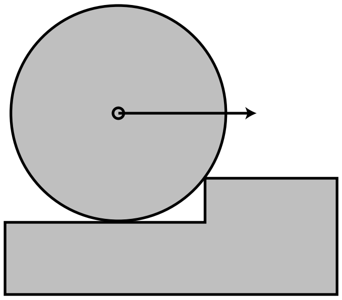
Problems 8 and 9.

8. A uniform ladder of mass m and length L leans against a smooth wall, making an angle θ with respect to the ground. The dirt exerts a normal force and a frictional force on the ladder, producing a force vector with magnitude F_1 at an angle ϕ with respect to the ground. Since the wall is smooth, it exerts only a normal force on the ladder; let its magnitude be F_2.

(a) Explain why ϕ must be greater than θ. No math is needed.

(b) Choose any numerical values you like for m and L, and show that the ladder can be in equilibrium (zero torque and zero total force vector) for θ=45.00° and ϕ=63.43°.

9 ★. Continuing the previous problem, find an equation for ϕ in terms of θ, and show that m and L do not enter into the equation. Do not assume any numerical values for any of the variables. You will need the trig identity $\sin(a\text{-}b) = \sin a \cos b - \sin b \cos a$. (As a numerical check on your result, you may wish to check that the angles given in part b of the previous problem satisfy your equation.)



Problem 10.

10. (a) Find the minimum horizontal force which, applied at the axle, will pull a wheel over a step. Invent algebra symbols for whatever quantities you find to be relevant, and give your answer in symbolic form. [Hints: There are three forces on the wheel at first, but only two when it lifts off. Normal forces are always perpendicular to the surface of contact. Note that the corner of the step cannot be perfectly sharp, so the surface of contact for this force really coincides with the surface of the wheel.]

(b) Under what circumstances does your result become infinite? Give a physical interpretation.

11 ★. A yo-yo of total mass m consists of two solid cylinders of radius R, connected by a small spindle of negligible mass and radius r. The top of the string is held motionless while the string unrolls from the spindle. Show that the acceleration of the yo-yo is $g/(1+R^2/2r^2)$. [Hint: The acceleration and the tension in the string are unknown. Use $\tau=\Delta L/\Delta t$ and $F=ma$ to determine these two unknowns.]

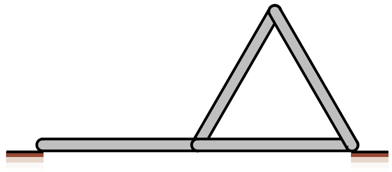
12. A ball is connected by a string to a vertical post. The ball is set in horizontal motion so that it starts winding the string around the post. Assume that the motion is confined to a horizontal plane, i.e. ignore gravity. Michelle and Astrid are trying to predict the final velocity of the ball when it reaches the post. Michelle says that according to conservation of angular momentum, the ball has to speed up as it approaches the post. Astrid says that according to conservation of energy, the ball has to keep a constant speed. Who is right? [Hint: How is this different from the case where you whirl a rock in a circle on a string and gradually pull in the string?]

13. In the 1950's, serious articles began appearing in magazines like *Life* predicting that world domination would be achieved by the nation that could put nuclear bombs in orbiting space stations, from which they could be dropped at will. In fact it can be quite difficult to get an orbiting object to come down. Let the object have energy $E=KE+PE$ and angular momentum L. Assume that the energy is negative, i.e. the object is moving at less than escape velocity. Show that it can never reach a radius less than

$$r_{\text{min}} = \frac{GMm}{2E}\left(-1+\sqrt{1+\frac{2EL^2}{G^2M^2m^3}}\right) .$$

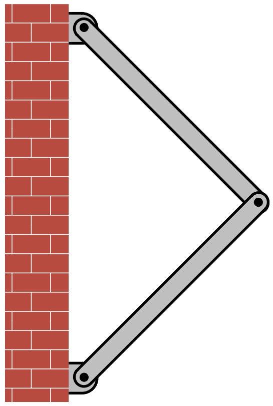
[Note that both factors are negative, giving a positive result.]

14. The figure shows a bridge made out of four identical trusses, each of weight W. How much force must be supplied by each pier to hold up the bridge?



Problem 14.

15★. Two bars slanted at 45 degrees are attached to each other and to a wall as shown in the figure. Each bar has weight W. The goal is to show that the middle joint must be able to handle a strain $W/2$, and the top and bottom joints $\sqrt{5}W/4$. [The problem could be set up with three equations of equilibrium for the top bar and three for the bottom bar, giving a total of six equations in six unknowns (two unknowns for the components of each of the three forces). A less tedious approach is as follows. First prove that the horizontal forces at all three joints have the same magnitude, X, and figure out which are to the right and which are to the left. Next, choose the top joint as an axis, and use the fact that the total torque equals zero to prove $X=W/2$. Finally, find the vertical forces.]

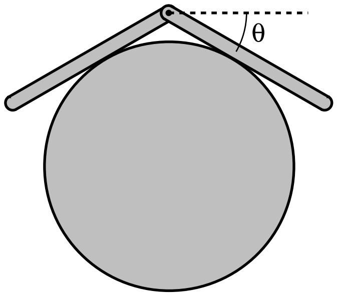


Problem 15.

16★. Two bars of length L are connected with a hinge and placed on a frictionless cylinder of radius r. (a) Show that the angle θ shown in the figure is related to the unitless ratio r/L by the equation

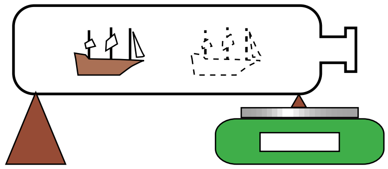
$$\frac{r}{L} = \frac{\cos^2\theta}{2\tan\theta} .$$

(b) Discuss the physical behavior of this equation for very large and very small values of r/L.



Problem 16.

17. You wish to determine the mass of a ship in a bottle without taking it out. Show that this can be done with the setup shown in the figure, with a scale supporting the bottle at one end, provided that it is possible to take readings with the ship slid to two different locations.



Problem 17.

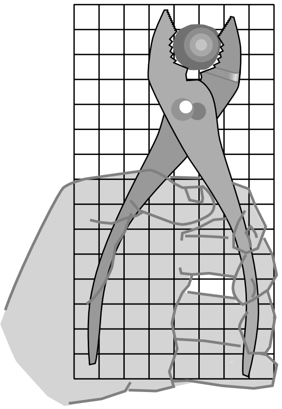
18 ∫. Two atoms will interact via electrical forces between their protons and electrons. One fairly good approximation to the potential energy is the Lennard-Jones potential,

$$PE(r) = k\left[\left(\frac{a}{r}\right)^{12} - 2\left(\frac{a}{r}\right)^{6}\right] \quad ,$$

where r is the center-to-center distance between the atoms.

Show that (a) there is an equilibrium point at $a=r$, (b) the equilibrium is stable, and (c) the energy required to bring the atoms from their equilibrium separation to infinity is k. Hint: The first two parts can be done more easily by setting $a=1$, since the value of a only changes the distance scale.

19. Suppose that we lived in a universe in which Newton's law of gravity gave forces proportional to r^{-7} rather than r^{-2}. Which, if any, of Kepler's laws would still be true? Which would be completely false? Which would be different, but in a way that could be calculated with straightforward algebra?



Problem 18.

18 S. The figure shows scale drawing of a pair of pliers being used to crack a nut, with an appropriately reduced centimeter grid. Warning: do not attempt this at home; it is bad manners. If the force required to crack the nut is 300 N, estimate the force required of the person's hand.

19. Show that a sphere of radius R that is rolling without slipping has angular momentum and momentum in the ratio $L/p=(2/5)R$.

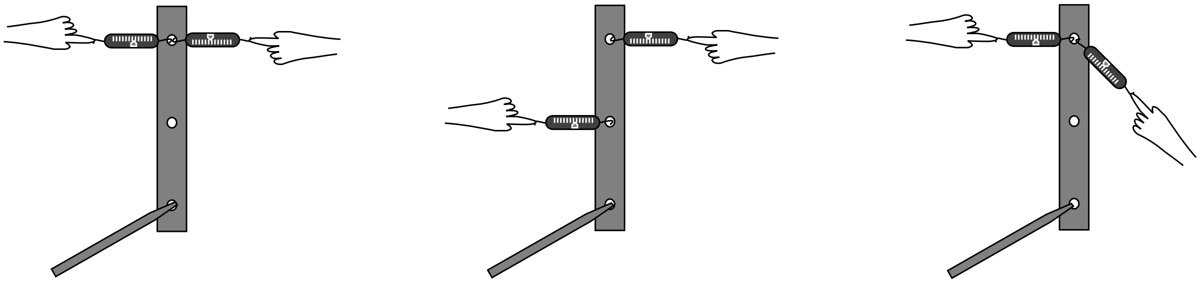
20. Suppose a bowling ball is initially thrown so that it has no angular momentum at all, i.e. it is initially just sliding down the lane. Eventually kinetic friction will bring its angular velocity up to the point where it is rolling without slipping. Show that the final velocity of the ball equals 5/7 of its initial velocity. [Hint: You'll need the result of problem 19.]

Exercises

Exercise 5A: Torque

Equipment:

- rulers with holes in them
- spring scales (two per group)



While one person holds the pencil which forms an axle for the ruler, the other members of the group pull on the scales and take readings. In each case, determine whether the total torque on the ruler appears to equal zero to roughly within the aqccuracy of the measurement.

Solutions to Selected Problems

Chapter 1

7. A force is an interaction between two objects, so while the bullet is in the air, there is no force. There is only a force while the bullet is in contact with the book. There is energy the whole time, and the total amount doesn't change. The bullet has some kinetic energy, and transfers some of it to the book as heat, sound, and the energy required to tear a hole through the book.

8. (a) The energy stored in the gasoline is being changed into heat via frictional heating, and also probably into sound and into energy of water waves. Note that the kinetic energy of the propeller and the boat are changing, so they are not involved in the energy transformation. (b) The crusing speed would be greater by a factor of the cube root of 2, or about a 26% increase.

9. We don't have actual masses and velocities to plug in to the equation, but that's OK. We just have to reason in terms of ratios and proportionalities. Kinetic energy is proportional to mass and to the square of velocity, so B's kinetic energy equals

$$(13.4\ \text{J})(3.77)/(2.34)^2 = 9.23\ \text{J}$$

11. Room temperature is about 20°C. The fraction of the power that actually goes into heating the water is

$$\frac{(250\ \text{g}) / (0.24\ \text{J/g¡C}) \times (100\text{¡C} - 20\text{¡C}) / 126\ \text{s}}{1.25\times10^3\ \text{J/s}} = 0.53$$

So roughly half of the energy is wasted. The wasted energy might be in several forms: heating of the cup, heating of the oven itself, or leakage of microwaves from the oven.

Chapter 2

5.

$$\begin{aligned}
E_{\text{total,i}} &= E_{\text{total,f}} \\
PE_{\text{i}} + heat_{\text{i}} &= PE_{\text{f}} + KE_{\text{f}} + heat_{\text{f}} \\
\tfrac{1}{2}mv^2 &= PE_{\text{i}} - PE_{\text{f}} + heat_{\text{i}} - heat_{\text{f}} \\
&= -\Delta PE - \Delta heat \\
v &= \sqrt{2\left(\frac{-\Delta PE - \Delta heat}{m}\right)} \\
&= 6.4\ \text{m/s}
\end{aligned}$$

7. Let θ be the angle she by which she has progressed around the pipe. Conservation of energy gives

$$\begin{aligned}
E_{\text{total,i}} &= E_{\text{total,f}} \\
PE_{\text{i}} &= PE_{\text{f}} + KE_{\text{f}}
\end{aligned}$$

Let's make PE=0 at the top, so

$$0 = mgr(\cos\theta - 1) + \tfrac{1}{2}mv^2 \quad .$$

While she is still in contact with the pipe, the radial component of her acceleration is

$$a_{\text{r}} = v^2/r \quad ,$$

and making use of the previous equation we find

$$a_{\text{r}} = 2g(1-\cos\theta) \quad .$$

There are two forces on her, a normal force from the pipe and a downward gravitation force from the earth. At the moment when she loses contact with the pipe, the normal force is zero, so the radial component, $mg \cos\theta$, of the gravitational force must equal ma_{r},

$$mg\cos\theta = 2mg(1-\cos\theta) \quad ,$$

which gives

$$\cos\theta = 2/3 \quad .$$

The amount by which she has dropped is $r(1-\cos\theta)$, which equals $r/3$ at this moment.

9. (a) Example: As one child goes up on one side of a see-saw, another child on the other side comes down. (b) Example: A pool ball hits another pool ball, and transfers some KE.

11. Suppose the river is 1 m deep, 100 m wide, and flows at a speed of 10 m/s, and that the falls are 100 m tall. In 1 second, the volume of water flowing over the falls is 10^3 m^3, with a mass of 10^6 kg. The potential energy released in one second is (10^6 kg)(g)(100 m)=10^9 J, so the power is 10^9 W. A typical household might have 10 hundred-watt applicances turned on at any given time, so it consumes about 10^3 watts on the average. The plant could supply a about million households with electricity.

Chapter 3

18. No. Work describes how energy was transferred by some process. It isn't a measurable property of a system.

Chapter 4

8. Let m be the mass of the little puck and M=2.3m be the mass of the little one. All we need to do is find the direction of the total momentum vector before the collision, because the total momentum vector is the same after the collision. Given the two components of the momentum vector $p_x=mv$ and $p_y=Mv$, the direction of the vector is $\tan^{-1}(p_y/p_x)$=23° counterclockwise from the big puck's original direction of motion.

11. Momentum is a vector. The total momentum of the molecules is always zero, since the momenta in different directions cancal out on the average. Cooling changes individual molecular momenta, but not the total.

15) (a) Particle i had velocity $\boldsymbol{v}_i$ in the center-of-mass frame, and has velocity $\boldsymbol{v}_i+\boldsymbol{u}$ in the new frame. The total kinetic energy is

$$\frac{1}{2}m_1\left(\boldsymbol{v}_1+\boldsymbol{u}\right)^2+\ldots$$

where "..." indicates that the sum continues for all the particles. Rewriting this in terms of the vector dot product, we have

$$\frac{1}{2}m_1\left(\boldsymbol{v}_1+\boldsymbol{u}\right)\cdot\left(\boldsymbol{v}_1+\boldsymbol{u}\right)+\ldots$$

$$=\frac{1}{2}m_1\left(\boldsymbol{v}_1\cdot\boldsymbol{v}_1+2\boldsymbol{u}\cdot\boldsymbol{v}_1+\boldsymbol{u}\cdot\boldsymbol{u}\right)+\ldots$$

When we add up all the terms like the first one, we get K_{cm}. Adding up all the terms like the third one, we get $M|\boldsymbol{u}|^2/2$. The terms like the second term cancel out:

$$m_1\boldsymbol{u}\cdot\boldsymbol{v}_1+\ldots$$

$$=\boldsymbol{u}\cdot\left(m_1\boldsymbol{v}_1+\ldots\right)\quad,$$

where the sum in brackets equals the total momentum in the center-of-mass frame, which is zero by definition. (b) Changing frames of reference doesn't change the distances between the particles, so the potential energies are all unaffected by the change of frames of reference. Suppose that in a given frame of reference,frame 1, energy is conserved in some process: the initial and final energies add up to be the same. First let's transform to the center-of-mass frame. The potential energies are unaffected by the transformation, and the total kinetic energy is simply reduced by the quantity $M|\boldsymbol{u}_1|^2/2$, where $\boldsymbol{u}_1$ is the velocity of frame 1 relative to the center of mass. Subtracting the same constant from the initial and final energies still leaves them equal. Now we transform to frame 2. Again, the effect is simply to change the initial and final energies by adding the same constant.

Chapter 5

18. The pliers are not moving, so their angular momentum remains constant at zero, and the total torque on them must be zero. Not only that, but each half of the pliers must have zero total torque on it. This tells us that the magnitude of the torque at one end must be the same as that at the other end. The distance from the axis to the nut is about 2.5 cm, and the distance from the axis to the centers of the palm and fingers are about 8 cm. The angles are close enough to 90° that we can pretend they're 90 degrees, considering the rough nature of the other assumptions and measurements. The result is (300 N)(2.5 cm)=(F)(8 cm), or F=90 N.

Glossary

Angular momentum. A measure of rotational motion; a conserved quantity for a closed system.

Axis. An arbitrarily chosen point used in the definition of angular momentum. Any object whose direction changes relative to the axis is considered to have angular momentum. No matter what axis is chosen, the angular momentum of a closed system is conserved.

Center of mass. The balance point or average position of the mass in a system.

Collision. An interaction between moving objects that lasts for a certain time.

Energy. A numerical scale used to measure the heat, motion, or other properties th.at would require fuel or physical effort to put into an object; a scalar quantity with units of joules (J).

Equilibrium. A state in which an object's momentum and angular momentum are constant.

Heat. The energy that an object has because of its temperature. Heat is different from temperature (q.v.) because an object with twice as much mass requires twice as much heat to increase its temperature by the same amount. There is a further distinction in the terminology, not emphasized in this book, between heat and thermal energy. See the entry under thermal energy for a discussion of this distinction.

Kinetic energy. The energy an object posesses because of its motion. Cf. potential energy.

Momentum. A measure of motion, equal to *mv* for material objects.

Potential energy. The energy having to do with the distance between to objects that interact via a noncontact force. Cf. Kinetic energy.

Power. The rate of transferring energy; a scalar quantity with units of watts (W).

Stable equilibrium. One in which a force always acts to bring the object back to a certain point.

Temperature. What a thermometer measures. Objects left in contact with each other tend to reach the same temperature. Roughly speaking, temperature measures the average kinetic energy per molecule. For the distinction between temperature and heat, see the glossary entry for heat.

Thermal energy.Careful writers make a distinction between heat and thermal energy, but the distinction is often ignored in casual speech, even among physicists. Properly, thermal energy is used to mean the total amount of energy posessed by an object, while heat indicates the amount of thermal energy transferred in or out. The term heat is used in this book to include both meanings.

Torque. The rate of change of angular momentum; a numerical measure of a force's ability to twist on an object.

Unstable equilibrium. One in which any deviation of the object from its equilibrium position results in a force pushing it even farther away.

Work. The amount of energy transferred into or out of a system, excluding energy transferred by heat conduction.

Index

Photo Credits

All photographs are by Benjamin Crowell, except as noted below.

Cover
Sun: Image from the SOHO space probe, NASA/European Space Agency.
Hurricane Andrew, 1992: National Oceanic and Atmospheric Administration.

Chapter 1
Jupiter: Images from the Hubble Space Telescope, NASA.
Skateboarder in pool: J.D. Rogge, from www.sonic.net/~shawn.
Hoover Dam: U.S. Department of the Interior, Bureau of Reclamation, Lower Colorado Region.

Chapter 3
Breaking Trail: Art by Walter E. Bohl. Image courtesy of the University of Michigan Museum of Art/School of Information and Library Studies.

Chapter 4
Nucleus of Halley's comet: NASA
Halley's comet with tail: W. Liller.
Chadwick's apparatus: The original drawing from Chadwick's paper.
Wrench: Uncopyrighted photograph from PSSC Physics.

Chapter 5
Tornado: Rev. Robert Alexander.
Longjump: Thomas Eakins.
Diver: Uncopyrighted photograph from PSSC Physics.
Windmill farm: Tehachapi, CA. U.S. Department of Energy.
White dwarf: Image of NGC 2440 from the Hubble Space Telescope, H. Bond and R. Ciardullo.

Useful Data

Metric Prefixes

M-	mega-	10^{6}
k-	kilo-	10^{3}
m-	milli-	10^{-3}
μ- (Greek mu)	micro-	10^{-6}
n-	nano-	10^{-9}

(Centi-, 10^{-2}, is used only in the centimeter.)

Notation and Units

quantity	unit	symbol
distance	meter, m	x, Δx
time	second, s	t, Δt
mass	kilogram, kg	m
area	m^2 (square meters)	A
volume	m^3 (cubic meters)	V
density	kg/m^3	ρ
force	newton, 1 N=1 $kg\cdot m/s^2$	$\boldsymbol{F}$
velocity	m/s	$\boldsymbol{v}$
acceleration	m/s^2	$\boldsymbol{a}$
energy	joule, J	E
momentum	$kg\cdot m/s$	$\boldsymbol{p}$
angular momentum	$kg\cdot m^2/s$	L

symbol	meaning
$\propto$	is proportional to
$\approx$	is approximately equal to
$\sim$	on the order of

The Greek Alphabet

α	Α	alpha	ν	Ν	nu
β	Β	beta	ξ	Ξ	xi
γ	Γ	gamma	ο	Ο	omicron
δ	Δ	delta	π	Π	pi
ε	Ε	epsilon	ρ	Ρ	rho
ζ	Ζ	zeta	σ	Σ	sigma
η	Η	eta	τ	Τ	tau
θ	Θ	theta	υ	Υ	upsilon
ι	Ι	iota	φ	Φ	phi
κ	Κ	kappa	χ	Χ	chi
λ	Λ	lambda	ψ	Ψ	psi
μ	Μ	mu	ω	Ω	omega

Conversions

Conversions between SI and other units:

1 inch	=	2.54 cm (exactly)
1 mile	=	1.61 km
1 pound	=	4.45 N
$(1\ \mathrm{kg})\cdot g$	=	2.2 lb
1 gallon	=	$3.78\text{x}10^{3}\ \mathrm{cm}^{3}$
1 horsepower	=	746 W
1 kcal*	=	$4.18\text{x}10^{3}$ J

*When speaking of food energy, the word “Calorie” is used to mean 1 kcal, i.e. 1000 calories. In writing, the capital C may be used to indicate 1 Calorie=1000 calories.

Conversions between U.S. units:

1 foot	=	12 inches
1 yard	=	3 feet
1 mile	=	5280 ft

Earth, Moon, and Sun

body	mass (kg)	radius (km)	radius of orbit (km)
earth	$5.97\text{x}10^{24}$	$6.4\text{x}10^{3}$	$1.49\text{x}10^{8}$
moon	$7.35\text{x}10^{22}$	$1.7\text{x}10^{3}$	$3.84\text{x}10^{5}$
sun	$1.99\text{x}10^{30}$	$7.0\text{x}10^{5}$	

The radii and radii of orbits are average values. The moon orbits the earth and the earth orbits the sun.

Subatomic Particles

particle	mass (kg)	radius (m)
electron	$9.109\text{x}10^{-31}$	? – less than about 10^{-17}
proton	$1.673\text{x}10^{-27}$	about $1.1\text{x}10^{-15}$
neutron	$1.675\text{x}10^{-27}$	about $1.1\text{x}10^{-15}$

The radii of protons and neutrons can only be given approximately, since they have fuzzy surfaces. For comparison, a typical atom is about 10^{-9} m in radius.

Fundamental Constants

speed of light	$c=3.00\text{x}10^{8}$ m/s
gravitational constant	$G=6.67\text{x}10^{-11}\ \mathrm{N\cdot m^{2}\cdot kg^{-2}}$

Vibrations and Waves

Benjamin Crowell

Book 3 in the **Light and Matter** series of introductory physics textbooks
www.lightandmatter.com

Vibrations and Waves

The **Light and Matter** series of introductory physics textbooks:

1 Newtonian Physics

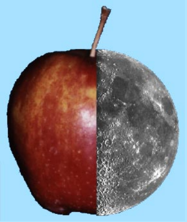
2 Conservation Laws

3 Vibrations and Waves

4 Electricity and Magnetism

5 Optics

6 The Modern Revolution in Physics



Vibrations and Waves

Benjamin Crowell

www.lightandmatter.com

Light and Matter

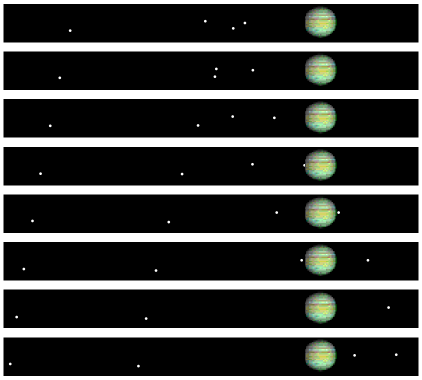
Fullerton, California
www.lightandmatter.com

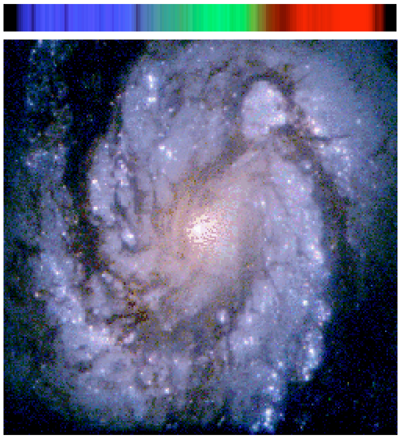
Edition 2.0
rev. 2000-09-16

ISBN 0-9704670-3-6

To Diz and Bird.

Contents





Supplement 3-5: Complex-Number Techniques

Supplement 3-5 is available online in electronic format at www.lightandmatter.com.

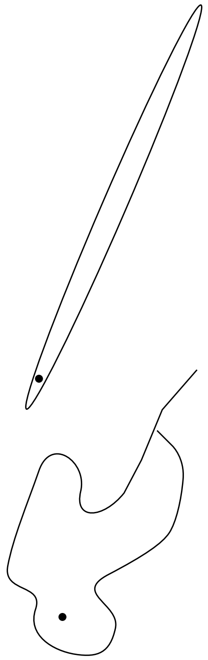
The vibrations of this electric bass string are converted to electrical vibrations, then to sound vibrations, and finally to vibrations of our eardrums.

1 Vibrations

Dandelion. Cello. Read those two words, and your brain instantly conjures a stream of associations, the most prominent of which have to do with vibrations. Our mental category of “dandelion-ness” is strongly linked to the color of light waves that vibrate about half a million billion times a second: yellow. The velvety throb of a cello has as its most obvious characteristic a relatively low musical pitch — the note you are spontaneously imagining right now might be one whose sound vibrations repeat at a rate of a hundred times a second.

Evolution has designed our two most important senses around the assumption that not only will our environment be drenched with information-bearing vibrations, but in addition those vibrations will often be repetitive, so that we can judge colors and pitches by the rate of repetition. Granting that we do sometimes encounter nonrepeating waves such as the consonant “sh,” which has no recognizable pitch, why was Nature’s assumption of repetition nevertheless so right in general?

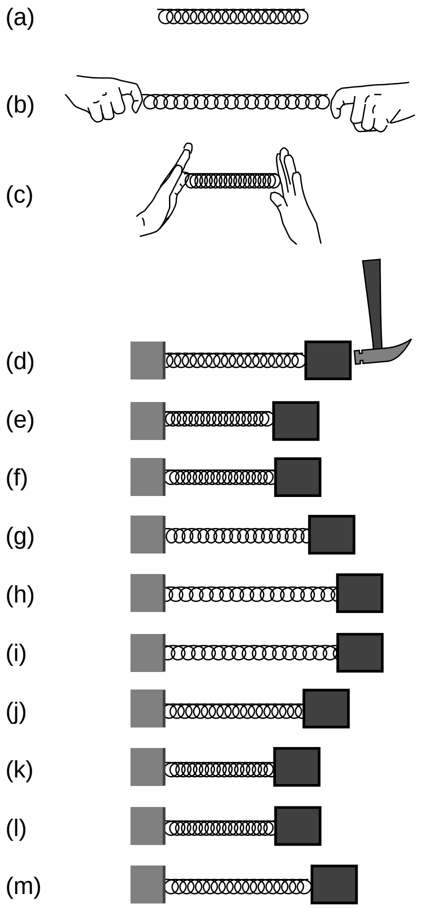
Repeating phenomena occur throughout nature, from the orbits of electrons in atoms to the reappearance of Halley’s Comet every 75 years. Ancient cultures tended to attribute repetitious phenomena like the seasons to the cyclical nature of time itself, but we now have a less mystical explanation. Suppose that instead of Halley’s Comet’s true, repeating elliptical orbit that closes seamlessly upon itself with each revolution, we decide to take a pen and draw a whimsical alternative path that never repeats. We will not be able to draw for very long without having the path cross itself. But at such a crossing point, the comet has returned to a place it visited once before, and since its potential energy is the same as it was on the last visit, conservation of energy proves that it must again have the same kinetic energy and therefore the same speed. Not only that, but the comet's direction of motion cannot be randomly chosen, because angular momentum must be conserved as well. Although this falls short of being an ironclad proof that the comet's orbit must repeat, it no longer seems surprising that it does.



If we try to draw a non-repeating orbit for Halley's Comet, it will inevitably end up crossing itself.

Conservation laws, then, provide us with a good reason why repetitive motion is so prevalent in the universe. But it goes deeper than that. Up to this point in your study of physics, we have been indoctrinating you with a mechanistic vision of the universe as a giant piece of clockwork. Breaking the clockwork down into smaller and smaller bits, we end up at the atomic level, where the electrons circling the nucleus resemble — well, little clocks! From this point of view, particles of matter are the fundamental building blocks of everything, and vibrations and waves are just a couple of the tricks that groups of particles can do. But at the beginning of the 20th century, the tabled were turned. A chain of discoveries initiated by Albert Einstein led to the realization that the so-called subatomic "particles" were in fact waves. In this new world-view, it is vibrations and waves that are fundamental, and the formation of matter is just one of the tricks that waves can do.

1.1 Period, Frequency, and Amplitude

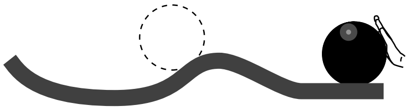


The figure shows our most basic example of a vibration. With no forces on it, the spring assumes its equilibrium length, (a). It can be stretched, (b), or compressed, (c). We attach the spring to a wall on the left and to a mass on the right. If we now hit the mass with a hammer, (d), it oscillates as shown in the series of snapshots, (d)-(m). If we assume that the mass slides back and forth without friction and that the motion is one-dimensional, then conservation of energy proves that the motion must be repetitive. When the block comes back to its initial position again, (g), its potential energy is the same again, so it must have the same kinetic energy again. The motion is in the opposite direction, however. Finally, at (j), it returns to its initial position with the same kinetic energy and the same direction of motion. The motion has gone through one complete cycle, and will now repeat forever in the absence of friction.

The usual physics terminology for motion that repeats itself over and over is **periodic motion**, and the time required for one repetition is called the **period**, T. (The symbol P is not used because of the possible confusion with momentum.) One complete repetition of the motion is called a cycle.

We are used to referring to short-period sound vibrations as "high" in pitch, and it sounds odd to have to say that high pitches have low periods. It is therefore more common to discuss the rapidity of a vibration in terms of the number of vibrations per second, a quantity called the **frequency**, f. Since the period is the number of seconds per cycle and the frequency is the number of cycles per second, they are reciprocals of each other,

$$f = 1/T \quad .$$



Example: a carnival game

In the carnival game shown in the figure, the rube is supposed to push the bowling ball on the track just hard enough so that it goes over the hump and into the valley, but does not come back out again. If the only types of energy involved are kinetic and potential, this is impossible. Suppose you expect the ball to come back to a point such as the one shown with the dashed outline, then stop and turn around. It would already have passed through this point once before, going to the left on its way into the valley. It was moving then, so conservation of energy tells us that it cannot be at rest when it comes back to the same point. The motion that the customer hopes for is physically impossible. There is a physically possible periodic motion in which the ball rolls back and forth, staying confined within the valley, but there is no way to get the ball into that motion beginning from the place where we start. There is a way to beat the game, though. If you put enough spin on the ball, you can create enough kinetic friction so that a significant amount of heat is generated. Conservation of energy then allows the ball to be at rest when it comes back to a point like the outlined one, because kinetic energy has been converted into heat.

Example: Period and frequency of a fly's wing-beats

A Victorian parlor-trick was to listen to the pitch of a fly's buzz, reproduce the musical note on the piano, and announce how many times the fly's wings had flapped in one second. If the fly's wings flap, say, 200 times in one second, then the frequency of their motion is f=200/1 s=200 s^{-1}. The period is one 200th of a second, T=1/f=(1/200) s=0.005 s.

Units of inverse second, s^{-1}, are awkward in speech, so an abbreviation has been created. One Hertz, named in honor of a pioneer of radio technology, is one cycle per second. In abbreviated form, 1 Hz=1 s^{-1}. This is the familiar unit used for the frequencies on the radio dial.

Example: frequency of a radio station

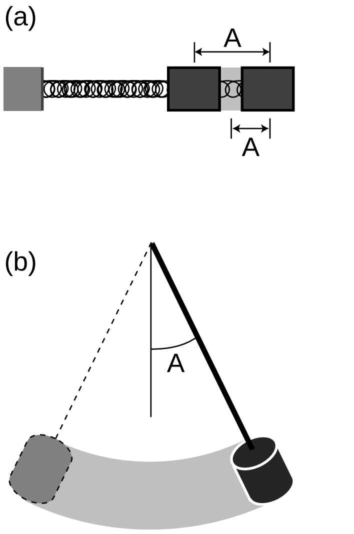
Question: KLON's frequency is 88.1 MHz. What does this mean, and what period does this correspond to?

Solution: The metric prefix M- is mega-, i.e. millions. The radio waves emitted by KLON's transmitting antenna vibrate 88.1 million times per second. This corresponds to a period of

$$T = 1/f = 1.14\text{x}10^{-8}\ \text{s} \quad .$$

This example shows a second reason why we normally speak in terms of frequency rather than period: it would be painful to have to refer to such small time intervals routinely. I could abbreviate by telling people that KLON's period was 11.4 nanoseconds, but most people are more familiar with the big metric prefixes than with the small ones.

Units of frequency are also commonly used to specify the speeds of computers. The idea is that all the little circuits on a computer chip are synchronized by the very fast ticks of an electronic "clock," so that the circuits can all cooperate on a task without any circuit's getting ahead or behind. Adding two numbers might require, say, 30 clock cycles. Microcomputers these days operate at clock frequencies of a few hundred MHz.



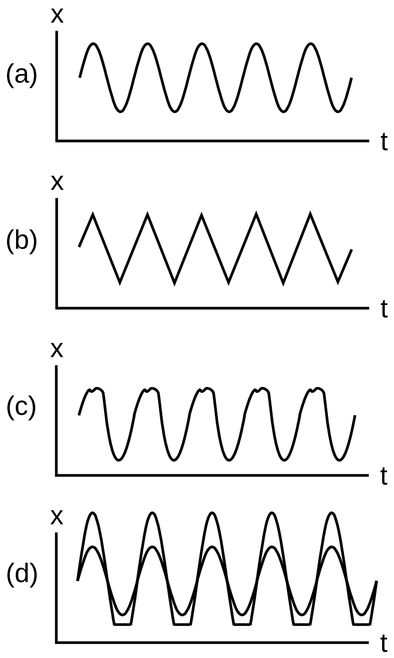
The amplitude of the vibrations of the mass on a spring could be defined in two different ways, (a). It would have units of distance. The amplitude of a swinging pendulum would more naturally be defined as an angle, (b).

We have discussed how to measure how fast something vibrates, but not how big the vibrations are. The general term for this is **amplitude**, A. The definition of amplitude depends on the system being discussed, and two people discussing the same system may not even use the same definition. In the example of the block on the end of the spring, the amplitude will be measured in distance units such as cm. One could work in terms of the distance traveled by the block from the extreme left to the extreme right, but it would be somewhat more common in physics to use the distance from the center to one extreme. The former is usually referred to as the peak-to-peak amplitude, since the extremes of the motion look like mountain peaks or upside-down mountain peaks on a graph of position versus time.

In other situations we would not even use the same units for amplitude. The amplitude of a child on a swing would most conveniently be measured as an angle, not a distance, since her feet will move a greater distance than her head. The electrical vibrations in a radio receiver would be measured in electrical units such as volts or amperes.

1.2 Simple Harmonic Motion

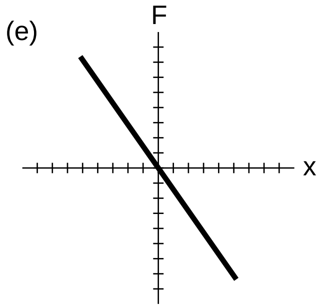
Why are sine-wave vibrations so common?



If we actually construct the mass-on-a-spring system discussed in the previous section and measure its motion accurately, we will find that its *x-t* graph is nearly a perfect sine-wave shape, as shown in figure (a). (We call it a “sine wave” or “sinusoidal” even if it is a cosine, or a sine or cosine shifted by some arbitrary horizontal amount.) It may not be surprising that it is a wiggle of this general sort, but why is it a specific mathematically perfect shape? Why is it not a sawtooth shape like (b) or some other shape like (c)? The mystery deepens as we find that a vast number of apparently unrelated vibrating systems show the same mathematical feature. A tuning fork, a sapling pulled to one side and released, a car bouncing on its shock absorbers, all these systems will exhibit sine-wave motion under one condition: the amplitude of the motion must be small.

It is not hard to see intuitively why extremes of amplitude would act differently. For example, a car that is bouncing lightly on its shock absorbers may behave smoothly, but if we try to double the amplitude of the vibrations the bottom of the car may begin hitting the ground, (d). (Although we are assuming for simplicity in this chapter that energy is never dissipated, this is clearly not a very realistic assumption in this example. Each time the car hits the ground it will convert quite a bit of its potential and kinetic energy into heat and sound, so the vibrations would actually die out quite quickly, rather than repeating for many cycles as shown in the figure.)

The key to understanding how an object vibrates is to know how the force on the object depends on the object’s position. If an object is vibrating to the right and left, then it must have a leftward force on it when it is on the right side, and a rightward force when it is on the left side. In one dimension, we can represent the direction of the force using a positive or negative sign, and since the force changes from positive to negative there must be a point in the middle where the force is zero. This is the equilibrium point, where the object would stay at rest if it was released at rest. For convenience of notation throughout this chapter, we will define the origin of our coordinate system so that *x* equals zero at equilibrium.



The simplest example is the mass on a spring, for which force on the mass is given by Hooke's law,

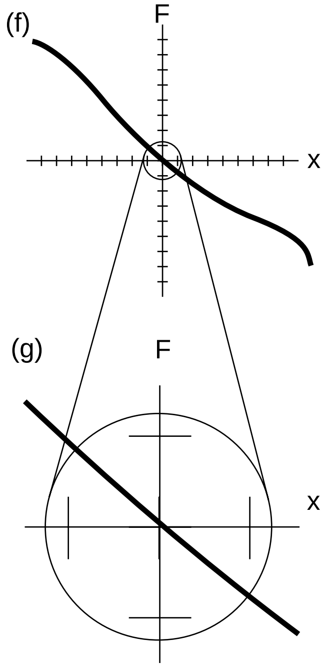
$$F = -kx \quad .$$

We can visualize the behavior of this force using a graph of *F* versus *x*, fig. (e). The graph is a line, and the spring constant, *k*, is equal to minus its slope. A stiffer spring has a larger value of *k* and a steeper slope. Hooke's law is only an approximation, but it works very well for most springs in real life, as long as the spring isn't compressed or stretched so much that it is permanently bent damaged.

The following important theorem, whose proof is given in optional section 1.3, relates the motion graph to the force graph.

> **Theorem: A linear force graph makes a sinusoidal motion graph.**
> If the total force on a vibrating object depends only on the object's position, and is related to the objects displacement from equilibrium by an equation of the form $F=-kx$, then the object's motion displays a sinusoidal graph with period $T = 2\pi\sqrt{m/k}$.

Even if you do not read the proof, it is not too hard to understand why the equation for the period makes sense. A greater mass causes a greater period, since the force will not be able to whip a massive object back and forth very rapidly. A larger value of *k* causes a shorter period, because a stronger force can whip the object back and forth more rapidly.



This may seem like only an obscure theorem about the mass-on-a-spring system, but figures (f) and (g) show it to be far more general than that. Figure (f) depicts a force curve that is not a straight line. A system with this *F-x* curve would have large-amplitude vibrations that were complex and not sinusoidal. But the same system would exhibit sinusoidal small-amplitude vibrations. This is because any curve looks linear from very close up. If we magnify the *F-x* graph as shown in (g), it becomes very difficult to tell that the graph is not a straight line. If the vibrations were confined to the region shown in (g), they would be very nearly sinusoidal. This is the reason why sinusoidal vibrations are a universal feature of all vibrating systems, if we restrict ourselves to small amplitudes. The theorem is therefore of great general significance. It applies throughout the universe, to objects ranging from vibrating stars to vibrating nuclei. A sinusoidal vibration is known as **simple harmonic motion**.

Period is independent of amplitude.

Until now we have not even mentioned the most counterintuitive aspect of the equation $T = 2\pi\sqrt{m/k}$: it does not depend on amplitude at all. Intuitively, most people would expect the mass-on-a-spring system to take longer to complete a cycle if the amplitude was larger. (We are comparing amplitudes that are different from each other, but both small enough that the theorem applies.) In fact the larger-amplitude vibrations take the same amount of time as the small-amplitude ones

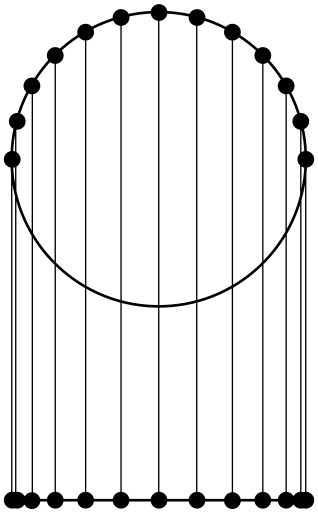
Legend has it that this fact was first noticed by Galileo during what was apparently a less than enthralling church service. A gust of wind would now and then start one of the chandeliers in the cathedral swaying back and forth, and he noticed that regardless of the amplitude of the vibrations, the period of oscillation seemed to be the same. Up until that time, he had been carrying out his physics experiments with such crude time-measuring techniques as feeling his own pulse or singing a tune to keep a musical beat. But after going home and testing a pendulum, he convinced himself that he had found a superior method of measuring time. Even without a fancy system of pulleys to keep the pendulum's vibrations from dying down, he could get very accurate time measurements, because the gradual decrease in amplitude due to friction would have no effect on the pendulum's period. (Galileo never produced a modern-style pendulum clock with pulleys, a minute hand, and a second hand, but within a generation the device had taken on the form that persisted for hundreds of years after.)

Example: the pendulum

Question: Compare the periods of pendula having bobs with different masses.

Solution: From the equation $T=2\pi\sqrt{m/k}$, we might expect that a larger mass would lead to a longer period. However, increasing the mass also increases the forces that act on the pendulum: gravity and the tension in the string. This increases *k* as well as *m*, so the period of a pendulum is independent of *m*.

1.3* Proofs



The object moves along the circle at constant speed, but even though its over-all speed is constant, the x and y components of its velocity are continuously changing, as shown by the unequal spacing of the points when projected onto the line below. Projected onto the line, its motion is the same as that of an object experiencing a force *F*=–*kx*.

In this section we prove (1) that a linear *F-x* graph gives sinusoidal motion, (2) that the period of the motion is $T=2\pi\sqrt{m/k}$, and (3) that the period is independent of the amplitude. You may omit this section without losing the continuity of the chapter.

The basic idea of the proof can be understood by imagining that you are watching a child on a merry-go-round from far away. Because you are in the same horizontal plane as her motion, she appears to be moving from side to side along a line. Circular motion viewed edge-on doesn't just look like any kind of back-and-forth motion, it looks like motion with a sinusoidal *x-t* graph, because the sine and cosine functions can be defined as the *x* and *y* coordinates of a point at angle θ on the unit circle. The idea of the proof, then, is to show that an object acted on by a force that varies as *F*=–*kx* has motion that is identical to circular motion projected down to one dimension. The equation $T=2\pi\sqrt{m/k}$ will also fall out nicely at the end.

For an object performing uniform circular motion, we have

$$|\mathbf{a}| = v^2/r \quad .$$

The *x* component of the acceleration is therefore

$$a_x = -\frac{v^2}{r}\cos\theta \quad ,$$

where θ is the angle measured counterclockwise from the *x* axis. Applying Newton's second law,

$$F_x / m = -\frac{v^2}{r}\cos\theta \quad \text{, so}$$

$$F_x = -m\frac{v^2}{r}\cos\theta \quad .$$

Since our goal is an equation involving the period, it is natural to eliminate the variable v = circumference/T = $2\pi r/T$, giving

$$F_x = -\frac{4\pi^2 mr}{T^2}\cos\theta \quad .$$

The quantity $r\cos\theta$ is the same as x, so we have

$$F_x = -\frac{4\pi^2 m}{T^2}x \quad .$$

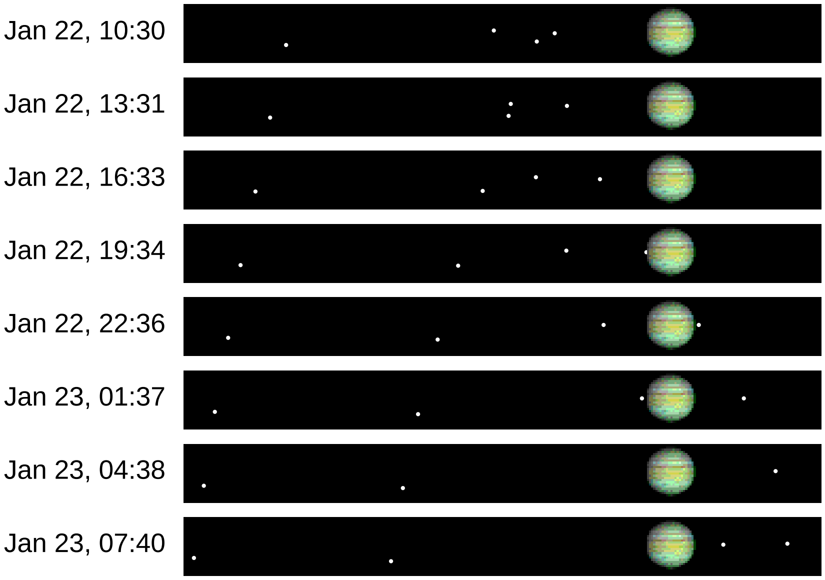
Since everything is constant in this equation except for x, we have proven that motion with force proportional to x is the same as circular motion projected onto a line, and therefore that a force proportional to x gives sinusoidal motion. Finally, we identify the constant factor of $4\pi^2 m / T^2$ with k, and solving for T gives the desired equation for the period,

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

Since this equation is independent of r, T is independent of the amplitude.

Example: The moons of Jupiter.

The idea behind this proof is aptly illustrated by the moons of Jupiter. Their discovery by Galileo was an epochal event in astronomy, because it proved that not everything in the universe had to revolve around the earth as had been believed. Galileo's telescope was of poor quality by modern standards, but the figure below shows a simulation of how Jupiter and its moons might appear at intervals of three hours through a large present-day instrument. Because we see the moons' circular orbits edge-on, they appear to perform sinusoidal vibrations. Over this time period, the innermost moon, Io, completes half a cycle.



Summary

Selected Vocabulary

periodic motion	motion that repeats itself over and over
period	the time required for one cycle of a periodic motion
frequency	the number of cycles per second, the inverse of the period
amplitude	the amount of vibration, often measured from the center to one side; may have different units depending on the nature of the vibration
simple harmonic motion	motion whose *x*-*t* graph is a sine wave

Notation

T	period
f	frequency
A	amplitude
k	the slope of the graph of F versus x, where F is the total force acting on an object and x is the object's position; For a spring, this is known as the spring constant.

Notation Used in Other Books

ν	The Greek letter ν, nu, is used in many books for frequency.
ω	The Greek letter ω, omega, is often used as an abbreviation for $2\pi f$.

Summary

Periodic motion is common in the world around us because of conservation laws. An important example is one-dimensional motion in which the only two forms of energy involved are potential and kinetic; in such a situation, conservation of energy requires that an object repeat its motion, because otherwise when it came back to the same point, it would have to have a different kinetic energy and therefore a different total energy.

Not only are periodic vibrations very common, but small-amplitude vibrations are always sinusoidal as well. That is, the *x*-*t* graph is a sine wave. This is because the graph of force versus position will always look like a straight line on a sufficiently small scale. This type of vibration is called simple harmonic motion. In simple harmonic motion, the period is independent of the amplitude, and is given by

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

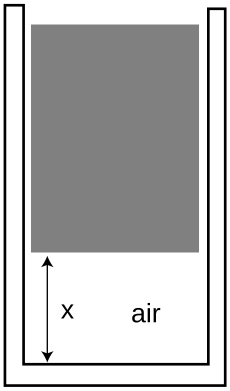
Homework Problems

1. Find an equation for the frequency of simple harmonic motion in terms of k and m.

2. Many single-celled organisms propel themselves through water with long tails, which they wiggle back and forth. (The most obvious example is the sperm cell.) The frequency of the tail's vibration is typically about 10-15 Hz. To what range of periods does this range of frequencies correspond?

3. (a) Pendulum 2 has a string twice as long as pendulum 1. If we define x as the distance traveled by the bob along a circle away from the bottom, how does the k of pendulum 2 compare with the k of pendulum 1? Give a numerical ratio. [Hint: the total force on the bob is the same if the angles away from the bottom are the same, but equal angles do not correspond to equal values of x.]

(b) Based on your answer from part (a), how does the period of pendulum 2 compare with the period of pendulum 1? Give a numerical ratio.



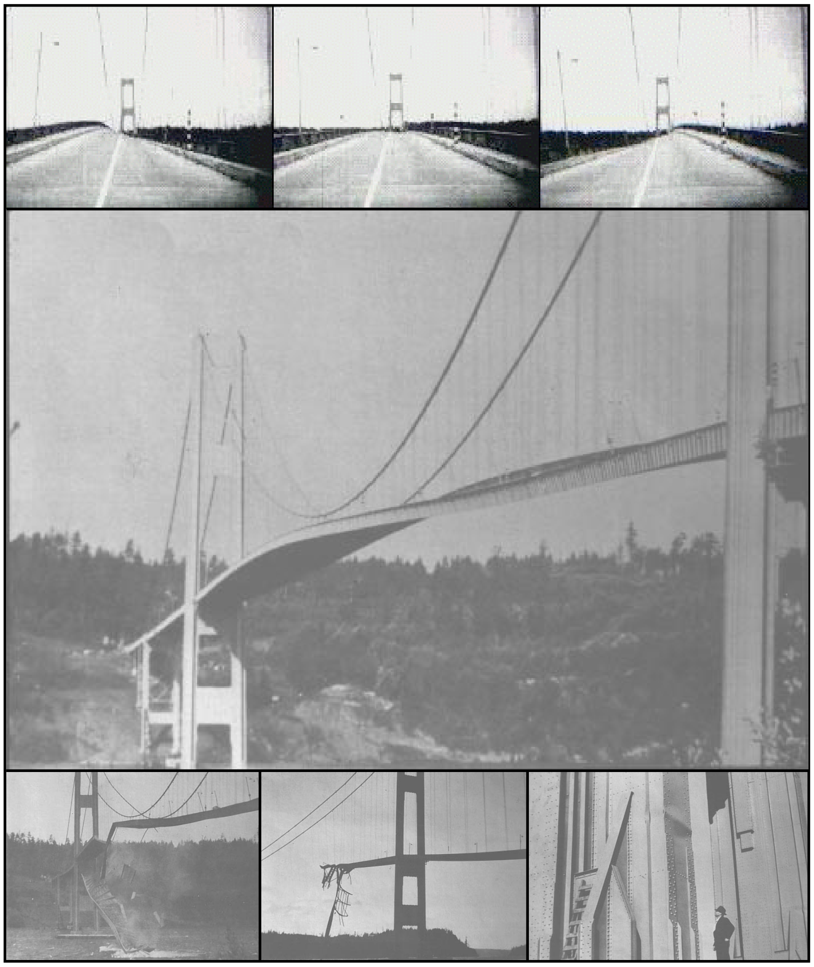
Problems 4 and 5.

4 ✓. A pneumatic spring consists of a piston riding on top of the air in a cylinder. The upward force of the air on the piston is given by $F_{\text{air}}=ax^{-1.4}$, where a is a constant with funny units of $\text{N}\cdot\text{m}^{1.4}$. For simplicity, assume the air only supports the weight, F_W, of the piston itself, although in practice this device is used to support some other object. The equilibrium position, x_0, is where F_W equals $-F_{\text{air}}$. (Note that in the main text I have assumed the equilibrium position to be at x=0, but that is not the natural choice here.) Assume friction is negligible, and consider a case where the amplitude of the vibrations is very small. Let a=1 $\text{N}\cdot\text{m}^{1.4}$, x_0=1.00 m, and F_W=–1.00 N. The piston is released from x=1.01 m. Draw a neat, accurate graph of the total force, F, as a function of x, on graph paper, covering the range from x=0.98 m to 1.02 m. Over this small range, you will find that the force is very nearly proportional to x-x_0. Approximate the curve with a straight line, find its slope, and derive the approximate period of oscillation.

5. Consider the same pneumatic piston described in the previous problem, but now imagine that the oscillations are not small. Sketch a graph of the total force on the piston as it would appear over this wider range of motion. For a wider range of motion, explain why the vibration of the piston about equilibrium is not simple harmonic motion, and sketch a graph of x vs t, showing roughly how the curve is different from a sine wave. [Hint: Acceleration corresponds to the curvature of the x-t graph, so if the force is greater, the graph should curve around more quickly.]

S A solution is given in the back of the book.
✓ A computerized answer check is available.
★ A difficult problem.
∫ A problem that requires calculus.

6. Archimedes' principle states that an object partly or wholly immersed in fluid experiences a buoyant force equal to the weight of the fluid it displaces. For instance, if a boat is floating in water, the upward pressure of the water (vector sum of all the forces of the water pressing inward and upward on every square inch of its hull) must be equal to the weight of the water displaced, because if the boat was instantly removed and the hole in the water filled back in, the force of the surrounding water would be just the right amount to hold up this new "chunk" of water. (a) Show that a cube of mass *m* with edges of length *b* floating upright (not tilted) in a fluid of density ρ will have a draft (depth to which it sinks below the waterline) *h* given at equilibrium by $h_o = m / b^2\rho$. (b) Find the total force on the cube when its draft is *h*, and verify that plugging in $h = h_o$ gives a total force of zero. (c) Find the cube's period of oscillation as it bobs up and down in the water, and show that can be expressed in terms of h_o and *g* only.



Top: A series of images from a silent movie of the bridge vibrating on the day it was to collapse, taken by an unknown amateur photographer.
Middle: The bridge immediately before the collapse, with the sides vibrating 8.5 meters (28 feet) up and down. Note that the bridge is over a mile long.
Bottom: During and after the final collapse. The right-hand picture gives a sense of the massive scale of the construction.

2 Resonance

Soon after the mile-long Tacoma Narrows Bridge opened in July 1940, motorists began to notice its tendency to vibrate frighteningly in even a moderate wind. Nicknamed “Galloping Gertie,” the bridge collapsed in a steady 42-mile-per-hour wind on November 7 of the same year. The following is an eyewitness report from a newspaper editor who found himself on the bridge as the vibrations approached the breaking point.

"Just as I drove past the towers, the bridge began to sway violently from side to side. Before I realized it, the tilt became so violent that I lost control of the car... I jammed on the brakes and got out, only to be thrown onto my face against the curb.

"Around me I could hear concrete cracking. I started to get my dog Tubby, but was thrown again before I could reach the car. The car itself began to slide from side to side of the roadway.

"On hands and knees most of the time, I crawled 500 yards or more to the towers... My breath was coming in gasps; my knees were raw and bleeding, my hands bruised and swollen from gripping the concrete curb... Toward the last, I risked rising to my feet and running a few yards at a time... Safely back at the toll plaza, I saw the bridge in its final collapse and saw my car plunge into the Narrows."

The ruins of the bridge formed an artificial reef, one of the world's largest. It was not replaced for ten years. The reason for its collapse was not substandard materials or construction, nor was the bridge underdesigned: the piers were hundred-foot blocks of concrete, the girders massive and made of carbon steel. The bridge was destroyed because of the physical phenomenon of resonance, the same effect that allows an opera singer to break a wine glass with her voice and that lets you tune in the radio station you want. The replacement bridge, which has lasted half a century so far, was built smarter, not stronger. The engineers learned their lesson and simply included some slight modifications to avoid the resonance phenomenon that spelled the doom of the first one.

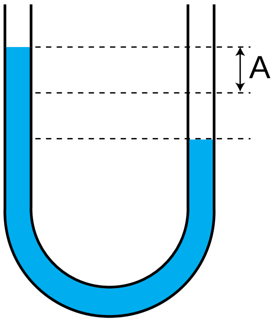
2.1 Energy in Vibrations

One way of describing the collapse of the bridge is that the bridge kept taking energy from the steadily blowing wind and building up more and more energetic vibrations. In this section, we discuss the energy contained in a vibration, and in the subsequent sections we will move on to the loss of energy and the adding of energy to a vibrating system, all with the goal of understanding the important phenomenon of resonance.

Going back to our standard example of a mass on a spring, we find that there are two forms of energy involved: the potential energy stored in the spring and the kinetic energy of the moving mass. We may start the system in motion either by hitting the mass to put in kinetic energy by pulling it to one side to put in potential energy. Either way, the subsequent behavior of the system is identical. It trades energy back and forth between kinetic and potential energy. (We are still assuming there is no friction, so that no energy is converted to heat, and the system never runs down.)

The most important thing to understand about the energy content of vibrations is that the total energy is proportional to the amplitude. Although the total energy is constant, it is instructive to consider two specific moments in the motion of the mass on a spring as examples. When the mass is all the way to one side, at rest and ready to reverse directions, all its energy is potential. We have already seen that the potential energy stored in a spring equals $\frac{1}{2}kx^2$, so the energy is proportional to the square of the amplitude. Now consider the moment when the mass is passing through the equilibrium point at x=0. At this point it has no potential energy, but it does have kinetic energy. The velocity is proportional to the amplitude of the motion, and the kinetic energy, $\frac{1}{2}mv^2$, is proportional to the square of the velocity, so again we find that the energy is proportional to the square of the amplitude. The reason for singling out these two points is merely instructive; proving that energy is proportional to A^2 at any point would suffice to prove that energy is proportional to A^2 in general, since the energy is constant.

Are these conclusions restricted to the mass-on-a-spring example? No. We have already seen that F=–kx is a valid approximation for any vibrating object, as long as the amplitude is small. We are thus left with a very general conclusion: the energy of any vibration is approximately proportional to the square of the amplitude, provided that the amplitude is small.



Example: water in a U-tube
If water is poured into a U-shaped tube as shown in the figure, it can undergo vibrations about equilibrium. The energy of such a vibration is most easily calculated by considering the “turnaround point” when the water has stopped and is about to reverse directions. At this point, it has only potential energy and no kinetic energy, so by calculatings its potential energy we can find the energy of the vibration. This potential energy is the same as the work that would have to be done to take the water out of the right-hand side down to a depth A below the equilibrium level, raise it through a height A, and place it in the left-hand side. The weight of this chunk of water is proportional to A, and so is the height through which it must be lifted, so the energy is proportional to A^2.

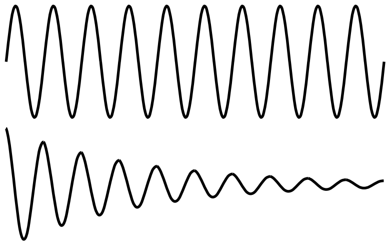
Example: the range of energies of sound waves
Question: The amplitude of vibration of your eardrum at the threshold of pain is about 10^6 times greater than the amplitude with which it vibrates in response to the softest sound you can hear. How many times greater is the energy with which your ear has to cope for the painfully loud sound, compared to the soft sound?
Solution: The amplitude is 10^6 times greater, and energy is proportional to the square of the amplitude, so the energy is greater by a factor of 10^{12}. This is a phenomenally large factor!

We are only studying vibrations right now, not waves, so we are not yet concerned with how a sound wave works, or how the energy gets to us through the air. Note that because of the huge range of energies that our ear can sense, it would not be reasonable to have a sense of loudness that was additive. Consider, for instance, the following three levels of sound:

barely audible, gentle wind	
quiet conversation	10^5 times more energy than the wind
heavy metal concert..................	10^{12} times more energy than the wind

In terms of addition and subtraction, the difference between the wind and the quiet conversation is nothing compared to the difference between the quiet conversation and the heavy metal concert. Evolution wanted our sense of hearing to be able to encompass all these sounds without collapsing the bottom of the scale so that anything softer than the crack of doom would sound the same. So rather than making our sense of loudness additive, mother nature made it multiplicative. We sense the difference between the wind and the quiet conversation as spanning a range of about 5/12 as much as the whole range from the wind to the heavy metal concert. Although a detailed discussion of the decibel scale is not relevant here, the basic point to note about the decibel scale is that it is logarithmic. The zero of the decibel scale is close to the lower limit of human hearing, and adding 1 unit to the decibel measurement corresponds to *multiplying* the energy level (or actually the power per unit area) by a certain factor.

2.2 Energy Lost From Vibrations



Friction has the effect of pinching the x-t graph of a vibrating object.

Until now, we have been making the relatively unrealistic assumption that a vibration would never die out. For a realistic mass on a spring, there will be friction, and the kinetic and potential energy of the vibrations will therefore be gradually converted into heat. Similarly, a guitar string will slowly convert its kinetic and potential energy into sound. In all cases, the effect is to "pinch" the sinusoidal *x-t* graph more and more with passing time. Friction is not necessarily bad in this context — a musical instrument that never got rid of any of its energy would be completely silent! The dissipation of the energy in a vibration is known as damping.

Self-Check

Most people who try to draw graphs like those shown on the left will tend to shrink their wiggles horizontally as well as vertically. Why is this wrong?

In the graphs on the left, I have not shown any point at which the damped vibration finally stops completely. Is this realistic? Yes and no. If energy is being lost due to friction between two solid surfaces, then we expect the force of friction to be nearly independent of velocity. This constant friction force puts an upper limit on the total distance that the vibrating object can ever travel without replenishing its energy, since work equals force times distance, and the object must stop doing work when its energy is all converted into heat. (The friction force does reverse directions when the object turns around, but reversing the direction of the motion at the same time that we reverse the direction of the force makes it certain that the object is always doing positive work, not negative work.)

Damping due to a constant friction force is not the only possibility however, or even the most common one. A pendulum may be damped mainly by air friction, which is approximately proportional to v^2, while other systems may exhibit friction forces that are proportional to v. It turns out that friction proportional to v is the simplest case to analyze mathematically, and anyhow all the important physical insights can be gained by studying this case.

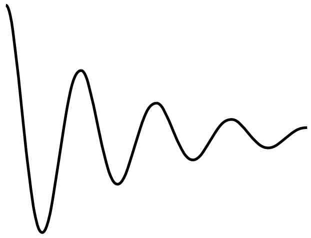
If the friction force is proportional to v, then as the vibrations die down, the frictional forces get weaker due to the lower speeds. The less energy is left in the system, the more miserly the system becomes with giving away any more energy. Under these conditions, the vibrations theoretically never die out completely, and mathematically, the loss of energy from the system is exponential: the system loses a fixed percentage of its energy per cycle. This is referred to as exponential decay.

A nonrigorous proof is as follows. The force of friction is proportional to v, and v is proportional to how far the objects travels in one cycle, so the frictional force is proportional to amplitude. The amount of work done by friction is proportional to the force and to the distance traveled, so the work done in one cycle is proportional to the square of the amplitude. Since both the work and the energy are proportional to A^2, the amount of energy taken away by friction in one cycle is a fixed percentage of the amount of energy the system has.

The horizontal axis is a time axis, and the period of the vibrations is independent of amplitude. Shrinking the amplitude does not make the cycles any faster.

Self-Check

The figure shows an x-t graph for a strongly damped vibration, which loses half of its amplitude with every cycle. What fraction of the energy is lost in each cycle?



It is customary to describe the amount of damping with a quantity called the **quality factor**, *Q*, defined as the number of cycles required for the energy to fall off by a factor of 535. (The origin of this obscure numerical factor is $e^{2\pi}$, where e=2.71828... is the base of natural logarithms.) The terminology arises from the fact that friction is often considered a bad thing, so a mechanical device that can vibrate for many oscillations before it loses a significant fraction of its energy would be considered a high-quality device.

Example: exponential decay in a trumpet

Question: The vibrations of the air column inside a trumpet have a *Q* of about 10. This means that even after the trumpet player stops blowing, the note will keep sounding for a short time. If the player suddenly stops blowing, how will the sound intensity 20 cycles later compare with the sound intensity while she was still blowing?

Solution: The trumpet's *Q* is 10, so after 10 cycles the energy will have fallen off by a factor of 535. After another 10 cycles we lose another factor of 535, so the sound intensity is reduced by a factor of 535x535=2.9x10^5.

The decay of a musical sound is part of what gives it its character, and a good musical instrument should have the right *Q*, but the *Q* that is considered desirable is different for different instruments. A guitar is meant to keep on sounding for a long time after a string has been plucked, and might have a *Q* of 1000 or 10000. One of the reasons why a cheap synthesizer sounds so bad is that the sound suddenly cuts off after a key is released.

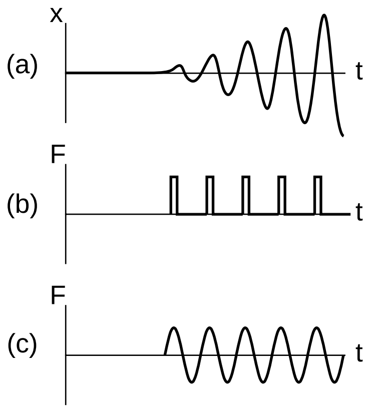
Example: Q of a stereo speaker

Stereo speakers are not supposed to reverberate or "ring" after an electrical signal that stops suddenly. After all, the recorded music was made by musicians who knew how to shape the decays of their notes correctly. Adding a longer "tail" on every note would make it sound wrong. We therefore expect that stereo speaker will have a very low *Q*, and indeed, most speakers are designed with a *Q* of about 1. (Low-quality speakers with larger *Q* values are referred to as "boomy.")

We will see later in the chapter that there are other reasons why a speaker should not have a high *Q*.

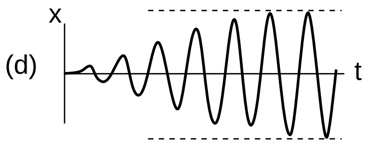
Energy is proportional to the square of amplitude, so its energy is four times smaller after every cycle. It loses three quarters of its energy with each cycle.

2.3 Putting Energy Into Vibrations



When pushing a child on a swing, you cannot just apply a constant force. A constant force will move the swing out to a certain angle, but will not allow the swing to start swinging. Nor can you give short pushes at randomly chosen times. That type of random pushing would increase the child's kinetic energy whenever you happened to be pushing in the same direction as her motion, but it would reduce her energy when your pushing happened to be in the opposite direction compared to her motion. To make her build up her energy, you need to make your pushes rhythmic, pushing at the same point in each cycle. In other words, your force needs to form a repeating pattern with the same frequency as the normal frequency of vibration of the swing. Graph (a) shows what the child's *x-t* graph would look like as you gradually put more and more energy into her vibrations. A graph of your *force* versus time would probably look something like graph (b). It turns out, however, that it is much simpler mathematically to consider a vibration with energy being pumped into it by a driving force that is itself a sine-wave, (c). A good example of this is your eardrum being driven by the force of a sound wave.

Now we know realistically that the child on the swing will not keep increasing her energy forever, nor does your eardrum end up exploding because a continuing sound wave keeps pumping more and more energy into it. In any realistic system, there is energy going out as well as in. As the vibrations increase in amplitude, there is an increase in the amount of energy taken away by damping with each cycle. This occurs for two reasons. Work equals force times distance (or, more accurately, the area under the force-distance curve). As the amplitude of the vibrations increases, the damping force is being applied over a longer distance. Furthermore, the damping force usually increases with velocity (we usually assume for simplicity that it is proportional to velocity), and this also serves to increase the rate at which damping forces remove energy as the amplitude increases. Eventually (and small children and our eardrums are thankful for this!), the amplitude approaches a maximum value (d) at which energy is removed by the damping force just as quickly as it is being put in by the driving force.



This process of approaching a maximum amplitude happens extremely quickly in many cases, e.g. the ear or a radio receiver, and we don't even notice that it took a millisecond or a microsecond for the vibrations to "build up steam." We are therefore mainly interested in predicting the behavior of the system once it has had enough time to reach essentially its maximum amplitude. This is known as the steady-state behavior of a vibrating system.

Now comes the interesting part: what happens if the frequency of the driving force is mismatched to the frequency at which the system would naturally vibrate on its own? We all know that a radio station doesn't have to be tuned in exactly, although there is only a small range over which a given station can be received. The designers of the radio had to make the range fairly small to make it possible eliminate unwanted stations that happened to be nearby in frequency, but it couldn't be too small or you wouldn't be able to adjust the knob accurately enough. (Even a digital radio can be tuned to 88.0 MHz and still bring in a station at 88.1 MHz.) The ear also has some natural frequency of vibration, but in this case the range of frequencies to which it can respond is quite broad. Evolution has made the ear's frequency response as broad as possible because it was to our ancestors' advantage to be able to hear everything from a low roars to a high-pitched shriek.

The remainder of this section develops four important facts about the response of a system to a driving force whose frequency is not necessarily the same is the system's natural frequency of vibration. The style is approximate and intuitive, but proofs are given in the subsequent optional section.

First, although we know the ear has a frequency — about 4000 Hz — at which it would vibrate naturally, it does not vibrate at 4000 Hz in response to a low-pitched 200 Hz tone. It always responds at the frequency at which it is driven. Otherwise all pitches would sound like 4000 Hz to us. This is a general fact about driven vibrations:

> (1) The steady-state response to a sinusoidal driving force occurs at the frequency of the force, not at the system's own natural frequency of vibration.

Now let's think about the amplitude of the steady-state response. Imagine that a child on a swing has a natural frequency of vibration of 1 Hz, but we are going to try to make her swing back and forth at 3 Hz. We intuitively realize that quite a large force would be needed to achieve an amplitude of even 30 cm, i.e. the amplitude is less in proportion to the force. When we push at the natural frequency of 1 Hz, we are essentially just pumping energy back into the system to compensate for the loss of energy due to the damping (friction) force. At 3 Hz, however, we are not just counteracting friction. We are also providing an extra force to make the child's momentum reverse itself more rapidly than it would if gravity and the tension in the chain were the only forces acting. It is as if we are artificially increasing the k of the swing, but this is wasted effort because we spend just as much time decelerating the child (taking energy out of the system) as accelerating her (putting energy in).

Now imagine the case in which we drive the child at a very low frequency, say 0.02 Hz or about one vibration per minute. We are essentially just holding the child in position while very slowly walking back and forth. Again we intuitively recognize that the amplitude will be very small in proportion to our driving force. Imagine how hard it would be to hold the child at our own head-level when she is at the end of her swing! As in the too-fast 3 Hz case, we are spending most of our effort in artificially changing the k of the swing, but now rather than reinforcing the gravity and tension forces we are working against them, effectively reducing k. Only a very small part of our force goes into counteracting friction, and the rest is used in repetitively putting potential energy in on the upswing and taking it back out on the downswing, without any long-term gain.

We can now generalize to make the following statement, which is true for all driven vibrations:

(2) A vibrating system **resonates** at its own natural frequency. That is, the amplitude of the steady-state response is greatest in proportion to the amount of driving force when the driving force matches the natural frequency of vibration.

Example: an opera singer breaking a wineglass
In order to break a wineglass by singing, an opera singer must first tap the glass to find its natural frequency of vibration, and then sing the same note back.

Example: collapse of the Nimitz Freeway in an earthquake
I led off the chapter with the dramatic collapse of the Tacoma Narrows Bridge, mainly because a it was well documented by a local physics professor, and an unknown person made a movie of the collapse. The collapse a section of the Nimitz Freeway in Oakland, CA, during a 1989 earthquake is however a simpler example to analyze.

The collapsed section of the Nimitz Freeway.

An earthquake consists of many low-frequency vibrations that occur simultaneously, which is why it sounds like a rumble of indeterminate pitch rather than a low hum. The frequencies that we can hear are not even the strongest ones; most of the energy is in the form of vibrations in the range of frequencies from about 1 Hz to 10 Hz.

Now all the structures we build are resting on geological layers of dirt, mud, sand, or rock. When an earthquake wave comes along, the topmost layer acts like a system with a certain natural frequency of vibration, sort of like a cube of jello on a plate being shaken from side to side. The resonant frequency of the layer depends on how stiff it is and also on how deep it is. The ill-fated section of the Nimitz freeway was built on a layer of mud, and analysis by geologist Susan E. Hough of the U.S. Geological Survey shows that the mud layer's resonance was centered on about 2.5 Hz, and had a width covering a range from about 1 Hz to 4 Hz.

When the earthquake wave came along with its mixture of frequencies, the mud responded strongly to those that were close to its own natural 2.5 Hz frequency. Unfortunately, an engineering analysis after the quake showed that the overpass itself had a resonant frequency of 2.5 Hz as well! The mud responded strongly to the earthquake waves with frequencies close to 2.5 Hz, and the bridge responded strongly to the 2.5 Hz vibrations of the mud, causing sections of it to collapse.

Example: Collapse of the Tacoma Narrows Bridge
Let's now examine the more conceptually difficult case of the Tacoma Narrows Bridge. The surprise here is that the wind was steady. If the wind was blowing at constant velocity, why did it shake the bridge back and forth? The answer is that once the bridge started twisting a little, it began acting like a kite or an airplane wing. It established swirling patterns of air flow around itself, of the kind that you can see in a moving cloud of smoke. Thus even though the wind velocity measured far away from any structures was constant, the bridge's vibration at its own natural frequency of 0.2 Hz set up an alternating pattern of wind gusts in the air immediately around it, which then increased the amplitude of the bridge's vibrations. This vicious cycle fed upon itself, increasing the amplitude of the vibrations until the bridge finally collapsed.

As long as we're on the subject of collapsing bridges, it is worth bringing up the reports of bridges falling down when soldiers marching over them happened to step in rhythm with the bridge's natural frequency of oscillation. This is supposed to have happened in 1831 in Manchester, England, and again in 1849 in Anjou, France. Many modern engineers and scientists, however, are suspicious of the analysis of these reports. It is possible that the collapses had more to do with poor construction and overloading than with resonance. The Nimitz Freeway and Tacoma Narrows Bridge are far better documented, and occurred in an era when engineers' abilities to analyze the vibrations of a complex structure were much more advanced.

Example: emission and absorption of light waves by atoms

In a very thin gas, the atoms are sufficiently far apart that they can act as individual vibrating systems. Although the vibrations are of a very strange and abstract type described by the theory of quantum mechanics, they nevertheless obey the same basic rules as ordinary mechanical vibrations. When a thin gas made of a certain element is heated, it emits light waves with certain specific frequencies, which are like a fingerprint of that element. As with all other vibrations, these atomic vibrations respond most strongly to a driving force that matches their own natural frequency. Thus if we have a relatively cold gas with light waves of various frequencies passing through it, the gas will absorb light at precisely those frequencies at which it would emit light if heated.

(3) When a system is driven at resonance, the steady-state vibrations have an amplitude that is proportional to Q.

This is fairly intuitive. The steady-state behavior is an equilibrium between energy input from the driving force and energy loss due to damping. A low-Q oscillator, i.e. one with strong damping, dumps its energy faster, resulting in lower-amplitude steady-state motion.

Self-Check

If an opera singer is shopping for a wine glass that she can impress her friends by breaking, what should she look for?

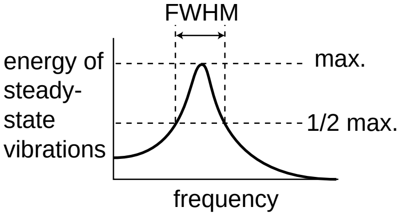
Example: Piano strings ringing in sympathy with a sung note

Question: A sufficiently loud musical note sung near a piano with the lid raised can cause the corresponding strings in the piano to vibrate. (A piano has a set of three strings for each note, all struck by the same hammer.) Why would this trick be unlikely to work with a violin?

Solution: If you have heard the sound of a violin being plucked (the pizzicato effect), you know that the note dies away very quickly. In other words, a violin's Q is much lower than a piano's. This means that its resonances are much weaker in amplitude.

She should tap the wineglasses she finds in the store and look for one with a high Q, i.e. one whose vibrations die out very slowly. The one with the highest Q will have the highest-amplitude response to her driving force, making it more likely to break.

Our fourth and final fact about resonance is perhaps the most surprising. It gives us a way to determine numerically how wide a range of driving frequencies will produce a strong response. As shown in the graph, resonances do not suddenly fall off to zero outside a certain frequency range. It is usual to describe the width of a resonance by its full width at half-maximum (FWHM) as illustrated on the graph.



(4) The FWHM of a resonance is related to its Q and its resonant frequency f_{res} by the equation

$$\text{FWHM} = \frac{f_{\text{res}}}{Q} \ .$$

(This equation is only a good approximation when Q is large.)

Why? It is not immediately obvious that there should be any logical relationship between Q and the FWHM. Here's the idea. As we have seen already, the reason why the response of an oscillator is smaller away from resonance is that much of the driving force is being used to make the system act as if it had a different k. Roughly speaking, the half-maximum points on the graph correspond to the places where the amount of the driving force being wasted in this way is the same as the amount of driving force being used productively to replace the energy being dumped out by the damping force. If the damping force is strong, then a large amount of force is needed to counteract it, and we can waste quite a bit of driving force on changing k before it becomes comparable to the to it. If, on the other hand, the damping force is weak, then even a small amount of force being wasted on changing k will become significant in proportion, and we cannot get very far from the resonant frequency before the two are comparable.

Example: Changing the pitch of a wind instrument
Question: A saxophone player normally selects which note to play by choosing a certain fingering, which gives the saxophone a certain resonant frequency. The musician can also, however, change the pitch significantly by altering the tightness of her lips. This corresponds to driving the horn slightly off of resonance. If the pitch can be altered by about 5% up or down (about one musical half-step) without too much effort, roughly what is the Q of a saxophone?
Solution: Five percent is the width on one side of the resonance, so the full width is about 10%, FWHM / f_{res}=0.1. This implies a Q of about 10, i.e. once the musician stops blowing, the horn will continue sounding for about 10 cycles before its energy falls off by a factor of e. (Blues and jazz saxophone players will typically choose a mouthpiece that produces a low Q, so that they can produce the bluesy pitch-slides typical of their style. "Legit," i.e. classically oriented players, use a higher-Q setup because their style only calls for enough pitch variation to produce a vibrato.)

Example: decay of a saxophone tone
Question: If a typical saxophone setup has a Q of about 10, how long will it take for a 100-Hz tone played on a baritone saxophone to die down by a factor of 535 in energy, after the player suddenly stops blowing?
Solution: A Q of 10 means that it takes 10 cycles for the vibrations to die down in energy by a factor of 535. Ten cycles at a

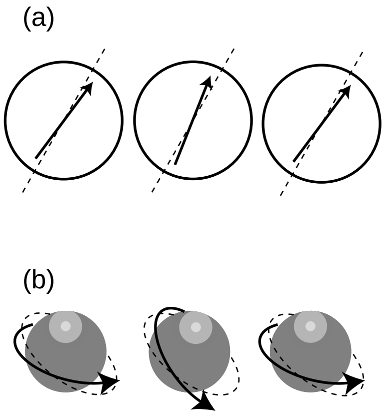
frequency of 100 Hz would correspond to a time of 0.1 seconds, which is not very long. This is why a saxophone note doesn't "ring" like a note played on a piano or an electric guitar.

Example: Q of a radio receiver
Question: A radio receiver used in the FM band needs to be tuned in to within about 0.1 MHz for signals at about 100 MHz. What is its *Q*?
Solution: $Q = f_{\text{res}}$ / FWHM =1000. This is an extremely high *Q* compared to most mechanical systems.

Example: Q of a stereo speaker
We have already given one reason why a stereo speaker should have a low *Q:* otherwise it would continue ringing after the end of the musical note on the recording. The second reason is that we want it to be able to respond to a large range of frequencies.

Example: Nuclear magnetic resonance

If you have ever played with a magnetic compass, you have undoubtedly noticed that if you shake it, it takes some time to settle down. As it settles down, it acts like a damped oscillator of the type we have been discussing. The compass needle is simply a small magnet, and the planet earth is a big magnet. The magnetic forces between them tend to bring the needle to an equilibrium position in which it lines up with the planet-earth-magnet.

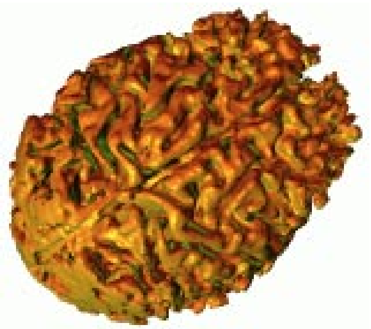


(a) A compass needle vibrates about its equilibrium position under the influence of the earth's magnetic forces. (b) The orientation of a proton's spin vibrates about its equilibrium direction under the influence of the magnetic forces coming from the surrounding electrons and nuclei.

Essentially the same physics lies behind the technique called Nuclear Magnetic Resonance (NMR). NMR is a technique used to deduce the molecular structure of unknown chemical substances, and it is also used for making medical images of the inside of people's bodies. If you ever have an NMR scan, they will actually tell you you are undergoing "magnetic resonance imaging" or "MRI," because people are scared of the word "nuclear." In fact, the nuclei being referred to are simply the nonradioactive nuclei of atoms found naturally in your body.

Here's how NMR works. Your body contains large numbers of hydrogen atoms, each consisting of a small, lightweight electron orbiting around a large, heavy proton. That is, the nucleus of a hydrogen atom is just one proton. A proton is always spinning on its own axis, and the combination of its spin and its electrical charge cause it to behave like a tiny magnet. The principle identical to that of an electromagnet, which consists of a coil of wire through which electrical charges pass; the circling motion of the charges in the coil of wire makes it magnetic, and in the same way, the circling motion of the proton's charge makes it magnetic.

Now a proton in one of your body's hydrogen atoms finds itself surrounded by many other whirling, electrically charged particles: its own electron, plus the electrons and nuclei of the other nearby atoms. These neighbors act like magnets, and exert magnetic forces on the proton. The *k* of the vibrating proton is simply a measure of the total strength of these magnetic forces. Depending on the structure of the molecule in which the hydrogen atom finds itself, there will be a particular set of magnetic forces acting on the proton and a particular value of *k*. The NMR apparatus bombards the sample with radio waves, and if the frequency of the radio waves matches the resonant frequency of the proton, the proton will absorb radio-wave energy strongly and oscillate wildly. Its vibrations are damped not by friction, because there is no friction inside an atom, but by the reemission of radio waves.



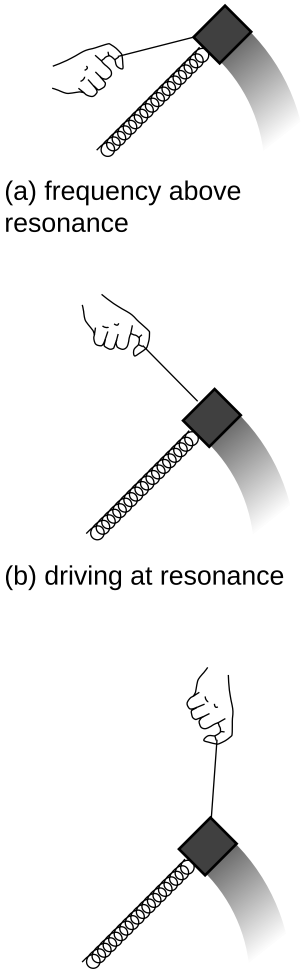
A three-dimensional computer reconstruction of the shape of a human brain, based on magnetic resonance data. R. Malladi, LBNL.

By working backward through this chain of reasoning, one can determine the geometric arrangement of the hydrogen atom's neighboring atoms. It is also possible to locate atoms in space, allowing medical images to be made.

Finally, it should be noted that the behavior of the proton cannot be described entirely correctly by Newtonian physics. Its vibrations are of the strange and spooky kind described by the laws of quantum mechanics. It is impressive, however, that the few simple ideas we have learned about resonance can still be applied successfully to describe many aspects of this exotic system.

2.4* Proofs

Our first goal is to predict the amplitude of the steady-state vibrations as a function of the frequency of the driving force and the amplitude of the driving force. With that equation in hand, we will then prove statements 2, 3, and 4 from the previous section. We assume without proof statement 1, that the steady-state motion occurs at the same frequency as the driving force.


(a) frequency above resonance

(b) driving at resonance

(c) below resonance

As with the proof in the previous chapter, we make use of the fact that a sinusoidal vibration is the same as the projection of circular motion onto a line. We visualize the system shown in the figure, in which the mass swings in a circle on the end of a spring. The spring does not actually change its length at all, but it appears to from the flattened perspective of a person viewing the system edge-on. The radius of the circle is the amplitude, A, of the vibrations as seen edge-on. The damping force can be imagined as a backward drag force supplied by some fluid through which the mass is moving. As usual, we assume that the damping is proportional to velocity, and we use the symbol b for the proportionality constant, $|F_d|=bv$. The driving force, represented by a hand towing the mass with a string, has a tangential component $|F_t|$ which counteracts the damping force, $|F_t|=|F_d|$, and a radial component F_r which works either with or against the spring's force, depending on whether we are driving the system above or below its resonant frequency.

The speed of the rotating mass is the circumference of the circle divided by the period, $v=2\pi A/T$, its acceleration (which is directly inward) is $a=v^2/r$, and Newton's second law gives $a=F/m=(kA+F_r)/m$. We write f_{res} for $(1/2\pi)\sqrt{k/m}$. Straightforward algebra yields

$$\frac{F_r}{F_t} = \frac{2\pi m}{bf}\left(f^2 - f_{res}^2\right) \qquad (1)$$

This is the ratio of the wasted force to the useful force, and we see that it becomes zero when the system is driven at resonance.

The amplitude of the vibrations can be found by attacking the equation $|F_t|=bv=2\pi bAf$, which gives

$$A = \frac{|F_t|}{2\pi bf} \quad . \qquad (2)$$

However, we wish to know the amplitude in terms of $|\boldsymbol{F}|$, not $|F_t|$. From now on, let's drop the cumbersome magnitude symbols. With the Pythagorean theorem, it is easily proven that

$$F_t = \frac{F}{\sqrt{1+\left(\frac{F_r}{F_t}\right)^2}} \quad , \qquad (3)$$

and equations 1-3 are readily combined to give the final result

$$A = \frac{F}{2\pi\sqrt{4\pi^2 m^2\left(f^2 - f_{res}^2\right)^2 + b^2 f^2}} \quad . \quad (4)$$

Statement 2: maximum amplitude at resonance

Equation 4 shows directly that the amplitude is maximized when the system is driven at its resonant frequency. At resonance, the first term inside the square root vanishes, and this makes the denominator as small as possible, causing the amplitude to be as big as possible. (Actually this is only approximately true, because it is possible to make A a little bigger by decreasing f a little below f_{res}, which makes the second term smaller. This technical issue is addressed in the homework problems.)

Statement 3: amplitude at resonance proportional to Q

Equation 4 shows that the amplitude at resonance is proportional to $1/b$, and the Q of the system is inversely proportional to b, so the amplitude at resonance is proportional to Q.

Statement 4: FWHM related to Q

We will satisfy ourselves by proving only the proportionality FWHM$\propto f_{res}/Q$, not the actual equation FWHM$=f_{res}/Q$. The energy is proportional to A^2, i.e. to the inverse of the quantity inside the square root in equation 4. At resonance, the first term inside the square root vanishes, and the half-maximum points occur at frequencies for which the whole quantity inside the square root is double its value at resonance, i.e. when the two terms are equal. At the half-maximum points, we have

$$\begin{aligned} f^2 - f_{res}^2 &= \left(f_{res} \pm \frac{FWHM}{2}\right)^2 - f_{res}^2 \\ &= \pm f_{res} FWHM + \tfrac{1}{4} FWHM^2 \qquad (5) \end{aligned}$$

If we assume that the width of the resonance is small compared to the resonant frequency, then the $FWHM^2$ term in equation 5 is negligible compared to the $f_{res}FWHM$ term, and setting the terms in equation 4 equal to each other gives

$$4\pi^2 m^2\left(f_{res} FWHM\right)^2 = b^2 f^2 \quad .$$

We are assuming that the width of the resonance is small compared to the resonant frequency, so f and f_{res} can be taken as synonyms. Thus,

$$FWHM = \frac{b}{2\pi m} \quad .$$

We wish to connect this to Q, which can be interpreted as the energy of the free (undriven) vibrations divided by the work done by damping in one cycle. The former equals $kA^2/2$, and the latter is proportional to the force, $bv \propto bAf_{\text{res}}$, multiplied by the distance traveled, A. (This is only a proportionality, not an equation, since the force is not constant.) We therefore find that Q is proportional to k/bf_{res}. The equation for the FWHM can then be restated as a proportionality $FWHM \propto k/Qf_{\text{res}}m \propto f_{\text{res}}/Q$.

Summary

Selected Vocabulary

damping .... the dissipation of a vibration’s energy into heat energy, or the frictional force that causes the loss of energy

quality factor .... the number of oscillations required for a system’s energy to fall off by a factor of 535 due to damping

driving force .... an external force that pumps energy into a vibrating system

resonance .... the tendency of a vibrating system to respond most strongly to a driving force whose frequency is close to its own natural frequency of vibration

steady state .... the behavior of a vibrating system after it has had plenty of time to settle into a steady response to a driving force

Notation

Q .... the quality factor

f_{res} .... the natural (resonant) frequency of a vibrating system, i.e. the frequency at which it would vibrate if it was simply kicked and left alone

f .... the frequency at which the system actually vibrates, which in the case of a driven system is equal to the frequency of the driving force, not the natural frequency

Summary

The energy of a vibration is always proportional to the square of the amplitude, assuming the amplitude is small. Energy is lost from a vibrating system for various reasons such as the conversion to heat via friction or the emission of sound. This effect, called damping, will cause the vibrations to decay exponentially unless energy is pumped into the system to replace the loss. A driving force that pumps energy into the system may drive the system at its own natural frequency or at some other frequency. When a vibrating system is driven by an external force, we are usually interested in its *steady-state behavior*, i.e. its behavior after it has had time to settle into a steady response to a driving force. In the steady state, the same amount of energy is pumped into the system during each cycle as is lost to damping during the same period.

The following are four important facts about a vibrating system being driven by an external force:

(1) The steady-state response to a sinusoidal driving force occurs at the frequency of the force, not at the system’s own natural frequency of vibration.

(2) A vibrating system *resonates* at its own natural frequency. That is, the amplitude of the steady-state response is greatest in proportion to the amount of driving force when the driving force matches the natural frequency of vibration.

(3) When a system is driven at resonance, the steady-state vibrations have an amplitude that is proportional to Q.

(4) The FWHM of a resonance is related to its Q and its resonant frequency f_{res} by the equation

$$FWHM = \frac{F_{res}}{Q} \ .$$

(This equation is only a good approximation when Q is large.)

Homework Problems

1. If one stereo system is capable of producing 20 watts of sound power and another can put out 50 watts, how many times greater is the amplitude of the sound wave that can be created by the more powerful system? (Assume they are playing the same music.)

2. Many fish have an organ known as a swim bladder, an air-filled cavity whose main purpose is to control the fish's buoyancy an allow it to keep from rising or sinking without having to use its muscles. In some fish, however, the swim bladder (or a small extension of it) is linked to the ear and serves the additional purpose of amplifying sound waves. For a typical fish having such an anatomy, the bladder has a resonant frequency of 300 Hz, the bladder's Q is 3, and the maximum amplification is about a factor of 100 in energy. Over what range of frequencies would the amplification be at least a factor of 50?

3 ∫. As noted in section 2.4, it is only approximately true that the amplitude has its maximum at $f=2\pi\sqrt{k/m}$. Being more careful, we should actually define two different symbols, $f_o=2\pi\sqrt{k/m}$ and f_{res} for the slightly different frequency at which the amplitude is a maximum, i.e. the actual resonant frequency. In this notation, the amplitude as a function of frequency is

$$A = \frac{F}{2\pi\sqrt{4\pi^2m^2\left(f^2-f_o^2\right)^2+b^2f^2}} \quad .$$

Show that the maximum occurs not at f_o but rather at the frequency

$$f_{res} = \sqrt{f_o^2-\frac{b^2}{8\pi^2m^2}} = \sqrt{f_o^2-\tfrac{1}{2}\mathrm{FWHM}^2}$$

Hint: Finding the frequency that minimizes the quantity inside the square root is equivalent to, but much easier than, finding the frequency that maximizes the amplitude.

S A solution is given in the back of the book. ★ A difficult problem.
✓ A computerized answer check is available. ∫ A problem that requires calculus.

4. (a) Let W be the amount of work done by friction per cycle of oscillation, i.e. the amount of energy lost to heat. Find the fraction of energy *remaining* in the oscillations after n cycles of motion.

(b) From this prove the equation $\left(1 - W/E\right)^Q = e^{-2\pi}$.

(c) Use this to prove the approximation $1/Q \approx (1/2\pi) W/E$. [Hint: Use the approximation $\ln(1+x) \approx x$, which is valid for small values of x.]

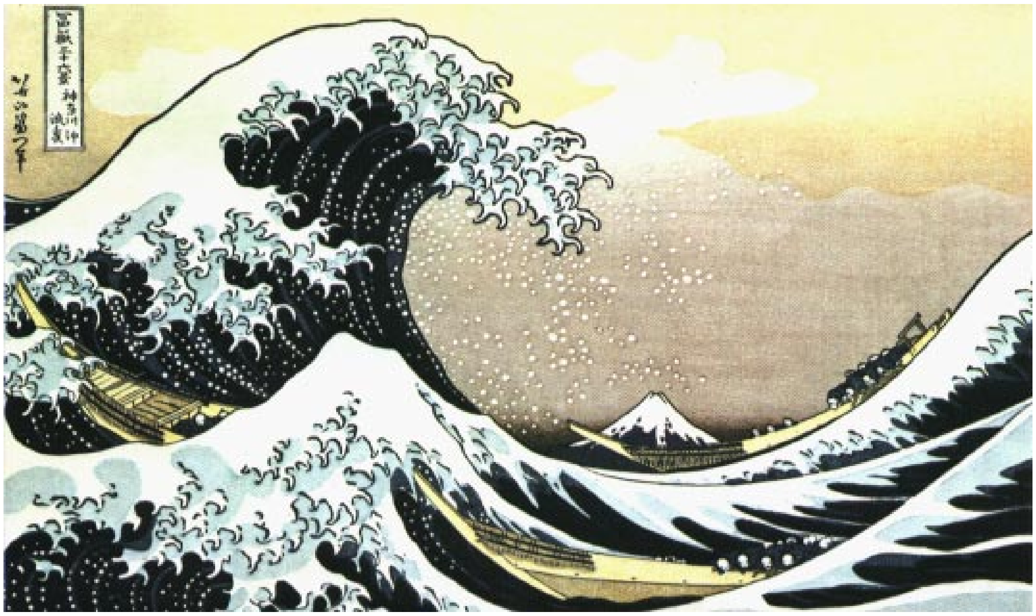
5★ ∫. The goal of this problem is to refine the proportionality FWHM $\propto f_{res}/Q$ into the equation FWHM$=f_{res}/Q$, i.e. to prove that the constant of proportionality equals 1.

(a) Show that the work done by a damping force $F=-bv$ over one cycle of steady-state motion equals $W_{damp}=-2\pi^2 bfA^2$. Hint: It is less confusing to calculate the work done over half a cycle, from $x=-A$ to $x=+A$, and then double it.

(b) Show that the fraction of the undriven oscillator's energy lost to damping over one cycle is $|W_{damp}| / E = 4\pi^2 bf / k$.

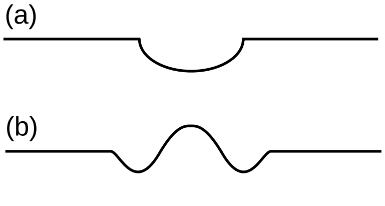
(c) Use the previous result, combined with the result of problem 4, to prove that Q equals $k/2\pi bf$.

(d) Combine the preceding result for Q with the equation FWHM$=b/2\pi m$ from section 2.4 to prove the equation FWHM$=f_{res}/Q$.



3 Free Waves

Your vocal cords or a saxophone reed can vibrate, but being able to vibrate wouldn't be of much use unless the vibrations could be transmitted to the listener's ear by sound waves. What are waves and why do they exist? Put your fingertip in the middle of a cup of water and then remove it suddenly. You will have noticed two results that are surprising to most people. First, the flat surface of the water does not simply sink uniformly to fill in the volume vacated by your finger. Instead, ripples spread out, and the process of flattening out occurs over a long period of time, during which the water at the center vibrates above and below the normal water level. This type of wave motion is the topic of the present chapter. Second, you have found that the ripples bounce off of the walls of the cup, in much the same way that a ball would bounce off of a wall. In the next chapter we discuss what happens to waves that have a boundary around them. Until then, we confine ourselves to wave phenomena that can be analyzed as if the medium (e.g. the water) was infinite and the same everywhere.



It isn't hard to understand why removing your fingertip creates ripples rather than simply allowing the water to sink back down uniformly. The initial crater, (a), left behind by your finger has sloping sides, and the water next to the crater flows downhill to fill in the hole. The water far away, on the other hand, initially has no way of knowing what has happened, because there is no slope for it to flow down. As the hole fills up, the rising water at the center gains upward momentum, and overshoots, creating a little hill where there had been a hole originally. The area just outside of this region has been robbed of some of its water in order to build the hill, so a depressed "moat" is formed, (b). This effect cascades outward, producing ripples.

The two circular patterns of ripples pass through each other. Unlike material objects, wave patterns can overlap in space, and then this happens they combine by addition.

3.1 Wave Motion

There are three main ways in which wave motion differs from the motion of objects made of matter.

1. Superposition

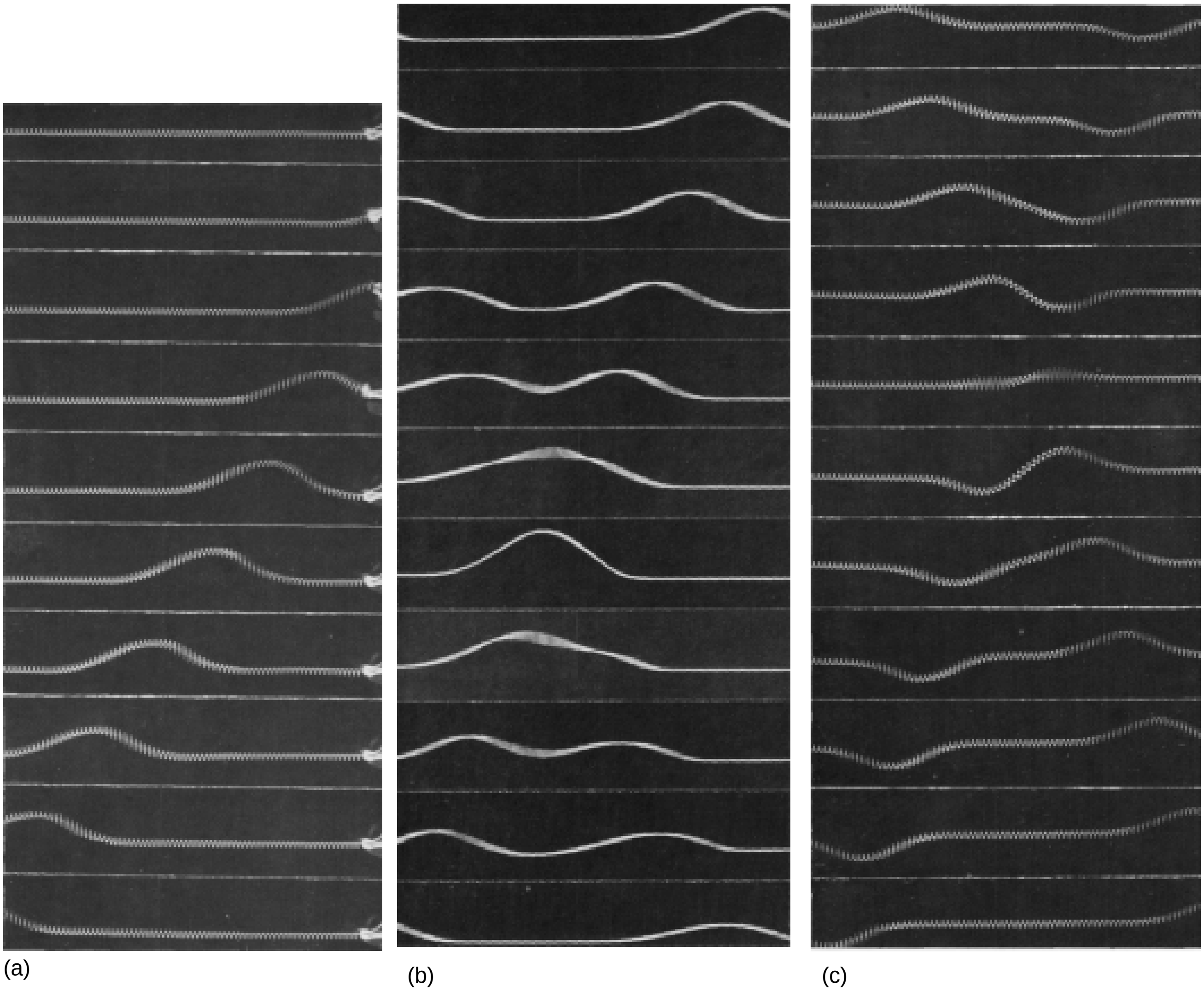
The first, and most profound, difference between wave motion and the motion of objects is that waves do not display any repulsion of each other analogous to the normal forces between objects that come in contact. Two wave patterns can therefore overlap in the same region of space, as shown in the figure at the top of the page. Where the two waves coincide, they add together. For instance, suppose that at a certain location in at a certain moment in time, each wave would have had a crest 3 cm above the normal water level. The waves combine at this point to make a 6-cm crest. We use negative numbers to represent depressions in the water. If both waves would have had a troughs measuring –3 cm, then they combine to make an extra-deep –6 cm trough. A +3 cm crest and a –3 cm trough result in a height of zero, i.e. the waves momentarily cancel each other out at that point. This additive rule is referred to as the *principle of superposition*, "superposition" being merely a fancy word for "adding."

Superposition can occur not just with sinusoidal waves like the ones in the figure above but with waves of any shape. The figures on the following page show superposition of wave *pulses*. A pulse is simply a wave of very short duration. These pulses consist only of a single hump or trough. If you hit a clothesline sharply, you will observe pulses heading off in both directions. This is analogous to the way ripples spread out in all directions when you make a disturbance at one point on water. The same occurs when the hammer on a piano comes up and hits a string.

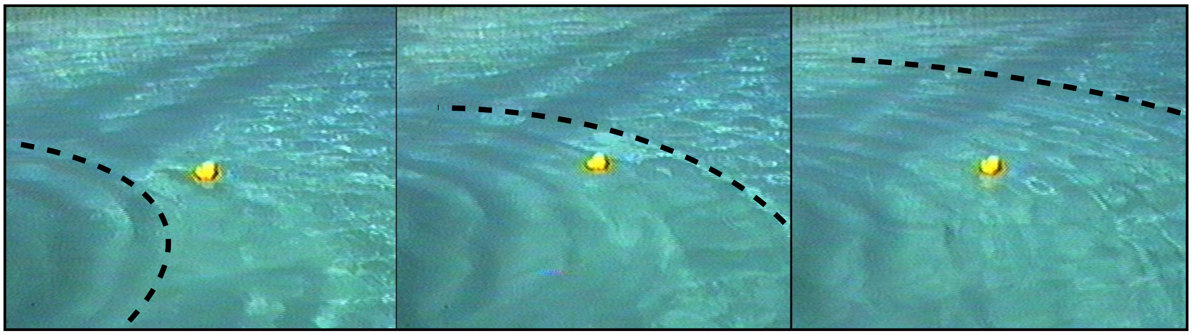
Experiments to date have not shown any deviation from the principle of superposition in the case of light waves. For other types of waves, it is typically a very good approximation for low-energy waves.

Discussion Question

A. In figure (c) below, the fifth frame shows the spring just about perfectly flat. If the two pulses have essentially canceled each other out perfectly, then why does the motion pick up again? Why doesn’t the spring just stay flat?



These pictures show the motion of wave pulses along a spring. To make a pulse, one end of the spring was shaken by hand. Movies were filmed, and a series of frames chosen to show the motion.
(a) A pulse travels to the left. (b) Superposition of two colliding positive pulses. (c) Superposition of two colliding pulses, one positive and one negative.
Uncopyrighted photographs from PSSC Physics.



As the wave pattern passes the rubber duck, the duck stays put. The water isn't moving with the wave.

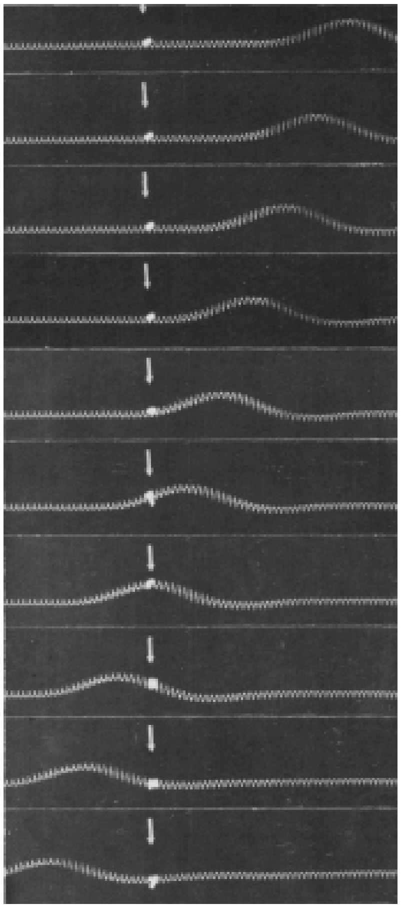
2. The medium is not transported with the wave.

The sequence of three photos above shows a series of water waves before it has reached a rubber duck (left), having just passed the duck (middle) and having progressed about a meter beyond the duck (right). The duck bobs around its initial position, but is not carried along with the wave. This shows that the water itself does not flow outward with the wave. If it did, we could empty one end of a swimming pool simply by kicking up waves! We must distinguish between the motion of the medium (water in this case) and the motion of the wave pattern through the medium. The medium vibrates; the wave progresses through space.

Self-Check

In the photos on the left, you can detect the side-to-side motion of the spring because the spring appears blurry. At a certain instant, represented by a single photo, how would you describe the motion of the different parts of the spring? Other than the flat parts, do any parts of the spring have zero velocity?

The incorrect belief that the medium moves with the wave is often reinforced by garbled secondhand knowledge of surfing. Anyone who has actually surfed knows that the front of the board pushes the water to the sides, creating a wake. If the water was moving along with the wave and the surfer, this wouldn't happen. The surfer is carried forward because forward is downhill, not because of any forward flow of the water. If the water was flowing forward, then a person floating in the water up to her neck would be carried along just as quickly as someone on a surfboard. In fact, it is even possible to surf down the back side of a wave, although the ride wouldn't last very long because the surfer and the wave would quickly part company.



As the wave pulse goes by, the ribbon tied to the spring is not carried along. The motion of the wave pattern is to the right, but the medium (spring) is moving from side to side, not to the right. Uncopyrighted photos from PSSC Physics.

The leading edge is moving up, the trailing edge is moving down, and the top of the hump is motionless for one instant.

3. A wave's velocity depends on the medium.

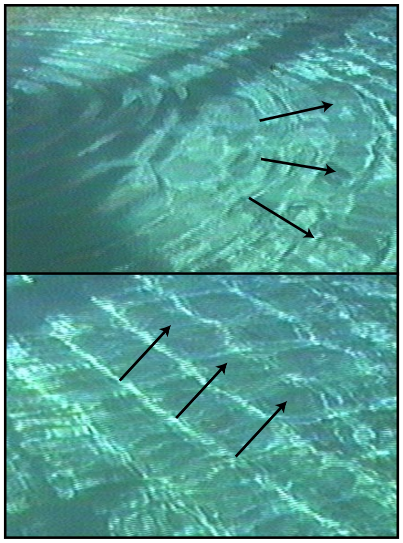
A material object can move with any velocity, and can be sped up or slowed down by a force that increases or decreases its kinetic energy. Not so with waves. The magnitude of a wave's velocity depends on the properties of the medium (and perhaps also on the shape of the wave, for certain types of waves). Sound waves travel at about 340 m/s in air, 1000 m/s in helium. If you kick up water waves in a pool, you will find that kicking harder makes waves that are taller (and therefore carry more energy), not faster. The sound waves from an exploding stick of dynamite carry a lot of energy, but are no faster than any other waves. In the following section we will give an example of the physical relationship between the wave speed and the properties of the medium.

Once a wave is created, the only reason its speed will change is if it enters a different medium or if the properties of the medium change. It is not so surprising that a change in medium can slow down a wave, but the reverse can also happen. A sound wave traveling through a helium balloon will slow down when it emerges into the air, but if it enters another balloon it will speed back up again! Similarly, water waves travel more quickly over deeper water, so a wave will slow down as it passes over an underwater ridge, but speed up again as it emerges into deeper water.

Example: Hull speed

The speeds of most boats (and of some surface-swimming animals) are limited by the fact that they make a wave due to their motion through the water. A fast motor-powered boat can go faster and faster, until it is going at the same speed as the waves it creates. It may then be unable to go any faster, because it cannot climb over the wave crest that builds up in front of it. Increasing the power to the propeller may not help at all. Putting more energy into the waves doesn't make them go any faster, it just makes them taller and more energetic, and that much more difficult to climb over.

A water wave, unlike many other types of wave, has a speed that depends on its shape: a broader wave moves faster. The shape of the wave made by a boat tends to mold itself to the shape of the boat's hull, so a boat with a longer hull makes a broader wave that moves faster. The maximum speed of a boat whose speed is limited by this effect is therefore closely related to the length of its hull, and the maximum speed is called the hull speed. Small racing boats ("cigarette boats") are not just long and skinny to make them more streamlined — they are also long so that their hull speeds will be high.



Circular and linear wave patterns, with velocity vectors shown at selected points.

Wave patterns

If the magnitude of a wave's velocity vector is preordained, what about its direction? Waves spread out in all directions from every point on the disturbance that created them. If the disturbance is small, we may consider it as a single point, and in the case of water waves the resulting wave pattern is the familiar circular ripple. If, on the other hand, we lay a pole on the surface of the water and wiggle it up and down, we create a linear wave pattern. For a three-dimensional wave such as a sound wave, the analogous patterns would be spherical waves (visualize concentric spheres) and plane waves (visualize a series of pieces of paper, each separated from the next by the same gap).

Infinitely many patterns are possible, but linear or plane waves are often the simplest to analyze, because the velocity vector is in the same direction no matter what part of the wave we look at. Since all the velocity vectors are parallel to one another, the problem is effectively one-dimensional. Throughout this chapter and the next, we will restrict ourselves mainly to wave motion in one dimension, while not hesitating to broaden our horizons when it can be done without too much complication.

Discussion Questions

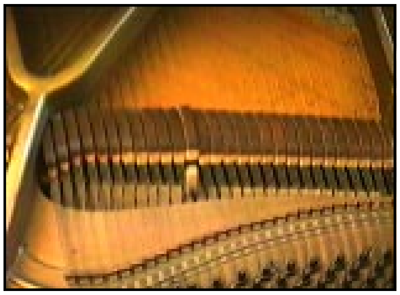
A. [see above]

B. Sketch two positive wave pulses on a string that are overlapping but not right on top of each other, and draw their superposition. Do the same for a positive pulse running into a negative pulse.

C. A traveling wave pulse is moving to the right on a string. Sketch the velocity vectors of the various parts of the string. Now do the same for a pulse moving to the left.

D. In a spherical sound wave spreading out from a point, how would the energy of the wave fall off with distance?

3.2 Waves on a String

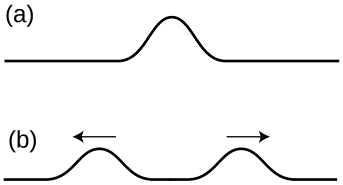


Hitting a key on a piano causes a hammer to come up from underneath and hit a string (actually a set of three). The result is a pair of pulses moving away from the point of impact.

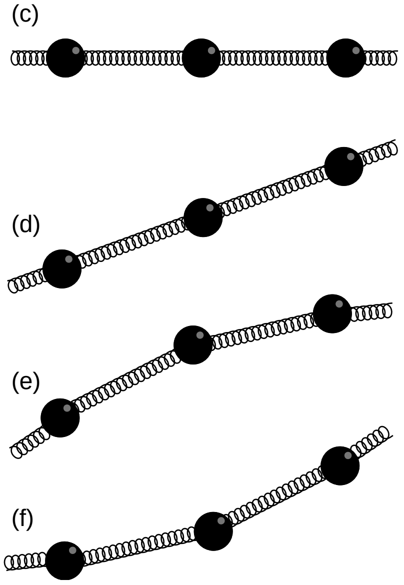
So far you have learned some counterintuitive things about the behavior of waves, but intuition can be trained. The first half of this section aims to build your intuition by investigating a simple, one-dimensional type of wave: a wave on a string. If you have ever stretched a string between the bottoms of two open-mouthed cans to talk to a friend, you were putting this type of wave to work. Stringed instruments are another good example. Although we usually think of a piano wire simply as vibrating, the hammer actually strikes it quickly and makes a dent in it, which then ripples out in both directions. Since this chapter is about free waves, not bounded ones, we pretend that our string is infinitely long.

After the qualitative discussion, we will use simple approximations to investigate the speed of a wave pulse on a string. This quick and dirty treatment is then followed by a rigorous attack using the methods of calculus, which may be skipped by the student who has not studied calculus. How far you penetrate in this section is up to you, and depends on your mathematical self-confidence. If you skip the later parts and proceed to the next section, you should nevertheless be aware of the important result that the speed at which a pulse moves does not depend on the size or shape of the pulse. This is a fact that is true for many other types of waves.

Intuitive ideas



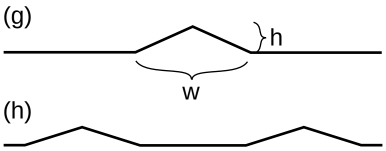
Consider a string that has been struck, (a), resulting in the creation of two wave pulses, (b), one traveling to the left and one to the right. This is analogous to the way ripples spread out in all directions from a splash in water, but on a one-dimensional string, “all directions” becomes “both directions.”



We can gain insight by modeling the string as a series of masses connected by springs. (In the actual string the mass and the springiness are both contributed by the molecules themselves.) If we look at various microscopic portions of the string, there will be some areas that are flat, (c), some that are sloping but not curved, (d), and some that are curved, (e) and (f). In example (c) it is clear that both the forces on the central mass cancel out, so it will not accelerate. The same is true of (d), however. Only in curved regions such as (e) and (f) is an acceleration produced. In these examples, the vector sum of the two forces acting on the central mass is not zero. The important concept is that *curvature makes force:* the curved areas of a wave tend to experience forces resulting in an acceleration toward the mouth of the curve. Note, however, that an uncurved portion of the string need not remain motionless. It may move at constant velocity to either side.

Approximate treatment

We now carry out an approximate treatment of the speed at which two pulses will spread out from an initial indentation on a string. For simplicity, we imagine a hammer blow that creates a triangular dent, (g). We will estimate the amount of time, t, required until each of the pulses has traveled a distance equal to the width of the pulse itself. The velocity of the pulses is then $\pm w/t$.



As always, the velocity of a wave depends on the properties of the medium, in this case the string. The properties of the string can be summarized by two variables: the tension, T, and the mass per unit length, μ (Greek letter mu).

If we consider the part of the string encompassed by the initial dent as a single object, then this object has a mass of approximately μw (mass/length x length=mass). (Here, and throughout the derivation, we assume that h is much less than w, so that we can ignore the fact that this segment of the string has a length slightly greater than w.) Although the downward acceleration of this segment of the string will be neither constant over time nor uniform across the string, we will pretend that it is constant for the sake of our simple estimate. Roughly speaking, the time interval between (g) and (h) is the amount of time required for the initial dent to accelerate from rest and reach its normal, flattened position. Of course the tip of the triangle has a longer distance to travel than the edges, but again we ignore the complications and simply assume that the segment as a whole must travel a distance h. Indeed, it might seem surprising that the triangle would so neatly spring back to a perfectly flat shape. It is an experimental fact that it does, but our analysis is too crude to address such details.

The string is kinked, i.e. tightly curved, at the edges of the triangle, so it is here that there will be large forces that do not cancel out to zero. There are two forces acting on the triangular hump, one of magnitude T acting down and to the right, and one of the same magnitude acting down and to the left. If the angle of the sloping sides is θ, then the total force on the segment equals $2T\sin\theta$. Dividing the triangle into two right triangles, we see that $\sin\theta$ equals h divided by the length of one of the sloping sides. Since h is much less than w, the length of the sloping side is essentially the same as $w/2$, so we have $\sin\theta = 2h/w$, and $F=4Th/w$. The acceleration of the segment (actually the acceleration of its center of mass) is

$$
\begin{aligned}
a &= F/m \\
&= 4Th/\mu w^2 \quad .
\end{aligned}
$$

The time required to move a distance h under constant acceleration a is found by solving $h=\frac{1}{2}at^2$ to yield

$$
\begin{aligned}
t &= \sqrt{2h\ /\ a} \\
&= w\sqrt{\frac{\mu}{2T}} \quad .
\end{aligned}
$$

Our final result for the velocity of the pulses is

$$
\begin{aligned}
|v| &= w/t \\
&= \sqrt{\frac{2T}{\mu}} \quad .
\end{aligned}
$$

The velocity of a wave on a string does not depend on the shape of the wave. The same is true for many other types of waves.

The remarkable feature of this result is that the velocity of the pulses does not depend at all on w or h, i.e. any triangular pulse has the same speed. It is an experimental fact (and we will also prove rigorously in the following subsection) that any pulse of any kind, triangular or otherwise, travels along the string at the same speed. Of course, after so many approximations we cannot expect to have gotten all the numerical factors right. The correct result for the velocity of the pulses is

$$
v = \sqrt{\frac{T}{\mu}} \quad .
$$

The importance of the above derivation lies in the insight it brings —that all pulses move with the same speed — rather than in the details of the numerical result. The reason for our too-high value for the velocity is not hard to guess. It comes from the assumption that the acceleration was constant, when actually the total force on the segment would diminish as it flattened out.

Rigorous derivation using calculus (optional)

After expending considerable effort for an approximate solution, we now display the power of calculus with a rigorous and completely general treatment that is nevertheless much shorter and easier. Let the flat position of the string define the x axis, so that y measures how far a point on the string is from equilibrium. The motion of the string is characterized by $y(x,t)$, a function of two variables. Knowing that the force on any small segment of string depends on the curvature of the string in that area, and that the second derivative is a measure of curvature, it is not surprising to find that the infinitesimal force dF acting on an infinitesimal segment dx is given by

$$
dF = T\frac{\mathrm{d}^2y}{\mathrm{d}x^2}\mathrm{d}x \quad .
$$

proof of the principle of superposition, in the case of waves on a string

(This can be proven by vector addition of the two infinitesimal forces acting on either side.) The acceleration is then $a = \mathrm{d}F/\mathrm{d}m$, or, substituting $\mathrm{d}m=\mu\mathrm{d}x$,

$$\frac{\mathrm{d}^2 y}{\mathrm{d}t^2} = \frac{T}{\mu}\frac{\mathrm{d}^2 y}{\mathrm{d}x^2} \quad .$$

The second derivative with respect to time is related to the second derivative with respect to position. This is no more than a fancy mathematical statement of the intuitive fact developed above, that the string accelerates so as to flatten out its curves.

proof that all wave shapes travel at the same speed, in the case of waves on a string

Before even bothering to look for solutions to this equation, we note that it already proves the principle of superposition, because the derivative of a sum is the sum of the derivatives. Therefore the sum of any two solutions will also be a solution.

Based on experiment, we expect that this equation will be satisfied by any function $y(x,t)$ that describes a pulse or wave pattern moving to the left or right at the correct speed v. In general, such a function will be of the form $y=f(x-vt)$ or $y=f(x+vt)$, where f is any function of one variable. Because of the chain rule, each derivative with respect to time brings out a factor of $\pm v$. Evaluating the second derivatives on both sides of the equation gives

$$\left(\pm v\right)^2 f'' = \frac{T}{\mu}f'' \quad .$$

Squaring gets rid of the sign, and we find that we have a valid solution for any function f, provided that v is given by

$$v = \sqrt{\frac{T}{\mu}} \quad .$$

3.3 Sound and Light Waves

Sound waves

The phenomenon of sound is easily found to have all the characteristics we expect from a wave phenomenon:

- Sound waves obey superposition. Sounds do not knock other sounds out of the way when they collide, and we can hear more than one sound at once if they both reach our ear simultaneously.
- The medium does not move with the sound. Even standing in front of a titanic speaker playing earsplitting music, we do not feel the slightest breeze.
- The velocity of sound depends on the medium. Sound travels faster in helium than in air, and faster in water than in helium. Putting more energy into the wave makes it more intense, not faster. For example, you can easily detect an echo when you clap your hands a short distance from a large, flat wall, and the delay of the echo is no shorter for a louder clap.

Although not all waves have a speed that is independent of the shape of the wave, and this property therefore is irrelevant to our collection of evidence that sound is a wave phenomenon, sound does nevertheless have this property. For instance, the music in a large concert hall or stadium may take on the order of a second to reach someone seated in the nosebleed section, but we do not notice or care, because the delay is the same for every sound. Bass, drums, and vocals all head outward from the stage at 340 m/s, regardless of their differing wave shapes.

If sound has all the properties we expect from a wave, then what type of wave is it? It is a series of compressions and expansions of the air. Even for a very loud sound, the increase or decrease compared to normal atmospheric pressure is no more than a part per million, so our ears are apparently very sensitive instruments. In a vacuum, there is no medium for the sound waves, and so they cannot exist. The roars and whooshes of space ships in Hollywood movies are fun, but scientifically wrong.

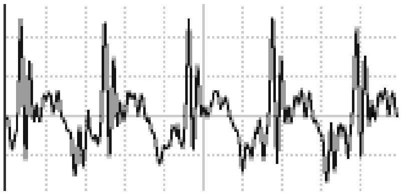
Light waves

Entirely similar observations lead us to believe that light is a wave, although the concept of light as a wave had a long and tortuous history. It is interesting to note that Isaac Newton very influentially advocated a contrary idea about light. The belief that matter was made of atoms was stylish at the time among radical thinkers (although there was no experimental evidence for their existence), and it seemed logical to Newton that light as well should be made of tiny particles, which he called corpuscles (Latin for "small objects"). Newton's triumphs in the science of mechanics, i.e. the study of matter, brought him such great prestige that nobody bothered to question his incorrect theory of light for 150 years. One persuasive proof that light is a wave is that according to Newton's theory, two intersecting beams of light should experience at least some disruption because of collisions between their corpuscles. Even if the corpuscles were extremely small, and collisions therefore very infrequent, at least some dimming should have been measurable. In fact, very delicate experiments have shown that there is no dimming.

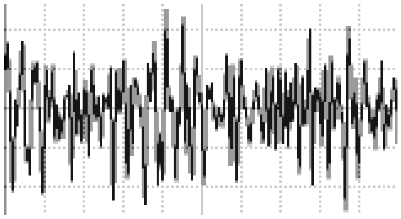
The wave theory of light was entirely successful up until the 20th century, when it was discovered that not all the phenomena of light could be explained with a pure wave theory. It is now believed that both light and matter are made out of tiny chunks which have *both* wave and particle properties. For now, we will content ourselves with the wave theory of light, which is capable of explaining a great many things, from cameras to rainbows.

If light is a wave, what is waving? What is the medium that wiggles when a light wave goes by? It isn't air. A vacuum is impenetrable to sound, but light from the stars travels happily through zillions of miles of empty space. Light bulbs have no air inside them, but that doesn't prevent the light waves from leaving the filament. For a long time, physicists assumed that there must be a mysterious medium for light waves, and they called it the ether (not to be confused with the chemical). Supposedly the ether existed everywhere in space, and was immune to vacuum pumps. The details of the story are more fittingly reserved for later in this course, but the end result was that a long series of experiments failed to detect any evidence for the ether, and it is no longer believed to exist. Instead, light can be explained as a wave pattern made up of electrical and magnetic fields.

3.4 Periodic Waves



(a) A graph of pressure versus time for a periodic sound wave, the vowel “ah.”



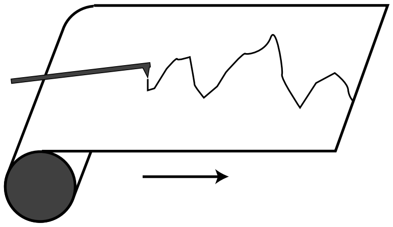
(b) A similar graph for a nonperiodic wave, “sh.”

Period and frequency of a periodic wave

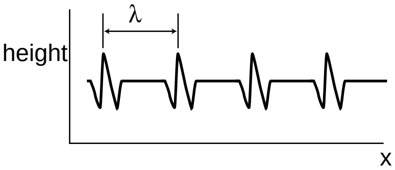
You choose a radio station by selecting a certain frequency. We have already defined period and frequency for vibrations, but what do they signify in the case of a wave? We can recycle our previous definition simply by stating it in terms of the vibrations that the wave causes as it passes a receiving instrument at a certain point in space. For a sound wave, this receiver could be an eardrum or a microphone. If the vibrations of the eardrum repeat themselves over and over, i.e. are periodic, then we describe the sound wave that caused them as periodic. Likewise we can define the period and frequency of a wave in terms of the period and frequency of the vibrations it causes. As another example, a periodic water wave would be one that caused a rubber duck to bob in a periodic manner as they passed by it.

The period of a sound wave correlates with our sensory impression of musical pitch. A high frequency (short period) is a high note. The sounds that really define the musical notes of a song are only the ones that are periodic. It is not possible to sing a nonperiodic sound like “sh” with a definite pitch.

The frequency of a light wave corresponds to color. Violet is the high-frequency end of the rainbow, red the low-frequency end. A color like brown that does not occur in a rainbow is not a periodic light wave. Many phenomena that we do not normally think of as light are actually just forms of light that are invisible because they fall outside the range of frequencies our eyes can detect. Beyond the red end of the visible rainbow, there are infrared and radio waves. Past the violet end, we have ultraviolet, x-rays, and gamma rays.



A strip chart recorder.



A water wave profile created by a series of repeating pulses.

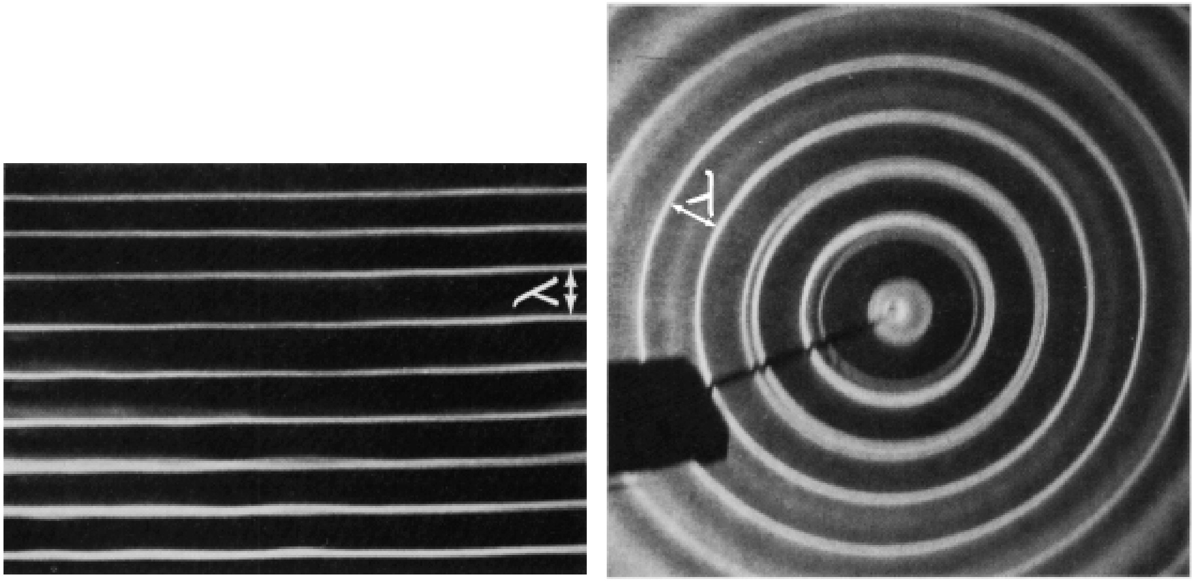
Graphs of waves as a function of position

Some waves, light sound waves, are easy to study by placing a detector at a certain location in space and studying the motion as a function of time. The result is a graph whose horizontal axis is time. With a water wave, on the other hand, it is simpler just to look at the wave directly. This visual snapshot amounts to a graph of the height of the water wave as a function of *position*. Any wave can be represented in either way.

An easy way to visualize this is in terms of a strip chart recorder, an obsolescing device consisting of a pen that wiggles back and forth as a roll of paper is fed under it. It can be used to record a person's electrocardiogram, or seismic waves too small to be felt as a noticeable earthquake but detectable by a seismometer. Taking the seismometer as an example, the chart is essentially a record of the ground's wave motion as a function of time, but if the paper was set to feed at the same velocity as the motion of an earthquake wave, it would also be a full-scale representation of the profile of the actual wave pattern itself. Assuming, as is usually the case, that the wave velocity is a constant number regardless of the wave's shape, knowing the wave motion as a function of time is equivalent to knowing it as a function of position.

Wavelength

Any wave that is periodic will also display a repeating pattern when graphed as a function of position. The distance spanned by one repetition is referred to as one *wavelength*. The usual notation for wavelength is λ, the Greek letter lambda. Wavelength is to space as period is to time.



Wavelengths of linear and circular waves. Uncopyrighted photos from PSSC Physics.

Wave velocity related to frequency and wavelength

Suppose that we create a repetitive disturbance by kicking the surface of a swimming pool. We are essentially making a series of wave pulses. The wavelength is simply the distance a pulse is able to travel before we make the next pulse. The distance between pulses is λ, and the time between pulses is the period, T, so the speed of the wave is the distance divided by the time,

$$v = \lambda/T \quad .$$

This important and useful relationship is more commonly written in terms of the frequency,

$$v = f\lambda \quad .$$

A note on dispersive waves
The discussion of wave velocity given here is actually a little bit of an oversimplification for a wave whose velocity depends on its frequency and wavelength. Such a wave is called a dispersive wave. Nearly all the waves we deal with in this course are nondispersive, but the issue becomes important in book 6 of this series, where it is discussed in detail in optional section 5.2.

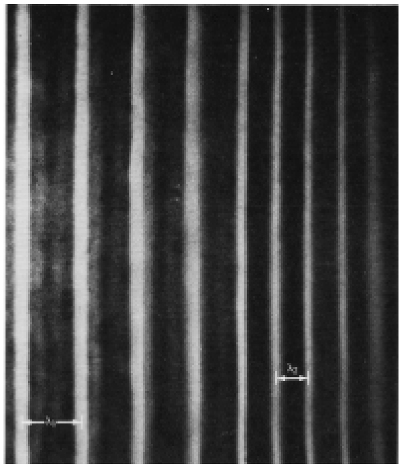
Example: Wavelength of radio waves
Question: The speed of light is $3.0\text{x}10^8$ m/s. What is the wavelength of the radio waves emitted by KLON, a station whose frequency is 88.1 MHz?
Solution: Solving for wavelength, we have

$$\begin{aligned} \lambda &= v/f \\ &= (3.0\text{x}10^8\ \text{m/s})/(88.1\text{x}10^6\ \text{s}^{-1}) \\ &= 3.4\ \text{m} \end{aligned}$$

The size of a radio antenna is closely related to the wavelength of the waves it is intended to receive. The match need not be exact (since after all one antenna can receive more than one wavelength!), but the ordinary “whip” antenna such as a car’s is 1/4 of a wavelength. An antenna optimized to receive KLON’s signal (which is the only one my car radio is ever tuned to) would have a length of 3.4 m/4 = 0.85 m.

Ultrasound, i.e. sound with frequencies higher than the range of human hearing, was used to make this image of a fetus. The resolution of the image is related to the wavelength, since details smaller than about one wavelength cannot be resolved. High resolution therefore requires a short wavelength, corresponding to a high frequency.

The equation $v=f\lambda$ defines a fixed relationship between any two of the variables if the other is held fixed. The speed of radio waves in air is almost exactly the same for all wavelengths and frequencies (it is exactly the same if they are in a vacuum), so there is a fixed relationship between their frequency and wavelength. Thus we can say either "Are we on the same wavelength?" or "Are we on the same frequency?"



A water wave traveling into a region with different depth will change its wavelength.

A different example is the behavior of a wave that travels from a region where the medium has one set of properties to an area where the medium behaves differently. The frequency is now fixed, because otherwise the two portions of the wave would otherwise get out of step, causing a kink or discontinuity at the boundary, which would be unphysical. (A more careful argument is that a kink or discontinuity would have infinite curvature, and waves tend to flatten out their curvature. An infinite curvature would flatten out infinitely fast, i.e. it could never occur in the first place.) Since the frequency must stay the same, any change in the velocity that results from the new medium must cause a change in wavelength.

The velocity of water waves depends on the depth of the water, so based on $\lambda=v/f$, we see that water waves that move into a region of different depth must change their wavelength, as shown in the figure on the left. This effect can be observed when ocean waves come up to the shore. If the deceleration of the wave pattern is sudden enough, the tip of the wave can curl over, resulting in a breaking wave.

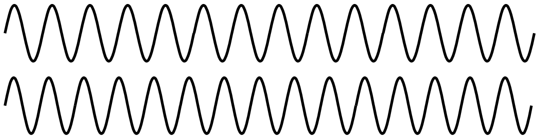
Sinusoidal waves

Sinusoidal waves are the most important special case of periodic waves. In fact, many scientists and engineers would be uncomfortable with defining a waveform like the "ah" vowel sound as having a definite frequency and wavelength, because they consider only sine waves to be pure examples of a certain frequency and wavelengths. Their bias is not unreasonable, since the French mathematician Fourier showed that any periodic wave with frequency f can be constructed as a superposition of sine waves with frequencies f, $2f$, $3f$, ... In this sense, sine waves are the basic, pure building blocks of all waves. (Fourier's result so surprised the mathematical community of France that he was ridiculed the first time he publicly presented his theorem.)

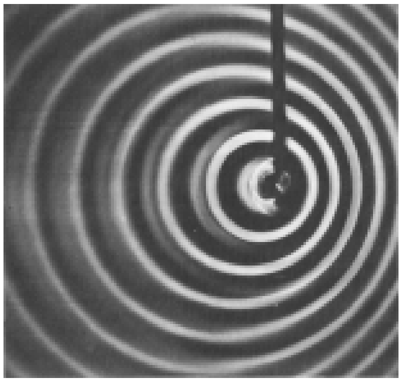
However, what definition to use is really a matter of convenience. Our sense of hearing perceives any two sounds having the same period as possessing the same pitch, regardless of whether they are sine waves or not. This is undoubtedly because our ear-brain system evolved to be able to interpret human speech and animal noises, which are periodic but not sinusoidal. Our eyes, on the other hand, judge a color as pure (belonging to the rainbow set of colors) only if it is a sine wave.

Discussion Question

Suppose we superimpose two sine waves with equal amplitudes but slightly different frequencies, as shown in the figure. What will the superposition look like? What would this sound like if they were sound waves?



3.5 The Doppler Effect



The pattern of waves made by a point source moving to the right across the water. Note the shorter wavelength of the forward-emitted waves and the longer wavelength of the backward-going ones.

The figure shows the wave pattern made by the tip of a vibrating rod which is moving across the water. If the rod had been vibrating in one place, we would have seen the familiar pattern of concentric circles, all centered on the same point. But since the source of the waves is moving, the wavelength is shortened on one side and lengthened on the other. This is known as the *Doppler effect*.

Note that the velocity of the waves is a fixed property of the medium, so for example the forward-going waves do not get an extra boost in speed as would a material object like a bullet being shot forward from an airplane.

We can also infer a change in frequency. Since the velocity is constant, the equation $v=f\lambda$ tells us that the change in wavelength must be matched by an opposite change in frequency: higher frequency for the waves emitted forward, and lower for the ones emitted backward. The frequency Doppler effect is the reason for the familiar dropping-pitch sound of a race car going by. As the car approaches us, we hear a higher pitch, but after it passes us we hear a frequency that is lower than normal.

The Doppler effect will also occur if the observer is moving but the source is stationary. For instance, an observer moving toward a stationary source will perceive one crest of the wave, and will then be surrounded by the next crest sooner than she otherwise would have, because she has moved toward it and hastened her encounter with it. Roughly speaking, the Doppler effect depends only the relative motion of the source and the observer, not on their absolute state of motion (which is not a well-defined notion in physics) or on their velocity relative to the medium.

Restricting ourselves to the case of a moving source, and to waves emitted either directly along or directly against the direction of motion, we can easily calculate the wavelength, or equivalently the frequency, of the Doppler-shifted waves. Let v be the velocity of the waves, and v_s the velocity of the source. The wavelength of the forward-emitted waves is shortened by an amount v_sT equal to the distance traveled by the source over the course of one period. Using the definition $f=1/T$ and the equation $v=f\lambda$, we find for the wavelength λ' of the Doppler-shifted wave the equation

$$\lambda' = \left(1-\frac{v_s}{v}\right)\lambda \quad .$$

A similar equation can be used for the backward-emitted waves, but with a plus sign rather than a minus sign.

Example: Doppler-shifted sound from a race car

Question: If a race car moves at a velocity of 50 m/s, and the velocity of sound is 340 m/s, by what percentage are the wavelength and frequency of its sound waves shifted for an observer lying along its line of motion?

Solution: For an observer whom the car is approaching, we find

$$1-\frac{v_s}{v} \quad = \quad 0.85 \quad ,$$

so the shift in wavelength is 15%. Since the frequency is inversely proportional to the wavelength for a fixed value of the speed of sound, the frequency is shifted upward by

$$1/0.85 \quad = \quad 1.18 \quad ,$$

i.e. a change of 18%. (For velocities that are small compared to the wave velocities, the Doppler shifts of the wavelength and frequency are about the same.)

Example: Doppler shift of the light emitted by a race car

Question: What is the percent shift in the wavelength of the light waves emitted by a race car's headlights?

Solution: Looking up the speed of light in the front of the book, v=3.0x10^8 m/s, we find

$$1-\frac{v_s}{v} \quad = \quad 0.99999983 \quad ,$$

i.e. the percentage shift is only 0.000017%.

The second example shows that under ordinary earthbound circumstances, Doppler shifts of light are negligible because ordinary things go so much slower than the speed of light. It's a different story, however, when it comes to stars and galaxies, and this leads us to a story that has profound implications for our understanding of the origin of the universe.

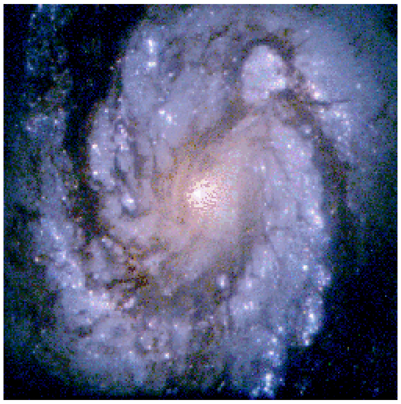
Optional Topic: A Note on Doppler Shifts of Light

If Doppler shifts depend only on the relative motion of two the source and receiver, then there is no way for a person moving with the source and another person moving with the receiver to determine who is moving and who isn't. Either can blame the Doppler shift entirely on the other's motion and claim to be at rest herself. This is entirely in agreement with the principle stated originally by Galileo that all motion is relative.

On the other hand, a careful analysis of the Doppler shifts of water or sound waves shows that it is only approximately true, at low speeds, that the shifts just depend on the relative motion of the source and observer. For instance, it is possible for a jet plane to keep up with its own sound waves, so that the sound waves appear to stand still to the pilot of the plane. The pilot then knows she is moving at exactly the speed of sound. The reason this doesn't disprove the relativity of motion is that the pilot is not really determining her absolute motion but rather her motion *relative to the air*, which is the medium of the sound waves.

Einstein realized that this solved the problem for sound or water waves, but would not salvage the principle of relative motion in the case of light waves, since light is not a vibration of any physical medium such as water or air. Beginning by imagining what a beam of light would look like to a person riding a motorcycle alongside it, Einstein eventually came up with a radical new way of describing the universe, in which space and time are distorted as measured by observers in different states of motion. As a consequence of this Theory of Relativity, he showed that light waves would have Doppler shifts that would exactly, not just approximately, depend only on the relative motion of the source and receiver.

The Big Bang



The galaxy M100 in the constellation Coma Berenices. Under higher magnification, the milky clouds reveal themselves to be composed of trillions of stars.

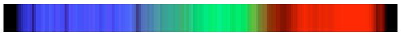
As soon as astronomers began looking at the sky through telescope, they began noticing certain objects that looked like clouds in deep space. The fact that they looked the same night after night meant that they were beyond the earth's atmosphere. Not knowing what they really were, but wanting to sound official, they called them "nebulae," a Latin word meaning "clouds" but sounding more impressive. In the early 20th century, astronomers realized that although some really were clouds of gas (e.g. the middle "star" of Orion's sword, which is visibly fuzzy even to the naked eye when conditions are good), others were what we now call galaxies: virtual island universes consisting of trillions of stars (for example the Andromeda Galaxy, which is visible as a fuzzy patch through binoculars). Three hundred years after Galileo had resolved the Milky Way into individual stars through his telescope, astronomers realized that the universe is made of galaxies of stars, and the Milky Way is simply the visible part of the flat disk of our own galaxy, seen from inside.

This opened up the scientific study of cosmology, the structure and history of the universe as a whole, a field that had not been seriously attacked since the days of Newton. Newton had realized that if gravity was always attractive, never repulsive, the universe would have a tendency to collapse. His solution to the problem was to posit a universe that was infinite and uniformly populated with matter, so that it would have no geometrical center. The gravitational forces in such a universe would always tend to cancel out by symmetry, so there would be no collapse. By the 20th century, the belief in an unchanging and infinite universe had become conventional wisdom in science, partly as a reaction against the time that had been wasted trying to find explanations of ancient geological phenomena based on catastrophes suggested by biblical events like Noah's flood.

Edwin Hubble

In the 1920's astronomer Edwin Hubble began studying the Doppler shifts of the light emitted by galaxies. A former college football player with a serious nicotine addiction, Hubble did not set out to change our image of the beginning of the universe. His autobiography seldom even mentions the cosmological discovery for which he is now remembered. When astronomers began to study the Doppler shifts of galaxies, they expected that each galaxy's direction and velocity of motion would be essentially random. Some would be approaching us, and their light would therefore be Doppler-shifted to the blue end of the spectrum, while an equal number would be expected to have red shifts. What Hubble discovered instead was that except for a few very nearby ones, all the galaxies had red shifts, indicating that they were receding from us at a hefty fraction of the speed of light. Not only that, but the ones farther away were receding more quickly. The speeds were directly proportional to their distance from us.

Did this mean that the earth (or at least our galaxy) was the center of the universe? No, because Doppler shifts of light only depend on the relative motion of the source and the observer. If we see a distant galaxy moving away from us at 10% of the speed of light, we can be assured that the astronomers who live in that galaxy will see ours receding from them at the same speed in the opposite direction. The whole universe can be envisioned as a rising loaf of raisin bread. As the bread expands, there is more and more space between the raisins. The farther apart two raisins are, the greater the speed with which they move apart.



How do astronomers know what mixture of wavelengths a star emitted originally, so that they can tell how much the Doppler shift was? This image (obtained by the author with equipment costing about $5, and no telescope) shows the mixture of colors emitted by the star Sirius. (If you have the book in black and white, blue is on the left and red on the right.) The star appears white or bluish-white to the eye, but any light looks white if it contains roughly an equal mixture of the rainbow colors, i.e. of all the pure sinusoidal waves with wavelengths lying in the visible range. Note the black "gap teeth." These are the fingerprint of hydrogen in the outer atmosphere of Sirius. These wavelengths are selectively absorbed by hydrogen. Sirius is in our own galaxy, but similar stars in other galaxies would have the whole pattern shifted toward the red end, indicating they are moving away from us.

The universe's expansion is presumably decelerating because of gravitational attraction among the galaxies. We do not presently know whether there is enough mass in the universe to cause enough attraction to halt the expansion eventually. But perhaps more interesting than the distant future of the universe is what its present expansion implies about its past. Extrapolating backward in time using the known laws of physics, the universe must have been denser and denser at earlier and earlier times. At some point, it must have been extremely dense and hot, and we can even detect the radiation from this early fireball, in the form of microwave radiation that permeates space. The phrase Big Bang was originally coined by the doubters of the theory to make it sound ridiculous, but it stuck, and today essentially all astronomers accept the Big Bang theory based on the very direct evidence of the red shifts and the cosmic microwave background radiation.

What the Big Bang is not

Finally it should be noted what the Big Bang theory is not. It is not an explanation of *why* the universe exists. Such questions belong to the realm of religion, not science. Science can find ever simpler and ever more fundamental explanations for a variety of phenomena, but ultimately science takes the universe as it is according to observations.

Furthermore, there is an unfortunate tendency, even among many scientists, to speak of the Big Bang theory was a description of the very first event in the universe, which caused everything after it. Although it is true that time may have had a beginning (Einstein's theory of general relativity admits such a possibility), the methods of science can only work within a certain range of conditions such as temperature and density. Beyond a temperature of about 10^9 degrees C, the random thermal motion of subatomic particles becomes so rapid that its velocity is comparable to the speed of light. Early enough in the history of the universe, when these temperatures existed, Newtonian physics becomes less accurate, and we must describe nature using the more general description given by Einstein's theory of relativity, which encompasses Newtonian physics as a special case. At even higher temperatures, beyond about 10^{33} degrees, physicists know that Einstein's theory as well begins to fall apart, but we don't know how to construct the even more general theory of nature that would work at those temperatures. No matter how far physics progresses, we will never be able to describe nature at infinitely high temperatures, since there is a limit to the temperatures we can explore by experiment and observation in order to guide us to the right theory. We are confident that we understand the basic physics involved in the evolution of the universe starting a few minutes after the Big Bang, and we may be able to push back to milliseconds or microseconds after it, but we cannot use the methods of science to deal with the beginning of time itself.

Discussion Questions

A. If an airplane travels at exactly the speed of sound, what would be the wavelength of the forward-emitted part of the sound waves it emitted? How should this be interpreted, and what would actually happen?
B. If bullets go slower than the speed of sound, why can a supersonic fighter plane catch up to its own sound, but not to its own bullets?
C. If someone inside a plane is talking to you, should their speech be Doppler shifted?

Summary

Selected Vocabulary

superposition	the adding together of waves that overlap with each other
medium	a physical substance whose vibrations constitute a wave
wavelength	the distance in space between repetitions of a periodic wave
Doppler effect	the change in a wave's frequency and wavelength due to the motion of the source or the observer or both

Notation

λ	wavelength (Greek letter lambda)

Summary

Wave motion differs in three important ways from the motion of material objects:

(1) Waves obey the principle of superposition. When two waves collide, they simply add together.

(2) The medium is not transported along with the wave. The motion of any given point in the medium is a vibration about its equilibrium location, not a steady forward motion.

(3) The velocity of a wave depends on the medium, not on the amount of energy in the wave. (For some types of waves, notably water waves, the velocity may also depend on the shape of the wave.)

Sound waves consist of increases and decreases (typically very small ones) in the density of the air. Light is a wave, but it is a vibration of electric and magnetic fields, not of any physical medium. Light can travel through a vacuum.

A periodic wave is one that creates a periodic motion in a receiver as it passes it. Such a wave has a well-defined period and frequency, and it will also have a wavelength, which is the distance in space between repetitions of the wave pattern. The velocity, frequency, and wavelength of a periodic wave are related by the equation

$$v = f\lambda \quad .$$

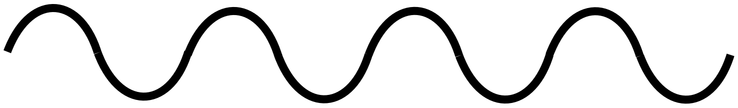
A wave emitted by a moving source will be shifted in wavelength and frequency. The shifted wavelength is given by the equation

$$\lambda' = \left(1 - \frac{v_s}{v}\right)\lambda \quad ,$$

where v is the velocity of the waves and v_s is the velocity of the source, taken to be positive or negative so as to produce a Doppler-lengthened wavelength if the source is receding and a Doppler-shortened one if it approaches. A similar shift occurs if the observer is moving, and in general the Doppler shift depends approximately only on the relative motion of the source and observer if their velocities are both small compared to the waves' velocity. (This is not just approximately but exactly true for light waves, and this fact forms the basis of Einstein's Theory of Relativity.)

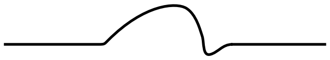
Homework Problems

1. The following is a graph of the height of a water wave as a function of *position*, at a certain moment in time.



Trace this graph onto another piece of paper, and then sketch below it the corresponding graphs that would be obtained if
(a) the amplitude and frequency were doubled while the velocity remained the same;
(b) the frequency and velocity were both doubled while the amplitude remained unchanged;
(c) the wavelength and amplitude were reduced by a factor of three while the velocity was doubled.

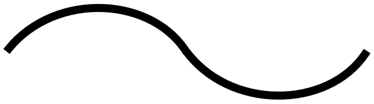
[Problem by Arnold Aarons.]



Problem 2.

2. (a) The graph shows the height of a water wave pulse as a function of position. Draw a graph of height as a function of time for a specific point on the water. Assume the pulse is traveling to the right.
(b) Repeat part a, but assume the pulse is traveling to the left.
(c) Now assume the original graph was of height as a function of time, and draw a graph of height as a function of position, assuming the pulse is traveling to the right.
(d) Repeat part c, but assume the pulse is traveling to the left.

[Problem by Arnold Aarons.]



Problem 3.

3. The figure shows one wavelength of a steady sinusoidal wave traveling to the right along a string. Define a coordinate system in which the positive *x* axis points to the right and the positive *y* axis up, such that the flattened string would have *y*=0. Copy the figure, and label with "*y*=0" all the appropriate parts of the string. Similarly, label with "*v*=0" all parts of the string whose instantaneous velocities are zero, and with "*a*=0" all parts whose instantaneous acceleration is zero. There is more than one point whose velocity is of the greatest magnitude. Pick one of these, and indicate the direction of its velocity vector. Do the same for a point having the maximum magnitude of acceleration.

[Problem by Arnold Aarons.]

4. Find an equation for the relationship between the Doppler-shifted frequency of a wave and the frequency of the original wave, for the case of a stationary observer and a source moving directly toward or away from the observer.

5. Suggest an experiment to look for any deviation from the principle of superposition for surface waves in water. You will receive more credit if your experiment is simple and practical.

S	A solution is given in the back of the book.	★	A difficult problem.
✓	A computerized answer check is available.	∫	A problem that requires calculus.

6✓. The musical note middle C has a frequency of 262 Hz. What are its period and wavelength?

7✓. Singing that is off-pitch by more than about 1% sounds bad. How fast would a singer have to be moving relative to a the rest of a band to make this much of a change in pitch due to the Doppler effect?

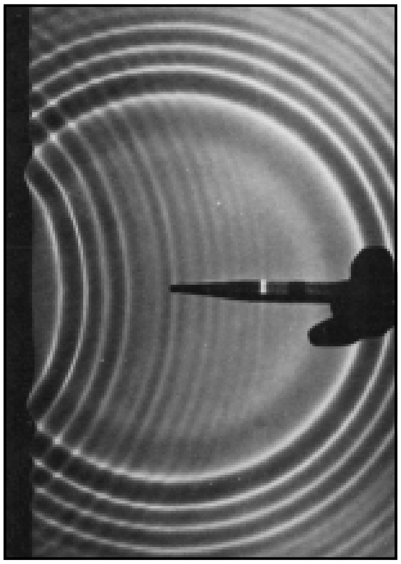
A cross-sectional view of a human body, showing the vocal tract.

4 Bounded Waves

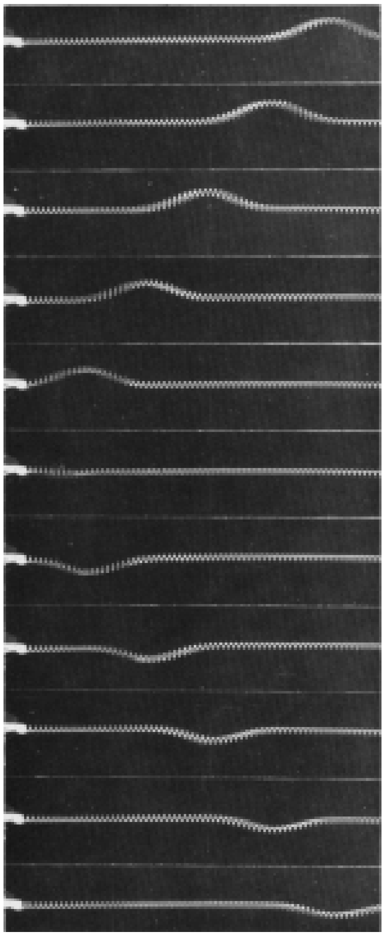
Speech is what separates humans most decisively from animals. No other species can master syntax, and even though chimpanzees can learn a vocabulary of hand signs, there is an unmistakable difference between a human infant and a baby chimp: starting from birth, the human experiments with the production of complex speech sounds.

Since speech sounds are instinctive for us, we seldom think about them consciously. How do we do control sound waves so skillfully? Mostly we do it by changing the shape of a connected set of hollow cavities in our chest, throat, and head. Somehow by moving the boundaries of this space in and out, we can produce all the vowel sounds. Up until now, we have been studying only those properties of waves that can be understood as if they existed in an infinite, open space with no boundaries. In this chapter we address what happens when a wave is confined within a certain space, or when a wave pattern encounters the boundary between two different media, such as when a light wave moving through air encounters a glass windowpane.

4.1 Reflection, Transmission, and Absorption



(a) Circular water waves are reflected from a boundary on the left.
Uncopyrighted photo from PSSC Physics.



(b) A wave on a coil spring, initially traveling to the left, is reflected from the fixed end.
Uncopyrighted photo from PSSC Physics.

Reflection and transmission

Sound waves can echo back from a cliff, and light waves are reflected from the surface of a pond. We use the word *reflection*, normally applied only to light waves in ordinary speech, to describe any such case of a wave rebounding from a barrier. Figure (a) shows a circular water wave being reflected from a straight wall. In this chapter, we will concentrate mainly on reflection of waves that move in one dimension, as in figure (b).

Wave reflection does not surprise us. After all, a material object such as a rubber ball would behave in the same way. But waves are not objects, and there are some surprises in store.

First, only part of the wave is usually reflected. Looking out through a window, we see light waves that passed through it, but a person standing outside would also be able to see her reflection in the glass. A light wave that strikes the glass is partly *reflected* and partly *transmitted* (passed) by the glass. The energy of the original wave is split between the two. This is different from the behavior of the rubber ball, which must go one way or the other, not both.

Second, consider what you see if you are swimming underwater and you look up at the surface. You see your own reflection. This is utterly counterintuitive, since we would expect the light waves to burst forth to freedom in the wide-open air. A material projectile shot up toward the surface would never rebound from the water-air boundary!

What is it about the difference between two media that causes waves to be partly reflected at the boundary between them? Is it their density? Their chemical composition? Ultimately all that matters is the speed of the wave in the two media. *A wave is partially reflected and partially transmitted at the boundary between media in which it has different speeds.* For example, the speed of light waves in window glass is about 30% less than in air, which explains why windows always make reflections. Figures (c) and (d) show examples of wave pulses being reflected at the boundary between two coil springs of different weights, in which the wave speed is different.

Reflections such as (a) and (b), where a wave encounters a massive fixed object, can usually be understood on the same basis as cases like (c) and (d) where two media meet. Example (b), for instance, is like a more extreme version of example (c). If the heavy coil spring in (c) was made heavier and heavier, it would end up acting like the fixed wall to which the light spring in (b) has been attached.

Self-Check

A. In figure (b), the reflected pulse is upside-down, but its depth is just as big as the original pulse's height. How does the energy of the reflected pulse compare with that of the original?

Example: Fish have internal ears.
Why don't fish have ear-holes? The speed of sound waves in a fish's body is not much different from their speed in water, so sound waves are not strongly reflected from a fish's skin. They pass right through its body, so fish can have internal ears.

Example: Whale songs traveling long distances
Sound waves travel at drastically different speeds through rock, water, and air. Whale songs are thus strongly reflected both at both the bottom and the surface. The sound waves can travel hundreds of miles, bouncing repeatedly between the bottom and the surface, and still be detectable. Sadly, noise pollution from ships has nearly shut down this cetacean version of the internet.

Example: Long-distance radio communication.
Radio communication can occur between stations on opposite sides of the planet. The mechanism is entirely similar to the one explained in the previous example, but the three media involved are the earth, the atmosphere, and the ionosphere.

Self-Check

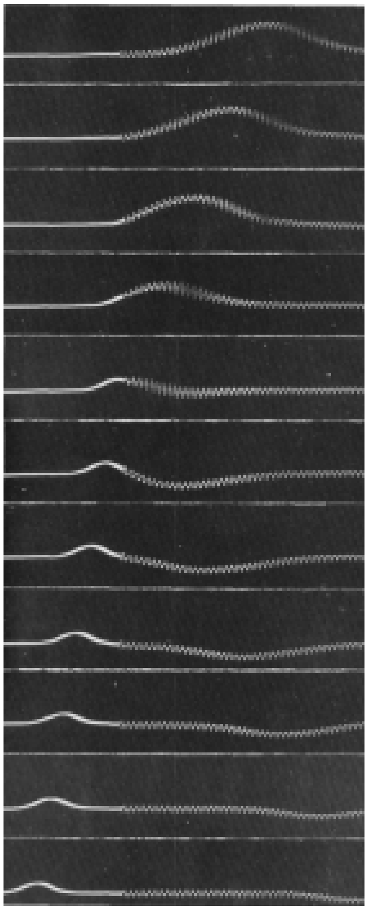
B. Sonar is a method for ships and submarines to detect each other by producing sound waves and listening for echoes. What properties would an underwater object have to have in order to be invisible to sonar?

A. A substance is invisible to sonar if the speed of sound waves in it is the same as in water. Reflections occur only at boundaries between media in which the wave speed is different. **B**. The energy of a wave is usually proportional to the square of the amplitude. Squaring a negative number gives a positive result, so the energy is the same.

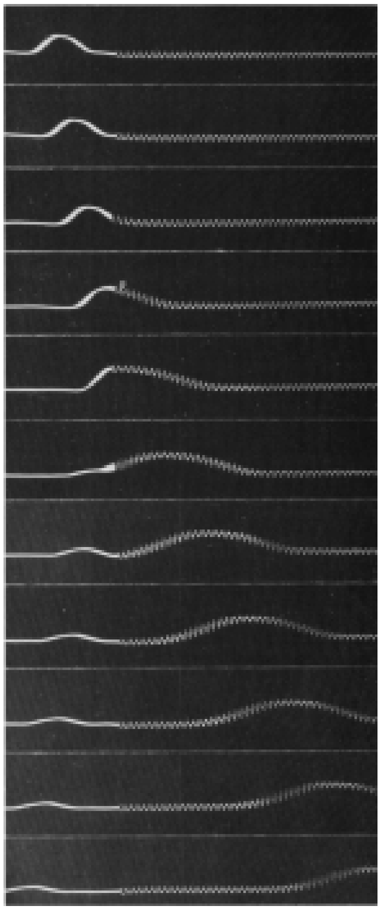
The use of the word “reflection” naturally brings to mind the creation of an image by a mirror, but this might be confusing, because we do not normally refer to “reflection” when we look at surfaces that are not shiny. Nevertheless, reflection is how we see the surfaces of all objects, not just polished ones. When we look at a sidewalk, for example, we are actually seeing the reflecting of the sun from the concrete. The reason we don’t see an image of the sun at our feet is simply that the rough surface blurs the image so drastically.

Inverted and uninverted reflections

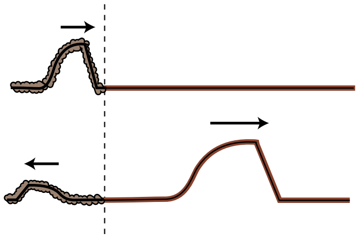
Notice how the pulse reflected back to the right in example (c) comes back upside-down, whereas the one reflected back to the left in (d) returns in its original upright form. This is true for other waves as well. In general, there are two possible types of reflections, a reflection back into a faster medium and a reflection back into a slower medium. One type will always be an inverting reflection and one noninverting.



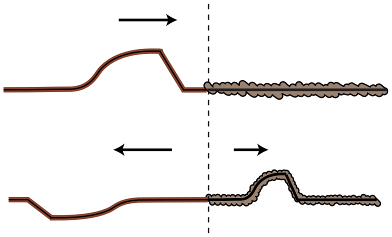
(c) A wave in the lighter spring, where the wave speed is greater, travels to the left and is then partly reflected and partly transmitted at the boundary with the heavier coil spring, which has a lower wave speed. The reflection is inverted. Uncopyrighted figure from PSSC Physics.



(d) A wave moving to the right in the heavier spring is partly reflected at the boundary with the lighter spring. The reflection is uninverted. Uncopyrighted figure from PSSC Physics.



(e) An uninverted reflection. The reflected pulse is reversed front to back, but is not upside-down.



(f) An inverted reflection. The reflected pulse is reversed both front to back and top to bottom.

It's important to realize that when we discuss inverted and uninverted reflections on a string, we are talking about whether the wave is flipped across the direction of motion (i.e. upside-down in these drawings). The reflected pulse will always be reversed front to back, as shown in figures (e) and (f). This is because it is traveling in the other direction. The leading edge of the pulse is what gets reflected first, so it is still ahead when it starts back to the left — it's just that "ahead" is now in the opposite direction.

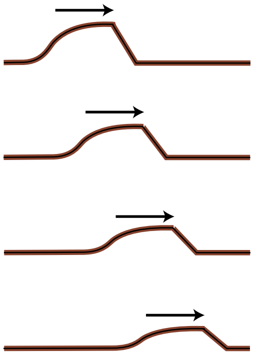
Absorption

So far we have tacitly assumed that wave energy remains as wave energy, and is not converted to any other form. If this was true, then the world would become more and more full of sound waves, which could never escape into the vacuum of outer space. In reality, any mechanical wave consists of a traveling pattern of vibrations of some physical medium, and vibrations of matter always produce heat, as when you bend a coathangar back and forth and it becomes hot. We can thus expect that in mechanical waves such as water waves, sound waves, or waves on a string, the wave energy will gradually be converted into heat. This is referred to as *absorption.*

The wave suffers a decrease in amplitude, as shown in figure (g). The decrease in amplitude amounts to the same fractional change for each unit of distance covered. For example, if a wave decreases from amplitude 2 to amplitude 1 over a distance of 1 meter, then after traveling another meter it will have an amplitude of 1/2. That is, the reduction in amplitude is exponential. This can be proven as follows. By the principle of superposition, we know that a wave of amplitude 2 must behave like the superposition of two identical waves of amplitude 1. If a single amplitude-1 wave would die down to amplitude 1/2 over a certain distance, then two amplitude-1 waves superposed on top of one another to make amplitude 1+1=2 must die down to amplitude 1/2+1/2=1 over the same distance.

Self-Check

As a wave undergoes absorption, it loses energy. Does this mean that it slows down?



(g) A pulse traveling through a highly absorptive medium.

In many cases, this frictional heating effect is quite weak. Sound waves in air, for instance, dissipate into heat extremely slowly, and the sound of church music in a cathedral may reverberate for as much as 3 or 4 seconds before it becomes inaudible. During this time it has traveled over a kilometer! Even this very gradual dissipation of energy occurs mostly as heating of the church's walls and by the leaking of sound to the outside (where it will eventually end up as heat). Under the right conditions (humid air and low frequency), a sound wave in a straight pipe could theoretically travel hundreds of kilometers before being noticeable attenuated.

In general, the absorption of mechanical waves depends a great deal on the chemical composition and microscopic structure of the medium. Ripples on the surface of antifreeze, for instance, die out extremely rapidly compared to ripples on water. For sound waves and surface waves in liquids and gases, what matters is the viscosity of the substance, i.e. whether it flows

No. A material object that loses kinetic energy slows down, but a wave is not a material object. The velocity of a wave ordinarily only depends on the medium, not on the amplitude. The speed of soft sound, for example, is the same as the speed of loud sound.

easily like water or mercury or more sluggishly like molasses or antifreeze. This explains why our intuitive expectation of strong absorption of sound in water is incorrect. Water is a very weak absorber of sound (viz. whale songs and sonar), and our incorrect intuition arises from focusing on the wrong property of the substance: water's high density, which is irrelevant, rather than its low viscosity, which is what matters.

Light is an interesting case, since although it can travel through matter, it is not itself a vibration of any material substance. Thus we can look at the star Sirius, 10^{14} km away from us, and be assured that none of its light was absorbed in the vacuum of outer space during its 9-year journey to us. The Hubble Space Telescope routinely observes light that has been on its way to us since the early history of the universe, billions of years ago. Of course the energy of light can be dissipated if it does pass through matter (and the light from distant galaxies is often absorbed if there happen to be clouds of gas or dust in between).

Example: soundproofing

Typical amateur musicians setting out to soundproof their garages tend to think that they should simply cover the walls with the densest possible substance. In fact, sound is not absorbed very strongly even by passing through several inches of wood. A better strategy for soundproofing is to create a sandwich of alternating layers of materials in which the speed of sound is very different, to encourage reflection.

The classic design is alternating layers of fiberglass and plywood. The speed of sound in plywood is very high, due to its stiffness, while its speed in fiberglass is essentially the same as its speed in air. Both materials are fairly good sound absorbers, but sound waves passing through a few inches of them are still not going to be absorbed sufficiently. The point of combining them is that a sound wave that tries to get out will be strongly reflected at each of the fiberglass-plywood boundaries, and will bounce back and forth many times like a ping pong ball. Due to all the back-and-forth motion, the sound may end up traveling a total distance equal to ten times the actual thickness of the soundproofing before it escapes. This is the equivalent of having ten times the thickness of sound-absorbing material.

Example: radio transmission

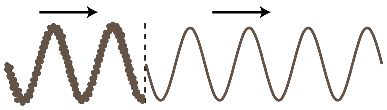
A radio transmitting station, such as a commercial station or an amateur "ham" radio station, must have a length of wire or cable connecting the amplifier to the antenna. The cable and the antenna act as two different media for radio waves, and there will therefore be partial reflection of the waves as they come from the cable to the antenna. If the waves bounce back and forth many times between the amplifier and the antenna, a great deal of their energy will be absorbed. There are two ways to attack the problem. One possibility is to design the antenna so that the speed of the waves in it is the same as the speed of the waves in the cable. There is then no reflection. The other method is to connect the amplifier to the antenna using a type of wire or cable that does not strongly absorb the waves. Partial reflection then becomes irrelevant, since all the wave energy will eventually exit through the antenna.

Discussion Question

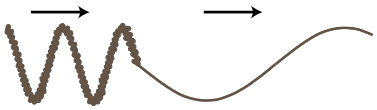
A sound wave that underwent a pressure-inverting reflection would have its compressions converted to expansions and vice versa. How would its energy and frequency compare with those of the original sound? Would it sound any different? What happens if you swap the two wires where they connect to a stereo speaker, resulting in waves that vibrate in the opposite way?

4.2* Quantitative Treatment of Reflection

In this optional section we analyze the reasons why reflections occur at a speed-changing boundary, predict quantitatively the intensities of reflection and transmission, and discuss how to predict for any type of wave which reflections are inverting and which are uninverting. The gory details are likely to be of interest mainly to students with concentrations in the physical sciences, but all readers are encouraged at least to skim the first two subsections for physical insight.



(a) A change in frequency without a change in wavelength would produce a discontinuity in the wave.



(b) A simple change in wavelength without a reflection would result in a sharp kink in the wave.

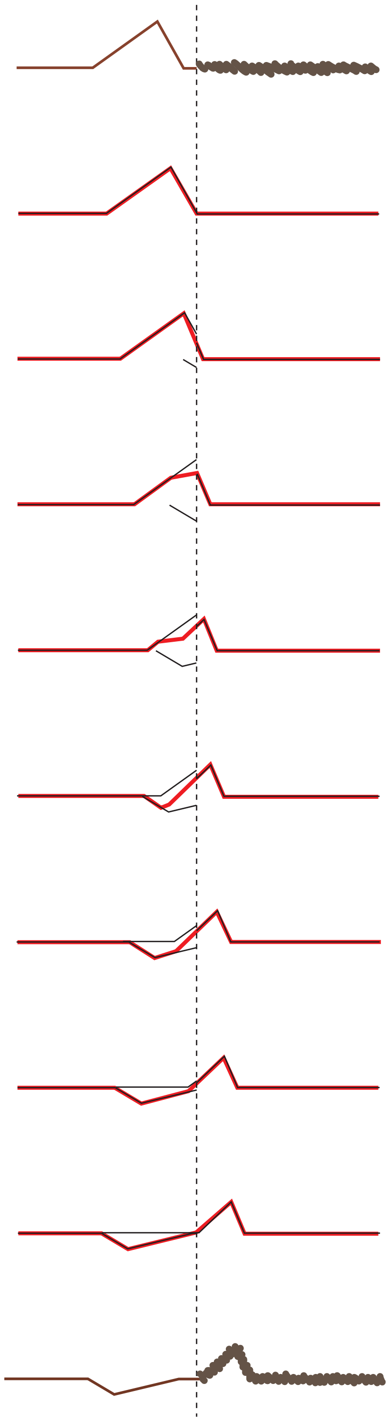
Why reflection occurs

To understand the fundamental reasons for what does occur at the boundary between media, let’s first discuss what doesn’t happen. For the sake of concreteness, consider a sinusoidal wave on a string. If the wave progresses from a heavier portion of the string, in which its velocity is low, to a lighter-weight part, in which it is high, then the equation $v=f\lambda$ tells us that it must change its frequency, or its wavelength, or both. If only the frequency changed, then the parts of the wave in the two different portions of the string would quickly get out of step with each other, producing a discontinuity in the wave, (a). This is unphysical, so we know that the wavelength must change while the frequency remains constant, (b).

But there is still something unphysical about figure (b). The sudden change in the shape of the wave has resulted in a sharp kink at the boundary. This can’t really happen, because the medium tends to accelerate in such a way as to eliminate curvature. A sharp kink corresponds to an infinite curvature at one point, which would produce an infinite acceleration, which would not be consistent with the smooth pattern of wave motion envisioned in fig. (b). Waves can have kinks, but not stationary kinks.

We conclude that without positing partial reflection of the wave, we cannot simultaneously satisfy the requirements of (1) continuity of the wave, and (2) no sudden changes in the slope of the wave. (The student who has studied calculus will recognize this as amounting to an assumption that both the wave and its derivative are continuous functions.)

Does this amount to a proof that reflection occurs? Not quite. We have only proven that certain types of wave motion are not valid solutions. In the following subsection, we prove that a valid solution can always be found in which a reflection occurs. Now in physics, we normally assume (but seldom prove formally) that the equations of motion have a unique solution, since otherwise a given set of initial conditions could lead to different behavior later on, but the Newtonian universe is supposed to be deterministic. Since the solution must be unique, and we derive below a valid solution involving a reflected pulse, we will have ended up with what amounts to a proof of reflection.



A pulse being partially reflected and partially transmitted at the boundary between two strings in which the wave speed is different. The top drawing shows the pulse heading to the right, toward the heaver string. For clarity, all but the first and last drawings are schematic. Once the reflected pulse begins to emerge from the boundary, it adds together with the trailing parts of the incident pulse. Their sum, shown as a wider line, is what is actually observed.

Intensity of reflection

We will now show, in the case of waves on a string, that it is possible to satisfy the physical requirements given above by constructing a reflected wave, and as a bonus this will produce an equation for the proportions of reflection and transmission and a prediction as to which conditions will lead to inverted and which to uninverted reflection. We assume only that the principle of superposition holds, which is a good approximations for waves on a string of sufficiently small amplitude.

Let the unknown amplitudes of the reflected and transmitted waves be R and T, respectively. An inverted reflection would be represented by a negative value of R. We can without loss of generality take the incident (original) wave to have unit amplitude. Superposition tells us that if, for instance, the incident wave had double this amplitude, we could immediately find a corresponding solution simply by doubling R and T.

Just to the left of the boundary, the height of the wave is given by the height 1 of the incident wave, plus the height R of the part of the reflected wave that has just been created and begun heading back, for a total height of $1+R$. On the right side immediately next to the boundary, the transmitted wave has a height T. To avoid a discontinuity, we must have

$$1+R = T \quad .$$

Next we turn to the requirement of equal slopes on both sides of the boundary. Let the slope of the incoming wave be s immediately to the left of the junction. If the wave was 100% reflected, and without inversion, then the slope of the reflected wave would be $-s$, since the wave has been reversed in direction. In general, the slope of the reflected wave equals $-sR$, and the slopes of the superposed waves on the left side add up to $s-sR$. On the right, the slope depends on the amplitude, T, but is also changed by the stretching or compression of the wave due to the change in speed. If, for example, the wave speed is twice as great on the right side, then the slope is cut in half by this effect. The slope on the right is therefore $s(v_1/v_2)T$, where v_1 is the velocity in the original medium and v_2 the velocity in the new medium. Equality of slopes gives $s-sR = s(v_1/v_2)T$, or

$$1 - R = \frac{v_1}{v_2}T \quad .$$

Solving the two equations for the unknowns R and T gives

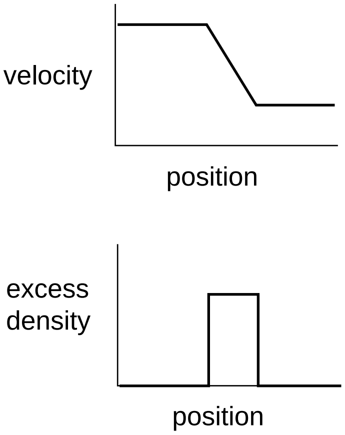
$$R = \frac{v_2 - v_1}{v_2 + v_1} \quad \text{and} \quad T = \frac{2v_2}{v_2 + v_1} \quad .$$

The first equation shows that there is no reflection unless the two wave speeds are different, and that the reflection is inverted in reflection back into a fast medium.

The energies of the transmitted and reflected wavers always add up to the same as the energy of the original wave. There is never any abrupt loss (or gain) in energy when a wave crosses a boundary; conversion of wave energy to heat occurs for many types of waves, but it occurs throughout the medium. The equation for T, surprisingly, allows the amplitude of the transmitted wave to be greater than 1, i.e. greater than that of the incident wave. This does not violate conservation of energy, because this occurs when the second string is less massive, reducing its kinetic energy, and the transmitted pulse is broader and less strongly curved, which lessens its potential energy.

Inverted and uninverted reflections in general

For waves on a string, reflections back into a faster medium are inverted, while those back into a slower medium are uninverted. Is this true for all types of waves? The rather subtle answer is that it depends on what property of the wave you are discussing.



Let’s start by considering wave disturbances of freeway traffic. Anyone who has driven frequently on crowded freeways has observed the phenomenon in which one driver taps the brakes, starting a chain reaction that travels backward down the freeway as each person in turn exercises caution in order to avoid rear-ending anyone. The reason why this type of wave is relevant is that it gives a simple, easily visualized example of our description of a wave depends on which aspect of the wave we have in mind. In steadily flowing freeway traffic, both the density of cars and their velocity are constant all along the road. Since there is no disturbance in this pattern of constant velocity and density, we say that there is no wave. Now if a wave is touched off by a person tapping the brakes, we can either describe it as a region of high density or as a region of decreasing velocity.

The freeway traffic wave is in fact a good model of a sound wave, and a sound wave can likewise be described either by the density (or pressure) of the air or by its speed. Likewise many other types of waves can be described by either of two functions, one of which is often the derivative of the other with respect to position.

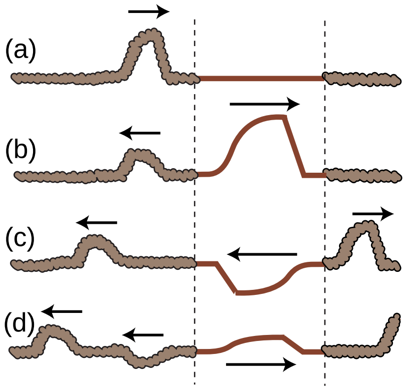
Now let’s consider reflections. If we observe the freeway wave in a mirror, the high-density area will still appear high in density, but velocity in the opposite direction will now be described by a negative number. A person observing the mirror image will draw the same density graph, but the velocity graph will be flipped across the x axis, and its original region of negative slope will now have positive slope. Although I don’t know any physical situation that would correspond to the reflection of a traffic wave, we can immediately apply the same reasoning to sound waves, which often do get reflected, and determine that a reflection can either be density-inverting and velocity-uninverting or density-uninverting and velocity-inverting.

This same type of situation will occur over and over as one encounters new types of waves, and to apply the analogy we need only determine which quantities, like velocity, become negated in a mirror image and which, like density, stay the same.

A light wave, for instance consists of a traveling pattern of electric and magnetic fields. All you need to know in order to analyze the reflection of light waves is how electric and magnetic fields behave under reflection; you don’t need to know any of the detailed physics of electricity and magnetism. An electric field can be detected, for example, by the way one’s hair stands on end. The direction of the hair indicates the direction of the electric field. In a mirror image, the hair points the other way, so the electric field is apparently reversed in a mirror image. The behavior of magnetic fields, however, is a little tricky. The magnetic properties of a bar magnet, for instance, are caused by the aligned rotation of the outermost orbiting electrons of the atoms. In a mirror image, the direction of rotation is reversed, say from clockwise to counterclockwise, and so the magnetic field is reversed twice: once simply because the whole picture is flipped and once because of the reversed rotation of the electrons. In other words, magnetic fields do not reverse themselves in a mirror image. We can thus predict that there will be two possible types of reflection of light waves. In one, the electric field is inverted and the magnetic field uninverted. In the other, the electric field is uninverted and the magnetic field inverted.

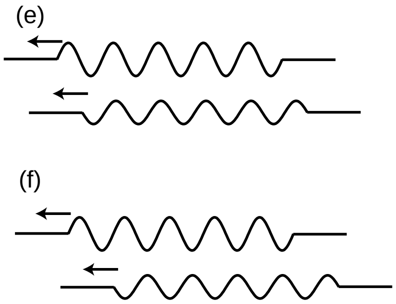
4.3 Interference Effects

If you look at the front of a pair of high-quality binoculars, you will notice a greenish-blue coating on the lenses. This is advertized as a coating to prevent reflection. Now reflection is clearly undesirable — we want the light to go *in* the binoculars — but so far I've described reflection as an unalterable fact of nature, depending only on the properties of the two wave media. The coating can't change the speed of light in air or in glass, so how can it work? The key is that the coating itself is a wave medium. In other words, we have a three-layer sandwich of materials: air, coating, and glass. We will analyze the way the coating works, not because optical coatings are an important part of your education but because it provides a good example of the general phenomenon of wave interference effects.



There are two different interfaces between media: an air-coating boundary and a coating-glass boundary. Partial reflection and partial transmission will occur at each boundary. For ease of visualization let's start by considering an equivalent system consisting of three dissimilar pieces of string tied together, and a wave pattern consisting initially of a single pulse. Figure (a) shows the incident pulse moving through the heavy rope, in which its velocity is low. When it encounters the lighter-weight rope in the middle, a faster medium, it is partially reflected and partially transmitted. (The transmitted pulse is bigger, but nevertheless has only part of the original energy.) The pulse transmitted by the first interface is then partially reflected and partially transmitted by the second boundary, (c). In figure (d), two pulses are on the way back out to the left, and a single pulse is heading off to the right. (There is still a weak pulse caught between the two boundaries, and this will rattle back and forth, rapidly getting too weak to detect as it leaks energy to the outside with each partial reflection.)

Note how of the two reflected pulses in (d), one is inverted and one uninverted. One underwent reflection at the first boundary (a reflection back into a slower medium is uninverted), but the other was reflected at the second boundary (reflection back into a faster medium is inverted).



Now let's imagine what would have happened if the incoming wave pattern had been a long sinusoidal wave train instead of a single pulse. The first two waves to reemerge on the left could be in phase, (e), or out of phase, (f), or anywhere in between. The amount of lag between them depends entirely on the width of the middle segment of string. If we choose the width of the middle string segment correctly, then we can arrange for destructive interference to occur, (f), with cancellation resulting in a very weak reflected wave.

This whole analysis applies directly to our original case of optical coatings. Visible light from most sources does consist of a stream of short sinusoidal wave-trains such as the ones drawn above. The only real difference between the waves-on-a-rope example and the case of an optical coating is that the first and third media are air and glass, in which light does not have the same speed. However, the general result is the same as long as the air and the glass have light-wave speeds that either both greater than the coating's or both less than the coating's.

The business of optical coatings turns out to be a very arcane one, with a plethora of trade secrets and "black magic" techniques handed down from master to apprentice. Nevertheless, the ideas you have learned about waves in general are sufficient to allow you to come to some definite conclusions without any further technical knowledge. The self-check and discussion questions will direct you along these lines of thought.

Self-Check

Color corresponds to wavelength of light waves. Is it possible to choose a thickness for an optical coating that will produce destructive interference for all colors of light?

This example was typical of a wide variety of wave interference effects. With a little guidance, you are now ready to figure out for yourself other examples such as the rainbow pattern made by a compact disc or by a layer of oil on a puddle.

Discussion Questions

A. Is it possible to get *complete* destructive interference in an optical coating, at least for light of one specific wavelength?

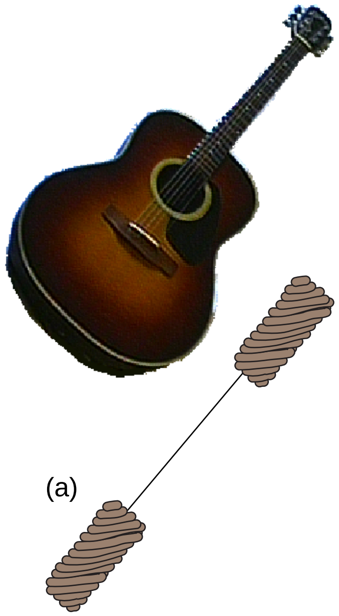
B. Sunlight consists of sinusoidal wave-trains containing on the order of a hundred cycles back-to-back, for a length of something like a tenth of a millimeter. What happens if you try to make an optical coating thicker than this?

C. Suppose you take two microscope slides and lay one on top of the other so that one of its edges is resting on the corresponding edge of the bottom one. If you insert a sliver of paper or a hair at the opposite end, a wedge-shaped layer of air will exist in the middle, with a thickness that changes gradually from one end to the other. What would you expect to see if the slides were illuminated from above by light of a single color? How would this change if you gradually lifted the lower edge of the top slide until the two slides were finally parallel?

D. An observation like the one described in the previous discussion question was used by Newton as evidence *against* the wave theory of light! If Newton didn't know about inverting and noninverting reflections, what would have seemed inexplicable to him about the region where the air layer had zero or nearly zero thickness?

No. To get the best possible interference, the thickness of the coating must be such that the second reflected wave train lags behind the first by an integer number of wavelengths. Optimal performance can therefore only be produced for one specific color of light. The typical greenish color of the coatings shows that it does the worst job for green light.

4.4 Waves Bounded on Both Sides

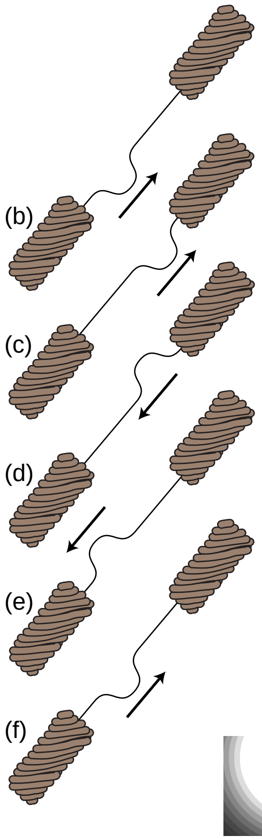


We model a guitar string attached to the guitar's body at both ends as a light-weight string attached to extremely heavy strings at its ends.

In the example of the previous section, it was theoretically true that a pulse would be trapped permanently in the middle medium, but that pulse was not central to our discussion, and in any case it was weakening severely with each partial reflection. Now consider a guitar string. At its ends it is tied to the body of the instrument itself, and since the body is very massive, the behavior of the waves when they reach the end of the string can be understood in the same way as if the actual guitar string was attached on the ends to strings that were extremely massive. Reflections are most intense when the two media are very dissimilar. Because the wave speed in the body is so radically different from the speed in the string, we should expect nearly 100% reflection.

Although this may seem like a rather bizarre physical model of the actual guitar string, it already tells us something interesting about the behavior of a guitar that we would not otherwise have understood. The body, far from being a passive frame for attaching the strings to, is actually the exit path for the wave energy in the strings. With every reflection, the wave pattern on the string loses a tiny fraction of its energy, which is then conducted through the body and out into the air. (The string has too little cross-section to make sound waves efficiently by itself.) By changing the properties of the body, moreover, we should expect to have an effect on the manner in which sound escapes from the instrument. This is clearly demonstrated by the electric guitar, which has an extremely massive, solid wooden body. Here the dissimilarity between the two wave media is even more pronounced, with the result that wave energy leaks out of the string even more slowly. This is why an electric guitar with no electric pickup can hardly be heard at all, and it is also the reason why notes on an electric guitar can be sustained for longer than notes on an acoustic guitar.

If we initially create a disturbance on a guitar string, how will the reflections behave? In reality, the finger or pick will give the string a triangular shape before letting it go, and we may think of this triangular shape as a very broad "dent" in the string which will spread out in both directions. For simplicity, however, let's just imagine a wave pattern that initially consists of a simplicity, narrow pulse traveling up the neck, (b). After reflection from the top end, it is inverted, (d). Now something interesting happens: figure (f) is identical to figure (b). After two reflections, the pulse has been inverted twice and has changed direction twice. It is now back where it started. The motion is periodic. This is why a guitar produces sounds that have a definite sensation of pitch.

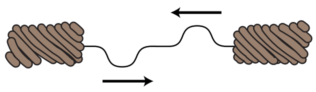


Self-Check

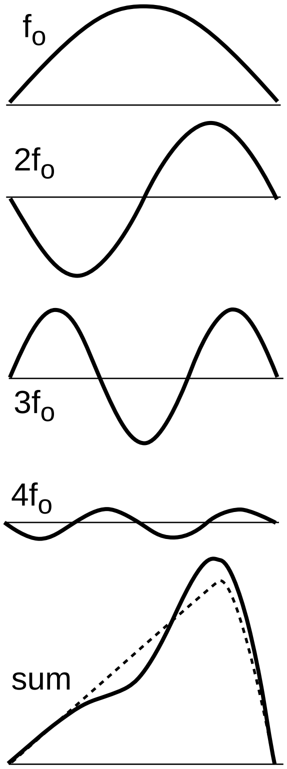
Notice that from (b) to (f), the pulse has passed by every point on the string exactly twice. This means that the total distance it has traveled equals 2L, where L is the length of the string. Given this fact, what are the period and frequency of the sound it produces, expressed in terms of L and v, the velocity of the wave? [answer on next page]

Note that if the waves on the string obey the principle of superposition, then the velocity must be independent of amplitude, and the guitar will produce the same pitch regardless of whether it is played loudly or softly. In reality, waves on a string obey the principle of superposition approximately, but not exactly. The guitar, like just about any acoustic instrument, is a little out of tune when played loudly. (The effect is more pronounced for wind instruments than for strings, but wind players are able to compensate for it.)

(g)



Now there is only one hole in our reasoning. Suppose we somehow arrange to have an initial setup consisting of two identical pulses heading toward each other, as in figure (g). They will pass through each other, undergo a single inverting reflection, and come back to a configuration in which their positions have been exactly interchanged. This means that the period of vibration is half as long. The frequency is twice as high.

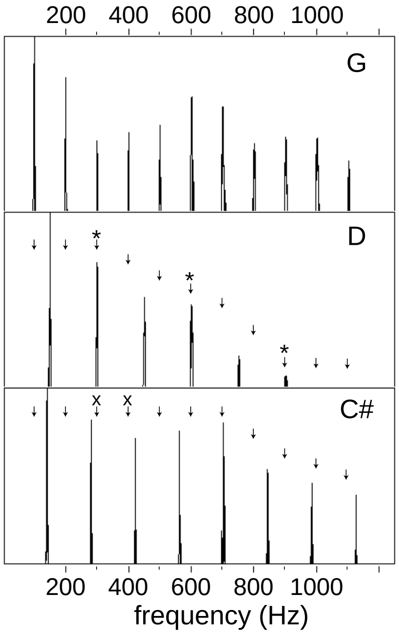


This might seem like a purely academic possibility, since nobody actually plays the guitar with two picks at once! But in fact it is an example of a very general fact about waves that are bounded on both sides. A mathematical theorem called Fourier's theorem states that any wave can be created by superposing sine waves. The figure on the left shows how even by using only four sine waves with appropriately chosen amplitudes, we can arrive at a sum which is a decent approximation to the realistic triangular shape of a guitar string being plucked. The one-hump wave, in which half a wavelength fits on the string, will behave like the single pulse we originally discussed. We call its frequency f_o. The two-hump wave, with one whole wavelength, is very much like the two-pulse example. For the reasons discussed above, its frequency is $2f_o$. Similarly, the three-hump and four-hump waves have frequencies of $3f_o$ and $4f_o$.

Theoretically we would need to add together infinitely many such wave patterns to describe the initial triangular shape of the string exactly, although the amplitudes required for the very high frequency parts would be very small, and an excellent approximation could be achieved with as few as ten waves.

We thus arrive at the following very general conclusion. Whenever a wave pattern exists in a medium bounded on both sides by media in which the wave speed is very different, the motion can be broken down into the motion of a (theoretically infinite) series of sine waves, with frequencies f_o, $2f_o$, $3f_o$, ... Except for some technical details, to be discussed below, this analysis applies to a vast range of sound-producing systems, including the air column within the human vocal tract. Because sounds composed of this kind of pattern of frequencies are so common, our ear-brain system has evolved so as to perceive them as a single, fused sensation of tone.

[answer to self-check on previous page] The period is the time required to travel a distance $2L$ at speed v, $T=2L/v$. The frequency is $f=1/T=v/2L$.



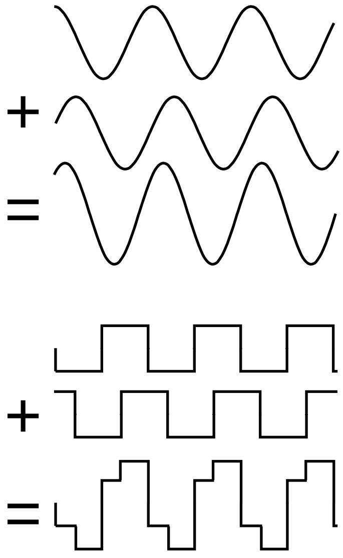
Graphs of loudness versus frequency for the vowel “ah,” sung as three different musical notes. G is consonant with D, since every overtone of G that is close to an overtone of D (marked “*”) is at exactly the same frequency. G and C# are dissonant together, since some of the overtones of G (marked “x”) are close to, but not right on top of, those of C#.

Musical applications

Many musicians claim to be able to identify individually several of the frequencies above the first one, called *overtones* or *harmonics*, but they are kidding themselves. In reality, the overtone series has two important roles in music, neither of which depends on this fictitious ability to “hear out” the individual overtones.

First, the relative strengths of the overtones is an important part of the personality of a sound, called its *timbre* (rhymes with “amber”). The characteristic tone of the brass instruments, for example, is a sound that starts out with a very strong harmonic series extending up to very high frequencies, but whose higher harmonics die down drastically as the attack changes to the sustained portion of the note.

Second, although the ear cannot separate the individual harmonics of a single musical tone, it is very sensitive to clashes between the overtones of notes played simultaneously, i.e. in harmony. We tend to perceive a combination of notes as being dissonant if they have overtones that are close but not the same. Roughly speaking, strong overtones whose frequencies differ by more than 1% and less than 10% cause the notes to sound dissonant. It is important to realize that the term “dissonance” is not a negative one in music. No matter how long you search the radio dial, you will never hear more than three seconds of music without at least one dissonant combination of notes. Dissonance is a necessary ingredient in the creation of a musical cycle of tension and release. Musically knowledgeable people do not usually use the word “dissonant” as a criticism of music, and if they do, what they really saying is that the dissonance has been used in a clumsy way, or without providing any contrast between dissonance and consonance.



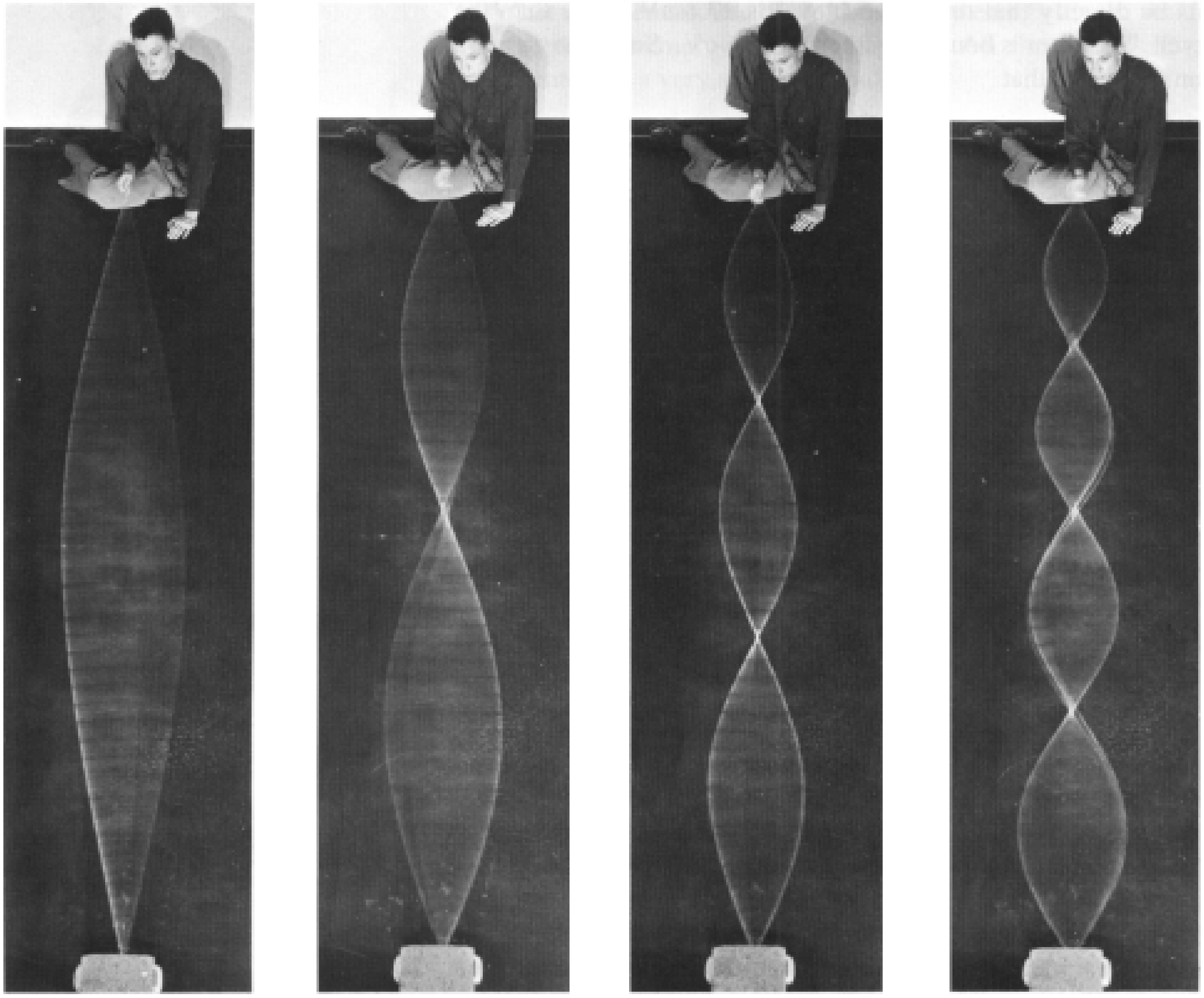
If you take a sine wave and make a copy of it shifted over, their sum is still a sine wave. The same is not true for a square wave.

Standing waves

The photos below show sinusoidal wave patterns made by shaking a rope. I used to enjoy doing this at the bank with the pens on chains, back in the days when people actually went to the bank. You might think that I and the person in the photos had to practice for a long time in order to get such nice sine waves. In fact, a sine wave is the only shape that can create this kind of wave pattern, called a *standing wave*, which simply vibrates back and forth in one place without moving. The sine wave just creates itself automatically when you find the right frequency, because no other shape is possible.

If you think about it, it's not even obvious that sine waves should be able to do this trick. After all, waves are supposed to travel at a set speed, aren't they? The speed isn't supposed to be zero! Well, we can actually think of a standing wave as a superposition of a moving sine wave with its own reflection, which is moving the opposite way. Sine waves have the unique mathematical property that the sum of sine waves of equal wavelength is simply a new sine wave with the same wavelength. As the two sine waves go back and forth, they always cancel perfectly at the ends, and their sum appears to stand still.

Standing wave patterns are rather important, since atoms are really standing-wave patterns of electron waves. You are a standing wave!



Uncopyrighted photos from PSSC Physics.

Standing-wave patterns of air columns

The air column inside a wind instrument or the human vocal tract behaves very much like the wave-on-a-string example we've been concentrating on so far, the main difference being that we may have either inverting or noninverting reflections at the ends.

Inverting reflection at one end and uninverting at the other

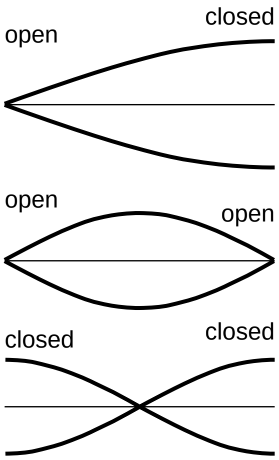
If you blow over the mouth of a beer bottle to produce a tone, the bottle outlines an air column that is closed at the bottom and open at the top. Sound waves will be reflected at the bottom because of the difference in the speed of sound in air and glass. The speed of sound is greater in glass (because its stiffness more than compensates for its higher density compared to air). Using the type of reasoning outlined in optional section 4.2, we find that this reflection will be density-uninverting: a compression comes back as a compression, and a rarefaction as a rarefaction. There will be strong reflection and very weak transmission, since the difference in speeds is so great. But why do we get a reflection at the *mouth* of the bottle? There is no change in medium there, and the air inside the bottle is connected to the air in the room. This is a type of reflection that has to do with the three-dimensional shape of the sound waves, and cannot be treated the same way as the other types of reflection we have encountered. Since this chapter is supposed to be confined mainly to wave motion in one dimension, and it would take us too far afield here to explain it in detail, but a general justification is given in the caption of the figure.

Surprisingly, sound waves undergo partial reflection at the open ends of tubes as well as closed ones. The reason has to do with the readjustment of the wave pattern from a plane wave to a spherical wave. If the readjustment was as sudden as that shown in the figure, then there would be kinks in the wave. Waves don't like to develop kinks. In section 4.2 we deduced the strength of the reflected wave at a change in medium from the requirement that the wave would not have discontinuities or kinks. Here there is no change in medium, but a reflected wave is still required in order to avoid kinks.

The important point is that whereas the reflection at the bottom was density-uninverting, the one at the top is density-inverting. This means that at the top of the bottle, a compression superimposes with its own reflection, which is a rarefaction. The two nearly cancel, and so the wave has almost zero amplitude at the mouth of the bottle. The opposite is true at the bottom — here we have a peak in the standing-wave pattern, not a stationary point. The standing wave with the lowest frequency, i.e. the longest wave length, is therefore one in which 1/4 of a wavelength fits along the length of the tube.

Both ends the same

If both ends are open (as in the flute) or both ends closed (as in some organ pipes), then the standing wave pattern must be symmetric. The lowest-frequency wave fits half a wavelength in the tube.



Graphs of excess density versus position for the lowest-frequency standing waves of three types of air columns. Points on the axis have normal air density.

Self-Check

Draw a graph of pressure versus position for the first overtone of the air column in a tube open at one end and closed at the other. This will be the next-to-longest possible wavelength that allows for a point of maximum vibration at one end and a point of no vibration at the other. How many times shorter will its wavelength be compared to the frequency of the lowest-frequency standing wave, shown in the figure? Based on this, how many times greater will its frequency be? [Answer on next page.]

Summary

Selected Vocabulary

reflection	the bouncing back of part of a wave from a boundary
transmission	the continuation of part of a wave through a boundary
absorption	the gradual conversion of wave energy into heating of the medium
standing wave	a wave pattern that stays in one place

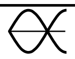
Notation

λ	wavelength (Greek letter lambda)

Summary

Whenever a wave encounters the boundary between two media in which its speeds are different, part of the wave is reflected and part is transmitted. The reflection is always reversed front-to-back, but may also be inverted in amplitude. Whether the reflection is inverted depends on how the wave speeds in the two media compare, e.g. a wave on a string is uninverted when it is reflected back into a segment of string where its speed is lower. The greater the difference in wave speed between the two media, the greater the fraction of the wave energy that is reflected. Surprisingly, a wave in a dense material like wood will be strongly reflected back into the wood at a wood-air boundary.

A one-dimensional wave confined by highly reflective boundaries on two sides will display motion which is periodic. For example, if both reflections are inverting, the wave will have a period equal to twice the time required to traverse the region, or to that time divided by an integer. An important special case is a sinusoidal wave; in this case, the wave forms a stationary pattern composed of a superposition of sine waves moving in opposite direction.

[Answer to self-check on previous page.] The wave pattern will look like this:  .Three quarters of a wavelength fit in the tube, so the wavelength is three times shorter than that of the lowest-frequency mode, in which one quarter of a wave fits. Since the wavelength is smaller by a factor of three, the frequency is three times higher. Instead of f_o, $2f_o$, $3f_o$, $4f_o$, ..., the pattern of wave frequencies of this air column goes f_o, $3f_o$, $5f_o$, $7f_o$, ...

Homework Problems

1. Light travels faster in warmer air. Use this fact to explain the formation of a mirage appearing like the shiny surface of a pool of water when there is a layer of hot air above a road.

2. (a) Using the equations from optional section 4.2, compute the amplitude of light that is reflected back into air at an air-water interface, relative to the amplitude of the incident wave. The speeds of light in air and water are 3.0x10^{8} and 2.2x10^{8} m/s, respectively.

(b✓) Find the energy of the reflected wave as a fraction of the incident energy. [Hint: The answers to the two parts are not the same.]

3. A B-flat clarinet (the most common kind) produces its lowest note, at about 230 Hz, when half of a wavelength fits inside its tube. Compute the length of the clarinet. [Check: The actual length of a clarinet is about 67 cm from the tip of the mouthpiece to the end of the bell. Because the behavior of the clarinet and its coupling to air outside it is a little more complex than that of a simple tube enclosing a cylindrical air column, your answer will be close to this value, but not exactly equal to it.]

4. (a) A good tenor saxophone player can play all of the following notes without changing her fingering, simply by altering the tightness of her lips: Eb (150 Hz), Eb (300 Hz), Bb (450 Hz), and Eb (600 Hz). How is this possible? (b) Some saxophone players are known for their ability to use this technique to play "freak notes," i.e. notes above the normal range of the instrument. Why isn't it possible to play notes below the normal range using this technique?

C	261.6 Hz
D	293.7
E	329.6
F	349.2
G	392.0
A	440.0
B flat	466.2

Problem 5.

5. The table gives the frequencies of the notes that make up the key of F major, starting from middle C and going up through all seven notes. (a) Calculate the first five or six harmonics of C and G, and determine whether these two notes will be consonant or dissonant. (b) Do the same for C and B flat. [Hint: Remember that harmonics that differ by about 1-10% cause dissonance.]

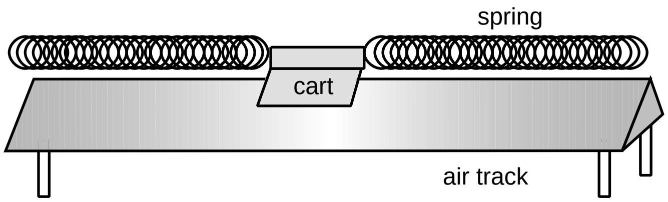
S A solution is given in the back of the book.　★ A difficult problem.
✓ A computerized answer check is available.　∫ A problem that requires calculus.

Exercises

Exercise 1A: Vibration

Equipment:

air track and carts of two different masses
springs
alligator clips
meter sticks
spring scales
stopwatches



Place the cart on the air track and attach springs so that it can vibrate.

1. Test whether the period of vibration depends on amplitude. Try at least two moderate amplitudes, for which the springs do not go slack, and at least one amplitude that is large enough so that they dop go slack.

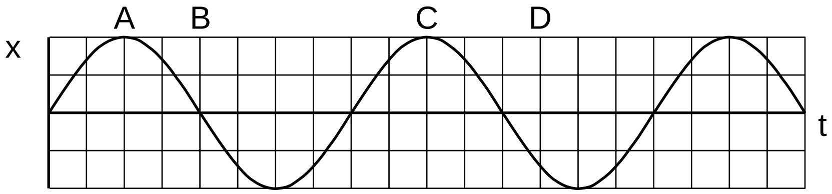
2. Try a cart with a different mass. Does the period change by the expected factor, based on the equation $T=2\pi\sqrt{m/k}$?

3. Use a spring scale to pull the cart away from equilibrium, and make a graph of force versus position. Is it linear? If so, what is its slope?

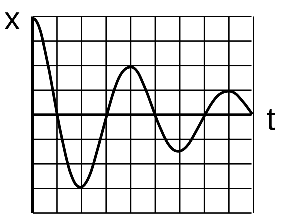
4. Test the equation $T=2\pi\sqrt{m/k}$ numerically.

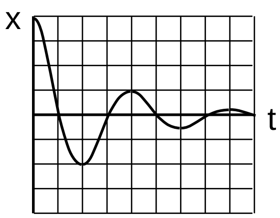
Exercise 2A: Worksheet on Resonance

1. Compare the oscillator’s energies at A, B, C, and D.

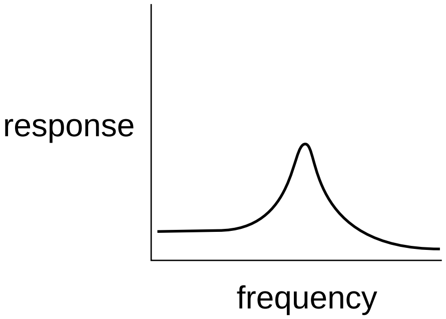


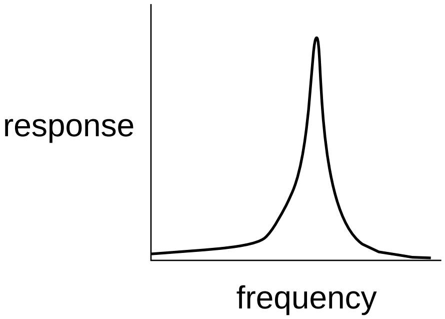
2. Compare the Q values of the two oscillators.





3. Match the *x*-*t* graphs in #2 with the amplitude-frequency graphs below.





Glossary

Amplitude. The amount of vibration, often measured from the center to one side; may have different units depending on the nature of the vibration.

Damping. the dissipation of a vibration's energy into heat energy, or the frictional force that causes the loss of energy

Driving force. An external force that pumps energy into a vibrating system.

Frequency. The number of cycles per second, the inverse of the period (q.v.).

Period. The time required for one cycle of a periodic motion (q.v.).

Periodic motion. Motion that repeats itself over and over.

Resonance. The tendency of a vibrating system to respond most strongly to a driving force whose frequency is close to its own natural frequency of vibration.

Simple harmonic motion. Motion whose *x-t* graph is a sine wave.

Steady state. The behavior of a vibrating system after it has had plenty of time to settle into a steady response to a driving force. In the steady state, the same amount of energy is pumped into the system during each cycle as is lost to damping during the same period.

Quality factor. The number of oscillations required for a system's energy to fall off by a factor of 535 due to damping.

Index

Photo Credits

All photographs are by Benjamin Crowell, except as noted below.

Chapter 2
Tacoma Narrows Bridge: Uncopyrighted still photos and a movie of the bridge's collapse were taken by an unknown photographer.
Nimits Freeway: Uncopyrighted photo by an unknown photographer, courtesy of the UC Berkeley Earth Sciences and Map Library.
Brain: R. Malladi, LBNL.

Chapter 3
Painting of ocean waves: Hokusai
M100: Hubble Space Telescope.

Chapter 4
Human Cross-Section: Courtesy of the Visible Human Project, National Library of Medicine, US NIH.

Useful Data

Metric Prefixes

M-	mega-	10^6
k-	kilo-	10^3
m-	milli-	10^{-3}
μ- (Greek mu)	micro-	10^{-6}
n-	nano-	10^{-9}

(Centi-, 10^{-2}, is used only in the centimeter.)

Notation and Units

quantity	unit	symbol
distance	meter, m	x, Δx
time	second, s	t, Δt
mass	kilogram, kg	m
density	kg/m^3	ρ
force	newton, 1 N=1 kg·m/s^2	$\boldsymbol{F}$
velocity	m/s	$\boldsymbol{v}$
acceleration	m/s^2	$\boldsymbol{a}$
energy	joule, J	E
momentum	kg·m/s	$\boldsymbol{p}$
angular momentum	kg·m^2/s	L
period	s	T
wavelength	m	λ
frequency	s^{-1} or Hz	f

symbol	meaning
$\propto$	is proportional to
$\approx$	is approximately equal to
~	on the order of

The Greek Alphabet

α	Α	alpha	ν	Ν	nu
β	Β	beta	ξ	Ξ	xi
γ	Γ	gamma	ο	Ο	omicron
δ	Δ	delta	π	Π	pi
ε	Ε	epsilon	ρ	Ρ	rho
ζ	Ζ	zeta	σ	Σ	sigma
η	Η	eta	τ	Τ	tau
θ	Θ	theta	υ	Υ	upsilon
ι	Ι	iota	φ	Φ	phi
κ	Κ	kappa	χ	Χ	chi
λ	Λ	lambda	ψ	Ψ	psi
μ	Μ	mu	ω	Ω	omega

Conversions

Conversions between SI and other units:

1 inch	=	2.54 cm (exactly)
1 mile	=	1.61 km
1 pound	=	4.45 N
(1 kg)·g	=	2.2 lb
1 gallon	=	3.78x10^3 cm^3
1 horsepower	=	746 W
1 kcal*	=	4.18x10^3 J

*When speaking of food energy, the word “Calorie” is used to mean 1 kcal, i.e. 1000 calories. In writing, the capital C may be used to indicate 1 Calorie=1000 calories.

Conversions between U.S. units:

1 foot	=	12 inches
1 yard	=	3 feet
1 mile	=	5280 ft

Earth, Moon, and Sun

body	mass (kg)	radius (km)	radius of orbit (km)
earth	5.97x10^{24}	6.4x10^3	1.49x10^8
moon	7.35x10^{22}	1.7x10^3	3.84x10^5
sun	1.99x10^{30}	7.0x10^5	

The radii and radii of orbits are average values. The moon orbits the earth and the earth orbits the sun.

Subatomic Particles

particle	mass (kg)	radius (m)
electron	9.109x10^{-31}	? – less than about 10^{-17}
proton	1.673x10^{-27}	about 1.1x10^{-15}
neutron	1.675x10^{-27}	about 1.1x10^{-15}

The radii of protons and neutrons can only be given approximately, since they have fuzzy surfaces. For comparison, a typical atom is about 10^{-9} m in radius.

Speeds of Light and Sound

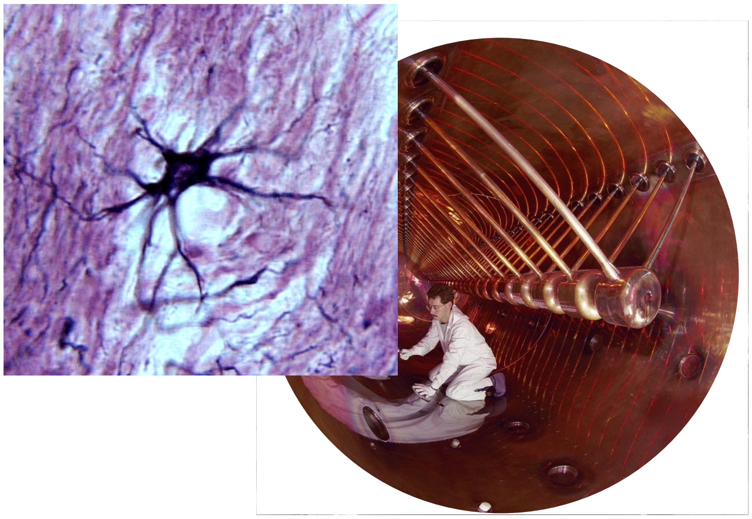
speed of light	c=3.00x10^8 m/s
speed of sound	340 m/s

Electricity

and

Magnetism

Benjamin Crowell



Book 4 in the **Light and Matter** series of introductory physics textbooks
www.lightandmatter.com

Electricity and Magnetism

The **Light and Matter** series of introductory physics textbooks:

1 Newtonian Physics

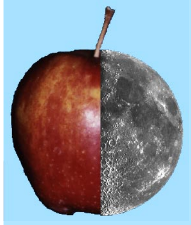
2 Conservation Laws

3 Vibrations and Waves

4 Electricity and Magnetism

5 Optics

6 The Modern Revolution in Physics



Electricity and Magnetism

Benjamin Crowell

www.lightandmatter.com

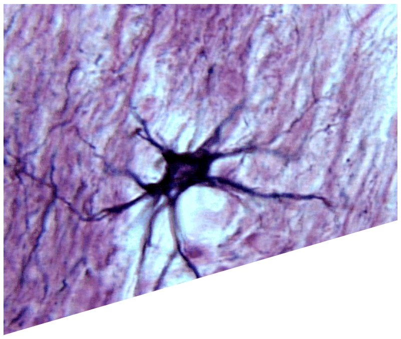
Light and Matter

Fullerton, California
www.lightandmatter.com

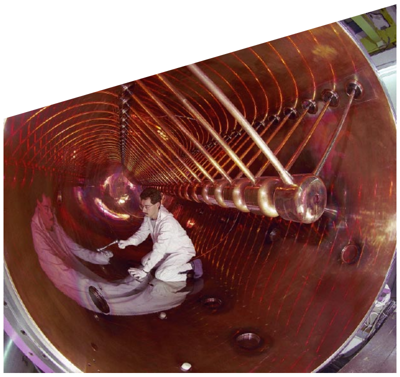
Edition 2.0
rev. 2000-09-17

ISBN 0-9704670-4-4

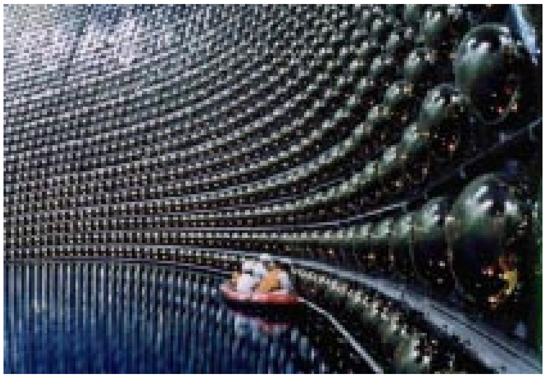
To Arnold Arons.

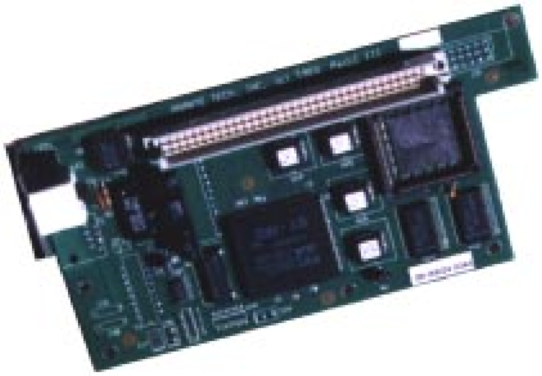


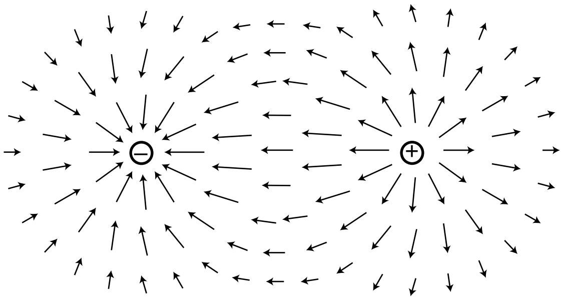
Brief Contents

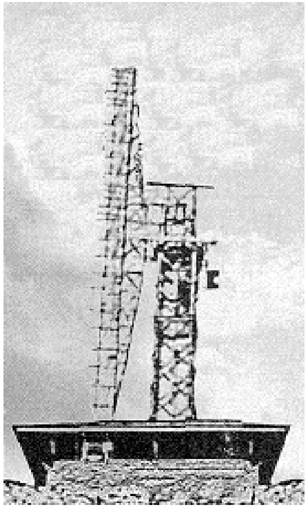


Contents









Supplement 4-7 is available online in electronic format at www.lightandmatter.com.

Preface

Who are you? However much you relate your identity to your physical appearance, you know that your personality ultimately resides in the unique arrangement of your brain's electrical network. Mary Shelley may have conceived of electricity as a mystical life force that could jerk the leg of a dead frog or animate Dr. Frankenstein's monster, but we now know the truth is both more subtle and more wonderful. Electricity is not the stuff of life but of consciousness.

Evidence is mounting that the universe has produced vast numbers of suitable habitats for life — including, within our own solar system, a watery ancient Mars and the oceans that lie under the icy surface of Jupiter's moon Europa. But even as we debate claims of fossilized Martian bacteria, a third generation of radio astronomers has found nothing but a wasteland of static in the search for extraterrestrial intelligence.

Is life ubiquitous in the universe but consciousness rare? In terms of geologic time, it took a mere wink of an eye for life to come into being on Earth once conditions were suitable, so there is every reason to believe that it exists elsewhere. Large-brained mammals, however, appear as a virtual afterthought in the record of our biosphere, which remains dominated by single-celled life. Now you begin your study of electricity and magnetism, the phenomena of which your own mind is made. Give some though to this image of awesome loneliness: there may be no other planet in our galaxy of ten billion stars where a collection of electric charges and fields can ponder its own existence.

1 Electricity and the Atom

Where the telescope ends, the microscope begins. Which of the two has the grander view?

Victor Hugo

His father died during his mother's pregnancy. Rejected by her as a boy, he was packed off to boarding school when she remarried. He himself never married, but in middle age he formed an intense relationship with a much younger man, a relationship that he terminated when he underwent a psychotic break. Following his early scientific successes, he spent the rest of his professional life mostly in frustration over his inability to unlock the secrets of alchemy.

The man being described is Isaac Newton, but not the triumphant Newton of the standard textbook hagiography. Why dwell on the sad side of his life? To the modern science educator, Newton's lifelong obsession with alchemy may seem an embarrassment, a distraction from his main achievement, the creation the modern science of mechanics. To Newton, however, his alchemical researches were naturally related to his investigations of force and motion. What was radical about Newton's analysis of motion was its universality: it succeeded in describing both the heavens and the earth with the same equations, whereas previously it had been assumed that the sun, moon, stars, and planets were fundamentally different from earthly objects. But Newton realized that if science was to describe all of nature in a unified way, it was not enough to unite the human scale with the scale of the universe: he would not be satisfied until he fit the microscopic universe into the picture as well.

It should not surprise us that Newton failed. Although he was a firm believer in the existence of atoms, there was no more experimental evidence for their existence than there had been when the ancient Greeks first posited them on purely philosophical grounds. Alchemy labored under a tradition of secrecy and mysticism. Newton had already almost single-handedly transformed the fuzzyheaded field of "natural philosophy" into something we would recognize as the modern science of physics, and it would be unjust to criticize him for failing to change alchemy into modern chemistry as well. The time was not ripe. The microscope was a new invention, and it was cutting-edge science when Newton's contemporary Hooke discovered that living things were made out of cells.

1.1 The Quest for the Atomic Force

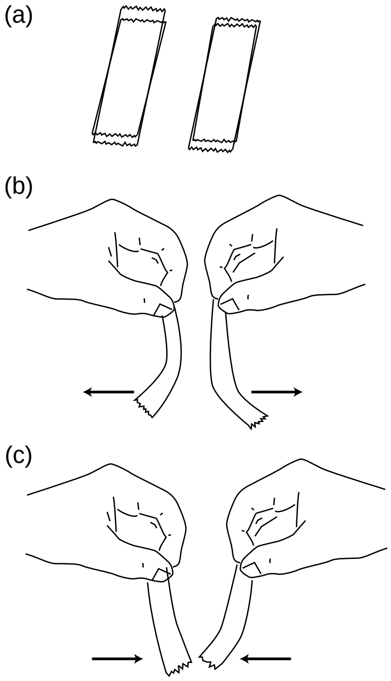
> Newton was not the first of the age of reason. He was the last of the magicians...
>
> John Maynard Keynes

Newton's quest

Nevertheless it will be instructive to pick up Newton's train of thought and see where it leads us with the benefit of modern hindsight. In uniting the human and cosmic scales of existence, he had reimagined both as stages on which the actors were objects (trees and houses, planets and stars) that interacted through attractions and repulsions. He was already convinced that the objects inhabiting the microworld were atoms, so it remained only to determine what kinds of forces they exerted on each other.

His next insight was no less brilliant for his inability to bring it to fruition. He realized that the many human-scale forces — friction, sticky forces, the normal forces that keep objects from occupying the same space, and so on — must all simply be expressions of a more fundamental force acting between atoms. Tape sticks to paper because the atoms in the tape attract the atoms in the paper. My house doesn't fall to the center of the earth because its atoms repel the atoms of the dirt under it.

Here he got stuck. It was tempting to think that the atomic force was a form of gravity, which he knew to be universal, fundamental, and mathematically simple. Gravity, however, is always attractive, so how could he use it to explain the existence of both attractive and repulsive atomic forces? The gravitational force between objects of ordinary size is also extremely small, which is why we never notice cars and houses attracting us gravitationally. It would be hard to understand how gravity could be responsible for anything as vigorous as the beating of a heart or the explosion of



gunpowder. Newton went on to write a million words of alchemical notes filled with speculation about some other force, perhaps a “divine force” or “vegetative force” that would for example be carried by the sperm to the egg.

Luckily, we now know enough to investigate a different suspect as a candidate for the atomic force: electricity. Electric forces are often observed between objects that have been prepared by rubbing (or other surface interactions), for instance when clothes rub against each other in the dryer. A useful example is shown in figure (a): stick two pieces of tape on a tabletop, and then put two more pieces on top of them. Lift each pair from the table, and then separate them. The two top pieces will then repel each other, (b), as will the two bottom pieces. A bottom piece will attract a top piece, however, (c). Electrical forces like these are similar in certain ways to gravity, the other force that we already know to be fundamental:

- Electrical forces are *universal*. Although some substances, such as fur, rubber, and plastic, respond more strongly to electrical preparation than others, all matter participates in electrical forces to some degree. There is no such thing as a “nonelectric” substance. Matter is both inherently gravitational and inherently electrical.

- Experiments show that the electrical force, like the gravitational force, is an *inverse square* force. That is, the electrical force between two spheres is proportional to $1/r^2$, where r is the center-to-center distance between them.

Furthermore, electrical forces make more sense than gravity as candidates for the fundamental force between atoms, because we have observed that they can be either attractive or repulsive.

1.2 Charge, Electricity and Magnetism

Charge

"Charge" is the technical term used to indicate that an object has been prepared so as to participate in electrical forces. This is to be distinguished from the common usage, in which the term is used indiscriminately for anything electrical. For example, although we speak colloquially of "charging" a battery, you may easily verify that a battery has no charge in the technical sense, e.g. it does not exert any electrical force on a piece of tape that has been prepared as described in the previous section.

Two types of charge

We can easily collect reams of data on electrical forces between different substances that have been charged in different ways. We find for example that cat fur prepared by rubbing against rabbit fur will attract glass that has been rubbed on silk. How can we make any sense of all this information? A vast simplification is achieved by noting that there are really only two types of charge. Suppose we pick cat fur rubbed on rabbit fur as a representative of type A, and glass rubbed on silk for type B. We will now find that there is no "type C." Any object electrified by any method is either A-like, attracting things A attracts and repelling those it repels, or B-like, displaying the same attractions and repulsions as B. The two types, A and B, always display opposite interactions. If A displays an attraction with some charged object, then B is guaranteed to undergo repulsion with it, and vice-versa.

The coulomb

Although there are only two types of charge, each type can come in different amounts. The metric unit of charge is the coulomb (rhymes with "drool on"), defined as follows:

> One Coulomb (C) is defined as the amount of charge such that a force of 9.0x10^9 N occurs between two pointlike objects with charges of 1 C separated by a distance of 1 m.

The notation for an amount of charge is *q*. The numerical factor in the definition is historical in origin, and is not worth memorizing. The definition is stated for pointlike, i.e. very small, objects, because otherwise different parts of them would be at different distances from each other.

A model of two types of charged particles

Experiments show that all the methods of rubbing or otherwise charging objects involve two objects, and both of them end up getting charged. If one object acquires a certain amount of one type of charge, then the other ends up with an equal amount of the other type. Various interpretations of this are possible, but the simplest is that the basic building blocks of matter come in two flavors, one with each type of charge. Rubbing objects together results in the transfer of some of these particles from one object to the other. In this model, an object that has not been electrically prepared may actually possesses a great deal of *both* types of charge, but the amounts are equal and they are distributed in the same way throughout it. Since type A repels anything that type B attracts, and vice versa, the object will make a total force of zero on any other object. The rest of this chapter fleshes out this model and discusses how these mysterious particles can be understood as being internal parts of atoms.

Use of positive and negative signs for charge

Because the two types of charge tend to cancel out each other's forces, it makes sense to label them using positive and negative signs, and to discuss the *total* charge of an object. It is entirely arbitrary which type of charge to call negative and which to call positive. Benjamin Franklin decided to describe the one we've been calling "A" as negative, but it really doesn't matter as long as everyone is consistent with everyone else. An object with a total charge of zero (equal amounts of both types) is referred to as electrically *neutral.*

Self-Check

Criticize the following statement: "There are two types of charge, attractive and repulsive."

Coulomb's law

A large body of experimental observations can be summarized as follows:

Coulomb's law

The magnitude of the force acting between pointlike charged objects at a center-to-center distance *r* is given by the equation

$$\left|\boldsymbol{F}\right| = k\frac{\left|q_1\right|\left|q_2\right|}{r^2} \quad ,$$

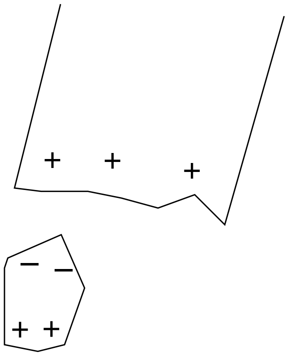
where the constant *k* equals 9.0x10^9 N·m^2/C^2. The force is attractive if the charges are of different signs, and repulsive if they have the same sign.

Clever modern techniques have allowed the 1/r^2 form of Coulomb's law to be tested to incredible accuracy, showing that the exponent is in the range from 1.9999999999999998 to 2.0000000000000002.

Note that Coulomb's law is closely analogous to Newton's law of gravity, where the magnitude of the force is $Gm_1 m_2 / r^2$, except that there is only one type of mass, not two, and gravitational forces are never repulsive. Because of this close analogy between the two types of forces, we can recycle

Either type can be involved in either an attraction or a repulsion. A positive charge could be involved in either an attraction (with a negative charge) or a repulsion (with another positive), and a negative could participate in either an attraction (with a positive) or a repulsion (with a negative).

(a) A charged piece of tape attracts uncharged pieces of paper from a distance, and they leap up to it.



(b) The paper has zero total charge, but it does have charged particles in it that can move.

a great deal of our knowledge of gravitational forces. For instance, there is an electrical equivalent of the shell theorem: the electrical forces exerted externally by a uniformly charged spherical shell are the same as if all the charge was concentrated at its center, and the forces exerted internally are zero.

Conservation of charge

An even more fundamental reason for using positive and negative signs for electrical charge is that experiments show that charge is conserved according to this definition: in any closed system, the total amount of charge is a constant. This is why we observe that rubbing initially uncharged substances together always has the result that one gains a certain amount of one type of charge, while the other acquires an equal amount of the other type. Conservation of charge seems natural in our model in which matter is made of positive and negative particles. If the charge on each particle is a fixed property of that type of particle, and if the particles themselves can be neither created nor destroyed, then conservation of charge is inevitable.

Electrical forces involving neutral objects

As shown in figure (a), an electrically charged object can attract objects that are uncharged. How is this possible? The key is that even though each piece of paper has a total charge of zero, it has at least some charged particles in it that are free to move. Suppose that the tape is positively charged, (b). Mobile particles in the paper will respond to its forces, causing one end of the paper to become negatively charged and the other to become positive. The attraction is between the paper and the tape is now stronger than the repulsion, because the negatively charge end is closer to the tape.

Self-Check

What would have happened if the tape was negatively charged?

The path ahead

We have begun to encounter complex electrical behavior that we would never have realized was occurring just from the evidence of our eyes. Unlike the pulleys, blocks, and inclined planes of mechanics, the actors on the stage of electricity and magnetism are invisible phenomena alien to our everyday experience. For this reason, the flavor of the second half of your physics education is dramatically different, focusing much more on experiments and techniques. Even though you will never actually see charge moving through a wire, you can learn to use an ammeter to measure the flow.

Students also tend to get the impression from their first semester of physics that it is a dead science. Not so! We are about to pick up the historical trail that leads directly to the cutting-edge physics research you read about in the newspaper. The atom-smashing experiments that began around 1900, which we will be studying in chapters 1 and 2, were not that different from the ones of the year 2000 — just smaller, simpler, and much cheaper.

It wouldn't make any difference. The roles of the positive and negative charges in the paper would be reversed, but there would still be a net attraction.

Magnetic forces

A detailed mathematical treatment of magnetism won't come until much later in this book, but we need to develop a few simple ideas about magnetism now because magnetic forces are used in the experiments and techniques we come to next. Everyday magnets come in two general types. Permanent magnets, such as the ones on your refrigerator, are made of iron or substances like steel that contain iron atoms. (Certain other substances also work, but iron is the cheapest and most common.) The other type of magnet, an example of which is the ones that make your stereo speakers vibrate, consist of coils of wire through which electric charge flows. Both types of magnets are able to attract iron that has not been magnetically prepared, for instance the door of the refrigerator.

A single insight makes these apparently complex phenomena much simpler to understand: magnetic forces are interactions between moving charges, occurring in addition to the electric forces. Suppose a permanent magnet is brought near a magnet of the coiled-wire type. The coiled wire has moving charges in it because we force charge to flow. The permanent magnet also has moving charges in it, but in this case the charges that naturally swirl around inside the iron. (What makes a magnetized piece of iron different from a block of wood is that the motion of the charge in the wood is random rather than organized. The moving charges in the coiled-wire magnet exert a force on the moving charges in the permanent magnet, and vice-versa.)

The mathematics of magnetism is significantly more complex than the Coulomb force law for electricity, which is why we will wait until chapter 6 before delving deeply into it. Two simple facts will suffice for now:

(1) If a charged particle is moving in a region of space near where other charged particles are also moving, their magnetic force on it is directly proportional to its velocity.

(2) The magnetic force on a moving charged particle is always perpendicular to the direction the particle is moving.

Example: A magnetic compass

The Earth is molten inside, and like a pot of boiling water, it roils and churns. To make a drastic oversimplification, electric charge can get carried along with the churning motion, so the Earth contains moving charge. The needle of a magnetic compass is itself a small permanent magnet. The moving charge inside the earth interacts magnetically with the moving charge inside the compass needle, causing the compass needle to twist around and point north.

Example: A television tube

A TV picture is painted by a stream of electrons coming from the back of the tube to the front. The beam scans across the whole surface of the tube like a reader scanning a page of a book. Magnetic forces are used to steer the beam. As the beam comes from the back of the tube to the front, up-down and left-right forces are needed for steering. But magnetic forces cannot be used to get the beam up to speed in the first place, since they can only push perpendicular to the electrons' direction of motion, not forward along it.

Discussion Questions

A. If the electrical attraction between two pointlike objects at a distance of 1 m is $9\text{x}10^9$ N, why can't we infer that their charges are +1 and –1 C? What further observations would we need to do in order to prove this?

B. If an electrically prepared piece of tape attracts a piece of paper, does that allow us to tell whether the mobile charged particles in the paper are positive or negative, or both?

1.3 Atoms

I was brought up to look at the atom as a nice, hard fellow, red or grey in color according to taste.

Rutherford

Atomism

The Greeks have been kicked around a lot in the last couple of millennia: dominated by the Romans, bullied during the crusades by warlords going to and from the Holy Land, and occupied by Turkey until recently. It's no wonder they prefer to remember their salad days, when their best thinkers came up with concepts like democracy and atoms. Greece is democratic again after a period of military dictatorship, and an atom is proudly pictured on one of their coins. That's why it hurts me to have to say that the ancient Greek hypothesis that matter is made of atoms was pure guesswork. There was no real experimental evidence for atoms, and the 18th-century revival of the atom concept by Dalton owed little to the Greeks other than the name, which means "unsplittable." Subtracting even more cruelly from Greek glory, the name was shown to be inappropriate in 1899 when physicist J.J. Thomson proved experimentally that atoms had even smaller things inside them, which could be extracted. (Thomson called them "electrons.") The "unsplittable" was splittable after all.

But that's getting ahead of our story. What happened to the atom concept in the intervening two thousand years? Educated people continued to discuss the idea, and those who were in favor of it could often use it to give plausible explanations for various facts and phenomena. One fact that was readily explained was conservation of mass. For example, if you mix 1 kg of water with 1 kg of dirt, you get exactly 2 kg of mud, no more and no less. The same is true for the a variety of processes such as freezing of water, fermenting beer, or pulverizing sandstone. If you believed in atoms, conservation of mass made perfect sense, because all these processes could be interpreted as mixing and rearranging atoms, without changing the total number of atoms. Still, this is nothing like a proof that atoms exist.

If atoms did exist, what types of atoms were there, and what distinguished the different types from each other? Was it their sizes, their shapes, their weights, or some other quality? The chasm between the ancient and modern atomisms becomes evident when we consider the wild speculations that existed on these issues until the present century. The ancients decided that there were four types of atoms, earth, water, air and fire; the most

Are you now or have you ever been an atomist?

"You are what you eat." The glib modern phrase more or less assumes the atomic explanation of digestion. After all, digestion was pretty mysterious in ancient times, and premodern cultures would typically believe that eating allowed you to extract some kind of mysterious "life force" from the food. Myths abound to the effect that abstract qualities such as bravery or ritual impurity can enter your body via the food you eat. In contrast to these supernatural points of view, the ancient atomists had an entirely naturalistic interpretation of digestion. The food was made of atoms, and when you digested it you were simply extracting some atoms from it and rearranging them into the combinations required for your own body tissues. The more progressive medieval and renaissance scientists loved this kind of explanation. They were anxious to drive a stake through the heart of Aristotelian physics (and its embellished, Church-friendly version, scholasticism), which in their view ascribed too many occult properties and "purposes" to objects. For instance, the Aristotelian explanation for why a rock would fall to earth was that it was its "nature" or "purpose" to come to rest on the ground.

The seemingly innocent attempt to explain digestion naturalistically, however, ended up getting the atomists in big trouble with the Church. The problem was that the Church's most important sacrament involves eating bread and wine and thereby receiving the supernatural effect of forgiveness of sin. In connection with this ritual, the doctrine of transubstantiation asserts that the blessing of the eucharistic bread and wine literally transforms it into the blood and flesh of Christ. Atomism was perceived as contradicting transubstantiation, since atomism seemed to deny that the blessing could change the nature of the atoms. Although the historical information given in most science textbooks about Galileo represents his run-in with the Inquisition as turning on the issue of whether the earth moves, some historians believe his punishment had more to do with the perception that his advocacy of atomism subverted transubstantiation. (Other issues in the complex situation were Galileo's confrontational style, Pope Urban's military problems, and rumors that the stupid character in Galileo's dialogues was meant to be the Pope.) For a long time, belief in atomism served as a badge of nonconformity for scientists, a way of asserting a preference for natural rather than supernatural interpretations of phenomena. Galileo and Newton's espousal of atomism was an act of rebellion, like later generations' adoption of Darwinism or Marxism.

Another conflict between scholasticism and atomism came from the question of what was between the atoms. If you ask modern people this question, they will probably reply "nothing" or "empty space." But Aristotle and his scholastic successors believed that there could be no such thing as empty space, i.e. a vacuum. That was not an unreasonable point of view, because air tends to rush in to any space you open up, and it wasn't until the renaissance that people figured out how to make a vacuum.

popular view was that they were distinguished by their shapes. Water atoms were spherical, hence water's ability to flow smoothly. Fire atoms had sharp points, which was why fire hurt when it touched one's skin. (There was no concept of temperature until thousands of years later.) The drastically different modern understanding of the structure of atoms was achieved in the course of the revolutionary decade stretching 1895 to 1905. The main purpose of chapters 1 and 2 is to describe those momentous experiments.

Atoms, light, and everything else

Although I tend to ridicule ancient Greek philosophers like Aristotle, let's take a moment to praise him for something. If you read Aristotle's writings on physics (or just skim them, which is all I've done), the most striking thing is how careful he is about classifying phenomena and analyzing relationships among phenomena. The human brain seems to naturally make a distinction between two types of physical phenomena: objects and motion of objects. When a phenomenon occurs that does not immediately present itself as one of these, there is a strong tendency to conceptualize it as one or the other, or even to ignore its existence completely. For instance, physics teachers shudder at students' statements that "the dynamite exploded, and force came out of it in all directions." In these examples, the nonmaterial concept of force is being mentally categorized as if it was a physical substance. The statement that "winding the clock stores motion in the spring" is a miscategorization of potential energy as a form of motion. An example of ignoring the existence of a phenomenon altogether can be elicited by asking people why we need lamps. The typical response that "the lamp illuminates the room so we can see things," ignores the necessary role of light coming into our eyes from the things being illuminated.

If you ask someone to tell you briefly about atoms, the likely response is that "everything is made of atoms," but we've now seen that it's far from obvious which "everything" this statement would properly refer to. For the scientists of the early 1900s who were trying to investigate atoms, this was not a trivial issue of definitions. There was a new gizmo called the vacuum tube, of which the only familiar example today is the picture tube of a TV. In short order, electrical tinkerers had discovered a whole flock of new phenomena that occurred in and around vacuum tubes, and given them picturesque names like "x-rays," "cathode rays," "Hertzian waves," and "N-rays." These were the types of observations that ended up telling us that we know about matter, but fierce controversies ensued over whether these were themselves forms of matter.

Let's bring ourselves up to the level of classification of phenomena employed by physicists in the year 1900. They basically recognized three categories:

- **Matter** has mass, can have kinetic energy, and can travel through a vacuum, transporting its mass and kinetic energy with it. Matter is conserved, both in the sense of conservation of mass and conservation of the number of atoms of each element. Atoms can't occupy the same space as other atoms, so a convenient way to prove something is not a form of matter is to show that it can pass through a solid material, in which the atoms are packed together closely.

- **Light** has no mass, always has energy, and can travel through a vacuum, transporting its energy with it. Two light beams can penetrate through each other and emerge from the collision without being weakened, deflected, or affected in any other way. Light can penetrate certain kinds of matter, e.g. glass.

- The third category is everything that doesn't fit the definition of light or matter. This catch-all category includes, for example, time, velocity, heat, and force.

The chemical elements

How would one find out what types of atoms there were? Today, it doesn't seem like it should have been very difficult to work out an experimental program to classify the types of atoms. For each type of atom, there should be a corresponding element, i.e. a pure substance made out of nothing but that type of atom. Atoms are supposed to be unsplittable, so a substance like milk could not possibly be elemental, since churning it vigorously causes it to split up into two separate substances: butter and whey. Similarly, rust could not be an element, because it can be made by combining two substances: iron and oxygen. Despite its apparent reasonableness, no such program was carried out until the eighteenth century. The ancients presumably did not do it because observation was not universally agreed on as the right way to answer questions about nature, and also because they lacked the necessary techniques or the techniques were the province of laborers with low social status, such as smiths and miners. Premodern European alchemists were hindered by atomism's reputation for subversiveness and a tendency toward mysticism and secretiveness. (The most celebrated challenge facing the alchemists, that of converting lead into gold, is one we now know to be impossible, since lead and gold are both elements.)

$$\frac{m_{He}}{m_H} = 3.97$$

$$\frac{m_{Ne}}{m_H} = 20.01$$

$$\frac{m_{Sc}}{m_H} = 44.60$$

Examples of masses of atoms compared to that of hydrogen. Note how some, but not all, are close to integers.

By 1900, however, chemists had done a reasonably good job of finding out what the elements were. They also had determined the ratios of the different atoms' masses fairly accurately. A typical technique would be to measure how many grams of sodium (Na) would combine with one gram of chlorine (Cl) to make salt (NaCl). (This assumes you've already decided based on other evidence that salt consisted of equal numbers of Na and Cl atoms.) The masses of individual atoms, as opposed to the mass ratios, were known only to within a few orders of magnitude based on indirect evidence, and plenty of physicists and chemists denied that individual atoms were anything more than convenient symbols.

Making sense of the elements

As the information accumulated, the challenge was to find a way of systematizing it; the modern scientist's aesthetic sense rebels against complication. This hodgepodge of elements was an embarrassment. One contemporary observer, William Crookes, described the elements as extending "before us as stretched the wide Atlantic before the gaze of Columbus, mocking, taunting and murmuring strange riddles, which no man has yet been able to solve." It wasn't long before people started recognizing that many atoms' masses were nearly integer multiples of the mass of hydrogen, the lightest element. A few excitable types began speculating that hydrogen was the basic building block, and that the heavier elements were made of clusters of hydrogen. It wasn't long, however, before their parade was rained on by more accurate measurements, which showed that not all of the elements had atomic masses that were near integer multiples of hydrogen, and even the ones that were close to being integer multiples were off by one percent or so.

Chemistry professor Dmitri Mendeleev, preparing his lectures in 1869, wanted to find some way to organize his knowledge for his students to make it more understandable. He wrote the names of all the elements on cards and began arranging them in different ways on his desk, trying to find an arrangement that would make sense of the muddle. The row-and-column scheme he came up with is essentially our modern periodic table. The columns of the modern version represent groups of elements with similar chemical properties, and each row is more massive than the one above it. Going across each row, this almost always resulted in placing the atoms in sequence by weight as well. What made the system significant was its predictive value. There were three places where Mendeleev had to leave gaps in his checkerboard to keep chemically similar elements in the same column. He predicted that elements would exist to fill these gaps, and extrapolated or interpolated from other elements in the same column to predict their numerical properties, such as masses, boiling points, and

1 H																	2 He
3 Li	4 Be											5 B	6 C	7 N	8 O	9 F	10 Ne
11 Na	12 Mg											13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
55 Cs	56 Ba	57 La *	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
87 Fr	88 Ra	89 Ac **	104 Rf	105 Ha	106	107	108	109	110	111	112	113	114	115	116	117	118

*	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb	71 Lu
**	Th	Pa	U	Np	Pu	Am	Cm	Bk	Cf	Es	Fm	Md	No	Lr

A modern periodic table. Elements in the same column have similar chemical properties. The modern atomic numbers, discussed in ch. 2, were not known in Mendeleev's time, since the table could be flipped in various ways.

densities. Mendeleev's professional stock skyrocketed when his three elements (later named gallium, scandium and germanium) were discovered and found to have very nearly the properties he had predicted.

One thing that Mendeleev's table made clear was that mass was not the basic property that distinguished atoms of different elements. To make his table work, he had to deviate from ordering the elements strictly by mass. For instance, iodine atoms are lighter than tellurium, but Mendeleev had to put iodine after tellurium so that it would lie in a column with chemically similar elements.

Direct proof that atoms existed

The success of the kinetic theory of heat was taken as strong evidence that, in addition to the motion of any object as a whole, there is an invisible type of motion all around us: the random motion of atoms within each object. But many conservatives were not convinced that atoms really existed. Nobody had ever seen one, after all. It wasn't until generations after the kinetic theory of heat was developed that it was demonstrated conclusively that atoms really existed and that they participated in continuous motion that never died out.

The smoking gun to prove atoms were more than mathematical abstractions came when some old, obscure observations were reexamined by an unknown Swiss patent clerk named Albert Einstein. A botanist named Brown, using a microscope that was state of the art in 1827, observed tiny grains of pollen in a drop of water on a microscope slide, and found that they jumped around randomly for no apparent reason. Wondering at first if the pollen he'd assumed to be dead was actually alive, he tried looking at particles of soot, and found that the soot particles also moved around. The same results would occur with any small grain or particle suspended in a liquid. The phenomenon came to be referred to as Brownian motion, and its existence was filed away as a quaint and thoroughly unimportant fact, really just a nuisance for the microscopist.

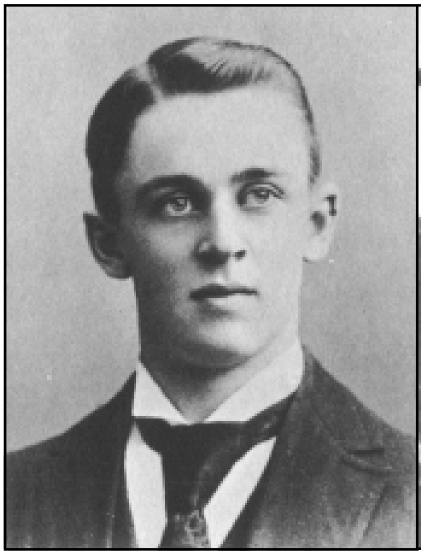
It wasn't until 1906 that Einstein found the correct interpretation for Brown's observation: the water molecules were in continuous random motion, and were colliding with the particle all the time, kicking it in random directions. After all the millennia of speculation about atoms, at last there was solid proof. Einstein's calculations dispelled all doubt, since he was able to make accurate predictions of things like the average distance traveled by the particle in a certain amount of time. (Einstein received the Nobel Prize not for his theory of relativity but for his papers on Brownian motion and the photoelectric effect.)

Discussion Questions

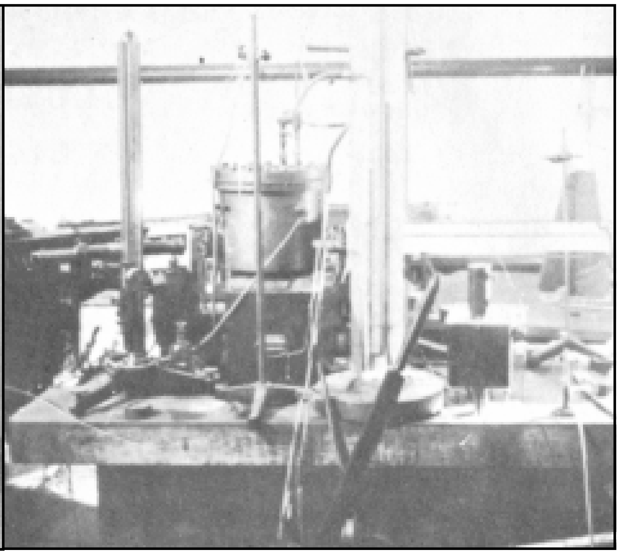
A. Based on Franklin's data, how could one estimate the size of an oil molecule?

B. How could knowledge of the size of an individual aluminum atom be used to infer an estimate of its mass, or vice versa?

C. How could one test Einstein's interpretation by observing Brownian motion at different temperatures?



A young Robert Millikan.

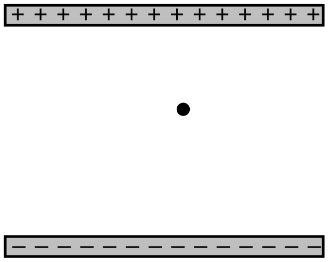


Millikan's workbench, with the oil-drop apparatus.

1.4 Quantization of Charge

Proving that atoms actually existed was a big accomplishment, but demonstrating their existence was different from understanding their properties. Note that the Brown-Einstein observations had nothing at all to do with electricity, and yet we know that matter is inherently electrical, and we have been successful in interpreting certain electrical phenomena in terms of mobile positively and negatively charged particles. Are these particles atoms? Parts of atoms? Particles that are entirely separate from atoms? It is perhaps premature to attempt to answer these questions without any conclusive evidence in favor of the charged-particle model of electricity.

Strong support for the charged-particle model came from a 1911 experiment by physicist Robert Millikan at the University of Chicago. Consider a jet of droplets of perfume or some other liquid made by blowing it through a tiny pinhole. The droplets emerging from the pinhole must be smaller than the pinhole, and in fact most of them are even more microscopic than that, since the turbulent flow of air tends to break them up. Millikan reasoned that the droplets would acquire a little bit of electric charge as they rubbed against the channel through which they emerged, and if the charged-particle model of electricity was right, the charge might be split up among so many minuscule liquid drops that a single drop might have a total charge amounting to an excess of only a few charged particles — perhaps an excess of one positive particle on a certain drop, or an excess of two negative ones on another.



A simplified diagram of Millikan's apparatus.

Millikan's ingenious apparatus, shown in the figure, consisted of two metal plates, which could be electrically charged as needed. He sprayed a cloud of oil droplets into the space between the plates, and selected one drop through a microscope for study. First, with no charge on the plates, he would determine the drop's mass by letting it fall through the air and measuring its terminal velocity, i.e. the velocity at which the force of air friction canceled out the force of gravity. The force of air drag on a slowly moving sphere had already been found by experiment to be bvr^2, where b was a constant, independent of the size of the drop, that was determined by studying spheres that were larger, large enough to be weighed on a balance. Setting the total force equal to zero when the drop is at terminal velocity gives

$$bvr^2 - mg = 0 \quad ,$$

and setting the known density of oil equal to the drop's mass divided by its volume gives a second equation,

$$\rho = \frac{m}{\frac{4}{3}\pi r^3} \quad .$$

Everything in these equations can be measured directly except for *m* and *r*, so these are two equations in two unknowns, which can be solved in order to determine how big the drop is.

Next Millikan charged the metal plates, adjusting the amount of charge so as to exactly counteract gravity and levitate the drop. If, for instance, the drop being examined happened to have a total charge that was negative, then positive charge put on the top plate would attract it, pulling it up, and negative charge on the bottom plate would repel it, pushing it up. (Theoretically only one plate would be necessary, but in practice a two-plate arrangement like this gave electrical forces that were more uniform in strength throughout the space where the oil drops were.) The amount of charge on the plates required to levitate the charged drop gave Millikan a handle on the amount of charge the drop carried. The more charge the drop had, the stronger the electrical forces on it would be, and the less charge would have to be put on the plates to do the trick. Unfortunately, expressing this relationship using Coulomb's law would have been impractical, because it would require a perfect knowledge of how the charge was distributed on each plate, plus the ability to perform vector addition of all the forces being exerted on the drop by all the charges on the plate. Instead, Millikan made use of the fact that the electrical force experienced by a pointlike charged object at a certain point in space is proportional to its charge,

$$\frac{F}{q} = \text{constant} \quad .$$

With a given amount of charge on the plates, this constant could be determined for instance by discarding the oil drop, inserting between the plates a larger and more easily handled object with a known charge on it, and measuring the force with conventional methods. (Millikan actually used a slightly different set of techniques for determining the constant, but the concept is the same.) The amount of force on the actual oil drop had to equal *mg*, since it was just enough to levitate it, and once the calibration constant had been determined, the charge of the drop could then be found based on its previously determined mass.

q (C)	q / 1.64 x10^{-19} C
-1.970×10^{-18}	–12.02
-0.987×10^{-18}	–6.02
-2.773×10^{-18}	–16.93

The table on the left shows a few of the results from Millikan's 1911 paper. (Millikan took data on both negatively and positively charged drops, but in his paper he gave only a sample of his data on negatively charged drops, so these numbers are all negative.) Even a quick look at the data leads to the suspicion that the charges are not simply a series of random numbers. For instance, the second charge is almost exactly equal to half the first one. Millikan explained the observed charges as all being integer multiples of a single number, 1.64x10^{-19} C. In the second column, dividing by this constant gives numbers that are essentially integers, allowing for the random errors present in the experiment. Millikan states in his paper that these results were a

> ...direct and tangible demonstration...of the correctness of the view advanced many years ago and supported by evidence from many sources that all electrical charges, however produced, are exact multiples of one definite, elementary electrical charge, or in other words, that an electrical charge instead of being spread uniformly over the charged surface has a definite granular structure, consisting, in fact, of ... specks, or atoms of electricity, all precisely alike, peppered over the surface of the charged body.

In other words, he had provided direct evidence for the charged-particle model of electricity and against models in which electricity was described as some sort of fluid. The basic charge is notated e, and the best modern value is e=1.60x10^{-19} C

The word "*quantized*" is used in physics to describe a quantity that can only have certain numerical values, and cannot have any of the values between those. In this language, we would say that Millikan discovered that charge is quantized. The charge e is referred to as the quantum of charge.

Historical Note: Millikan's Fraud

Every undergraduate physics textbook I've ever seen fails to note the well documented fact that although Millikan's conclusions were correct, he was guilty of scientific fraud. His technique was difficult and painstaking to perform, and his original notebooks, which have been preserved, show that the data were far less perfect than he claimed in his published scientific papers. In his publications, he stated categorically that every single oil drop observed had had a charge that was a multiple of e, with no exceptions or omissions. But his notebooks are replete with notations such as "beautiful data, keep," and "bad run, throw out." Millikan, then, appears to have earned his Nobel Prize by advocating a correct position with dishonest descriptions of his data.

Why do textbook authors fail to mention Millikan's fraud? It's an interesting sociological question. I don't think it's because of a lack of space: most of these texts take a slavishly historical approach in introducing modern physics, devoting entire sections to discussions of topics like black body radiation, which are historically important but not particularly helpful to students. It may be that they think students are too unsophisticated to correctly evaluate the implications of the fact that scientific fraud has sometimes existed and even been rewarded by the scientific establishment. Maybe they are afraid students will reason that fudging data is OK, since Millikan got the Nobel Prize for it. But falsifying history in the name of encouraging truthfulness is more than a little ironic. English teachers don't edit Shakespeare's tragedies so that the bad characters are always punished and the good ones never suffer!

Another possible explanation is simply a lack of originality; it's possible that some venerated textbook was uncritical of Millikan's fraud, and later authors simply followed suit. Biologist Stephen Jay Gould has written an essay tracing an example of how authors of biology textbooks tend to follow a certain traditional treatment of a topic, using the giraffe's neck to discuss the nonheritability of acquired traits. Yet another interpretation is that scientists derive status from their popular images as impartial searchers after the truth, and they don't want the public to realize how human and imperfect they can be.

Self-Check

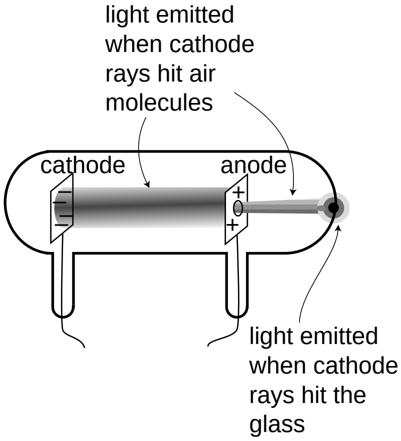
Is money quantized? What is the quantum of money?

1.5 The Electron

Cathode rays

Nineteenth-century physicists didn't just spend a lot of time coming up with wild, random speculations about things like atoms being made out of knots. They also spent a lot of time trying to come up with wild, random ways to play with electricity. The best experiments of this kind were the ones that made big sparks or pretty colors of light.

One such parlor trick was the cathode ray. To produce it, you first had to hire a good glassblower and find a good vacuum pump. The glassblower would create a hollow tube and embed two pieces of metal in it, called the electrodes, which were connected to the outside via metal wires passing through the glass. Before letting him seal up the whole tube, you would hook it up to a vacuum pump, and spend several hours huffing and puffing away at the pump's hand crank to get a good vacuum inside. Then, while you were still pumping on the tube, the glassblower would melt the glass and seal the whole thing shut. Finally, you would put a large amount of positive charge on one wire and a large amount of negative charge on the other. Metals have the property of letting charge move through them easily, so the charge deposited on one of the wires would quickly spread out because of the repulsion of each part of it for every other part. This spreading-out process would result in nearly all the charge ending up in the electrodes, where there is more room to spread out than there is in the wire. For obscure historical reasons a negative electrode is called a cathode and a positive one is an anode.


light emitted when cathode rays hit air molecules
cathode
anode
−
−
+
+
light emitted when cathode rays hit the glass

The figure shows the light-emitting stream that was observed. If, as shown in this figure, a hole was made in the anode, the beam would extend on through the hole until it hit the glass. Drilling a hole in the cathode, however would not result in any beam coming out on the left side, and this indicated that the stuff, whatever it was, was coming from the cathode. The rays were therefore christened "cathode rays." (The terminology is still used today in the term "cathode ray tube" or "CRT" for the picture tube of a TV or computer monitor.)

Yes. In U.S. currency, the quantum of money is the penny.

Were cathode rays a form of light, or of matter?

Were cathode rays a form of light, or matter? At first no one really cared what they were, but as their scientific importance became more apparent, the light-versus-matter issue turned into a controversy along nationalistic lines, with the Germans advocating light and the English holding out for matter. The supporters of the material interpretation imagined the rays as consisting of a stream of atoms ripped from the substance of the cathode.

One of our defining characteristics of matter is that material objects cannot pass through each other. Experiments showed that cathode rays could penetrate at least some small thickness of matter, such as a metal foil a tenth of a millimeter thick, would imply they were a form of light.

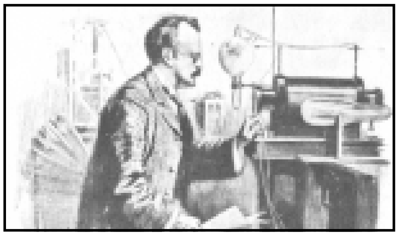
Other experiments, however, pointed to the contrary conclusion. Light is a wave phenomenon, and one distinguishing property of waves is demonstrated speaking into one end of a paper towel roll. The sound waves do not emerge from the other end of the tube as a focused beam. Instead, they begin spreading out in all directions as soon as they emerge. This shows that waves do not necessarily travel in straight lines. If a piece of metal foil in the shape of a star or a cross was placed in the way of the cathode ray, then a "shadow" of the same shape would appear on the glass, showing that the rays traveled in straight lines. This straight-line motion suggested that they were a stream of small particles of matter.

These observations were inconclusive, so what was really needed was a determination of whether the rays had mass and weight. The trouble was that cathode rays could not simply be collected in a cup and put on a scale. When the cathode ray tube is in operation, one does not observe any loss of material from the cathode, or any crust being deposited on the anode.

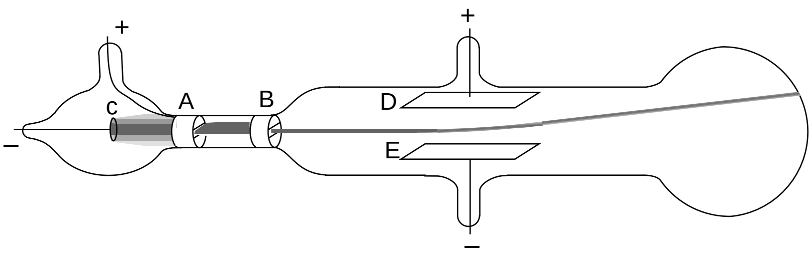
Nobody could think of a good way to weigh cathode rays, so the next most obvious way of settling the light/matter debate was to check whether the cathode rays possessed electrical charge. Light was known to be uncharged. If the cathode rays carried charge, they were definitely matter and not light, and they were presumably being made to jump the gap by the simultaneous repulsion of the negative charge in the cathode and attraction of the positive charge in the anode. The rays would overshoot the anode because of their momentum. (Although electrically charged particles do not normally leap across a gap of vacuum, very large amounts of charge were being used, so the forces were unusually intense.)

Thomson's experiments

Physicist J.J. Thomson at Cambridge carried out a series of definitive experiments on cathode rays around the year 1897. By turning them slightly off course with electrical forces, as shown in the figure, he showed that they were indeed electrically charged, which was strong evidence that they were material. Not only that, but he proved that they had mass, and measured the ratio of their mass to their charge, m/q. Since their mass was not zero, he concluded that they were a form of matter, and presumably made up of a stream of microscopic, negatively charged particles. When Millikan published his results fourteen years later, it was reasonable to assume that the charge of one such particle equaled minus one fundamental charge, $q=-e$, and from the combination of Thomson's and Millikan's results one could therefore determine the mass of a single cathode ray particle.



J.J. Thomson in the lab.



Thomson's experiment proving cathode rays had electric charge (redrawn from his original paper). The cathode, c, and anode, A, are as in any cathode ray tube. The rays pass through a slit in the anode, and a second slit, B, is interposed in order to make the beam thinner and eliminate rays that were not going straight. Charging plates D and E shows that cathode rays have charge: they are attracted toward the positive plate D and repelled by the negative plate E.

The basic technique for determining *m*/*q* was simply to measure the angle through which the charged plates bent the beam. The electric force acting on a cathode ray particle while it was between the plates would be proportional to its charge,

$$F_{\text{elec}} = \left(\text{known constant}\right) \cdot q \quad .$$

Application of Newton's second law, *a*=*F*/*m*, would allow *m*/*q* to be determined:

$$\frac{m}{q} = \frac{\text{known constant}}{a}$$

There was just one catch. Thomson needed to know the cathode ray particles' velocity in order to figure out their acceleration. At that point, however, nobody had even an educated guess as to the speed of the cathode rays produced in a given vacuum tube. The beam appeared to leap across the vacuum tube practically instantaneously, so it was no simple matter of timing it with a stopwatch!

Thomson's clever solution was to observe the effect of both electric and magnetic forces on the beam. The magnetic force exerted by a particular magnet would depend on both the cathode ray's charge and its velocity:

$$F_{\text{mag}} = \left(\text{known constant \#2}\right) \cdot qv$$

Thomson played with the electric and magnetic forces until either one would produce an equal effect on the beam, allowing him to solve for the velocity,

$$v = \frac{\left(\text{known constant}\right)}{\left(\text{known constant \#2}\right)} \quad .$$

Knowing the velocity (which was on the order of 10% of the speed of light for his setup), he was able to find the acceleration and thus the mass-to-charge ratio *m*/*q*. Thomson's techniques were relatively crude (or perhaps more charitably we could say that they stretched the state of the art of the time), so with various methods he came up with *m*/*q* values that ranged over about a factor of two, even for cathode rays extracted from a cathode made of a single material. The best modern value is m/q= 5.69x10 $^{-12}$ kg/C, which is consistent with the low end of Thomson's range.

The cathode ray as a subatomic particle: the electron

What was significant about Thomson's experiment was not the actual numerical value of *m*/*q*, however, so much as the fact that, combined with Millikan's value of the fundamental charge, it gave a mass for the cathode ray particles that was thousands of times smaller than the mass of even the lightest atoms. Even without Millikan's results, which were 14 years in the future, Thomson recognized that the cathode rays' *m*/*q* was thousands of times smaller than the *m*/*q* ratios that had been measured for electrically charged atoms in chemical solutions. He correctly interpreted this as evidence that the cathode rays were smaller building blocks — he called them *electrons* — out of which atoms themselves were formed. This was an extremely radical claim, coming at a time when atoms had not yet been proven to exist! Even those who used the word "atom" often considered them no more than mathematical abstractions, not literal objects. The idea of searching for structure inside of "unsplittable" atoms was seen by some as lunacy, but within ten years Thomson's ideas had been amply verified by many more detailed experiments.

Discussion Questions

A. Thomson started to become convinced during his experiments that the "cathode rays" observed coming from the cathodes of vacuum tubes were building blocks of atoms — what we now call electrons. He then carried out observations with cathodes made of a variety of metals, and found that *m*/*q* was roughly the same in every case, considering his limited accuracy. Given his suspicion, why did it make sense to try different metals? How would the consistent values of *m*/*q* serve to test his hypothesis?

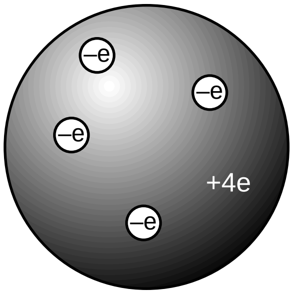
B. My students have frequently asked whether the *m*/*q* that Thomson measured was the value for a single electron, or for the whole beam. Can you answer this question?

C. Thomson found that the *m*/*q* of an electron was thousands of times smaller than that of charged atoms in chemical solutions. Would this imply that the electrons had more charge? Less mass? Would there be no way to tell? Explain. Remember that Millikan's results were still many years in the future, so *q* was unknown.

D. Can you guess any practical reason why Thomson couldn't just let one electron fly across the gap before disconnecting the battery and turning off the beam, and then measure the amount of charge deposited on the anode, thus allowing him to measure the charge of a single electron directly?

E. Why is it not possible to determine *m* and *q* themselves, rather than just their ratio, by observing electrons' motion in electric and magnetic fields?

1.6 The Raisin Cookie Model of the Atom



The raisin cookie model of the atom with four units of charge, which we now know to be beryllium.

Based on his experiments, Thomson proposed a picture of the atom which became known as the raisin cookie model. In the neutral atom shown in the figure, there are four electrons with a total charge of -4*e*, sitting in a sphere (the "cookie") with a charge of +4*e* spread throughout it. It was known that chemical reactions could not change one element into another, so in Thomson's scenario, each element's cookie sphere had a permanently fixed radius, mass, and positive charge, different from those of other elements. The electrons, however, were not a permanent feature of the atom, and could be tacked on or pulled out to make charged ions. Although we now know, for instance, that a neutral atom with four electrons is the element beryllium, scientists at the time did not know how many electrons the various neutral atoms possessed.

This model is clearly different from the one you've learned in grade school or through popular culture, where the positive charge is concentrated in a tiny nucleus at the atom's center. An equally important change in ideas about the atom has been the realization that atoms and their constituent subatomic particles behave entirely differently from objects on the human scale. For instance, we'll see later that an electron can be in more than one place at one time. The raisin cookie model was part of a long tradition of attempts to make mechanical models of phenomena, and Thomson and his contemporaries never questioned the appropriateness of building a mental model of an atom as a machine with little parts inside. Today, mechanical models of atoms are still used (for instance the tinker-toy-style molecular modeling kits like the ones used by Watson and Crick to figure out the double helix structure of DNA), but scientists realize that the physical objects are only aids to help our brains' symbolic and visual processes think about atoms.

Although there was no clear-cut experimental evidence for many of the details of the raisin cookie model, physicists went ahead and started working out its implications. For instance, suppose you had a four-electron atom. All four electrons would be repelling each other, but they would also all be attracted toward the center of the "cookie" sphere. The result should be some kind of stable, symmetric arrangement in which all the forces canceled out. People sufficiently clever with math soon showed that the electrons in a four-electron atom should settle down at the vertices of a pyramid with one less side than the Egyptian kind, i.e. a regular tetrahedron. This deduction turns out to be wrong because it was based on incorrect features of the model, but the model also had many successes, a few of which we will now discuss.

Example: flow of electrical charge in wires

One of my former students was the son of an electrician who had become an electrician himself. He related to me how his father had remained refused to believe all his life that electrons really flowed through wires. If they had, he reasoned, the metal would have gradually become more and more damaged, eventually crumbling to dust.

His opinion is not at all unreasonable based on the fact that electrons are material particles, and that matter cannot normally pass through matter without making a hole through it. Nineteenth-century physicists would have shared his objection to a charged-particle model of the flow of electrical charge. In the raisin-cookie model, however, the electrons are very low in mass, and therefore presumably very small in size as well. It is not surprising that they can slip between the atoms without damaging them.

Example: flow of electrical charge across cell membranes

Your nervous system is based on signals carried by charge moving from nerve cell to nerve cell. Your body is essentially all liquid, and atoms in a liquid are mobile. This means that, unlike the case of charge flowing in a solid wire, entire charged atoms can flow in your nervous system

Example: emission of electrons in a cathode ray tube

Why do electrons detach themselves from the cathode of a vacuum tube? Certainly they are encouraged to do so by the repulsion of the negative charge placed on the cathode and the attraction from the net positive charge of the anode, but these are not strong enough to rip electrons out of atoms by main force — if they were, then the entire apparatus would have been instantly vaporized as every atom was simultaneously ripped apart!

The raisin cookie model leads to a simple explanation. We know that heat is the energy of random motion of atoms. The atoms in any object are therefore violently jostling each other all the time, and a few of these collisions are violent enough to knock electrons out of atoms. If this occurs near the surface of a solid object, the electron may can come loose. Ordinarily, however, this loss of electrons is a self-limiting process; the loss of electrons leaves the object with a net positive charge, which attracts the lost sheep home to the fold. (For objects immersed in air rather than vacuum, there will also be a balanced exchange of electrons between the air and the object.)

This interpretation explains the warm and friendly yellow glow of the vacuum tubes in an antique radio. To encourage the emission of electrons from the vacuum tubes' cathodes, the cathodes are intentionally warmed up with little heater coils.

Discussion Questions

A. Today many people would define an ion as an atom (or molecule) with missing electrons or extra electrons added on. How would people have defined the word "ion" before the discovery of the electron?

B. Since electrically neutral atoms were known to exist, there had to be positively charged subatomic stuff to cancel out the negatively charged electrons in an atom. Based on the state of knowledge immediately after the Millikan and Thomson experiments, was it possible that the positively charged stuff had an unquantized amount of charge? Could it be quantized in units of +e? In units of +2e? In units of +5/7e?

Summary

Selected Vocabulary

atom	the basic unit of one of the chemical elements
molecule	a group of atoms stuck together
electrical force	one of the fundamental forces of nature; a noncontact force that can be either repulsive or attractive
charge	a numerical rating of how strongly an object participates in electrical forces
coulomb (C)	the unit of electrical charge
ion	an electrically charged atom or molecule
cathode ray	the mysterious ray that emanated from the cathode in a vacuum tube; shown by Thomson to be a stream of particles smaller than atoms
electron	Thomson's name for the particles of which a cathode ray was made
quantized	describes quantity such as money or electrical charge, that can only exist in certain amounts

Notation

q	charge
e	the quantum of charge

Summary

All the forces we encounter in everyday life boil down to two basic types: gravitational forces and electrical forces. A force such as friction or a "sticky force" arises from electrical forces between individual atoms.

Just as we use the word "mass" to describe how strongly an object participates in gravitational forces, we use the word "charge" for the intensity of its electrical forces. There are two types of charge. Two charges of the same type repel each other, but objects whose charges are different attract each other. Charge is measured in units of coulombs (C).

Mobile charged particle model: A great many phenomena are easily understood if we imagine matter as containing two types of charged particles, which are at least partially able to move around.

Positive and negative charge: Ordinary objects that have not been specially prepared have both types of charge spread evenly throughout them in equal amounts. The object will then tend not to exert electrical forces on any other object, since any attraction due to one type of charge will be balanced by an equal repulsion from the other. (We say "tend not to" because bringing the object near an object with unbalanced amounts of charge could cause its charges to separate from each other, and the force would no longer cancel due to the unequal distances.) It therefore makes sense to describe the two types of charge using positive and negative signs, so that an unprepared object will have zero *total* charge.

The Coulomb force law states that the magnitude of the electrical force between two charged particles is given by $|\boldsymbol{F}| = k\,|q_1|\,|q_2|\,/\,r^2$.

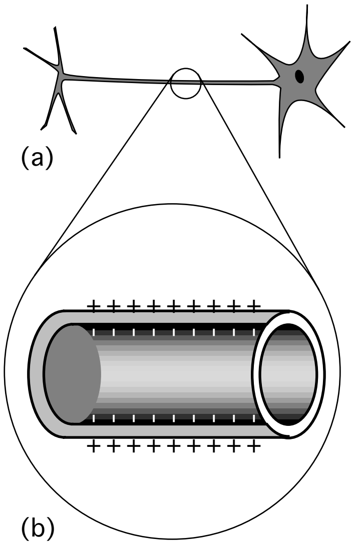
Conservation of charge: An even more fundamental reason for using positive and negative signs for charge is that with this definition the total charge of a closed system is a conserved quantity.

Quantization of charge: Millikan's oil drop experiment showed that the total charge of an object could only be an integer multiple of a basic unit of charge (e). This supported the idea the the "flow" of electrical charge was the motion of tiny particles rather than the motion of some sort of mysterious electrical fluid.

Einstein's analysis of Brownian motion was the first definitive proof of the existence of atoms. Thomson's experiments with vacuum tubes demonstrated the existence of a new type of microscopic particle with a very small ratio of mass to charge. Thomson correctly interpreted these as building blocks of matter even smaller than atoms: the first discovery of subatomic particles. These particles are called electrons.

The above experimental evidence led to the first useful model of the interior structure of atoms, called the raisin cookie model. In the raisin cookie model, an atom consists of a relatively large, massive, positively charged sphere with a certain number of negatively charged electrons embedded in it.

Homework Problems



Pronblem 1. (a) Realistic picture of a neuron. (b) Simplified diagram of one segment of the tail (axon).

1. The figure shows a neuron, which is the type of cell your nerves are made of. Neurons serve to transmit sensory information to the brain, and commands from the brain to the muscles. All this data is transmitted electrically, but even when the cell is resting and not transmitting any information, there is a layer of negative electrical charge on the inside of the cell membrane, and a layer of positive charge just outside it. This charge is in the form of various ions dissolved in the interior and exterior fluids. Why would the negative charge remain plastered against the inside surface of the membrane, and likewise why doesn't the positive charge wander away from the outside surface?

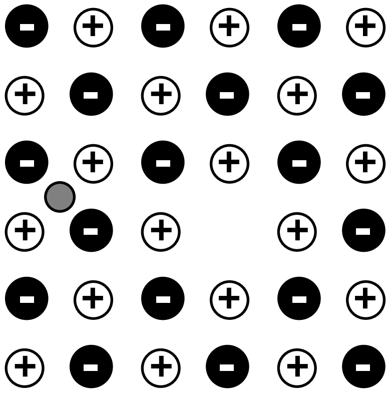
2. Use the nutritional information on some packaged food to make an order-of-magnitude estimate of the amount of chemical energy stored in one atom of food, in units of Joules. Assume that a typical atom has a mass of 10^{-26} kg. This constitutes a rough estimate of the amounts of energy there are on the atomic scale. [See chapter 1 of book 1, Newtonian Physics, for help on how to do order-of-magnitude estimates. Note that a nutritional "calorie" is really a kilocalorie.]

3. (a) Recall that the potential energy of two gravitationally interacting spheres is given by $PE = -Gm_1m_2/r$, where r is the center-to-center distance. What would be the analogous equation for two electrically interacting spheres? Justify your choice of a plus or minus sign on physical grounds, considering attraction and repulsion. (b) Use this expression to estimate the energy required to pull apart a raisin-cookie atom of the one-electron type, assuming a radius of 10^{-10} m. (c) Compare this with the result of the previous problem.

4. A neon light consists of a long glass tube full of neon, with metal caps on the ends. Positive charge is placed on one end of the tube, and negative charge on the other. The electric forces generated can be strong enough to strip electrons off of a certain number of neon atoms. Assume for simplicity that only one electron is ever stripped off of any neon atom. When an electron is stripped off of an atom, both the electron and the neon atom (now an ion) have electric charge, and they are accelerated by the forces exerted by the charged ends of the tube. (They do not feel any significant forces from the other ions and electrons within the tube, because only a tiny minority of neon atoms ever gets ionized.) Light is finally produced when ions are reunited with electrons. Compare the magnitudes and directions of the accelerations of the electrons and ions. (A numerical answer is not necessary.)

5. If you put two hydrogen atoms near each other, they will feel an attractive force, and they will pull together to form a molecule. (Molecules consisting of two hydrogen atoms are the normal form of hydrogen gas.) Why do they feel a force if they are near each other, since each is electrically neutral? Shouldn't the attractive and repulsive forces all cancel out exactly?

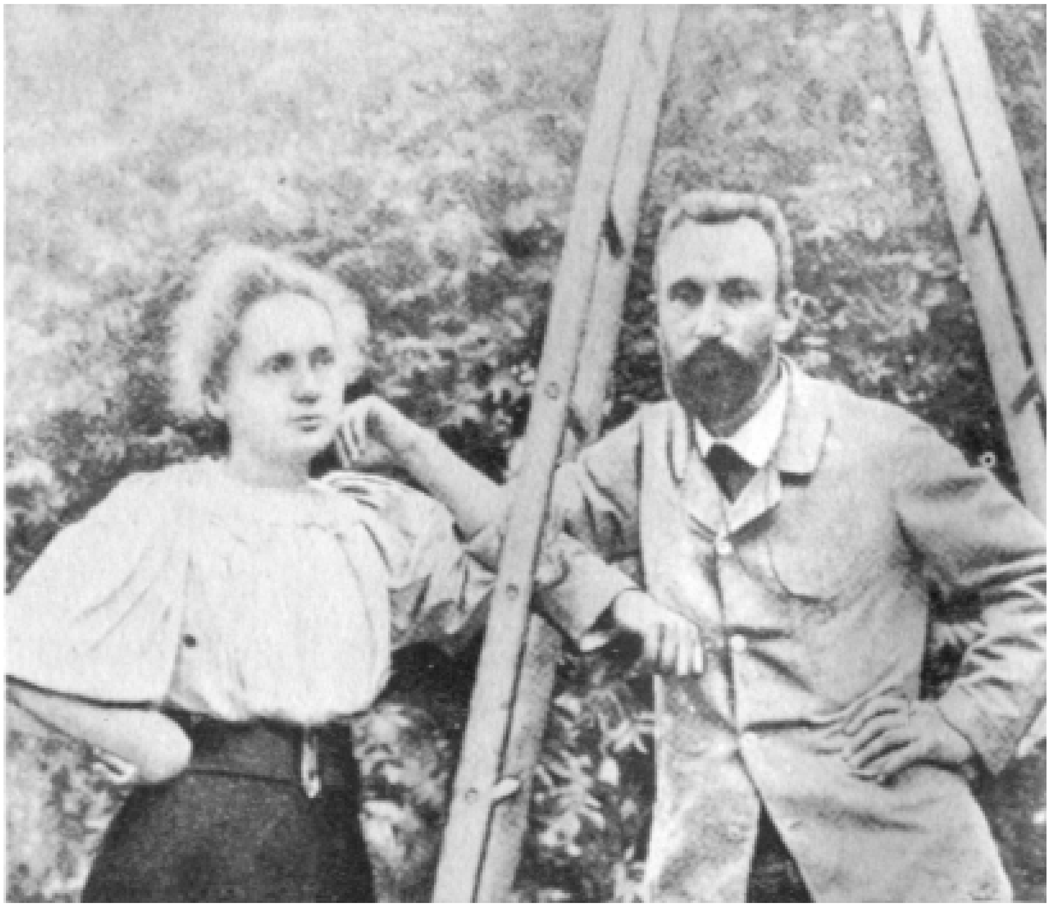
S A solution is given in the back of the book. ★ A difficult problem.
✓ A computerized answer check is available. ∫ A problem that requires calculus.



Problem 6.

6 ✓. The figure shows one layer of the three-dimensional structure of a salt crystal. The atoms extend much farther off in all directions, but only a six-by-six square is shown here. The larger circles are the chlorine ions, which have charges of *-e*. The smaller circles are sodium ions, with charges of *+e*. The distance between neighboring ions is about 0.3 nm. Real crystals are never perfect, and the crystal shown here has two defects: a missing atom at one location, and an extra lithium atom, shown as a grey circle, inserted in one of the small gaps. If the lithium atom has a charge of *+e*, what is the direction and magnitude of the force on it? Assume there are no other defects nearby in the crystal besides the two shown here. [Hints: The force on the lithium ion is the vector sum of all the forces of all the quadrillions of sodium and chlorine atoms, which would obviously be too laborious to calculate. Approach the problem by comparing with the vector sum of the forces that would be exerted on the lithium if the hole nearby was occupied by a chlorine ion.]

7 ✓. The Earth and Moon are bound together by gravity. If, instead, the force of attraction were the result of each having a charge of the same magnitude but opposite in sign, find the quantity of charge that would have to be placed on each to produce the required force.



Marie and Pierre Curie were the first to purify radium in significant quantities. Radium's intense radioactivity made possible the experiments that led to the modern planetary model of the atom, in which electrons orbit a nucleus made of protons and neutrons.

2 The Nucleus

2.1 Radioactivity

Becquerel's discovery of radioactivity

How did physicists figure out that the raisin cookie model was incorrect, and that the atom's positive charge was concentrated in a tiny, central nucleus? The story begins with the discovery of radioactivity by the French chemist Becquerel. Up until radioactivity was discovered, all the processes of nature were thought to be based on chemical reactions, which were rearrangements of combinations of atoms. Atoms exert forces on each other when they are close together, so sticking or unsticking them would either release or store potential energy. That energy could be converted to and from other forms, as when a plant uses the energy in sunlight to make sugars and carbohydrates, or when a small child eats sugar, releasing the energy in the form of kinetic energy.

Becquerel discovered a process that seemed to release energy from an unknown new source that was not chemical. Becquerel, whose father and grandfather had also been physicists, spent the first twenty years of his professional life as a successful civil engineer, teaching physics only on a part-time basis. He was awarded the chair of physics at the Musée d'Histoire Naturelle in Paris after the death of his father, who had previously occupied it. Having now a significant amount of time to devote to physics, he began studying the interaction of light and matter. He became interested in the phenomenon of phosphorescence, in which a substance absorbs energy from light, then releases the energy via a glow that only gradually goes away. One of the substances he investigated was a uranium compound, the salt $UKSO_5$. One day in 1896, cloudy weather interfered with his plan to expose this substance to sunlight in order to observe its fluorescence. He stuck it in a drawer, coincidentally on top of a blank photographic plate — the old-fashioned glass-backed counterpart of the modern plastic roll of film. The plate had been carefully wrapped, but several days later when Becquerel checked it in the darkroom before using it, he found that it was ruined, as if it had been completely exposed to light.

History provides many examples of scientific discoveries that occurred in this way: an alert and inquisitive mind decides to investigate a phenomenon that most people would not have worried about explaining. He first determined by further experiments that the effect was produced by the uranium salt, despite a thick wrapping of paper around the plate that blocked out all light. He tried a variety of compounds, and found that it was the uranium that did it: the effect was produced by any uranium compound, but not by any compound that didn't include uranium atoms. The effect could be at least partially blocked by a sufficient thickness of metal, and he was able to produce silhouettes of coins by interposing them between the uranium and the plate. This indicated that the effect traveled in a straight line., so that it must have been some kind of ray rather than, e.g., the seepage of chemicals through the paper. He used the word "radiations," since the effect radiated out from the uranium salt.

At this point Becquerel still believed that the uranium atoms were absorbing energy from light and then gradually releasing the energy in the form of the mysterious rays, and this was how he presented it in his first published lecture describing his experiments. Interesting, but not earth-shattering. But he then tried to determine how long it took for the uranium to use up all the energy that had been supposedly been stored in it by light, and he found that it never seemed to become inactive, no matter how long he waited. Not only that, but a sample that had been exposed to intense sunlight for a whole afternoon was no more or less effective than a sample that had always been kept inside. Was this a violation of conservation of energy? If the energy didn't come from exposure to light, where did it come from?

Three kinds of “radiations”

Even though turn-of-the-century physicists knew little about Becquerel’s new phenomenon, they decided to go ahead and study how it behaved. Becquerel had already shown that the radioactivity could penetrate through cloth and paper, so the first obvious thing to do was to investigate in more detail what thickness of material the radioactivity could get through. They soon learned that a certain fraction of the radioactivity’s intensity would be eliminated by even a few inches of air, but the remainder was not eliminated by passing through more air. Apparently, then, the radioactivity was a mixture of more than one type, of which one was blocked by air. They then found that of the part that could penetrate air, a further fraction could be eliminated by a piece of paper or a very thin metal foil. What was left after that, however, was a third, extremely penetrating type, some of whose intensity would still remain even after passing through a brick wall. They decided that this showed there were three types of radioactivity, and without having the faintest idea of what they really were, they made up names for them. The least penetrating type was arbitrarily labeled α (alpha), the first letter of the alphabet, and so on through β (beta) and finally γ (gamma), the third Greek letter, for the most penetrating type.

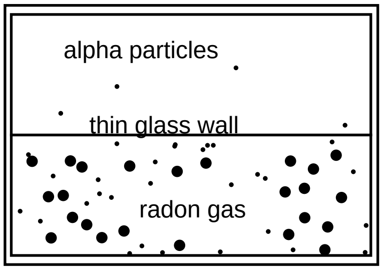
Radium: a more intense source of radioactivity

The measuring devices used to detect radioactivity were crude: photographic plates or even human eyeballs (radioactivity makes flashes of light inside the jelly-like fluid inside the eye, which can be seen by the eyeball’s owner if it is otherwise very dark). Because the ways of detecting radioactivity were so crude and insensitive, further progress was hindered by the fact that the amount of radioactivity emitted by uranium was not really very great. The vital contribution of physicist/chemist Marie Curie and her husband Pierre was to discover the element radium, and to purify and isolate significant quantities it. Radium emits about a million times more radioactivity per unit mass than uranium, making it possible to do the experiments that were needed to learn the true nature of radioactivity. The dangers of radioactivity to human health were then unknown, and Marie died of leukemia thirty years later. (Pierre was run over and killed by a horsecart.)

Tracking down the nature of alphas, betas, and gammas

As radium was becoming available, an apprentice scientist named Ernest Rutherford arrived in England from his native New Zealand and began studying radioactivity at the Cavendish Laboratory. The young colonial's first success was to measure the mass-to-charge ratio of beta rays. The technique was essentially the same as the one Thomson had used to measure the charge-to-mass ratio of cathode rays by measuring their deflections in electric and magnetic fields. The only difference was that instead of the cathode of a vacuum tube, a nugget of radium was used to supply the beta rays. Not only was the technique the same, but so was the result. Beta rays had the same m/q ratio as cathode rays, which suggested they were one and the same. Nowadays, it would make sense simply to use the term "electron," and avoid the archaic terms "cathode ray" and "beta particle," but the old terms are still widely used, and it is unfortunately necessary for physics students to memorize all three names for the same thing.

At first, it seemed that neither alphas or gammas could be deflected in electric or magnetic fields, making it appear that neither was electrically charged. But soon Rutherford obtained a much more powerful magnet, and was able to use it to deflect the alphas but not the gammas. The alphas had a much larger value of m/q than the betas (about 4000 times greater), which was why they had been so hard to deflect. Gammas are uncharged, and were later found to be a form of light.



A simplified version of Rutherford's 1908 experiment, showing that alpha particles were doubly ionized helium atoms.

The m/q ratio of alpha particles turned out to be the same as those of two different types of ions, He^{++} (a helium atom with two missing electrons) and H_2^+ (two hydrogen atoms bonded into a molecule, with one electron missing) , so it seemed likely that they were one or the other of those. The diagram shows a simplified version of Rutherford's ingenious experiment proving that they were He^{++} ions. The gaseous element radon, an alpha emitter, was introduced into one half of a double glass chamber. The glass wall dividing the chamber was made extremely thin, so that some of the rapidly moving alpha particles were able to penetrate it. The other chamber, which was initially evacuated, gradually began to accumulate a population of alpha particles (which would quickly pick up electrons from their surroundings and become electrically neutral). Rutherford then determined that it was helium gas that had appeared in the second chamber. Thus alpha particles were proved to be He^{++} ions. The nucleus was yet to be discovered, but in modern terms, we would describe a He^{++} ion as the nucleus of a He atom.

To summarize, here are the three types of radiation emitted by radioactive elements, and their descriptions in modern terms:

alpha particle	stopped by a few inches of air	He nucleus
beta particle	stopped by a piece of paper	electron
gamma ray	penetrates thick shielding	a type of light

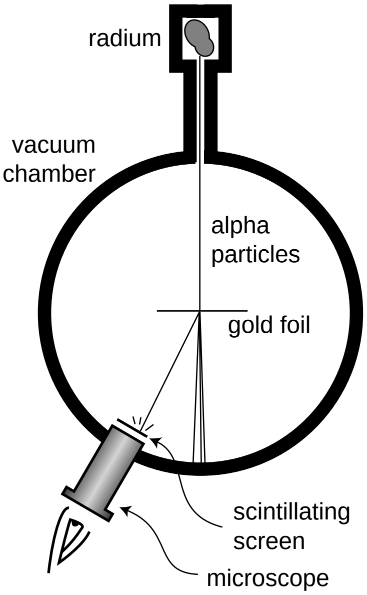
Discussion Question

Most sources of radioactivity emit alphas, betas, and gammas, not just one of the three. In the radon experiment, how did Rutherford know that he was studying the alphas?

2.2 The Planetary Model of the Atom

The stage was now set for the unexpected discovery that the positively charged part of the atom was a tiny, dense lump at the atom's center rather than the "cookie dough" of the raisin cookie model. By 1909, Rutherford was an established professor, and had students working under him. For a raw undergraduate named Marsden, he picked a research project he thought would be tedious but straightforward.

It was already known that although alpha particles would be stopped completely by a sheet of paper, they could pass through a sufficiently thin metal foil. Marsden was to work with a gold foil only 1000 atoms thick. (The foil was probably made by evaporating a little gold in a vacuum chamber so that a thin layer would be deposited on a glass microscope slide. The foil would then be lifted off the slide by submerging the slide in water.)

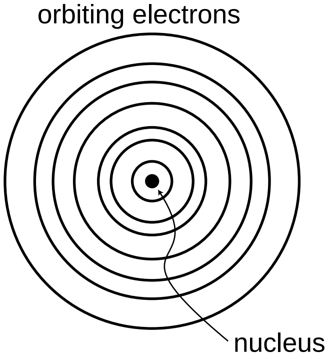


Rutherford had already determined in his previous experiments the speed of the alpha particles emitted by radium, a fantastic 1.5x10^7 m/s. The experimenters in Rutherford's group visualized them as very small, very fast cannonballs penetrating the "cookie dough" part of the big gold atoms. A piece of paper has a thickness of a hundred thousand atoms or so, which would be sufficient to stop them completely, but crashing through a thousand would only slow them a little and turn them slightly off of their original paths.

Marsden's supposedly ho-hum assignment was to use the apparatus shown in the figure to measure how often alpha particles were deflected at various angles. A tiny lump of radium in a box emitted alpha particles, and a thin beam was created by blocking all the alphas except those that happened to pass out through a tube. Typically deflected in the gold by only a small amount, they would reach a screen very much like the screen of a TV's picture tube, which would make a flash of light when it was hit. Here is the first example we have encountered of an experiment in which a beam of particles is detected one at a time. This was possible because each alpha particle carried so much kinetic energy; they were moving at about the same speed as the electrons in the Thomson experiment, but had ten thousand times more mass.

Marsden sat in a dark room, watching the apparatus hour after hour and recording the number of flashes with the screen moved to various angles. The rate of the flashes was highest when he set the screen at an angle close to the line of the alphas' original path, but if he watched an area farther off to the side, he would also occasionally see an alpha that had been deflected through a larger angle. After seeing a few of these, he got the crazy idea of moving the screen to see if even larger angles ever occurred, perhaps even angles larger than 90 degrees.

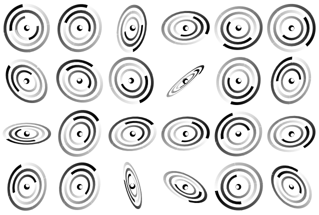
The crazy idea worked: a few alpha particles were deflected through angles of up to 180 degrees, and the routine experiment had become an epoch-making one. Rutherford said, "We have been able to get some of the alpha particles coming backwards. It was almost as incredible as if you fired a 15-inch shell at a piece of tissue paper and it came back and hit you." Explanations were hard to come by in the raisin cookie model. What intense electrical forces could have caused some of the alpha particles, moving at such astronomical speeds, to change direction so drastically? Since each gold atom was electrically neutral, it would not exert much force on an alpha particle outside it. True, if the alpha particle was very near to or inside of a particular atom, then the forces would not necessarily cancel out perfectly; if the alpha particle happened to come very close to a particular electron, the $1/r^2$ form of the Coulomb force law would make for a very strong force. But Marsden and Rutherford knew that an alpha particle was 8000 times more massive than an electron, and it is simply not possible for a more massive object to rebound backwards from a collision with a less massive object while conserving momentum and energy. It might be possible in principle for a particular alpha to follow a path that took it very close to one electron, and then very close to another electron, and so on, with the net result of a large deflection, but careful calculations showed that such multiple "close encounters" with electrons would be millions of times too rare to explain what was actually observed.



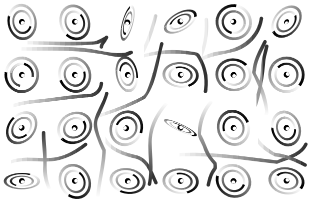
The planetary model of the atom.

At this point, Rutherford and Marsden dusted off an unpopular and neglected model of the atom, in which all the electrons orbited around a small, positively charged core or "nucleus," just like the planets orbiting around the sun. All the positive charge and nearly all the mass of the atom would be concentrated in the nucleus, rather than spread throughout the atom as in the raisin cookie model. The positively charged alpha particles would be repelled by the gold atom's nucleus, but most of the alphas would not come close enough to any nucleus to have their paths drastically altered. The few that did come close to a nucleus, however, could rebound backwards from a single such encounter, since the nucleus of a heavy gold atom would be fifty times more massive than an alpha particle. It turned out that it was not even too difficult to derive a formula giving the relative frequency of deflections through various angles, and this calculation agreed with the data well enough (to within 15%), considering the difficulty in getting good experimental statistics on the rare, very large angles.

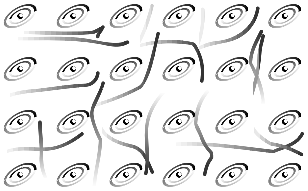
What had started out as a tedious exercise to get a student started in science had ended as a revolution in our understanding of nature. Indeed, the whole thing may sound a little too much like a moralistic fable of the scientific method with overtones of the Horatio Alger genre. The skeptical reader may wonder why the planetary model was ignored so thoroughly until Marsden and Rutherford's discovery. Is science really more of a sociological enterprise, in which certain ideas become accepted by the establishment, and other, equally plausible explanations are arbitrarily discarded? Some social scientists are currently ruffling a lot of scientists' feathers with critiques very much like this, but in this particular case, there were very sound reasons for rejecting the planetary model. As you'll learn in more detail later in this course, any charged particle that undergoes an acceleration dissipate energy in the form of light. In the planetary model, the electrons were orbiting the nucleus in circles or ellipses, which meant



A nonmetal.



A nonmagnetic metal.



A magnetized metal.

Note that all these figures are simplified in several ways. For one thing, the electrons of an individual atom do not all revolve around the nucleus in the same plane. It is also very unusual for a metal to become so strongly magnetized that 100% of its atoms have their rotations aligned as shown in this figure.

they were undergoing acceleration, just like the acceleration you feel in a car going around a curve. They should have dissipated energy as light, and eventually they should have lost all their energy. Atoms don't spontaneously collapse like that, which was why the raisin cookie model, with its stationary electrons, was originally preferred. There were other problems as well. In the planetary model, the one-electron atom would have to be flat, which would be inconsistent with the success of molecular modeling with spherical balls representing hydrogen and atoms. These molecular models also seemed to work best if specific sizes were used for different atoms, but there is no obvious reason in the planetary model why the radius of an electron's orbit should be a fixed number. In view of the conclusive Marsden-Rutherford results, however, these became fresh puzzles in atomic physics, not reasons for disbelieving the planetary model.

Some phenomena explained with the planetary model

The planetary model may not be the ultimate, perfect model of the atom, but don't underestimate its power. It already allows us to visualize correctly a great many phenomena.

As an example, let's consider the distinctions among nonmetals, metals that are magnetic, and metals that are nonmagnetic. As shown in the figures, a metal differs from a nonmetal because its outermost electrons are free to wander rather than owing their allegiance to a particular atom. A metal that can be magnetized is one that is willing to line up the rotations of some of its electrons so that their axes are parallel. Recall that magnetic forces are forces made by moving charges; we have not yet discussed the mathematics and geometry of magnetic forces, but it is easy to see how random orientations of the atoms in the nonmagnetic substance would lead to cancellation of the forces.

Even if the planetary model does not immediately answer such questions as why one element would be a metal and another a nonmetal, these ideas would be difficult or impossible to conceptualize in the raisin cookie model.

Discussion questions

A. The diagram showing alpha particles being deflected by a gold nucleus was drawn with the assumption that alpha particles came in on lines at many different distances from the nucleus. Why wouldn't they all come in along the same line, since they all came out through the same tube?

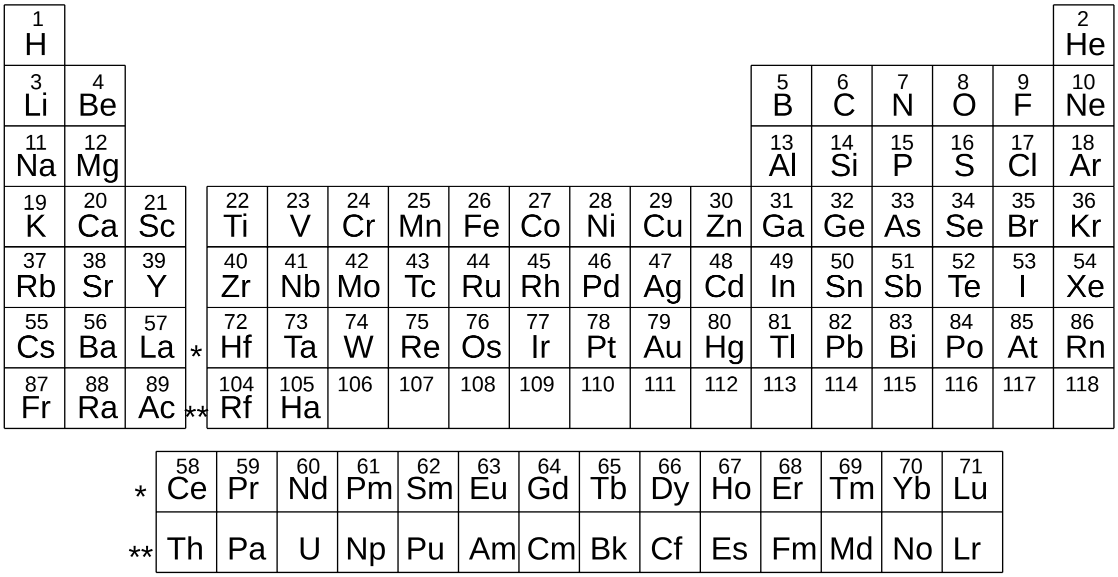
B. In reality, charges of the same type repel one another and charges of different types are attracted. Suppose the rules were the other way around, giving repulsion between opposite charges and attraction between similar ones. What would the universe be like?

2.3 Atomic Number

As alluded to in a discussion question in the previous section, scientists of this period had only a very approximate idea of how many units of charge resided in the nuclei of the various chemical elements. Although we now associate the number of units of nuclear charge with the element's position on the periodic table, and call it the *atomic number*, they had no idea that such a relationship existed. Mendeleev's table just seemed like an organizational tool, not something with any necessary physical significance. And everything Mendeleev had done seemed equally valid if you turned the table upside-down or reversed its left and right sides, so even if you wanted to number the elements sequentially with integers, there was an ambiguity as to how to do it. Mendeleev's original table was in fact upside-down compared to the modern one.

In the period immediately following the discovery of the nucleus, physicists only had rough estimates of the charges of the various nuclei. In the case of the very lightest nuclei, they simply found the maximum number of electrons they could strip off by various methods: chemical reactions, electric sparks, ultraviolet light, and so on. For example they could easily strip of one or two electrons from helium, making He^{+} or He^{++}, but nobody could make He^{+++}, presumably because the nuclear charge of helium was only $+2e$. Unfortunately only a few of the lightest elements could be stripped completely, because the more electrons were stripped off, the greater the positive net charge remaining, and the more strongly the rest of the negatively charged electrons would be held on. The heavy elements' atomic numbers could only be roughly extrapolated from the light elements, where the atomic number was about half the atom's mass expressed in units of the mass of a hydrogen atom. Gold, for example, had a mass about 197 times that of hydrogen, so its atomic number was estimated to be about half that, or somewhere around 100. We now know it to be 79.

How did we finally find out? The riddle of the nuclear charges was at last successfully attacked using two different techniques, which gave consis-



1 H																	2 He
3 Li	4 Be											5 B	6 C	7 N	8 O	9 F	10 Ne
11 Na	12 Mg											13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
55 Cs	56 Ba	57 La *	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
87 Fr	88 Ra	89 Ac **	104 Rf	105 Ha	106	107	108	109	110	111	112	113	114	115	116	117	118

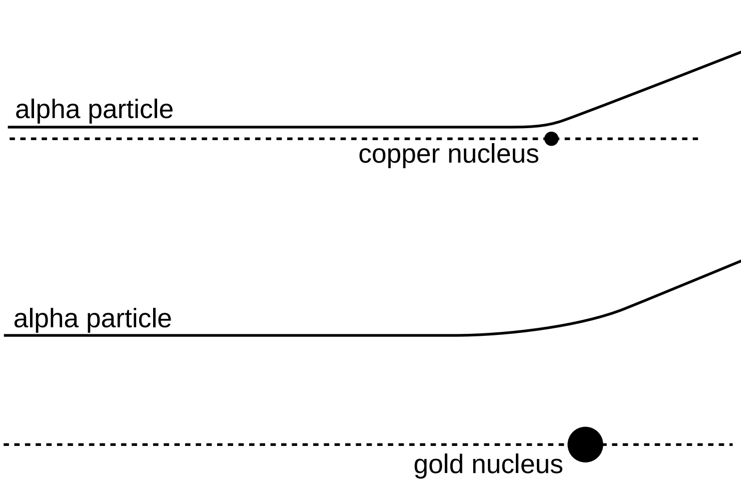
*	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb	71 Lu
**	Th	Pa	U	Np	Pu	Am	Cm	Bk	Cf	Es	Fm	Md	No	Lr

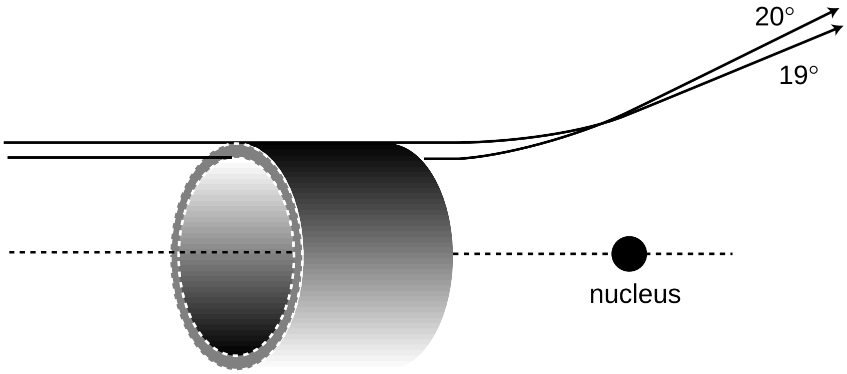
A modern periodic table, labeled with atomic numbers. Mendeleev's original table was upside-down compared to this one.

tent results. One set of experiments, involving x-rays, was performed by the young Henry Mosely, whose scientific brilliance was soon to be sacrificed in a battle between European imperialists over who would own the Dardanelles, during that pointless conflict then known as the War to End All Wars, and now referred to as World War I.

Since Mosely's analysis requires several concepts with which you are not yet familiar, we will instead describe the technique used by James Chadwick at around the same time. An added bonus of describing Chadwick's experiments is that they presaged the important modern technique of studying *collisions* of subatomic particles. In grad school, I worked with a professor whose thesis adviser's thesis adviser was Chadwick, and he related some interesting stories about the man. Chadwick was apparently a little nutty and a complete fanatic about science, to the extent that when he was held in a German prison camp during World War II, he managed to cajole his captors into allowing him to scrounge up parts from broken radios so that he could attempt to do physics experiments.

Chadwick's experiment worked like this. Suppose you perform two Rutherford-type alpha scattering measurements, first one with a gold foil as a target as in Rutherford's original experiment, and then one with a copper foil. It is possible to get large angles of deflection in both cases, but as shown in the figure, the alpha particle must be heading almost straight for the copper nucleus to get the same angle of deflection that would have occurred with an alpha that was much farther off the mark; the gold nucleus' charge is so much greater than the copper's that it exerts a strong force on the alpha particle even from far off. The situation is very much like that of a blindfolded person playing darts. Just as it is impossible to aim an alpha particle at an individual nucleus in the target, the blindfolded person cannot really aim the darts. Achieving a very close encounter with the copper atom would be akin to hitting an inner circle on the dartboard. It's





An alpha particle must be headed for the ring on the front of the imaginary cylindrical pipe in order to produce scattering at an angle between 19 and 20 degrees. The area of this ring is called the "cross-section" for scattering at 19-20° because it is the cross-sectional area of a cut through the pipe.

much more likely that one would have the luck to hit the outer circle, which covers a greater number of square inches. By analogy, if you measure the frequency with which alphas are scattered by copper at some particular angle, say between 19 and 20 degrees, and then perform the same measurement at the same angle with gold, you get a much higher percentage for gold than for copper.

In fact, the numerical ratio of the two nuclei's charges can be derived from this same experimentally determined ratio. Using the standard notation Z for the atomic number (charge of the nucleus divided by e), the following equation can be proven:

$$\frac{Z^2_{\text{gold}}}{Z^2_{\text{copper}}} = \frac{\text{number of alphas scattered by gold at 19–20°}}{\text{number of alphas scattered by copper at 19–20°}}$$

By making such measurements for targets constructed from all the elements, one can infer the ratios of all the atomic numbers, and since the atomic numbers of the light elements were already known, atomic numbers could be assigned to the entire periodic table. According to Mosely, the atomic numbers of copper, silver and platinum were 29, 47, and 78, which corresponded well with their positions on the periodic table. Chadwick's figures for the same elements were 29.3, 46.3, and 77.4, with error bars of about ± 1.5 times the fundamental charge, so the two experiments were in good agreement.

The point here is absolutely not that you should be ready to plug numbers into the above equation for a homework or exam question! My overall goal in this chapter is to explain how we know what we know about atoms. An added bonus of describing Chadwick's experiment is that the approach is very similar to that used in modern particle physics experiments, and the ideas used in the analysis are closely related to the now-ubiquitous concept of a "cross-section." In the dartboard analogy, the cross-section would be the area of the circular ring you have to hit. The reasoning behind the invention of the term "cross-section" can be visualized as shown in the figure. In this language, Rutherford's invention of the planetary model came from his unexpected discovery that there was a nonzero cross-section for alpha scattering from gold at large angles, and Chadwick confirmed Mosely's determinations of the atomic numbers by measuring cross-sections for alpha scattering.

Proof of the relationship between atomic number and scattering

The equation above can be derived by the following not very rigorous proof. To deflect the alpha particle by a certain angle requires that it acquire a certain momentum component in the direction perpendicular to its original momentum. Although the nucleus' force on the alpha particle is not constant, we can pretend that it is approximately constant during the time when the alpha is within a distance equal to, say, 150% of its distance of closest approach, and that the force is zero before and after that part of the motion. (If we chose 120% or 200%, it shouldn't make any difference in the final result, because the final result is a ratio, and the effects on the numerator and denominator should cancel each other.) In the approximation of constant force, the change in the alpha's perpendicular momentum component is then equal to $F\Delta t$. The Coulomb force law says the force is proportional to Z/r^2. Although r does change somewhat during the time interval of interest, it's good enough to treat it as a constant number, since we're only computing the ratio between the two experiments' results. Since we are approximating the force as acting over the time during which the distance is not too much greater than the distance of closest approach, the time interval Δt must be proportional to r, and the sideways momentum imparted to the alpha, $F\Delta t$, is proportional to $\left(Z/r^2\right)r$, or Z/r. If we're comparing alphas scattered at the same angle from gold and from copper, then Δp is the same in both cases, and the proportionality $\Delta p \propto Z/r$ tells us that the ones scattered from copper at that angle had to be headed in along a line closer to the central axis by a factor equaling $Z_{\text{gold}}/Z_{\text{copper}}$. If you imagine a "dartboard ring" that the alphas have to hit, then the ring for the gold experiment has the same proportions as the one for copper, but it is enlarged by a factor equal to $Z_{\text{gold}}/Z_{\text{copper}}$. That is, not only is the radius of the ring greater by that factor, but unlike the rings on a normal dartboard, the thickness of the outer ring is also greater in proportion to its radius. When you take a geometric shape and scale it up in size like a photographic enlargement, its area is increased in proportion to the square of the enlargement factor, so the area of the dartboard ring in the gold experiment is greater by a factor equal to $\left(Z_{\text{gold}}/Z_{\text{copper}}\right)^2$. Since the alphas are aimed entirely randomly, the chances of an alpha hitting the ring are in proportion to the area of the ring, which proves the equation given above.

As an example of the modern use of scattering experiments and cross-section measurements, you may have heard of the recent experimental evidence for the existence of a particle called the top quark. Of the twelve subatomic particles currently believed to be the smallest constituents of matter, six form a family called the quarks, distinguished from the other six by the intense attractive forces that make the quarks stick to each other. (The other six consist of the electron plus five other, more exotic particles.) The only two types of quarks found in naturally occurring matter are the "up quark" and "down quark," which are what protons and neutrons are made of, but four other types were theoretically predicted to exist, for a total of six. (The whimsical term "quark" comes from a line by James Joyce reading "Three quarks for master Mark.") Until recently, only five types of quarks had been proven to exist via experiments, and the sixth, the top quark, was only theorized. There was no hope of ever detecting a top quark directly, since it is radioactive, and only exists for a zillionth of a second before evaporating. Instead, the researchers searching for it at the Fermi National Accelerator Laboratory near Chicago measured cross-sections for scattering of nuclei off of other nuclei. The experiment was much like those of Rutherford and Chadwick, except that the incoming nuclei had to be boosted to much higher speeds in a particle accelerator. The resulting encounter with a target nucleus was so violent that both nuclei were completely demolished, but, as Einstein proved, energy can be converted into matter, and the energy of the collision creates a spray of exotic, radioactive particles, like the deadly shower of wood fragments produced by a cannon ball in an old naval battle. Among those particles were some top quarks. The cross-sections being measured were the cross-sections for the production of certain combinations of these secondary particles. However different the details, the principle was the same as that employed at the turn of the century: you smash things together and look at the fragments that fly off to see what was inside them. The approach has been compared to shooting a clock with a rifle and then studying the pieces that fly off to figure out how the clock worked.

Discussion questions

A. Why does it make sense that, as shown in the figure, the trajectories that result in 19° and 20° scattering cross each other?

B. Rutherford knew the velocity of the alpha particles emitted by radium, and guessed that the positively charged part of a gold atom had a charge of about +100e (we now know it is +79e). Considering the fact that some alpha particles were deflected by 180°, how could he then use conservation of energy to derive an upper limit on the size of a gold nucleus? (For simplicity, assume the size of the alpha particle is negligible compared to that of the gold nucleus, and ignore the fact that the gold nucleus recoils a little from the collision, picking up a little kinetic energy.)

2.4 The Structure of Nuclei

The proton

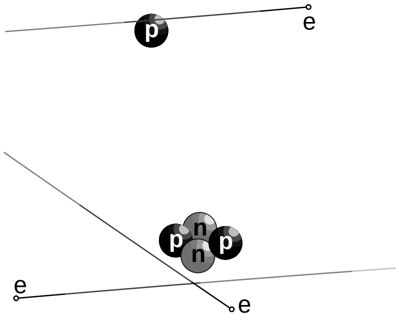
The fact that the nuclear charges were all integer multiples of e suggested to many physicists that rather than being a pointlike object, the nucleus might contain smaller particles having individual charges of $+e$. Evidence in favor of this idea was not long in arriving. Rutherford reasoned that if he bombarded the atoms of a very light element with alpha particles, the small charge of the target nuclei would give a very weak repulsion. Perhaps those few alpha particles that happened to arrive on head-on collision courses would get so close that they would physically crash into some of the target nuclei. An alpha particle is itself a nucleus, so this would be a collision between two nuclei, and a violent one due to the high speeds involved. Rutherford hit pay dirt in an experiment with alpha particles striking a target containing nitrogen atoms. Charged particles were detected flying out of the target like parts flying off of cars in a high-speed crash. Measurements of the deflection of these particles in electric and magnetic fields showed that they had the same charge-to-mass ratio as singly-ionized hydrogen atoms. Rutherford concluded that these were the conjectured singly-charged particles that held the charge of the nucleus, and they were later named protons. The hydrogen nucleus consists of a single proton, and in general, an element's atomic number gives the number of protons contained in each of its nuclei. The mass of the proton is about 1800 times greater than the mass of the electron.

The neutron

It would have been nice and simple if all the nuclei could have been built only from protons, but that couldn't be the case. If you spend a little time looking at a periodic table, you will soon notice that although some of the atomic masses are very nearly integer multiples of hydrogen's mass, many others are not. Even where the masses are close whole numbers, the masses of an element other than hydrogen is always greater than its atomic number, not equal to it. Helium, for instance, has two protons, but its mass is four times greater than that of hydrogen.

Chadwick cleared up the confusion by proving the existence of a new subatomic particle. Unlike the electron and proton, which are electrically charged, this particle is electrically neutral, and he named it the neutron. Chadwick's experiment has been described in detail in chapter 4 of book 2 of this series, but briefly the method was to expose a sample of the light element beryllium to a stream of alpha particles from a lump of radium. Beryllium has only four protons, so an alpha that happens to be aimed directly at a beryllium nucleus can actually hit it rather than being stopped short of a collision by electrical repulsion. Neutrons were observed as a new form of radiation emerging from the collisions, and Chadwick correctly inferred that they were previously unsuspected components of the nucleus that had been knocked out. As described in book 2, Chadwick also determined the mass of the neutron; it is very nearly the same as that of the proton.

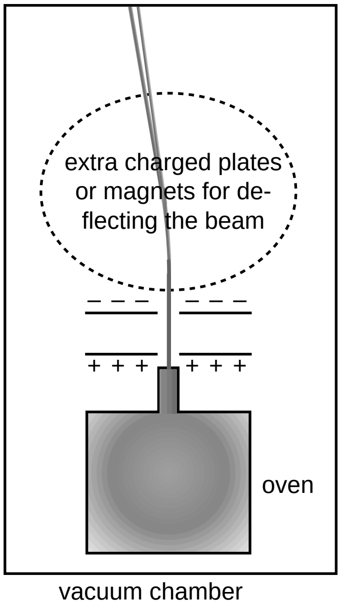
To summarize, atoms are made of three types of particles:



Examples of the construction of atoms: hydrogen (top) and helium (bottom). On this scale, the electrons' orbits would be the size of a college campus.

		mass in units of	
	charge	**the proton's mass**	**location in atom**
proton	$+e$	1	in nucleus
neutron	0	1.001	in nucleus
electron	$-e$	1/1836	orbiting around nucleus

The existence of neutrons explained the mysterious masses of the elements. Helium, for instance, has a mass very close to four times greater than that of hydrogen. This is because it contains two neutrons in addition to its two protons. The mass of an atom is essentially determined by the total number of neutrons and protons. The total number of neutrons plus protons is therefore referred to as the atom's *mass number*.



A version of the Thomson apparatus modified for measuring the mass-to-charge ratios of ions rather than electrons. A small sample of the element in question, copper in our example, is boiled in the oven to create a thin vapor. (A vacuum pump is continuously sucking on the main chamber to keep it from accumulating enough gas to stop the beam of ions.) Some of the atoms of the vapor are ionized by a spark or by ultraviolet light. Ions that wander out of the nozzle and into the region between the charged plates are then accelerated toward the top of the figure. As in the Thomson experiment, mass-to-charge ratios are inferred from the deflection of the beam.

Isotopes

We now have a clear interpretation of the fact that helium is close to four times more massive than hydrogen, and similarly for all the atomic masses that are close to an integer multiple of the mass of hydrogen. But what about copper, for instance, which had an atomic mass 63.5 times that of hydrogen? It didn't seem reasonable to think that it possessed an extra half of a neutron! The solution was found by measuring the mass-to-charge ratios of singly-ionized atoms (atoms with one electron removed). The technique is essentially that same as the one used by Thomson for cathode rays, except that whole atoms do not spontaneously leap out of the surface of an object as electrons sometimes do. The figure shows an example of how the ions can be created and injected between the charged plates for acceleration.

Injecting a stream of copper ions into the device, we find a surprise — the beam splits into two parts! Chemists had elevated to dogma the assumption that all the atoms of a given element were identical, but we find that 69% of copper atoms have one mass, and 31% have another. Not only that, but both masses are very nearly integer multiples of the mass of hydrogen (63 and 65, respectively). Copper gets its chemical identity from the number of protons in its nucleus, 29, since chemical reactions work by electric forces. But apparently some copper atoms have 63-29=34 neutrons while others have 65-29=36. The atomic mass of copper, 63.5, reflects the proportions of the mixture of the mass-63 and mass-65 varieties. The different mass varieties of a given element are called *isotopes* of that element.

Isotopes can be named by giving the mass number as a subscript to the left of the chemical symbol, e.g. ^{65}Cu. Examples:

	protons	**neutrons**	**mass number**
^{1}H	1	0	0+1=1
^{4}He	2	2	2+2=4
^{12}C	6	6	6+6=12
^{14}C	6	8	6+8=14
^{262}Ha	105	157	105+157=262

Self-Check

Why are the positive and negative charges of the accelerating plates reversed in the isotope-separating apparatus compared to the Thomson apparatus?

Chemical reactions are all about the exchange and sharing of electrons: the nuclei have to sit out this dance because the forces of electrical repulsion prevent them from ever getting close enough to make contact with each other. Although the protons do have a vitally important effect on chemical processes because of their electrical forces, the neutrons can have no effect on the atom's chemical reactions. It is not possible, for instance, to separate ^{63}Cu from ^{65}Cu by chemical reactions. This is why chemists had never realized that different isotopes existed. (To be perfectly accurate, different isotopes do behave slightly differently because the more massive atoms move more sluggishly and therefore react with a tiny bit less intensity. This tiny difference is used, for instance, to separate out the isotopes of uranium needed to build a nuclear bomb. The smallness of this effect makes the separation process a slow and difficult one, which is what we have to thank for the fact that nuclear weapons have not been built by every terrorist cabal on the planet.)

Sizes and shapes of nuclei

Matter is nearly all nuclei if you count by weight, but in terms of volume nuclei don't amount to much. The radius of an individual neutron or proton is very close to 1 fm (1 fm=10^{-15} m), so even a big lead nucleus with a mass number of 208 still has a diameter of only about 13 fm, which is ten thousand times smaller than the diameter of a typical atom. Contrary to the usual imagery of the nucleus as a small sphere, it turns out that many nuclei are somewhat elongated, like an American football, and a few have exotic asymmetric shapes like pears or kiwi fruits.

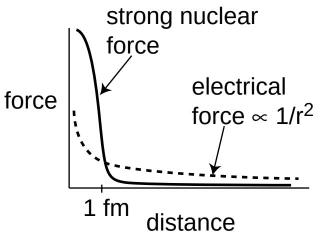
Discussion questions

Suppose the entire universe was in a (very large) cereal box, and the nutritional labeling was supposed to tell a godlike consumer what percentage of the contents was nuclei. Roughly what would the percentage be like if the labeling was according to mass? What if it was by volume?

Thomson was accelerating electrons, which are negatively charged. This apparatus is supposed to accelerate atoms with one electron stripped off, which have positive net charge. In both cases, a particle that is between the plates should be attracted by the forward plate and repelled by the plate behind it.

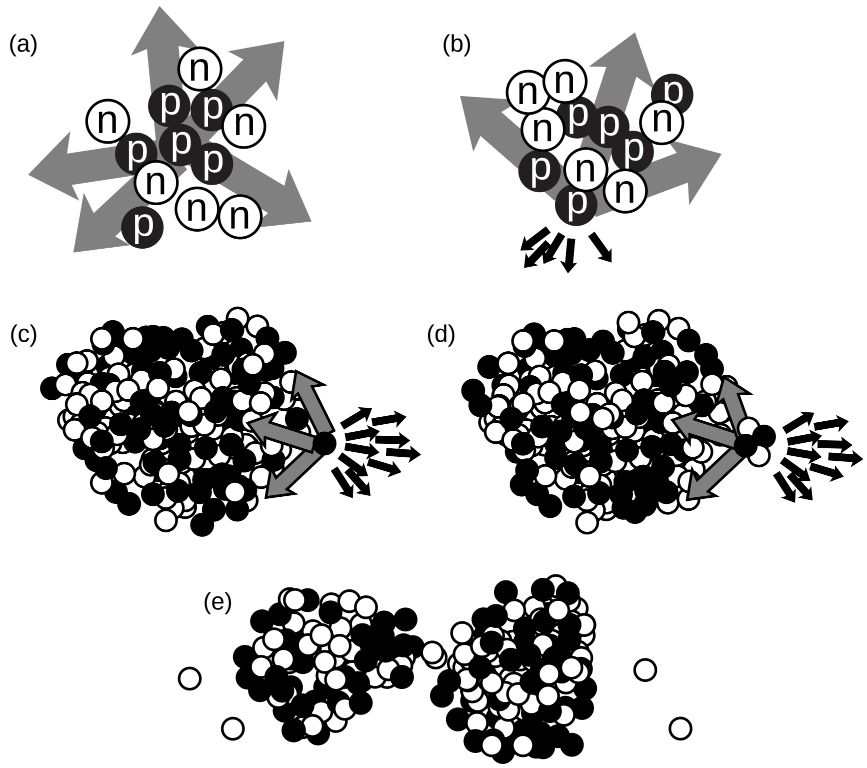
2.5 The Strong Nuclear Force, Alpha Decay and Fission

Once physicists realized that nuclei consisted of positively charged protons and uncharged neutrons, they had a problem on their hands. The electrical forces among the protons are all repulsive, so the nucleus should simply fly apart! The reason all the nuclei in your body are not spontaneously exploding at this moment is that there is another force acting. This force, called the *strong nuclear force*, is always attractive, and acts between neutrons and neutrons, neutrons and protons, and protons and protons with roughly equal strength. The strong nuclear force does not have any effect on electrons, which is why it does not influence chemical reactions.



Unlike the electric forces, whose strengths are given by the simple Coulomb force law, there is no simple formula for how the strong nuclear force depends on distance. Roughly speaking, it is effective over ranges of ~1 fm, but falls off extremely quickly at larger distances (much faster than $1/r^2$). Since the radius of a neutron or proton is about 1 fm, that means that when a bunch of neutrons and protons are packed together to form a nucleus, the strong nuclear force is effective only between neighbors.

The figure illustrates how the strong nuclear force acts to keep ordinary nuclei together, but is not able to keep very heavy nuclei from breaking apart. In (a), a proton in the middle of a carbon nucleus feels an attractive strong nuclear force (arrows) from each of its nearest neighbors. The forces are all in different directions, and tend to cancel out. The same is true for the repulsive electrical forces (not shown). (b) A proton at the edge of the nucleus has neighbors only on one side, and therefore all the strong nuclear forces acting on it are tending to pull it back in. Although all the electrical forces from the other five protons (dark arrows) are all pushing it out of the



nucleus, they are not sufficient to overcome the strong nuclear forces.

In a very heavy nucleus, (c), a proton that finds itself near the edge has only a few neighbors close enough to attract it significantly via the strong nuclear force, but every other proton in the nucleus exerts a repulsive electrical force on it. If the nucleus is large enough, the total electrical repulsion may be sufficient to overcome the attraction of the strong force, and the nucleus may spit out a proton. Proton emission is fairly rare, however; a more common type of radioactive decay in heavy nuclei is alpha decay, shown in (d). The imbalance of the forces is similar, but the chunk that is ejected is an alpha particle (two protons and two neutrons) rather than a single proton.

It is also possible for the nucleus to split into two pieces of roughly equal size, (e), a process known as fission.

When a nucleus is able to undergo one of these processes, it is said to be radioactive, and to undergo radioactive decay. Some of the naturally occurring nuclei on earth are radioactive. The term "radioactive" comes from Becquerel's image of rays radiating out from something, not from radio waves, which are a whole different phenomenon. The term "decay" can also be a little misleading, since it implies that the nucleus turns to dust or simply disappears -- actually it is splitting into two new nuclei with an the same total number of neutrons and protons, so the term "radioactive transformation" would have been more appropriate. Although the original atom's electrons are mere spectators in the process of weak radioactive decay, we often speak loosely of "radioactive atoms" rather than "radioactive nuclei."

Randomness in physics

How does an atom decide when to decay? We might imagine that it is like a termite-infested house that gets weaker and weaker, until finally it reaches the day on which it is destined to fall apart. Experiments, however, have not succeeded in detecting such “ticking clock” hidden below the surface; the evidence is that all atoms of a given isotope are absolutely identical. Why, then, would one uranium atom decay today while another lives for another million years? The answer appears to be that it is entirely random. We can make general statements about the average time required for a certain isotope to decay, or how long it will take for half the atoms in a sample to decay (its half-life), but we can never predict the behavior of a particular atom.

This is the first example we have encountered of an inescapable randomness in the laws of physics. If this kind of randomness makes you uneasy, you're in good company. Einstein's famous quote is “...I am convinced that He [God] does not play dice." Einstein's distaste for randomness, and his association of determinism with divinity, goes back to the Enlightenment conception of the universe as a gigantic piece of clockwork that only had to be set in motion initially by the Builder. Physics had to be entirely rebuilt in the 20th century to incorporate the fundamental randomness of physics, and this modern revolution is the topic of book 6 of this series. In particular, we will delay the mathematical development of the half-life concept until then.

2.6 The Weak Nuclear Force; Beta Decay

All the nuclear processes we've discussed so far have involved rearrangements of neutrons and protons, with no change in the total number of neutrons or the total number of protons. Now consider the proportions of neutrons and protons in your body and in the planet earth: neutrons and protons are roughly equally numerous in your body's carbon and oxygen nuclei, and also in the nickel and iron that make up most of the earth. The proportions are about 50-50. But the only chemical elements produced in any significant quantities by the big bang were hydrogen (about 90%) and helium (about 10%). If the early universe was almost nothing but hydrogen atoms, whose nuclei are protons, where did all those neutrons come from?

The answer is that there is another nuclear force, the weak nuclear force, that is capable of transforming neutrons into protons and vice-versa. Two possible reactions are

$n \rightarrow p + e^{-} + \bar{\nu}$ (electron decay)

and

$p \rightarrow n + e^{+} + \nu$. (positron decay)

(There is also a third type called electron capture, in which a proton grabs one of the atom's electrons and they produce a neutron and a neutrino.)

Whereas alpha decay and fission are just a redivision of the previously existing particles, these reactions involve the destruction of one particle and the creation of three new particles that did not exist before.

There are three new particles here that you have never previously encountered. The symbol e^{+} stands for an antielectron, which is a particle just like the electron in every way, except that its electric charge is positive rather than negative. Antielectrons are also known as positrons. Nobody knows why electrons are so common in the universe and antielectrons are scarce. When an antielectron encounters an electron, they annihilate each other, and this is the fate of all the antielectrons that are produced by natural radioactivity on earth.

The notation ν stands for a particle called a neutrino, and $\bar{\nu}$ means an antineutrino. Neutrinos and antineutrinos have no electric charge (hence the name)

We can now list all four of the known fundamental forces of physics:

gravity
electromagnetism
strong nuclear force
weak nuclear force

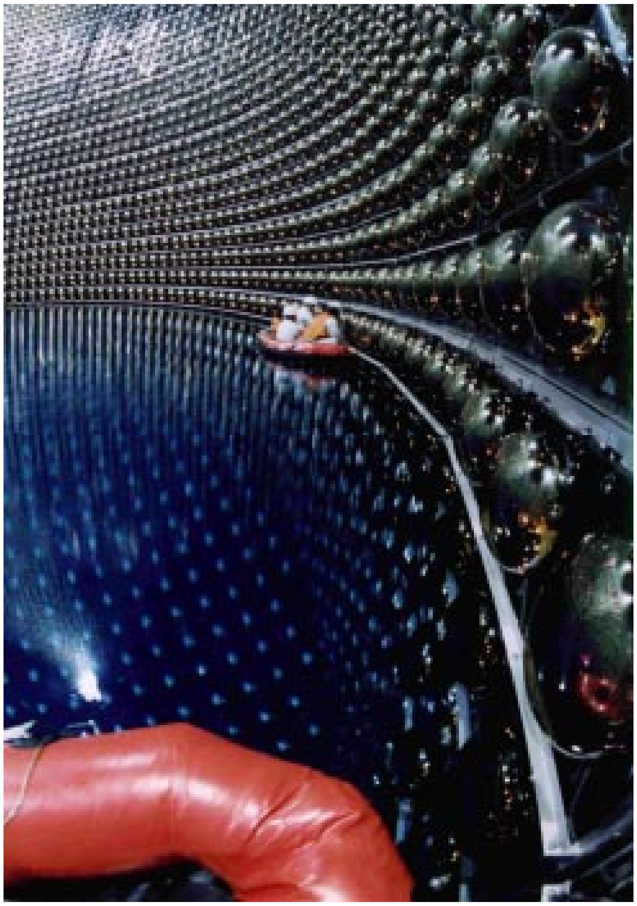
The other forces we have learned about, such as friction and the normal

A billion of them pass through your body every microsecond, but until recently almost nothing was known about the particles called neutrinos. Produced as a side-effect of the nuclear reactions that power our sun and other stars, these ghostlike bits of matter are believed to be the most numerous particles in the universe. But they interact so weakly with ordinary matter that nearly all the neutrinos that enter the earth on one side will emerge from the other side of our planet without even slowing down.

Our first real peek at the properties of the elusive neutrino has come from a huge detector in a played-out Japanese zinc mine. An international team of physicists outfitted the mineshaft with wall-to-wall light sensors, and then filled the whole thing with water so pure that you can see through it for a hundred meters, compared to only a few meters for typical tap water. Neutrinos stream through the 50 million liters of water continually, just as they flood everything else around us, and the vast majority never interact with a water molecule. A very small percentage, however, do annihilate themselves in the water, and the tiny flashes of light they produce can be detected by the beachball-sized vacuum tubes that line the darkened mineshaft. Most of the neutrinos around us come from the sun, but for technical reasons this type of water-based detector is more sensitive to the less common but more energetic neutrinos produced when cosmic ray particles strike the earth's atmosphere.

Neutrinos were already known to come in three "flavors," which can be distinguished from each other by the particles created when they collide with matter. An "electron-flavored neutrino" creates an ordinary electron when it is annihilated, while the two other types create more exotic particles called mu and tau particles. Think of the three types of neutrinos as chocolate, vanilla, and strawberry. When you buy a chocolate ice cream cone, you expect that it will keep being chocolate as you eat it. The unexpected finding from the Japanese experiment is that some of the neutrinos are changing flavor between the time when they are produced by a cosmic ray and the moment when they wink out of existence in the water. It's as though your chocolate ice cream cone transformed itself magically into strawberry while your back was turned.

The drawing shows the clue that tipped off the experimenters to the change in flavor. The experiment detects some neutrinos originating in the atmosphere above Japan, and also many neutrinos coming from distant parts of the earth. A neutrino created above the Atlantic Ocean arrives in Japan from underneath, and the experiment can distinguish these upward-traveling neutrinos from the downward-moving local variety. They found that the mixture of neutrinos coming from below was different from the mixture arriving from above, with some of the electron-flavored and tau-flavored neutrinos having apparently changed into mu-flavored neutrinos during their voyage through the earth. The ones coming from above didn't have time to change flavors on their much shorter journey.



This is interpreted as evidence that the neutrinos are vibrating back and forth among the three flavors, like a rope vibrating back and forth as a wave passes through it. On theoretical grounds, it is believed that such a vibration can only occur if neutrinos have mass. Only a rough estimate of the mass is possible at this point: it appears that neutrinos have a mass somewhere in the neighborhood of one billionth of the mass of an electron, or about 10^{-39} kg.

If the neutrino's mass is so tiny, does it even matter? The answer from cosmologists is a resounding yes. Although a single neutrino's mass may not amount to much, they are so numerous that they may have had a decisive effect on the gravitational forces that have molded the evolution of the universe from the big bang to the present time.

force, all arise from electromagnetic interactions between atoms, and therefore are not considered to be fundamental forces of physics.

Example: decay of ^{212}Pb
As an example, consider the radioactive isotope of lead ^{212}Pb. It contains 82 protons and 130 neutrons. It decays by the process $n \rightarrow p + e^{-} + \bar{\nu}$. The newly created proton is held inside the nucleus by the strong nuclear force, so the new nucleus contains 83 protons and 129 neutrons. Having 83 protons makes it the element bismuth, so it will be an atom of ^{212}Bi.

In a reaction like this one, the electron flies off at high speed (typically close to the speed of light), and the escaping electrons are the things that make large amounts of this type of radioactivity dangerous. The outgoing electron was the first thing that tipped off scientists in the early 1900s to the existence of this type of radioactivity. Since they didn't know that the outgoing particles were electrons, they called them beta particles, and this type of radioactive decay was therefore known as beta decay. A clearer but less common terminology is to call the two processes electron decay and positron decay.

The antineutrino pretty much ignores all matter, because its lack of charge makes it immune to electrical forces, and it also remains aloof from strong nuclear interactions. Even if it happens to fly off going straight down, it is almost certain to make it through the entire earth without interacting with any atoms in any way. It ends up flying through outer space forever. The neutrino's behavior makes it exceedingly difficult to detect, and when beta decay was first discovered nobody realized that neutrinos even existed. We now know that the neutrino carries off some of the energy produced in the reaction, but at the time it seemed that the total energy afterwards (not counting the unsuspected neutrino's energy) was greater than the total energy before the reaction, violating conservation of energy. Physicists were getting ready to throw conservation of energy out the window as a basic law of physics when indirect evidence led them to the conclusion that neutrinos existed.

Discussion questions

A. In the reactions $n \rightarrow p + e^- +$ and $p \rightarrow n + e^+ + n$, verify that charge is conserved. In beta decay, when one of these reactions happens to a neutron or proton within a nucleus, one or more gamma rays may also be emitted. Does this effect conservation of charge? Would it be possible for some extra electrons to be released without violating charge conservation?

B. When an antielectron and an electron annihilate each other, they produce two gamma rays. Is charge conserved in this reaction?

2.7 Fusion

As we have seen, heavy nuclei tend to fly apart because each proton is being repelled by every other proton in the nucleus, but is only attracted by its nearest neighbors. The nucleus splits up into two parts, and as soon as those two parts are more than about 1 fm apart, the strong nuclear force no longer causes the two fragments to attract each other. The electrical repulsion then accelerates them, causing them to gain a large amount of kinetic energy. This release of kinetic energy is what powers nuclear reactors and fission bombs.

It might seem, then, that the lightest nuclei would be the most stable, but that is not the case. Let's compare an extremely light nucleus like ^{4}He with a somewhat heavier one, ^{16}O. A neutron or proton in ^{4}He can be attracted by the three others, but in ^{16}O, it might have five or six neighbors attracting it. The ^{16}O nucleus is therefore more stable.

It turns out that the most stable nuclei of all are those around nickel and iron, having about 30 protons and 30 neutrons. Just as a nucleus that is too heavy to be stable can release energy by splitting apart into pieces that are closer to the most stable size, light nuclei can release energy if you stick them together to make bigger nuclei that are closer to the most stable size. Fusing one nucleus with another is called nuclear fusion. Nuclear fusion is what powers our sun and other stars.

This array of gamma-ray detectors, called GAMMASPHERE, is currently housed at Argonne National Laboratory, in Illinois. During operation, the array is closed up, and a beam of ions produced by a particle accelerator strikes a target at its center, producing nuclear fusion reactions. The gamma rays can be studied for information about the structure of the fused nuclei, which are typically varieties not found in nature. The barrel-shaped part behind the scientist is a mass separator used for identifying the type of nucleus formed in the reaction after it recoils out of GAMMASPHERE.

2.8 Nuclear Energy and Binding Energies

In the same way that chemical reactions can be classified as exothermic (releasing energy) or endothermic (requiring energy to react), so nuclear reactions may either release or use up energy. The energies involved in nuclear reactions are greater by a huge factor. Thousands of tons of coal would have to be burned to produce as much energy as would be produced in a nuclear power plant by one kg of fuel.

Although nuclear reactions that use up energy (endothermic reactions) can be initiated in accelerators, where one nucleus is rammed into another at high speed, they do not occur in nature, not even in the sun. The amount of kinetic energy required is simply not available.

To find the amount of energy consumed or released in a nuclear reaction, you need to know how much potential energy was stored or released. Experimentalists have determined the amount of potential energy stored in the nucleus of every stable element, as well as many unstable elements. This is the amount of mechanical work that would be required to pull the nucleus apart into its individual neutrons and protons, and is known as the nuclear *binding energy*.

Example: a reaction occurring in the sun

The sun produces its energy through a series of nuclear fusion reactions. One of the reactions is

$$^{1}H + ^{2}H \rightarrow ^{3}He + \gamma$$

The excess energy is almost all carried off by the gamma ray (not by the kinetic energy of the helium-3 atom). The binding energies in units of pJ (picojoules) are:

^{1}H	0 J
^{2}H	0.35593 pJ
^{3}He	1.23489 pJ

The total initial potential energy is 0 pJ+0.35593 pJ, and the final potential energy is 1.23489 pJ, so by conservation of energy, the gamma ray must carry off 0.87896 pJ of energy. The gamma ray is then absorbed by the sun and converted to heat.

Self-Check

Why is the binding energy of ^{1}H exactly equal to zero?

The hydrogen-1 nucleus is simply a proton. The binding energy is the energy required to tear a nucleus apart, but for a nucleus this simple there is nothing to tear apart.

Optional topic: conversion of mass to energy and energy to mass

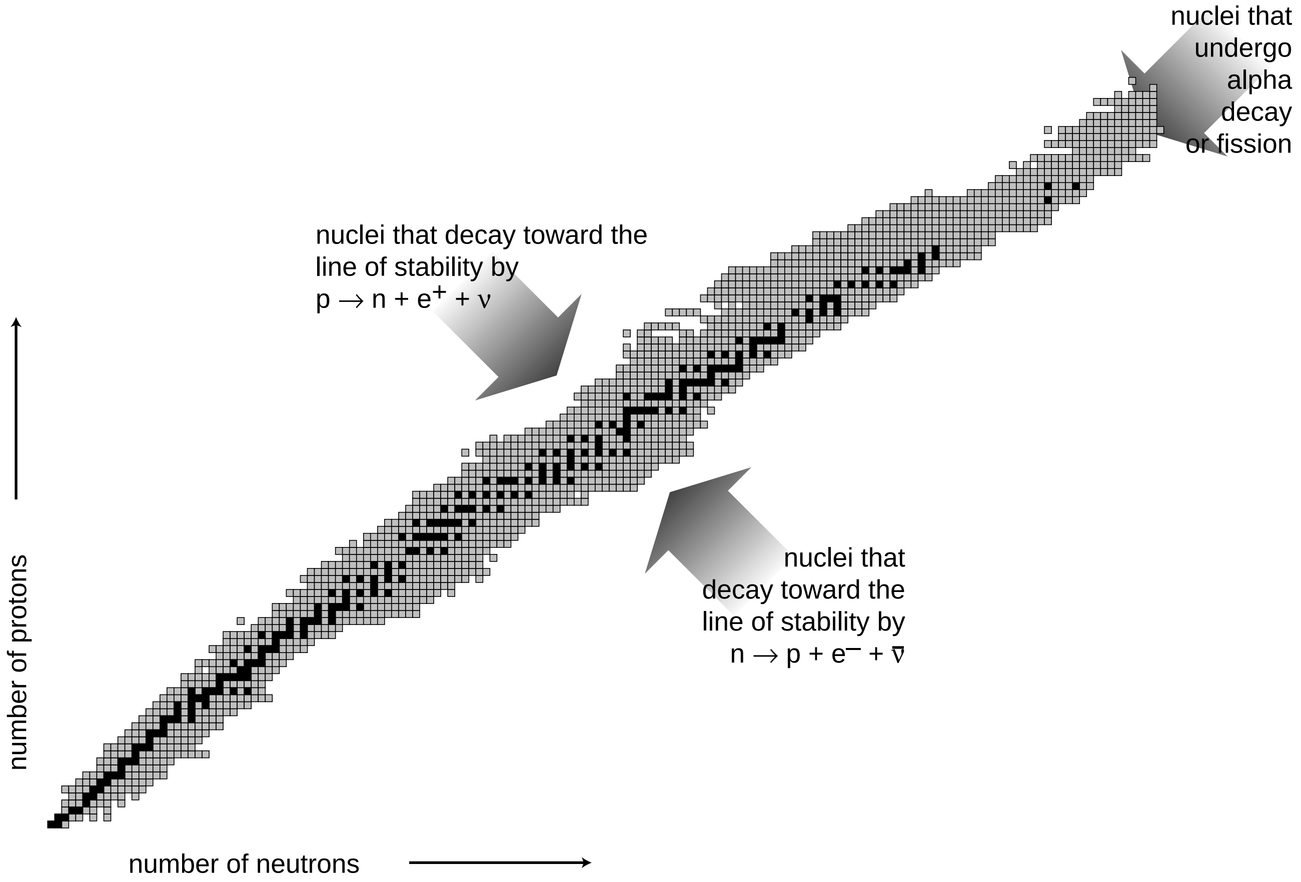
If you add up the masses of the three particles produced in the reaction $n \rightarrow p + e^- + \bar{\nu}$, you will find that they do not equal the mass of the neutron, so mass is not conserved. An even more blatant example is the annihilation of an electron with a positron, $e^- + e^+ \rightarrow 2\gamma$, in which the original mass is completely destroyed, since gamma rays have no mass. Nonconservation of mass is not just a property of nuclear reactions. It also occurs in chemical reactions, but the change in mass is too small to detect with ordinary laboratory balances.

The reason why mass is not being conserved is that mass is being converted to energy, according to Einstein's celebrated equation $E=mc^2$, in which c stands for the speed of light. In the reaction $e^- + e^+ \rightarrow 2\gamma$, for instance, imagine for simplicity that the electron and positron are moving very slowly when they collide, so there is no significant amount of energy to start with. We are starting with mass and no energy, and ending up with two gamma rays that possess energy but no mass. Einstein's $E=mc^2$ tells us that the conversion factor between mass and energy is equal to the square of the speed of light. Since c is a big number, the amount of energy consumed or released by a chemical reaction only shows up as a tiny change in mass. But in nuclear reactions, which involve large amounts of energy, the change in mass may amount to as much as one part per thousand. Note that in this context, c is not necessarily the speed of any of the particles. We are just using its numerical value as a conversion factor. Note also that $E=mc^2$ does not mean that an object of mass m has a kinetic energy equal to mc^2; the energy being described by $E=mc^2$ is the energy you could release if you destroyed the particle and converted its mass entirely into energy, and that energy would be in addition to any kinetic or potential energy the particle had.

Have we now been cheated out of two perfectly good conservation laws, the laws of conservation of mass and of energy? No, it's just that according to Einstein, the conserved quantity is $E+mc^2$, not E or m individually. The quantity $E+mc^2$ is referred to as the mass-energy, and no violation of the law of conservation of mass-energy has yet been observed. In most practical situations, it is a perfectly reasonable to treat mass and energy as separately conserved quantities.

It is now easy to explain why isolated protons (hydrogen nuclei) are found in nature, but neutrons are only encountered in the interior of a nucleus, not by themselves. In the process $n \rightarrow p + e^- + \bar{\nu}$, the total final mass is less than the mass of the neutron, so mass is being converted into energy. In the beta decay of a proton, $p \rightarrow n + e^+ + \nu$, the final mass is greater than the initial mass, so some energy needs to be supplied for conversion into mass. A proton sitting by itself in a hydrogen atom cannot decay, since it has no source of energy. Only protons sitting inside nuclei can decay, and only then if the difference in potential energy between the original nucleus and the new nucleus would result in a release of energy. But any isolated neutron that is created in natural or artificial reactions will decay within a matter of seconds, releasing some energy.

The equation $E=mc^2$ occurs naturally as part of Einstein's theory of special relativity, which is not what we are studying right now. This brief treatment is only meant to clear up the issue of where the mass was going in some of the nuclear reactions we were discussing.



The figure above is a compact way of showing the vast variety of the nuclei. Each box represents a particular number of neutrons and protons. The black boxes are nuclei that are stable, i.e. that would require an input of energy in order to change into another. The gray boxes show all the unstable nuclei that have been studied experimentally. Some of these last for billions of years on the average before decaying and are found in nature, but most have much shorter average lifetimes, and can only be created and studied in the laboratory.

The curve along which the stable nuclei lie is called the line of stability. Nuclei along this line have the most stable proportion of neutrons to protons. For light nuclei the most stable mixture is about 50-50, but we can see that stable heavy nuclei have two or three times more neutrons than protons. This is because the electrical repulsions of all the protons in a heavy nucleus add up to a powerful force that would tend to tear it apart. The presence of a large number of neutrons increases the distances among the protons, and also increases the number of attractions due to the strong nuclear force.

2.9 Biological Effects of Ionizing Radiation

As a science educator, I find it frustrating that nowhere in the massive amount of journalism devoted to the Chernobyl disaster does one ever find any numerical statements about the amount of radiation to which people have been exposed. Anyone mentally capable of understanding sports statistics or weather reports ought to be able to understand such measurements, as long as something like the following explanatory text was inserted somewhere in the article:

Radiation exposure is measured in units of millirems. The average person is exposed to about 100 millirems each year from natural background sources.

With this context, people would be able to come to informed conclusions based on statements such as, "Children in Finland received an average dose of ___________ millirems above natural background levels because of the Chernobyl disaster."

A millirem, or mrem, is, of course, a thousandth of a rem, but what is a rem? It measures the amount of energy per kilogram deposited in the body by ionizing radiation, multiplied by a "quality factor" to account for the different health hazards posed by alphas, betas, gammas, neutrons, and other types of radiation. Only ionizing radiation is counted, since nonionizing radiation simply heats one's body rather than killing cells or altering DNA. For instance, alpha particles are typically moving so fast that their kinetic energy is sufficient to ionize thousands of atoms, but it is possible for an alpha particle to be moving so slowly that it would not have enough kinetic energy to ionize even one atom.

Notwithstanding the pop culture images of the Incredible Hulk and Godzilla, it is not possible for a multicellular animal to become “mutated” as a whole. In most cases, a particle of ionizing radiation will not even hit the DNA, and even if it does, it will only affect the DNA of a single cell, not every cell in the animal’s body. Typically, that cell is simply killed, because the DNA becomes unable to function properly. Once in a while, however, the DNA may be altered so as to make that cell cancerous. For instance, skin cancer can be caused by UV light hitting a single skin cell in the body of a sunbather. If that cell becomes cancerous and begins reproducing uncontrollably, she will end up with a tumor twenty years later.

Other than cancer, the only other dramatic effect that can result from altering a single cell’s DNA is if that cell happens to be a sperm or ovum, which can result in nonviable or mutated offspring. Men are relatively immune to reproductive harm from radiation, because their sperm cells are replaced frequently. Women are more vulnerable because they keep the same set of ova as long as they live.

A whole-body exposure of 500,000 mrem will kill a person within a week or so. Luckily, only a small number of humans have ever been exposed to such levels: one scientist working on the Manhattan Project, some victims of the Nagasaki and Hiroshima explosions, and 31 workers at Chernobyl. Death occurs by massive killing of cells, especially in the blood-producing cells of the bone marrow.

Lower levels, on the order of 100,000 mrem, were inflicted on some people at Nagasaki and Hiroshima. No acute symptoms result from this level of exposure, but certain types of cancer are significantly more common among these people. It was originally expected that the radiation would cause many mutations resulting in birth defects, but very few such inherited effects have been observed.

A great deal of time has been spent debating the effects of very low levels of ionizing radiation. A medical x-ray, for instance, may result in a dose on the order of 100 mrem above background, i.e. a doubling of the normal background level. Similar doses in excess of the average background level may be received by people living at high altitudes or people with high concentrations of radon gas in their houses. Unfortunately (or fortunately, depending on how you look at it), the added risks of cancer or birth defects resulting from these levels of exposure are extremely small, and therefore nearly impossible to measure. As with many suspected carcinogenic chemicals, the only practical method of estimating risks is to give laboratory animals doses many orders of magnitude greater, and then assume that the health risk is directly proportional to the dose. Under these assumptions, the added risk posed by a dental x-ray or radon in one's basement is negligible on a personal level, and is only significant in terms of a slight increase in cancer throughout the population. As a matter of social policy, excess radiation exposure is not a significant public health problem compared to car accidents or tobacco smoking.

Discussion questions

A. Should the quality factor for neutrinos be very small, because they mostly don't interact with your body?

B. Would an alpha source be likely to cause different types of cancer depending on whether the source was external to the body or swallowed in contaminated food? What about a gamma source?

2.10* The Creation of the Elements

Creation of hydrogen and helium in the Big Bang

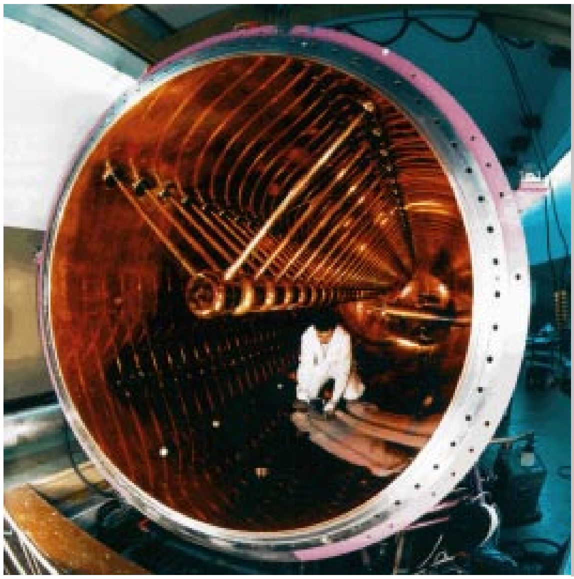
We have discussed in book 3 of this series the evidence for the Big Bang theory of the origin of the universe. Did all the chemical elements we're made of come into being in the Big Bang? The answer is definitely no, since the temperatures in the first microseconds after the Big Bang were so high that atoms and nuclei could not hold together at all. Even after things had cooled down enough for nuclei and atoms to exist, theorists are sure that the only elements created were hydrogen and helium.

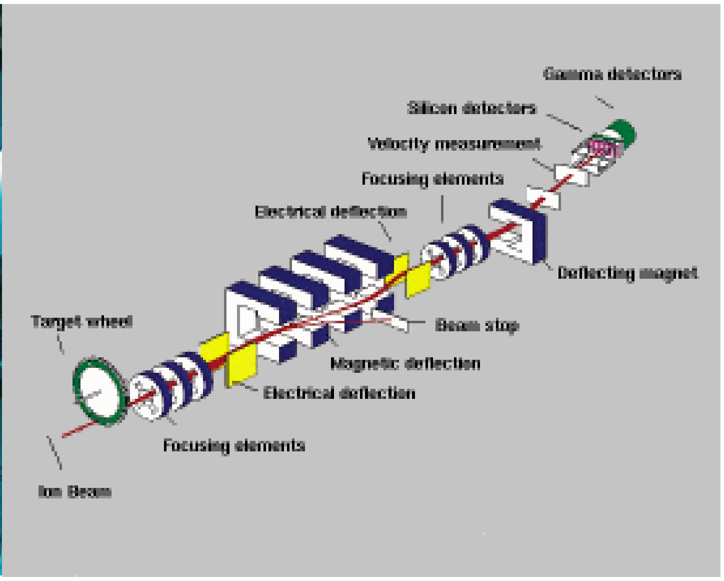
We are stardust

In that case, where did all the other elements come from? Astronomers came up with the answer. By studying the combinations of wavelengths of light, called spectra, emitted by various stars, they had been able to determine what kinds of atoms they contained. (We will have more to say about spectra in book 6.) They found that the stars fell into two groups. One type was nearly 100% hydrogen and helium, while the other contained 99% hydrogen and helium and 1% other elements. They interpreted these as two generations of stars. The first generation had formed out of clouds of gas that came fresh from the big bang, and their composition reflected that of the early universe. The nuclear fusion reactions by which they shine have mainly just increased the proportion of helium relative to hydrogen, without making any heavier elements.

The members of the first generation that we see today, however, are only those that lived a long time. Small stars are more miserly with their fuel than large stars, which have short lives. The large stars of the first generation have already finished their lives. Near the end of its lifetime, a star runs out hydrogen fuel and undergoes a series of violent and spectacular reorganizations as it fuses heavier and heavier elements. Very large stars finish this sequence of events by undergoing supernova explosions, in which some of their material is flung off into the void while the rest collapses into an exotic object such as a black hole or neutron star.

The second generation of stars, of which our own sun is an example, condensed out of clouds of gas that had been enriched in heavy elements due to supernova explosions. It is those heavy elements that make up our planet and our bodies.





Top left: Construction of the UNILAC accelerator in Germany, one of whose uses is for experiments to create very heavy artificial elements.

Top right: This formidable-looking apparatus, called SHIP, is really nothing more than a glorified version of the apparatus used by Thomson to determine the velocity and mass-to-charge ratios of a beam of unknown particles. Nuclei from a beam of ions produced by UNILAC strike a metal foil target, and the nuclei produced in the resulting fusion reaction recoil into ship, which is connected to the "downstream" end of the accelerator. A typical experiment runs for several months, and out of the billions of fusion reactions induced during this time, only one or two may result in the production of superheavy atoms. In all the rest, the fused nucleus breaks up immediately. SHIP is used to identify the small number of "good" reactions and separate them from this intense background.

Artificial synthesis of heavy elements

Elements up to uranium, atomic number 92, were created by these astronomical processes. Beyond that, the increasing electrical repulsion of the protons leads to shorter and shorter half-lives. Even if a supernova a billion years ago did create some quantity of an element such as Berkelium, number 97, there would be none left in the Earth's crust today. The heaviest elements have all been created by artificial fusion reactions in accelerators. The heaviest element that has been reported in a published scientific paper is 112, but as of 1999 scientists at Berkeley and Dubna have announced the creation of 114 and 118 as well.

Although the creation of a new element, i.e. an atom with a novel number of protons, has historically been considered a glamorous accomplishment, to the nuclear physicist the creation of an atom with a hitherto unobserved number of neutrons is equally important. The greatest neutron number reached so far is 179. One tantalizing goal of this type of research is the theoretical prediction that there might be an island of stability beyond the previously explored tip of the chart of the nuclei shown in section 2.8. Just as certain numbers of electrons lead to the chemical stability of the noble gases (helium, neon, argon, ...), certain numbers of neutrons and protons lead to a particularly stable packing of orbits. Calculations dating back to the 1960's have hinted that there might be relatively stable nuclei having approximately 114 protons and 184 neutrons. Proton number 114 has been achieved, and indeed displays an amazingly long half-life of 30 seconds. This may not seem like very long, but lifetimes in the microsecond range are more typical for the superheavy elements that have previously been discovered. There is even speculation that certain superheavy elements would be stable enough to be produced in quantities that could for instance be weighed and used in chemical reactions.

Summary

Selected Vocabulary

alpha particle	a form of radioactivity consisting of helium nuclei
beta particle	a form of radioactivity consisting of electrons
gamma ray	a form of radioactivity consisting of a very high-frequency form of light
proton	a positively charged particle, one of the types that nuclei are made of
neutron	an uncharged particle, the other types that nuclei are made of
isotope	one of the possible varieties of atoms of a given element, having a certain number of neutrons
atomic number	the number of protons in an atom's nucleus; determines what element it is
atomic mass	the mass of an atom
mass number	the number of protons plus the number of neutrons in a nucleus; approximately proportional to its atomic mass
strong nuclear force	the force that holds nuclei together against electrical repulsion
weak nuclear force	the force responsible for beta decay
beta decay	the radioactive decay of a nucleus via the reaction $n \rightarrow p + e^- + \bar{\nu}$ or $p \rightarrow n + e^+ + \nu$; so called because an electron or antielectron is also known as a beta particle
alpha decay	the radioactive decay of a nucleus via emission of an alpha particle
fission	the radioactive decay of a nucleus by splitting into two parts
fusion	a nuclear reaction in which two nuclei stick together to form one bigger nucleus
millirem	a unit for measuring a person's exposure to radioactivity

Notation

e^-	an electron
e^+	an antielectron; just like an electron, but with positive charge
n	a neutron
p	a proton
ν	a neutrino
$\bar{\nu}$	an antineutrino

Notation Used in Other Books

Z	atomic number (number of protons in a nucleus)
N	number of neutrons in a nucleus
A	mass number (N+Z)

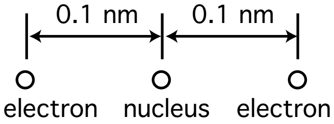
Summary

Rutherford and Marsden observed that some alpha particles from a beam striking a thin gold foil came back at angles up to 180 degrees. This could not be explained in the then-favored raisin-cookie model of the atom, and led to the adoption of the planetary model of the atom, in which the electrons orbit a tiny, positively-charged nucleus. Further experiments showed that the nucleus itself was a cluster of positively-charged protons and uncharged neutrons.

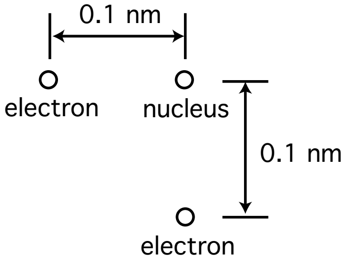
Radioactive nuclei are those that can release energy. The most common types of radioactivity are alpha decay (the emission of a helium nucleus), beta decay (the transformation of a neutron into a proton or vice-versa), and gamma decay (the emission of a type of very-high-frequency light). Stars are powered by nuclear fusion reactions, in which two light nuclei collide and form a bigger nucleus, with the release of energy.

Human exposure to ionizing radiation is measured in units of millirem. The typical person is exposed to about 100 mrem worth of natural background radiation per year.

Homework Problems



Problem 1.



Problem 2.

1 ✓. A helium atom finds itself momentarily in this arrangement. Find the direction and magnitude of the force acting on the right-hand electron. The two protons in the nucleus are so close together (~1 fm) that you can consider them as being right on top of each other.

2 ✓. The helium atom of problem 1 has some new experiences, goes through some life changes, and later on finds itself in the configuration shown here. What are the direction and magnitude of the force acting on the bottom electron? (Draw a sketch to make clear the definition you are using for the angle that gives direction.)

3. Suppose you are holding your hands in front of you, 10 cm apart.

(a) Estimate the total number of electrons in each hand.

(b) Estimate the total repulsive force of all the electrons in one hand on all the electrons in the other.

(c) Why don't you feel your hands repelling each other?

(d) Estimate how much the charge of a proton could differ in magnitude from the charge of an electron without creating a noticeable force between your hands.

4 ✓. Suppose that a proton in a lead nucleus wanders out to the surface of the nucleus, and experiences a strong nuclear force of about 8 kN from the nearby neutrons and protons pulling it back in. Compare this numerically to the repulsive electrical force from the other protons, and verify that the net force is attractive. A lead nucleus is very nearly spherical, and is about 6.5 fm in radius.

5. The subatomic particles called muons behave exactly like electrons, except that a muon's mass is greater by a factor of 206.77. Muons are continually bombarding the Earth as part of the stream of particles from space known as cosmic rays. When a muon strikes an atom, it can displace one of its electrons. If the atom happens to be a hydrogen atom, then the muon takes up an orbit that is on the average 206.77 times closer to the proton than the orbit of the ejected electron. How many times greater is the electric force experienced by the muon than that previously felt by the electron?

S A solution is given in the back of the book.	★ A difficult problem.
✓ A computerized answer check is available.	∫ A problem that requires calculus.

3 Circuits, Part 1

Madam, what good is a baby?

Michael Faraday, when asked by Queen Victoria
what the electrical devices in his lab were good for

A few years ago, my wife and I bought a house with Character, Character being a survival mechanism that houses have evolved in order to convince humans to agree to much larger mortgage payments than they'd originally envisioned. Anyway, one of the features that gives our house Character is that it possesses, built into the wall of the family room, a set of three pachinko machines. These are Japanese gambling devices sort of like vertical pinball machines. (The legal papers we got from the sellers hastened to tell us that they were "for amusement purposes only.") Unfortunately, only one of the three machines was working when we moved in, and it soon died on us. Having become a pachinko addict, I decided to fix it, but that was easier said than done. The inside is a veritable Rube Goldberg mechanism of levers, hooks, springs, and chutes. My hormonal pride, combined with my Ph.D. in physics, made me certain of success, and rendered my eventual utter failure all the more demoralizing.

Contemplating my defeat, I realized how few complex mechanical devices I used from day to day. Apart from our cars and my saxophone, every technological tool in our modern life-support system was electronic rather than mechanical.

3.1 Current

Gymnotus carapo, a knifefish, uses electrical signals to sense its environment and to communicate with others of its species.

Unity of all types of electricity

We are surrounded by things we have been *told* are "electrical," but it's far from obvious what they have in common to justify being grouped together. What relationship is there between the way socks cling together and the way a battery lights a lightbulb? We have been told that both an electric eel and our own brains are somehow electrical in nature, but what do they have in common?

British physicist Michael Faraday (1791-1867) set out to address this problem. He investigated electricity from a variety of sources — including electric eels! — to see whether they could all produce the same effects, such as shocks and sparks, attraction and repulsion. "Heating" refers, for example, to the way a lightbulb filament gets hot enough to glow and emit light. Magnetically induction is an effect discovered by Faraday himself that connects electricity and magnetism. We will not study this effect, which is the basis for the electric generator, in detail until later in the book.

source of electricity	shocks	sparks	attraction and repulsion	heating
rubbing	✓	✓	✓	✓
battery	✓	✓	✓	✓
animal	✓	✓	(✓)	✓
magnetically induced	✓	✓	✓	✓

The table shows a summary of some of Faraday's results. Check marks indicate that Faraday or his close contemporaries were able to verify that a particular source of electricity was capable of producing a certain effect. (They evidently failed to demonstrate attraction and repulsion between objects charged by electric eels, although modern workers have studied these species in detail and been able to understand all their electrical characteristics on the same footing as other forms of electricity.)

Faraday's results indicate that there is nothing fundamentally different about the types of electricity supplied by the various sources. They are all able to produce a wide variety of identical effects. Wrote Faraday, "The general conclusion which must be drawn from this collection of facts is that electricity, whatever may be its source, is identical in its nature."

If the types of electricity are the same thing, what thing is that? The answer is provided by the fact that all the sources of electricity can cause objects to repel or attract each other. We use the word “charge” to describe the property of an object that allows it to participate in such electrical forces, and we have learned that charge is present in matter in the form of nuclei and electrons. Evidently all these electrical phenomena boil down to the motion of charged particles in matter.

Electric current

If the fundamental phenomenon is the motion of charged particles, then how can we define a useful numerical measurement of it? We might describe the flow of a river simply by the velocity of the water, but velocity will not be appropriate for electrical purposes because we need to take into account how much charge the moving particles have, and in any case there are no practical devices sold at Radio Shack that can tell us the velocity of charged particles. Experiments show that the intensity of various electrical effects is related to a different quantity: the number of coulombs of charge that pass by a certain point per second. By analogy with the flow of water, this quantity is called the electric *current*, *I*. Its units of coulombs/second are more conveniently abbreviated as amperes, 1 A=1 C/s. (In informal speech, one usually says “amps.”)

The main subtlety involved in this definition is how to account for the two types of charge. The stream of water coming from a hose is made of atoms containing charged particles, but it produces none of the effects we associate with electric currents. For example, you do not get an electrical shock when you are sprayed by a hose. This type of experiment shows that the effect created by the motion of one type of charged particle can be canceled out by the motion of the opposite type of charge in the same direction. In water, every oxygen atom with a charge of $+8e$ is surrounded by eight electrons with charges of $-e$, and likewise for the hydrogen atoms.

We therefore refine our definition of current as follows:

definition of electric current

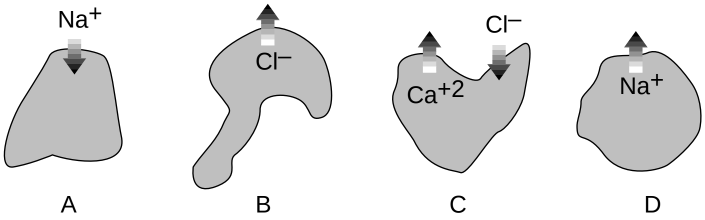
When charged particles are exchanged between regions of space A and B, the electric current flowing from A to B is defined as

$$I = \Delta q \,/\, \Delta t \quad ,$$

where Δq is the change in the total charge of region B.

In the garden hose example, your body picks up equal amounts of positive and negative charge, resulting in no change in your total charge, so the electrical current flowing into you is zero.

Mathematically, this is a definition involving a rate of change, very similar to examples such as the rate of change of velocity, $a=\Delta v/\Delta t$, or the rate of change of angular momentum, $\tau=\Delta L/\Delta t$. You can therefore recycle the hard lessons you learned in those cases: this definition will only make sense when the rate of change is constant, and when the rate of change is not constant, we have to change the definition to refer to the slope of the tangent line on a graph of q versus t.



Example: Ions moving across a cell membrane
Question: The figures show ions, labeled with their charges, moving in or out through the membranes of three cells. If the ions all cross the membranes during the same interval of time, how would the currents into the cells compare with each other?
Solution:

Cell A has positive current going into it because its charge is increased, i.e. has a positive value of Δq.

Cell B has the same current as cell A, because by losing one unit of negative charge it also ends up increasing its own total charge by one unit.

Cell C's total charge is reduced by three units, so it has a large negative current going into it.

Cell D loses one unit of charge, so it has a small negative current into it.

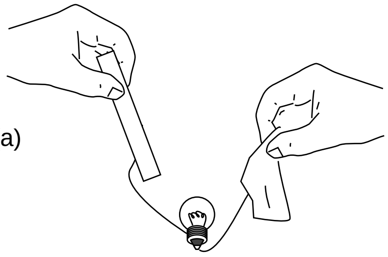
It may seem strange to say that a negatively charged particle going one way creates a current going the other way, but this is quite ordinary. As we will see, currents flow through metal wires via the motion of electrons, which are negatively charged, so the direction of motion of the electrons in a circuit is always opposite to the direction of the current. Of course it would have been convenient of Benjamin Franklin had defined the positive and negative signs of charge the opposite way, since so many electrical devices are based on metal wires.

Example: Number of electrons flowing through a lightbulb
Question: If a lightbulb has 1.0 A flowing through it, how many electrons will pass through the filament in 1.0 s?
Solution: We are only calculating the number of electrons that flow, so we can ignore the positive and negative signs. Solving the definition of current, $I = \Delta q / \Delta t$, for $\Delta q = I \Delta t$ gives a charge of 1.0 C flowing in this time interval. The number of electrons is

$$\begin{aligned}\text{number of electrons} &= \text{colombs} \times \frac{\text{electrons}}{\text{coulomb}} \\ &= \text{colombs} / \left(\frac{\text{coulombs}}{\text{electron}}\right) \\ &= 1.0\ \text{C} / e \\ &= 6.2\text{x}10^{18}\end{aligned}$$

3.2 Circuits

(a)



How can we put electric currents to work? The only method of controlling electric charge we have studied so far is to charge different substances, e.g. rubber and fur, by rubbing them against each other. Figure (a) shows an attempt to use this technique to light a lightbulb. This method is unsatisfactory. True, current will flow through the bulb, since electrons can move through metal wires, and the excess electrons on the rubber rod will therefore come through the wires and bulb due to the attraction of the positively charged fur and the repulsion of the other electrons. The problem is that after a zillionth of a second of current, the rod and fur will both have run out of charge. No more current will flow, and the lightbulb will go out.

(b)

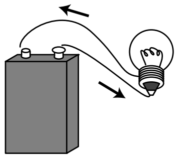
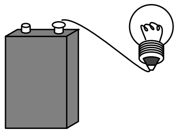


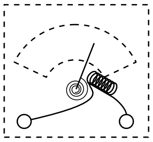
Figure (b) shows a setup that works. The battery pushes charge through the circuit, and recycles it over and over again. (We will have more to say later in this chapter about how batteries work.) This is called a *complete circuit*. Today, the electrical use of the word "circuit" is the only one the springs to mind for most people, but the original meaning was to travel around and make a round trip, as when a circuit court judge would ride around the boondocks, dispensing justice in each town on a certain date.

(c)

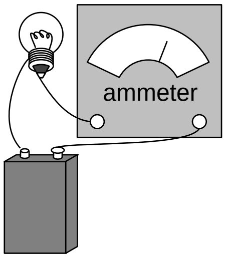


Note that an example like (c) does not work. The wire will quickly begin acquiring a net charge, because it has no way to get rid of the charge flowing into it. The repulsion of this charge will make it more and more difficult to send any more charge in, and soon the electrical forces exerted by the battery will be canceled out completely. The whole process would be over so quickly that the filament would not even have enough time to get hot and glow. This is known as an *open circuit*. Exactly the same thing would happen if the complete circuit of figure (b) was cut somewhere with a pair of scissors, and in fact that is essentially how an ordinary light switch works: by opening up a gap in the circuit.

(d)



(e)



The definition of electric current we have developed has the great virtue that it is easy to measure. In practical electrical work, one almost always measures current, not charge. The instrument used to measure current is called an *ammeter*. A simplified ammeter, (d), simply consists of a coiled-wire magnet whose force twists an iron needle against the resistance of a spring. The greater the current, the greater the force. Although the construction of ammeters may differ, their use is always the same. We break into the path of the electric current and interpose the meter like a tollbooth on a road, (e). There is still a complete circuit, and as far as the battery and bulb are concerned, the ammeter is just another segment of wire.

Does it matter where in the circuit we place the ammeter? Could we, for instance, have put it in the left side of the circuit instead of the right? Conservation of charge tells us that this can make no difference. Charge is not destroyed or "used up" by the lightbulb, so we will get the same current reading on either side of it. What is "used up" is energy stored in the battery, which is being converted into heat and light energy.

3.3 Voltage

The volt unit

Electrical circuits can be used for sending signals, storing information, or doing calculations, but their most common purpose by far is to manipulate energy, as in the battery-and-bulb example of the previous section. We know that lightbulbs are rated in units of watts, i.e. how many joules per second of energy they can convert into heat and light, but how would this relate to the flow of charge as measured in amperes? By way of analogy, suppose your friend, who didn't take physics, can't find any job better than pitching bales of hay. The number of calories he burns per hour will certainly depend on how many bales he pitches per minute, but it will also be proportional to how much mechanical work he has to do on each bale. If his job is to toss them up into a hayloft, he will got tired a lot more quickly than someone who merely tips bales off a loading dock into trucks. In metric units,

$$\frac{\text{joules}}{\text{second}} = \frac{\text{haybales}}{\text{second}} \times \frac{\text{joules}}{\text{haybale}} \quad .$$

Similarly, the rate of energy transformation by a battery will not just depend on how many coulombs per second it pushes through a circuit but also on how much mechanical work it has to do on each coulomb of charge:

$$\frac{\text{joules}}{\text{second}} = \frac{\text{coulombs}}{\text{second}} \times \frac{\text{joules}}{\text{coulomb}}$$

or

$$\text{power} = \text{current} \times \text{work per unit charge} \quad .$$

Units of joules per coulomb are abbreviated as *volts*, 1 V=1 J/C, named after the Italian physicist Count Volta. Everyone knows that batteries are rated in units of volts, but the voltage concept is more general than that; it turns out that voltage is a property of every point in space. To gain more insight, let's think more carefully about what goes on in the battery and bulb circuit.

The voltage concept in general

To do work on a charged particle, the battery apparently must be exerting forces on it. How does it do this? Well, the only thing that can exert an electrical force on a charged particle is another charged particle. It's as though the haybales were pushing and pulling each other into the hayloft! This is potentially a horribly complicated situation. Even if we knew how much excess positive or negative charge there was at every point in the circuit (which realistically we don't) we would have to calculate zillions of forces using Coulomb's law, perform all the vector additions, and finally calculate how much work was being done on the charges as they moved along. To make things even more scary, there is more than one type of charged particle that moves: electrons are what move in the wires and the bulb's filament, but ions are the moving charge carriers inside the battery. Luckily, there are two ways in which we can simplify things:

The situation is unchanging. Unlike the imaginary setup in which we attempted to light a bulb using a rubber rod and a piece of fur, this circuit maintains itself in a steady state (after perhaps a microsecond-long period of settling down after the circuit is first assembled). The current is steady, and as charge flows out of any area of the circuit it is replaced by the same amount of charge flowing in. The amount of excess positive or negative charge in any part of the circuit therefore stays constant. Similarly, when we watch a river flowing, the water goes by but the river doesn't disappear.

Force depends only on position. Since the charge distribution is not changing, the total electrical force on a charged particle depends only on its own charge and on its location. If another charged particle of the same type visits the same location later on, it will feel exactly the same force.

The second observation tells us that there is nothing all that different about the experience of one charged particle as compared to another's. If we single out one particle to pay attention to, and figure out the amount of work done on it by electrical forces as it goes from point A to point B along a certain path, then this is the same amount of work that will be done on any other charged particles of the same type as it follows the same path. For the sake of visualization, let's think about the path that starts at one terminal of the battery, goes through the light bulb's filament, and ends at the other terminal. When an object experiences a force that depends only on its position (and when certain other, technical conditions are satisfied), we can define a potential energy associated with the position of that object. The amount of work done on the particle by electrical forces as it moves from A to B equals the drop in electrical potential energy between A and B. This potential energy is what is being converted into other forms of energy such as heat and light. We therefore define voltage in general as electrical potential energy per unit charge:

definition of voltage difference
The difference in voltage between two points in space is defined as

$$\Delta V = \Delta PE_{\text{elec}} / q \quad ,$$

where ΔPE_{elec} is the change in the potential energy of a particle with charge q as it moves from the initial point to the final point.

The amount of power dissipated (i.e. rate at which energy is transformed by the flow of electricity) is then given by the equation

$$P = I\Delta V \quad .$$

Example: Energy stored in a battery

Question: My old camcorder runs off of a big lead-acid battery that is labeled 12 volts, 4 AH. The “AH” stands for ampere-hours. What is the maximum amount of energy the battery can store?

Solution: An ampere-hour is a unit of current multiplied by a unit of time. Current is charge per unit time, so an ampere-hour is in fact a funny unit of *charge*:

$$\begin{aligned}(1\text{ A})(1\text{ hour}) &= (1\text{ C/s})(3600\text{ s})\\ &= 3600\text{ C}\end{aligned}$$

Now 3600 C is a huge number of charged particles, but the total loss of potential energy will just be their total charge multiplied by the voltage difference across which they move:

$$\begin{aligned}\Delta PE_{\text{elec}} &= q\,\Delta V\\ &= (3600\text{ C})(12\text{ V})\\ &= 43\text{ kJ}\end{aligned}$$

Example: Units of volt-amps

Question: Doorbells are often rated in volt-amps. What does this combination of units mean?

Solution: Current times voltage gives units of power, $P{=}I\,\Delta V$, so volt-amps are really just a nonstandard way of writing watts. They are telling you how much power the doorbell requires.

Example: Power dissipated by a battery and bulb

Question: If a 9.0-volt battery causes 1.0 A to flow through a lightbulb, how much power is dissipated?

Solution: The voltage rating of a battery tells us what voltage difference ΔV it is designed to maintain between its terminals.

$$\begin{aligned}P &= I\;\Delta V\\ &= 9.0\text{ A}\cdot\text{V}\\ &= 9.0\,\frac{\text{C}}{\text{s}}\cdot\frac{\text{J}}{\text{C}}\\ &= 9.0\text{ J/s}\\ &= 9.0\text{ W}\end{aligned}$$

The only nontrivial thing in this problem was dealing with the units. One quickly gets used to translating common combinations like A·V into simpler terms.

Here are a few questions and answers about the voltage concept.

Question: OK, so what *is* voltage, really?

Answer: A device like a battery has positive and negative charges inside it that push other charges around the outside circuit. A higher-voltage battery has denser charges in it, which will do more work on each charged particle that moves through the outside circuit.

To use a gravitational analogy, we can put a paddlewheel at the bottom of either a tall waterfall or a short one, but a kg of water that falls through the greater gravitation potential energy difference will have more energy to give up to the paddlewheel at the bottom.

Question: Why do we define voltage as electrical potential energy divided by charge, instead of just defining it as electrical potential energy?
Answer: One answer is that it's the only definition that makes the equation $P = I\ \Delta V$ work. A more general answer is that we want to be able to define a voltage difference between any two points in space without having to know in advance how much charge the particles moving between them will have. If you put a nine-volt battery on your tongue, then the charged particles that move across your tongue and give you that tingly sensation are not electrons but ions, which may have charges of $+e$, $-2e$, or practically anything. The manufacturer probably expected the battery to be used mostly in circuits with metal wires, where the charged particles that flowed would be electrons with charges of $-e$. If the ones flowing across your tongue happen to have charges of $-2e$, the potential energy difference for them will be twice as much, but dividing by their charge of $-2e$ in the definition of voltage will still give a result of 9 V.

Question: Are there two separate roles for the charged particles in the circuit, a type that sits still and exerts the forces, and another that moves under the influence of those forces?
Answer: No. Every charged particle simultaneously plays both roles. Newton's third law says that any particle that has an electrical forces acting on it must also be exerting an electrical force back on the other particle. There are no "designated movers" or "designated force-makers."

Question: Why does the definition of voltage only refer to voltage *differences*?
Answer: It's perfectly OK to define voltage as $V = PE_{elec}/q$. But recall that it is only *differences* in potential energy that have direct physical meaning in physics. Similarly, voltage differences are really more useful than absolute voltages. A voltmeter measures voltage differences, not absolute voltages.

Discussion Questions

A. A roller coaster is sort of like an electric circuit, but it uses gravitational forces on the cars instead of electric ones. What would a high-voltage roller coaster be like? What would a high-current roller coaster be like?
B. Criticize the following statements:
"He touched the wire, and 10000 volts went through him."
"That battery has a charge of 9 volts."
"You used up the charge of the battery."
C. When you touch a 9-volt battery to your tongue, both positive and negative ions move through your saliva. Which ions go which way?
D. I once touched a piece of physics apparatus that had been wired incorrectly, and got a several-thousand-volt voltage difference across my hand. I was not injured. For what possible reason would the shock have had insufficient power to hurt me?

3.4 Resistance

Resistance

So far we have simply presented it as an observed fact that a battery-and-bulb circuit quickly settles down to a steady flow, but why should it? Newton's second law, $a=F/m$, would seem to predict that the steady forces on the charged particles should make them whip around the circuit faster and faster. The answer is that as charged particles move through matter, there are always forces, analogous to frictional forces, that resist the motion. These forces need to be included in Newton's second law, which is really $a=F_{total}/m$, not $a=F/m$. If, by analogy, you push a crate across the floor at constant speed, i.e. with zero acceleration, the total force on it must be zero. After you get the crate going, the floor's frictional force is exactly canceling out your force. The chemical energy stored in your body is being transformed into heat in the crate and the floor, and no longer into an increase in the crate's kinetic energy. Similarly, the battery's internal chemical energy is converted into heat, not into perpetually increasing the charged particles' kinetic energy. Changing energy into heat may be a nuisance in some circuits, such as a computer chip, but it is vital in a lightbulb, which must get hot enough to glow. Whether we like it or not, this kind of heating effect is going to occur any time charged particles move through matter.

What determines the amount of heating? One flashlight bulb designed to work with a 9-volt battery might be labeled 1.0 watts, another 5.0. How does this work? Even without knowing the details of this type of friction at the atomic level, we can relate the heat dissipation to the amount of current that flows via the equation $P=I\Delta V$. If the two flashlight bulbs can have two different values of P when used with a battery that maintains the same ΔV, it must be that the 5.0-watt bulb allows five times more current to flow through it.

For many substances, including the tungsten from which lightbulb filaments are made, experiments show that the amount of current that will flow through it is directly proportional to the voltage difference placed across it. For an object made of such a substance, we define its electrical *resistance* as

definition of resistance

If an object inserted in a circuit displays a current flow which is proportional to the voltage difference across it, we define its *resistance* as the constant ratio

$$R = \Delta V / I \quad .$$

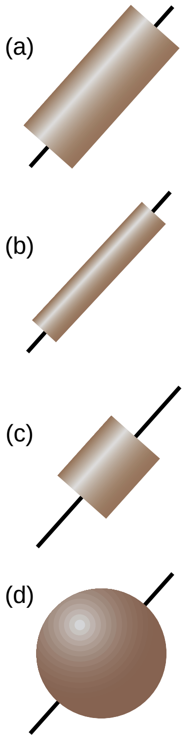
The units of resistance are volts/ampere, usually abbreviated as ohms, symbolized with the capital Greek letter omega, **Ω**.

Example: Resistance of a lightbulb
Question: A flashlight bulb powered by a 9-volt battery has a resistance of 10 Ω. How much current will it draw?
Solution: Solving the definition of resistance for *I*, we find

$$\begin{aligned} I &= \Delta V / R \\ &= 0.9\ \mathrm{V}/\Omega \\ &= 0.9\ \mathrm{V}/(\mathrm{V}/\mathrm{A}) \\ &= 0.9\ \mathrm{A} \end{aligned}$$

Ohm's law states that many substances, including many solids and some liquids, display this kind of behavior, at least for voltages that are not too large. The fact that Ohm's law is called a "law" should not be taken to mean that all materials obey it, or that it has the same fundamental importance as Newton's laws, for example. Materials are called *ohmic* or *nonohmic*, depending on whether they obey Ohm's law. If objects of the same size and shape made from two different ohmic materials have different resistances, we can say that one material is more resistive than the other, or equivalently that it is less conductive. Materials, such as metals, that are very conductive are said to be good *conductors*. Those that are extremely poor conductors, for example wood or rubber, are classified as *insulators*. There is no sharp distinction between the two classes of materials. Some, such as silicon, lie midway between the two extremes, and are called semiconductors.

On an intuitive level, we can understand the idea of resistance by making the sounds "hhhhhh" and "ffffff." To make air flow out of your mouth, you use your diaphragm to compress the air in your chest. The pressure difference between your chest and the air outside your mouth is analogous to a voltage difference. When you make the "h" sound, you form your mouth and throat in a way that allows air to flow easily. The large flow of air is like a large current. Dividing by a large current in the definition of resistance means that we get a small resistance. We say that the small resistance of your mouth and throat allows a large current to flow. When you make the "f" sound, you increase the resistance and cause a smaller current to flow.



Note that although the resistance of an object depends on the substance it is made of, we cannot speak simply of the "resistance of gold" or the "resistance of wood." The figures show four examples of objects that have had wires attached at the ends as electrical connections. If they were made of the same substance, they would all nevertheless have different resistances because of their different sizes and shapes. A more detailed discussion will be more natural in the context of the following chapter, but it should not be too surprising that the resistance of (b) will be greater than that of (a) — the image of water flowing through a pipe, however incorrect, gives us the right intuition. Object (c) will have a smaller resistance than (a) because the charged particles have less of it to get through.

A superconducting segment of the ATLAS accelerator at Argonne National Laboratory near Chicago. It is used to accelerate beams of ions to a few percent of the speed of light for nuclear physics reasearch. The shiny silver-colored surfaces are made of the element niobium, which is a superconductor at relatively high temperatures compared to other metals — relatively high meaning the temperature of liquid helium! The beam of ions passes through the holes in the two small cylinders on the ends of the curved rods. Charge is shuffled back and forth between them at a frequency of 12 million cycles per second, so that they take turns being positive and negative. The positively charged beam consists of short spurts, each timed so that when it is in one of the segments it will be pulled forward by negative charge on the cylinder in front of it and pushed forward by the positively charged one behind. The huge currents involved would quickly melt any metal that was not superconducting, but in a superconductor they produce no heat at all. My own PhD thesis was based on data from this accelerator.

Superconductors

All materials display some variation in resistance according to temperature (a fact that is used in thermostats to make a thermometer that can be easily interfaced to an electric circuit). More spectacularly, most metals have been found to exhibit a sudden change to *zero* resistance when cooled to a certain critical temperature. They are then said to be superconductors. Theoretically, superconductors should make a great many exciting devices possible, for example coiled-wire magnets that could be used to levitate trains. In practice, the critical temperatures of all metals are very low, and the resulting need for extreme refrigeration has made their use uneconomical except for such specialized applications as particle accelerators for physics research.

But scientists have recently made the surprising discovery that certain ceramics are superconductors at less extreme temperatures. The technological barrier is now in finding practical methods for making wire out of these brittle materials. Wall Street is currently investing billions of dollars in developing superconducting devices for cellular phone relay stations based on these materials.

There is currently no satisfactory theory of superconductivity in general, although superconductivity in metals is understood fairly well. Unfortunately I have yet to find an explanation of superconductivity in metals that works at the introductory level.

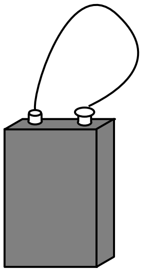
Constant voltage throughout a conductor

The idea of a superconductor leads us to the question of how we should expect an object to behave if it is made of a very good conductor. Superconductors are an extreme case, but often a metal wire can be thought of as a perfect conductor, for example if the parts of the circuit other than the wire are made of much less conductive materials. What happens if R equals zero in the equation $R=\Delta V/I$? The result of dividing two numbers can only be zero if the number on top equals zero. This tells us that if we pick any two points in a perfect conductor, the voltage difference between them must be zero. In other words, the entire conductor must be at the same voltage.

Constant voltage means that no work would be done on a charge as it moved from one point in the conductor to another. If zero work was done only along a certain path between two specific points, it might mean that positive work was done along part of the path and negative work along the rest, resulting in a cancellation. But there is no way that the work could come out to be zero for all possible paths unless the electrical force on a charge was in fact zero at every point. Suppose, for example, that you build up a static charge by scuffing your feet on a carpet, and then you deposit some of that charge onto a doorknob, which is a good conductor. How can all that charge be in the doorknob without creating any electrical force at any point inside it? The only possible answer is that the charge moves around until it has spread itself into just the right configuration so that the forces exerted by all the little bits of excess surface charge on any charged particle within the doorknob exactly canceled out.

We can explain this behavior if we assume that the charge placed on the doorknob eventually settles down into a stable equilibrium. Since the doorknob is a conductor, the charge is free to move through it. If it was free to move and any part of it did experience a nonzero total force from the rest of the charge, then it would move, and we would not have an equilibrium.

It also turns out that charge placed on a conductor, once it reaches its equilibrium configuration, is entirely on the surface, not on the interior. We will not prove this fact formally, but it is intuitively reasonable. Suppose, for instance, that the net charge on the conductor is negative, i.e. it has an excess of electrons. These electrons all repel each other, and this repulsion will tend to push them onto the surface, since being on the surface allows them to be as far apart as possible.



Short-circuiting a battery. Warning: you can burn yourself this way or start a fire! If you want to try this, try making the connection only very briefly, use a low-voltage battery, and avoid touching the battery or the wire, both of which will get hot.

Short circuits

So far we have been assuming a perfect conductor. What if it is a good conductor, but not a perfect one? Then we can solve for $\Delta V=IR$. An ordinary-sized current will make a very small result when we multiply it by the resistance of a good conductor such as a metal wire. The voltage throughout the wire will then be nearly constant. If, on the other hand, the current is extremely large, we can have a significant voltage difference. This is what happens in a *short-circuit:* a circuit in which a low-resistance pathway connects the two sides of a voltage source. Note that this is much more specific than the popular use of the term to indicate any electrical malfunction at all. If, for example, you short-circuit a 9-volt battery as shown in the figure, you will produce perhaps a thousand amperes of current, leading to a very large value of $P=I\Delta V$. The wire gets hot!

Self-Check

What would happen to the battery in this kind of short circuit?

The large amount of power means a high rate of conversion of the battery's chemical energy into heat. The battery will quickly use up all its energy, i.e. "burn out."

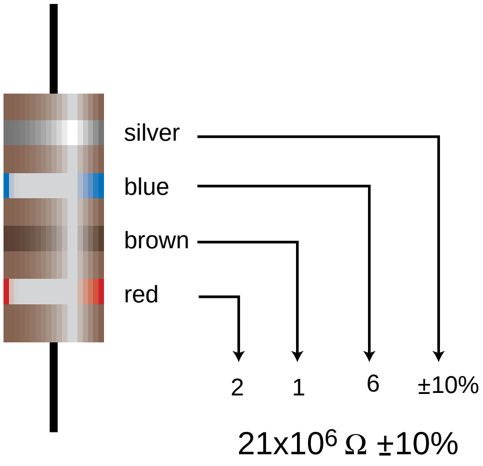
Resistors

Inside any electronic gadget you will see quite a few little circuit elements like the one shown below. These *resistors* are simply a cylinder of ohmic material with wires attached to the end.

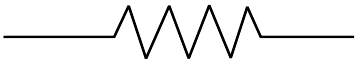
At this stage, most students have a hard time understanding why resistors would be used inside a radio or a computer. We obviously want a lightbulb or an electric stove to have a circuit element that resists the flow of electricity and heats up, but heating is undesirable in radios and computers. Without going too far afield, let's use a mechanical analogy to get a general idea of why a resistor would be used in a radio.

The main parts of a radio receiver are an antenna, a tuner for selecting the frequency, and an amplifier to strengthen the signal sufficiently to drive a speaker. The tuner resonates at the selected frequency, just as in the examples of mechanical resonance discussed in book 3. The behavior of a mechanical resonator depends on three things: its inertia, its stiffness, and the amount of friction or damping. The first two parameters locate the peak of the resonance curve, while the damping determines the width of the resonance. In the radio tuner we have an electrically vibrating system that resonates at a particular frequency. Instead of a physical object moving back and forth, these vibrations consist of electrical currents that flow first in one direction and then in the other. In a mechanical system, damping means taking energy out of the vibration in the form of heat, and exactly the same idea applies to an electrical system: the resistor supplies the damping, and therefore controls the width of the resonance. If we set out to eliminate all resistance in the tuner circuit, by not building in a resistor and by somehow getting rid of all the inherent electrical resistance of the wires, we would

color	meaning
black	0
brown	1
red	2
orange	3
yellow	4
green	5
blue	6
violet	7
gray	8
white	9
silver	±10%
gold	±5%



Color codes for resistors (left), and an example of their use (top).



The symbol used in schematics to represent a resistor.

have a useless radio. The tuner's resonance would be so narrow that we could never get close enough to the right frequency to bring in the station. The roles of inertia and stiffness are played by other circuit elements we have not discusses (a capacitor and a coil).

Applications

Many electrical devices are based on electrical resistance and Ohm's law, even if they do not have little components in them that look like the usual resistor. The following are some examples.

Lightbulb

There is nothing special about a lightbulb filament — you can easily make a lightbulb by cutting a narrow waist into a metallic gum wrapper and connecting the wrapper across the terminals of a 9-volt battery. The trouble is that it will instantly burn out. Edison solved this technical challenge by encasing the filament in an evacuated bulb, which prevented burning, since burning requires oxygen.

Polygraph

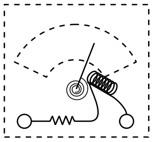
The polygraph, or "lie detector," is really just a set of meters for recording physical measures of the subject's psychological stress, such as sweating and quickened heartbeat. The real-time sweat measurement works on the principle that dry skin is a good insulator, but sweaty skin is a conductor. Of course a truthful subject may become nervous simply because of the situation, and a practiced liar may not even break a sweat. The method's practitioners claim that they can tell the difference, but you should think twice before allowing yourself to be polygraph tested. Most U.S. courts exclude all polygraph evidence, but some employers attempt to screen out dishonest employees by polygraph testing job applicants, an abuse that ranks with such pseudoscience as handwriting analysis.

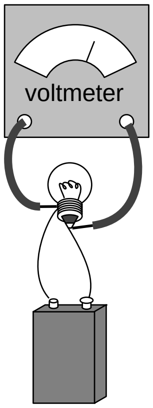
Fuse

A fuse is a device inserted in a circuit tollbooth-style in the same manner as an ammeter. It is simply a piece of wire made of metals having a relatively low melting point. If too much current passes through the fuse, it melts, opening the circuit. The purpose is to make sure that the building's wires do not carry so much current that they themselves will get hot enough to start a fire. Most modern houses use circuit breakers instead of fuses, although fuses are still common in cars and small devices. A circuit breaker is a switch operated by a coiled-wire magnet, which opens the circuit when enough current flows. The advantage is that once you turn off some of the appliances that were sucking up too much current, you can immediately flip the switch closed. In the days of fuses, one might get caught without a replacement fuse, or even be tempted to stuff aluminum foil in as a replacement, defeating the safety feature.

Voltmeter

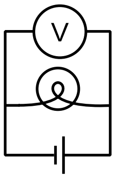
A voltmeter is nothing more than an ammeter with an additional high-value resistor through which the current is also forced to flow. Ohm's law relates the current through the resistor is related directly to the voltage difference across it, so the meter can be calibrated in units of volts based on the known value of the resistor. The voltmeter's two probes are touched to the two locations in a circuit between which we wish to measure the voltage difference, (b). Note how cumbersome this type of drawing is, and how difficult it can be to tell what is connected to what. This is why electrical drawing are usually shown in schematic form. Figure (c) is a schematic representation of figure (b).


(a) A simplified diagram of how a voltmeter works.

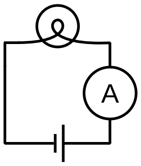


(b) Measuring the voltage difference across a lightbulb.

The setups for measuring current and voltage are different. When we are measuring current, we are finding "how much stuff goes through," so we place the ammeter where all the current is forced to go through it. Voltage, however, is not "stuff that goes through," it is a measure of potential energy. If an ammeter is like the meter that measures your water use, a voltmeter is like a measuring stick that tells you how high a waterfall is, so that you can determine how much energy will be released by each kilogram of falling water. We do not want to force the water to go through the measuring stick! The arrangement in figure (c) is a *parallel* circuit: one in there are "forks in the road" where some of the current will flow one way and some will flow the other. Figure (d) is said to be wired in *series:* all the current will visit all the circuit elements one after the other. We will deal with series and parallel circuits in more detail in the following chapter.



(c) The same setup drawn in schematic form.



(d) The setup for measuring current is different.

If you inserted a voltmeter incorrectly, in series with the bulb and battery, its large internal resistance would cut the current down so low that the bulb would go out. You would have severely disturbed the behavior of the circuit by trying to measure something about it.

Incorrectly placing an ammeter in parallel is likely to be even more disconcerting. The ammeter has nothing but wire inside it to provide resistance, so given the choice, most of the current will flow through it rather than through the bulb. So much current will flow through the ammeter, in fact, that there is a danger of burning out the battery or the meter or both! For this reason, most ammeters have fuses or circuit breakers inside. Some models will trip their circuit breakers and make an audible alarm in this situation, while others will simply blow a fuse and stop working until you replace it.

Discussion Question

A. In figure (b), would it make any difference in the voltage measurement if we touched the voltmeter's probes to different points along the same segments of wire?

B. Explain why it would be incorrect to define resistance as the amount of charge the resistor allows to flow.

3.5 Current-Conducting Properties of Materials

Ohm's law has a remarkable property, which is that current will flow even in response to a voltage difference that is as small as we care to make it. In the analogy of pushing a crate across a floor, it is as though even a flea could slide the crate across the floor, albeit at some very low speed. The flea cannot do this because of static friction, which we can think of as an effect arising from the tendency of the microscopic bumps and valleys in the crate and floor to lock together. The fact that Ohm's law holds for nearly all solids has an interesting interpretation: at least some of the electrons are not "locked down" at all to any specific atom.

More generally we can ask how charge actually flows in various solids, liquids, and gases. This will lead us to the explanations of many interesting phenomena, including lightning, the bluish crust that builds up on the terminals of car batteries, and the need for electrolytes in sports drinks.

Solids

In atomic terms, the defining characteristic of a solid is that its atoms are packed together, and the nuclei cannot move very far from their equilibrium positions. It makes sense, then, that electrons, not ions, would be the charge carriers when currents flow in solids. This fact was established experimentally by Tolman and Stewart, in an experiment in which they spun a large coil of wire and then abruptly stopped it. They observed a current in the wire immediately after the coil was stopped, which indicated that charged particles that were not permanently locked to a specific atom had continued to move because of their own inertia, even after the material of the wire in general stopped. The direction of the current showed that it was negatively charged particles that kept moving. The current only lasted for an instant, however; as the negatively charged particles collected at the downstream end of the wire further particles were prevented joining them due to their electrical repulsion, as well as the attraction from the upstream end, which was left with a net positive charge. Tolman and Stewart were even able to determine the mass-to-charge ratio of the particles. We need not go into the details of the analysis here, but a particles with high mass would be difficult to decelerate, leading to a stronger and longer pulse of current, while particles with high charge would feel stronger electrical forces decelerating them, which would cause a weaker and shorter pulse. The mass-to-charge ratio thus determined was consistent with the m/q of the electron to within the accuracy of the experiment, which essentially established that the particles were electrons.

The fact that only electrons carry current in solids, not ions, has many important implications. For one thing, it explains why wires don't fray or turn to dust after carrying current for a long time. Electrons are very small (perhaps even pointlike), and it is easy to imagine them passing between the cracks among the atoms without creating holes or fractures in the atomic framework. For those who know a little chemistry, it also explains why all the best conductors are on the left side of the periodic table. The elements in that area are the ones that have only a very loose hold on their outermost electrons.

Gases

The molecules in a gas spend most of their time separated from each other by significant distances, so it is not possible for them to conduct electricity the way solids do, by handing off electrons from atom to atom. It is therefore not surprising that gases are good insulators.

Gases are also usually nonohmic. As opposite charges build up on a stormcloud and the ground below, the voltage difference becomes greater and greater. Virtually zero current flows, however, until finally the voltage reaches a certain threshold and we have an impressive example of what is known as a spark or electrical discharge. If air was ohmic, the current between the cloud and the ground would simply increase steadily as the voltage difference increased, rather than being zero until a threshold was reached. This behavior can be explained as follows. At some point, the electrical forces on the air electrons and nuclei of the air molecules become so strong that electrons are ripped right off of some of the molecules. The electrons then accelerate toward either the cloud or the ground, whichever is positively charged, and the positive ions accelerate the opposite way. As these charge carriers accelerate, they strike and ionize other molecules, which produces a rapidly growing cascade.

Liquids

Molecules in a liquid are able to slide past each other, so ions as well as electrons can carry currents. Pure water is a poor conductor because the water molecules tend to hold onto their electrons strongly, and there are therefore not many electrons or ions available to move. Water can become quite a good conductor, however, with the addition of even a small amount of certain substances called electrolytes, which are typically salts. For example, if we add table salt, NaCl, to water, the NaCl molecules dissolve into Na^+ and Cl^- ions, which can then move and create currents. This is why electric currents can flow among the cells in our bodies: cellular fluid is quite salty. When we sweat, we lose not just water but electrolytes, so dehydration plays havoc with our cells' electrical systems. It is for this reason that electrolytes are included in sports drinks and formulas for rehydrating infants who have diarrhea.

Since current flow in liquids involves entire ions, it is not surprising that we can see physical evidence when it has occurred. For example, after a car battery has been in use for a while, the H_2SO_4 battery acid becomes depleted of hydrogen ions, which are the main charge carriers that complete the circuit on the inside of the battery. The leftover SO_4 then forms a visible blue crust on the battery posts.

Speed of currents and electrical signals

When I talk on the phone to my mother in law two thousand miles away, I do not notice any delay while the signal makes its way back and forth. Electrical signals therefore must travel very quickly, but how fast exactly? The answer is rather subtle. For the sake of concreteness, let's restrict ourselves to currents in metals, which consist of electrons.

The electrons themselves are only moving at speeds of perhaps a few thousand miles per hour, and their motion is mostly random thermal motion. This shows that the electrons in my phone cannot possibly be zipping back and forth between California and New York fast enough to carry the signals. Even if their thousand-mile-an-hour motion was organized rather than random, it would still take them many minutes to get there. Realistically, it will take the average electron even longer than that to make the trip. The current in the wire consists only of a slow overall drift, at a speed on the order of a few centimeters per second, superimposed on the more rapid random motion. We can compare this with the slow westward drift in the population of the U.S. If we could make a movie of the motion of all the people in the U.S. from outer space, and could watch it at high speed so that the people appeared to be scurrying around like ants, we would think that the motion was fairly random, and we would not immediately notice the westward drift. Only after many years would we realize that the number of people heading west over the Sierras had exceeded the number going east, so that California increased its share of the country's population.

So why are electrical signals so fast if the average drift speed of electrons is so slow? The answer is that a disturbance in an electrical system can move much more quickly than the charges themselves. It is as though we filled a pipe with golf balls and then inserted an extra ball at one end, causing a ball to fall out at the other end. The force propagated to the other end in a fraction of a second, but the balls themselves only traveled a few centimeters in that time.

Because the reality of current conduction is so complex, we often describe things using mental shortcuts that are technically incorrect. This is OK as long as we know that they are just shortcuts. For example, suppose the presidents of France and Russia shake hands, and the French politician has inadvertently picked up a positive electrical charge, which shocks the Russian. We may say that the excess positively charged particles in the French leader's body, which all repel each other, take the handshake as an opportunity to get farther apart by spreading out into two bodies rather than one. In reality, it would be a matter of minutes before the ions in one person's body could actually drift deep into the other's. What really happens is that throughout the body of the recipient of the shock there are already various positive and negative ions which are free to move. Even before the perpetrator's charged hand touches the victim's sweaty palm, the charges in the shocker's body begin to repel the positive ions and attract the negative ions in the other person. The split-second sensation of shock is caused by the sudden jumping of the victim's ions by distances of perhaps a micrometer, this effect occurring simultaneously throughout the whole body, although more violently in the hand and arm, which are closer to the other person.

3.6∫ Applications of Calculus

The definition of current as the rate of change of charge with respect to time can clearly be reexpressed in terms of a derivative in the case where the rate of change is not constant,

$$I = \frac{\mathrm{d}q}{\mathrm{d}t} \quad .$$

This immediately allows us to use the bag of tricks from calculus.

Example
Question: A charged balloon falls to the ground, and its charge begins leaking off to the Earth. Suppose that the charge on the balloon is given by $q=ae^{-bt}$. Find the current as a function of time, and interpret the answer.
Solution: Taking the derivative, we have

$$\begin{aligned} I &= \frac{\mathrm{d}}{\mathrm{d}t}\left(ae^{-bt}\right) \\ &= -abe^{-bt} \end{aligned}$$

The exponential function approaches zero as the exponent gets more and more negative. This means that both the charge and the current are decreasing in magnitude with time. It makes sense that the charge approaches zero, since the balloon is losing its charge. It also makes sense that the current is decreasing in magnitude, since charge cannot flow at the same rate forever without overshooting zero.

Example
Question: In the segment of the ATLAS accelerator shown in section 3.4, suppose the current flowing back and forth between the two cylinders is given by $I=a \cos bt$. What is the charge on one of the cylinders as a function of time?
Solution: We are given the current and want to find the charge, i.e. we are given the derivative and we want to find the original function that would give that derivative. This means we need to integrate:

$$\begin{aligned} q &= \int I dt \\ &= \int a \cos bt\, dt \\ &= \frac{a}{b} \sin bt + q_o \quad , \end{aligned}$$

where q_o is a constant of integration.

We can interpret this in order to explain why a superconductor needs to be used. The constant b must be very large, since the current is supposed to oscillate back and forth millions of times a second. Looking at the final result, we see that if b is a very large number, and q is to be a significant amount of charge, then a must be a very large number as well. If a is numerically large, then the current must be very large, so it would heat the accelerator too much if it was flowing through an ordinary conductor.

Summary

Selected Vocabulary

current	the rate at which charge crosses a certain boundary
ampere	the metric unit of current, one coulomb pe second; also "amp"
ammeter	a device for measuring electrical current
circuit	an electrical device in which charge can come back to its starting point and be recycled rather than getting stuck in a dead end
open circuit	a circuit that does not function because it has a gap in it
short circuit	a circuit that does not function because charge is given a low-resistance "shortcut" path that it can follow, instead of the path that makes it do something useful
voltage	electrical potential energy per unit charge that will be possessed by a charged particle at a certain point in space
volt	the metric unit of voltage, one joule per coulomb
voltmeter	a device for measuring voltage differences
ohmic	describes a substance in which the flow of current between two points is proportional to the voltage difference between them
resistance	the ratio of the voltage difference to the current in an object made of an ohmic substance
ohm	the metric unit of electrical resistance, one volt per ampere

Notation

I	current
A	units of amperes
V	voltage
V	units of volts
R	resistance
Ω	units of ohms

Notation and Terminology Used in Other Books

electric potential	rather than the more informal "voltage" used here; despite the misleading name, it is not the same as electric potential energy
eV	a unit of energy, equal to e multiplied by 1 volt; $1.6\text{x}10^{-19}$ joules

Summary

All electrical phenomena are alike in that that arise from the presence or motion of charge. Most practical electrical devices are based on the motion of charge around a complete circuit, so that the charge can be recycled and does not hit any dead ends. The most useful measure of the flow of charge is current, $I = \Delta q / \Delta t$.

An electrical device whose job is to transform energy from one form into another, e.g. a lightbulb, uses power at a rate which depends both on how rapidly charge is flowing through it and on how much work is done on each unit of charge. The latter quantity is known as the voltage difference between the point where the current enters the device and the point where the current leaves it. Since there is a type of potential energy associated with electrical forces, the amount of work they do is equal to the difference in potential energy between the two points, and we therefore define voltage differences directly in terms of potential energy, $\Delta V = \Delta PE_{elec} / q$. The rate of power dissipation is $P = I\Delta V$.

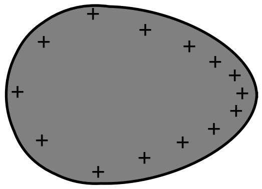
Many important electrical phenomena can only be explained if we understand the mechanisms of current flow at the atomic level. In metals, currents are carried by electrons, in liquids by ions. Gases are normally poor conductors unless their atoms are subjected to such intense electrical forces that the atoms become ionized.

Many substances, including all solids, respond to electrical forces in such a way that the flow of current between two points is proportional to the voltage difference between those points. Such a substance is called ohmic, and an object made out of an ohmic substance can be rated in terms of its resistance, $R = \Delta V / I$. An important corollary is that a perfect conductor, with $R=0$, must have constant voltage everywhere within it.

Homework Problems

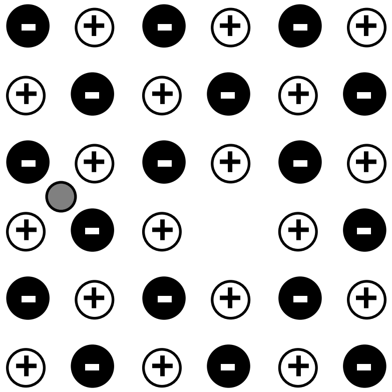
1. A resistor has a voltage difference ΔV across it, causing a current I to flow. (a) Find an equation for the power it dissipates as heat in terms of the variables I and R only, eliminating ΔV. (b) If an electrical line coming to your house is to carry a given amount of current, interpret your equation from part a to explain whether the wire's resistance should be small or large.

2. (a) Express the power dissipated by a resistor in terms of R and ΔV only, eliminating I. (b ✓) Electrical receptacles in your home are mostly 110 V, but circuits for electric stoves, air conditioners, and washers and driers are usually 220 V. The two types of circuits have differently shaped receptacles. Suppose you rewire the plug of a drier so that it can be plugged in to a 110 V receptacle. The resistor that forms the heating element of the drier would normally draw 200 W. How much power does it actually draw now?



Problem 3.

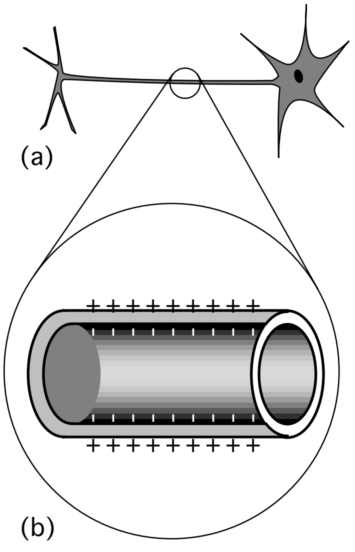
3. As discussed in the text, when a conductor reaches an equilibrium where its charge is at rest, there is always zero electric force on a charge in its interior, and any excess charge concentrates itself on the surface. The surface layer of charge arranges itself so as to produce zero this zero force. (Otherwise the free charge in the interior could not be at rest.) Suppose you have a teardrop-shaped conductor like the one shown in the figure. Explain why the surface layer of charge must be denser in the pointed part of the teardrop. (Similar reasoning shows why lightning rods are made with points. The charged stormclouds induce positive and negative charges to move to opposite ends of the rod. At the pointed upper end of the rod, the charge tends to concentrate at the point, and this charge attracts the lightning.)



Problem 5.

4. Use the result of problem 3 in ch. 1 to find an equation for the potential at a point in space at a distance r from a point charge Q. (Take your V=0 distance to be anywhere you like.)

5★✓. Referring back to the homework problem in chapter 1 about the sodium chloride crystal, suppose the lithium ion is going to jump from the gap it is occupying to one of the four closest neighboring gaps. Which one will it jump to, and if it starts from rest, how fast will it be going by the time it gets there? (It will keep on moving and accelerating after that, but that does not concern us.) [Hint: The approach is similar to the one used for the previous problem, but you want to work with addition and comparison of voltage rather than force.]



Problem 6.

6 ✓. Referring back to our old friend the neuron from the homework problem in chapter 1, let's now consider what happens when the nerve is stimulated to transmit information. When the blob at the top (the cell body) is stimulated, it causes Na$^+$ ions to rush into the top of the tail (axon). This electrical pulse will then travel down the axon, like a flame burning down from the end of a fuse, with the Na$^+$ ions at each point first going out and then coming back in. If 10^{10} Na$^+$ ions cross the cell membrane in 0.5 ms, what amount of current is created?

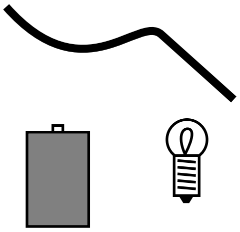
S A solution is given in the back of the book. ★ A difficult problem.
✓ A computerized answer check is available. ∫ A problem that requires calculus.

7 ✓. If a typical light bulb draws about 900 mA from a 110-V household circuit, what is its resistance? (Don't worry about the fact that it's alternating current.)

8✓ . Today, even a big luxury car like a Cadillac can have an electrical system that is relatively low in power, since it doesn't need to do much more than run headlights, power windows, etc. In the near future, however, manufacturers plan to start making cars with electrical systems about five times more powerful. This will allow certain energy-wasting parts like the water pump to be run on electrical motors and turned off when they're not needed — currently they're run directly on shafts from the motor, so they can't be shut off. It may even be possible to make an engine that can shut off at a stoplight and then turn back on again without cranking, since the valves can be electrically powered. Current cars' electrical systems have 12-volt batteries (with 14-volt chargers), but the new systems will have 36-volt batteries (with 42-volt chargers). (a) Based on the sample figures above, how would the currents handled by the wires in one of the new cars compare with the currents in the old ones? (b) Can you guess why they decided to change to 36-volt batteries rather than increasing the power without increasing the voltage?

9★. (a✓) You take an LP record out of its sleeve, and it acquires a static charge of 1 nC. You play it at the normal speed of $33^1/_3$ r.p.m., and the charge moving in a circle creates an electric current. What is the current, in amperes?

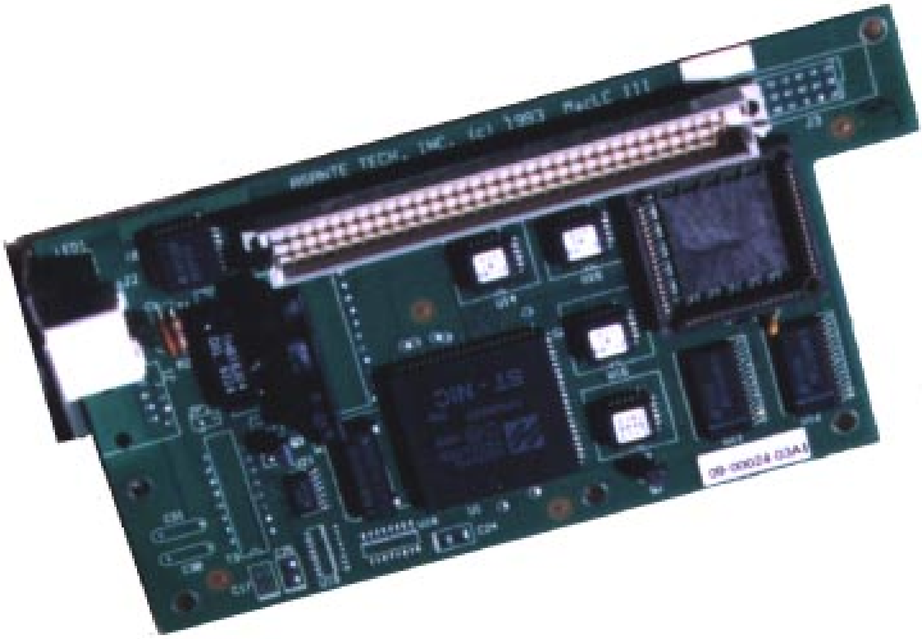
(b✓) Although the planetary model of the atom can be made to work with any value for the radius of the electrons' orbits, more advanced models that we will study later in this course predict definite radii. If the electron is imagined as circling around the proton at a speed of $2.2\text{x}10^6$ m/s, in an orbit with a radius of 0.05 nm, what electric current is created?



Problem 10.

10. We have referred to resistors *dissipating* heat, i.e. we have assumed that $P=I\Delta V$ is always greater than zero. Could $I\Delta V$ come out to be negative for a resistor? If so, could one make a refrigerator by hooking up a resistor in such a way that it absorbed heat instead of dissipating it?

11. You are given a battery, a flashlight bulb, and a single piece of wire. Draw at least two configurations of these items that would result in lighting up the bulb, and at least two that would not light it. (Don't draw schematics.) If you're not sure what's going on, borrow the materials from your instructor and try it.



4 Circuits, Part 2

In the previous chapter we limited ourselves to relatively simple circuits, essentially nothing more than a battery and a single lightbulb. The purpose of this chapter is to introduce you to more complex circuits, containing multiple resistors or voltage sources in series, in parallel, or both.

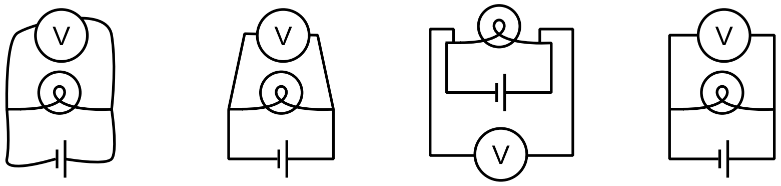
Why do you need to know this stuff? After all, if you were planning on being an electrical engineer you probably wouldn't be learning physics from this book. Consider, however, that every time you plug in a lamp or a radio you are adding a circuit element to a household circuit and making it more complex. Electrical safety, as well, cannot really be understood without understanding multiple-element circuits, since getting shocked usually involves at least two parts: the device that is supposed to use power plus the body of the person in danger. If you are a student majoring in the life sciences, you should realize as well that all cells are inherently electrical, and in any multicellular organism there will therefore be various series and parallel circuits.

Even apart from these practical purposes, there is a very fundamental reason for reading this chapter: to understand the previous chapter better. At this point in their studies, I always observe students using words and modes of reasoning that show they have not yet become completely comfortable and fluent with the concepts of voltage and current. They ask, "aren't voltage and current sort of the same idea?" They speak of voltage "going through" a lightbulb. Once they begin honing their skills on more complicated circuits I always see their confidence and understanding increase immeasurably.

4.1 Schematics

I see a chess position; Kasparov sees an interesting Ruy Lopez variation. To the uninitiated a schematic may look as unintelligible as Mayan hieroglyphs, but even a little bit of eye training can go a long way toward making its meaning leap off the page. A schematic is a stylized and simplified drawing of a circuit. The purpose is to eliminate as many irrelevant features as possible, so that the relevant ones are easier to pick out.

An example of an irrelevant feature is the physical shape, length, and diameter of a wire. In nearly all circuits, it is a good approximation to assume that the wires are perfect conductors, so that any piece of wire uninterrupted by other components has constant voltage throughout it. Changing the length of the wire, for instance, does not change this fact. (Of course if we used miles and miles of wire, as in a telephone line, the wire's resistance would start to add up, and its length would start to matter.) The shapes of the wires are likewise irrelevant, so we draw them with standardized, stylized shapes made only of vertical and horizontal lines with right-angle bends in them. This has the effect of making similar circuits look more alike and helping us to recognize familiar patterns, just as words in a newspaper are easier to recognize than handwritten ones. Figures a-d show some examples of these concepts.

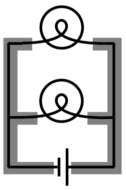


(a) wrong: Shapes of wires are irrelevant.

(b) wrong: Should use right angles.

(c) wrong: Makes a simple pattern look unfamliar.

(d) right

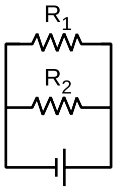


(e)

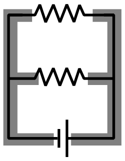
The most important first step in learning to read schematics is to learn to recognize contiguous pieces of wire which must have constant voltage throughout. In figure (e), for example, the two shaded E-shaped pieces of wire must each have constant voltage. This focuses our attention on two of the main unknowns we'd like to be able to predict: the voltage of the left-hand E and the voltage of the one on the right.

4.2 Parallel Resistances and the Junction Rule

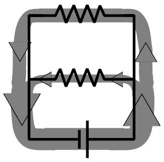
(f)



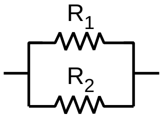
(g)



(h)



(i)



One of the simplest examples to analyze is the parallel resistance circuit, of which figure (e) was an example. In general we may have unequal resistances R_1 and R_2, as in (f). Since there are only two constant-voltage areas in the circuit, (g), all three components have the same voltage difference across them. A battery normally succeeds in maintaining the voltage differences across itself for which it was designed, so the voltage drops ΔV_1 and ΔV_2 across the resistors must both equal the voltage of the battery:

$$\Delta V_1 = \Delta V_2 = \Delta V_{\text{battery}} .$$

Each resistance thus feels the same voltage difference as if it was the only one in the circuit, and Ohm's law tells us that the amount of current flowing through each one is also the same as it would have been in a one-resistor circuit. This is why household electrical circuits are wired in parallel. We want every appliance to work the same, regardless of whether other appliances are plugged in or unplugged, turned on or switched off. (The electric company doesn't use batteries of course, but our analysis would be the same for any device that maintains a constant voltage.)

Of course the electric company can tell when we turn on every light in the house. How do they know? The answer is that we draw more current. Each resistance draws a certain amount of current, and the amount that has to be supplied is the sum of the two individual currents. The current is like a river that splits in half, (h), and then reunites. The total current is

$$I_{\text{total}} = I_1 + I_2 \quad .$$

This is an example of a general fact called the junction rule:

> **the junction rule**
> In any circuit that is not storing or releasing charge, conservation of charge implies that the total current flowing out of any junction must be the same as the total flowing in.

Coming back to the analysis of our circuit, we apply Ohm's law to each resistance, resulting in

$$\begin{aligned} I_{\text{total}} &= \Delta V/R_1 + \Delta V/R_2 \\ &= \Delta V\left(\frac{1}{R_1} + \frac{1}{R_2}\right) \quad . \end{aligned}$$

As far as the electric company is concerned, your whole house is just one resistor with some resistance R, called the *equivalent resistance*. They would write Ohm's law as

$$I_{\text{total}} = \Delta V/R \quad ,$$

from which we can determine the equivalent resistance by comparison with the previous expression:

$$\frac{1}{R} = \frac{1}{R_1} + \frac{1}{R_2}$$

$$R = \left(\frac{1}{R_1} + \frac{1}{R_2}\right)^{-1}$$

[equivalent resistance of two resistors in parallel]

Two resistors in parallel, (i), are equivalent to a single resistor with a value given by the above equation.

Example: two lamps on the same household circuit

Question: You turn on two lamps that are on the same household circuit. Each one has a resistance of 1 ohm. What is the equivalent resistance, and how does the power dissipation compare with the case of a single lamp?

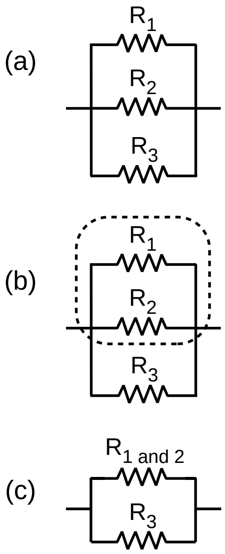
Solution: The equivalent resistance of the two lamps in parallel is

$$\begin{aligned} R &= \left(\frac{1}{R_1} + \frac{1}{R_2}\right)^{-1} \\ &= \left(\frac{1}{1\ \Omega} + \frac{1}{1\ \Omega}\right)^{-1} \\ &= (1\ \Omega^{-1} + 1\ \Omega^{-1})^{-1} \\ &= (2\ \Omega^{-1})^{-1} \\ &= 0.5\ \Omega \end{aligned}$$

The voltage difference across the whole circuit is always the 110 V set by the electric company (it's alternating current, but that's irrelevant). The resistance of the whole circuit has been cut in half by turning on the second lamp, so a fixed amount of voltage will produce twice as much current. Twice the current flowing across the same voltage difference means twice as much power dissipation, which makes sense.

The cutting in half of the resistance surprises many students, since we are "adding more resistance" to the circuit by putting in the second lamp. Why does the equivalent resistance come out to be less than the resistance of a single lamp? This is a case where purely verbal reasoning can be misleading. A resistive circuit element, such as the filament of a lightbulb, is neither a perfect insulator nor a perfect conductor. Instead of analyzing this type of circuit in terms of "resistors," i.e. partial insulators, we could have spoken of "conductors." This example would then seem reasonable, since we "added more conductance," but one would then have the incorrect expectation about the case of resistors in series, discussed in the following section.

Perhaps a more productive way of thinking about it is to use mechanical intuition. By analogy, your nostrils resist the flow of air through them, but having two nostrils makes it twice as easy to breathe.



Example: three resistors in parallel

Question: What happens if we have three or more resistors in parallel?

Solution: This is an important example, because the solution involves an important technique for understanding circuits: breaking them down into smaller parts and them simplifying those parts. In the circuit (a), with three resistors in parallel, we can think of two of the resistors as forming a single big resistor, (b), with equivalent resistance

$$R_{1\text{ and }2} = \left(\frac{1}{R_1}+\frac{1}{R_2}\right)^{-1} \quad .$$

We can then simplify the circuit as shown in (c), so that it contains only two resistances. The equivalent resistance of the whole circuit is then given by

$$R_{1,\,2\text{ and }3} = \left(\frac{1}{R_{1\text{ and }2}}+\frac{1}{R_3}\right)^{-1} \quad .$$

Substituting for $R_{1\text{ and }2}$ and simplifying, we find the result

$$R_{1,\,2\text{ and }3} = \left(\frac{1}{R_1}+\frac{1}{R_2}+\frac{1}{R_3}\right)^{-1} \quad ,$$

which you probably could have guessed. The interesting point here is the divide-and-conquer concept, not the mathematical result.

Example: an arbitrary number of identical resistors in parallel

Question: What is the resistance of N identical resistors in parallel?

Solution: Generalizing the results for two and three resistors, we have

$$R_N = \left(\frac{1}{R_1}+\frac{1}{R_2}+\ldots\right)^{-1} \quad ,$$

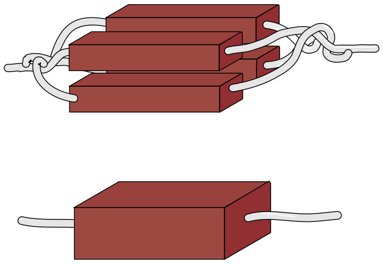
where "..." means that the sum includes all the resistors. If all the resistors are identical, this becomes

$$R_N = \left(\frac{N}{R}\right)^{-1} = \frac{R}{N}$$

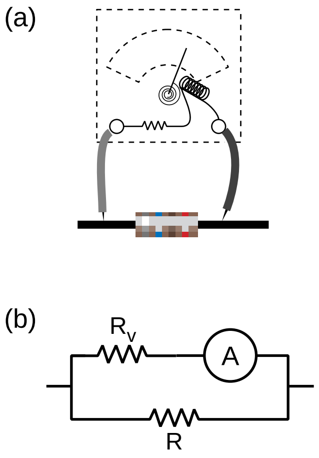
Example: dependence of resistance on cross-sectional area

We have alluded briefly to the fact that an object's electrical resistance depends on its size and shape, but now we are ready to begin making more mathematical statements about it. As suggested by the figure, increasing a resistors's cross-sectional area is equivalent to adding more resistors in parallel, which will lead to an overall decrease in resistance. Any real resistor with straight, parallel sides can be sliced up into a large number of pieces, each with cross-sectional area of, say, 1 μm^2. The number, N, of such slices is proportional to the total cross-sectional area of the resistor, and by application of the result of the previous example we therefore find that the resistance of an object is inversely proportional to its cross-sectional area.

An analogous relationship holds for water pipes, which is why high-flow trunk lines have to have large cross-sectional areas. To make lots of water (current) flow through a skinny pipe, we'd need an impractically large pressure (voltage) difference.



Uniting four resistors in parallel is equivalent to making a single resistor with the same length but four times the cross-sectional area. The result is to make a resistor with one quarter the resistance.



A voltmeter is really an ammeter with an internal resistor. When we measure the voltage difference across a resistor, (a), we are really constructing a parallel resistance circuit, (b).

Example: incorrect readings from a voltmeter

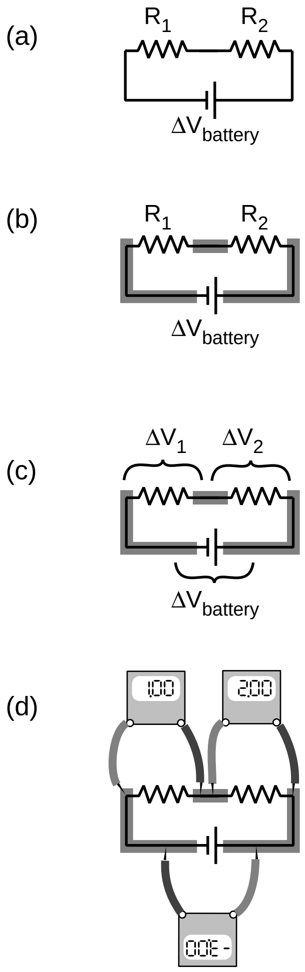
A voltmeter is really just an ammeter with an internal resistor, and we use a voltmeter in parallel with the thing that we're trying to measure the voltage difference across. This means that any time we measure the voltage drop across a resistor, we're essentially putting two resistors in series. The ammeter inside the voltmeter can be ignored for the purpose of analyzing what how current flows in the circuit, since it is essentially just some coiled-up wire with a very low resistance.

Now if we are carrying out this measurement on a resistor that is part of a larger circuit, we have changed the behavior of the circuit through our act of measuring. It is as though we had modified the circuit by replacing the resistance R with the smaller equivalent resistance of R and R_v in parallel. It is for this reason that voltmeters are built with the largest possible internal resistance. As a numerical example, if we use a voltmeter with an internal resistance of 1 MΩ to measure the voltage drop across a one-ohm resistor, the equivalent resistance is 0.999999 Ω, which is not different enough to make any difference. But if we tried to use the same voltmeter to measure the voltage drop across a 2-MΩ resistor, we would be reducing the resistance of that part of the circuit by a factor of three, which would produce a drastic change in the behavior of the whole circuit.

This is the reason why you can't use a voltmeter to measure the voltage difference between two different points in mid-air, or between the ends of a piece of wood. This is by no means a stupid thing to want to do, since the world around us is not a constant-voltage environment, the most extreme example being when an electrical storm is brewing. But it will not work with an ordinary voltmeter because the resistance of the air or the wood is many gigaohms. The effect of waving a pair of voltmeter probes around in the air is that we provide a reuniting path for the positive and negative charges that have been separated — through the voltmeter itself, which is a good conductor compared to the air. This reduces to zero the voltage difference we were trying to measure.

In general, a voltmeter that has been set up with an open circuit (or a very large resistance) between its probes is said to be "floating." An old-fashioned analog voltmeter of the type described here will read zero when left floating, the same as when it was sitting on the shelf. A floating digital voltmeter usually shows an error message.

4.3 Series Resistances



The two basic circuit layouts are parallel and series, so a pair of resistors is series, (a), is another of the most basic circuits we can make. By conservation of charge, all the current that flows through one resistor must also flow through the other (as well as through the battery):

$$I_1 = I_2 \quad .$$

The only way the information about the two resistance values is going to be useful is if we can apply Ohm's law, which will relate the resistance of each resistor to the current flowing through it and the voltage difference across it. Figure (b) shows the three constant-voltage areas. Voltage differences are more physically significant than voltages, so we define symbols for the voltage differences across the two resistors in figure (c).

We have three constant-voltage areas, with symbols for the difference in voltage between every possible pair of them. These three voltage differences must be related to each other. It is as though I tell you that Fred is a foot taller than Ginger, Ginger is a foot taller than Sally, and Fred is two feet taller than Sally. The information is redundant, and you really only needed two of the three pieces of data to infer the third. In the case of our voltage differences, we have

$$\left|\Delta V_1\right| + \left|\Delta V_2\right| = \left|\Delta V_{\text{battery}}\right| \quad .$$

The absolute value signs are because of the ambiguity in how we define our voltage differences. If we reversed the two probes of the voltmeter, we would get a result with the opposite sign. Digital voltmeters will actually provide a minus sign on the screen if the wire connected to the "V" plug is lower in voltage than the one connected to the "COM" plug. Analog voltmeters pin the needle against a peg if you try to use them to measure negative voltages, so you have to fiddle to get the leads connected the right way, and then supply any necessary minus sign yourself.

Figure (d) shows a standard way of taking care of the ambiguity in signs. For each of the three voltage measurements around the loop, we keep the same probe (the darker one) on the clockwise side. It is as though the voltmeter was sidling around the circuit like a crab, without ever "crossing its legs." With this convention, the relationship among the voltage drops becomes

$$\Delta V_1 + \Delta V_2 = -\Delta V_{\text{battery}} \quad ,$$

or, in more symmetrical form,

$$\Delta V_1 + \Delta V_2 + \Delta V_{\text{battery}} = 0$$

More generally, this is known as the loop rule for analyzing circuits:

the loop rule
Assuming the standard convention for plus and minus signs, the sum of the voltage drops around any closed loop in a circuit must be zero.

Looking for an exception to the loop rule would be like asking for a hike that would be downhill all the way and that would come back to its starting point!

For the circuit we set out to analyze, the equation

$$\Delta V_1 + \Delta V_2 + \Delta V_{\text{battery}} = 0$$

can now be rewritten by applying Ohm's law to each resistor:

$$I_1 R_1 + I_2 R_2 + \Delta V_{\text{battery}} = 0 \quad .$$

The currents are the same, so we can factor them out:

$$I\left(R_1 + R_2\right) + \Delta V_{\text{battery}} = 0 \quad ,$$

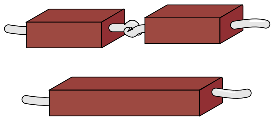
and this is the same result we would have gotten if we had been analyzing a one-resistor circuit with resistance R_1+R_2. Thus the equivalent resistance of resistors in series equals the sum of their resistances.

Example: two lightbulbs in series
Question: If two identical lightbulbs are placed in series, how do their brightnesses compare with the brightness of a single bulb?
Solution: Taken as a whole, the pair of bulbs act like a doubled resistance, so they will draw half as much current from the wall. Each bulb will be dimmer than a single bulb would have been.

The total power dissipated by the circuit is $I\Delta V$. The voltage drop across the whole circuit is the same as before, but the current is halved, so the two-bulb circuit draws half as much total power as the one-bulb circuit. Each bulb draws one-quarter of the normal power.

Roughly speaking, we might expect this to result in one quarter the light being produced by each bulb, but in reality lightbulbs waste quite a high percentage of their power in the form of heat and wavelengths of light that are not visible (infrared and ultraviolet). Less light will be produced, but it's hard to predict exactly how much less, since the efficiency of the bulbs will be changed by operating them under different conditions.

Example: more than two equal resistances in series
By straightforward application of the divide-and-conquer technique discussed in the previous section, we find that the equivalent resistance of N identical resistances R in series will beNR.



Doubling the length of a resistor is like putting two resistors in series. The resistance is doubled.

Example: dependence of resistance on length
In the previous section, we proved that resistance is inversely proportional to cross-sectional area. By equivalent reason about resistances in series, we find that resistance is proportional to length. Analogously, it is harder to blow through a long straw than through a short one.

Putting the two arguments together, we find that the resistance of an object with straight, parallel sides is given by

$$R = (\text{constant}) \cdot L/A$$

The proportionality constant is called the resistivity, and it depends only on the substance of which the object is made. A resistivity measurement could be used, for instance, to help identify a sample of an unknown substance.

Example: choice of high voltage for power lines

Thomas Edison got involved in a famous technological controversy over the voltage difference that should be used for electrical power lines. At this time, the public was unfamiliar with electricity, and easily scared by it. The president of the United States, for instance, refused to have electrical lighting in the White House when it first became commercially available because he considered it unsafe, preferring the known fire hazard of oil lamps to the mysterious dangers of electricity. Mainly as a way to overcome public fear, Edison believed that power should be transmitted using small voltages, and he publicized his opinion by giving demonstrations at which a dog was lured into position to be killed by a large voltage difference between two sheets of metal on the ground. (Edison's opponents also advocated alternating current rather than direct current, and AC is more dangerous than DC as well. As we will discuss later, AC can be easily stepped up and down to the desired voltage level using a device called a transformer.)

Now if we want to deliver a certain amount of power P_L to a load such as an electric lightbulb, we are constrained only by the equation $P_L = I\Delta V_L$. We can deliver any amount of power we wish, even with a low voltage, if we are willing to use large currents. Modern electrical distribution networks, however, use dangerously high voltage differences of tens of thousands of volts. Why did Edison lose the debate?

It boils down to money. The electric company must deliver the amount of power P_L desired by the customer through a transmission line whose resistance R_T is fixed by economics and geography. The same current flows through both the load and the transmission line, dissipating power usefully in the former and wastefully in the latter. The efficiency of the system is

$$\begin{aligned}\text{efficiency} &= \frac{\text{power paid for by the customer}}{\text{power paid for by the utility}}\\ &= \frac{P_L}{P_L + P_T}\\ &= \frac{1}{1 + P_T/P_L}\end{aligned}$$

Putting ourselves in the shoes of the electric company, we wish to get rid of the variable P_T, since it is something we control only indirectly by our choice of ΔV_T and I. Substituting $P_T = I\Delta V_T$, we find

$$\text{efficiency} = \frac{1}{1 + \frac{I\Delta V_T}{P_L}}$$

We assume the transmission line (but not necessarily the load) is ohmic, so substituting $\Delta V_T = IR_T$ gives

$$\text{efficiency} = \frac{1}{1 + \frac{I^2 R_T}{P_L}}$$

This quantity can clearly be maximized by making I as small as possible, since we will then be dividing by the smallest possible quantity on the bottom of the fraction. A low-current circuit can only deliver significant amounts of power if it uses high voltages, which is why electrical transmission systems use dangerous high voltages.

Example: getting killed by your ammeter

As with a voltmeter, an ammeter can give erroneous readings if it is used in such a way that it changes the behavior the circuit. An ammeter is used in series, so if it is used to measure the current through a resistor, the resistor's value will effectively be changed to $R+R_a$, where R_a is the resistance of the ammeter. Ammeters are designed with very low resistances in order to make it unlikely that $R+R_a$ will be significantly different from R.

In fact, the real hazard is death, not a wrong reading! Virtually the only circuits whose resistances are significantly less than that of an ammeter are those designed to carry huge currents. An ammeter inserted in such a circuit can easily melt. When I was working at a laboratory funded by the Department of Energy, we got periodic bulletins from the DOE safety office about serious accidents at other sites, and they held a certain ghoulish fascination. One of these was about a DOE worker who was completely incinerated by the explosion created when he inserted an ordinary Radio Shack ammeter into a high-current circuit. Later estimates showed that the heat was probably so intense that the explosion was a ball of plasma — a gas so hot that the atoms have been ionized.

Discussion Question

We have stated the loop rule in a symmetric form where a series of voltage drops adds up to zero. To do this, we had to define a standard way of connecting the voltmeter to the circuit so that the plus and minus signs would come out right. Suppose we wish to restate the junction rule in a similar symmetric way, so that instead of equating the current coming in to the current going out, it simply states that a certain sum of currents at a junction adds up to zero. What standard way of inserting the ammeter would we have to use to make this work?

Summary

Notation and Terminolog Used in Other Books

eV .... a unit of energy, equal to *e* multiplied by 1 volt; $1.6\text{x}10^{-19}$ joules

Summary

A schematic is a drawing of a circuit that standardizes and stylizes its features to make it easier to understand. Any circuit can be broken down into smaller parts. For instance, one big circuit may be understood as two small circuits in series, another as three circuits in parallel. When circuit elements are combined in parallel and in series, we have two basic rules to guide us in understanding how the parts function as a whole:

> **the junction rule**: In any circuit that is not storing or releasing charge, conservation of charge implies that the total current flowing out of any junction must be the same as the total flowing in.
>
> **the loop rule**: Assuming the standard convention for plus and minus signs, the sum of the voltage drops around any closed loop in a circuit must be zero.

The simplest application of these rules is to pairs of resistors combined in series or parallel. In such cases, the pair of resistors acts just like a single unit with a certain resistance value, called their equivalent resistance. Resistances in series add to produce a larger equivalent resistance,

$$R_{series} = R_1 + R_2 \quad ,$$

because the current has to fight its way through both resistances. Parallel resistors combine to produce an equivalent resistance that is smaller than either individual resistance,

$$R_{parallel} = \left(\frac{1}{R_1} + \frac{1}{R_2}\right)^{-1} \quad ,$$

because the current has two different paths open to it.

An important example of resistances in parallel and series is the use of voltmeters and ammeters in resistive circuits. A voltmeter acts as a large resistance in parallel with the resistor across which the voltage drop is being measured. The fact that its resistance is not infinite means that it alters the circuit it is being used to investigate, producing a lower equivalent resistance. An ammeter acts as a small resistance in series with the circuit through which the current is to be determined. Its resistance is not quite zero, which leads to an increase in the resistance of the circuit being tested.

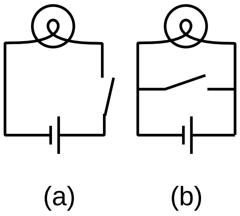
Homework Problems

1. (a) Many battery-operated devices take more than one battery. If you look closely in the battery compartment, you will see that the batteries are wired in series. Consider a flashlight circuit. What does the loop rule tell you about the effect of putting several batteries in series in this way? (b) The cells of an electric eel's nervous system are not that different from ours — each cell can develop a voltage difference across it of somewhere on the order of one volt. How, then, do you think an electric eel can create voltages of thousands of volts between different parts of its body?

2. The heating element of an electric stove is connected in series with a switch that opens and closes many times per second. When you turn the knob up for more power, the fraction of the time that the switch is closed increases. Suppose someone suggests a simpler alternative for controlling the power by putting the heating element in series with a variable resistor controlled by the knob. (With the knob turned all the way, the variable resistor's resistance is nearly zero, and when it's turned all the way the other way, its resistance is essentially infinite.) Why would the simpler design be undesirable?

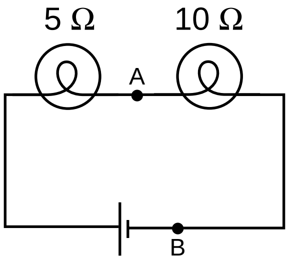
3. A one-ohm toaster and a two-ohm lamp are connected in parallel with the 110-V supply of your house. (Ignore the fact that the voltage is AC rather than DC.) (a) Draw a schematic of the circuit. (b✓). For each of the three components in the circuit, find the current passing through it and the voltage drop across it. (c✓) Suppose they were instead hooked up in series. Draw a schematic and calculate the same things.

4. Wire is sold in a series of standard diameters, called "gauges." The difference in diameter between one gauge and the next in the series is about 20%. How would the resistance of a given length of wire of a certain gauge compare with the resistance of the same length of wire in the next gauge in the series?



Problem 5.

5. The figure shows two possible ways of wiring a flashlight with a switch. Both will serve to turn the bulb on and off, although the switch functions in the opposite sense. Why is the method shown in (a) preferable?



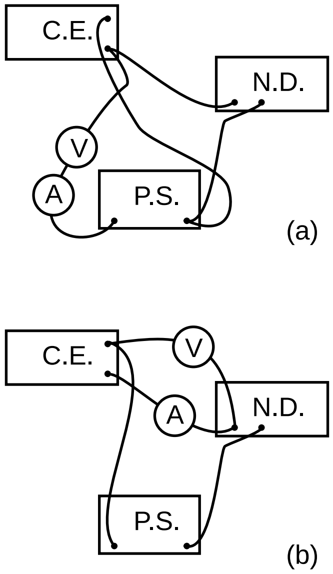
Problem 6.

6. In the figure, the battery is 9 V. (a) What are the voltage differences across each light bulb? (b) What current flows through each of the three components of the circuit? (c) If a new wire is added to connect points A and B, how will the appearances of the bulbs change? What will be the new voltages and currents? (d) Suppose no wire is connected from A to B, but the two bulbs are switched. How will the results compare with the results from the original setup as drawn?

S A solution is given in the back of the book. ★ A difficult problem.
✓ A computerized answer check is available. ∫ A problem that requires calculus.

7. You have a circuit consisting of two unknown resistors in series, and a second circuit consisting of two unknown resistors in parallel. (a) What, if anything, would you learn about the resistors in the series circuit by finding that the currents through them were equal? (b) What if you found out the voltage differences across the resistors in the series circuit were equal? (c) What would you learn about the resistors in the parallel circuit from knowing that the currents were equal? (d) What if the voltages in the parallel circuit were equal?

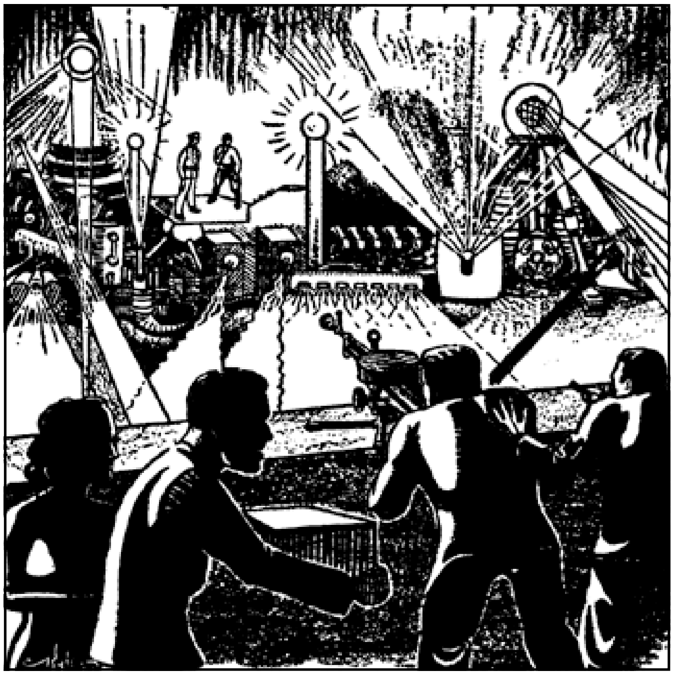
8. A student in a biology lab is given the following instructions: “Connect the cerebral eraser (C.E.) and the neural depolarizer (N.D.) in parallel with the power supply (P.S.). (Under no circumstances should you ever allow the cerebral eraser to come within 20 cm of your head.) Connect a voltmeter to measure the voltage across the cerebral eraser, and also insert an ammeter in the circuit so that you can make sure you don’t put more than 100 mA through the neural depolarizer.” The diagrams show two lab groups’ attempts to follow the instructions. (a) Translate diagram (a) into a standard-style schematic. What is incorrect and incorrect about this group’s setup? (b) Do the same for diagram (b).



Problem 8.

9. How many different resistance values can be created by combining three unequal resistors?

10 ∫. A person in a rural area who has no electricity runs an extremely long extension cord to a friend’s house down the road so she can run an electric light. The cord is so long that its resistance, *x*, is not negligible. Show that the lamp’s brightness is greatest if its resistance, *y*, is equal to *x*. Explain physically why the lamp is dim for values of *y* that are too small or too large.



"Okay. Your duties are as follows: Get Breen. I don't care how you get him, but get him soon. That faker! He posed for twenty years as a scientist without ever being apprehended. Well, I'm going to do some apprehending that'll make all previous apprehending look like no apprehension at all. You with me?"

"Yes," said Battle, very much confused. "What's that thing you have?"

"Piggy-back heat-ray. You transpose the air in its path into an unstable isotope which tends to carry all energy as heat. Then you shoot your juice light, or whatever along the isotopic path and you burn whatever's on the receiving end. You want a few?"

"No," said Battle. "I have my gats. What else have you got for offense and defense?"

Underbottam opened a cabinet and proudly waved an arm. "Everything," he said. "Disintegraters, heat-rays, bombs of every type. And impenetrable shields of energy, massive and portable. What more do I need?"

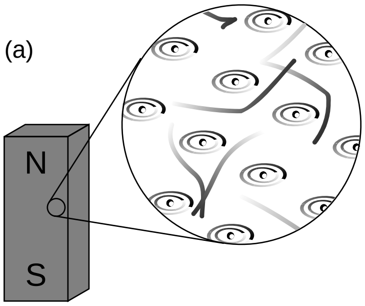
From THE REVERSIBLE REVOLUTIONS by Cecil Corwin, Cosmic Stories, March 1941. Art by Morey, Bok, Kyle, Hunt, Forte. Copyright expired.

5 Fields of Force

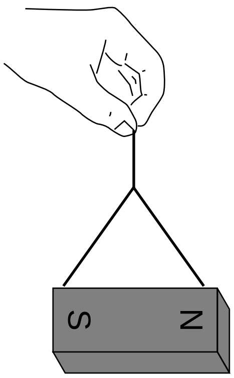
Cutting-edge science readily infiltrates popular culture, though sometimes in garbled form. The Newtonian imagination populated the universe mostly with that nice solid stuff called matter, which was made of little hard balls called atoms. In the early twentieth century, consumers of pulp fiction and popularized science began to hear of a new image of the universe, full of x-rays, N-rays, and Hertzian waves. What they were beginning to soak up through their skins was a drastic revision of Newton's concept of a universe made of chunks of matter which happened to interact via forces. In the newly emerging picture, the universe was *made* of force, or, to be more technically accurate, of ripples in universal fields of force. Unlike the average reader of Cosmic Stories in 1941, you now possess enough technical background to be understand what a "force field" really is.

5.1 Why Fields?

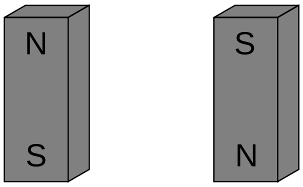
(a)



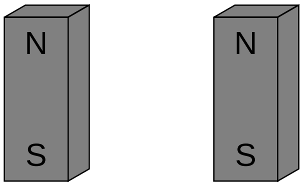
(b)



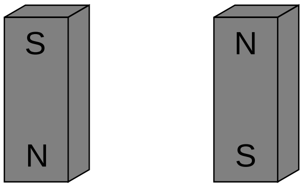
(c)



(d)



(e)



Time delays in forces exerted at a distance

What convinced physicists that they needed this new concept of a field of force? Although we have been dealing mostly with electrical forces, let's start with a magnetic example. (In fact the main reason I've delayed a detailed discussion of magnetism for so long is that mathematical calculations of magnetic effects are handled much more easily with the concept of a field of force.) First a little background leading up to our example. A bar magnet, (a), has an axis about which many of the electrons' orbits are oriented. The earth itself is also a magnet, although not a bar-shaped one. The interaction between the earth-magnet and the bar magnet, (b), makes them want to line up their axes in opposing directions (in other words such that their electrons rotate in parallel planes, but with one set rotating clockwise and the other counterclockwise as seen looking along the axes). On a smaller scale, any two bar magnets placed near each other will try to align themselves head-to-tail, (c).

Now we get to the relevant example. It is clear that two people separated by a paper-thin wall could use a pair of bar magnets to signal to each other. Each person would feel her own magnet trying to twist around in response to any rotation performed by the other person's magnet. The practical range of communication would be very short for this setup, but a sensitive electrical apparatus could pick up magnetic signals from much farther away. In fact, this is not so different from what a radio does: the electrons racing up and down the transmitting antenna create forces on the electrons in the distant receiving antenna. (Both magnetic and electric forces are involved in real radio signals, but we don't need to worry about that yet.)

A question now naturally arises as to whether there is any time delay in this kind of communication via magnetic (and electric) forces. Newton would have thought not, since he conceived of physics in terms of instantaneous action at a distance. We now know, however, that there is such a time delay. If you make a long-distance phone call that is routed through a communications satellite, you should easily be able to detect a delay of about half a second over the signal's round trip of 50,000 miles. Modern measurements have shown that electric, magnetic, and gravitational forces all travel at the speed of light, $3\text{x}10^{8}$ m/s. (In fact, we will soon discuss how light itself is made of electricity and magnetism.)

If it takes some time for forces to be transmitted through space, then apparently there is some *thing* that travels *through* space. The fact that the phenomenon travels outward at the same speed in all directions strongly evokes wave metaphors such as ripples on a pond.

More evidence that fields of force are real: they carry energy.

The smoking-gun argument for this strange notion of traveling force ripples comes from the fact that they carry energy.

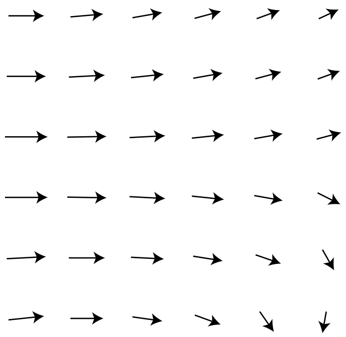
First suppose that the person holding the bar magnet on the right decides to reverse hers, resulting in configuration (d). She had to do mechanical work to twist it, and if she releases the magnet, energy will be released as it flips back to (c). She has apparently stored energy by going from (c) to (d). So far everything is easily explained without the concept of a field of force.

But now imagine that the two people start in position (c) and then simultaneously flip their magnets extremely quickly to position (e), keeping them lined up with each other the whole time. Imagine, for the sake of argument, that they can do this so quickly that each magnet is reversed while the force signal from the other is still in transit. (For a more realistic example, we'd have to have two radio antennas, not two magnets, but the magnets are easier to visualize.) During the flipping, each magnet is still feeling the forces arising from the way the other magnet *used* to be oriented. Even though the two magnets stay aligned during the flip, the time delay causes each person to feel resistance as she twists her magnet around. How can this be? Both of them are apparently doing mechanical work, so they must be storing magnetic energy somehow. But in the traditional Newtonian conception of matter interacting via instantaneous forces at a distance, potential energy arises from the relative positions of objects that are interacting via forces. If the magnets never changed their orientations relative to each other, how can any potential energy have been stored?

The only possible answer is that the energy must have gone into the magnetic force ripples crisscrossing the space between the magnets. Fields of force apparently carry energy across space, which is strong evidence that they are real things.

This is perhaps not as radical an idea to us as it was to our ancestors. We are used to the idea that a radio transmitting antenna consumes a great deal of power, and somehow spews it out into the universe. A person working around such an antenna needs to be careful not to get too close to it, since all that energy can easily cook flesh (a painful phenomenon known as an "RF burn").

5.2 The Gravitational Field



The wind patterns in a certain area of the ocean could be charted in a “sea of arrows” representation like this. Each arrow represents both the wind’s strength and its direction at a certain location.

Given that fields of force are real, how do we define, measure, and calculate them? A fruitful metaphor will be the wind patterns experienced by a sailing ship. Wherever the ship goes, it will feel a certain amount of force from the wind, and that force will be in a certain direction. The weather is ever-changing, of course, but for now let’s just imagine steady wind patterns. Definitions in physics are operational, i.e. they describe how to measure the thing being defined. The ship’s captain can measure the wind’s “field of force” by going to the location of interest and determining both the direction of the wind and the strength with which it is blowing. Charting all these measurements on a map leads to a depiction of the field of wind force like the one shown in the figure. This is known as the “sea of arrows” method of visualizing a field.

Now let’s see how these concepts are applied to the fundamental force fields of the universe. We’ll start with the gravitational field, which is the easiest to understand. As with the wind patterns, we’ll start by imagining gravity as a static field, even though the existence of the tides proves that there are continual changes in the gravity field in our region of space. Defining the direction of the gravitational field is easy enough: we simply go to the location of interest and measure the direction of the gravitational force on an object, such as a weight tied to the end of a string.

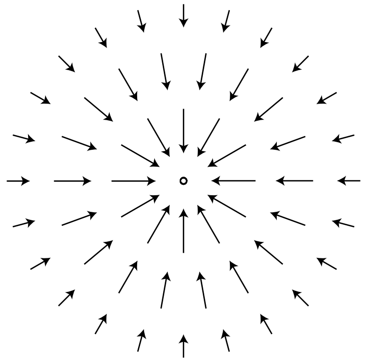
But how should we define the strength of the gravitational field? Gravitational forces are weaker on the moon than on the earth, but we cannot specify the strength of gravity simply by giving a certain number of newtons. The number of newtons of gravitational force depends not just on the strength of the local gravitational field but also on the mass of the object on which we’re testing gravity, our “test mass.” A boulder on the moon feels a stronger gravitational force than a pebble on the earth. We can get around this problem by defining the strength of the gravitational field as the force acting on an object, *divided by the object’s mass.*

definition of the gravitational field

The gravitational field vector, $\boldsymbol{g}$, at any location in space is found by placing a test mass m_t at that point. The field vector is then given by $\boldsymbol{g}=\boldsymbol{F}/m_t$, where $\boldsymbol{F}$ is the gravitational force on the test mass.

The magnitude of the gravitational field near the surface of the earth is about 9.8 N/kg, and it’s no coincidence that this number looks familiar, or that the symbol $\boldsymbol{g}$ is the same as the one for gravitational acceleration. The force of gravity on a test mass will equal $m_t\boldsymbol{g}$, where $\boldsymbol{g}$ is the gravitational acceleration. Dividing by m_t simply gives the gravitational acceleration. Why define a new name and new units for the same old quantity? The main reason is that it prepares us with the right approach for defining other fields.

The most subtle point about all this is that the gravitational field tells us about what forces *would* be exerted on a test mass by the earth, sun, moon, and the rest of the universe, *if* we inserted a test mass at the point in question. The field still exists at all the places where we didn’t measure it, and the field doesn’t take into account forces made by the test mass itself. Newton’s third law guarantees that no object can make a force on itself.



(a) The gravitational field surrounding a clump of mass such as the earth.

Example: gravitational field of the earth

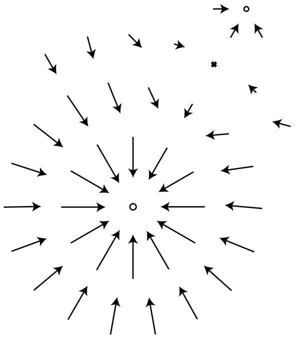
Question: What is the magnitude of the earth's gravitational field, in terms of its mass, *M*, and the distance *r* from its center?

Solution: Substituting $|\boldsymbol{F}|=GMm_t/r^2$ into the definition of the gravitational field, we find $|\boldsymbol{g}|=GM/r^2$.

Sources and sinks

If we make a sea-of-arrows picture of the gravitational fields surrounding the earth, (a), the result is evocative of water going down a drain. For this reason, anything that creates an inward-pointing field around itself is called a sink. The earth is a gravitational sink. The term "source" is used to either refer to things that make outward-pointing field patterns or often as a more general term for both "outies" and "innies." However confusing the terminology, we know that gravitational fields are only attractive, so we will never find a region of space with an outward-pointing field pattern.

Knowledge of the field is interchangeable with knowledge of its sources (at least in the case of a static, unchanging field). If aliens saw the earth's gravitational field pattern they could immediately infer the existence of the planet, and conversely if they knew the mass of the earth they could predict its influence on the surrounding gravitational field.



(b) The gravitational fields of the earth and moon superpose. Note how the fields cancel at one point, and how there is no boundary between the interpenetrating fields surrounding the two bodies.

Superposition of fields

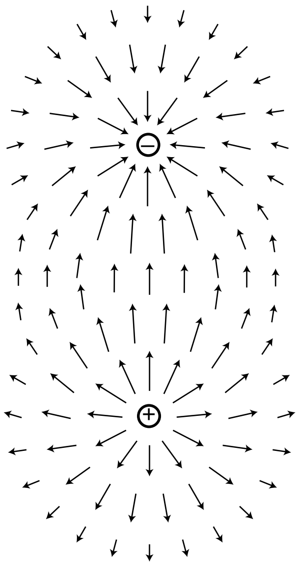
A simple, general, and very important fact about all fields of force is that they represent the vector sum of the fields contributed by all the objects creating them. The gravitational field certainly will have this property, since it is defined in terms of the force on a test mass, and that force equals the vector sum of all the gravitational attractions exerted by all the other masses in the universe. Superposition is an important characteristics of waves, so the superposition property of fields is consistent with the idea that disturbances can propagate outward as waves in a field.

Gravitational waves

A source that sits still will create a static field pattern, like a steel ball sitting peacefully on a sheet of rubber. A moving source will create a spreading wave pattern in the field, like a bug thrashing on the surface of a pond. Although we have started with the gravitational field as the simplest example of a static field, stars and planets do more stately gliding than thrashing, so gravitational waves are not easy to detect. Newton's theory of gravity does not describe gravitational waves, but they are predicted by Einstein's general theory of relativity. J.H. Taylor and R.A. Hulse were awarded the Nobel Prize in 1993 for giving indirect evidence that Einstein's waves actually exist. They discovered a pair of exotic, ultra-dense stars called neutron stars orbiting one another very closely, and showed that they were losing orbital energy at the rate predicted by Einstein's theory. A CalTech-MIT collaboration is currently building a pair of gravitational wave detectors called LIGO to search for more direct evidence of gravitational waves. Since they are essentially the most sensitive vibration detectors ever made, they are being built in quiet rural areas, and signals will be compared between them to make sure that they were not due to passing trucks. The project only has enough funding to keep the detectors operating for a few years after they become operational in 2000, and they can only hope that a sufficiently violent cataclysm will occur somewhere in the universe to make a detectable gravitational wave during that time.

The LIGO detector in Louisiana. More information about LIGO is available on the world wide web at www.ligo.caltech.edu.

5.3 The Electric Field



(a) A dipole field. Electric fields diverge from a positive charge and converge on a negative charge.

Definition

The definition of the electric field is directly analogous to, and has the same motivation as, the definition of the gravitational field:

> **definition of the electric field**
>
> The electric field vector, $\boldsymbol{E}$, at any location in space is found by placing a test charge q_t at that point. The electric field vector is then given by $\boldsymbol{E}=\boldsymbol{F}/q_t$, where $\boldsymbol{F}$ is the electric force on the test charge.

Charges are what create electric fields. Unlike gravity, which is always attractive, electricity displays both attraction and repulsion. A positive charge is a source of electric fields, and a negative one is a sink.

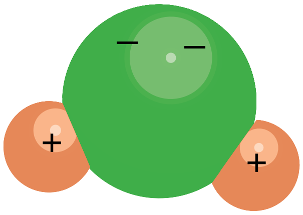
The most difficult point about the definition of the electric field is that the force on a negative charge is in the opposite direction compared to the field. This follows from the definition, since dividing a vector by a negative number reverses its direction. It is as though we had some objects that fell upward instead of down.

Self-Check

Find an equation for the field of a single point charge Q.

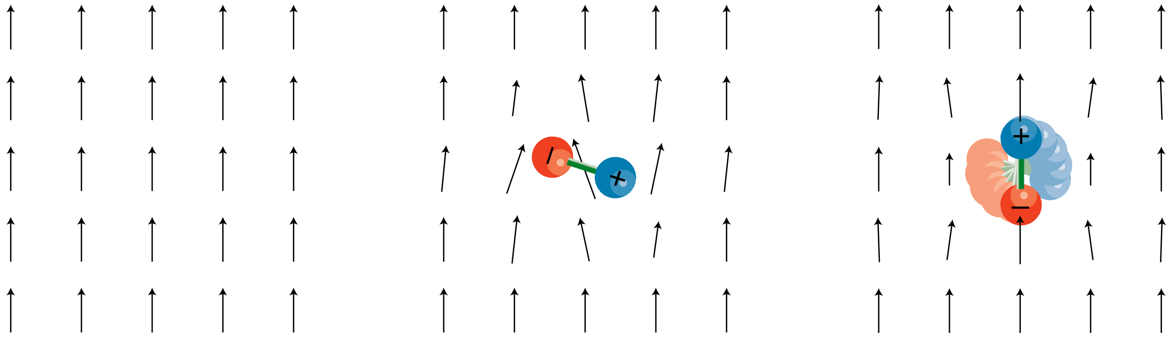
Dipoles

The simplest set of sources that can occur with electricity but not with gravity is the *dipole*, (a), consisting of a positive charge and a negative charge of equal magnitudes. More generally, an electric dipole can be any object with an imbalance of positive charge on one side and negative on the other. A water molecule, (b), is a dipole because the electrons tend to shift away from the hydrogen atoms and onto the oxygen atom.



(b) A water molecule is a dipole.

Your microwave oven works by acting on water molecules with electric fields. Let us imagine what happens if we start with a uniform electric field, (c), made by some external charges, and then insert a dipole, (d), consisting of two charges connected rigidly by a rod. The dipole disturbs the field pattern, but more important for our present purposes is that it experiences a torque. In this example, the positive charge feels an upward force, but the negative charge is pulled down. The result is that the dipole wants to align itself with the field, (e). The microwave oven heats food with electrical (and



(c) (d) (e)

The reasoning is exactly analogous to that used in the previous section to derive an equation for the gravitational field of the Earth. The field is $F/q_t = (kQq_t/r^2)/q_t = kQ/r^2$.

magnetic) waves. The alternation of the torque causes the molecules to wiggle and increase the amount of random motion. The slightly vague definition of a dipole given above can be improved by saying that a dipole is any object that experiences a torque in an electric field.

What determines the torque on a dipole placed in an externally created field? Torque depends on the amount of force, the distance from the axis at which the force is applied, and the angle between the force and the line from the axis to the point of application. Let a dipole consisting of charges $+q$ and $-q$ separated by a distance ℓ be placed in an external field of magnitude $|\mathbf{E}|$, at an angle θ with respect to the field. The total torque on the dipole is

$$\begin{aligned}\tau &= \frac{\ell}{2}q|E|\sin\theta + \frac{\ell}{2}q|E|\sin\theta \\ &= \ell q|E|\sin\theta \quad .\end{aligned}$$

(Note that even though the two forces are in opposite directions, the torques do not cancel, because they are trying to twist the dipole in the same direction.) The quantity ℓq is called the dipole moment, notated D. (More complex dipoles can also be assigned a dipole moment — they are defined as having the same dipole moment as the two-charge dipole that would experience the same torque.)

Example: dipole moment of a molecule of NaCl gas
Question: In a molecule of NaCl gas, the center-to-center distance between the two atoms is about 0.6 nm. Assuming that the chlorine completely steals one of the sodium's electrons, compute this molecule's dipole moment.
Solution: The dipole moment is

$$\begin{aligned}\ell q &= (6\text{x}10^{-10}\ \text{m})(\text{e}) \\ &= (6\text{x}10^{-10}\ \text{m})(1.6\text{x}10^{-19}\ \text{C}) \\ &= 1\text{x}10^{-28}\ \text{C}\cdot\text{m}\end{aligned}$$

Alternative definition of the electric field

The behavior of a dipole in an externally created field leads us to an alternative definition of the electric field:

alternative definition of the electric field
The electric field vector, $\mathbf{E}$, at any location in space is defined by observing the torque exerted on a test dipole D_t placed there. The direction of the field is the direction in which the field tends to align a dipole (from – to +), and the field's magnitude is $|\mathbf{E}| = \tau/D_t \sin\theta$.

The main reason for introducing a second definition for the same concept is that the magnetic field is most easily defined using a similar approach.

Voltage related to electric field

Voltage is potential energy per unit charge, and electric field is force per unit charge. We can therefore relate voltage and field if we start from the relationship between potential energy and force,

$$\Delta PE = -Fd \qquad \text{[assuming constant force and motion parallel to the force],}$$

and divide by charge,

$$\frac{\Delta PE}{q} = -\frac{Fd}{q} \qquad \text{[assuming constant force and motion parallel to the force],}$$

giving

$$\Delta V = -Ed \qquad \text{[assuming constant field and motion parallel to the field].}$$

In other words, the difference in voltage between two points equals the electric field strength multiplied by the distance between them.

Example: field generated by an electric eel

Question: Suppose an electric eel is 1 m long, and generates a voltage difference of 1000 volts between its head and tail. What is the electric field in the water around it?

Solution: We are only calculating the amount of field, not its direction, so we ignore positive and negative signs. Subject to the possibly inaccurate assumption of a constant field parallel to the eel's body, we have

$$\begin{aligned} |\mathbf{E}| &= \Delta V/d \\ &= 1000\ \text{V/m} \quad . \end{aligned}$$

Example: relating the units of electric field and voltage

From our original definition of the electric field, we expect it to have units of newtons per coulomb, N/C. The example above, however, came out in volts per meter, V/m. Are these inconsistent? Let's reassure ourselves that this all works. In this kind of situation, the best strategy is usually to simplify the more complex units so that they involve only mks units and coulombs. Since voltage is defined as potential energy per unit charge, it has units of J/C:

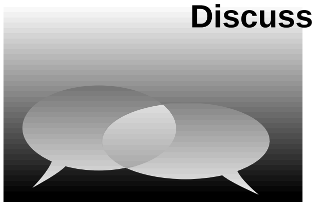
$$\begin{aligned} \frac{V}{m} &= \frac{(J/C)}{m} \\ &= \frac{J}{C \cdot m} \quad . \end{aligned}$$

To connect joules to newtons, we recall that work equals force times distance, so J=N·m, and

$$\begin{aligned} \frac{V}{m} &= \frac{N \cdot m}{C \cdot m} \\ &= \frac{N}{C} \end{aligned}$$

As with other such difficulties with electrical units, one quickly begins to recognize frequently occurring combinations.

Discussion Questions



A. In the definition of the electric field, does the test charge need to be 1 coulomb? Does it need to be positive?

B. Does a charged particle such as an electron or proton feel a force from its own electric field?

C. Is there an electric field surrounding a wall socket that has nothing plugged into it, or a battery that is just sitting on a table?

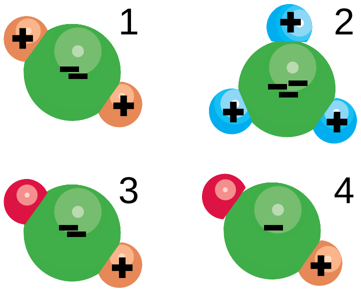
D. In a flashlight powered by a battery, which way do the electric fields point? What would the fields be like inside the wires? Inside the filament of the bulb?

E. Criticize the following statement: “An electric field can be represented by a sea of arrows showing how current is flowing.”

F. The field of a point charge, $|\mathbf{E}|=kQ/r^2$, was derived in the self-check above. How would the field pattern of a uniformly charged sphere compare with the field of a point charge?

G. The interior of a perfect electrical conductor in equilibrium must have zero electric field, since otherwise the free charges within it would be drifting in response to the field, and it would not be in equilibrium. What about the field right at the surface of a perfect conductor? Consider the possibility of a field perpendicular to the surface or parallel to it.

H. Compare the dipole moments of the molecules and molecular ions shown in the figure.



Discussion question H.

I. Small pieces of paper that have not been electrically prepared in any way can be picked up with a charged object such as a charged piece of tape. In our new terminology, we could describe the tape’s charge as inducing a dipole moment in the paper. Can a similar technique be used to induce not just a dipole moment but a charge?

5.4∫ Voltage for Nonuniform Electric Fields

The calculus-savvy reader will have no difficulty generalizing the field-voltage relationship to the case of a varying field. The potential energy associated with a varying force is

$$\Delta PE = -\int |F| \mathbf{dx} \quad \text{[assuming motion parallel to the field],}$$

so for electric fields we divide by q to find

$$\Delta V = -\int |E| \mathbf{dx} \quad \text{[assuming motion parallel to the field].}$$

Those who have studied vector calculus will note that this can be generalized even further as a path integral,

$$\Delta V = -\int E \cdot dx \quad ,$$

where the quantity inside the integral is a vector dot product.

Example: voltage associated with a point charge
Question: What is the voltage associated with a point charge?
Solution: First we need to find the field. We have a field created by a point charge, into which we insert a second charge, the test charge. The force is given by Coulomb's law,

$$|\boldsymbol{F}| = k\frac{qq_t}{r^2} \quad ,$$

resulting in a field

$$|\boldsymbol{E}| = k\frac{q}{r^2} \quad .$$

The difference in voltage between two points on the same radius line is

$$\Delta V = -\int |\boldsymbol{E}| dx$$

In the general discussion above, x was just a generic name for distance traveled along the line from one point to the other, so in this case *x* really means *r*.

$$\begin{aligned}\Delta V &= -\int_{r_1}^{r_2} k\frac{q}{r^2} dr \\ &= -kq\left[-\frac{1}{r_2} - \left(-\frac{1}{r_1}\right)\right] \\ &= kq\left(\frac{1}{r_2} - \frac{1}{r_1}\right) \quad .\end{aligned}$$

The standard convention is to use $r_1=\infty$ as a reference point, so that the voltage is

$$V = \frac{kq}{r} \quad .$$

The interpretation is that if you bring a positive test charge closer to a positive charge, its potential energy is increased; if it was released, it would spring away, releasing this as kinetic energy.

Summary

Selected Vocabulary

field	a property of a point in space describing the forces that would be exerted on a particle if it was there
sink	a point at which field vectors converge
source	a point from which field vectors diverge; often used more inclusively to refer to points of either convergence or divergence
electric field	the force per unit charge exerted on a test charge at a given point in space
gravitational field	the force per unit mass exerted on a test mass at a given point in space
electric dipole	an object that has an imbalance between positive charge on one side and negative charge on the other; an object that will experience a torque in an electric field

Notation

$\boldsymbol{g}$	the gravitational field
$\boldsymbol{E}$	the electric field
D	an electric dipole moment

Notation Used in Other Books

d,p,m	other notations for the electric dipole moment

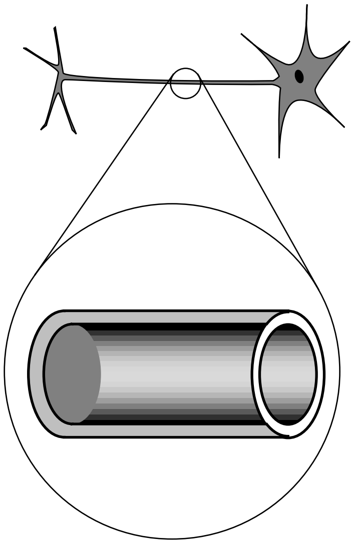
Summary

Newton conceived of a universe where forces reached across space instantaneously, but we now know that there is a delay in time before a change in the configuration of mass and charge in one corner of the universe will make itself felt as a change in the forces experienced far away. We imagine the outward spread of such a change as a ripple in an invisible universe-filling *field of force*.

We define the *gravitational field* at a given point as the force per unit mass exerted on objects inserted at that point, and likewise the *electric field* is defined as the force per unit charge. These fields are vectors, and the fields generated by multiple sources add according to the rules of vector addition.

When the electric field is constant, the voltage difference between two points lying on a line parallel to the field is related to the field by the equation $\Delta V = -Ed$, where d is the distance between the two points.

Homework Problems


Problem 1.

1 ✓. In our by-now-familiar neuron, the voltage difference between the inner and outer surfaces of the cell membrane is about V_{out}-V_{in} = –70 mV in the resting state, and the thickness of the membrane is about 6.0 nm (i.e. only about a hundred atoms thick). What is the electric field inside the membrane?

2. (a✓) The gap between the electrodes in an automobile engine's spark plug is 0.060 cm. To produce an electric spark in a gasoline-air mixture, an electric field of 3.0x10^6 V/m must be achieved. On starting a car, what minimum voltage must be supplied by the ignition circuit? Assume the field is constant. (b) The small size of the gap between the electrodes is inconvenient because it can get blocked easily, and special tools are needed to measure it. Why don't they design spark plugs with a wider gap?

3. (a) At time t=0, a small, positively charged object is placed, at rest, in a uniform electric field of magnitude E. Write an equation giving its speed, v, in terms of t, E, and its mass and charge m and q.

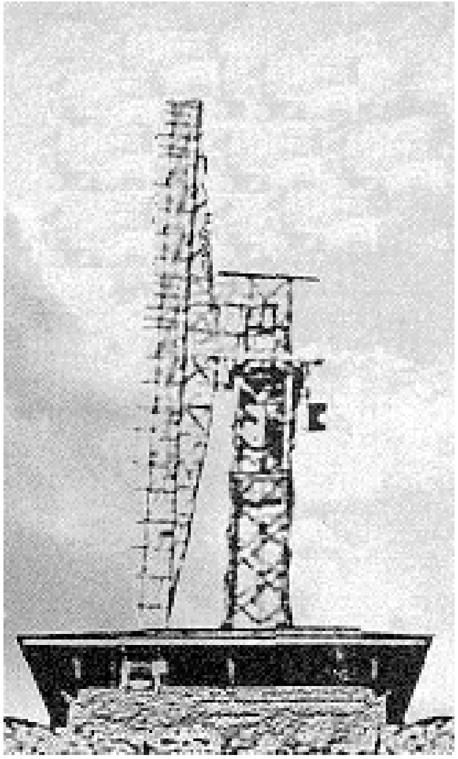
(b) If this is done with two different objects and they are observed to have the same motion, what can you conclude about their masses and charges? (For instance, when radioactivity was discovered, it was found that one form of it had the same motion as an electron in this type of experiment.)

4★. Show that the magnitude of the electric field produced by a simple two-charge dipole, at a distant point along the dipole's axis, is to a good approximation proportional to D/r^3, where r is the distance from the dipole. [Hint: Use the approximation $\left(1+\varepsilon\right)^p \approx 1 + p\varepsilon$, which is valid for small ε.]

5∫. Given that the field of a dipole is proportional to D/r^3 (see previous problem), show that its potential varies as D/r^2. (Ignore positive and negative signs and numerical constants of proportionality.)

6∫. A carbon dioxide molecule is structured like O-C-O, with all three atoms along a line. The oxygen atoms grab a little bit of extra negative charge, leaving the carbon positive. The molecule's symmetry, however, means that it has no overall dipole moment, unlike a V-shaped water molecule, for instance. Whereas the potential of a dipole of magnitude D is proportional to D/r^2 (see previous problem), it turns out that the potential of carbon dioxide molecule along its axis equals k/r^3, where r is the distance from the molecule and k is a constant. What would be the electric field of a carbon dioxide molecule at a distance r?

S A solution is given in the back of the book.	★ A difficult problem.
✓ A computerized answer check is available.	∫ A problem that requires calculus.



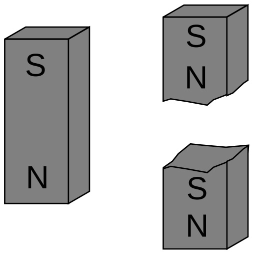
A World Ward II radar installation in Sitka, Alaska.

6 Electromagnetism

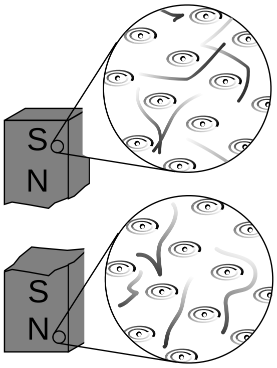
Everyone knows the story of how the nuclear bomb influenced the end of World War II. It was dramatic. It went boom. The Hiroshima bombing, however, came when Germany was already defeated and Japan's surrender was only a matter of time. Far less well known, but more important to the outcome of the war, was a different contribution from physics: radar. This new technology, based on the work of theoretical physicists James Clerk Maxwell in the previous century, played a decisive role in the air war known as the Battle of Britain. If not for the English radar defenses, Hitler might have succeeded in conquering Britain and gone on to win the war.

In this chapter, we discuss the magnetic field, its intimate relationship to the electric field, and the existence of waves made of electric and magnetic fields linked to each other.

6.1 The Magnetic Field



(a)



(b)

No magnetic monopoles

If you could play with a handful of electric dipoles and a handful of bar magnets, they would appear very similar. For instance, a pair of bar magnets wants to align themselves head-to-tail, and a pair of electric dipoles does the same thing. (It is unfortunately not that easy to make a permanent electric dipole that can be handled like this, since the charge tends to leak.)

You would eventually notice an important difference between the two types of objects, however. The electric dipoles can be broken apart to form isolated positive charges and negative charges. The two-ended device can be broken into parts that are not two-ended. But if you break a bar magnet in half, (a), you will find that you have simply made two smaller two-ended objects.

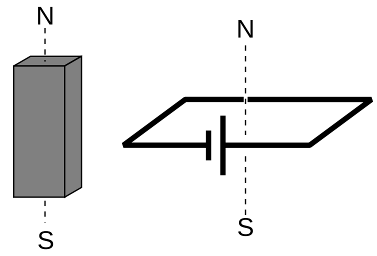
The reason for this behavior is not hard to divine from our microscopic picture of permanent iron magnets. An electric dipole has extra positive "stuff" concentrated in one end and extra negative in the other. The bar magnet, on the other hand, gets its magnetic properties not from an imbalance of magnetic "stuff" at the two ends but from the orientation of the rotation of its electrons. One end is the one from which we could look down the axis and see the electrons rotating clockwise, and the other is the one from which they would appear to go counterclockwise. There is no difference between the "stuff" in one end of the magnet and the other, (b).

Nobody has ever succeeded in isolating a single magnetic pole. In technical language, we say that magnetic *monopoles* do not seem to exist. Electric monopoles *do* exist — that's what charges are.

Electric and magnetic forces seem similar in many ways. Both act at a distance, both can be either attractive or repulsive, and both are intimately related to the property of matter called charge. (Recall that magnetism is an interaction between moving charges.) Physicists's aesthetic senses have been offended for a long time because this seeming symmetry is broken by the existence of electric monopoles and the absence of magnetic ones. Perhaps some exotic form of matter exists, composed of particles that are magnetic monopoles. If such particles could be found in cosmic rays or moon rocks, it would be evidence that the apparent asymmetry was only an asymmetry in the composition of the universe, not in the laws of physics. For these admittedly subjective reasons, there have been several searches for magnetic monopoles. Experiments have been performed, with negative results, to look for magnetic monopoles embedded in ordinary matter. Soviet physicists in the 1960s made exciting claims that they had created and detected magnetic monopoles in particle accelerators, but there was no success in attempts to reproduce the results there or at other accelerators. The most recent search for magnetic monopoles, done by reanalyzing data from the search for the top quark at Fermilab, turned up no candidates, which shows that either monopoles don't exist in nature or they are extremely massive and thus hard to create in accelerators.

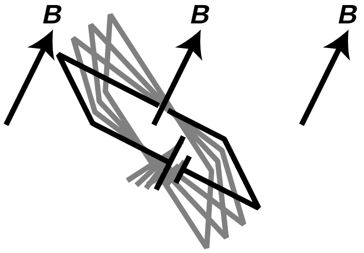
Definition of the magnetic field

Since magnetic monopoles don't seem to exist, it would not make much sense to define a magnetic field in terms of the force on a test monopole. Instead, we follow the philosophy of the alternative definition of the electric field, and define the field in terms of the torque on a magnetic test dipole. This is exactly what a magnetic compass does: the needle is a little iron magnet which acts like a magnetic dipole and shows us the direction of the earth's magnetic field.



(c) A standard dipole made from a square loop of wire shorting across a battery. It acts very much like a bar magnet, but its strength is more easily quantified.

To define the strength of a magnetic field, however, we need some way of defining the strength of a test dipole, i.e. we need a definition of the magnetic dipole moment. We could use an iron permanent magnet constructed according to certain specifications, but such an object is really an extremely complex system consisting of many iron atoms, only some of which are aligned. A more fundamental standard dipole is a square current loop. This could be little resistive circuit consisting of a square of wire shorting across a battery.



(d) A dipole tends to align itself to the surrounding magnetic field.

We will find that such a loop, when placed in a magnetic field, experiences a torque that tends to align plane so that its interior "face" points in a certain direction. Since the loop is symmetric, it doesn't care if we rotate it like a wheel without changing the plane in which it lies. It is this preferred facing direction that we will end up defining as the direction of the magnetic field.

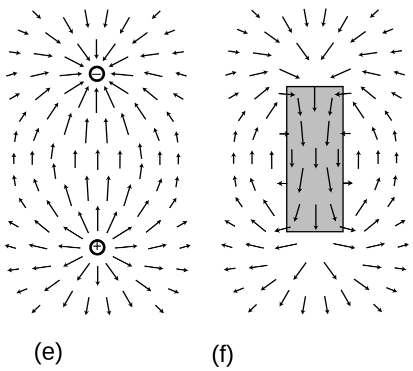
Experiments show if the loop is out of alignment with the field, the torque on it is proportional to the amount of current, and also to the interior area of the loop. The proportionality to current makes sense, since magnetic forces are interactions between moving charges, and current is a measure of the motion of charge. The proportionality to the loop's area is also not hard to understand, because increasing the length of the sides of the square increases both the amount of charge contained in this circular "river" and the amount of leverage supplied for making torque. Two separate physical reasons for a proportionality to length result in an overall proportionality to length squared, which is the same as the area of the loop. For these reasons, we define the magnetic dipole moment of a square current loop as

$$D_{\mathrm{m}} = IA \quad \text{[definition of the magnetic dipole moment of a square loop]} \quad .$$

We now define the magnetic field in a manner entirely analogous to the second definition of the electric field:

> **definition of the magnetic field**
> The magnetic field vector, $\boldsymbol{B}$, at any location in space is defined by observing the torque exerted on a magnetic test dipole D_{mt} consisting of a square current loop. The field's magnitude is $|\boldsymbol{B}| = \tau / D_{\mathrm{mt}} \sin\theta$, where θ is the angle by which the loop is misaligned. The direction of the field is perpendicular to the loop; of the two perpendiculars, we choose the one such that if we look along it, the loop's current is counterclockwise.

We find from this definition that the magnetic field has mks units of $\mathrm{N{\cdot}m/A{\cdot}m^2}$=$\mathrm{N/A{\cdot}m}$. This unwieldy combination of units is abbreviated as the tesla, 1 T=1 $\mathrm{N/A{\cdot}m}$. Refrain from memorizing the part about the counter-



clockwise direction at the end; in section 6.4 we'll see how to understand this in terms of more basic principles.

The nonexistence of magnetic monopoles means that unlike an electric field, (e), a magnetic one, (f), can never have sources or sinks. The magnetic field vectors lead in paths that loop back on themselves, without ever converging or diverging at a point.

6.2 Calculating Magnetic Fields and Forces

Magnetostatics

Our study of the electric field built on our previous understanding of electric forces, which was ultimately based on Coulomb's law for the electric force between two point charges. Since magnetism is ultimately an interaction between currents, i.e. between moving charges, it is reasonable to wish for a magnetic analog of Coulomb's law, an equation that would tell us the magnetic force between any two moving point charges.

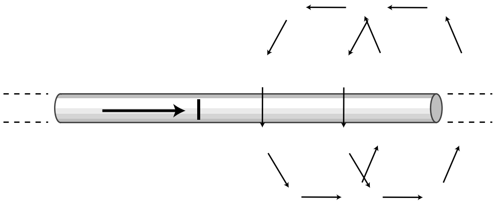
Such a law, unfortunately, does not exist. Coulomb's law describes the special case of electrostatics: if a set of charges is sitting around and not moving, it tells us the interactions among them. Coulomb's law fails if the charges are in motion, since it does not incorporate any allowance for the time delay in the outward propagation of a change in the locations of the charges.

A pair of moving point charges will certainly exert magnetic forces on one another, but their magnetic fields are like the v-shaped bow waves left by boats. Each point charge experiences a magnetic field that originated from the other charge when it was at some previous position. There is no way to construct a force law that tells us the force between them based only on their current positions in space.

There is, however, a science of magnetostatics that covers a great many important cases. Magnetostatics describes magnetic forces among currents in the special case where the currents are steady and continuous, leading to magnetic fields throughout space that do not change over time.

If we cannot build a magnetostatics from a force law for point charges, then where do we start? It can be done, but the level of mathematics required (vector calculus) is inappropriate for this course. Luckily there is an alternative that is more within our reach. Physicists of generations past have used the fancy math to derive simple equations for the fields created by various static current distributions, such as a coil of wire, a circular loop, or a straight wire. Virtually all practical situations can be treated either directly using these equations or by doing vector addition, e.g. for a case like the field of two circular loops whose fields add onto one another.

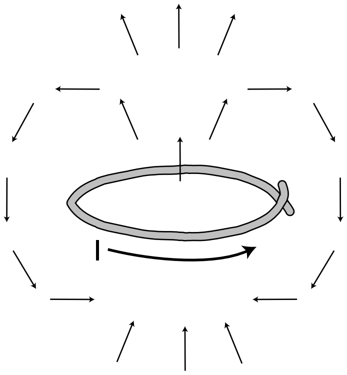
The figure shows the equations for some of the more commonly encountered configurations, with illustrations of their field patterns. Do not memorize the equations! The symbol μ_o is an abbreviation for the constant $4\pi\text{x}10^{-7}$ T·m/A. It is the magnetic counterpart of the Coulomb force constant *k*. The Coulomb constant tells us how much electric field is produced by a given amount of charge, while μ_o relates currents to magnetic fields. Unlike *k*, μ_o has a definite numerical value because of the design of the metric system.



Field created by a long, straight wire carrying current I:

$$B = \frac{\mu_o I}{2\pi r}$$

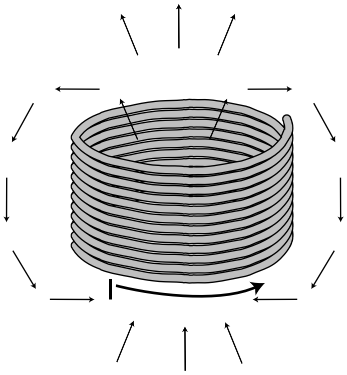
Here *r* is the distance from the center of the wire. The field vectors trace circles in planes perpendicular to the wire, going clockwise when viewed from along the direction of the current.



Field created by a single circular loop of current:
The field vectors form a dipole-like pattern, coming through the loop and back around on the outside. Each oval path traced out by the field vectors appears clockwise if viewed from along the direction the current is going when it punches through it. There is no simple equation for the field at an arbitrary point in space, but for a point lying *along the central axis* perpendicular to the loop, the field is

$$B = \tfrac{1}{2}\mu_o I b^2\left(b^2 + z^2\right)^{-3/2} ,$$

where *b* is the radius of the loop and *z* is the distance of the point from the plane of the loop.



Field created by a solenoid (cylindrical coil):
The field pattern is similar to that of a single loop, but for a long solenoid the paths of the field vectors become very straight on the inside of the coil and on the outside immediately next to the coil. For a sufficiently long solenoid, the interior field also becomes very nearly uniform, with a magnitude of

$$B = \mu_o I N / \ell ,$$

where *N* is the number of turns of wire and ℓ is the length of the solenoid. (Textbooks often give this as $\mu_o In$, where $n = N / \ell$ is the number of turns *per unit length*.) The field near the mouths or outside the coil is not constant and more difficult to calculate. (There is a fairly simple equation for the field along the axis, inside and out, but we will not concern ourselves with it.) For a long solenoid, the exterior field is much smaller than the interior field.

Force on a charge moving through a magnetic field

We now know how to calculate magnetic fields in some typical situations, but one might also like to be able to calculate magnetic forces, such as the force of a solenoid on a moving charged particle, or the force between two parallel current-carrying wires.

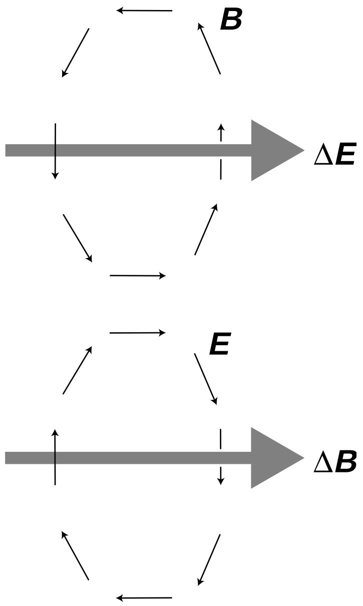
We will restrict ourselves to the case of the force on a charged particle moving through a magnetic field, which allows us to calculate the force between two objects when one is a moving charged particle and the other is one whose magnetic field we know how to find. An example is the use of solenoids inside a TV tube to guide the electron beam as it paints a picture.

Experiments show that the magnetic force on a moving charged particle has a magnitude given by

$$\left|\boldsymbol{F}\right| = q\left|\boldsymbol{v}\right|\left|\boldsymbol{B}\right|\sin\theta \quad ,$$

where ***v*** is the velocity vector of the particle, and θ is the angle between the ***v*** and ***B*** vectors. Unlike electric ana gravitational forces, magnetic forces do not lie along the same line as the field vector. The force is always *perpendicular* to both ***v*** and***B***. Given two vectors, there is only one line perpendicular to both of them, so the force vector points in one of the two possible directions along this line. For a positively charged particle, the direction of the force vector is the one such that if you sight along it, the ***B*** vector is clockwise from the ***v*** vector; for a negatively charged particle the direction of the force is reversed. Note that since the force is perpendicular the particle's motion, the magnetic field never does work on it.

6.3 Induction



(a) The geometry of induced fields. The induced field tends to form a whirlpool pattern around the change in the vector producing it. Note how they circulate in opposite directions.

The theory of electric and magnetic fields constructed up to this point contains a paradox. One of the most basic principles of physics, dating back to Newton and Galileo and still going strong today, states that motion is relative, not absolute. Thus the laws of physics should not function any differently in a moving frame of reference, or else we would be able to tell which frame of reference was the one in an absolute state of rest.

The problem with this was first clearly articulated by Einstein: if we state that magnetism is an interaction between moving charges, we have apparently created a law of physics that violates the principle that motion is relative, since different observers in different frames would disagree about how fast the charges were moving, or even whether they were moving at all.

The story of this puzzle is a long and complex one, but its resolution by Einstein and his successors in the 1950's was based on an idea from the work of Maxwell and others all the way back in the 1800's. Maxwell described the electric and magnetic fields as closely related phenomena, two sides of the same coin. Stripped of the mathematics, the relationship is as follows:

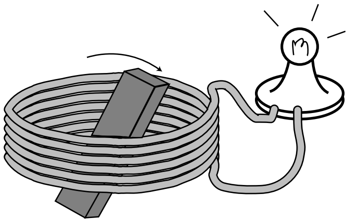
the principle of induction
Any electric field that changes over time will produce a magnetic field in the space around it.
Any magnetic field that changes over time will produce an electric field in the space around it.

The induced field tends to have a whirlpool pattern, as shown in figure (a).

A full discussion of how this solves the relative motion problem is beyond the scope of this book, but on a more pedestrian level, the induction principle has some important applications. We will restrict ourselves to a nonmathematical discussion.

Example: the generator

A generator, (b), consists of a permanent magnet that rotates within a coil of wire. The magnet is turned by a motor or crank, (not shown). As it spins, the nearby magnetic field changes. According to the principle of induction, this changing magnetic field results in an electric field, which has a whirlpool pattern. This electric field pattern creates a current that whips around the coils of wire, and we can tape this current to light the lightbulb.

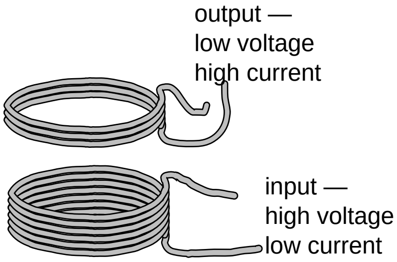


(b) A generator.

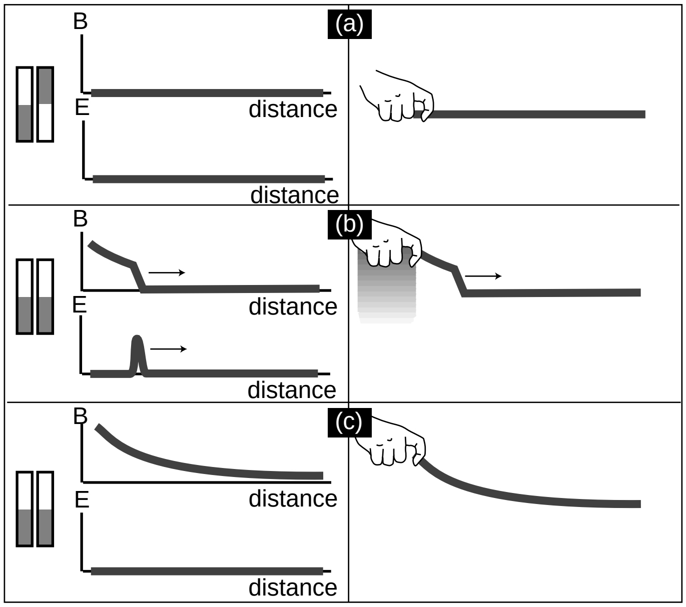
Example: the transformer

In section 4.3 we discussed the advantages of transmitting power over electrical lines using high voltages and low currents. However, we don’t want our wall sockets to operate at 10000 volts! For this reason, the electric company uses a device called a transformer, (c), to convert to lower voltages and higher currents inside your house. The coil on the input side creates a magnetic field. Transformers work with alternating current, so the magnetic field surrounding the input coil is always changing. This induces an electric field, which drives a current around the output coil.

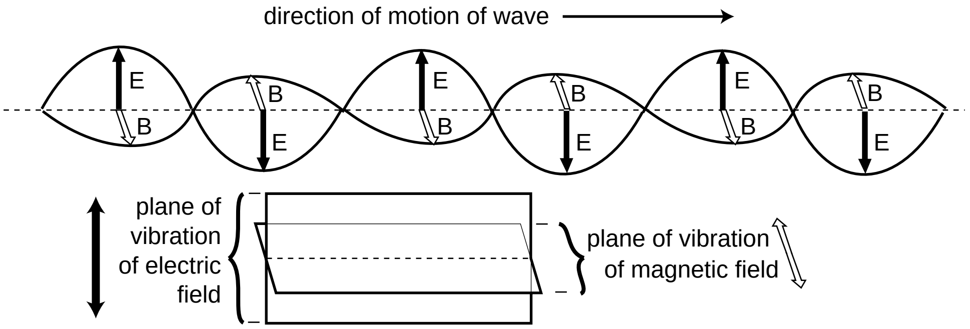
If both coils were the same, the arrangement would be symmetric, and the output would be the same as the input, but an output coil with a smaller number of coils gives the electric forces a smaller distance through which to push the electrons. Less mechanical work per unit charge means a lower voltage. Conservation of energy, however, guarantees that the amount of power on the output side must equal the amount put in originally, $I_{in}V_{in}=I_{out}V_{out}$, so this reduced voltage must be accompanied by an increased current.



(c) A transformer.



An example of induction (left) with a mechanical analogy (right). The two bar magnets are initially pointing in opposite directions, (a), and their magnetic fields cancel out. If one magnet is flipped, (b), their fields reinforce, but the change in the magnetic field takes time to spread through space. Eventually, (c), the field becomes what you would expect from the theory of magnetostatics. In the mechanical analogy, the sudden motion of the hand produces a violent kink or wave pulse in the rope, the pulse travels along the rope, and it takes some time for the rope to settle down. An electric field is also induced in (b) by the changing magnetic field, eventhough there is no net charge anywhere to to act as a source. (These simplified drawings are not meant to be accurate representations of the complete three-dimensional pattern of electric and magnetic fields.)


direction of motion of wave
E
B
E
B
E
B
E
B
E
B
E
B
plane of vibration of electric field
plane of vibration of magnetic field

6.4 Electromagnetic Waves

The most important consequence of induction is the existence of electromagnetic waves. Whereas a gravitational wave would consist of nothing more than a rippling of gravitational fields, the principle of induction tells us that there can be no purely electrical or purely magnetic waves. Instead, we have waves in which there are both electric and magnetic fields, such as the sinusoidal one shown in the figure. Maxwell proved that such waves were a direct consequence of his equations, and derived their properties mathematically. The derivation would be beyond the mathematical level of this book, so we will just state the results.

A sinusoidal electromagnetic wave has the geometry shown in the figure above. The ***E*** and ***B*** fields are perpendicular to the direction of motion, and are also perpendicular to each other. If you look along the direction of motion of the wave, the ***B*** vector is always 90 degrees clockwise from the ***E*** vector. The magnitudes of the two fields are related by the equation $|\boldsymbol{E}|=c|\boldsymbol{B}|$.

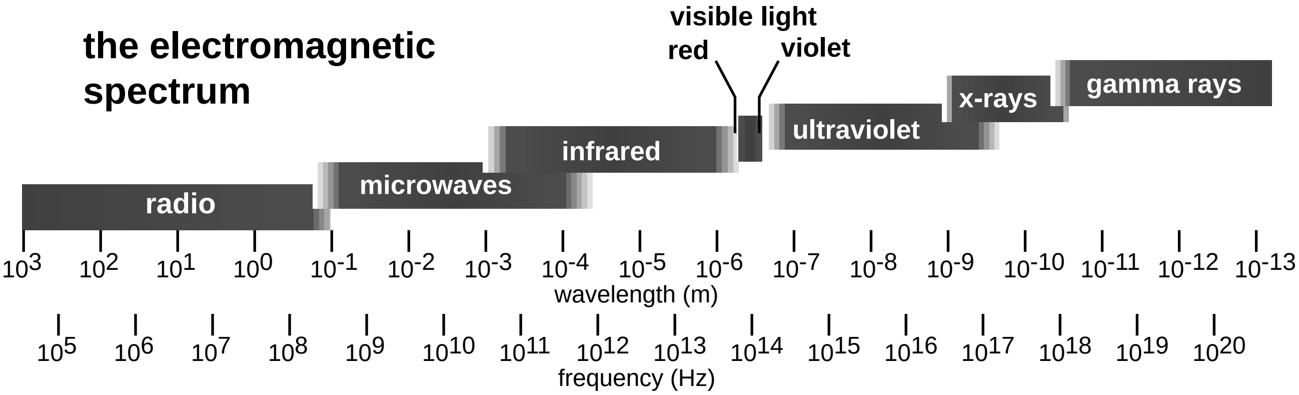
How is an electromagnetic wave created? It could be emitted, for example, by an electron orbiting an atom or currents going back and forth in a transmitting antenna. In general any accelerating charge will create an electromagnetic wave, although only a current that varies sinusoidally with time will create a sinusoidal wave. Once created, the wave spreads out through space without any need for charges or currents along the way to keep it going. As the electric field oscillates back and forth, it induces the magnetic field, and the oscillating magnetic field in turn creates the electric field. The whole wave pattern propagates through empty space at a velocity c=3.0x10^8 m/s, which is related to the constants k and μ_o by $c=\sqrt{4\pi k/\mu_o}$.

Polarization

Two electromagnetic waves traveling in the same direction through space can differ by having their electric and magnetic fields in different directions, a property of the wave called its polarization.

Light is an electromagnetic wave

Once Maxwell had derived the existence of electromagnetic waves, he became certain that they were the same phenomenon as light. Both are transverse waves (i.e. the vibration is perpendicular to the direction the wave is moving), and the velocity is the same. He is said to have gone for a walk with his wife one night and told her that she was the only other person in the world who knew what starlight really was.



Heinrich Hertz (for whom the unit of frequency is named) verified Maxwell's ideas experimentally. Hertz was the first to succeed in producing, detecting, and studying electromagnetic waves in detail using antennas and electric circuits. To produce the waves, he had to make electric currents oscillate very rapidly in a circuit. In fact, there was really no hope of making the current reverse directions at the frequencies of 10^{15} Hz possessed by visible light. The fastest electrical oscillations he could produce were 10^9 Hz, which would give a wavelength of about 30 cm. He succeeded in showing that the waves he produced were polarizable like light, and could be reflected and refracted just like light, and he built devices such as parabolic mirrors that worked according to the same optical principles as those employing light. Hertz's results were convincing evidence that light and electromagnetic waves were one and the same.

The electromagnetic spectrum

Today, electromagnetic waves with frequencies in the range employed by Hertz are known as radio waves. Any remaining doubts that the "Hertzian waves," as they were then called, were the same type of wave as light waves were soon dispelled by experiments in the whole range of frequencies in between, as well as the frequencies outside that range. In analogy to the spectrum of visible light, we speak of the entire electromagnetic spectrum, of which the visible spectrum is one segment.

The terminology for the various parts of the spectrum is worth memorizing, and is most easily learned by recognizing the logical relationships between the wavelengths and the properties of the waves with which you are already familiar. Radio waves have wavelengths that are comparable to the size of a radio antenna, i.e. meters to tens of meters. Microwaves were named that because they have much shorter wavelengths than radio waves; when food heats unevenly in a microwave oven, the small distances between neighboring hot and cold spots is half of one wavelength of the standing wave the oven creates. The infrared, visible, and ultraviolet obviously have much shorter wavelengths, because otherwise the wave nature of light would have been as obvious to humans as the wave nature of ocean waves. To remember that ultraviolet, x-rays, and gamma rays all lie on the short-wavelength side of visible, recall that all three of these can cause cancer. (As we'll discuss later in the course, there is a basic physical reason why the cancer-causing disruption of DNA can only be caused by very short-wavelength electromagnetic waves. Contrary to popular belief, microwaves cannot cause cancer, which is why we have microwave ovens and not x-ray ovens!)

6.5 Calculating Energy in Fields

We have seen that the energy stored in a wave (actually the energy density) is typically proportional to the square of the wave's amplitude. Fields of force can make wave patterns, for which we might expect the same to be true. This turns out to be true not only for wave-like field patterns but for all fields:

$$\text{energy stored in the gravitational field per m}^3 = -\frac{1}{8\pi G}\left|\boldsymbol{g}\right|^2$$

$$\text{energy stored in the electric field per m}^3 = \frac{1}{8\pi k}\left|\boldsymbol{E}\right|^2$$

$$\text{energy stored in the magnetic field per m}^3 = \frac{1}{2\mu_o}\left|\boldsymbol{B}\right|^2$$

Although funny factors of 8π and the plus and minus signs may have initially caught your eye, they are not the main point. The important idea is that the energy density is proportional to the square of the field strength in all three cases. We first give a simple numerical example and work a little on the concepts and then turn our attention to the factors out in front.

Example: Getting killed by a solenoid

Solenoids are very common electrical devices, but they can be a hazard to someone who is working on them. Imagine a solenoid that initially has a DC current passing through it. The current creates a magnetic field inside and around it, which contains energy. Now suppose that we break the circuit. Since there is no longer a complete circuit, current will quickly stop flowing, and the magnetic field will collapse very quickly. The field had energy stored in it, and even a small amount of energy can create a dangerous power surge if released over a short enough time interval. It is prudent not to fiddle with a solenoid that has current flowing through it, since breaking the circuit could be hazardous to your health.

As a typical numerical estimate, let's assume a 40 cm x 40 cm x 40 cm solenoid with an interior magnetic field of 1.0 T (quite a strong field) For the sake of this rough estimate, we ignore the exterior field, which is weak, and assume that the solenoid is cubical in shape. The energy stored in the field is

(energy per unit volume)(volume)

$$= \frac{1}{2\mu_o}\left|\boldsymbol{B}\right|^2 V$$

$$= 3\mathrm{x}10^{4}\ \mathrm{J}$$

That's a lot of energy!

In chapter 5 when we discussed the original reason for introducing the concept of a field of force, a prime motivation was that otherwise there was no way to account for the energy transfers involved when forces were delayed by an intervening distance. We used to think of the universe's energy as consisting of

kinetic energy

+ gravitational potential energy based on the distances between objects that interact gravitationally

+ electric potential energy based on the distances between objects that interact electrically

+ magnetic potential energy based on the distances between objects that interact magnetically

but in nonstatic situations we must use a different method:

kinetic energy

+ gravitational potential energy stored in gravitational fields

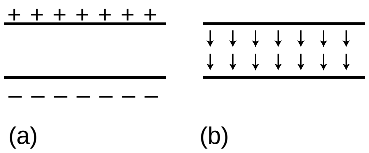
+ electric potential energy stored in electric fields

+ magnetic potential stored in magnetic fields

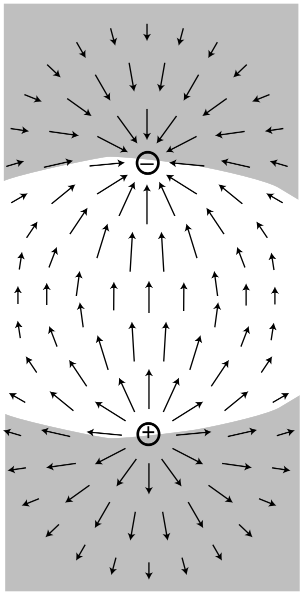
Surprisingly, the new method still gives the same answers for the static cases.

Example: energy stored in a capacitor

A pair of parallel metal plates, seen from the side in figures (a) and (b), can be used to store electrical energy by putting positive charge on one side and negative charge on the other. Such a device is called a capacitor. (We have encountered such an arrangement previously, but there its purpose was to deflect a beam of electrons, not to store energy.)



In the old method of describing potential energy, (a), we think in terms of the mechanical work that had to be done to separate the positive and negative charges onto the two plates, working against their electrical attraction. The new description, (b), attributes the storage of energy to the newly created electric field occupying the volume between the plates. Since this is a static case, both methods give the same, correct answer.



(c)

Example: potential energy of a pair of opposite charges

Imagine taking two opposite charges, (c), that were initially far apart and allowing them to come together under the influence of their electrical attraction.

According to the old method, potential energy is lost because the electric force did positive work as it brought the charges together. (This makes sense because as they come together and accelerate it is their potential energy that is being lost and converted to kinetic energy.)

By the new method, we must ask how the energy stored in the electric field has changed. In the region indicated approximately by the shading in the figure, the superposing fields of the two charges undergo partial cancellation because they are in opposing directions. The energy in the shaded region is reduced by this effect. In the unshaded region, the fields reinforce, and the energy is increased.

It would be quite a project to do an actual numerical calculation of the energy gained and lost in the two regions (this is a case where the old method of finding energy gives greater ease of computation), but it is fairly easy to convince oneself that the energy is less when the charges are closer. This is because bringing the charges shrinks the high-energy unshaded region and enlarges the low-energy shaded region.

Example: energy in an electromagnetic wave

The old method would give zero energy for a region of space containing an electromagnetic wave but no charges. That would be wrong! We can only use the old method in static cases.

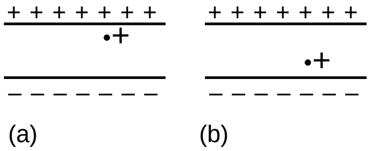
Now let's give at least some justification for the other features of the three expressions for energy density, $-\frac{1}{8\pi G}\left|\boldsymbol{g}\right|^2$, $\frac{1}{8\pi k}\left|\boldsymbol{E}\right|^2$, and $\frac{1}{2\mu_o}\left|\boldsymbol{B}\right|^2$, besides the proportionality to the square of the field strength.

First, why the different plus and minus signs? The basic idea is that the signs have to be opposite in the gravitational and electric cases because there is an attraction between two positive masses (which are the only kind that exist), but two positive charges would repel. Since we've already seen examples where the positive sign in the electric energy makes sense, the gravitational energy equation must be the one with the minus sign.

It may also seem strange that the constants G, k, and μ_o are in the denominator. They tell us how strong the three different forces are, so shouldn't they be on top? No. Consider, for instance, an alternative universe in which gravity is twice as strong as in ours. The numerical value of G is doubled. Because G is doubled, all the gravitational field strengths are doubled as well, which quadruples the quantity $\left|\mathbf{g}\right|^2$. In the expression $-\frac{1}{8\pi G}\left|\mathbf{g}\right|^2$, we have quadrupled something on top and doubled something on the bottom, which makes the energy twice as big. That makes perfect sense.

Discussion Question

A. The figure shows a positive charge in the gap between two capacitor plates. Compare the energy of the electric fields in the two cases. Does this agree with what you would have expected based on your knowledge of electrical forces?



B. Criticize the following statement: "A solenoid makes a charge in the space surrounding it, which dissipates when you release the energy."

Summary

Selected Vocabulary

magnetic field .... a field of force, defined in terms of the torque exerted on a test dipole

magnetic dipole .... an object, such as a current loop, an atom, or a bar magnet, that experiences torques due to magnetic forces; the strength of magnetic dipoles is measured by comparison with a standard dipole consisting of a square loop of wire of a given size and carrying a given amount of current

induction .... the production of an electric field by a changing magnetic field, or vice-versa

Notation

$\boldsymbol{B}$ .... the magnetic field

D_m .... magnetic dipole moment

Summary

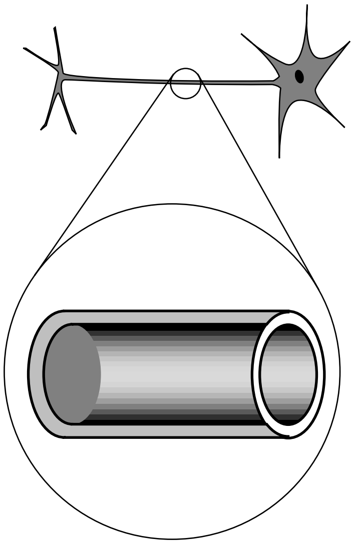
Magnetism is an interaction of moving charges with other moving charges. The magnetic field is defined in terms of the torque on a magnetic test dipole. It has no sources or sinks; magnetic field patterns never converge on or diverge from a point.

The magnetic and electric fields are intimately related. The principle of induction states that any changing electric field produces a magnetic field in the surrounding space, and vice-versa. These induced fields tend to form whirlpool patterns.

The most important consequence of the principle of induction is that there are no purely magnetic or purely electric waves. Disturbances in the electrical and magnetic fields propagate outward as combined magnetic and electric waves, with a well-defined relationship between their magnitudes and directions. These electromagnetic waves are what light is made of, but other forms of electromagnetic waves exist besides visible light, including radio waves, x-rays, and gamma rays.

Fields of force contain energy. The density of energy is proportional to the square of the magnitude of the field. In the case of static fields, we can calculate potential energy either using the previous definition in terms of mechanical work or by calculating the energy stored in the fields. If the fields are not static, the old method gives incorrect results and the new one must be used.

Homework Problems



Problem 2.

1. In an electrical storm, the cloud and the ground act like a parallel-plate capacitor, which typically charges up due to frictional electricity in collisions of ice particles in the cold upper atmosphere. Lightning occurs when the magnitude of the electric field builds up to a critical value, E_c, at which air is ionized.
(a) Treat the cloud as a flat square with sides of length L. If it is at a height h above the ground, find the amount of energy released in the lightning strike.
(b) Based on your answer from part (a), which is more dangerous, a lightning strike from a high-altitude cloud or a low-altitude one?
(c) Make an order-of-magnitude estimate of the energy released by a typical lightning bolt, assuming reasonable values for its size and altitude. E_c is about 10^4 V/m.

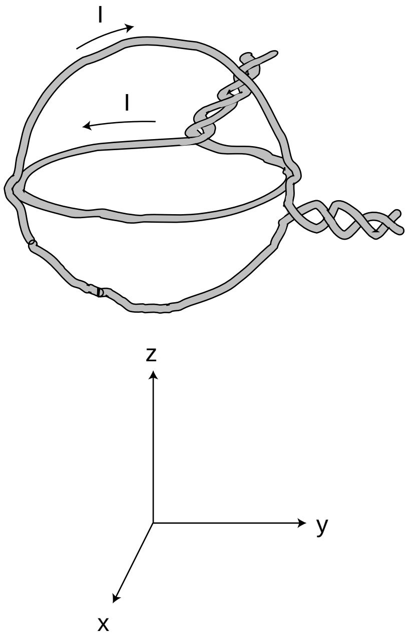
2. The neuron in the figure has been drawn fairly short, but some neurons in your spinal cord have tails (axons) up to a meter long. The inner and outer surfaces of the membrane act as the "plates" of a capacitor. (The fact that it has been rolled up into a cylinder has very little effect.) In order to function, the neuron must create a voltage difference V between the inner and outer surfaces of the membrane. Let the membrane's thickness, radius, and length be t, r, and L.

(a) Calculate the energy that must be stored in the electric field for the neuron to do its job. (In real life, the membrane is made out of a substance called a dielectric, whose electrical properties increase the amount of energy that must be stored. For the sake of this analysis, ignore this fact.) [Hint: The volume of the membrane is essentially the same as if it was unrolled and flattened out.]

(b) An organism's evolutionary fitness should be better if it needs less energy to operate its nervous system. Based on your answer to part (a), what would you expect evolution to do to the dimensions t and r? What other constraints would keep these evolutionary trends from going too far?

3. Consider two solenoids, one of which is smaller so that it can be put inside the other. Assume they are long enough so that each one only contributes significantly to the field inside itself, and the interior fields are nearly uniform. Consider the configuration where the small one is inside the big one with their currents circulating in the same direction, and a second configuration in which the currents circulate in opposite directions. Compare the energies of these configurations with the energy when the solenoids are far apart. Based on this reasoning, which configuration is stable, and in which configuration will the little solenoid tend to get twisted around or spit out? [Hint: A stable system has low energy; energy would have to be added to change its configuration.]

S A solution is given in the back of the book.	★ A difficult problem.
✓ A computerized answer check is available.	∫ A problem that requires calculus.



Problem 4.

4. The figure shows a nested pair of circular wire loops used to create magnetic fields. (The twisting of the leads is a practical trick for reducing the magnetic fields they contribute, so the fields are very nearly what we would expect for an ideal circular current loop.) The coordinate system below is to make it easier to discuss directions in space. One loop is in the y-z plane, the other in the x-y plane. Each of the loops has a radius of 1.0 cm, and carries 1.0 A in the direction indicated by the arrow.
(a✓) Using the equation in optional section 6.2, calculate the magnetic field that would be produced by *one* such loop, at its center.
(b) Describe the direction of the magnetic field that would be produced, at its center, by the loop in the x-y plane alone.
(c) Do the same for the other loop.
(d✓) Calculate the magnitude of the magnetic field produced by the two loops in combination, at their common center. Describe its direction.

5. (a) Show that the quantity $\sqrt{4\pi k / \mu_o}$ has units of velocity.

(b) Calculate it numerically and show that it equals the speed of light.

(c) Prove that in an electromagnetic wave, half the energy is in the electric field and half in the magnetic field.

6. One model of the hydrogen atom has the electron circling around the proton at a speed of $2.2\text{x}10^6$ m/s, in an orbit with a radius of 0.05 nm. (Although the electron and proton really orbit around their common center of mass, the center of mass is very close to the proton, since it is 2000 times more massive. For this problem, assume the proton is stationary.) In that previous homework problem, you calculated the electric current created.
(a✓) Now estimate the magnetic field created at the center of the atom by the electron. We are treating the circling electron as a current loop, even though it's only a single particle.
(b) Does the proton experience a nonzero force from the electron's magnetic field? Explain.
(c) Does the electron experience a magnetic field from the proton? Explain.
(d) Does the electron experience a magnetic field created by its own current? Explain.
(e✓) Is there an electric force acting between the proton and electron? If so, calculate it.
(f) Is there a gravitational force acting between the proton and electron? If so, calculate it.
(g) An inward force is required to keep the electron in its orbit -- otherwise it would obey Newton's first law and go straight, leaving the atom. Based on your answers to the previous parts, which force or forces (electric, magnetic and gravitational) contributes significantly to this inward force?

7. [You need to have read optional section 6.2 to do this problem.] Suppose a charged particle is moving through a region of space in which there is an electric field perpendicular to its velocity vector, and also a magnetic field perpendicular to both the particle's velocity vector and the electric field. Show that there will be one particular velocity at which the particle can be moving that results in a total force of zero on it. Relate this velocity to the magnitudes of the electric and magnetic fields. (Such an arrangement, called a velocity filter, is one way of determining the speed of an unknown particle.)

8. If you put four times more energy through a solenoid, how many times more energy is stored in its magnetic field?

9 ★. Suppose we are given a permanent magnet with a complicated, asymmetric shape. Describe how a series of measurements with a magnetic compass could be used to determine the strength and direction of its magnetic field at some point of interest. Assume that you are only able to see the direction to which the compass needle settles; you cannot measure the torque acting on it.

10. Consider two solenoids, one of which is smaller so that it can be put inside the other. Assume they are long enough so that each one only contributes significantly to the field inside itself, and the interior fields are nearly uniform. Consider the configuration where the small one is partly inside and partly hanging out of the big one, with their currents circulating in the same direction.

(a) Find the magnetic potential energy as a function of the length *x* of the part of the small solenoid that is inside the big one. (Your equation will include other relevant variables describing the two solenoids.)

(b) Based on your answer to part (a), find the force acting between the solenoids.

Glossary

Alpha decay. The radioactive decay of a nucleus via emission of an alpha particle.

Alpha particle. A form of radioactivity consisting of helium nuclei.

Ammeter. A device for measurin electrical current.

Ampere. The metric unit of current, one coulomb pe second; also "amp."

Atom. The basic unit of one of the chemical elements.

Atomic mass. The mass of an atom.

Atomic number. The number of protons in an atom's nucleus; determines what element it is.

Beta decay. The radioactive decay of a nucleus via the reaction n Æ p + e$^-$ + $\bar{\nu}$ or p Æ n + e$^+$ + n; so called because an electron or antielectron is also known as a beta particle.

Beta particle. A form of radioactivity consisting of electrons.

Cathode ray. The mysterious ray that emanated from the cathode in a vacuum tube; shown by Thomson to be a stream of particles smaller than atoms.

Charge. A numerical rating of how strongly an object participates in electrical forces.

Circuit. An electrical device in which charge can come back to its starting point and be recycled rather than getting stuck in a dead end.

Coulomb (C). The unit of electrical charge.

Current. The rate at which charge crosses a certain boundary.

Electric dipole. An object that has an imbalance between positive charge on one side and negative charge on the other; an object that will experience a torque in an electric field.

Electric field. The force per unit charge exerted on a test charge at a given point in space.

Electrical force. One of the fundamental forces of nature; a noncontact force that can be either repulsive or attractive.

Electron. Thomson's name for the particles of which a cathode ray was made; a subatomic particle.

Field. A property of a point in space describing the forces that would be exerted on a particle if it was there.

Fission. The radioactive decay of a nucleus by splitting into two parts.

Fusion. A nuclear reaction in which two nuclei stick together to form one bigger nucleus.

Gamma ray. Aform of radioactivity consisting of a very high-frequency form of light.

Gravitational field. The force per unit mass exerted on a test mass at a given point in space.

Induction. The production of an electric field by a changing magnetic field, or vice-versa.

Ion. An electrically charged atom or molecule.

Isotope. One of the possible varieties of atoms of a given element, having a certain number of neutrons.

Magnetic dipole. An object, such as a current loop, an atom, or a bar magnet, that experiences torques due to magnetic forces; the strength of magnetic dipoles is measured by comparison with a standard dipole consisting of a square loop of wire of a given size and carrying a given amount of current.

Magnetic field. A field of force, defined in terms of the torque exerted on a test dipole.

Mass number. The number of protons plus the number of neutrons in a nucleus; approximately proportional to its atomic mass.

Millirem. A unit for measuring a person's exposure to radioactivity; cf rem.

Molecule. A group of atoms stuck together.

Neutron. An uncharged particle, the other types that nuclei are made of.

Ohm. The metric unit of electrical resistance, one volt per ampere.

Ohmic. Describes a substance in which the flow of current between two points is proportional to the voltage difference between them.

Open circuit. A circuit that does not function because it has a gap in it.

Proton. A positively charged particle, one of the types that nuclei are made of.

Quantized. Describes quantity such as money or electrical charge, that can only exist in certain amounts.

Rem. A unit for measuring a person's exposure to radioactivity; cf millirem.

Resistance. The ratio of the voltage difference to the current in an object made of an ohmic substance.

Short circuit. A circuit that does not function because charge is given a low-resistance "shortcut" path that it can follow, instead of the path that makes it do something useful.

Sink. A point at which field vectors converge.

Source. A point from which field vectors diverge; often used more inclusively to refer to points of either convergence or divergence.

Strong nuclear force. The force that holds nuclei together against electrical repulsion.

Volt. The metric unit of voltage, one joule per coulomb.

Voltage. Electrical potential energy per unit charge that will be posessed by a charged particle at a certain point in space.

Voltmeter. A device for measuring voltage differences.

Weak nuclear force. The force responsible for beta decay.

Index

Photo Credits

All photographs are by Benjamin Crowell, except as noted below. Credits marked with an asterisk are photographs or illustrations from the turn of the 20th century whose copyrights have expired.

Cover
Cat neuron: Courtesy of Richard J. Harris, University of Western Ontario. *Accelerator:* Courtesy of A. Zachau, GSI.

Chapter 1
Lightning: C. Clark/NOAA photo library, uncopyrighted. *J.J. Thomson:* Harper's Monthly, 1904.* *Millikan's apparatus:* Caltech.* *Robert Millikan:* Clark Millikan, 1891.*

Chapter 2
Curies: Harper's Monthly, 1904.* *GAMMASPHERE*: C.J. Lister/R.V.F. Janssens, uncopyrighted.

Chapter 3
Knifefish: Courtesy of Greg DeGreef. *Superconducting accelerator segment:* Courtesy of Argonne National Laboratory, managed and operated by the University of Chicago for the U.S. Department of Energy under contract No. W-31-109-ENG-38.

Chapter 6
Radar: U.S. Forestry Service.

Useful Data

Metric Prefixes

M-	mega-	10^6
k-	kilo-	10^3
m-	milli-	10^{-3}
μ- (Greek mu)	micro-	10^{-6}
n-	nano-	10^{-9}
p-	pico-	10^{-12}
f-	femto-	10^{-15}

(Centi-, 10^{-2}, is used only in the centimeter.)

Notation and Units

quantity	**unit**	**symbol**
distance	meter, m	x, Δx
time	second, s	t, Δt
mass	kilogram, kg	m
density	kg/m^3	ρ
force	newton, 1 N=1 kg·m/s^2	$\boldsymbol{F}$
velocity	m/s	$\boldsymbol{v}$
acceleration	m/s^2	$\boldsymbol{a}$
energy	joule, J	E
momentum	kg·m/s	$\boldsymbol{p}$
angular momentum	kg·m^2/s	L
period	s	T
wavelength	m	λ
frequency	s^{-1} or Hz	f
charge	coulomb, C	q
current	ampere, A	I
voltage (potential)	volt, V	V
resistance	ohm, Ω	R
electric field	N/C	$\boldsymbol{E}$
magnetic field	tesla, 1 T=1 N·s/C·m	$\boldsymbol{B}$

Fundamental Constants

gravitational constant	G=6.67x10^{-11} N·m^2/kg^2
Coulomb constant	k=8.99x10^9 N·m^2/C^2
quantum of charge	e=1.60x10^{-19} C
speed of light	c=3.00x10^8 m/s
magnetic field constant	μ_o=4πx10^{-7} T·m/A

Conversions

Conversions between SI and other units:

1 inch	=	2.54 cm (exactly)
1 mile	=	1.61 km
1 pound	=	4.45 N
(1 kg)·g	=	2.2 lb
1 gallon	=	3.78x10^3 cm^3
1 horsepower	=	746 W
1 kcal*	=	4.18x10^3 J

*When speaking of food energy, the word “Calorie” is used to mean 1 kcal, i.e. 1000 calories. In writing, the capital C may be used to indicate 1 Calorie=1000 calories.

Conversions between U.S. units:

1 foot	=	12 inches
1 yard	=	3 feet
1 mile	=	5280 ft

Earth, Moon, and Sun

body	**mass (kg)**	**radius (km)**	**radius of orbit (km)**
earth	5.97x10^{24}	6.4x10^3	1.49x10^8
moon	7.35x10^{22}	1.7x10^3	3.84x10^5
sun	1.99x10^{30}	7.0x10^5	

The radii and radii of orbits are average values. The moon orbits the earth and the earth orbits the sun.

Subatomic Particles

particle	**mass (kg)**	**charge**	**radius (fm)**
electron	9.109x10^{-31}	$-e$	<~0.01
proton	1.673x10^{-27}	$+e$	~1.1
neutron	1.675x10^{-27}	0	~1.1
neutrino	~10^{-39} kg?	0	?

The radii of protons and neutrons can only be given approximately, since they have fuzzy surfaces. For comparison, a typical atom is about a million fm in radius.

Optics

Benjamin Crowell

Book 5 in the **Light and Matter** series of introductory physics textbooks
www.lightandmatter.com

Optics

The **Light and Matter** series of introductory physics textbooks:

1 Newtonian Physics

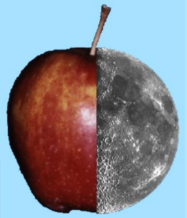
2 Conservation Laws

3 Vibrations and Waves

4 Electricity and Magnetism

5 Optics

6 The Modern Revolution in Physics



Optics

Benjamin Crowell

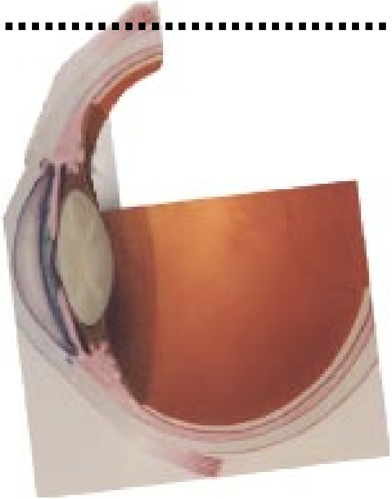
www.lightandmatter.com

Light and Matter
Fullerton, California
www.lightandmatter.com

Edition 2.0
rev. 2000-09-09

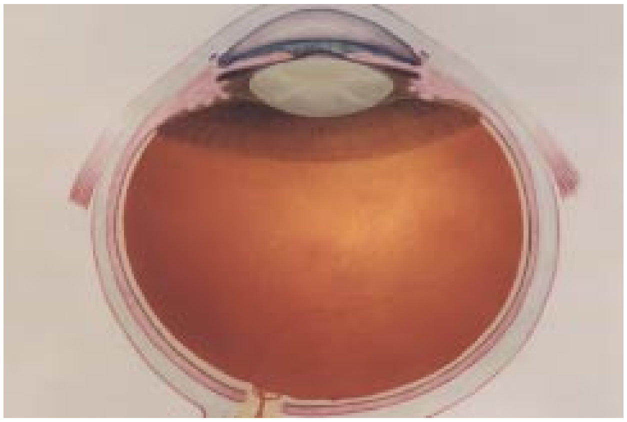
ISBN 0-9704670-5-2

Brief Contents



Contents

A Syncopated March and Two-Step
By
SCOTT JOPLIN
WRITER OF THE FAMOUS
MAPLE LEAF RAG
ETC.
SEARCHLIGHT RAG



4 Refraction and Images 43

5 Wave Optics 57

Exercises 73

Solutions to Selected Problems 79

Glossary 81

Index 83

Useful Data 90

1 The Ray Model of Light

Ads for the latest Macintosh computer brag that it can do an arithmetic calculation in less time than it takes for the light to get from the screen to your eye. We find this impressive because of the contrast between the speed of light and the speeds at which we interact with physical objects in our environment. Perhaps it shouldn't surprise us, then, that Newton succeeded so well in explaining the motion of objects, but was far less successful with the study of light.

This textbook series is billed as the Light and Matter series, but only now, in the fifth of the six volumes, are we ready to focus on light. If you are reading the series in order, then you know that the climax of our study of electricity and magnetism was discovery that light is an electromagnetic wave. Knowing this, however, is not the same as knowing everything about eyes and telescopes. In fact, the full description of light as a wave can be rather cumbersome. We will instead spend most of this book making use of a simpler model of light, the ray model, which does a fine job in most practical situations. Not only that, but we will even backtrack a little and start with a discussion of basic ideas about light and vision that predated the discovery of electromagnetic waves. Research in physics eduction has shown conclusively that the mere assertion that light is an electromagnetic wave is woefully insufficient to allow a student to interpret ordinary phenomena involving light.

1.1 The Nature of Light

The cause and effect relationship in vision

Despite its title, this chapter is far from your first look at light. That familiarity might seem like an advantage, but most people have never thought carefully about light and vision. Even smart people who have thought hard about vision have come up with incorrect ideas. The ancient Greeks, Arabs and Chinese had theories of light and vision, all of which were mostly wrong, and all of which were accepted for thousands of years.

One thing the ancients did get right is that there is a distinction between objects that emit light and objects that don't. When you see a leaf in the forest, it's because three different objects are doing their jobs: the leaf, the eye, and the sun. But luminous objects like the sun, a flame, or the filament of a light bulb can be seen by the eye without the presence of a third object. Emission of light is often, but not always, associated with heat. In modern times, we are familiar with a variety of objects that glow without being heated, including fluorescent lights and glow-in-the-dark toys.

How do we see luminous objects? The Greek philosophers Pythagoras (b. ca. 560 BC) and Empedocles of Acragas (b. ca. 492 BC), who unfortunately were very influential, claimed that when you looked at a candle flame, the flame and your eye were both sending out some kind of mysterious stuff, and when your eye's stuff collided with the candle's stuff, the candle would become evident to your sense of sight.

Bizarre as the Greek "collision of stuff theory" might seem, it had a couple of good features. It explained why both the candle and your eye had to be present for your sense of sight to function. The theory could also easily be expanded to explain how we see nonluminous objects. If a leaf, for instance, happened to be present at the site of the collision between your eye's stuff and the candle's stuff, then the leaf would be stimulated to express its green nature, allowing you to perceive it as green.

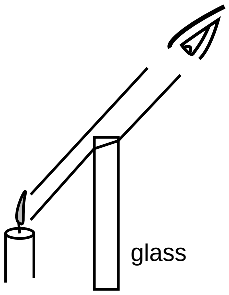
Modern people might feel uneasy about this theory, since it suggests that greenness exists only for our seeing convenience, implying a human precedence over natural phenomena. Nowadays, people would expect the cause and effect relationship in vision to be the other way around, with the leaf doing something to our eye rather than our eye doing something to the leaf. But how can you tell? The most common way of distinguishing cause from effect is to determine which happened first, but the process of seeing seems to occur too quickly to determine the order in which things happened. Certainly there is no obvious time lag between the moment when you move your head and the moment when your reflection in the mirror moves.

Today, photography provides the simplest experimental evidence that nothing has to be emitted from your eye and hit the leaf in order to make it "greenify." A camera can take a picture of a leaf even if there are no eyes anywhere nearby. Since the leaf appears green regardless of whether it is being sensed by a camera, your eye, or an insect's eye, it seems to make more sense to say that the leaf's greenness is the cause, and something happening in the camera or eye is the effect.

Light is a thing, and it travels from one point to another.

Another issue that few people have considered is whether a candle's flame simply affects your eye directly, or whether it sends out light which then gets into your eye. Again, the rapidity of the effect makes it difficult to tell what's happening. If someone throws a rock at you, you can see the rock on its way to your body, and you can tell that the person affected you by sending a material substance your way, rather than just harming you directly with an arm motion, which would be known as "action at a distance." It is not easy to do a similar observation to see whether there is some "stuff" that travels from the candle to your eye, or whether it is a case of action at a distance.

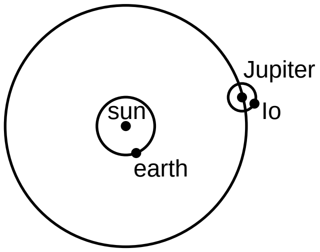
Newtonian physics includes both action at a distance (e.g. the earth's gravitational force on a falling object) and contact forces such as the normal force, which only allow distant objects to exert forces on each other by shooting some substance across the space between them (e.g. a garden hose spraying out water that exerts a force on a bush).



Light from a candle is bumped off course by a piece of glass. Inserting the glass causes the apparent location of the candle to shift. The same effect can be produced by taking off your eyeglasses and looking at what you see in the lens, but a flat piece of glass works just as well as a lens for this purpose.

One piece of evidence that the candle sends out stuff that travels to your eye is that intervening transparent substances can make the candle appear to be in the wrong location, suggesting that light is a thing that can be bumped off course. Many people would dismiss this kind of observation as an optical illusion, however. (Some optical illusions are purely neurological or psychological effects, although some others, including this one, turn out to be caused by the behavior of light itself.)

A more convincing way to decide in which category light belongs is to find out if it takes time to get from the candle to your eye; in Newtonian physics, action at a distance is supposed to be instantaneous. The fact that we speak casually today of "the speed of light" implies that at some point in history, somebody succeeded in showing that light did not travel infinitely fast. Galileo tried, and failed, to detect a finite speed for light, by arranging with a person in a distant tower to signal back and forth with lanterns. Galileo uncovered his lantern, and when the other person saw the light, he uncovered his lantern. Galileo was unable to measure any time lag that was significant compared to the limitations of human reflexes.



The first person to prove that light's speed was finite, and to determine it numerically, was Ole Roemer, in a series of measurements around the year 1675. Roemer observed Io, one of Jupiter's moons, over a period of several years. Since Io presumably took the same amount of time to complete each orbit of Jupiter, it could be thought of as a very distant, very accurate clock. A practical and accurate pendulum clock had recently been invented, so Roemer could check whether the ratio of the two clocks' cycles, about 42.5 hours to 1 orbit, stayed exactly constant or changed a little. If the process of seeing the distant moon was instantaneous, there would be no reason for the two to get out of step. Even if the speed of light was finite, you might expect that the result would be only to offset one cycle relative to the other. The earth does not, however, stay at a constant distance from Jupiter and its moons. Since the distance is changing gradually due to the two planets' orbit motions, a finite speed of light would make the "Io clock" appear to run faster as the planets drew near each other, and more slowly as their separation increased. Roemer did find a variation in the apparent speed of Io's orbits, which caused Io's eclipses by Jupiter (the moments when Io passed in front of or behind Jupiter) to occur about 7 minutes early when the earth was closest to Jupiter, and 7 minutes late when it was farthest. Based on these measurements, Roemer estimated the speed of light to be approximately $2\text{x}10^8$ m/s, which is in the right ballpark compared to modern measurements of $3\text{x}10^8$ m/s. (I'm not sure whether the fairly large experimental error was mainly due to imprecise knowledge of the radius of the earth's orbit or limitations in the reliability of pendulum clocks.)

Light can travel through a vacuum.

Many people are confused by the relationship between sound and light. Although we use different organs to sense them, there are some similarities. For instance, both light and sound are typically emitted in all directions by their sources. Musicians even use visual metaphors like "tone color," or "a bright timbre" to describe sound. One way to see that they are clearly different phenomena is to note their very different velocities. Sure, both are pretty fast compared to a flying arrow or a galloping horse, but as we have seen, the speed of light is so great as to appear instantaneous in most situations. The speed of sound, however, can easily be observed just by watching a group of schoolchildren a hundred feet away as they clap their hands to a song. There is an obvious delay between when you see their palms come together and when you hear the clap.

The fundamental distinction between sound and light is that sound is an oscillation in air pressure, so it requires air (or some other medium such as water) in which to travel. Today, we know that outer space is a vacuum, so the fact that we get light from the sun, moon and stars clearly shows that air is not necessary for the propagation of light. Also, a light bulb has a near vacuum inside, but that does not prevent the light from getting out. (The reason why the air is pumped out of light bulbs is to keep the oxygen from reacting violently with the hot filament and destroying it.)

Discussion Questions

A. If you observe thunder and lightning, you can tell how far away the storm is. Do you need to know the speed of sound, of light, or of both?

B. When phenomena like X-rays and cosmic rays were first discovered, suggest a way one could have tested whether they were forms of light.

C. Why did Roemer only need to know the radius of the earth's orbit, not Jupiter's, in order to find the speed of light?

1.2 Interaction of Light with Matter

Absorption of light

The reason why the sun feels warm on your skin is that the sunlight is being absorbed, and the light energy is being transformed into heat energy. The same happens with artificial light, so the net result of leaving a light turned on is to heat the room. It doesn't matter whether the source of the light is hot, like the sun, a flame, or an incandescent light bulb, or cool, like a fluorescent bulb. (If your house has electric heat, then there is absolutely no point in fastidiously turning off lights in the winter; the lights will help to heat the house at the same dollar rate as the electric heater.)

This process of heating by absorption is entirely different from heating by thermal conduction, as when an electric stove heats spaghetti sauce through a pan. Heat can only be conducted through matter, but there is vacuum between us and the sun, or between us and the filament of an incandescent bulb. Also, heat conduction can only transfer heat energy from a hotter object to a colder one, but a cool fluorescent bulb is perfectly capable of heating something that had already started out being warmer than the bulb itself.

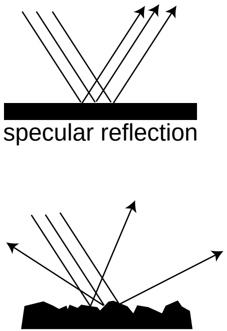
How we see nonluminous objects

Not all the light energy that hits an object is transformed into heat. Some is reflected, and this leads us to the question of how we see nonluminous objects. If you ask the average person how we see a light bulb, the most likely answer is "The light bulb makes light, which hits our eyes." But if you ask how we see a book, they are likely to say "The bulb lights up the room, and that lets me see the book." All mention of light actually entering our eyes has mysteriously disappeared.

Most people would disagree if you told them that light was reflected from the book to the eye, because they think of reflection as something that mirrors do, not something that a book does. They associate reflection with the formation of a reflected image, which does not seem to appear in a piece of paper.

Imagine that you are looking at your reflection in a nice smooth piece of aluminum foil, fresh off the roll. You perceive a face, not a piece of metal. Perhaps you also see the bright reflection of a lamp over your shoulder behind you. Now imagine that the foil is just a little bit less smooth. The different parts of the image are now a little bit out of alignment with each other. Your brain can still recognize a face and a lamp, but it's a little scrambled, like a Picasso painting. Now suppose you use a piece of aluminum foil that has been crumpled up and then flattened out again. The parts of the image are so scrambled that you cannot recognize an image. Instead, your brain tells you you're looking at a rough, silvery surface.

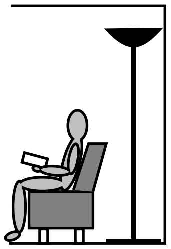
Mirrorlike reflection at a specific angle is known as specular reflection, and random reflection in many directions is called diffuse reflection. Diffuse reflection is how we see nonluminous objects. Specular reflection only allows us to see images of objects other than the one doing the reflecting. In top part of figure (a), imagine that the rays of light are coming from the sun. If you are looking down at the reflecting surface, there is no way for your eye-brain system to tell that the rays are not really coming from a sun down below you.



specular reflection

diffuse reflection

(a) Diffuse and specular reflection.



(b)

Figure (b) shows another example of how we can't avoid the conclusion that light bounces off of things other than mirrors. The lamp is one I have in my house. It has a bright bulb, housed in a completely opaque bowl-shaped metal shade. The only way light can get out of the lamp is by going up out of the top of the bowl. The fact that I can read a book in the position shown in the figure means that light must be bouncing off of the ceiling, then bouncing off of the book, then finally getting to my eye.

This is where the shortcomings of the Greek theory of vision become glaringly obvious. In the Greek theory, the light from the bulb and my mysterious "eye rays" are both supposed to go to the book, where they collide, allowing me to see the book. But we now have a total of four objects: lamp, eye, book, and ceiling. Where does the ceiling come in? Does it also send out its own mysterious "ceiling rays," contributing to a three-way collision at the book? That would just be too bizarre to believe!

The differences among white, black, and the various shades of grey in between is a matter of how what percentage of the light they absorb and what percentage they reflect. That's why light-colored clothing is more comfortable in the summer, and light-colored upholstery in a car stays cooler that dark upholstery.

Numerical measurement of the brightness of light

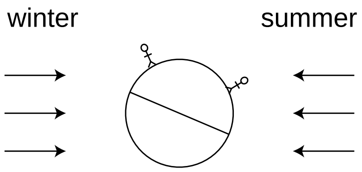
We have already seen that the physiological sensation of loudness relates to the sound's intensity (power per unit area), but is not directly proportional to it. If sound A has an intensity of 1 nW/m^2, sound B is 10 nW/m^2, and sound C is 100 nW/m^2, then the increase in loudness from C to B is perceived to be the same as the increase from A to B, not ten times greater. That is, the sensation of loudness if logarithmic.

The same is true for the brightness of light. Brightness is related to power per unit area, but the relationship is a logarithmic one rather than a proportionality. For doing physics, it's the power per unit area that we're interested in. The relevant SI unit is W/m^2, although various non-SI units are still used in photography, for example. One way to determine the brightness of light is to measure the increase in temperature of a black object exposed to the light. The light energy is being converted to heat energy, and the amount of heat energy absorbed in a given amount of time can be related to the power absorbed, using the known heat capacity of the object. More practical devices for measuring light intensity, such as the light meters built into some cameras, are based on the conversion of light into electrical energy, but these meters have to be calibrated somehow against heat measurements.

Discussion questions

A. The curtains in a room are drawn, but a small gap lets light through, illuminating a spot on the floor. It may or may not also be possible to see the beam of sunshine crossing the room, depending on the conditions. What's going on?

B. Laser beams are made of light. In science fiction movies, laser beams are often shown as bright lines shooting out of a laser gun on a spaceship. Why is this scientifically incorrect?



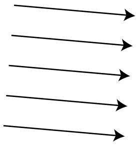
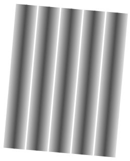
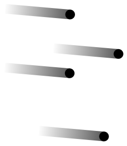
Discussion question C.

C. A documentary filmmaker went to Harvard's 1987 graduation ceremony and asked the graduates, on camera, to explain the cause of the seasons. Only two out of 23 were able to give a correct explanation, but you now have all the information needed to figure it out for yourself, assuming you didn't already know. The figure shows the earth in its winter and summer positions relative to the sun. Hint: Consider the units used to measure the brightness of light, and recall that the sun is lower in the sky in winter, so its rays are coming in at a shallower angle.

1.3 The Ray Model of Light

Models of light

Note how I've been casually diagramming the motion of light with pictures showing light rays as lines on the page. More formally, this is known as the ray model of light. The ray model of light seems natural once we convince ourselves that light travels through space, and observe phenomena like sunbeams coming through holes in clouds. Having already been introduced to the concept of light as an electromagnetic wave, you know that the ray model is not the ultimate truth about light, but the ray model is simpler, and in any case science always deals with models of reality, not the ultimate nature of reality. The following table summarizes three models of light.

ray model		*Advantage:* Simplicity.
wave model		*Advantage:* Color is described naturally in terms of wavelength. *Required* in order to explain the interaction of light with material objects with sizes comparable to a wavelength of light or smaller.
particle model		*Required* in order to explain the interaction of light with individual atoms. At the atomic level, it becomes apparent that a beam of light has a certain graininess to it.

The ray model is essentially a generic one. By using it we can discuss the path taken by the light, without committing ourselves to any specific description of what it is that is moving along that path. We will use the nice simple ray model for most of this book, and with it we can analyze a great many devices and phenomena. Not until the last chapter will we concern ourselves specifically with wave optics, although in the intervening chapters I will sometimes analyze the same phenomenon using both the ray model and the wave model.

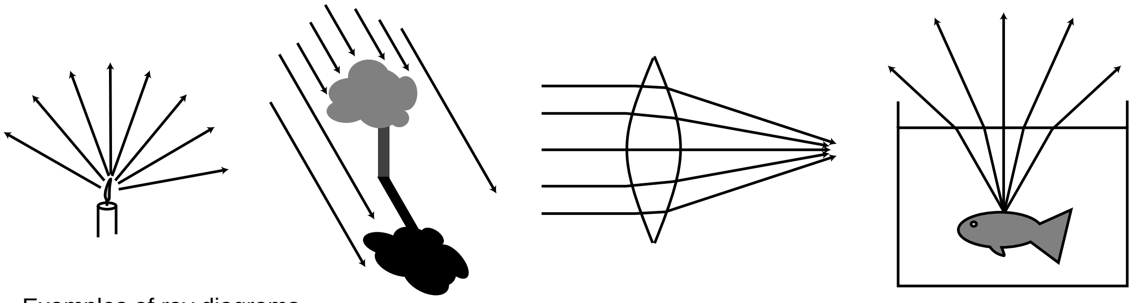
Note that the statements about the applicability of the various models are only rough guides. For instance, wave interference effects are often detectable, if small, when light passes around an obstacle that is quite a bit bigger than a wavelength. Also, the criterion for when we need the particle model really has more to do with energy scales than distance scales, although the two turn out to be related.

The alert reader may have noticed that the wave model is required at scales smaller than a wavelength of light (on the order of a micrometer for visible light), and the particle model is demanded on the atomic scale or lower (a typical atom being a nanometer or so in size). This implies that at the smallest scales we need *both* the wave model and the particle model. They appear incompatible, so how can we simultaneously use both? The answer is that they are not as incompatible as they seem. Light is both a wave and a particle, but a full understanding of this apparently nonsensical statement is a topic for the following book in this series.

Ray diagrams

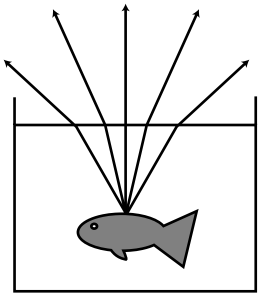
Without even knowing how to use the ray model to calculate anything numerically, we can learn a great deal by drawing ray diagrams. For instance, if you want to understand how eyeglasses help you to see in focus, a ray diagram is the right place to start. Many students underutilize ray diagrams in optics and instead rely on rote memorization or plugging into formulas. The trouble with memorization and plug-ins is that they can obscure what's really going on, and it is easy to get them wrong. Often the best plan is to do a ray diagram first, then do a numerical calculation, then check that your numerical results are in reasonable agreement with what you expected from the ray diagram.

Examples (a) through (c) show some guidelines for using ray diagrams effectively. The light rays bend when then pass out through the surface of the water (a phenomenon that we'll discuss in more detail later). The rays appear to have come from a point above the goldfish's actual location, an

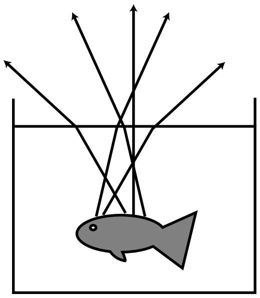


Examples of ray diagrams.

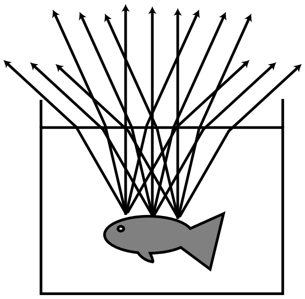
effect that is familiar to people who have tried spearfishing.



(a) correct



(b) Incorrect: implies that diffuse reflection only gives one ray from each reflecting point.



(c) Correct, but unnecessarily complicated.

- A stream of light is not really confined to a finite number of narrow lines. We just draw it that way. In (a), it has been necessary to choose a finite number of rays to draw (five), rather than the theoretically infinite number of rays that will diverge from that point.
- There is a tendency to conceptualize rays incorrectly as objects. In his Optics, Newton goes out of his way to caution the reader against this, saying that some people "consider ... the refraction of ... rays to be the bending or breaking of them in their passing out of one medium into another." But a ray is a record of the path traveled by light, not a physical thing that can be bent or broken.
- In theory, rays may continue infinitely far into the past and future, but we need to draw lines of finite length. In (a), a judicious choice has been made as to where to begin and end the rays. There is no point in continuing the rays any farther than shown, because nothing new and exciting is going to happen to them. There is also no good reason to start them earlier, before being reflected by the fish, because the direction of the diffusely reflected rays is random anyway, and unrelated to the direction of the original, incoming ray.
- When representing diffuse reflection in a ray diagram, many students have a mental block against drawing many rays fanning out from the same point. Often, as in example (b), the problem is the misconception that light can only be reflected in one direction from one point.
- Another difficulty associated with diffuse reflection, example (c), is the tendency to think that in addition to drawing many rays coming out of one point, we should also be drawing many rays coming from many points. In (a), drawing many rays coming out of one point gives useful information, telling us, for instance, that the fish can be seen from any angle. Drawing many sets of rays, as in (c), does not give us any more useful information, and just clutters up the picture in this example. The only reason to draw sets of rays fanning out from more than one point would be if different things were happening to the different sets.

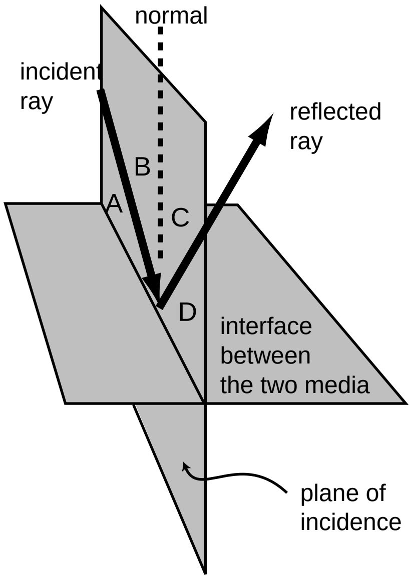
Discussion Question

Suppose an intelligent tool-using fish is spearhunting for humans. Draw a ray diagram to show how the fish has to correct its aim. Note that although the rays are now passing from the air to the water, the same rules apply: the rays are closer to being perpendicular to the surface when they are in the water, and rays that hit the air-water interface at a shallow angle are bent the most.

1.4 Geometry of Specular Reflection

To change the motion of a material object, we use a force. Is there any way to exert a force on a beam of light? Experiments show that electric and magnetic fields do not deflect light beams, so apparently light has no electric charge. Light also has no mass, so until the twentieth century it was believed to be immune to gravity as well. Einstein predicted that light beams would be very slightly deflected by strong gravitational fields, and he was proven correct by observations of rays of starlight that came close to the sun, but obviously that's not what makes mirrors and lenses work!

If we investigate how light is reflected by a mirror, we will find that the process is horrifically complex, but the final result is surprisingly simple. What actually happens is that the light is made of electric and magnetic fields, and these fields accelerate the electrons in the mirror. Energy from the light beam is momentarily transformed into extra kinetic energy of the electrons, but because the electrons are accelerating the reradiate more light, converting their kinetic energy back into light energy. We might expect this to result in a very chaotic situation, but amazingly enough, the electrons move together to produce a new, reflected beam of light, which obeys two simple rules:



- The angle of the reflected ray is the same as that of the incident ray.
- The reflected ray lies in the plane containing the incident ray and the normal (perpendicular) line. This plane is known as the plane of incidence.

The two angles can be defined either with respect to the normal, like angles B and C in the figure, or with respect to the reflecting surface, like angles A and D. There is a convention of several hundred years' standing that one measures the angles with respect to the normal, but the rule about equal angles can logically be stated either as B=C or as A=D.

The phenomenon of reflection occurs only at the boundary between two media, just like the change in the speed of light that passes from one medium to another.

Most people are surprised by the fact that light can be reflected back into a less dense medium. For instance, if you are diving and you look up at the surface of the water, you will see a reflection of yourself.

Reversibility of light rays

The fact that specular reflection displays equal angles of incidence and reflection means that there is a symmetry: if the ray had come in from the right instead of the left in the figure above, the angles would have looked exactly the same. This is not just a pointless detail about specular reflection. It's a manifestation of a very deep and important fact about nature, which is that the laws of physics do not distinguish between past and future. Cannonballs and planets have trajectories that are equally natural in reverse, and so do light rays. This type of symmetry is called time-reversal symmetry.

Typically, time-reversal symmetry is a characteristic of any process that does not involve heat. For instance, the planets do not experience any friction as they travel through empty space, so there is no frictional heating. We should thus expect the time-reversed versions of their orbits to obey the laws of physics, which they do. In contrast, a book sliding across a table does generate heat from friction as it slows down, and it is therefore not surprising that this type of motion does not appear to obey time-reversal symmetry. A book lying still on a flat table is never observed to spontaneously start sliding, sucking up heat energy and transforming it into kinetic energy.

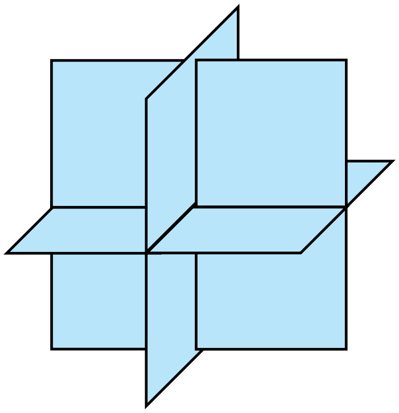
Similarly, the only situation we've observed so far where light does not obey time-reversal symmetry is absorption, which involves heat. Your skin absorbs visible light from the sun and heats up, but we never observe people's skin to glow, converting heat energy into visible light. People's skin does glow in infrared light, but that doesn't mean the situation is symmetric. You never absorb infrared and emit visible light.

Discussion Questions

A. If a light ray has a velocity vector with components c_x and c_y, what will happen when it is reflected from a surface in the *y-z* plane? Make sure your answer does not imply a change in the ray's speed.

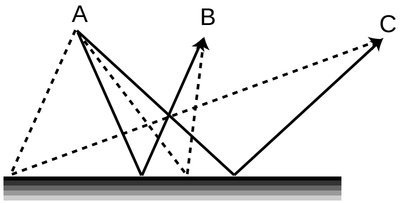
B. Radar works by bouncing radio waves off of objects and inferring the distance to the object from the time delay of the reflected "echo." As we've discussed in book 4, radio waves are really a form of light whose "color" is just way off the red end of the visible rainbow. The stealth bomber is flat on the bottom. How does this make the plane harder to detect with radar?

C. Three pieces of sheet metal arranged perpendicularly as shown in the figure form what is known as a radar corner. Explain why a radio ray that strikes it from any angle is guaranteed to go straight back and be detected. You may want to start by considering the case where the ray comes in parallel to one of the planes, in which case the problem can be treated two-dimensionally. For the full three-dimensional case, it may be helpful to consider the three components of the radio ray's velocity vector.

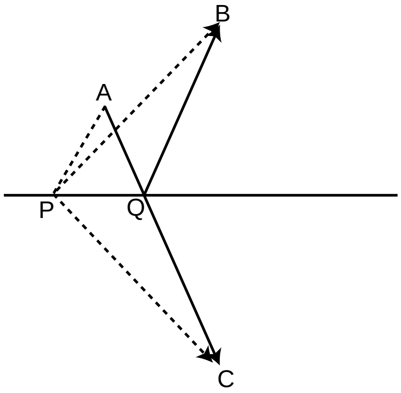


Discussion question C.

1.5* The Principle of Least Time for Reflection



The solid lines are physically possible paths for ight rays traveling from A to B and from A to C. They obey the principle of least time. The dashed lines do not obey the principle of least time, and are not physically possible.



Paths AQB and APB are two conceivable paths that a ray could travel to get from A to B with one reflection, but only AQB is physically possible. We wish to prove that path AQB, with equal angles of incidence and reflection, is shorter than any other path, such as APB. The trick is to contruct a third point, C, lying as far below the surface as B lies above it. Then path AQC is a straight line whose length is the same as AQB, and path APC has the same length as path APB. Since AQC is straight, it must be shorter than any other path such as APC that connects A and C, and therefore AQB must be shorter than any path such as APB.

We had to choose between an unwieldy explanation of reflection at the atomic level and a simpler geometric description that was not as fundamental. There is a third approach to describing the interaction of light and matter which is very deep and beautiful. Emphasized by the twentieth-century physicist Richard Feynman, it is called the principle of least time.

Let's start with the motion of light that is not interacting with matter at all. In a vacuum, a light ray moves in a straight line. This can be rephrased as follows: a light ray in vacuum that moves from point A to point B follows the path that takes the shortest time.

What about reflection? If light is going to go from one point to another, being reflected on the way, the quickest path is indeed the one with equal angles of incidence and reflection. If the starting and ending points are equally far from the reflecting surface, as shown in the top panel of the figure, it's not hard to convince yourself that this is true, just based on symmetry. There is also a tricky and simple proof, shown in the bottom panel of the figure, for the more general case where the points are at different distances from the surface.

Not only does the principle of least time work for light in a vacuum and light undergoing reflection, we will also see in a later chapter that it works for the bending of light when it passes from one medium into another.

Although it is beautiful that the laws concerning the motion of light can be reduced to one simple rule, the principle of least time, it may seem a little spooky to speak as if the ray of light is intelligent, and has carefully planned ahead to find the shortest route to its destination. How does it know in advance where it's going? What if we moved the mirror while the light was en route, so conditions along its planned path were not what it "expected"? The answer is that the principle of least time is really a shortcut for finding certain results of the wave model of light, which is the topic of the last chapter of this book.

Summary

Selected Vocabulary

absorption .... what happens when light hits matter and gives up some of its energy

reflection .... what happens when light hits matter and bounces off, retaining at least some of its energy

specular reflection............. reflection from a smooth surface, in which the light ray leaves at the same angle at which it came in

diffuse reflection .... reflection from a rough surface, in which a single ray of light is divided up into many weaker reflected rays going in many directions

Notation

c .... the speed of light

Summary

We can understand many phenomena involving light without having to use sophisticated models such as the wave model or the particle model. Instead, we simply describe light according to the path it takes, which we call a ray. The ray model of light is useful when light is interacting with material objects that are much larger than a wavelength of light. Since a wavelength of visible light is so short compared to the human scale of existence, the ray model is useful in many practical cases.

We see things because light comes from them to our eyes. Objects that glow may send light directly to our eyes, but we see an object that doesn't glow via light from another source that has been reflected by the object.

Many of the interactions of light and matter can be understood by considering what happens when light reaches the boundary between two different substances. In this situation, part of the light is reflected (bounces back) and part passes on into the new medium. This is not surprising — it is typical behavior for a wave, and light is a wave. Light energy can also be absorbed by matter, i.e. converted into heat.

A smooth surface produces specular reflection, in which the reflected ray exits at the same angle with respect to the normal as that of the incoming ray. A rough surface gives diffuse reflection, where a single ray of light is divided up into many weaker reflected rays going in many directions.

Homework Problems

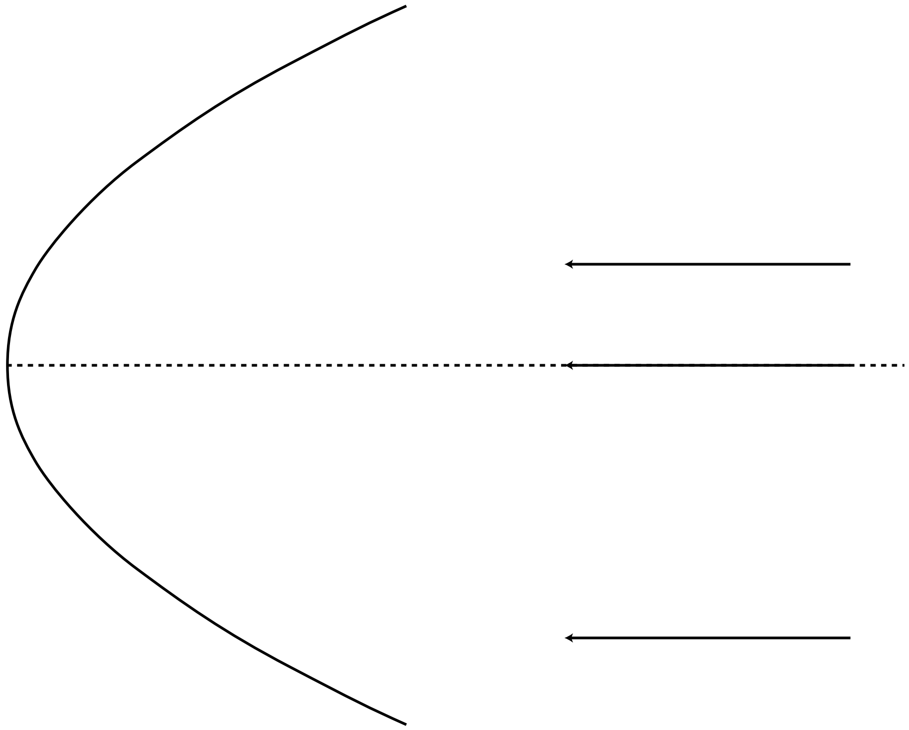
1. Draw a ray diagram showing why a small light source (a candle, say) produces sharper shadows than a large one (e.g. a long fluorescent bulb).

2. A Global Positioning System (GPS) receiver is a device that lets you figure out where you are by exchanging radio signals with satellites. It works by measuring the round-trip time for the signals, which is related to the distance between you and the satellite. By finding the ranges to several different satellites in this way, it can pin down your location in three dimensions to within a few meters. How accurate does the measurement of the time delay have to be to determine your position to this accuracy?

3. Estimate the frequency of an electromagnetic wave whose wavelength is similar in size to an atom (about a nm). Referring back to book 4, in what part of the electromagnetic spectrum would such a wave lie (infrared, gamma-rays,...)?

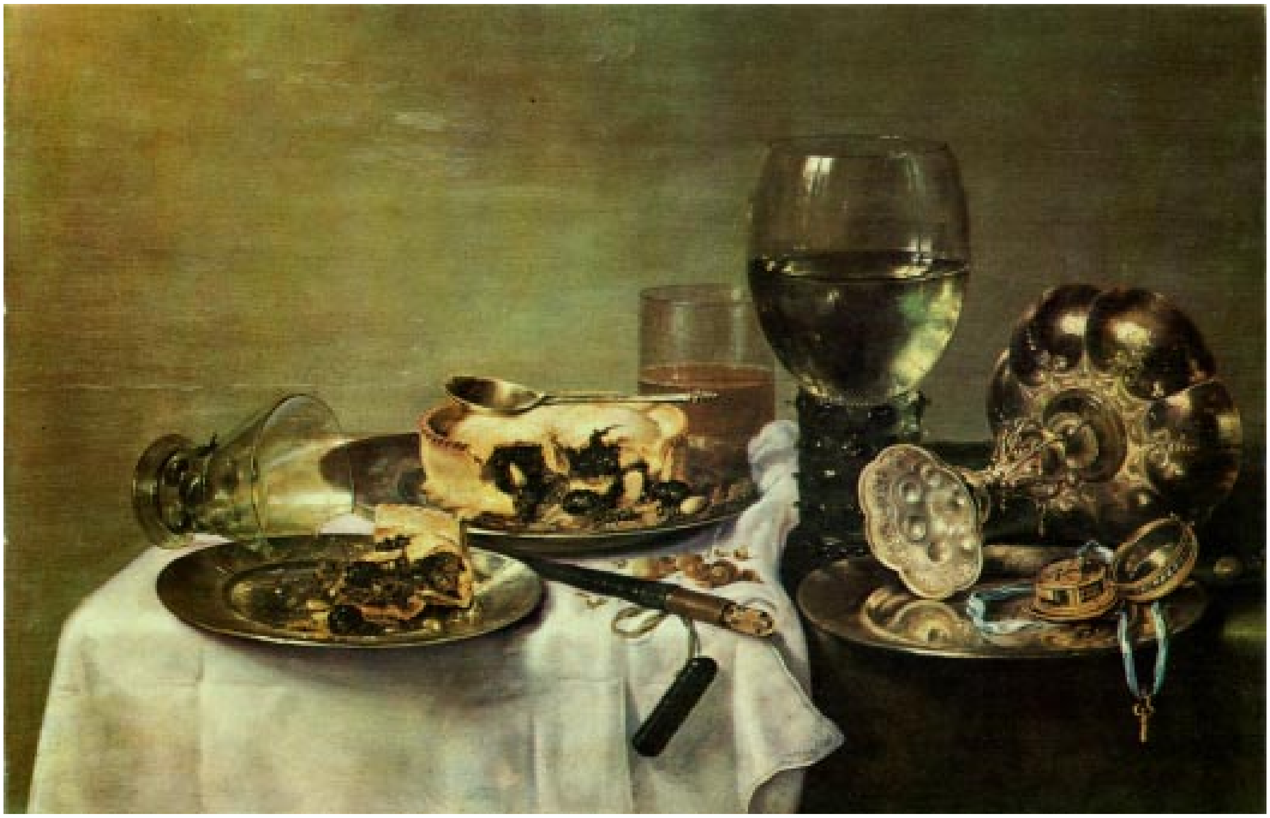
4. The Stealth bomber is designed with flat, smooth surfaces. Why would this make it difficult to detect via radar?

5. The large figure shows a curved (parabolic) mirror, with three parallel light rays coming toward it. One ray is approaching along the mirror's center line. (a) Trace the drawing accurately, and continue the light rays until they exit the diagram again on the right. What do you notice? (b) Make up an example of a practical use for this device. (c) How could you use this mirror with a small lightbulb to produce a parallel beam of light rays going off to the right?



Problem 5.

S A solution is given in the back of the book.
✓ A computerized answer check is available.
★ A difficult problem.
∫ A problem that requires calculus.



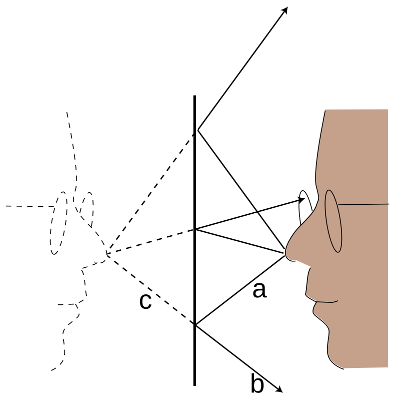
Breakfast Table, by Willem Claesz. de Heda, 17th century. A variety of images occur in the painting, some distorted, as a result of both reflection and refraction (ch. 4).

2 Images by Reflection, Part I

Infants are always fascinated by the antics of the Baby in the Mirror. Now if you want to know something about mirror images that most people don't understand, try this. First bring this page and closer to your eyes, until you can no longer focus on it without straining. Then go in the bathroom and see how close you can get your face to the surface of the mirror before you can no longer easily focus on the image of your own eyes. You will find that the shortest comfortable eye-mirror distance is much less than the shortest comfortable eye-paper distance. This demonstrates that the image of your face in the mirror acts as if it had depth and existed in the space *behind* the mirror. If the image was like a flat picture in a book, then you wouldn't be able to focus on it from such a short distance.

In this chapter we will study the images formed by flat and curved mirrors on a qualitative, conceptual basis. Although this type of image is not as commonly encountered in everyday life as images formed by lenses, images formed by reflection are simpler to understand, so we discuss them first. In chapter 3 we will turn to a more mathematical treatment of images made by reflection. Surprisingly, the same equations can also be applied to lenses, which are the topic of chapter 4.

2.1 A Virtual Image



(a) An image formed by a mirror.

We can understand a mirror image using a ray diagram. The figure shows several light rays, a, that originated by diffuse reflection at the person's nose. They bounce off the mirror, producing new rays, b. To anyone whose eye is in the right position to get one of these rays, they appear to have come from a behind the mirror, c, where they would have originated from a single point. This point is where the tip of the image-person's nose appears to be. A similar analysis applies to every other point on the person's face, so it looks as though there was an entire face behind the mirror. The customary way of describing the situation requires some explanation:

> **Customary description in physics**: There is an image of the face behind the mirror.
>
> **Translation**: The pattern of rays coming from the mirror is exactly the same as it would be if there was a face behind the mirror. Nothing is really behind the mirror.

This is referred to as a *virtual* image, because the rays do not actually cross at the point behind the mirror. They only appear to have originated there.

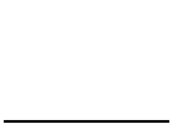
Self-Check

Imagine that the person in figure (a) moves his face down quite a bit — a couple of feet in real life, or a few inches on this scale drawing. Draw a new ray diagram. Will there still be an image? If so, where is it visible from?

The geometry of specular reflection tells us that rays a and b are at equal angles to the normal (the imaginary perpendicular line piercing the mirror at the point of reflection). This means that ray b's imaginary continuation, c, forms the same angle with the mirror as ray a. Since each ray of type c forms the same angles with the mirror as its partner of type a, we see that the distance of the image from the mirror is the same as the actual face from the mirror, and lies directly across from it. The image therefore appears to be the same size as the actual face.

Discussion Question

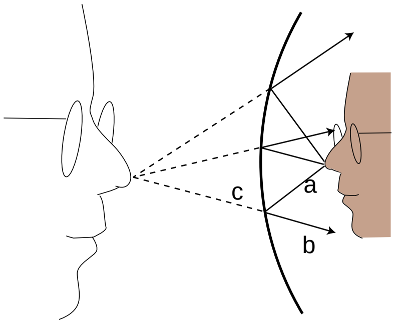
The figure shows an object that is off to one side of a mirror. Draw a ray diagram. Is an image formed? If so, where is it, and from which directions would it be visible?



You should have found from your ray diagram that an image is still formed, and it has simply moved down the same distance as the real face. However, this new image would only be visible from high up, and the person can no longer see his own image. If you couldn't draw a ray diagram that seemed to result in an image, the problem was probably that you didn't choose any rays that happened to go away from the face in the right direction to hit the mirror.

2.2 Curved Mirrors

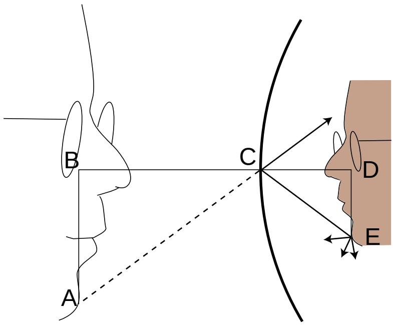
An image in a flat mirror is essentially a pretechnological example: even animals can look at their reflections in a calm pond. We now pass to our first nontrivial example of the manipulation of an image by technology: an image in a curved mirror. Before we dive in, let's consider why this is an important example. If it was just a question of memorizing a bunch of facts about curved mirrors, then you would rightly rebel against an effort to spoil the beauty of your liberally educated brain by force-feeding you technological trivia. The reason this is an important example is not that curved mirrors are so important in and of themselves, but that the results we derive for curved bowl-shaped mirrors turn out to be true for a large class of other optical devices, including mirrors that bulge outward rather than inward, and lenses as well. A complex optical device like a microscope or a telescope is simply a combination of lenses or mirrors or both. What you're really learning about here is the basic building block of all optical devices from movie projectors to octopus eyes.



(b) An image formed by a curved mirror.

Because the mirror in figure (b) is curved, it tends to bend the rays back closer together than a flat mirror would. It is an "inbending" mirror. (You may hear this referred to by other people as a concave mirror. The term is not all that hard to remember, because the hollowed-out interior of the mirror is like a cave.) It is surprising but true that all the rays like c really do converge on a point, forming a good image. We will not prove this fact, but it is true for any mirror whose curvature is gentle enough and that is symmetric with respect to rotation about the perpendicular line passing through its center. (The old-fashioned method of making mirrors and lenses is by grinding them in grit by hand, and this process automatically tends to produce an almost perfect spherical surface.)

Bending a ray like b inward implies bending its imaginary continuation c outward, in the same way that raising one end of a seesaw causes the other end to go down. The image therefore forms deeper behind the mirror. This doesn't just show that there is extra distance between the image-nose and the mirror; it also implies that the image itself is bigger from front to back. It has been *magnified* in the front-to-back direction.



(c) The image is magnified by the same factor in depth and in its other dimensions.

It is easy to prove that the same magnification also applies to the image's other dimensions. Consider a point like E in figure (c). The trick is that out of all the rays diffusely reflected by E, we pick the one that happens to head for the mirror's center, C. The equal-angle property of specular reflection plus a little straightforward geometry easily leads us to the conclusion that triangles ABC and CDE are the same shape, with ABC being simply a scaled-up version of CDE. The magnification of depth equals the ratio BC/CD, and the up-down magnification is AB/DE. A repetition of the same proof also shows that the magnification in the third dimension is also the same. This means that the image-head is simply a larger version of the real one, without any distortion. The scaling factor is called the magnification, M. The image in the figure is magnified by a factor M=1.9.

Note that we did explicitly specify whether the mirror was a sphere, a paraboloid, or some other shape. However, we assumed that a focused image would be formed, which would not necessarily be true, for instance, for a mirror that was asymmetric or very deeply curved.

2.3 A Real Image

If we start by placing an object very close to the mirror, (d), and then move it farther and farther away, the image at first behaves as we would expect from our everyday experience with flat mirrors, receding deeper and deeper behind the mirror. At a certain point, however, a dramatic change occurs. When the object is more than a certain distance from the mirror, (e), the image appears upside-down and in *front* of the mirror.

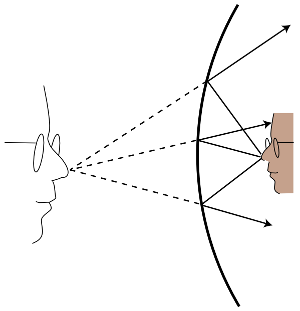
Here's what's happened. The mirror bends light rays inward, but when the object is very close to it, as in (d), the rays coming from a given point on the object are too strongly diverging (spreading) for the mirror to bring them back together. On reflection, the rays are still diverging, just not as strongly diverging. But when the object is sufficiently far away, (e), the mirror is only intercepting the rays that came out in a narrow cone, and it is able to bend these enough so that they will reconverge.

Note that the rays shown in the figure, which both originated at the same point on the object, reunite when they cross. The point where they cross is the image of the point on the original object. This type of image is called a *real image*, in contradistinction to the virtual images we've studied before. The use of the word "real" is perhaps unfortunate. It sounds as though we are saying the image was an actual material object, which of course it is not.

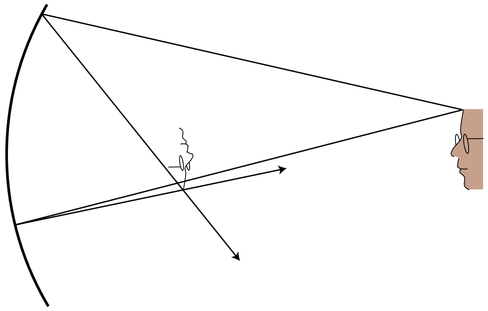
The distinction between a real image and a virtual image is an important one, because a real image can projected onto a screen or photographic film. If a piece of paper is inserted in figure (e) at the location of the image, the image will be visible to someone looking at the paper from the left. Your eye uses a lens to make a real image on the retina.

Self-Check

Sketch another copy of the face in figure (e), even farther from the mirror, and draw a ray diagram. What has happened to the location of the image?



(d) A virtual image.

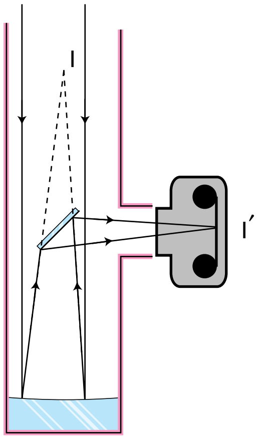


(e) A real image.

Increasing the distance from the face to the mirror has decreased the distance from the image to the mirror. This is the opposite of what happened with the virtual image.

2.4 Images of Images

If you are wearing glasses right now, then the light rays from the page are being manipulated first by your glasses and then by the lens of your eye. You might think that it would be extremely difficult to analyze this, but in fact it is quite easy. In any series of optical elements (mirrors or lenses or both), each element works on the rays furnished by the previous element in exactly the same manner as if the image formed by the previous element was an actual object.

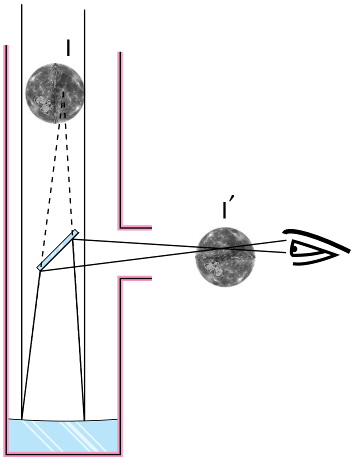


(f) A Newtonian telescope being used with a camera.

Figure (f) shows an example involving only mirrors. The Newtonian telescope, invented by Isaac Newton, consists of a large curved mirror, plus a second, flat mirror that brings the light out of the tube. (In very large telescopes, there may be enough room to put a camera or even a person inside the tube, in which case the second mirror is not needed.) The tube of the telescope is not vital; it is mainly a structural element, although it can also be helpful for blocking out stray light. The lens has been removed from the front of the camera body, and is not needed for this setup. Note that the two sample rays have been drawn parallel, because an astronomical telescope is used for viewing objects that are extremely far away. These two "parallel" lines actually meet at a certain point, say a crater on the moon, so they can't actually be perfectly parallel, but they are parallel for all practical purposes since we would have to follow them upward for a quarter of a million miles to get to the point where they intersect.

The large curved mirror by itself would form an image I, but the small flat mirror creates an image of the image, I′. The relationship between I and I′ is exactly the same as it would be if I was an actual object rather than an image: I and I′ are at equal distances from the plane of the mirror, and the line between them is perpendicular to the plane of the mirror.

One surprising wrinkle is that whereas a flat mirror used by itself forms a virtual image of an object that is real, here the mirror is forming a real image of virtual image I. This shows how pointless it would be to try to memorize lists of facts about what kinds of images are formed by various optical elements under various circumstances. You are better off simply drawing a ray diagram.



(g) A Newtonian telescope being used for visual rather than photographic observing. In real life, an eyepiece lens is normally used for additional magnification, but this simpler setup will also work.

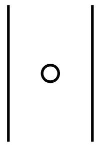
Although the main point here was to give an example of an image of an image, this is also an interesting case where we need to make the distinction between *magnification* and *angular magnification*. If you are looking at the moon through this telescope, then the images I and I′ are much *smaller* than the actual moon. Otherwise they would not fit inside the telescope or the camera! However, these images are very close to your eye compared to the actual moon. The small size of the image has been more than compensated for by the shorter distance. The important thing here is the amount of *angle* within your field of view that the image covers, and it is this angle that has been increased. The factor by which it is increased is called the *angular magnification*, M_a.

Discussion Questions

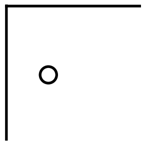
A. Describe the images that will be formed of you if you stand between two parallel mirrors.

B. Locate the images formed by two perpendicular mirrors, as in the figure. What happens if the mirrors are not perfectly perpendicular?

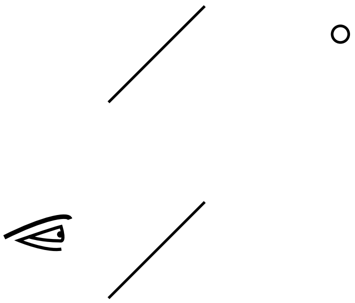
C. Locate the images formed by the periscope.



Discussion question A.



Discussion question B.



Discussion question C.

Summary

Selected Vocabulary

real image .... a place where an object appears to be, because the rays diffusely reflected from any given point on the object have been bent so that they come back together and then spread out again from the new point

virtual image .... like a real image, but the rays don't actually cross again; they only appear to have come from the point on the image

magnification .... the factor by which an image's linear size is increased (or decreased)

angular magnification .... the factor by which an image's apparent angular size is increased (or decreased)

Vocabulary Used in Other Books

concave .... describes a surface that is hollowed out like a cave

convex .... describes a surface that bulges outward

Notation

M .... the magnification of an image

M_a .... the angular magnification of an image

Summary

A large class of optical devices, including lenses and flat and curved mirrors, operates by bending light rays to form an image. A real image is one for which the rays actually cross at each point of the image. A virtual image, such as the one formed behind a flat mirror, is one for which the rays only appear to have crossed at a point on the image. A real image can be projected onto a screen; a virtual one cannot.

Mirrors and lenses will generally make an image that is either smaller than or larger than the original object. The scaling factor is called the magnification. In many situations, the angular magnification is more important than the actual magnification.

Homework Problems

1 ✓. A man is walking at 1.0 m/s directly towards a flat mirror. At what speed is his separation from his image reducing?

2. If a mirror on a wall is only big enough for you to see yourself from your head down to your waist, can you see your entire body by backing up? Test this experimentally and come up with an explanation for your observations. Note that it is easy to confuse yourself if the mirror is even a tiny bit off of vertical; check whether you are able to see more of yourself both above *and* below.

3. In this chapter we've only done ray examples of mirrors with hollowed-out shapes (called concave mirrors). Now draw a ray diagram for a curved mirror that has a bulging outward shape (called a convex mirror). (a) How does the image's distance from the mirror compare with the actual object's distance from the mirror? From this comparison, determine whether the magnification is greater than or less than one. (b) Is the image real or virtual? Could this mirror ever make the other type of image?

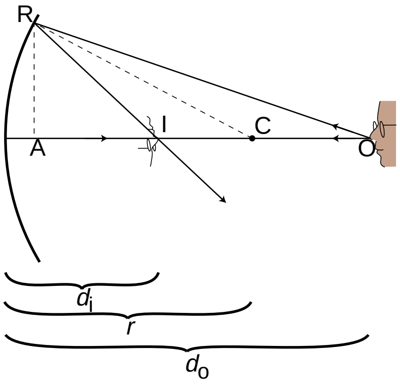
4. As discussed in question 3, there are two types of curved mirrors, concave and convex. Make a list of all the possible combinations of types of images (virtual or real) with types of mirrors (concave and convex). (Not all of the four combinations are physically possible.) Now for each one, use ray diagrams to determine whether increasing the distance of the object from the mirror leads to an increase or a decrease in the distance of the image from the mirror.

5. If the user of an astronomical telescope moves her head closer to or farther away from the image she is looking at, does the magnification change? Does the angular magnification change? Explain.

3 Images by Reflection, Part II

It sounds a bit odd when a scientist refers to a theory as “beautiful,” but to those in the know it makes perfect sense. One mark of a beautiful theory is that it surprises us by being simple. The mathematical theory of lenses and curved mirrors gives us just such a surprise. We expect the subject to be complex because there are so many cases: an inbending mirror forming a real image, an outbending lens that makes a virtual image, and so on for a total of six possibilities. If we want to predict the location of the images in all these situations, we might expect to need six different equations, and six more for predicting magnifications. Instead, it turns out that we can use just one equation for the location of the image and one equation for its magnification, and these two equations work in all the different cases with no changes except for plus and minus signs. This is the kind of thing the physicist Eugene Wigner referred to as “the unreasonable effectiveness of mathematics.” Sometimes we can find a deeper reason for this kind of unexpected simplicity, but sometimes it almost seems as if God went out of Her way to make the secrets of universe susceptible to attack by the human thought-tool called math.

3.1 A Real Image Formed by an In-Bending Mirror



Location of the image

We will now derive the equation for the location of a real image formed by an inbending mirror. We assume for simplicity that the mirror is spherical, but actually this isn't a restrictive assumption, because any shallow, symmetric curve can be approximated by a sphere. Given the radius, *r*, of the mirror, and the distance d_o from the mirror to the object, we wish to find the distance d_i from the mirror to the image. Note that although this is only one possible case of image formation, you will not have to plod through more such derivations. The equation will work in the other cases with a simple modification of plus and minus signs.

Since we have proven in the previous chapter that this type of image is not distorted, we can use an on-axis point, O, on the object. The equation we derive to relate d_o to d_i will also hold for off-axis points, since otherwise the image would have to be distorted, which we know is not true.

To locate the image we draw two rays emitted by the object and find where they cross. Ray ORI is chosen at random, but for the other ray we choose the special one that heads straight in along the axis. Point C is the center of the complete sphere we would get by extending the mirror's surface. We know that specular reflection gives equal angles with respect to the normal for the incoming and outgoing rays, so angle ACR must be equal to the average of angles AIR and AOR,

$$\measuredangle ACR = \frac{\measuredangle AIR + \measuredangle AOR}{2} \quad . \tag{1}$$

We want to relate this somehow to the distances d_i and d_o. For legibility I have drawn the figure with a very sharply curved mirror and the object and image quite close to it, but we're really assuming a mirror with a shallow curve, in which case the distance from point A to the mirror becomes negligible, and we have

$$r \approx AC, \quad d_i \approx AI, \quad \text{and} \quad d_o \approx AO \quad . \tag{2}$$

We can relate these to the angles via trigonometry:

$$\frac{RA}{AC} = \tan \measuredangle ACR\ , \quad \frac{RA}{d_i} \approx \tan \measuredangle AIR\ , \quad \text{and} \quad \frac{RA}{d_o} \approx \tan \measuredangle AOR \quad .(3)$$

We can simplify this by making use again of the assumption that the mirror is very shallow, and the triangles therefore very long and skinny. Under these circumstances, the small-angle approximation $\tan\theta \approx \theta$ applies (for θ measured in radians). Using this approximation on equations (3) and then substituting them into equation (1) yields

$$\frac{RA}{r} = \frac{1}{2}\left(\frac{RA}{d_i} + \frac{RA}{d_o}\right) \quad ,$$

which is readily simplified to

$$\frac{2}{r} = \frac{1}{d_i} + \frac{1}{d_o} \quad .$$

The quantity *r/2* is referred to as the *focal length*, *f*, of the mirror, so the equation is usually rewritten as

$$\frac{1}{f} = \frac{1}{d_i} + \frac{1}{d_o} \quad .$$

Self-Check

(1) If we work with a certain mirror, so that *f* is fixed, what happens to d_i if we increase d_o? If we decrease it?

Example: Solving for d_i

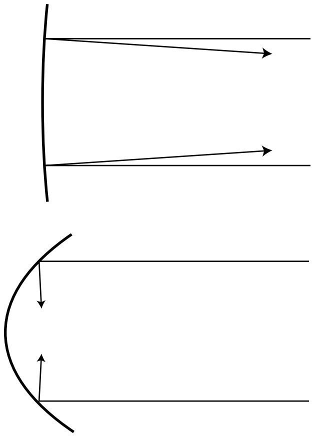
We encounter many cases in which *f* and d_o are known, and we wish to predict the location of the image, d_i. Straightforward algebra gives

$$d_i = \left(\frac{1}{f} - \frac{1}{d_o}\right)^{-1} \quad .$$

Example: Interpretation of the focal length

An inbending mirror can be used to cook food in locations where fuel and electricity are unavailable. The sun is essentially infinitely far away, do we have $d_o = \infty$ and $1/d_o=0$, so $d_i=f$. The focal length can thus be interpreted as the location of an image formed by an object infinitely far away.

Using this reinterpretation of the focal length, we see that a small focal length corresponds to a deeply curved, strongly focusing mirror, i.e. one that acts very different from a flat mirror.



A mirror with shallow curvature has a long focal length, and reflects rays at angles not much different from those that woould be produced with a flat mirror.

Self-Checks

(2) Suppose an object is placed at the center of the sphere made by extending the surface of the mirror. What would you expect to happen based on a ray diagram? How would you interpret this in relation to the equation for the position of the image?

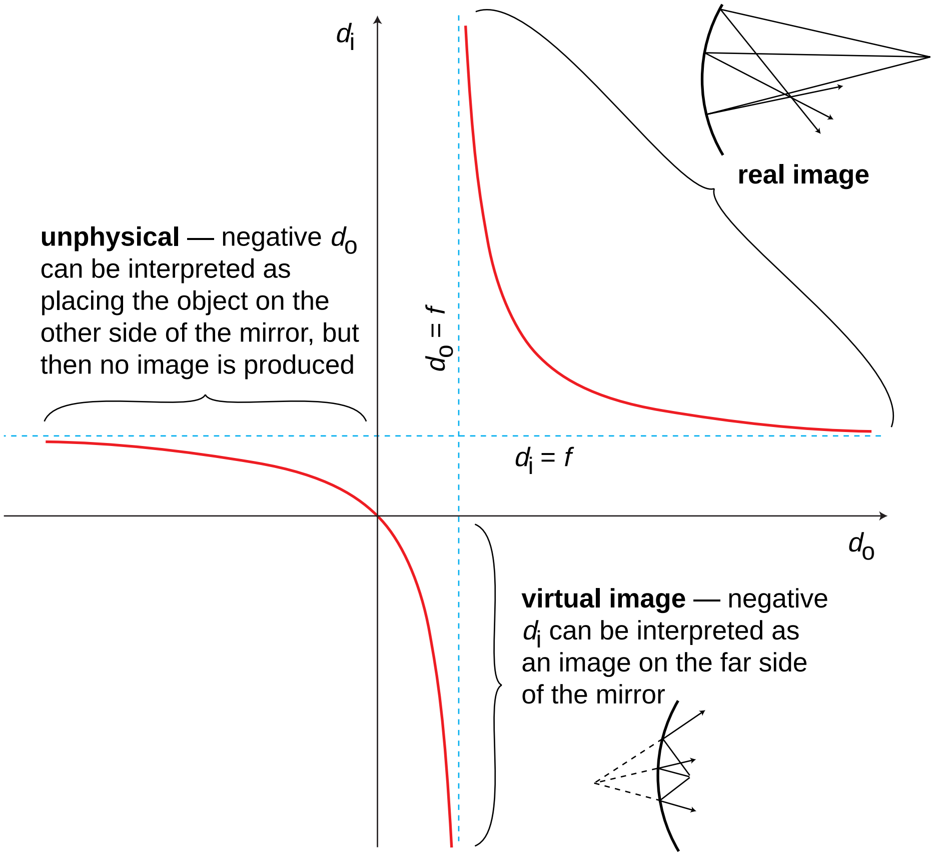
(3) Suppose a lightbulb is placed at a distance of one focal length from an inbending mirror. What will happen?

Magnification

We have already discussed in the previous chapter how to find the magnification of a virtual image made by a curved mirror. The result is the same for a real image, and we omit the proof, which is very similar. In our new notation, the result is

$$M = \frac{d_i}{d_o} \quad .$$

(1) The $1/d_i$ and $1/d_o$ terms have to add up to a fixed amount. If d_o gets smaller, then $1/d_o$ gets bigger, so $1/d_i$ must get smaller, meaning d_i becomes greater. Conversely, a bigger d_o leads to a smaller d_i. (2) Each ray will extend out along a radius of the sphere. The ones that strike the mirror will strike it along the normal and be reflected back along the same path. They will reconverge to form an image at the same location where the object is. This effect can be observed with sound in a spherical room. (3) This is the same as the example of the sun-oven, but with all the rays reversed in time. An image of the lightbulb will be formed infinitely far away, i.e. the rays will all be reflected in the same direction, so that they never cross. This is a searchlight.



3.2 Other Cases With Curved Mirrors

The equation $d_i = \left(1/f - 1/d_o\right)^{-1}$ can easily produce a negative result, but we have been thinking of d_i as a distance, and distances can't be negative. What's going on here?

Let's look at a graph of d_i as a function of d_o. The branch on the upper right corresponds to the case of a real image. Strictly speaking, this is the only part of the graph that we've proven corresponds to reality, since we never did any geometry for other cases, such as virtual images. As discussed in the self-check in the previous section, making d_o bigger causes d_i to become smaller, and vice-versa.

From drawing ray diagrams, we have already seen what happens when the object is very close to the mirror: the mirror intercepts a cone of rays from the object that is very broad, too broad to allow the mirror to bend the rays back together again. A virtual image is produced on the far side of the mirror. This is the first example of Wigner's "unreasonable effectiveness of mathematics" that we have encountered in optics. Even though our proof depended on the assumption that the image was real, the equation we derived turns out to be applicable to virtual images, provided that we either interpret the positive and negative signs in a certain way, or else modify the equation to have different positive and negative signs.

Self-Check

Interpret the three places where, in physically realistic parts of the graph, the graph approaches one of the dashed lines. [This will come more naturally if you have learned the concept of limits in a math class.]

Example: A flat mirror

We can even apply the equation to a flat mirror. As a sphere gets bigger and bigger, its surface is more and more gently curved. The planet Earth is so large, for example, that we cannot even perceive the curvature of its surface. To represent a flat mirror, we let the mirror's radius of curvature, and its focal length, become infinite. Dividing by infinity gives zero, so we have

$$1/d_o = -1/d_i \quad ,$$

or

$$d_o = -d_i \quad .$$

If we interpret the minus sign as indicating a virtual image on the far side of the mirror from the object, this makes sense.

It turns out that for any of the six possible combinations of real or virtual images formed by inbending or out-bending lenses or mirrors, we can apply an equation of the form

$$\frac{1}{f} = \frac{1}{d_\text{i}} + \frac{1}{d_\text{o}} \quad ,$$

with only a modification of plus or minus signs. There are two possible approaches here. The approach we have been using so far is the more popular approach in textbooks: leave the equation the same, but attach interpretations to the resulting negative or positive values of the variables. The trouble with this approach is that one is then forced to memorize tables of sign conventions, e.g. that the value of d_i should be negative when the image is a virtual image formed by an inbending mirror. Positive and negative signs also have to be memorized for focal lengths. Ugh! It's highly unlikely that any student has ever retained these lengthy tables in his or her mind for more than five minutes after handing in the final exam in a physics course. Of course one can always look such things up when they are needed, but the effect is to turn the whole thing into an exercise in blindly plugging numbers into formulas.

At the top of the graph, d_i approaches infinity when d_o approaches f; interpretation: the rays just barely converge to the right of the mirror. On the far right, d_i approaches f as d_o approaches infinity; this is the definition of the focal length. At the bottom, d_i approaches negative infinity when d_o approaches f from the other side; interpretation: the rays don't quite converge on the right side of the mirror, so they appear to have come from a virtual image point very far to the left of the mirror.

As you have gathered by now, I have a method I think is better. In this method, the variables f, d_i, and d_o are all positive by definition, and we put in positive and negative signs in the equation depending on the situations. Rather than memorizing these signs, we start with the generic equation

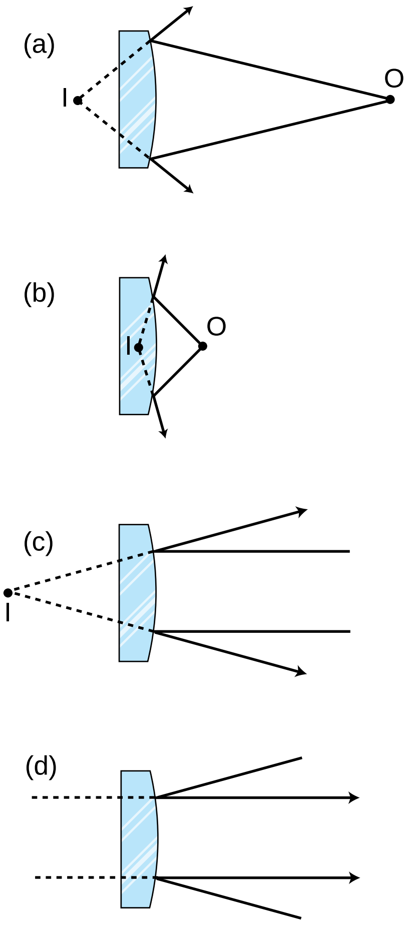
$$\frac{1}{f} = \pm\frac{1}{d_\text{i}} \pm \frac{1}{d_\text{o}}$$

and then determine the signs by a two-step method that depends on ray diagrams. The method is as follows:

1. Use ray diagrams to decide whether d_i and d_o vary in the same way or in opposite ways. (In other words, decide whether making d_o greater results in a greater value of d_i or a smaller one.) Based on this, decide whether the two signs in the equation are the same or opposite. If the signs are opposite, go on to step 2 to determine which is positive and which is negative.

2. It is normally only physically possible for either d_i or d_o to be infinite, not both. (Of course it is always possible to put an object at infinity, but that might for instance result in the formation of a real image, when you are interested in the case of a virtual image.) If we imagine the case where that variable is infinite, then dividing by infinity gives zero, so the only term on the right side of the equation would be the one that has the other variable in it. Since the left-hand side of the equation is positive by definition, the term on the right that we didn't eliminate must be the one that has a plus sign.

Although focal lengths are always positive in the method used in this book, you should be aware that out-bending mirrors and lenses are assigned negative focal lengths in the other method, so if you see a lens labeled f=–30 cm, you'll know what it means.



Example: An anti-shoplifting mirror

Question: Convenience stores often install an out-bending mirror so that the clerk has a view of the whole store and can catch shoplifters. Use a ray diagram to show that the image is reduced, bringing more into the clerk's field of view. If the focal length of the mirror is 3.0 m, and the mirror is 7.0 m from the farthest wall, how deep is the image of the store? (Note that in the other method of handling the signs, the focal length would have been given as –3.0 m.)

Solution: As shown in ray diagram (a), d_i is less than d_o. The magnification, $M = d_i/d_o$, will be less than one, i.e. the image is actually reduced rather than magnified.

We now apply the method outlined above for determining the plus and minus signs. Step 1: The object is the point on the opposite wall. As an experiment, (b), we try making the object closer — *much much* closer, so that even if our drawing isn't perfectly accurate we'll still get the right result for the change in the image's location. (I did these drawings using illustration software, but if you were doing them by hand, you'd also want to make much larger ones for greater accuracy.) Decreasing d_o has clearly made d_i smaller as well. There must be a cancellation of the effects of changing the two terms on the right in the same way, and the only way to get such a cancellation is if the two terms have *opposite signs*. (If they were both positive, for example, then increasing both terms would have to lead to an increase in the value of the whole right side, but that's impossible because the left side is a fixed constant.)

Step 2: Now which is the positive term and which is negative? Figure (c) shows a perfectly reasonable ray diagram of an image formed of an object at infinity, the moon, for example. Figure (d) shows an attempt to make an image at infinity. To get an image at infinity, we would have had to start with a converging set of rays, which is not physically possible, since diffuse reflection from a point on the wall creates diverging rays. If d_o can be infinite, then the sign of the d_i term must be positive.

We have now determined that the form of the equation must be

$$\frac{1}{f} = \frac{1}{d_i} - \frac{1}{d_o} \quad .$$

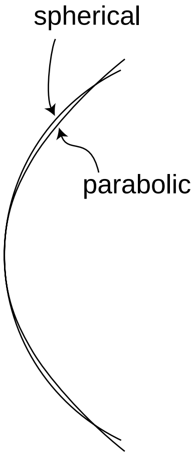
Solving for d_i, we find

$$\begin{aligned} d_i &= \left(\frac{1}{f} + \frac{1}{d_o}\right)^{-1} \\ &= 2.1 \text{ m} \quad . \end{aligned}$$

(The image of the store is reduced by a factor of 2.1/7.0=0.3, i.e. it is smaller by 70%.)

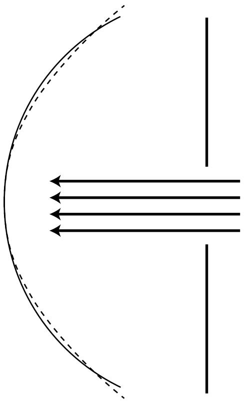
3.3* Aberrations

An imperfection or distortion in an image is called an aberration. An aberration can be produced by a flaw in a lens or mirror, but even with a perfect optical surface some degree of aberration is unavoidable. To see why, consider the mathematical approximation we've been making, which is that the depth of the mirror's curve is small compared to d_o and d_i. Since only a flat mirror can satisfy this shallow-mirror condition perfectly, any curved mirror will deviate somewhat from the mathematical behavior we derived by assuming that condition. There are two main types of aberration in curved mirrors, and these also occur with lenses.



Spherical mirrors are the cheapest to make, but parabolic mirrors are better for making images of objects at infinity. A sphere has equal curvature everywhere, but a parabola has tighter curvature at its center and gentler curvature at the sides.

(1) The image may be sharp when the object is at certain distances and blurry when it is at other distances. The blurriness occurs because the rays do not all cross at exactly the same point. If we know in advance the distance of the objects with which the mirror or lens will be used, then we can optimize the shape of the optical surface to make in-focus images in that situation. For instance, a spherical mirror will produce a perfect image of an object that is at the center of the sphere, because each ray is reflected directly onto the radius along which it was emitted. For objects at greater distances, however, the focus will be somewhat blurry. In astronomy the objects being used are always at infinity, so a spherical mirror is a poor choice for a telescope. A different shape (a parabola) is better specialized for astronomy.



Even though the spherical mirror (solid line) is not well adapted for viewing an object at infinity, we can improve its performance greatly by stopping it down. Now the only part of the mirror being used is the central portion, where its shape is virtually indistinguishable from a parabola (dashed line).

(2) An object on the axis of the lens or mirror may be imaged correctly, but off-axis objects may be out of focus. In a camera, this type of aberration would show up as a fuzziness near the sides of the picture when the center was perfectly focused.

One way of decreasing aberration is to use a small-diameter mirror or lens, or block most of the light with an opaque screen with a hole in it, so that only light that comes in close to the axis can get through. Either way, we are using a smaller portion of the lens or mirror whose curvature will be more shallow, thereby making the shallow-mirror (or thin-lens) approximation more accurate. Your eye does this by narrowing down the pupil to a smaller hole. In a camera, there is either an automatic or manual adjustment, and narrowing the opening is called "stopping down." The disadvantage of stopping down is that light is wasted, so the image will be dimmer or a longer exposure must be used.

What I would suggest you take away from this discussion for the sake of your general scientific education is simply an understanding of what an aberration is, why it occurs, and how it can be reduced, not detailed facts about specific types of aberrations.

Summary

Selected Vocabulary

focal length........................ a property of a lens or mirror, equal to the distance from the lens or mirror to the image it forms of an object that is infinitely far away

Notation

f... the focal length

d_o....................................... the distance of the object from the mirror (technically from the plane tangent to the center of the mirror, although this seldom matters much for a mirror whose curve is shallow)

d_i....................................... the distance of the image from the mirror

Notation Used in Other Books

$f>0$.................................... describes an inbending lens or mirror; in this book, all focal lengths are positive, so there is no such implication

$f<0$ .... describes an out-bending lens or mirror; in this book, all focal lengths are positive

$M<0$ .... indicates an inverted image

Summary

Every lens or mirror has a property called the focal length, which is defined as the distance from the lens or mirror to the image it forms of an object that is infinitely far away. A stronger lens or mirror has a shorter focal length.

The relationship between the locations of an object and its image formed by a lens or mirror can always be expressed by an equation of the form

$$\frac{1}{f} = \pm\frac{1}{d_i} \pm \frac{1}{d_o} \quad ,$$

The choice of plus and minus signs depends on whether we are dealing with a lens or a mirror, whether the lens or mirror is inbending or outbending, and whether the image is real or virtual. A method for determining the plus and minus signs is as follows:

1. Use ray diagrams to decide whether d_i and d_o vary in the same way or in opposite ways. Based on this, decide whether the two signs in the equation are the same or opposite. If the signs are opposite, go on to step 2 to determine which is positive and which is negative.

2. It is normally only physically possible for either d_i or d_o to be infinite, not both. If we imagine the case where that variable is infinite, then dividing by infinity gives zero, so the only term on the right side of the equation would be the one that has the other variable in it. Since the left-hand side of the equation is positive by definition, the term on the right that we didn't eliminate must be the one that has a plus sign.

Once the correct form of the equation has been determined, the magnification can be found via the equation

$$M = \frac{d_i}{d_o} \quad .$$

Homework Problems

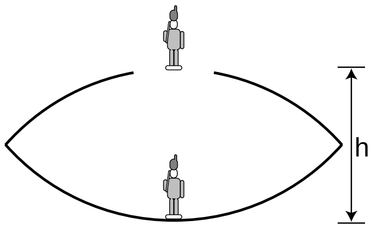
1. Apply the equation $M=d_i/d_o$ to the case of a flat mirror.

2 S. Use the method described in the text to derive the equation relating object distance to image distance for the case of a virtual image produced by an inbending mirror.

3. (a) Make up a numerical example of a virtual image formed by an inbending mirror with a certain focal length, and determine the magnification. (You will need the result of the previous problem.) Now change the location of the object a little bit and redetermine the magnification, showing that it changes. At my local department store, the cosmetics department sells mirrors advertised as giving a magnification of 5 times. How would you interpret this?

(b★) Suppose a Newtonian telescope is being used for astronomical observing. Assume for simplicity that no eyepiece is used, and assume a value for the focal length of the mirror that would be reasonable for an amateur instrument that is to fit in a closet. Is the angular magnification different for the moon than for a distant star?

4. (a) Find a case where the magnification of a curved mirror is infinite. Is the *angular* magnification infinite from any realistic viewing position? (b) Explain why infinite magnification can't be achieved by having d_o=0.



Problem 5.

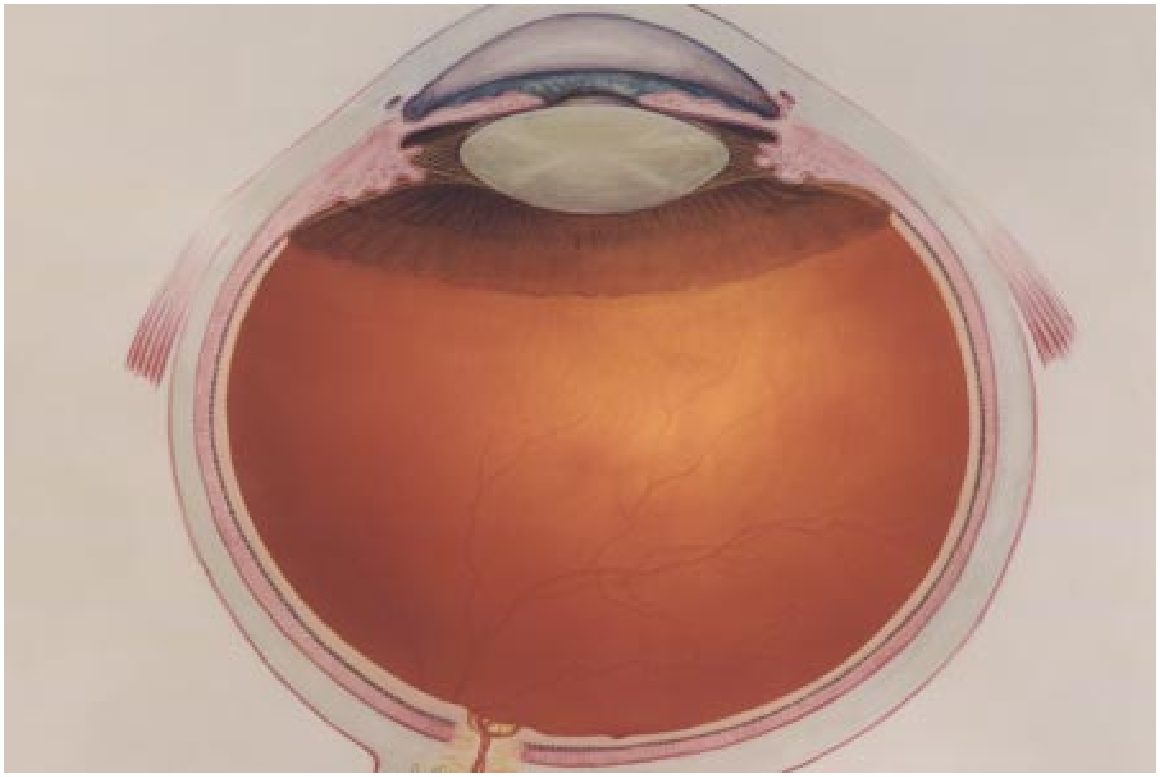
5★. The figure shows a device for constructing a realistic optical illusion. Two mirrors of equal focal length are put against each other with their silvered surfaces facing inward. A small object placed in the bottom of the cavity will have its image projected in the air above. The way it works is that the top mirror produces a virtual image, and the bottom mirror then creates a real image of the virtual image. (a) Show that if the image is to be positioned as shown, at the mouth of the cavity, then the focal length of the mirrors is related to the dimension h via the equation

$$\frac{1}{f} = \frac{1}{h} + \frac{1}{h + \left(\frac{1}{h} - \frac{1}{f}\right)^{-1}} \quad .$$

(b) Restate the equation in terms of a single variable $x=h/f$, and show that there are two solutions for x. Which solution is physically consistent with the assumptions of the calculation?

S A solution is given in the back of the book. ★ A difficult problem.

✓ A computerized answer check is available. ∫ A problem that requires calculus.

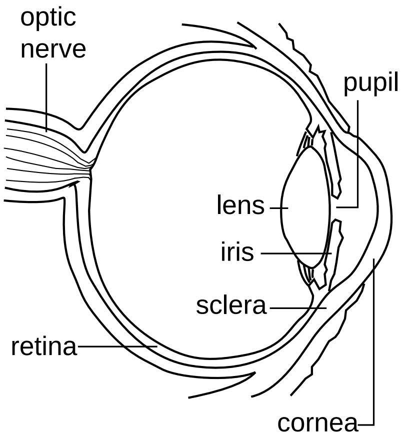


4 Refraction and Images

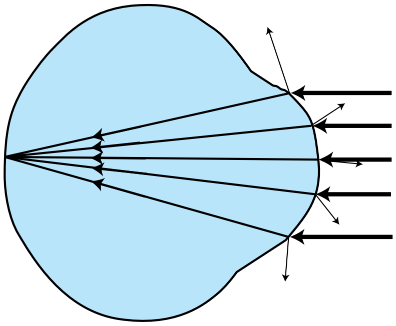
Economists normally consider free markets to be the natural way of judging the monetary value of something, but social scientists also use questionnaires to gauge the relative value of privileges, disadvantages, or possessions that cannot be bought or sold. They ask people to *imagine* that they could trade one thing for another and ask which they would choose. One interesting result is that the average light-skinned person in the U.S. would rather lose an arm than suffer the racist treatment routinely endured by African-Americans. Even more impressive is the value of sight. Many prospective parents can imagine without too much fear having a deaf child, but would have a far more difficult time coping with raising a blind one.

So great is the value attached to sight that some have imbued it with mystical aspects. Moses "had vision," George Bush did not. Christian fundamentalists who perceive a conflict between evolution and their religion have claimed that the eye is such a perfect device that it could never have arisen through a process as helter-skelter as evolution, or that it could not have evolved because half of an eye would be useless. In fact, the structure of an eye is fundamentally dictated by physics, and it has arisen separately by evolution somewhere between eight and 40 times, depending on which biologist you ask. We humans have a version of the eye that can be traced back to the evolution of a light-sensitive “eye spot” on the head of an ancient invertebrate. A sunken pit then developed so that the eye would only receive light from one direction, allowing the organism to tell where the light was coming from. (Modern flatworms have this type of eye.) The top of the pit then became partially covered, leaving a hole, for even greater directionality (as in the nautilus). At some point the cavity became filled with jelly, and this jelly finally became a lens, resulting in the general type of eye that we share with the bony fishes and other vertebrates. Far from being a perfect device, the vertebrate eye is marred by a serious design flaw due to the lack of planning or intelligent design in evolution: the nerve cells of the retina and the blood vessels that serve them are all in front of the light-sensitive cells, blocking part of the light. Squids and other molluscs, whose eyes evolved on a separate branch of the evolutionary tree, have a more sensible arrangement, with the light-sensitive cells out in front.

4.1 Refraction



(a) The anatomy of the human eye. (After an uncopyrighted diagram by the National Eye Institute, NIH.)



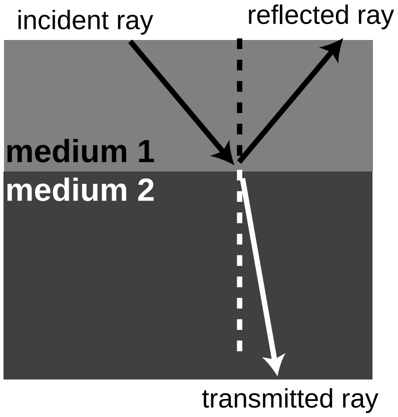
(b) A simplified optical diagram of the eye. Light rays are bent when they cross from the air into the eye.

Refraction

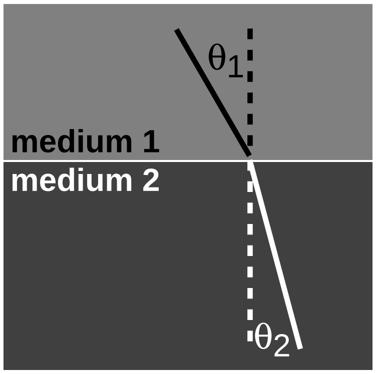
The fundamental physical phenomenon at work in the eye is that when light crosses a boundary between two media (such as air and the eye’s jelly), part of its energy is reflected, but part passes into the new medium. In the ray model of light, we describe the original ray as splitting into a reflected ray and a transmitted one (the one that gets through the boundary). Of course the reflected ray goes in a direction that is different from that of the original one, according to the rules of reflection we have already studied. More surprisingly — and this is the crucial point for making your eye focus light — the transmitted ray is bent somewhat as well. This bending phenomenon is called *refraction*. The origin of the word is the same as that of the word “fracture,” i.e. the ray is bent or “broken.” (Keep in mind, however, that light rays are not physical objects that can really be “broken.”) Refraction occurs with all waves, not just light waves.

The actual anatomy of the eye, (a), is quite complex, but in essence it is very much like every other optical device based on refraction. The rays are bent when they pass through the front surface of the eye. Rays that enter farther from the central axis are bent more, with the result that an image is formed on the retina. There is only one slightly novel aspect of the situation. In most human-built optical devices, such as a movie projector, the light is bent as it passes into a lens, bent again as it reemerges, and then reaches a focus beyond the lens. In the eye, however, the “screen” is inside the eye, so the rays are only refracted once, on entering the jelly, and never emerge again.

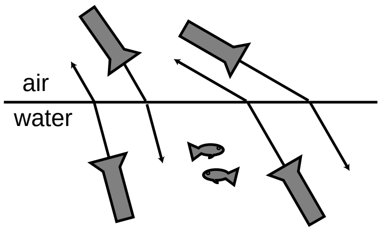
A common misconception is that the “lens” of the eye is what does the focusing. All the transparent parts of the eye are made of fairly similar stuff, so the dramatic change in medium is when a ray crosses from the air into the eye (at the outside surface of the cornea). This is where nearly all the refraction takes place. The lens medium differs only slightly in its optical properties from the rest of the eye, so very little refraction occurs as light enters and exits the lens. The lens, whose shape is adjusted by muscles attached to it, is only meant for fine-tuning the focus to form images of near or far objects.



(c) The incident, reflected, and refracted rays all lie in a plane that includes the normal (dashed line).



(d) The angles θ_1 and θ_2 are related to each other, and also depend on the properties of the two media. Because refraction is time-reversal symmetric, there is no need to label the rays with arrowheads.



(e) Refraction has time-reversal symmetry. Regardless of whether the light is going in or out of the water, the relationship between the two angles is the same, and the ray is closer to the normal while in the water.

Refractive properties of media

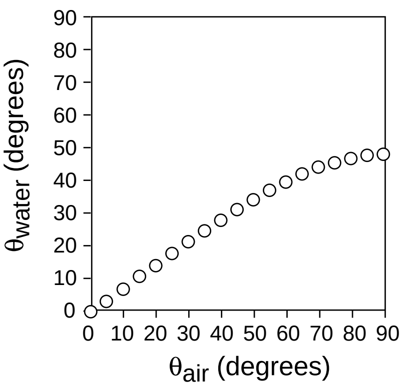
What are the rules governing refraction? The first thing to observe about is that just as with reflection, the new, bent part of the ray lies in the same plane as the normal (perpendicular) and the incident ray, (c).

If you try shooting a beam of light at the boundary between two substances, say water and air, you'll find that regardless of the angle at which you send in the beam, the part of the beam in the water is always closer to the normal line, (d). It doesn't matter if the ray is entering the water or leaving, so refraction is symmetric with respect to time-reversal, (e).

If, instead of water and air, you try another combination of substances, say plastic and gasoline, again you'll find that the ray's angle with respect to the normal is consistently smaller in one and larger in the other. Also, we find that if substance A has rays closer to normal than in B, and B has rays closer to normal than in C, then A has rays closer to normal than C. This means that we can rank-order all materials according to their refractive properties. Isaac Newton did so, including in his list many amusing substances, such as "Danzig vitriol" and "a pseudo-topazius, being a natural, pellucid, brittle, hairy stone, of a yellow color." Several general rules can be inferred from such a list:

- Vacuum lies at one end of the list. In refraction across the interface between vacuum and any other medium, the other medium has rays closer to the normal.
- Among gases, the ray gets closer to the normal if you increase the density of the gas by pressurizing it more.
- The refractive properties of liquid mixtures and solutions vary in a smooth and systematic manner as the proportions of the mixture are changed.
- Denser substances usually, but not always, have rays closer to the normal.

The second and third rules provide us with a method for measuring the density of an unknown sample of gas, or the concentration of a solution. The latter technique is very commonly used, and the CRC Handbook of Physics and Chemistry, for instance, contains extensive tables of the refractive properties of sugar solutions, cat urine, and so on.



Snell's law

The numerical rule governing refraction was discovered by Snell, who must have collected experimental data something like what is shown on this graph and then attempted by trial and error to find the right equation. The equation he came up with was

$$\frac{\sin\theta_1}{\sin\theta_2} = \text{constant} \quad .$$

The value of the constant would depend on the combination of media used. For instance, any one of the data points in the graph would have sufficed to show that the constant was 1.3 for an air-water interface (taking air to be substance 1 and water to be substance 2).

Snell further found that if media A and B gave a constant K_{AB} and media B and C gave a constant K_{BC}, then refraction at an interface between A and C would be described by a constant equal to the product, $K_{AC}=K_{AB}K_{BC}$. This is exactly what one would expect if the constant depended on the ratio of some number characterizing one medium to the number characteristic of the second medium. This number is called the *index of refraction* of the medium, written as "n" in equations. Since measuring the angles would only allow him to determine the *ratio* of the indices of refraction of two media, Snell had to pick some medium and define it as having n=1. He chose to define vacuum as having n=1. (The index of refraction of air at normal atmospheric pressure is 1.0003, so for most purposes it is a good approximation to assume that air has n=1.) He also had to decide which way to define the ratio, and he chose to define it so that media with their rays closer to the normal would have larger indices of refraction. This had the advantage that denser media would typically have higher indices of refraction, and for this reason the index of refraction is also referred to as the optical density. Written in terms of indices of refraction, Snell's equation becomes

$$\frac{\sin\theta_1}{\sin\theta_2} = \frac{n_2}{n_1} \quad ,$$

but rewriting it in the form

$$n_1 \sin\theta_1 = n_2 \sin\theta_2$$

[relationship between angles of rays at the interface between media with indices of refraction n_1 and n_2; angles are defined with respect to the normal]

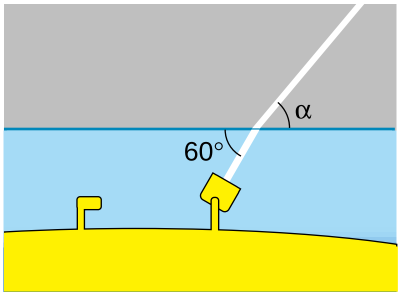
makes us less likely to get the 1's and 2's mixed up, so this the way most people remember Snell's law. A few indices of refraction are given in the front of the book.

Self-Check

(1) What would the graph look like for two substances with the same index of refraction?

(2) Based on the graph, when does refraction at an air-water interface change the direction of a ray most strongly?

[Answers on next page.]



Example

Question: A submarine shines its searchlight up toward the surface of the water. What is the angle θ shown in the figure?

Solution: The tricky part is that Snell's law refers to the angles with respect to the normal. Forgetting this is a very common mistake. The beam is at an angle of 30° with respect to the normal in the water. Let's call the air medium 1 and the water medium 2. Solving Snell's law for θ_1, we find

$$\theta_1 = \sin^{-1}\left(\frac{n_2}{n_1}\sin\theta_2\right)$$

As mentioned above, air has an index of refraction very close to 1, and water's is about 1.3, so we find θ_2=40°. The angle α is therefore 50°.

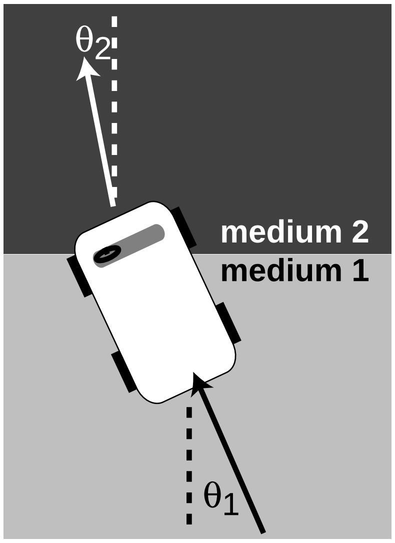
The index of refraction is related to the speed of light

What neither Snell nor Newton knew was that there is a very simple interpretation of the index of refraction. This may come as a relief to the reader who is taken aback by the complex reasoning involving proportionalities that led to its definition. Later experiments showed that the index of refraction of a medium was inversely proportional to the speed of light in that medium. Since c is defined as the speed of light in vacuum, and n=1 is defined as the index of refraction of vacuum, we have

$n = c/v$.   [n = medium's index of refraction, v = speed of light in that medium, c=speed of light in a vacuum]

Many textbooks start with this as the definition of the index of refraction, although that approach makes the quantity's name somewhat of a mystery, and leaves students wondering why c/v was used rather than v/c. It should also be noted that measuring angles of refraction is a far more practical method for determining n than direct measurement of the speed of light in the substance of interest.

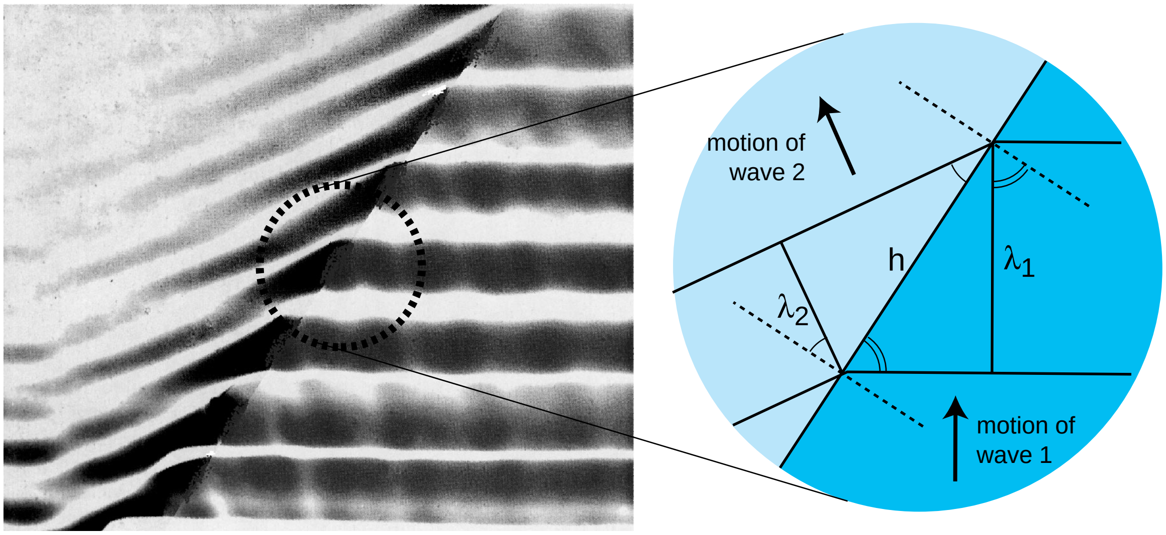
A mechanical model of Snell's law



Why should refraction be related to the speed of light? The mechanical model shown in the figure may help to make this more plausible. Suppose medium 2 is thick, sticky mud, which slows down the car. The car's right wheel hits the mud first, causing the right side of the car to slow down. This will cause the car to turn to the right until is moves far enough forward for the left wheel to cross into the mud. After that, the two sides of the car will once again be moving at the same speed, and the car will go straight.

Of course, light isn't a car. Why should a beam of light have anything resembling a "left wheel" and "right wheel?" After all, the mechanical model would predict that a motorcycle would go straight, and a motorcycle seems like a better approximation to a ray of light than a car. The whole thing is just a model, not a description of physical reality.

(1) If n_1 and n_2 are equal, Snell's law becomes sin θ_1=sin θ_2, which implies $\theta_1=\theta_2$. The graph would be a straight line along the diagonal of the graph.
(2) The graph is farthest from the diagonal when the angles are large, i.e. when the ray strikes the interface at an oblique or grazing angle.



(a) Refraction of a water wave. The water in the upper left part of the tank is shallower, so the speed of the waves is slower there, and their wavelength is shorter. The reflected part of the wave is also very faintly visible.

Retouched from an uncopyrighted PSSC College Physics photograph.

(b) A close-up view of what happens at the interface between the deeper medium and the shallower medium. The dashed lines are normals to the interface. The two marked angles on the right side are both equal to θ_1, and the two on the left equal θ_2.

A derivation of Snell's law

However intuitively appealing the mechanical model may be, light is a wave, and we should be using wave models to describe refraction. In fact Snell's law can be derived quite simply from wave concepts. In figure (b), simple trigonometry gives

$$\sin\theta_1 = \lambda_1/h \quad \text{and}$$

$$\sin\theta_2 = \lambda_2/h \quad .$$

Eliminating h by dividing the equations, we find

$$\frac{\sin\theta_1}{\sin\theta_2} = \frac{\lambda_1}{\lambda_2} \quad .$$

The frequencies of the two waves must be equal or else they would get out of step, so by $v=f\lambda$ we know that their wavelengths are proportional to their velocities. Combining $\lambda \propto v$ with $v \propto 1/n$ gives $\lambda \propto 1/n$, so we find

$$\frac{\sin\theta_1}{\sin\theta_2} = \frac{n_2}{n_1} \quad ,$$

which is one form of Snell's law.

Color and refraction

In general, the speed of light in a medium depends both on the medium and on the wavelength of the light. Another way of saying it is that a medium's index of refraction varies with wavelength. This is why a prism can be used to split up a beam of white light into a rainbow. Each wavelength of light is refracted through a different angle.

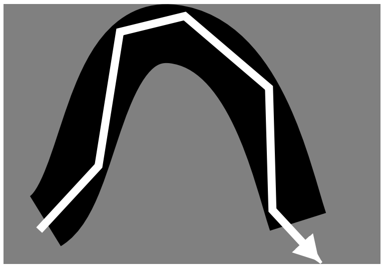
How much light is reflected, and how much is transmitted?

In book 3 we developed an equation for the percentage of the wave energy that is transmitted and the percentage reflected at a boundary between media. This was only done in the case of waves in one dimension, however, and rather than discuss the full three dimensional generalization it will be more useful to go into some qualitative observations about what happens. First, reflection happens only at the interface between two media, and two media with the same index of refraction act as if they were a single medium. Thus, at the interface between media with the same index of refraction, there is no reflection, and the ray keeps going straight. Continuing this line of thought, it is not surprising that we observe very little reflection at an interface between media with similar indices of refraction.

The next thing to note is that it is possible to have situations where no possible angle for the refracted ray can satisfy Snell's law. Solving Snell's law for θ_2, we find

$$\theta_2 = \sin^{-1}\left(\frac{n_1}{n_2}\sin\theta_1\right) \quad ,$$

and if n_1 is greater than n_2, then there will be large values of θ_1 for which the quantity $(n_1/n_2)\sin\theta$ is greater than one, meaning that your calculator will flash an error message at you when you try to take the inverse sine. What can happen physically in such a situation? The answer is that all the light is reflected, so there is no refracted ray. This phenomenon is known as *total internal reflection*, and is used in the fiber-optic cables that nowadays carry almost all long-distance telephone calls. The electrical signals from your phone travel to a switching center, where they are converted from electricity into light. From there, the light is sent across the country in a thin transparent fiber. The light is aimed straight into the end of the fiber, and as long as the fiber never goes through any turns that are too sharp, the light will always encounter the edge of the fiber at an angle sufficiently oblique to give total internal reflection. If the fiber-optic cable is thick enough, one can see an image at one end of whatever the other end is pointed at.



Total internal reflection in a fiber-optic cable.

Alternatively, a bundle of cables can be used, since a single thick cable is too hard to bend. This technique for seeing around corners is useful for making surgery less traumatic. Instead of cutting a person wide open, a surgeon can make a small "keyhole" incision and insert a bundle of fiber-optic cable (known as an endoscope) into the body.

Since rays at sufficiently large angles with respect to the normal may be completely reflected, it is not surprising that the relative amount of reflection changes depending on the angle of incidence, and is weakest for small angles of incidence.

Discussion questions

A. What index of refraction should a fish have in order to be invisible?

B. Does a surgeon using an endoscope need a source of light inside the body cavity? If so, how could this be done without inserting a light bulb through the incision?

C. A denser sample of a gas has a higher index of refraction than a less dense sample (i.e. a sample under lower pressure), but why would it not make sense for the index of refraction of a gas to be proportional to density?

D. The earth's atmosphere gets thinner and thinner as you go higher in altitude. If a ray of light comes from a star that is below the zenith, what will happen to it as it comes into the earth's atmosphere?

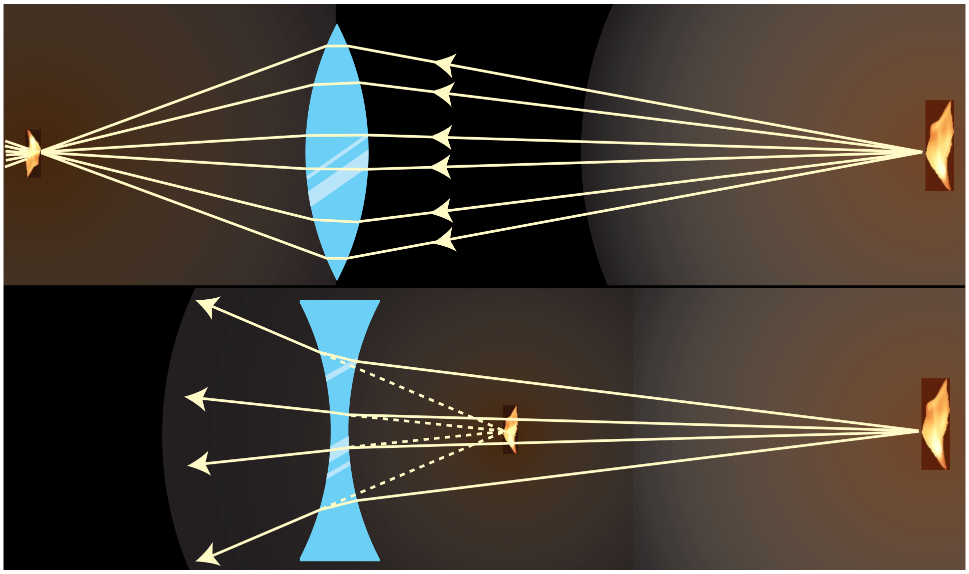
E. Does total internal reflection occur when light in a denser medium encounters a less dense medium, or the other way around? Or can it occur in either case?

4.2 Lenses

Figures (a) and (b) show examples of lenses forming images. There is essentially nothing for you to learn about imaging lenses that is truly new. You already know how to construct and use ray diagrams, and you know about real and virtual images. The definition of the focal length of a lens is the same as the one for a curved mirror: the location of an image formed of an object infinitely far away. The equations for locating images and determining magnifications are of the same form. It's really just a question of flexing your mental muscles on a few examples. The following self-checks and discussion questions will get you started.

Self-Checks

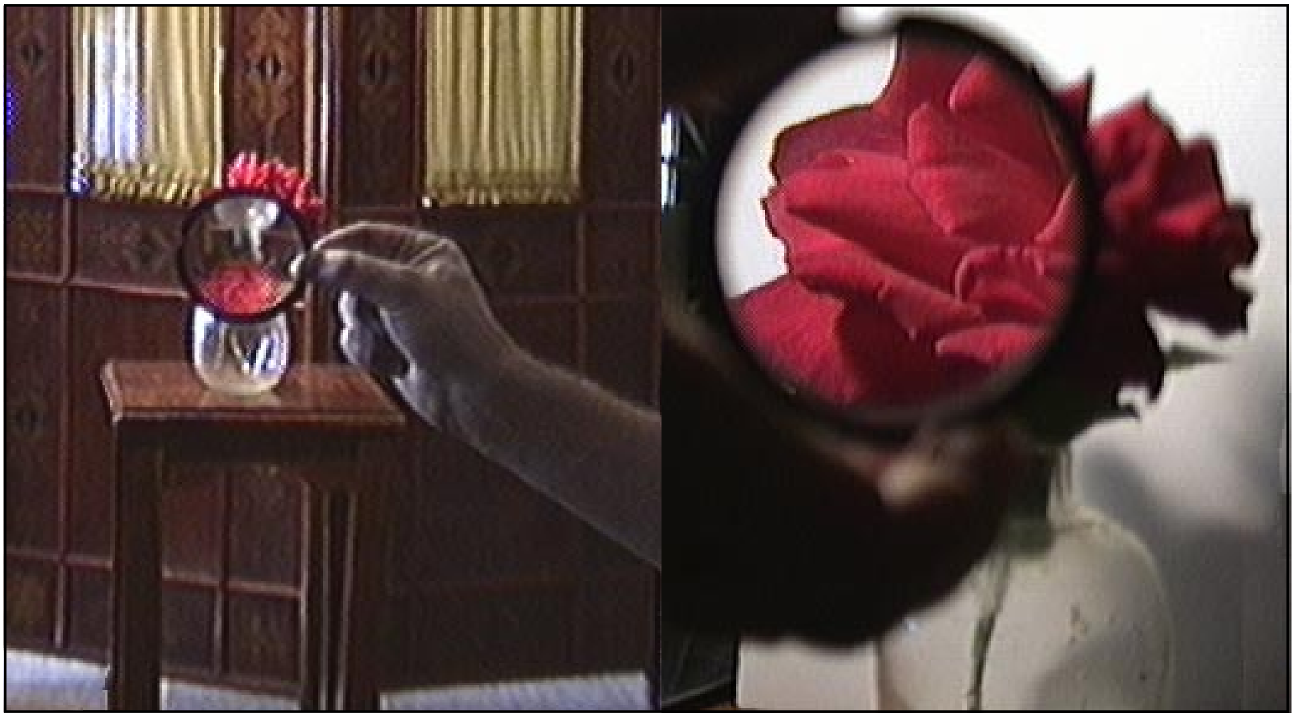
(1) In figures (a) and (b), classify the images as real or virtual.
(2) Glass has an index of refraction that is greater than that of air. Consider the topmost ray in figure (a). Explain why the ray makes a slight left turn upon entering the lens, and another left turn when it exits.
(3) If the flame in figure (b) was moved closer to the lens, what would happen to the location of the image?



(a) An inbending lens is making an image of a candle flame.

(b) Now an outbending lens is making an image of the flame.

(1) In (a), the rays cross at the image, so it is real. In (b), the rays only appear to have come from the image point, so the image is virtual.
(2) A ray is always closer to the normal in the medium with the higher index of refraction. The first left turn makes the ray closer to the normal, as it should be in glass. The second left turn makes the ray farther from the normal, which is how it should be in air.
(3) Take the topmost ray as an example. It will still take two right turns, but since it is entering the lens at a steeper angle, it will also leave at a steeper angle. Tracing backward to image, the steeper lines will meet closer to the lens.



(c) Two images of a rose created by the same lens and recorded with the same camera.

Discussion Questions

A. In figures (a) and (b), the front and back surfaces are parallel to each other at the center of the lens. What will happen to a ray that enters near the center?

B. Suppose you wanted to change the setup in figure (a) so that the location of the actual flame in the figure would instead be occupied by an image of a flame. Where would you have to move the candle to achieve this? What about in (b)?

C. There are three qualitatively different types of image formation that can occur with lenses, of which figures (a) and (b) exhaust only two. Figure out what the third possibility is. Which of the three possibilities result in a magnification greater than one?

D. Classify the examples shown in figure (c) according to the types of images delineated in the previous discussion question.

E. In figures (a) and (b), the only rays drawn were those that happened to enter the lenses. Discuss this in relation to figure (c).

F. In the right-hand side of figure (c), the image viewed through the lens is in focus, but the side of the rose that sticks out is not. Why?

G. In general, the index of refraction depends on the color of the light. What effect would this have on images formed by lenses?

4.3* The Lensmaker's Equation

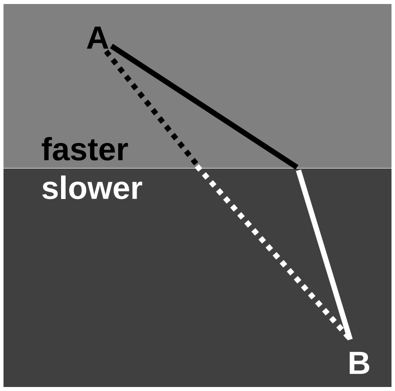
The focal length of a spherical mirror is simply $r/2$, but we cannot expect the focal length of a lens to be given by pure geometry, since it also depends on the index of refraction of the lens. Suppose we have a lens whose front and back surfaces are both spherical. (This is no great loss of generality, since any surface with a sufficiently shallow curvature can be approximated with a sphere.) Then if the lens is immersed in a medium with an index of refraction of 1, its focal length is given approximately by

$$f = \left[(n-1) \left| \frac{1}{r_1} \pm \frac{1}{r_2} \right| \right]^{-1} \quad .$$

This is known as the lensmaker's equation. In my opinion it is not particularly worthy of memorization. The positive sign is used when both surface are curved outward or both are curved inward; otherwise a negative sign applies. The proof of this equation is left as an exercise to those readers who are sufficiently brave and motivated.

4.4* Refraction and the Principle of Least Time

We seen previously how the rules governing straight-line motion of light and reflection of light can be derived from the principle of least time. What about refraction? In the figure, it is indeed plausible that the bending of the ray serves to minimize the time required to get from a point A to point B. If the ray followed the unbent path shown with a dashed line, it would have to travel a longer distance in the medium in which its speed is slower. By bending the correct amount, it can reduce the distance it has to cover in the slower medium without going too far out of its way. It is true that Snell's law gives exactly the set of angles that minimizes the time required for light to get from one point to another. The proof of this fact is left as an exercise.



Summary

Selected Vocabulary

refraction .... the change in direction that occurs when a wave encounters the interface between two media

index of refraction .... an optical property of matter; the speed of light in a vacuum divided by the speed of light in the substance in question

Notation

n .... the index of refraction

Summary

Refraction is change in direction that occurs when a wave encounters the interface between two media. Together, refraction and reflection account for the basic principles behind nearly all optical devices.

Snell discovered the equation for refraction,

$$n_1 \sin \theta_1 = n_2 \sin \theta_2 \qquad \text{[angles measured with respect to the normal]}$$

through experiments with light rays, long before light was proven to be a wave. Snell's law can be proven based on the geometrical behavior of waves. Here *n* is the index of refraction. Snell invented this quantity to describe the refractive properties of various substances, but it was later found to be related to the speed of light in the substance,

$$n = c/v \quad ,$$

where *c* is the speed of light in a vacuum. In general a material's index of refraction is different for different wavelengths of light.

As discussed in the third book of this series, any wave is partially transmitted and partially reflected at the boundary between two media in which its speeds are different. It is not particularly important to know the equation that tells what fraction is transmitted (and thus refracted), but important technologies such as fiber optics are based on the fact that this fraction becomes *zero* for sufficiently oblique angles. This phenomenon is referred to as total internal reflection. It occurs when there is no angle that satisfies Snell's law.

Homework Problems

1. Suppose an inbending lens is constructed of a type of plastic whose index of refraction is less than that of water. How will the lens's behavior be different if it is placed underwater?

2. There are two main types of telescopes, refracting (using lenses) and reflecting (using mirrors). (Some telescopes use a mixture of the two types of elements: the light first encounters a large curved mirror, and then goes through an eyepiece that is a lens.) What implications would the color-dependence of focal length have for the relative merits of the two types of telescopes? What would happen with white starlight, for example?

3. Based on Snell's law, explain why rays of light passing through the edges of an inbending lens are bent more than rays passing through parts closer to the center. It might seem like it should be the other way around, since the rays at the edge pass through less glass — shouldn't they be affected less?

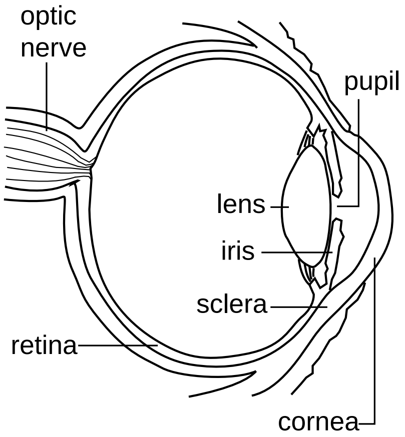
4. By changing the separation distance between lens and film, a camera can focus on subjects at a variety of distances. Suppose the proper lens-film separation for taking an in-focus picture of a distant object such as the moon is *x*. To take an in-focus picture of a nearby object, will the proper lens-film separation be greater than, equal to, or less than *x*? Explain using diagrams.

5 ★. (a) Light is being reflected diffusely from an object 1.000 m under water. The light that comes up to the surface is refracted at the water-air interface. If the refracted rays all appear to come from the same point, then there will be a virtual image of the object in the water, above the object's actual position, which will be visible to an observer above the water. Consider three rays, A, B and C, whose angles in the water with respect to the normal are θ_i=0.000°, 1.000° and 20.000° respectively. Find the depth of the point at which the refracted parts of A and B appear to have intersected, and do the same for A and C. Show that the intersections are at nearly the same depth, but not quite. [Check: The difference in depth should be about 4 cm.]

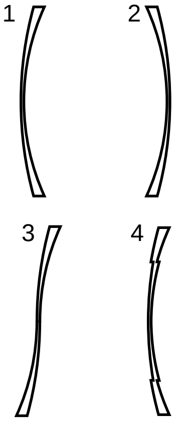
(b) Since all the refracted rays do not quite appear to have come from the same point, this is technically not a virtual image. In practical terms, what effect would this have on what you see?

(c) In the case where the angles are all small, use algebra and trig to show that the refracted rays do appear to come from the same point, and find an equation for the depth of the virtual image. Do not put in any numerical values for the angles or for the indices of refraction — just keep them as symbols. You will need the approximation $\sin\theta \approx \tan\theta \approx \theta$, which is valid for small angles measured in radians.

S A solution is given in the back of the book.	★ A difficult problem.
✓ A computerized answer check is available.	∫ A problem that requires calculus.



Problem 6.



Problem 8.

6 ★✓. The drawing shows the anatomy of the human eye, at twice life size. Find the radius of curvature of the outer surface of the cornea by measurements on the figure, and then derive the focal length of the air-cornea interface, where almost all the focusing of light occurs. You will need to use physical reasoning to modify the lensmaker's equation for the case where there is only a single refracting surface. Assume that the index of refraction of the cornea is essentially that of water.

7. When swimming underwater, why is your vision made much clearer by wearing goggles with flat pieces of glass that trap air behind them? [Hint: You can simplify your reasoning by considering the special case where you are looking at an object far away, and along the optic axis of the eye.]

8. The figure shows four lenses. Lens 1 has two spherical surfaces. Lens 2 is the same as lens 1 but turned around. Lens 3 is made by cutting through lens 1 and turning the bottom around. Lens 4 is made by cutting a central circle out of lens 1 and recessing it.

(a) A parallel beam of light enters lens 1 from the left, parallel to its axis. Reasoning based on Snell's law, will the beam emerging from the lens be bent inward or outward, or will it remain parallel to the axis? Explain your reasoning. [Hint: It may be helpful to make an enlarged drawing of one small part of the lens, and apply Snell's law at both interfaces. Recall that rays are bent more if they come to the interface at a larger angle with respect to the normal.]

(b) What will happen with lenses 2, 3, and 4? Explain. Drawings are not necessary.

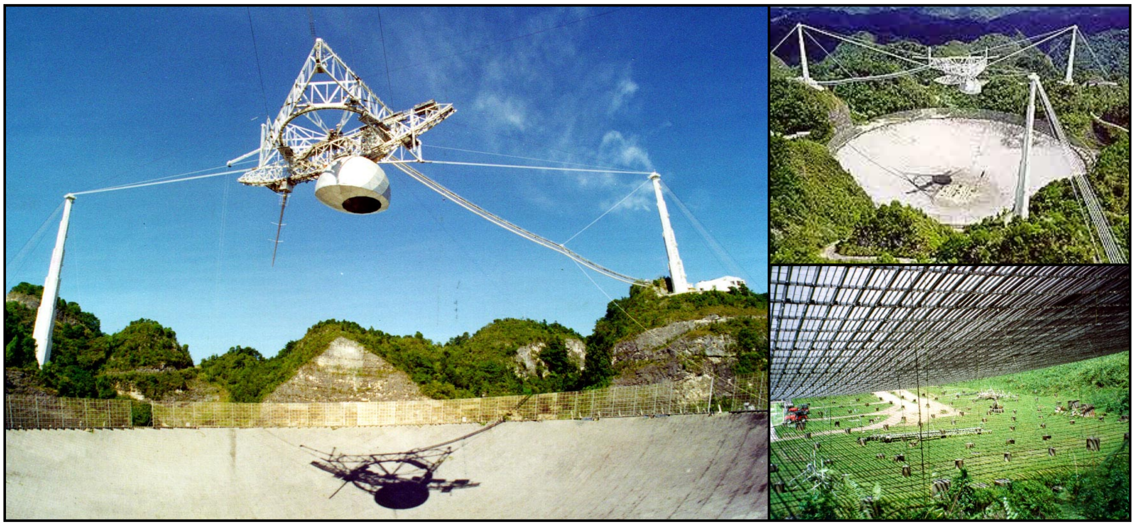
9 ★. Prove that the principle of least time leads to Snell's law.

10 ✓. An object is more than one focal length from an inbending lens. (a) Draw a ray diagram. (b) Using reasoning like that developed in the previous chapter, determine the positive and negative signs in the equation $\frac{1}{f}=\pm\frac{1}{d_\mathrm{i}}\pm\frac{1}{d_\mathrm{o}}$. (c) The images of the rose in section 4.2 were made using a lens with a focal length of 23 cm. If the lens is placed 80 cm from the rose, locate the image.

11 ✓. An object is less than one focal length from an inbending lens. (a) Draw a ray diagram. (b) Using reasoning like that developed in the previous chapter, determine the positive and negative signs in the equation $\frac{1}{f}=\pm\frac{1}{d_\mathrm{i}}\pm\frac{1}{d_\mathrm{o}}$. (c) The images of the rose in section 4.2 were made using a lens with a focal length of 23 cm. If the lens is placed 10 cm from the rose, locate the image.

12 ✓. Nearsighted people wear glasses whose lenses are outbending. (a) Draw a ray diagram. For simplicity pretend that there is no eye behind the glasses. (b) Using reasoning like that developed in the previous chapter, determine the positive and negative signs in the equation $\frac{1}{f}=\pm\frac{1}{d_\mathrm{i}}\pm\frac{1}{d_\mathrm{o}}$.

(c) If the focal length of the lens is 50 cm, and the person is looking at an object at a distance of 80 cm, locate the image.



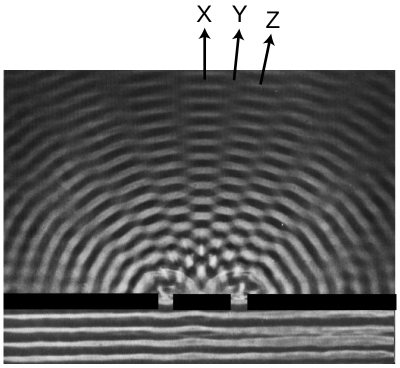
This huge radio dish nestles in a natural canyon at Arecibo, Puerto Rico. Its apparently solid surface is actually a scaffolding that hangs like a suspension bridge. It is a telescope that images the universe using radio waves rather than visible light, and it is also used to search for artificial radio signals from intelligent beings on other planets. Why does it have to be so huge? It's not mainly a question of picking up weak signals. The dish's sensitivity is overkill for most jobs, and for example it would have easily been able to pick up signals from a species on the other side of the galaxy that were no more intense than the ones humans themselves have transmitted out into space. (This is assuming that it was tuned to the right frequency at the right time, and that the signals came from within the limited field of view it sweeps out as the world spins.) No, the reason it has to be so big is a matter of wave optics. To make the antenna select signals from only a certain direction in space and reject those coming from other angles, it must be as large as possible compared to the wavelength of a radio wave.

5 Wave Optics

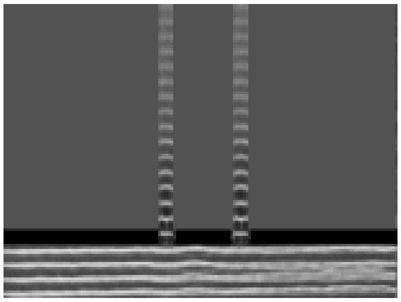
Electron microscopes can make images of individual atoms, but why will a visible-light microscope never be able to? Stereo speakers create the illusion of music that comes from a band arranged in your living room, but why doesn't the stereo illusion work with bass notes? Why are computer chip manufacturers investing billions of dollars in equipment to etch chips with x-rays instead of visible light?

The answers to all of these questions have to do with the subject of wave optics. So far this book has discussed the interaction of light waves with matter, and its practical applications to optical devices like mirrors, but we have used the ray model of light almost exclusively. Hardly ever have we explicitly made use of the fact that light is an electromagnetic wave. We were able to get away with the simple ray model because the chunks of matter we were discussing, such as lenses and mirrors, were thousands of times larger than a wavelength of light. We now turn to phenomena and devices that can only be understood using the wave model of light.

5.1 Diffraction



(a) In this view from overhead, a straight, sinusoidal water wave encounters a barrier with two gaps in it. Strong wave vibration occurs at angles X and Z, but there is none at all at angle Y.



(b) This doesn't happen.

Figures (a) and (b) were constructed as collages of uncopyrighted photos from PSSC College Physics. Figure (a), although essentially correct, is a little unrealistic because the waves beyond the barrier would be much weaker.

Figure (a) shows a typical problem in wave optics, enacted with water waves. It may seem surprising that we don't get a simple pattern like figure (b), but the pattern would only be that simple if the wavelength was hundreds of times shorter than the distance between the gaps in the barrier and the widths of the gaps.

Wave optics is a broad subject, but this example will help us to pick out a reasonable set of restrictions to make things more manageable:

(1) We restrict ourselves to cases in which a wave travels through a uniform medium, encounters a certain area in which the medium has different properties, and then emerges on the other side into a second uniform region.

(2) We assume that the incoming wave is a nice tidy sine-wave pattern with wavefronts that are lines (or, in three dimensions, planes).

(3) In figure (a) we can see that the wave pattern immediately beyond the barrier is rather complex, but farther on it sorts itself out into a set of wedges separated by gaps in which the water is still. We will restrict ourselves to studying the simpler wave patterns that occur farther away, so that the main question of interest is how intense the outgoing wave is at a given angle.

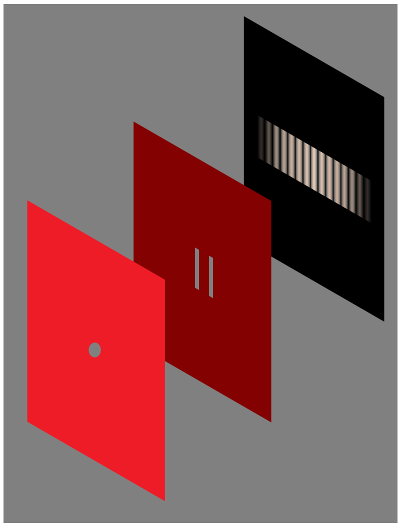
The kind of phenomenon described by restriction (1) is called *diffraction*. Diffraction can be defined as the behavior of a wave when it encounters an obstacle or a nonuniformity in its medium. In general, diffraction causes a wave to bend around obstacles and make patterns of strong and weak waves radiating out beyond the obstacle. Understanding diffraction is the central problem of wave optics. If you understand diffraction, even the subset of diffraction problems that fall within restrictions (2) and (3), the rest of wave optics is icing on the cake.

Diffraction can be used to find the structure of an unknown diffracting object: even if the object is too small to study with ordinary imaging, it may be possible to work backward from the diffraction pattern to learn about the object. The structure of a crystal, for example, can be determined by its x-ray diffraction pattern.

Diffraction can also be a bad thing. In a telescope, for example, light waves are diffracted by all the parts of the instrument. This will cause the image of a star, for example, to appear fuzzy even when the focus has been adjusted correctly. By understanding diffraction, one can learn how a telescope must be designed in order to reduce this problem — essentially, it should have the biggest possible diameter.

There are two ways in which restriction (2) might commonly be violated. First, the light might be a mixture of wavelengths. If we simply want to observe a diffraction pattern or to use diffraction as a technique for studying the object doing the diffracting (e.g. if the object is too small to see with a microscope), then we can pass the light through a colored filter before diffracting it.

A second issue is that light from sources such as the sun or a lightbulb



(c) A practical setup for observing diffraction of light.

does not consist of a nice neat plane wave, except over very small regions of space. Different parts of the wave are out of step with each other, and the wave is referred to as *incoherent*. One way of dealing with this is shown in figure (c). After filtering to select a certain wavelength of red light, we pass the light through a small pinhole. The region of the light that is intercepted by the pinhole is so small that one part of it is not out of step with another. Beyond the pinhole, light spreads out in a spherical wave; this is analogous to what happens when you speak into one end of a paper towel roll and the sound waves spread out in all directions from the other end. By the time the spherical wave gets to the double slit it has spread out and reduced its curvature, so that we can now think of it as a simple plane wave.

If this seems laborious, you may be relieved to know that modern technology gives us an easier way to produce a single-wavelength, coherent beam of light: the laser.

The parts of the final image on the screen in (c) are called diffraction fringes. The center of each fringe is a point of maximum brightness, and halfway between two fringes is a minimum.

Discussion Question

Why would x-rays rather than visible light be used to find the structure of a crystal? Sound waves are used to make images of fetuses in the womb. What would influence the choice of wavelength?

5.2 Scaling of Diffraction

This chapter has “optics” in its title, so it is nominally about light, but we started out with an example involving water waves. Water waves are certainly easier to visualize, but is this a legitimate comparison? In fact the analogy works quite well, despite the fact that a light wave has a wavelength about a million times shorter. This is because diffraction effects scale uniformly. That is, if we enlarge or reduce the whole diffraction situation by the same factor, including both the wavelengths and the sizes of the obstacles the wave encounters, the result is still a valid solution.

This is unusually simple behavior! In the first book of this series we saw many examples of more complex scaling, such as the impossibility of bacteria the size of dogs, or the need for an elephant to eliminate heat through its ears because of its small surface-to-volume ratio, whereas a tiny shrew’s life-style centers around conserving its body heat.

Of course water waves and light waves differ in many ways, not just in scale, but the general facts you will learn about diffraction are applicable to all waves. In some ways it might have been more appropriate to insert this chapter at the end of book 3, Vibrations and Waves, but many of the important applications are to light waves, and you would probably have found these much more difficult without any background in optics.

Another way of stating the simple scaling behavior of diffraction is that the type of diffraction we get depends only on the unitless ratio λ/d, where λ is the wavelength of the wave and d is some dimension of the diffracting objects, e.g. the center-to-center spacing between the slits in figures (a) and (b). If, for instance, we scale up both λ and d by a factor of 37, the ratio λ/d will be unchanged.

5.3 The Correspondence Principle

The only reason we don't usually notice diffraction of light in everyday life is that we don't normally deal with objects that are comparable in size to a wavelength of visible light, which is about a millionth of a meter. Does this mean that wave optics contradicts ray optics, or that wave optics sometimes gives wrong results? No. If you hold three fingers out in the sunlight and cast a shadow with them, *either* wave optics or ray optics can be used to predict the straightforward result: a shadow pattern with two bright lines where the light has gone through the gaps between your fingers. Wave optics is a more general theory than ray optics, so in any case where ray optics is valid, the two theories will agree. This is an example of a general idea enunciated by the physicist Niels Bohr, called the *correspondence principle:* when flaws in a physical theory lead to the creation of a new and more general theory, the new theory must still agree with the old theory within its more restricted area of applicability. After all, a theory is only created as a way of describing experimental observations. If the original theory had not worked in any cases at all, it would never have become accepted.

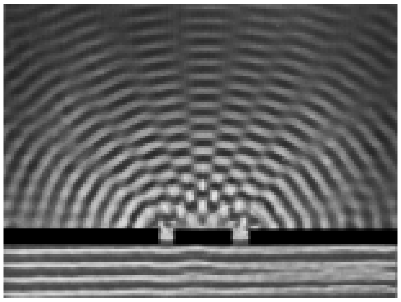
In the case of optics, the correspondence principle tells us that when λ/d is small, both the ray and the wave model of light must give approximately the same result. When you cast a shadow with your fingers, λ/d is about 10^{-4}, so the two models will agree very closely. (To be specific, the shadows of your fingers will be outlined by a series of light and dark fringes, but the angle subtended by a fringe will be on the order of 10^{-4} radians, so they will be invisible and washed out by the natural fuzziness of the edges of sun-shadows, caused by the finite size of the sun.)

Self-Check

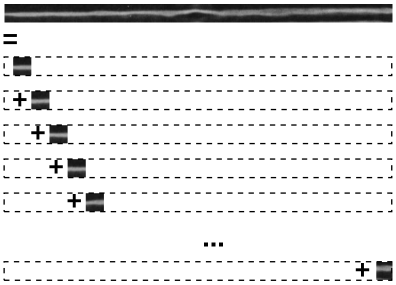
What kind of wavelength would an electromagnetic wave have to have in order to diffract dramatically around your body? Does this contradict the correspondence principle?

It would have to have a wavelength on the order of millimeters or meters, the same distance scale as that of your body. These would be microwaves or radio waves. (This effect can easily be noticed when a person affects a TV's reception by standing near the antenna.) None of this contradicts the correspondence principle, which only states that the wave model must agree with the ray model when the ray model is applicable. The ray model is not applicable here because λ/d is on the order of 1.

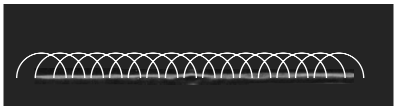
5.4 Huygens' Principle



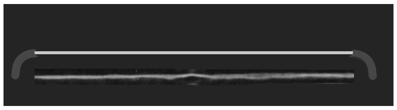
(a) Double-slit diffraction.



(b) A wavefront can be analyzed by the principle of superposition, breaking it down into many small parts.



(c) If it was by itself, each of the parts would spread out as a circular ripple.



(d) Adding up the ripples produces a new wavefront.

Returning to the example of double-slit diffraction, (a), note the strong visual impression of two overlapping sets of concentric semicircles. This is an example of *Huygens' principle*, named after the Dutch physicist who came up with it. (The first syllable rhymes with "boy.") Huygens' principle states that any wavefront can be broken down into many small side-by-side wave peaks, (b), which then spread out as circular ripples, (c), and by the principle of superposition, the result of adding up these sets of ripples must give the same result as allowing the wave to propagate forward, (d). In the case of sound or light waves, which propagate in three dimensions, the "ripples" are actually spherical rather than circular, but we can often imagine things in two dimensions for simplicity.

In double-slit diffraction the application of Huygens' principle is visually convincing: it is as though all the sets of ripples have been blocked except for two. It is a rather surprising mathematical fact, however, that Huygens' principle gives the right result in the case of an unobstructed linear wave, (c) and (d). A theoretically infinite number of circular wave patterns somehow conspire to add together and produce the simple linear wave motion with which we are familiar.

Since Huygens' principle is equivalent to the principle of superposition, and superposition is a property of waves, what Huygens had created was essentially the first wave theory of light. However, he imagined light as a series of pulses, like hand claps, rather than as a sinusoidal wave.

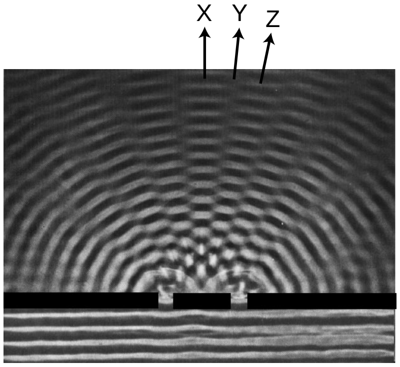
The history is interesting. Isaac Newton loved the atomic theory of matter so much that he searched enthusiastically for evidence that light was also made of tiny particles. The paths of his light particles would correspond to rays in our description; the only significant difference between a ray model and a particle model of light would occur if one could isolate individual particles and show that light had a "graininess" to it. Newton never did this, so although he thought of his model as a particle model, it is more accurate to say he was one of the builders of the ray model.

Almost all that was known about reflection and refraction of light could be interpreted equally well in terms of a particle model or a wave model, but Newton had one reason for strongly opposing Huygens' wave theory. Newton knew that waves exhibited diffraction, but diffraction of light is difficult to observe, so Newton believed that light did not exhibit diffraction, and therefore must not be a wave. Although Newton's criticisms were fair enough, the debate also took on the overtones of a nationalistic dispute between England and continental Europe, fueled by English resentment over Leibnitz's supposed plagiarism of Newton's calculus. Newton wrote a book on optics, and his prestige and political prominence tended to discourage questioning of his model.

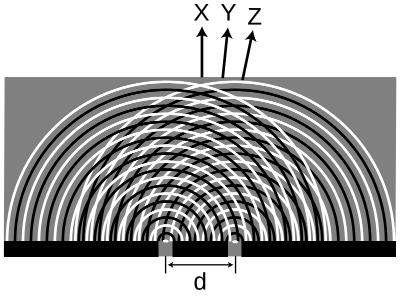
Thomas Young (1773-1829) was the person who finally, a hundred years later, did a careful search for wave interference effects with light and analyzed the results correctly. He observed double-slit diffraction of light as well as a variety of other diffraction effects, all of which showed that light exhibited wave interference effects, and that the wavelengths of visible light waves were extremely short. The crowning achievement was the demonstration by the experimentalist Heinrich Hertz and the theorist James Clerk

Maxwell that light was an *electromagnetic* wave. Maxwell is said to have related his discovery to his wife one starry evening and told her that she was the only person in the world who knew what starlight was.

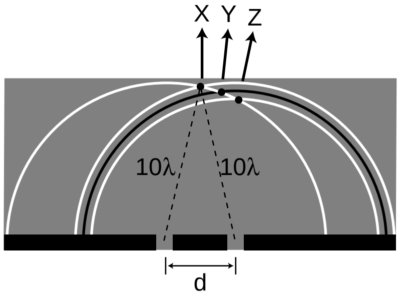
5.5 Double-Slit Diffraction



(a) Double-slit diffraction.



(b) Application of Huygens' principle. White lines represent peaks, black lines troughs, which we can refer to as positive and negative.



(c) Because both sets of ripples have ten wavelengths to cover in order to reach the point along direction X, they will be in step when they get there.

Let's now analyze double-slit diffraction, (a), using Huygens' principle. The most interesting question is how to compute the angles such as X and Z where the wave intensity is at a maximum, and the in-between angles like Y where it is minimized. Let us measure all our angles with respect to the vertical center line of the figure, which was the original direction of propagation of the wave.

If we assume that the width of the slits is small (on the order of the wavelength of the wave or less), then we can imagine only a single set of Huygens ripples spreading out from each one, (b). The only dimension of the diffracting slits that has any effect on the geometric pattern of the overlapping ripples is then the center-to-center distance, *d*, between the slits.

We know from our discussion of the scaling of diffraction that there must be some equation that relates an angle like θ_Z to the ratio λ/d,

$$\lambda / d \leftrightarrow \theta_Z \quad .$$

If the equation for θ_Z depended on some other expression such as $\lambda+d$ or λ^2/d, then it would change when we scaled λ and d by the same factor, which would violate what we know about the scaling of diffraction.

Along the central maximum line, X, we always have positive waves coinciding with positive ones and negative waves coinciding with negative ones. (I have arbitrarily chosen to take a snapshot of the pattern at a moment when the waves emerging from the slit are experiencing a positive peak.) The superposition of the two sets of ripples therefore results in a doubling of the wave amplitude along this line. There is constructive interference. This is easy to explain, because by symmetry, each wave has had to travel an equal number of wavelengths to get from its slit to the center line, (c).

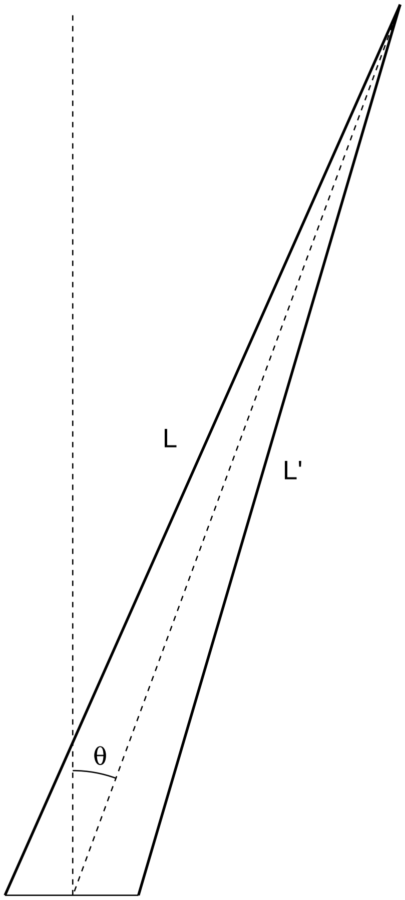
At the point along direction Y shown in the same figure, one wave has traveled ten wavelengths, and is therefore at a positive extreme, but the other has traveled only nine and a half wavelengths, so it at a negative extreme. There is perfect cancellation, so points along this line experience no wave motion.

But the distance traveled does not have to be equal in order to get constructive interference. At the point along direction Z, one wave has gone nine wavelengths and the other ten. They are both at a positive extreme.

Self-Check

At a point half a wavelength below the point marked along direction X, carry out a similar analysis.

To summarize, we will have perfect constructive interference at any point where the distance to one slit differs from the distance to the other slit by an integer number of wavelengths. Perfect destructive interference will occur when the number of wavelengths of path length difference equals an integer plus a half.



(d) The waves travel distances L_1 and L_2 from the two slits to get to the same point in space, at an angle θ from the center line.

Now we are ready to find the equation that predicts the angles of the maxima and minima. The waves travel different distances to get to the same point in space, (d). We need to find whether the waves are in phase (in step) or out of phase at this point in order to predict whether there will be constructive interference, destructive interference, or something in between.

One of our basic assumptions in this chapter is that we will only be dealing with the diffracted wave in regions very far away from the object that diffracts it, so the triangle is long and skinny. Most real-world examples with diffraction of light, in fact, would have triangles with even skinner proportions than this one. The two long sides are therefore very nearly parallel, and we are justified in drawing the right triangle shown in figure (e), labeling one leg of the right triangle as the difference in path length , *L-L'*, and labeling the acute angle as θ. (In reality this angle is a tiny bit greater than the one labeled θ in the previous figure.)

The difference in path length is related to *d* and θ by the equation

$$\frac{L - L'}{d} = \sin\theta \quad .$$

Constructive interference will result in a maximum at angles for which *L–L'* is an integer number of wavelengths,

$$L - L' = m\lambda \quad . \qquad \text{[condition for a maximum; } m \text{ is an integer]}$$

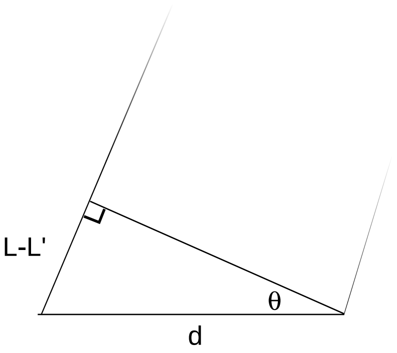
Here *m* equals 0 for the central maximum, –1 for the first maximum to its left, +2 for the second maximum on the right, etc. Putting all the ingredients together, we find $m\lambda/d = \sin\theta$, or

$$\frac{\lambda}{d} = \frac{\sin\theta}{m} \quad . \qquad \text{[condition for a maximum; } m \text{ is an integer]}$$

Similarly, the condition for a minimum is

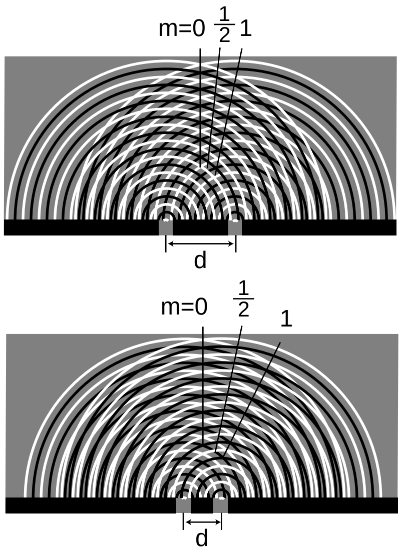
$$\frac{\lambda}{d} = \frac{\sin\theta}{m} \quad . \qquad \text{[a minimum if } m \text{ is an integer plus 1/2]}$$

That is, the minima are about halfway between the maxima.

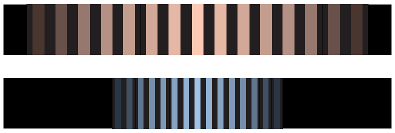


(e) A closeup of the previous figure, showing how the path length difference *L-L'* is related to *d* and to the angle φ.

At this point, both waves would have traveled nine and a half wavelengths. They would both be at a negative extreme, so there would be constructive interference.



(f) Cutting *d* in half doubles the angles of the diffraction fringes.



(g) Double-slit diffraction patterns of long-wavelength red light (top) and short-wavelength blue light (bottom).

As expected based on scaling, this equation relates angles to the unitless ratio λ/d. Alternatively, we could say that we have proven the scaling property in the special case of double-slit diffraction. It was inevitable that the result would have these scaling properties, since the whole proof was geometric, and would have been equally valid when enlarged or reduced on a photocopying machine!

Counterintuitively, this means that a diffracting object with smaller dimensions produces a bigger diffraction pattern, (f).

Example: Double-slit diffraction of blue and red light
Blue light has a shorter wavelength than red. For a given double-slit spacing *d*, the smaller value of λ/d for leads to smaller values of sin θ, and therefore to a more closely spaced set of diffraction fringes, (g)

Example: The correspondence principle
Let’s also consider how the equations for double-slit diffraction relate to the correspondence principle. When the ratio λ/d is very small, we should recover the case of simple ray optics. Now if λ/d is small, sin θ must be small as well, and the spacing between the diffraction fringes will be small as well. Although we have not proven it, the central fringe is always the brightest, and the fringes get dimmer and dimmer as we go farther from it. For small values of λ/d, the part of the diffraction pattern that is bright enough to be detectable covers only a small range of angles. This is exactly what we would expect from ray optics: the rays passing through the two slits would remain parallel, and would continue moving in the θ=0 direction. (In fact there would be images of the two separate slits on the screen, but our analysis was all in terms of angles, so we should not expect it to address the issue of whether there is structure within a set of rays that are all traveling in the θ=0 direction.)

Example: Spacing of the fringes at small angles

At small angles, we can use the approximation $\sin\theta \approx \theta$, which is valid if θ is measured in radians. The equation for double-slit diffraction becomes simply

$$\frac{\lambda}{d} = \frac{\theta}{m} \quad ,$$

which can be solved for θ to give

$$\theta = \frac{m\lambda}{d} \quad .$$

The difference in angle between successive fringes is the change in θ that results from changing m by plus or minus one,

$$\Delta\theta = \frac{\lambda}{d} \quad .$$

For example, if we write θ_7 for the angle of the seventh bright fringe on one side of the central maximum and θ_8 for the neighboring one, we have

$$\begin{aligned} \theta_8 - \theta_7 &= \frac{8\lambda}{d} - \frac{7\lambda}{d} \\ &= \frac{\lambda}{d} \quad , \end{aligned}$$

and similarly for any other neighboring pair of fringes.

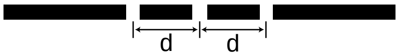
Although the equation $\lambda/d = \sin\theta/m$ is only valid for a double slit, it is still a valid guide to our thinking even if we want to make an image of a virus or a flea's leg: it is always true that

(1) large values of λ/d lead to a broad diffraction pattern, and

(2) diffraction patterns are repetitive.

In many cases the equation looks just like $\lambda/d = \sin\theta/m$ but with an extra numerical factor thrown in, and with d interpreted as some other dimension of the object, e.g. the diameter of a piece of wire.

5.6 Repetition



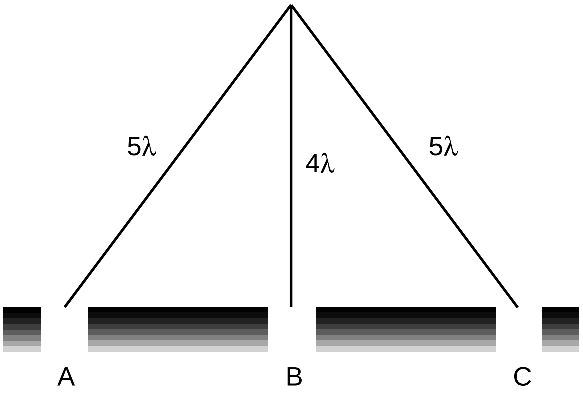
(a) A triple slit.

Suppose we replace a double slit with a triple slit, (a). We can think of this as a third *repetition* of the structures that were present in the double slit. Will this device be an improvement over the double slit for any practical reasons?

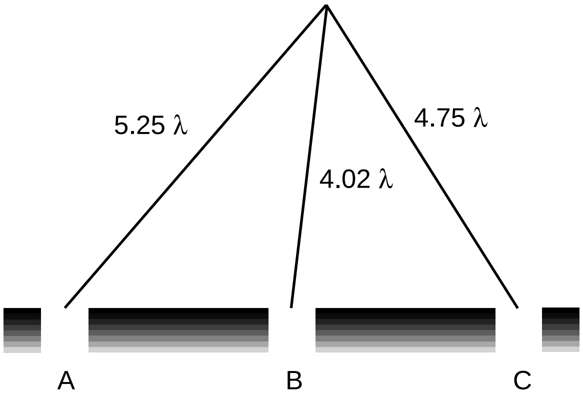
The answer is yes, as can be shown using figures (b) and (c). For ease of visualization, I have violated our usual rule of only considering points very far from the diffracting object. The scale of the drawing is such that a wavelengths is one cm. In (b), all three waves travel an integer number of wavelengths to reach the same point, so there is a bright central spot, as we would expect from our experience with the double slit. In figure (c), we show the path lengths to a new point. This point is farther from slit A by a quarter of a wavelength, and correspondingly closer to slit C. The distance from slit B has hardly changed at all. Because the paths lengths traveled from slits A and C differ from half a wavelength, there will be perfect destructive interference between these two waves. There is still some uncanceled wave intensity because of slit B, but the amplitude will be three times less than in figure (b), resulting in a factor of 9 decrease in brightness. Thus, by moving off to the right a little, we have gone from the bright central maximum to a point that is quite dark.

Now let's compare with what would have happened if slit C had been covered, creating a plain old double slit. The waves coming from slits A and B would have been out of phase by 0.23 wavelengths, but this would not have caused very severe interference. The point in figure (c) would have been quite brightly lit up.

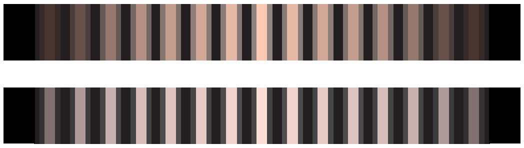
To summarize, we have found that adding a third slit narrows down the central fringe dramatically. The same is true for all the other fringes as well,



(b) There is a bright central maximum.



(c) At this point just off the central maximum, the path lengths traveled by the three waves have changed.



(d) A double-slit diffraction pattern (top), and a triple-slit pattern (bottom).

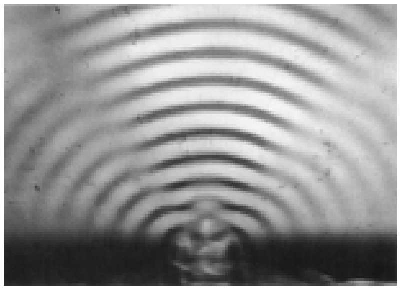
and since the same amount of energy is concentrated in narrower diffraction fringes, each fringe is brighter and easier to see, (d).

This is an example of a more general fact about diffraction: if some feature of the diffracting object is repeated, the locations of the maxima and minima are unchanged, but they become narrower.

Taking this reasoning to its logical conclusion, a diffracting object with thousands of slits would produce extremely narrow fringes. Such an object is called a diffraction grating.

5.7 Single-Slit Diffraction

If we use only a single slit, is there diffraction? If the slit is very narrow compared to a wavelength of light, then we can approximate its behavior by using only a single set of Huygens ripples. There are no other sets of ripples to add to it, so there are no constructive or destructive interference effects, and no maxima or minima. The result will be a uniform spherical wave of light spreading out in all directions, like what we would expect from a tiny lightbulb. We could call this a diffraction pattern, but it is a completely featureless one, and it could not be used, for instance, to determine the wavelength of the light, as other diffraction patterns could.



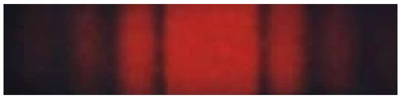
(a) Single-slit diffraction of water waves.

All of this, however, assumes that the slit is narrow compared to a wavelength of light. If, on the other hand, the slit is broader, there will indeed be interference among the sets of ripples spreading out from various points along the opening. Figure (a) shows an example with water waves, and figure (b) with light.

Self-Check

How does the wavelength of the waves compare with the width of the slit in figure (a)?

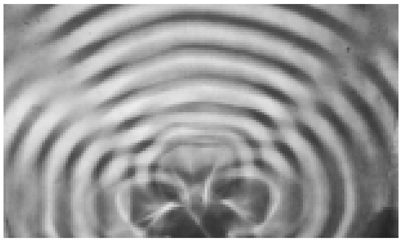
We will not go into the details of the analysis of single-slit diffraction, but let us see how its properties can be related to the general things we've learned about diffraction. We know based on scaling arguments that the angular sizes of features in the diffraction pattern must be related to the wavelength and the width, a, of the slit by some relationship of the form



(b) Single-slit diffraction of red light.

$$\frac{\lambda}{a} \leftrightarrow \theta \quad .$$

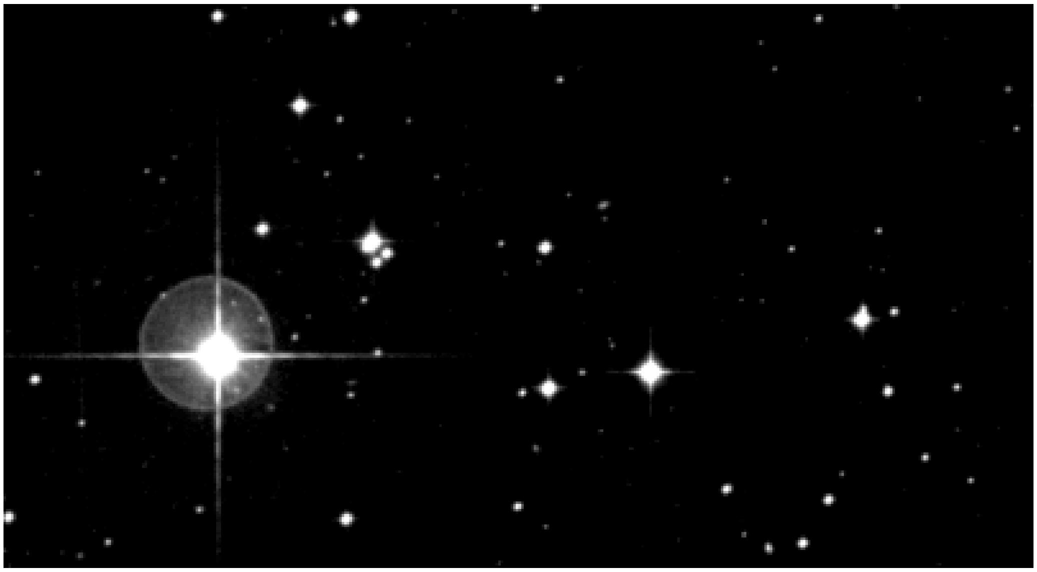
This is indeed true, and for instance the angle between the maximum of the central fringe and the maximum of the next fringe on one side equals $1.5\lambda/a$. Scaling arguments will never produce factors such as the 1.5, but they tell us that the answer must involve λ/a, so all the familiar qualitative facts are true. For instance, shorter-wavelength light will produce a more closely spaced diffraction pattern.



(c) A pretty good simulation of the single-slit pattern of figure (a), made by using three motors to produce overlapping ripples from three neighboring points in the water.

Figures a-c are uncopyrighted photos from PSSC College Physics.

Judging by the distance from one bright wave crest to the next, the wavelength appears to be about 2/3 or 3/4 as great as the width of the slit.



An important scientific example of single-slit diffraction is in telescopes. Images of individual stars, as in the figure above, are a good way to examine diffraction effects, because all stars except the sun are so far away that no telescope, even at the highest magnification, can image their disks or surface features. Thus any features of a star's image must be due purely to optical effects such as diffraction. A prominent cross appears around the brightest star, and dimmer ones surround the dimmer stars. Something like this is seen in most telescope photos, and indicates that inside the tube of the telescope there were two perpendicular struts or supports. Light diffracted around these struts. You might think that diffraction could be eliminated entirely by getting rid of all obstructions in the tube, but the circles around the stars are diffraction effects arising from single-slit diffraction at the mouth of the telescope's tube! (Actually we have not even talked about diffraction through a circular slit, but the idea is the same.) Since the angular sizes of the diffracted images depend on λ/a, the only way to improve the resolution of the images is to increase the diameter, a, of the tube. This is one of the main reasons (in addition to light-gathering power) why the best telescopes must be very large in diameter.

Self-Check

What would this imply about radio telescopes as compared with visible-light telescopes?

Discussion Question

Why is it optically impossible for bacteria to evolve eyes that use visible light to form images?

Since the wavelengths of radio waves are thousands of times longer, diffraction causes the resolution of a radio telescope to be thousands of times worse, all other things being equal.

Summary

Selected Vocabulary

diffraction .... the behavior of a wave when it encounters an obstacle or a nonuniformity in its medium; in general, diffraction causes a wave to bend around obstacles and make patterns of strong and weak waves radiating out beyond the obstacle.

coherent .... a light wave whose parts are all in phase with each other

Terminology Used in Other Books

wavelets .... the ripples in Huygens' principle

Summary

Wave optics is a more general theory of light than ray optics. When light interacts with material objects that are much larger then one wavelength of the light, the ray model of light is approximately correct, but in other cases the wave model is required.

Huygens' principle states that, given a wavefront at one moment in time, the future behavior of the wave can be found by breaking the wavefront up into a large number of small, side-by-side wave peaks, each of which then creates a pattern of circular or spherical ripples. As these sets of ripples add together, the wave evolves and moves through space. Since Huygens' principle is a purely geometrical construction, diffraction effects obey a simple scaling rule: the behavior is unchanged if the wavelength and the dimensions of the diffracting objects are both scaled up or down by the same factor. If we wish to predict the angles at which various features of the diffraction pattern radiate out, scaling requires that these angles depend only on the unitless ratio λ/d, where d is the size of some feature of the diffracting object.

Double-slit diffraction is easily analyzed using Huygens' principle if the slits are narrower than one wavelength. We need only construct two sets of ripples, one spreading out from each slit. The angles of the maxima (brightest points in the bright fringes) and minima (darkest points in the dark fringes) are given by the equation

$$\frac{\lambda}{d} = \frac{\sin\theta}{m} \quad ,$$

where d is the center-to-center spacing of the slits, and m is an integer at a maximum or an integer plus 1/2 at a minimum.

If some feature of a diffracting object is repeated, the diffraction fringes remain in the same places, but become narrower with each repetition. By repeating a double-slit pattern hundreds or thousands of times, we obtain a diffraction grating.

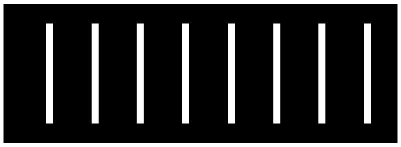
A single slit can produce diffraction fringes if it is larger than one wavelength. Many practical instances of diffraction can be interpreted as single-slit diffraction, e.g. diffraction in telescopes. The main thing to realize about single-slit diffraction is that it exhibits the same kind of relationship between λ, d, and angles of fringes as in any other type of diffraction.

Homework Problems

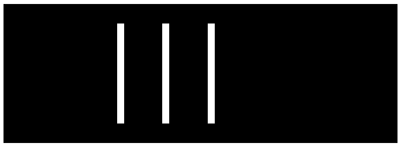
1. Why would blue or violet light be the best for microscopy?

2. Match gratings A-C with the diffraction patterns 1-3 that they produce. Explain.

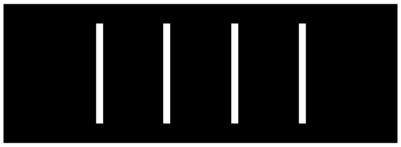
A



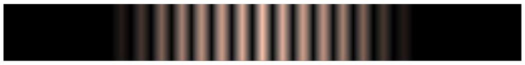
B



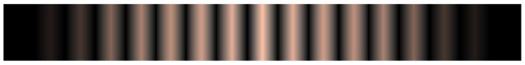
C



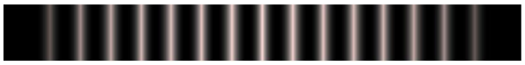
1



2



3



3 ✓. A transmission grating with 2000 lines/cm is illuminated by a beam of 694.3-nm light from a laser. Spots of light, on both sides of the undeflected beam, appear on a screen 2.0 m away. (a) How far from the central axis is either of the two nearest spots? (b) Find how much difference it makes whether you use the approximation $\sin\theta \approx \theta$.

4. When white light passes through a diffraction grating, what is the smallest value of *m* for which the visible spectrum of order *m* overlaps the next one, of order *m*+1? (The visible spectrum runs from about 400 nm to about 700 nm.)

5. Ultrasound, i.e. sound waves with frequencies too high to be audible, can be used for imaging fetuses in the womb or for breaking up kidney stones so that they can be eliminated by the body. Consider the latter application. Lenses can be built to focus sound waves, but because the wavelength of the sound is not all that small compared to the diameter of the lens, the sound will not be concentrated exactly at the geometrical focal point. Instead, a diffraction pattern will be created with an intense central spot surrounded by fainter rings. About 85% of the power is concentrated within the central spot. The angle of the first minimum (surrounding the central spot) is given by $\sin\theta = 1.22\,\lambda/b$, where *b* is the diameter of the lens. This is similar to the corresponding equation for a single slit, but with a factor of 1.22 in front which arises from the circular

S A solution is given in the back of the book. ★ A difficult problem.
✓ A computerized answer check is available. ∫ A problem that requires calculus.

shape of the aperture. Let the distance from the lens to the patient's kidney stone be L=20 cm. You will want f>20 kHz, so that the sound is inaudible. Find values of b and f that would result in a usable design, where the central spot is small enough to lie within a kidney stone 1 cm in diameter.

6. For star images such as the ones in the photo in section 5.6, estimate the angular width of the diffraction spot due to diffraction at the mouth of the telescope. Assume a telescope with a diameter of 10 meters (the largest currently in existence), and light with a wavelength in the middle of the visible range. Compare with the actual angular size of a star of diameter 10^9 m seen from a distance of 10^{17} m. What does this tell you?

Exercises

Exercise 2A: Exploring Images With a Curved Mirror

Equipment:

curved mirrors like the ones described in this chapter
curved mirrors that bulge outward (for part 6 only)

1. Obtain a curved mirror from your instructor. If it is silvered on both sides, make sure you're working with the hollowed-out side, which bends light rays inward. Look at your own face in the mirror. Now change the distance between your face and the mirror, and see what happens. How do you explain your observations?

2. With the mirror held far away from you, observe the image of something behind you, over your shoulder. Now bring your eye closer and closer to the mirror. Can you see the image with your eye very close to the mirror? Explain what's happening.

3. Now imagine the following new situation, but *don't actually do it yet*. Suppose you lay the mirror face-up on a piece of tissue paper, put your finger 5 or 10 cm or so above the mirror, and look at the image of your finger. As in part 2, you can bring your eye closer and closer to the mirror.

Write down a prediction of what will happen. Will you be able to see the image with your eye very close to the mirror?

Prediction:____

Now test your prediction. If your prediction was incorrect, can you explain your results?

4. Lay the mirror on the tissue paper, and use it to create an image of the overhead lights on a piece of paper above it. What do you have to do in order to make the image clear? Can you explain this observation?

5. Now imagine the following experiment, but *don't do it yet*. What will happen to the image on the paper if you cover half of the mirror with your hand?

Prediction:____

Test your prediction. If your prediction was incorrect, can you explain what happened?

6. Now imagine forming an image with a curved mirror that bulges outward, and that therefore bends light rays away from the central axis. Draw a typical ray diagram. Is the image real or virtual? Will there be more than one type of image?

Prediction:____

Test your prediction with the new type of mirror.

Exercise 3A: Object and Image Distances

Equipment:

- optical benches
- inbending mirrors
- illuminated objects

1. Set up the optical bench with the mirror at zero on the centimeter scale. Set up the illuminated object on the bench as well.

2. Each group will locate the image for their own value of the object distance, by finding where a piece of paper has to be placed in order to see the image on it. Note that you will have to tilt the mirror a little so that the paper on which you project the image doesn't block the light from the illuminated object.

Is the image real or virtual? How do you know? Draw a ray diagram.

3. Measure the image distance and write your result in the table on the board.

4. What do you notice about the trend of the data on the board? Draw a second ray diagram with a different object distance, and show why this makes sense.

Exercise 4A: How strong are your glasses?

This exercise was created by Dan MacIsaac.

Equipment:

eyeglasses
outbending lenses for students who don't wear glasses or use inbending glasses
rulers and metersticks
scratch paper
marking pens

Most people who wear glasses have glasses whose lenses are outbending, which allows them to focus on objects far away. Such a lens cannot form a real image, so its focal length cannot be measured as easily as that of an inbending lens. In this exercise you will determine the focal length of your own glasses by taking them off, holding them at a distance from your face, and looking through them at a set of parallel lines on a piece of paper. The lines will be reduced (the lens's magnification is less than one), and by adjusting the distance between the lens and the paper, you can make the magnification equal 1/2 exactly, so that two spaces between lines as seen through the lens fit into one space as seen simultaneously to the side of the lens. This object distance can be used in order to find the focal length of the lens.

1. Use a marker to draw three evenly spaced parallel lines on the paper. (A spacing of a few cm works well.)

2. Does this technique really measure magnification or does it measure angular magnification? What can you do in your experiment in order to make these two quantities nearly the same, so the math is simpler?

3. Before taking any numerical data, use algebra to find the focal length of the lens in terms of d_o, the object distance that results in a magnification of 1/2.

4. Measure the object distance that results in a magnification of 1/2, and determine the focal length of your lens.

Exercise 5: Single-slit diffraction

Equipment:

rulers
computer spreadsheet or computer program for adding sine waves

The following page is a diagram of a single slit and a screen onto which its diffraction pattern is projected. The class will make a numerical prediction of the intensity of the pattern at the different points on the screen. Each group will be responsible for calculating the intensity at one of the points. (Either 11 groups or six will work nicely -- in the latter case, only points a, c, e, g, i, and k are used.) The idea is to break up the wavefront in the mouth of the slit into nine parts, each of which is assumed to radiate semicircular ripples as in Huygens' principle. The wavelength of the wave is 1 cm, and we assume for simplicity that each set of ripples has an amplitude of 1 unit when it reaches the screen.

1. For simplicity, let's imagine that we were only to use two sets of ripples rather than nine. You could measure the distance from each of the two points inside the slit to your point on the screen. Suppose the distances were both 25.0 cm. What would be the amplitude of the superimposed waves at this point on the screen?

Suppose one distance was 24.0 cm and the other was 25.0 cm. What would happen?

What if one was 24.0 cm and the other was 26.0 cm?

What if one was 24.5 cm and the other was 25.0 cm?

In general, what combinations of distances will lead to completely destructive and completely constructive interference?

Can you estimate the answer in the case where the distances are 24.7 and 25.0 cm?

2. Although it is possible to calculate mathematically the amplitude of the sine wave that results from superimposing two sine waves with an arbitrary phase difference between them, the algebra is rather laborious, and it become even more tedious when we have more than two waves to superimpose. Instead, one can simply use a computer spreadsheet or some other computer program to add up the sine waves numerically at a series of points covering one complete cycle. This is what we will actually do. You just need to enter the relevant data into the computer, then examine the results and pick off the amplitude from the resulting list of numbers.

3. Measure all nine distances to your group's point on the screen, and write them on the board - that way everyone can see everyone else's data, and the class can try to make sense of why the results came out the way they did. Determine the amplitude of the combined wave, and write it on the board as well.

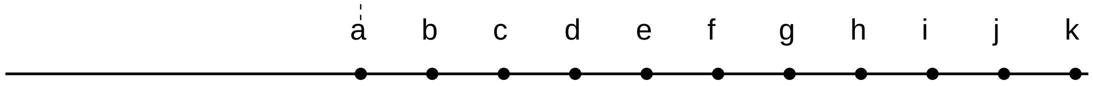
4. Why do you think the intensity at the center came out the way it did? Would it have mattered if we had used 900 sets of ripples rather than 9?

5. Looking at the raw data for the point that had the least intensity, can you see why it came out that way?

6. What do you notice about the width of the central maximum compared to the width of the first side maximum? How is this different from double-slit interference?

7. Although the pattern goes up and down, the general trend is that the farther away we get from the center, the weaker it gets. Why does it make sense that the intensity at some random angle far from the center would tend to be small?

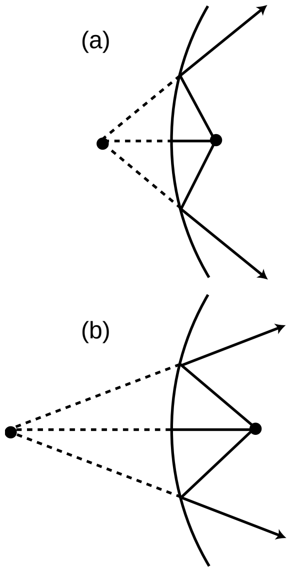
8. Single-slit diffraction can actually be calculated using equations in closed form rather than doing it numerically, and one result is that the intensity of the second maximum is always smaller than the intensity of the central maximum by a factor of $4/9\pi^2$. Note that the intensity (in units of watts per unit area) is proportional to the square of the wave's amplitude. Compare our results with the exact result.


a
b
c
d
e
f
g
h
i
j
k

Solutions to Selected Problems

Chapter 3

2. See the ray diagrams below. Increasing d_i increases d_o, so the equation $1/f = \pm 1/d_i \pm 1/d_o$ must have opposite signs on the right. Physically, we can have a virtual image with $d_i=\infty$, but not with $d_o=\infty$, so the positive sign has to be the one in front of d_o, giving $1/f = -1/d_i + 1/d_o$.



Problem 2.

Glossary

Absorption. What happens when light hits matter and gives up some of its energy.

Angular magnification. The factor by which an image's apparent angular size is increased (or decreased). Cf. magnification.

Coherent. A light wave whose parts are all in phase with each other.

Concave. Describes a surface that is hollowed out like a cave.

Convex. Describes a surface that bulges outward.

Diffraction. The behavior of a wave when it encounters an obstacle or a nonuniformity in its medium; in general, diffraction causes a wave to bend around obstacles and make patterns of strong and weak waves radiating out beyond the obstacle.

Diffuse reflection. Reflection from a rough surface, in which a single ray of light is divided up into many weaker reflected rays going in many directions.

Focal length. A property of a lens or mirror, equal to the distance from the lens or mirror to the image it forms of an object that is infinitely far away.

Image. A place where an object appears to be, because the rays diffusely reflected from any given point on the object have been bent so that they come back together and then spread out again from the image point, or spread apart as if they had originated from the image.

Index of refraction. An optical property of matter; the speed of light in a vacuum divided by the speed of light in the substance in question.

Magnification. The factor by which an image's linear size is increased (or decreased). Cf. angular magnification.

Real image. A place where an object appears to be, because the rays diffusely reflected from any given point on the object have been bent so that they come back together and then spread out again from the new point. Cf. virtual image.

Reflection. What happens when light hits matter and bounces off, retaining at least some of its energy.

Refraction. The change in direction that occurs when a wave encounters the interface between two media.

Specular reflection. Reflection from a smooth surface, in which the light ray leaves at the same angle at which it came in.

Virtual image. Like a real image, but the rays don't actually cross again; they only appear to have come from the point on the image. Cf. real image.

Index

Photo Credits

All photographs are by Benjamin Crowell, except as noted below.

Cover
Cross-section of eye: National Eye Institute, National Institutes of Health.

Chapter 4
Cross-section of eye: National Eye Institute, National Institutes of Health.
Diffraction of water wave: Uncopyrighted photo from PSSC College Physics (retouched).
Line drawing of eye: National Eye Institute, National Institutes of Health.

Chapter 5
Arecibo: Photos from above by Tony Acevedo, from below by David Parker. Courtesy of NAIC - Arecibo Observatory, a facility of thr NSF.
Star field: Space Telescope Science Institute/Digitized Sky Survey.

Useful Data

Metric Prefixes

M-	mega-	10^{6}
k-	kilo-	10^{3}
m-	milli-	10^{-3}
μ- (Greek mu)	micro-	10^{-6}
n-	nano-	10^{-9}
p-	pico-	10^{-12}
f-	femto-	10^{-15}

(Centi-, 10^{-2}, is used only in the centimeter.)

Notation and Units

quantity	**unit**	**symbol**
distance	meter, m	x, Δx
time	second, s	t, Δt
mass	kilogram, kg	m
density	kg/m^3	ρ
force	newton, 1 N=1 $\text{kg}\cdot\text{m/s}^2$	$\boldsymbol{F}$
velocity	m/s	$\boldsymbol{v}$
acceleration	m/s^2	$\boldsymbol{a}$
energy	joule, J	E
momentum	$\text{kg}\cdot\text{m/s}$	$\boldsymbol{p}$
angular momentum	$\text{kg}\cdot\text{m}^2\text{/s}$	L
period	s	T
wavelength	m	λ
frequency	s^{-1} or Hz	f
focal length	m	f
magnification	unitless	M
index of refraction	unitless	n

Fundamental Constants

gravitational constant	G=6.67x10^{-11} $\text{N}\cdot\text{m}^2/\text{kg}^2$
Coulomb constant	k=8.99x10^{9} $\text{N}\cdot\text{m}^2/\text{C}^2$
quantum of charge	e=1.60x10^{-19} C
speed of light	c=3.00x10^{8} m/s

Conversions

Conversions between SI and other units:

1 inch	=	2.54 cm (exactly)
1 mile	=	1.61 km
1 pound	=	4.45 N
(1 kg)·g	=	2.2 lb
1 gallon	=	3.78x10^{3} cm^3
1 horsepower	=	746 W
1 kcal*	=	4.18x10^{3} J

*When speaking of food energy, the word “Calorie” is used to mean 1 kcal, i.e. 1000 calories. In writing, the capital C may be used to indicate 1 Calorie=1000 calories.

Conversions between U.S. units:

1 foot	=	12 inches
1 yard	=	3 feet
1 mile	=	5280 ft

Some Indices of Refraction

substance	**index of refraction**
vacuum	1 by definition
air	1.0003
water	1.3
glass	1.5 to 1.9
diamond	2.4

Note that all indices of refraction depend on wavelength. These values are about right for the middle of the visible spectrum (yellow).

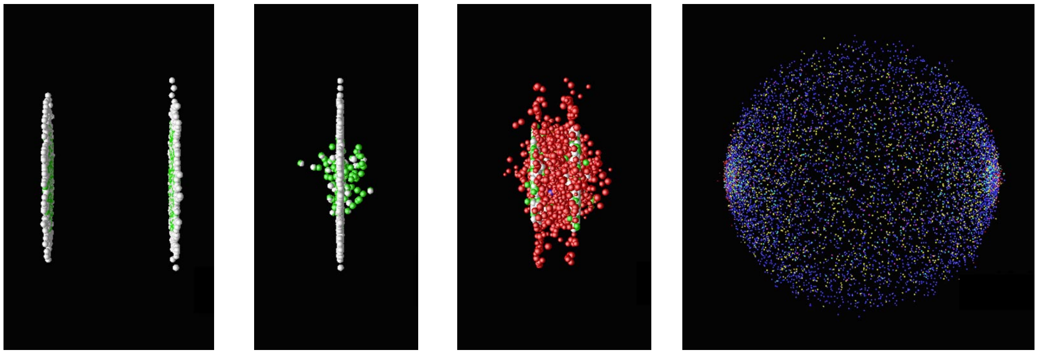
Subatomic Particles

particle	**mass (kg)**	**charge**	**radius (fm)**
electron	9.109x10^{-31}	$-e$	$<\sim$0.01
proton	1.673x10^{-27}	$+e$	~1.1
neutron	1.675x10^{-27}	0	~1.1
neutrino	~10^{-39} kg?	0	?

The radii of protons and neutrons can only be given approximately, since they have fuzzy surfaces. For comparison, a typical atom is about a million fm in radius.

The Modern Revolution in Physics

Benjamin Crowell



Book 6 in the **Light and Matter** series of introductory physics textbooks
www.lightandmatter.com

The Modern Revolution in Physics

The **Light and Matter** series of introductory physics textbooks:

1 Newtonian Physics

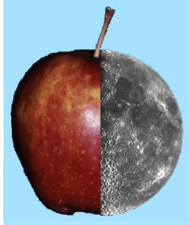
2 Conservation Laws

3 Vibrations and Waves

4 Electricity and Magnetism

5 Optics

6 The Modern Revolution in Physics



The Modern Revolution in Physics

Benjamin Crowell

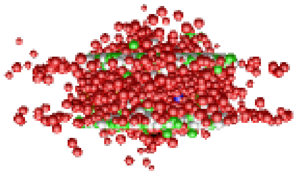
www.lightandmatter.com

Light and Matter

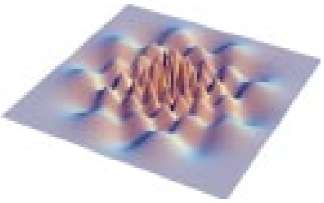
Fullerton, California
www.lightandmatter.com

Edition 2.0
rev. 2000-10-08

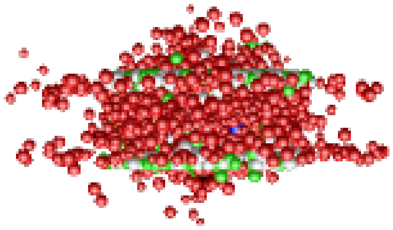
ISBN 0-9704670-6-0

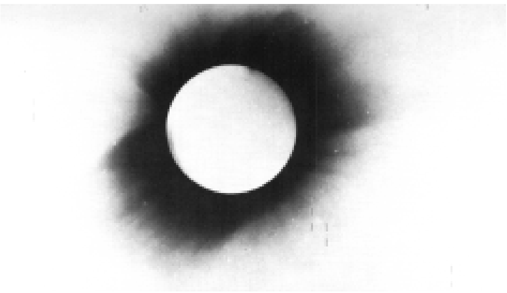


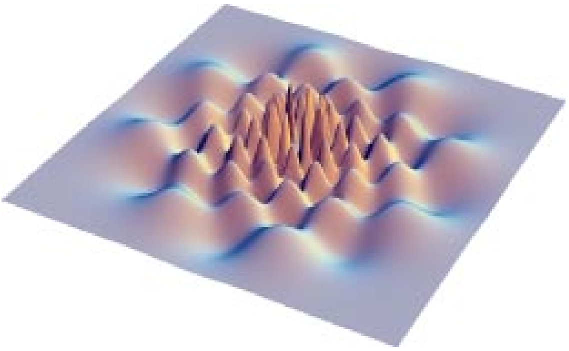
Brief Contents



Contents







Albert Einstein in his days as a Swiss patent clerk, when he developed his theory of relativity.

1 Relativity, Part I

Complaining about the educational system is a national sport among professors in the U.S., and I, like my colleagues, am often tempted to imagine a golden age of education in our country's past, or to compare our system unfavorably with foreign ones. Reality intrudes, however, when my immigrant students recount the overemphasis on rote memorization in their native countries and the philosophy that what the teacher says is always right, even when it's wrong.

Albert Einstein's education in late-nineteenth-century Germany was neither modern nor liberal. He did well in the early grades (the myth that he failed his elementary-school classes comes from a misunderstanding based on a reversal of the German numerical grading scale), but in high school and college he began to get in trouble for what today's edspeak calls "critical thinking."

Indeed, there was much that deserved criticism in the state of physics at that time. There was a subtle contradiction between Maxwell's theory of electromagnetism and Galileo's principle that all motion is relative. Einstein began thinking about this on an intuitive basis as a teenager, trying to imagine what a light beam would look like if you could ride along beside it on a motorcycle at the speed of light. Today we remember him most of all for his radical and far-reaching solution to this contradiction, his theory of relativity, but in his student years his insights were greeted with derision from his professors. One called him a "lazy dog." Einstein's distaste for authority was typified by his decision as a teenager to renounce his German citizenship and become a stateless person, based purely on his opposition to the militarism and repressiveness of German society. He spent his most productive scientific years in Switzerland and Berlin, first as a patent clerk but later as a university professor. He was an outspoken pacifist and a stubborn opponent of World War I, shielded from retribution by his eventual acquisition of Swiss citizenship.

As the epochal nature of his work began to become evident, some liberal Germans began to point to him as a model of the “new German,” but with the Nazi coup d’etat, staged public meetings began to be held at which Nazi scientists criticized the work of this ethnically Jewish (but spiritually nonconformist) giant of science. Einstein had the good fortune to be on a stint as a visiting professor at CalTech when Hitler was appointed chancellor, and so escaped the Holocaust. World War II convinced Einstein to soften his strict pacifist stance, and he signed a secret letter to President Roosevelt urging research into the building of a nuclear bomb, a device that could not have been imagined without his theory of relativity. He later wrote, however, that when Hiroshima and Nagasaki were bombed, it made him wish he could burn off his own fingers for having signed the letter.

This chapter and the next are specifically about Einstein’s theory of relativity, but Einstein also began a second, parallel revolution in physics known as the quantum theory, which stated, among other things, that certain processes in nature are inescapably random. Ironically, Einstein was an outspoken doubter of the new quantum ideas, being convinced that “the Old One [God] does not play dice with the universe,” but quantum and relativistic concepts are now thoroughly intertwined in physics. The remainder of this book beyond the present pair of chapters is an introduction to the quantum theory, but we will continually be led back to relativistic ideas.

1.1 The Principle of Relativity

> Absolute, true, and mathematical time...flows at a constant rate without relation to anything external... Absolute space...without relation to anything external, remains always similar and immovable.
>
> Isaac Newton (tr. Andrew Motte)

Relativity according to Galileo and Einstein

Galileo’s most important physical discovery was that motion is relative. With modern hindsight, we restate this in a way that shows what made the teenage Einstein suspicious:

> **The Principle of Galilean Relativity**
> Matter obeys the same laws of physics in any inertial frame of reference, regardless of the frame’s orientation, position, or constant-velocity motion.

If this principle was violated, then experiments would have different results in a moving laboratory than in one at rest. The results would allow us to decide which lab was in a state of absolute rest, contradicting the idea that motion is relative. The new way of saying it thus appears equivalent to the old one, and therefore not particularly revolutionary, but note that it only refers to matter, not light.

Einstein’s professors taught that light waves obeyed an entirely different set of rules than material objects. They believed that light waves were a vibration of a mysterious medium called the ether, and that the speed of light should be interpreted as a speed relative to this ether. Even though Maxwell's treatment of electromagnetism made no reference to any ether, they could not conceive of a wave that was not a vibration of some medium. Thus although the cornerstone of the study of matter had for two centuries been the idea that motion is relative, the science of light seemed to contain a concept that certain frames of reference were in an absolute state of rest with respect to the ether, and were therefore to be preferred over moving frames.

Now let's think about Albert Einstein's daydream of riding a motorcycle alongside a beam of light. In cyclist Albert's frame of reference, the light wave appears to be standing still. He can stick measuring instruments into the wave to monitor the electric and magnetic fields, and they will be constant at any given point. This, however, violates Maxwell's theory of electromagnetism: an electric field can only be caused by charges or by time-varying magnetic fields. Neither is present in the cyclist's frame of reference, so why is there an electric field? Likewise, there are no currents or time-varying electric fields that could serve as sources of the magnetic field.

Einstein could not tolerate this disagreement between the treatment of relative and absolute motion in the theories of matter on the one hand and light on the other. He decided to rebuild physics with a single guiding principle:

> **Einstein's Principle of Relativity**
> Both light and matter obey the same laws of physics in any inertial frame of reference, regardless of the frame's orientation, position, or constant-velocity motion.

Maxwell's equations are the basic laws of physics governing light, and Maxwell's equations predict a specific value for the speed of light, c=3.0x10^8 m/s, so this new principle implies that *the speed of light must be the same in all frames of reference*.

1.2 Distortion of Time and Space

This is hard to swallow. If a dog is running away from me at 5 m/s relative to the sidewalk, and I run after it at 3 m/s, the dog's velocity in my frame of reference is 2 m/s. According to everything we have learned about motion, the dog *must* have different speeds in the two frames: 5 m/s in the sidewalk's frame and 2 m/s in mine. How, then, can a beam of light have the same speed as seen by someone who is chasing the beam?

In fact the strange constancy of the speed of light had shown up in the now-famous Michelson-Morley experiment of 1887. Michelson and Morley set up a clever apparatus to measure any difference in the speed of light beams traveling east-west and north-south. The motion of the earth around the sun at 110,000 km/hour (about 0.01% of the speed of light) is to our west during the day. Michelson and Morley believed in the ether hypothesis, so they expected that the speed of light would be a fixed value relative to the ether. As the earth moved through the ether, they thought they would observe an effect on the velocity of light along an east-west line. For instance, if they released a beam of light in a westward direction during the day, they expected that it would move away from them at less than the normal speed because the earth was chasing it through the ether. They were surprised when they found that the expected 0.01% change in the speed of light did not occur.

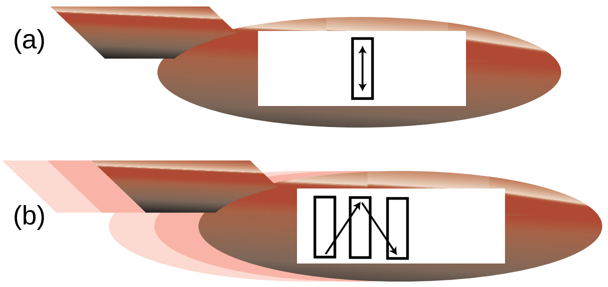
Although the Michelson-Morley experiment was nearly two decades in the past by the time Einstein published his first paper on relativity in 1905, he did not even know of the experiment until after submitting the paper. At this time he was still working at the Swiss patent office, and was isolated from the mainstream of physics.

How did Einstein explain this strange refusal of light waves to obey the usual rules of addition and subtraction of velocities due to relative motion? He had the originality and bravery to suggest a radical solution. He decided that space and time must be stretched and compressed as seen by observers in different frames of reference. Since velocity equals distance divided by time, an appropriate distortion of time and space could cause the speed of light to come out the same in a moving frame. This conclusion could have been reached by the physicists of two generations before, the day after Maxwell published his theory of light, but the attitudes about absolute space and time stated by Newton were so strongly ingrained that such a radical approach did not occur to anyone before Einstein.

An example of time distortion

Consider the situation shown in figures (a) and (b). Aboard a rocket ship we have a tube with mirrors at the ends. If we let off a flash of light at the bottom of the tube, it will be reflected back and forth between the top and bottom. It can be used as a clock: by counting the number of times the light goes back and forth we get an indication of how much time has passed. (This may not seem very practical, but a real atomic clock does work by essentially the same principle.) Now imagine that the rocket is cruising at a significant fraction of the speed of light relative to the earth. Motion is relative, so for a person inside the rocket, (a), there is no detectable change in the behavior of the clock, just as a person on a jet plane can toss a ball up and down without noticing anything unusual. But to an observer in the earth's frame of reference, the light appears to take a zigzag path through space, (b), increasing the distance the light has to travel.

If we didn't believe in the principle of relativity, we could say that the



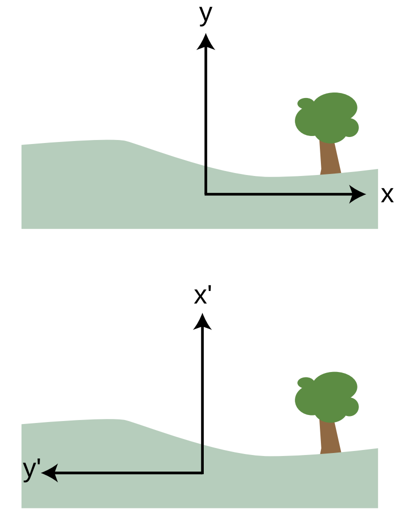
light just goes faster according to the earthbound observer. Indeed, this would be correct if the speeds were not close to the speed of light, and if the thing traveling back and forth was, say, a ping-pong ball. But according to the principle of relativity, the speed of light must be the same in both frames of reference. We are forced to conclude that time is distorted, and the light-clock appears to run more slowly than normal as seen by the earthbound observer. In general, a clock appears to run most quickly for observers who are in the same state of motion as the clock, and runs more slowly as perceived by observers who are moving relative to the clock.

Coordinate transformations

Speed relates to distance and time, so if the speed of light is the same in all frames of reference and time is distorted for different observers, presumably distance is distorted as well: otherwise the ratio of distance to time could not stay the same. Handling the two effects at the same time requires delicacy. Let's start with a couple of examples that are easier to visualize.

Rotation

For guidance, let's look at the mathematical treatment of a different part of the principle of relativity, the statement that the laws of physics are the same regardless of the orientation of the coordinate system. Suppose that two observers are in frames of reference that are at rest relative to each other, and they set up coordinate systems with their origins at the same point, but rotated by 90 degrees, as in figure (c). To go back and forth between the two systems, we can use the equations



(c) Two observers describe the same landscape with different coordinate systems.

$$\begin{aligned} x' &= y \\ y' &= -x \end{aligned}$$

A set of equations such as this one for changing from one system of coordinates to another is called a coordinate transformation, or just a transformation for short.

Similarly, if the coordinate systems differed by an angle of 5 degrees, we would have

$$\begin{aligned} x' &= (\cos 5°)\, x + (\sin 5°)\, y \\ y' &= (-\sin 5°)\, x + (\cos 5°)\, y \end{aligned}$$

Since cos 5°=0.997 is very close to one, and sin 5°=0.087 is close to zero, the rotation through a small angle has only a small effect, which makes sense. The equations for rotation are always of the form

$$\begin{aligned} x' &= (\text{constant \#1})\, x + (\text{constant \#2})\, y \\ y' &= (\text{constant \#3})\, x + (\text{constant \#4})\, y \quad . \end{aligned}$$

Galilean transformation for frames moving relative to each other

Einstein wanted to see if he could find a rule for changing between coordinate systems that were moving relative to each other. As a second warming-up example, let's look at the transformation between frames of reference in relative motion according to *Galilean* relativity, i.e. without any distortion of space and time. Suppose the x' axis is moving to the right at a speed v relative to the x axis. The transformation is simple:

$$\begin{aligned} x' &= x - vt \\ t' &= t \end{aligned}$$

Again we have an equation with constants multiplying the variables, but now the variables are distance and time. The interpretation of the $-vt$ term is the observer moving with the origin x' system sees a steady reduction in distance to an object on the right and at rest in the x system. In other words, the object appears to be moving according to the x' observer, but at rest according to x. The fact that the constant in front of x in the first equation equals one tells us that there is no distortion of space according to Galilean relativity, and similarly the second equation tells us there is no distortion of time.

Einstein's transformations for frames in relative motion

Guided by analogy, Einstein decided to look for a transformation between frames in relative motion that would have the form

$$\begin{aligned} x' &= Ax + Bt \\ t' &= Cx + Dt \ . \end{aligned}$$

(Any form more complicated than this, for example equations including x^2 or t^2 terms, would violate the part of the principle of relativity that says the laws of physics are the same in all locations.) The constants A, B, C, and D would depend only on the relative velocity, v, of the two frames. Galilean relativity had been amply verified by experiment for values of v much less than the speed of light, so at low speeds we must have $A\approx1$, $B\approx v$, $C\approx0$, and $D\approx1$. For high speeds, however, the constants A and D would start to become measurably different from 1, providing the distortions of time and space needed so that the speed of light would be the same in all frames of reference.

Self-Check

What units would the constants A, B, C, and D need to have?

Natural units

Despite the reputation for difficulty of Einstein's theories, the derivation of Einstein's transformations is fairly straightforward. The algebra, however, can appear more cumbersome than necessary unless we adopt a choice of units that is better adapted to relativity than the metric units of meters and seconds. The form of the transformation equations shows that time and

A relates distance to distance, so it is unitless, and similarly for D. Multiplying B by a time has to give a distance, so B has units of m/s. Multiplying C by distance has to give a time, so C has units of s/m.

space are not entirely separate entities. Life is easier if we adopt a new set of units:

> *Time* is measured in *seconds*.
>
> *Distance* is also measured in units of *seconds*. A distance of one second is how far light travels in one second of time.

In these units, the speed of light equals one by definition:

$$c = \frac{\text{1 second of distance}}{\text{1 second of time}} = 1$$

All velocities are represented by unitless numbers in this system, so for example v=0.5 would describe an object moving at half the speed of light.

Derivation of the transformations

To find how the constants A, B, C, and D in the transformation equations

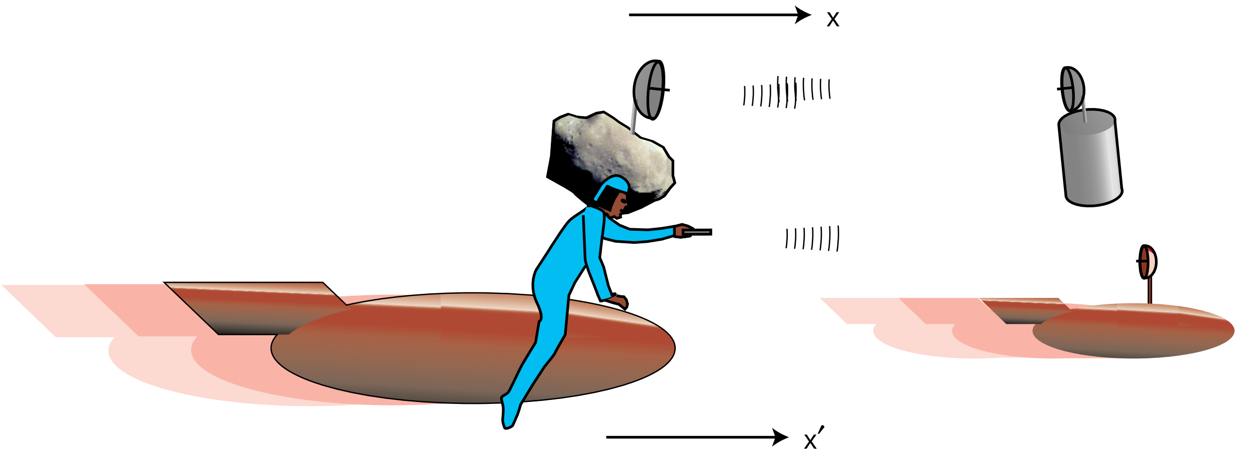
$$x' = Ax + Bt \qquad (1a)$$

$$t' = Cx + Dt \qquad (1b)$$

depend on velocity, we follow a strategy of relating the constants to one another by requiring that the transformation produce the right results in several different situations. By analogy, the rotation transformation for x and y coordinates has the same constants on the upper left and lower right, and the upper right and lower left constants are equal in absolute value but opposite in sign. We will look for similar rules for the frames-in-relative-motion transformations.

For vividness, we imagine that the x,t frame is defined by an asteroid at x=0, and the x',t' frame by a rocket ship at x'=0. The rocket ship is coasting at a constant speed v relative to the asteroid, and as it passes the asteroid they synchronize their clocks to read t=0 and t'=0.

We need to compare the perception of space and time by observers on the rocket and the asteroid, but this can be a bit tricky because our usual ideas about measurement contain hidden assumptions. If, for instance, we want to measure the length of a box, we imagine we can lay a ruler down on it, take in the scene visually, and take the measurement using the ruler's scale on the right side of the box while the left side of the box is simultaneously lined up with the butt of the ruler. The assumption that we can take in the whole scene at once with our eyes is, however, based on the



assumption that light travels with infinite speed to our eyes. Since we will be dealing with relative motion at speeds comparable to the speed of light, we have to spell out our methods of measuring distance.

We will therefore imagine an explicit procedure for the asteroid and the rocket pilot to make their distance measurements: they send electromagnetic signals (light or radio waves) back and forth to their own remote stations. For instance the asteroid's station will send it a message to tell it the time at which the rocket went by. The asteroid's station is at rest with respect to the asteroid, and the rocket's is at rest with respect to the rocket (and therefore in motion with respect to the asteroid).

The measurement of time is likewise fraught with danger if we are careless, which is why we have had to spell out procedures for the synchronization of clocks between the asteroid and the rocket. The asteroid must also synchronize its clock with its remote stations's clock by adjusting them until flashes of light released by both the asteroid and its station at equal clock readings are received on the opposite sides at equal clock readings. The rocket pilot must go through the same kind of synchronization procedure with her remote station.

Rocket's motion as seen by the asteroid

The origin of the rocket's x',t' frame is defined by the rocket itself, so the rocket always has $x'=0$. Let the asteroid's remote station be at position x in the asteroid's frame. The asteroid sees the rocket travel at speed v, so the asteroid's remote station sees the rocket pass it when x equals vt. Equation (1a) becomes $0=Avt+Bt$, which implies a relationship between A and B: $B/A=-v$. (In the Galilean version, we had $B=-v$ and $A=1$.) This restricts the transformation to the form

$$x' = Ax - Avt \qquad (2a)$$
$$t' = Cx + Dt \qquad (2b)$$

Asteroid's motion as seen by the rocket

Straightforward algebra can be used to reverse the transformation equations so that they give x and t in terms of x' and t'. The result for x is $x=(Dx'-Bt')/(AD-BC)$. The asteroid's frame of reference has its origin defined by the asteroid itself, so the asteroid is always at $x=0$. In the rocket's frame, the asteroid falls behind according to the equation $x'=-vt'$, and substituting this into the equation for x gives $0=(Dvt'-Bt')/(AD-BC)$. This requires us to have $B/D=-v$, i.e. D must be the same as A:

$$x' = Ax - Avt \qquad (3a)$$
$$t' = Cx + At \qquad (3b)$$

Agreement on the speed of light

Suppose the rocket pilot releases a flash of light in the forward direction as she passes the asteroid at $t=t'=0$. As seen in the asteroid's frame, we might expect this pulse to travel forward faster than normal because it was emitted by the moving rocket, but the principle of relativity tells us this is not so. The flash reaches the asteroid's remote station when x equals ct, and since we are working in natural units, this is equivalent to $x=t$. The speed of light must be the same in the rocket's frame, so we must also have $x'=t'$ when the flash gets there. Setting equations (3a) and (3b) equal to each other and substituting in $x=t$, we find $A-Av=C+A$, so we must have $C=-Av$:

$$x' = Ax - Avt \tag{4a}$$

$$t' = -Avx + At \tag{4b}$$

We have now determined the whole form of the transformation except for an overall multiplicative constant A.

Reversal of velocity

We can tie down this last unknown by considering what would have happened if the velocity of the rocket had been reversed. This would be entirely equivalent to reversing the direction of time, like playing a movie backwards, and it would also be equivalent to interchanging the roles of the rocket and the asteroid, since the rocket pilot sees the asteroid moving away from her to the left. The reversed transformation from the x',t' system to the x,t system must therefore be the one obtained by reversing the signs of t and t':

$$x = Ax' + Avt' \tag{5a}$$

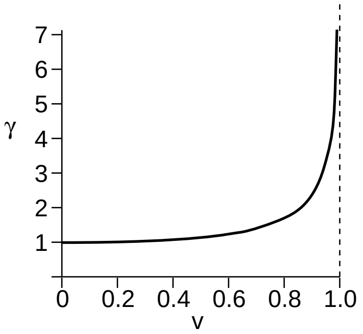
$$-t = -Avx' - At' \tag{5b}$$

We now substitute equations 4a and 4b into equation 5a to eliminate x' and t', leaving only x and t:

$$x = A(Ax-Avt) + Av(-Avx+At)$$

The t terms cancel out, and collecting the x terms we find

$$x = A^2(1-v^2)x \quad ,$$



which requires $A^2(1-v^2)=1$, or $A=1/\sqrt{1-v^2}$. Since this factor occurs so often, we give it a special symbol, γ, the Greek letter gamma,

$$\gamma = \frac{1}{\sqrt{1-v^2}} \qquad \text{[definition of the } \gamma \text{ factor]}$$

Its behavior is shown in the graph on the left.

We have now arrived at the correct relativistic equation for transforming between frames in relative motion. For completeness, I will include, without proof, the trivial transformations of the y and z coordinates.

$$\boxed{\begin{array}{rcrcr} x' & = & \gamma x & - & \gamma vt \\ t' & = & -\gamma vx & + & \gamma t \\ y' & = & y & & \\ z' & = & z & & \end{array}}$$

[transformation between frames in relative motion; v is the velocity of the x' frame relative to the x frame; the origins of the frames are assumed to have coincided at $x=x'=0$ and $t=t'=0$]

Self-Check

What is γ when v=0? Interpret the transformation equations in the case of v=0.

Discussion Question

A. If you were in a spaceship traveling at the speed of light (or extremely close to the speed of light), would you be able to see yourself in a mirror?

B. A person in a spaceship moving at 99.99999999% of the speed of light relative to Earth shines a flashlight forward through dusty air, so the beam is visible. What does she see? What would it look like to an observer on Earth?

Looking at the definition of γ, we see that γ=1 when v=0. The transformation equations then reduce to $x'=x$ and $t'=t$, which makes sense.

1.3 Applications

We now turn to the subversive interpretations of these equations.

Nothing can go faster than the speed of light.

Remember that these equations are expressed in natural units, so v=0.1 means motion at 10% of the speed of light, and so on. What happens if we want to send a rocket ship off at, say, twice the speed of light, v=2? Then γ will be $1/\sqrt{-3}$. But your math teacher has always cautioned you about the severe penalties for taking the square root of a negative number. The result would be physically meaningless, so we conclude that no object can travel faster than the speed of light. Even travel exactly at the speed of light appears to be ruled out for material objects, since then γ would be infinite.

Einstein had therefore found a solution to his original paradox about riding on a motorcycle alongside a beam of light, resulting in a violation of Maxwell's theory of electromagnetism. The paradox is resolved because it is impossible for the motorcycle to travel at the speed of light.

Most people, when told that nothing can go faster than the speed of light, immediately begin to imagine methods of violating the rule. For instance, it would seem that by applying a constant force to an object for a long time, we would give it a constant acceleration which would eventually result in its traveling faster than the speed of light. We will take up these issues in section 2.2.

No absolute time

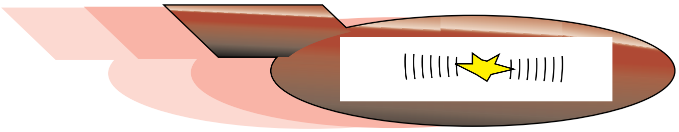
The fact that the equation for time is not just $t'=t$ tells us we're not in Kansas anymore — Newton's concept of absolute time is dead. One way of understanding this is to think about the steps described for synchronizing the four clocks:

(1) The asteroid's clock — call it A1 — was synchronized with the clock on its remote station, A2.

(2) The rocket pilot synchronized her clock, R1, with A1, at the moment when she passed the asteroid.

(3) The clock on the rocket's remote station, R2, was synchronized with R1.

Now if A2 matches A1, A1 matches R1, and R1 matches R2, we would expect A2 to match R2. This cannot be so, however. The rocket pilot released a flash of light as she passed the asteroid. In the asteroid's frame of reference, that light had to travel the full distance to the asteroid's remote station before it could be picked up there. In the rocket pilot's frame of reference, however, the asteroid's remote station is rushing at her, perhaps at a sizeable fraction of the speed of light, so the flash has less distance to travel before the asteroid's station meets it. Suppose the rocket pilot sets things up so that R2 has just enough of a head start on the light flash to reach A2 at the same time the flash of light gets there. Clocks A2 and R2 cannot agree, because the time required for the light flash to get there was different in the two frames. Thus, two clocks that were initially in agreement will disagree later on.

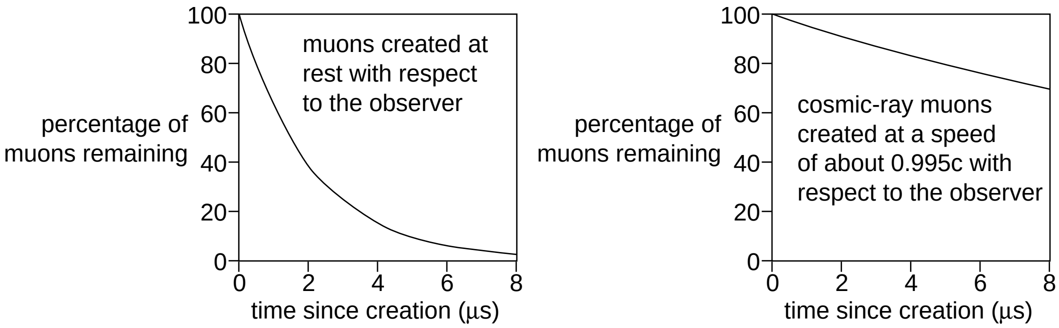


No simultaneity

Part of the concept of absolute time was the assumption that it was valid to say things like, “I wonder what my uncle in Beijing is doing right now.” In the nonrelativistic world-view, clocks in Los Angeles and Beijing could be synchronized and stay synchronized, so we could unambiguously define the concept of things happening simultaneously in different places. It is easy to find examples, however, where events that seem to be simultaneous in one frame of reference are not simultaneous in another frame. In the figure above, a flash of light is set off in the center of the rocket’s cargo hold. According to a passenger on the rocket, the flashes have equal distances to travel to reach the front and back walls, so they get there simultaneously. But an outside observer who sees the rocket cruising by at high speed will see the flash hit the back wall first, because the wall is rushing up to meet it, and the forward-going part of the flash hit the front wall later, because the wall was running away from it. Only when the relative velocity of two frames is small compared to the speed of light will observers in those frames agree on the simultaneity of events.

Time dilation

Let’s compare the rate at which time passes in two frames. A clock that stays on the asteroid will always have x=0, so the time transformation equation $t'=-v\gamma x+\gamma t$ becomes simply $t'=\gamma t$. If the rocket pilot monitors the ticking of a clock on the asteroid via radio (and corrects for the increasingly long delay for the radio signals to reach her as she gets farther away from it), she will find that the rate of increase of the time t' on her wristwatch is always greater than the rate at which the time t measured by the asteroid’s clock increases. It will seem to her that the asteroid’s clock is running too slowly by a factor of γ. This is known as the time dilation effect: clocks seem to run fastest when they are at rest relative to the observer, and more slowly when they are in motion. The situation is entirely symmetric: to people on the asteroid, it will appear that the rocket pilot’s clock is the one that is running too slowly.



Example: Cosmic-ray muons

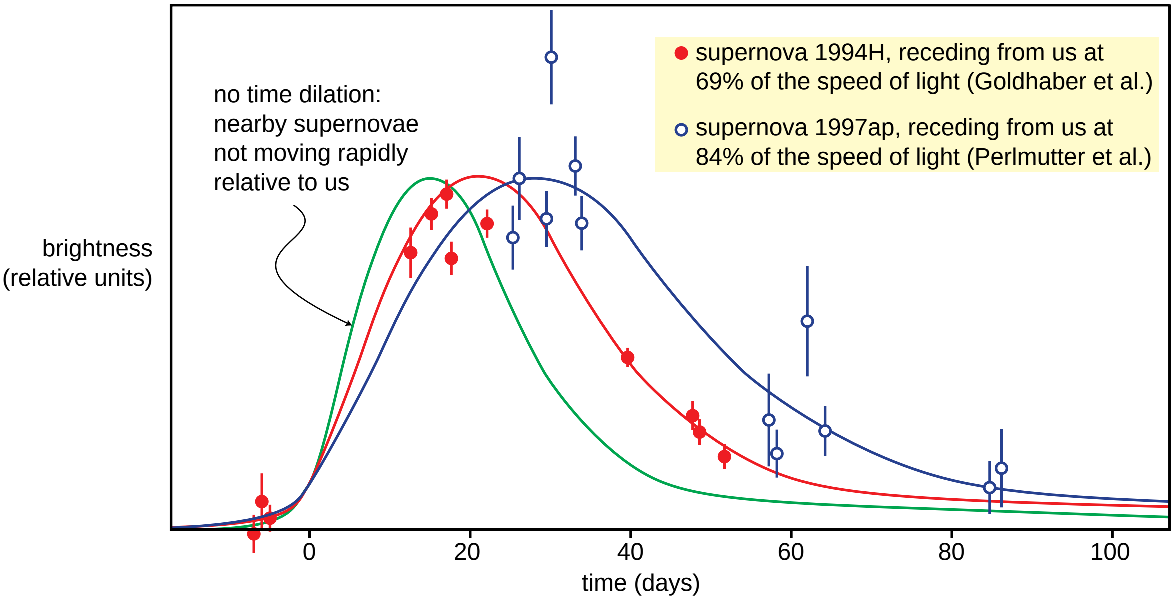
Cosmic rays are protons and other atomic nuclei from outer space. When a cosmic ray happens to come the way of our planet, the first earth-matter it encounters is an air molecule in the upper atmosphere. This collision then creates a shower of particles that cascade downward and can often be detected at the earth's surface. One of the more exotic particles created in these cosmic ray showers is the muon (named after the Greek letter mu, μ). The reason muons are not a normal part of our environment is that a muon is radioactive, lasting only 2.2 microseconds on the average before changing itself into an electron and two neutrinos. A muon can therefore be used as a sort of clock, albeit a self-destructing and somewhat random one! The graphs above show the average rate at which a sample of muons decays, first for muons created at rest and then for high-velocity muons created in cosmic-ray showers. The second graph is found experimentally to be stretched out by a factor of about ten, which matches well with the prediction of relativity theory:

$$\begin{aligned} \gamma &= 1 / \sqrt{1-v^2} \\ &= 1 / \sqrt{1-0.995^2} \\ &\approx 10 \end{aligned}$$

Since a muon takes many microseconds to pass through the atmosphere, the result is a marked increase in the number of muons that reach the surface.

Example: Time dilation for objects larger than the atomic scale

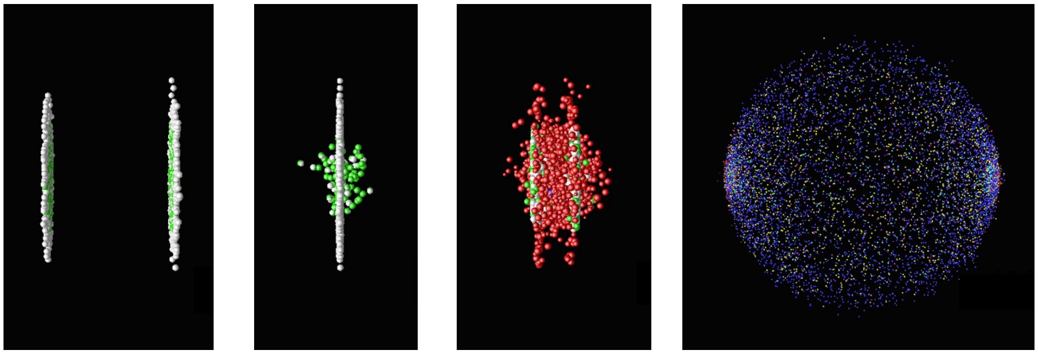
Our world is (fortunately) not full of human-scale objects moving at significant speeds compared to the speed of light. For this reason, it took over 80 years after Einstein's theory was published before anyone could come up with a conclusive example of drastic time dilation that wasn't confined to cosmic rays or particle accelerators. Recently, however, astronomers have found definitive proof that entire stars undergo time dilation. The universe is expanding in the aftermath of the Big Bang, so in general everything in the universe is getting farther away from everything else. One need only find an astronomical process that takes a standard amount of time, and then observe how long it appears to take when it occurs in a part of the universe that is receding from us rapidly. A type of exploding star called a type Ia supernova fills the bill, and technology is now sufficiently advanced to allow them to be detected across vast distances. The graph on the following page shows convincing evidence for time dilation in the brightening and dimming of two distant supernovae.



The twin paradox

A natural source of confusion in understanding the time-dilation effect is summed up in the so-called twin paradox, which is not really a paradox. Suppose there are two teenaged twins, and one stays at home on earth while the other goes on a round trip in a spaceship at relativistic speeds (i.e. speeds comparable to the speed of light, for which the effects predicted by the theory of relativity are important). When the traveling twin gets home, he has aged only a few years, while his brother is now old and gray. (Robert Heinlein even wrote a science fiction novel on this topic, although it is not one of his better stories.)

The paradox arises from an incorrect application of the theory of relativity to a description of the story from the traveling twin's point of view. From his point of view, the argument goes, his homebody brother is the one who travels backward on the receding earth, and then returns as the earth approaches the spaceship again, while in the frame of reference fixed to the spaceship, the astronaut twin is not moving at all. It would then seem that the twin on earth is the one whose biological clock should tick more slowly, not the one on the spaceship. The flaw in the reasoning is that the principle of relativity only applies to frames that are in motion at constant velocity relative to one another, i.e. inertial frames of reference. The astronaut twin's frame of reference, however, is noninertial, because his spaceship must accelerate when it leaves, decelerate when it reaches its destination, and then repeat the whole process again on the way home. What we have been studying is Einstein's special theory of relativity, which describes motion at constant velocity. To understand accelerated motion we would need the general theory of relativity (which is also a theory of gravity). A correct treatment using the general theory shows that it is indeed the traveling twin who is younger when they are reunited.



Length contraction

The treatment of space and time in the transformation between frames is entirely symmetric, so distance intervals as well as time intervals must be reduced by a factor of γ for an object in a moving frame. The figure above shows a an artist's rendering of this effect for the collision of two gold nuclei at relativistic speeds in the RHIC accelerator in Long Island, New York, scheduled to come on line in 2000. The gold nuclei would appear nearly spherical (or just slightly lengthened like an American football) in frames moving along with them, but in the laboratory's frame, they both appear drastically foreshortened as they approach the point of collision. The later pictures show the nuclei merging to form a hot soup, in which experimenters hope to observe a new form of matter.

Perhaps the most famous of all the so-called relativity paradoxes involves the length contraction. The idea is that one could take a schoolbus and drive it at relativistic speeds into a garage of ordinary size, in which it normally would not fit. Because of the length contraction, the bus would supposedly fit in the garage. The paradox arises when we shut the door and then quickly slam on the brakes of the bus. An observer in the garage's frame of reference will claim that the bus fit in the garage because of its contracted length. The driver, however, will perceive the garage as being contracted and thus even less able to contain the bus than it would normally be. The paradox is resolved when we recognize that the concept of fitting the bus in the garage "all at once" contains a hidden assumption, the assumption that it makes sense to ask whether the front and back of the bus can simultaneously be in the garage. Observers in different frames of reference moving at high relative speeds do not necessarily agree on whether things happen simultaneously. The person in the garage's frame can shut the door at an instant he perceives to be simultaneous with the front bumper's arrival at the opposite wall of the garage, but the driver would not agree about the simultaneity of these two events, and would perceive the door as having shut long after she plowed through the back wall.

Discussion Questions

A. A question that students often struggle with is whether time and space can really be distorted, or whether it just seems that way. Compare with optical illusions are magic tricks. How do we know that these illusions are not real? Are relativistic effects the same or not?

B. On a spaceship moving at relativistic speeds, would a lecture seem even longer and more boring than normal?

C. Mechanical clocks can be affected by motion. For example, it was a significant technological achievement to build a clock that could sail aboard a ship and still keep accurate time, allowing longitude to be determined. How is this similar to or different from relativistic time dilation?

D. What would the shapes of the two nuclei in the RHIC experiment look like to a microscopic observer riding on the left-hand nucleus? To an observer riding on the right-hand one? Can they agree on what is happening? If not, why not — after all, shouldn't they see the same thing if they both compare the two nuclei side-by-side at the same instant in time?

E. If you stick a piece of foam rubber out the window of your car while driving down the freeway, the wind may compress it a little. Does it make sense to interpret the relativistic length contraction as a type of strain that pushes an object's atoms together like this? How does this relate to the previous discussion question?

Summary

Selected Vocabulary

transformation .... the mathematical relationship between the variables such as x and t, as observed in different frames of reference

Terminology Used in Some Other Books

Lorentz transformation ...... the transformation between frames in relative motion

Notation

γ .... an abbreviation for $1/\sqrt{1-v^2}$

Summary

Einstein’s principle of relativity states that both light and matter obey the same laws of physics in any inertial frame of reference, regardless of the frame’s orientation, position, or constant-velocity motion. Maxwell’s equations are the basic laws of physics governing light, and Maxwell’s equations predict a specific value for the speed of light, c=3.0x10^8 m/s, so this new principle implies that *the speed of light must be the same in all frames of reference*, even when it seems intuitively that this is impossible because the frames are in relative motion. This strange constancy of the speed of light was experimentally supported by the 1887 Michelson-Morley experiment. Based only on this principle, Einstein showed that time and space as seen by one observer would be distorted compared to another observer’s perceptions if they were moving relative to each other. This distortion is spelled out in the transformation equations:

$$\begin{array}{ccccc} x' & = & \gamma x & - & \gamma v t \\ t & = & -\gamma v x & + & \gamma t \end{array} \quad ,$$

where v is the velocity of the x',t frame with respect to the x,t frame, and γ is an abbreviation for $1/\sqrt{1-v^2}$. Here, as throughout the chapter, we use the natural system of units in which the speed of light equals 1 by definition, and both times and distances are measured in units of seconds. One second of distance is how far light travels in one second. To change natural-unit equations back to metric units, we must multiply terms by factors of c as necessary in order to make the units of all the terms on both sides of the equation come out right.

Some of the main implications of these equations are:

(1) Nothing can move faster than the speed of light.

(2) The size of a moving object is shrunk. An object appears longest to an observer in a frame moving along with it (a frame in which the object appears is at rest).

(3) Moving clocks run more slowly. A clock appears to run fastest to an observer in a frame moving along with it (a frame in which the object appears is at rest).

(4) There is no well-defined concept of simultaneity for events occurring at different points in space.

Homework Problems

1.(a) Reexpress the transformation equations for frames in relative motion using ordinary units where $c\neq1$. (b) Show that for speeds that are small compared to the speed of light, they are identical to the Galilean equations.

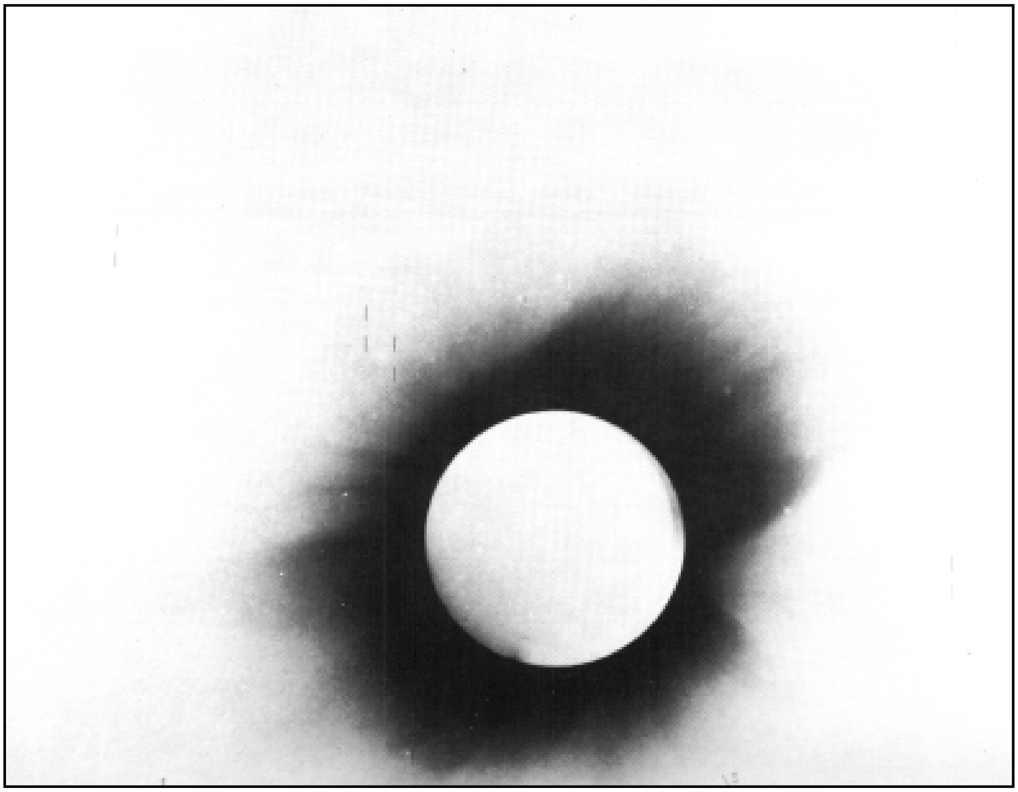
2. Atomic clocks can have accuracies of better than one part in 10^{13}. How does this compare with the time dilation effect produced if the clock takes a trip aboard a jet moving at 300 m/s? Would the effect be measurable? [Hint: Your calculator will round γ off to one. Use the low-velocity approximation $\gamma=1+v^2/2c^2$, which will be derived in chapter 2.]

3. (a) Find an expression for v in terms of γ in natural units. (b) Show that for very large values of γ, v gets close to the speed of light.

4 ★. Of the systems we ordinarily use to transmit information, the fastest ones — radio, television, phone conversations carried over fiber-optic cables — use light. Nevertheless, we might wonder whether it is possible to transmit information at speeds greater than c. The purpose of this problem is to show that if this was possible, then special relativity would have problems with *causality*, the principle that the cause should come earlier in time than the effect. Suppose an event happens at position and time x_1 and t_1 which causes some result at x_2 and t_2. Show that if the distance between x_1 and x_2 is greater than the distance light could cover in the time between t_1 and t_2, then there exists a frame of reference in which the event at x_2 and t_2 occurs *before* the one at x_1 and t_1.

5 ★. Suppose one event occurs at x_1 and t_1 and another at x_2 and t_2. These events are said to have a *spacelike* relationship to each other if the distance between x_1 and x_2 is greater than the distance light could cover in the time between t_1 and t_2, *timelike* if the time between t_1 and t_2 is greater than the time light would need to cover the distance between x_1 and x_2, and *lightlike* if the distance between x_1 and x_2 is the distance light could travel between t_1 and t_2. Show that spacelike relationships between events remain spacelike regardless of what coordinate system we transform to, and likewise for the other two categories. [It may be most elegant to do problem 9 from ch. 2 first and then use that result to solve this problem.]

S A solution is given in the back of the book.
✓ A computerized answer check is available.
★ A difficult problem.
∫ A problem that requires calculus.



Einstein’s famous equation $E=mc^2$ states that mass and energy are equivalent. The energy of a beam of light is equivalent to a certain amount of mass, and the beam is therefore deflected by a gravitational field. Einstein’s prediction of this effect was verified in 1919 by astronomers who photographed stars in the dark sky surrounding the sun during an eclipse. (This is a photographic negative, so the circle that appears bright is actually the dark face of the moon, and the dark area is really the bright corona of the sun.) The stars, marked by lines above and below them, appeared at positions slightly different than their normal ones, indicating that their light had been bent by the sun’s gravity on its way to our planet.

2 Relativity, Part II

So far we have said nothing about how to predict motion in relativity. Do Newton’s laws still work? Do conservation laws still apply? The answer is yes, but many of the definitions need to be modified, and certain entirely new phenomena occur, such as the conversion of mass to energy and energy to mass, as described by the famous equation $E=mc^2$. To cut down on the level of mathematical detail, I have relegated most of the derivations to optional section 2.6, presenting mainly the results and their physical explanations in this section.

2.1 Invariants

The discussion has the potential to become very confusing very quickly because some quantities, force for example, are perceived differently by observers in different frames, whereas in Galilean relativity they were the same in all frames of reference. To clear the smoke it will be helpful to start by identifying quantities that we can depend on *not* to be different in different frames. We have already seen how the principle of relativity requires that the speed of light is the same in all frames of reference. We say that c is invariant.

Another important invariant is mass. This makes sense, because the principle of relativity states that physics works the same in all reference frames. The mass of an electron, for instance, is the same everywhere in the universe, so its numerical value is one of the basic laws of physics. We should therefore expect it to be the same in all frames of reference as well. (Just to make things more confusing, about 50% of all books say mass is invariant, while 50% describe it as changing. It is possible to construct a self-consistent framework of physics according to either description. Neither way is right or wrong, the two philosophies just require different sets of definitions of quantities like momentum and so on. For what it's worth, Einstein eventually weighed in on the mass-as-an-invariant side of the argument. The main thing is just to be consistent.)

A second invariant is electrical charge. This has been verified to high precision because experiments show that an electric field does not produce any measurable force on a hydrogen atom. If charge varied with speed, then the electron, typically orbiting at about 1% of the speed of light, would not exactly cancel the charge of the proton, and the hydrogen atom would have a net charge.

2.2 Combination of Velocities

The impossibility of motion faster than light is the single most radical difference between relativistic and nonrelativistic physics, and we can get at most of the issues in this chapter by considering the flaws in various plans for going faster than light. The simplest argument of this kind is as follows. Suppose Janet takes a trip in a spaceship, and accelerates until she is moving at v=0.9 (90% of the speed of light in natural units) relative to the earth. She then launches a space probe in the forward direction at a speed u=0.2 relative to her ship. Isn't the probe then moving at a velocity of 1.1 times the speed of light relative to the earth?

The problem with this line of reasoning is that the distance covered by the probe in a certain amount of time is shorter as seen by an observer in the earthbound frame of reference, due to length contraction. Velocities are therefore combined not by simple addition but by a more complex method, which we derive in section 2.6 by performing two transformations in a row. In our example, the first transformation would be from the earth's frame to Janet's, the second from Janet's to the probe's. The result is

$$v_{\text{combined}} = \frac{u+v}{1+uv} \quad . \quad \text{[relativistic combination of velocities]}$$

Example: Janet's probe

Applying the equation to Janet's probe, we find

$$\begin{aligned} v_{\text{combined}} &= \frac{0.9+0.2}{1+(0.9)(0.2)} \\ &= 0.93 \quad , \end{aligned}$$

so it is still going quite a bit slower than the speed of light

Example: Combination of velocities in unnatural units

In a system of units, like the metric system, with c≠1, all our symbols for velocity should be replaced with velocities divided by *c*, so we have

$$\frac{v_{\text{combined}}}{c} = \frac{\frac{u}{c}+\frac{v}{c}}{1+\left(\frac{u}{c}\right)\left(\frac{v}{c}\right)} \quad ,$$

or

$$v_{\text{combined}} = \frac{u+v}{1+uv/c^2} \quad .$$

When u and v are both much less than the speed of light, the quantity uv/c^2 is very close to zero, and we recover the nonrelativistic approximation, $v_{\text{combined}}=u+v$.

The line of reasoning given in the second example shows the correspondence principle at work: when a new scientific theory replaces an old one, the two theories must agree within their common realm of applicability.

2.3 Momentum and Force

Momentum

We begin our discussion of relativistic momentum with another scheme for going faster than light. Imagine that a freight train moving at a velocity of 0.6 (v=0.6c in unnatural units) strikes a ping-pong ball that is initially at rest, and suppose that in this collision no kinetic energy is converted into other forms such as heat and sound. We can easily prove based on conservation of momentum that in a very unequal collision of this kind, the smaller object flies off with double the velocity with which it was hit. (This is because the center of mass frame of reference is essentially the same as the frame tied to the freight train, and in the center of mass frame both objects must reverse their initial momenta.) So doesn't the ping-pong ball fly off with a velocity of 1.2, i.e. 20% faster than the speed of light?

The answer is that since $p=mv$ led to this contradiction with the structure of relativity, $p=mv$ must not be the correct equation for relativistic momentum. Apparently $p=mv$ is only a low-velocity approximation to the correct relativistic result. We need to find a new expression for momentum that agrees approximately with $p=mv$ at low velocities, and that also agrees with the principle of relativity, so that if the law of conservation of momentum holds in one frame of reference, it also is obeyed in every other frame. A proof is given in section 2.6 that such an equation is

$$p = m\gamma v \quad , \qquad \text{[relativistic equation for momentum]}$$

which differs from the nonrelativistic version only by the factor of γ. At low velocities γ is very close to 1, so $p=mv$ is approximately true, in agreement with the correspondence principle. At velocities close to the speed of light, γ approaches infinity, and so an object would need infinite momentum to reach the speed of light.

Force

What happens if you keep applying a constant force to an object, causing it to accelerate at a constant rate until it exceeds the speed of light? The hidden assumption here is that Newton's second law, $a=F/m$, is still true. It isn't. Experiments show that at speeds comparable to the speed of light, $a=F/m$ is wrong. The equation that still *is* true is

$$F = \frac{\Delta p}{\Delta t} \quad .$$

You could apply a constant force to an object forever, increasing its momentum at a steady rate, but as the momentum approached infinity, the velocity would approach the speed of light. In general, a force produces an acceleration significantly *less* than F/m at relativistic speeds.

Would passengers on a spaceship moving close to the speed of light perceive every object as being more difficult to accelerate, as if it was more massive? No, because then they would be able to detect a change in the laws of physics because of their state of motion, which would violate the principle of relativity. The way out of this difficulty is to realize that force is not an invariant. What the passengers perceive as a small force causing a small change in momentum would look to a person in the earth's frame of reference like a large force causing a large change in momentum. As a practical matter, conservation laws are usually more convenient tools for relativistic problem solving than procedures based on the force concept.

2.4 Kinetic Energy

Since kinetic energy equals $\frac{1}{2}mv^2$, wouldn't a sufficient amount of energy cause v to exceed the speed of light? You're on to my methods by now, so you know this is motivation for a redefinition of kinetic energy. Section 2.6 derives the work-kinetic energy theorem using the correct relativistic treatment of force. The result is

$$KE = m(\gamma-1) \quad . \quad \text{[relativistic kinetic energy]}$$

Since γ approaches infinity as velocity approaches the speed of light, an infinite amount of energy would be required in order to make an object move at the speed of light.

Example: Kinetic energy in unnatural units

How can this equation be converted back into units in which the speed of light does not equal one? One approach would be to redo the derivation in section 2.6 in unnatural units. A far simpler approach is simply to add factors of c where necessary to make the metric units look consistent. Suppose we decide to modify the right side in order to make its units consistent with the energy units on the left. The ordinary nonrelativistic definition of kinetic energy as $\frac{1}{2}mv^2$ shows that the units on the left are

$$\text{kg}\cdot\frac{\text{m}^2}{\text{s}^2} \quad .$$

The factor of γ–1 is unitless, so the mass units on the right need to be multiplied by m^2/s^2 to agree with the left. This means that we need to multiply the right side by c^2:

$$KE = mc^2(\gamma-1)$$

This is beginning to resemble the famous $E=mc^2$ equation, which we will soon attack head-on.

Example: The correspondence principle for kinetic energy

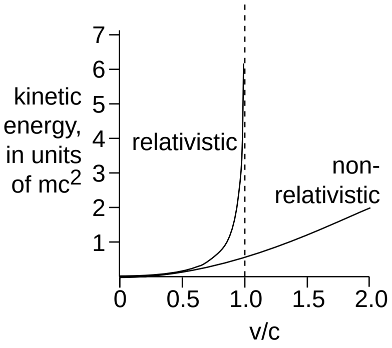
It is far from obvious that this result, even in its metric-unit form, reduces to the familiar $\frac{1}{2}mv^2$ at low speeds, as required by the correspondence principle. To show this, we need to find a low-velocity approximation for γ. In metric units, the equation for γ reads as

$$\gamma = \frac{1}{\sqrt{1-v^2/c^2}} \quad .$$

Reexpressing this as $\left(1-v^2/c^2\right)^{-1/2}$, and making use of the approximation $\left(1+\varepsilon\right)^p \approx 1+p\varepsilon$ for small ε, the equation for gamma becomes

$$\gamma \approx 1 + \frac{v^2}{2c^2} \quad ,$$

which can readily be used to show $mc^2(\gamma-1)\approx\frac{1}{2}mv^2$.



Echenevex
Versonnex
P5
P4
P6
Ornex
P7
Crozet
Forney-Voltaire
Prevessin
P8
Sergy
P2
Saint-Genis
P1
Meyrin

The Large Hadron Collider. The red circle shows the location of the underground tunnel which the LHC will share with a preexisting accelerator.

Example: the large hadron collider

Question: The Large Hadron Collider (LHC), being built in Switzerland, is a ring with a radius of 4.3 km, designed to accelerate two counterrotating beams of protons to energies of 7 TeV per proton. (The word “hadron” refers to any particle that participates in strong nuclear forces.) The TeV is a unit of energy equal to 10^{12} eV, where 1 eV=$1.60\text{x}10^{-19}$ J is the energy a particle with unit charge acquires by moving through a voltage difference of 1 V.) The ring has to be so big because the inward force from the accelerator’s magnets would not be great enough to make the protons curve more tightly at top speed.

(a) What inward force must be exerted on each proton?

(b) In a purely Newtonian world where there were no relativistic effects, how much smaller could the LHC be if it was to produce proton beams moving at speeds close to the speed of light?

Solution:

(a) Since the protons have velocity vectors with constant magnitudes, γ is constant, so let’s start by computing it. We’ll work the whole problem in mks, since none of the data are given in natural units. The kinetic energy of each proton is

$$\begin{aligned} KE &= 7\text{ TeV} \\ &= (7\text{ TeV})(10^{12}\text{ eV/TeV})(1.60\text{x}10^{-19}\text{ J/eV}) \\ &= 1.1\text{x}10^{-6}\text{ J} \quad . \end{aligned}$$

A microjoule is quite a healthy energy for a subatomic particle! Looking up the mass of a proton, we have

$$\begin{aligned} mc^2 &= (1.7\text{x}10^{-27}\text{ kg})(3.0\text{x}10^{8}\text{ m/s})^2 \\ &= 1.5\text{x}10^{-10}\text{ J} \quad . \end{aligned}$$

The kinetic energy is thousands of times greater than mc^2, so the protons go very close to the speed of light. Under these conditions there is no significant difference between γ and γ–1, so

$$\begin{aligned} \gamma &\approx KE\,/\,mc^2 \\ &= 7.3\text{x}10^{3} \end{aligned}$$

We analyze the circular motion in the laboratory frame of reference, since that is the frame of reference in which the LHC’s magnets sit, and their fields were calibrated by instruments at rest with respect to them. The inward force required is

$$\begin{aligned} \mathbf{F} &= \Delta\mathbf{p}/\Delta t \\ &= \Delta(m\gamma\mathbf{v})/\Delta t \\ &= m\gamma\,\Delta\mathbf{v}/\Delta t \\ &= m\gamma\,\mathbf{a} \quad . \end{aligned}$$

Except for the factor of γ, this is the same result we would have had in Newtonian physics, where we already know the equation $a=v^2/r$ for the inward acceleration in uniform circular motion. Since the velocity is essentially the speed of light, we have $a=c^2/r$. The force required is

$$\begin{aligned} \text{F} &= m\gamma\,c^2/r \\ &= KE\,/\,r \quad . \qquad [\text{since } \gamma \approx \gamma\text{–}1] \end{aligned}$$

This looks a little funny, but the units check out, since a joule is the same as a newton-meter. The result is

$$\text{F} = 2.6\text{x}10^{-10}\text{ N}$$

(b)
$$\begin{aligned} \text{F} &= mv^2/r \qquad [\text{nonrelativistic equation}] \\ &= mc^2/r \\ r &= mc^2/F \\ &= 0.59\text{ m} \end{aligned}$$

In a nonrelativistic world, it would be a table-top accelerator! The energies and momenta, however, would be smaller.

2.5 Equivalence of Mass and Energy

The treatment of relativity so far has been purely mechanical, so the only form of energy we have discussed is kinetic. For example, the storyline for the introduction of relativistic momentum was based on collisions in which no kinetic energy was converted to other forms. We know, however, that collisions can result in the production of heat, which is a form of kinetic energy at the molecular level, or the conversion of kinetic energy into entirely different forms of energy, such as light or potential energy.

Let's consider what happens if a blob of putty moving at velocity v hits another blob that is initially at rest, sticking to it, and that as much kinetic energy as possible is converted into heat. (It is not possible for all the KE to be converted to heat, because then conservation of momentum would be violated.) The nonrelativistic result is that to obey conservation of momentum the two blobs must fly off together at $v/2$.

Relativistically, however, an interesting thing happens. A hot object has more momentum than a cold object! This is because the relativistically correct expression for momentum is $p=m\gamma v$, and the more rapidly moving molecules in the hot object have higher values of γ. There is no such effect in nonrelativistic physics, because the velocities of the moving molecules are all in random directions, so the random motion's contribution to momentum cancels out.

In our collision, the final combined blob must therefore be moving a little more slowly than the expected $v/2$, since otherwise the final momentum would have been a little greater than the initial momentum. To an observer who believes in conservation of momentum and knows only about the overall motion of the objects and not about their heat content, the low velocity after the collision would have to seem to require a magical change in the mass, as if the mass of two combined, hot blobs of putty was more than the sum of their individual masses.

Heat energy is equivalent to mass.

Now we know that mass is invariant, and no molecules were created or destroyed, so the masses of all the molecules must be the same as they always were. The change is due to the change in γ with heating, not to a change in m. But how much does the mass *appear* to change? In section 2.6 we prove that the perceived change in mass exactly equals the change in heat energy between two temperatures, i.e. changing the heat energy by an amount E changes the effective mass of an object by E as well. This looks a bit odd because the natural units of energy and mass are the same. Converting back to ordinary units by our usual shortcut of introducing factors of c, we find that changing the heat energy by an amount E causes the apparent mass to change by $m=E/c^2$. Rearranging, we have the famous $E=mc^2$.

All energy is equivalent to mass.

But this whole argument was based on the fact that heat is a form of kinetic energy at the molecular level. Would $E=mc^2$ apply to other forms of energy as well? Suppose a rocket ship contains some electrical potential energy stored in a battery. If we believed that $E=mc^2$ applied to forms of kinetic energy but not to electrical potential energy, then we would have to expect that the pilot of the rocket could slow the ship down by using the battery to run a heater! This would not only be strange, but it would violate the principle of relativity, because the result of the experiment would be different depending on whether the ship was at rest or not. The only logical conclusion is that *all forms of energy are equivalent to mass*. Running the heater then has no effect on the motion of the ship, because the total energy in the ship was unchanged; one form of energy was simply converted to another.

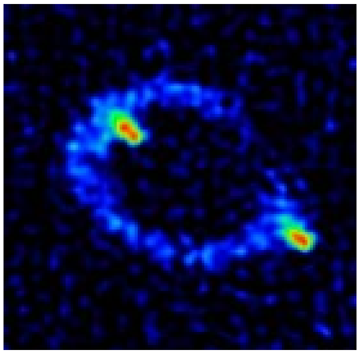
Example: A rusting nail

Question: A 50-gram iron nail is left in a cup of water until it turns entirely to rust. The energy released is about 0.5 MJ (mega-joules). In theory, would a sufficiently precise scale register a change in mass? If so, how much?

Solution: The energy will appear as heat, which will be lost to the environment. So the total mass-energy of the cup, water, and iron will indeed be lessened by 0.5 MJ. (If it had been perfectly insulated, there would have been no change, since the heat energy would have been trapped in the cup.) Converting to mass units, we have

$$\begin{aligned} m &= E/c^2 \\ &= (0.5\text{x}10^6\text{ J}) / (3.0\text{x}10^8\text{ m/s})^2 \\ &= 6\text{x}10^{-12}\text{ J/(m}^2\text{/s}^2) \\ &= 6\text{x}10^{-12}\text{ (kg}\cdot\text{m}^2\text{/s}^2\text{)/(m}^2\text{/s}^2) \\ &= 6\text{x}10^{-12}\text{ kg} \quad , \end{aligned}$$

so the change in mass is too small to measure with any practical technique. This is because the square of the speed of light is such a large number in metric units.



This telescope picture shows two images of the same distant object, an exotic, very luminous object called a quasar. This is interpreted as evidence that a massive, dark object, possibly a black hole, happens to be between us and it. Light rays that would otherwise have missed the earth on either side have been bent by the dark object's gravity so that they reach us. The actual direction to the quasar is presumably in the center of the image, but the light along that central line don't get to us because they are absorbed by the dark object. The quasar is known by its catalog number, MG1131+0456, or more informally as Einstein's Ring.

Energy participates in gravitational forces.

In the example we tacitly assumed that the increase in mass would show up on a scale, i.e. that its gravitational attraction with the earth would increase. Strictly speaking, however, we have only proven that energy relates to *inertial* mass, i.e. to phenomena like momentum and the resistance of an object to a change in its state of motion. Even before Einstein, however, experiments had shown to a high degree of precision that any two objects with the same inertial mass will also exhibit the same gravitational attractions, i.e. have the same *gravitational* mass. For example, the only reason that all objects fall with the same acceleration is that a more massive object's inertia is exactly in proportion to the greater gravitational forces in which it participates. We therefore conclude that energy participates in gravitational forces in the same way mass does. The total gravitational attraction between two objects is proportional not just to the product of their masses, m_1m_2, as in Newton's law of gravity, but to the quantity $(m_1+E_1)(m_2+E_2)$. (Even this modification does not give a complete, self-consistent theory of gravity, which is only accomplished through the general theory of relativity.)

Example: Gravity bending light

The first important experimental confirmation of relativity came when stars next to the sun during a solar eclipse were observed to have shifted a little from their ordinary position. (If there was no eclipse, the glare of the sun would prevent the stars from being observed.) Starlight had been deflected by gravity.

Example: Black holes

A star with sufficiently strong gravity can prevent light from leaving. Quite a few black holes have been detected via their gravitational forces on neighboring stars or clouds of dust.

Creation and destruction of particles

Since mass and energy are beginning to look like two sides of the same coin, it may not be so surprising that nature displays processes in which particles are actually destroyed or created; energy and mass are then converted back and forth on a wholesale basis. This means that in relativity there are no separate laws of conservation of energy and conservation of mass. There is only a law of conservation of mass plus energy (referred to as mass-energy). In natural units, $E+m$ is conserved, while in ordinary units the conserved quantity is $E+mc^2$.

Example: Electron-positron annihilation

Natural radioactivity in the earth produces positrons, which are like electrons but have the opposite charge. A form of antimatter, positrons annihilate with electrons to produce gamma rays, a form of high-frequency light. Such a process would have been considered impossible before Einstein, because conservation of mass and energy were believed to be separate principles, and the process eliminates 100% of the original mass. In metric units, the amount of energy produced by annihilating 1 kg of matter with 1 kg of antimatter is

$$\begin{aligned} E &= mc^2 \\ &= (2\ \text{kg})(3.0\text{x}10^8\ \text{m/s})^2 \\ &= 2\text{x}10^{17}\ \text{J} \quad , \end{aligned}$$

which is on the same order of magnitude as a day's energy consumption for the entire world!

Positron annihilation forms the basis for the medical imaging procedure called a PET (positron emission tomography) scan, in which a positron-emitting chemical is injected into the patient and mapped by the emission of gamma rays from the parts of the body where it accumulates.

Note that the idea of mass as an invariant is separate from the idea that mass is not separately conserved. Invariance is the statement that all observers agree on a particle's mass regardless of their motion relative to the particle. Mass may be created or destroyed if particles are created or destroyed, and in such a situation mass invariance simply says that all observers will agree on how much mass was created or destroyed.

2.6* Proofs

Combination of velocities

We proceed by transforming from the x,t frame to the x',t' frame moving relative to it at a velocity v_1, and then from that from to a third frame, x'',t'', moving with respect to the second at v_2. The result must be equivalent to a single transformation from x,t to x'',t'' using the combined velocity. Transforming from x,t to x',t' gives

$$\begin{aligned} x' &= \gamma x - v_1\gamma_1 t \\ t' &= -v_1\gamma_1 x + \gamma_1 t \quad , \end{aligned}$$

and plugging this into the second transformation results in

$$\begin{aligned} x'' &= \gamma_2(\gamma x - v_1\gamma_1 t) - v_2\gamma_2(-v_1\gamma_1 x + \gamma_1 t) \\ t'' &= \ldots + \ldots \quad , \end{aligned}$$

where "..." indicates terms that we don't need to complete the derivation. Collecting terms gives

$$x'' = (\ldots)x - (v_1+v_2)\gamma_1\gamma_2 t \quad ,$$

where the coefficient of t, $-(v_1+v_2)\gamma_1\gamma_2$, must be the same as it would have been in a direct transformation from x,t to x'',t'':

$$-v_{\text{combined}}\gamma_{\text{combined}} = -(v_1+v_2)\gamma_1\gamma_2$$

Straightforward algebra then produces the equation in section 2.2.

Relativistic momentum

We want to show that if $p=m\gamma v$, then any collision that conserves momentum in the center of mass frame will also conserve momentum in any other frame. The whole thing is restricted to two-body collisions in one dimension in which no kinetic energy is changed to any other form, so it is not a general proof that $p=m\gamma v$ forms a consistent part of the theory of relativity. This is just the minimum test we want the equation to pass.

Let the new frame be moving at a velocity u with respect to the center of mass and let Γ (capital gamma) be$1/\sqrt{1-u^2}$. Then the total momentum in the new frame (at any moment before or after the collision) is

$$p' = m_1\gamma_1'v_1' + m_2\gamma_2'v_2' \quad .$$

The velocities v_1' and v_2' result from combining v_1 and v_2 with u, so making use of the result from the previous proof,

$$\begin{aligned} p' &= m_1(v_1+u)\Gamma\gamma_1 + m_2(v_2+u)\Gamma\gamma_2 \\ &= (m_1\gamma_1 v_1+m_2\gamma_2 v_2)\Gamma + (m_1\gamma_1+m_2\gamma_2)\Gamma u \\ &= p\Gamma + (KE_1+m_1+KE_2+m_2)\Gamma u \quad . \end{aligned}$$

If momentum is conserved in the center of mass frame, then there is no change in p, the momentum in the center of mass frame, after the collision. The first term is therefore the same before and after, and the second term is also the same before and after because mass is invariant, and we have assumed no KE was converted to other forms of energy. (We shouldn't expect the proof to work if KE is changed to other forms, because we have not taken into account the effects of any other forms of mass-energy.)

Relativistic work-kinetic energy theorem

This is a straightforward application of calculus, albeit with a couple of tricks to make it easier to do without recourse to a table of integrals. The kinetic energy of an object of mass m moving with velocity v equals the work done in accelerating it to that speed from rest:

$$\begin{aligned}
KE &= \int_{v=0}^{v} F\,\mathrm{d}x \\
&= \int_{v=0}^{v} \frac{\mathrm{d}p}{\mathrm{d}t}\,\mathrm{d}x \\
&= \int_{v=0}^{v} \frac{\mathrm{d}\left(m\gamma v\right)}{\mathrm{d}t}\,\mathrm{d}x \\
&= m\int_{v=0}^{v} v\,\mathrm{d}\left(\gamma v\right) \\
&= m\int_{v=0}^{v} v^2\,\mathrm{d}\gamma + m\int_{v=0}^{v} v\gamma\,\mathrm{d}v \\
&= m\int_{v=0}^{v} \left(1-\gamma^{-2}\right)\mathrm{d}\gamma + m\int_{v=0}^{v} \frac{v\,\mathrm{d}v}{\sqrt{1-v^2}} \\
&= m\left(\gamma+\frac{1}{\gamma}\right)\Bigg]_{v=0}^{v} - m\sqrt{1-v^2}\,\Bigg]_{v=0}^{v} \\
&= m\left(\gamma+\frac{1}{\gamma}-\sqrt{1-v^2}\right)\Bigg]_{v=0}^{v} \\
&= m\gamma\Big]_{v=0}^{v} \quad = \quad m(\gamma–1)
\end{aligned}$$

Change in inertia with heating

Technical note
This proof really only applies to an ideal gas, which expresses all of its heat energy as kinetic energy. In general heat energy is expressed partly as kinetic energy and partly as electrical potential energy.

We prove here that the inertia of a heated object (its apparent mass) increases by an amount equal to the heat. Suppose an object moving with velocity v_{cm} consists of molecules with masses m_1, m_2, ..., which are moving relative to the origin at velocities v_{o1}, v_{o2}, ... and relative to the object's center of mass at velocities v_1, v_2, ... The total momentum is

$$\begin{aligned}
p_{\mathrm{total}} &= m_1 v_{\mathrm{o1}}\,\gamma_{\mathrm{o1}} + \ldots \\
&= m_1(v_{\mathrm{cm}}+v_1)\gamma_{\mathrm{cm}}\gamma_1 + \ldots
\end{aligned}$$

where we have used the result from the first subsection. Rearranging,

$$p_{\mathrm{total}} = \gamma_{\mathrm{cm}}\ [(m_1\gamma_1 v_{\mathrm{cm}}+\ldots) + (m_1\gamma_1 v_1+\ldots)]$$

The second term, which is the total momentum in the c.m. frame, vanishes.

$$p_{\mathrm{total}} = (m_1\gamma_1+\ldots)\gamma_{\mathrm{cm}} v_{\mathrm{cm}}$$

The quantity in parentheses is the total mass plus the total thermal energy.

Summary

Selected Vocabulary

invariant .... a quantity that does not change when transformed

Terminology Used in Some Other Books

rest mass .... referred to as mass in this book; written as m_0 in some books

mass .... What some books mean by “mass” is our $m\gamma$.

Summary

Other quantities besides space and time, including momentum, force, and energy, are distorted when transformed from one frame to another, just as time and space are. But some quantities, notably mass, electric charge, and the speed of light, are invariant: they are the same in all frames.

If object A moves at velocity u relative to object B, and B moves at velocity v relative to object C, the combination of the velocities, i.e. A’s velocity relative to C, is not given by $u+v$ but rather by

$$v_{\text{combined}} = \frac{u+v}{1+uv} \quad \text{[natural units]} \qquad = \frac{u+v}{1+uv/c^2} \quad \text{[ordinary units]} \quad .$$

Relativistic momentum is the same in either system of units,

$$p = m\gamma v \quad \text{[natural units]} \qquad = m\gamma v \quad \text{[ordinary units]} \quad ,$$

and kinetic energy is

$$KE = m(\gamma - 1) \quad \text{[natural units]} \qquad = mc^2(\gamma - 1) \quad \text{[ordinary units]} \quad .$$

A consequence of the theory of relativity is that mass and energy do not obey separate conservation laws. Instead, the conserved quantity is the mass-energy. Mass and energy may be converted into each other according to the famous equation

$$E = m \quad \text{[natural units]} \qquad = mc^2 \quad \text{[ordinary units]} \quad .$$

Homework Problems

1✓. (a) A spacecraft traveling at 1.0000x10^{7} m/s relative to the earth releases a probe in the forward direction at a relative speed of 2.0000x10^{7} m/s. How fast is the probe moving relative to the earth? How does this compare with the nonrelativistic result? (b) Repeat the calculation, but with both velocities equal to $c/2$. How does this compare with the nonrelativistic result?

2. (a) Show that when two velocities are combined relativistically, and one of them equals the speed of light, the result also equals the speed of light. (b) Explain why it has to be this way based on the principle of relativity. [Note that it doesn't work to say that it has to be this way because motion faster than c is impossible. That isn't what the principle of relativity says, and it also doesn't handle the case where the velocities are in opposite direction.]

3✓. Cosmic-ray particles with relativistic velocities are continually bombarding the earth's atmosphere. They are protons and other atomic nuclei. Suppose a carbon nucleus (containing six protons and six neutrons) arrives with an energy of 10^{-7} J, which is unusually high, but not unheard of. By what factor is its length shortened as seen by an observer in the earth's frame of reference? [Hint: You can just find γ, and avoid finding v.]

4✓. (a) A free neutron (as opposed to a neutron bound into an atomic nucleus) is unstable, and decays radioactively into a proton, an electron, and a particle called a neutrino. (This process can also occur for a neutron in a nucleus, but then other forms of mass-energy are involved as well.) The masses are as follows:

neutron	1.67495x10^{-27} kg
proton	1.67265x10^{-27} kg
electron	0.00091x10^{-27} kg
neutrino	negligible

Find the energy released in the decay of a free neutron.

(b) We might imagine that a proton could decay into a neutron, a positron, and a neutrino. Although such a process can occur within a nucleus, explain why it cannot happen to a free proton. (If it could, hydrogen would be radioactive!)

5. (a) Find a relativistic equation for the velocity of an object in terms of its mass and momentum (eliminating γ). Work in natural units. (b) Show that your result is approximately the same as the classical value, p/m, at low velocities. (c) Show that very large momenta result in speeds close to the speed of light.

S A solution is given in the back of the book.
✓ A computerized answer check is available.
★ A difficult problem.
∫ A problem that requires calculus.

6. (a) Prove the equation E^2-p^2=m^2 for a material object, where E=$m\gamma$ is the total mass-energy. (b) Using this result, show that an object with zero mass must move at the speed of light. (c) This equation can be applied more generally, to light for instance. Use it to find the momentum of a beam of light having energy E. (d) Convert your answer from the previous part into ordinary units. [answer: p=E/c]

7. Starting from the equation $v_{\text{combined}}\gamma_{\text{combined}} = (v_1+v_2)\gamma_1\gamma_2$ derived in section 2.6, complete the proof of $v_{\text{combined}} = (v_1+v_2)/(1+v_1v_2)$.

8★. A source of light with frequency f is moving toward an observer at velocity v (or away from the observer if v is negative). Find the relativistically correct equation for the Doppler shift of the light. [Hint: Write down an equation for the motion of one wavefront in the source's frame, and then a second equation for the motion of the next wavefront in the source's frame. Then transform to the observer's frame and find the separation in time between the arrival of the first and second wavefronts at the same point in the observer's frame.]

9 ★. Suppose one event occurs at x_1 and t_1 and another at x_2 and t_2. Prove that the quantity $(t_2-t_1)^2-(x_2-x_1)^2$ is the same even when we transform into another coordinate system. This quantity is therefore a kind of invariant, albeit an invariant of a more abstract kind than the ones discussed until now. [When the relationship between the events is timelike in the sense of problem 5 in ch. 1, the square root of $(t_2-t_1)^2-(x_2-x_1)^2$ can be interpreted as the amount of time that would be measured by a clock that moved from one event to the other at constant velocity. It is therefore known as the *proper time* between events 1 and 2. The way the proper time relates to space and time is very much like the way Pythagorean theorem relates distance to two space dimensions, the difference being the negative sign that occurs in the former. Proper time is unaffected by Einstein-style transformations, whereas distance is unaffected by rotations.]

10. An antielectron collides with an electron that is at rest. (An antielectron is a form of antimatter that is just like an electron, but with the opposite charge.) The antielectron and electron annihilate each other and produce two gamma rays. (A gamma ray is a form of light. It has zero mass.) Gamma ray 1 is moving in the same direction as the antielectron was initially going, and gamma ray 2 is going in the opposite direction. Throughout this problem, you should work in natural units and use the notation E to mean the total mass-energy of a particle, i.e. its mass plus its kinetic energy. Find the energies of the two gamma-rays, E_1 and E_2, in terms of m, the mass of an electron or antielectron, and E_o, the initial mass-energy of the antielectron. [Hint: See problem 6a.]

11 S. (a) Use the result of problem 6d to show that if light with power P is reflected perpendicularly from a perfectly reflective surface, the force on the surface is $2P/c$. (b) Estimate the maximum mass of a thin film that is to be levitated by a 100-watt lightbulb.

12 S. A solar sail is a propulsion system for a spacecraft that uses the sun's light pressure for propulsion. The Cosmos-1 solar sail, launched as a test in 2001, consisted of a 600 m^2 aluminized mylar sail attached to a 40 kg payload. The mylar was 5 μm thick. The density of mylar is 1.40 g/cm^3. The flux of light from the sun in the part of the solar system near the earth is about 1400 W/m^2. Find the acceleration of the vehicle due to light pressure, for the case where the sail is oriented for maximum thrust. (This acceleration is actually much smaller than the acceleration due to the sun's gravity. The earth, however, experiences this same gravitational acceleration, so what you're really calculating is the craft's acceleration relative to the earth.)

Left: In 1980, the continental U.S. got its first taste of active volcanism in recent memory with the eruption of Mount St. Helens.

Top: An eruption of the Hawaiian volcano Pu'u O'o.

3 Rules of Randomness

Given for one instant an intelligence which could comprehend all the forces by which nature is animated and the respective positions of the things which compose it...nothing would be uncertain, and the future as the past would be laid out before its eyes.

Pierre Simon de Laplace, 1776

The energy produced by the atom is a very poor kind of thing. Anyone who expects a source of power from the transformation of these atoms is talking moonshine.

Ernest Rutherford, 1933

The Quantum Mechanics is very imposing. But an inner voice tells me that it is still not the final truth. The theory yields much, but it hardly brings us nearer to the secret of the Old One. In any case, I am convinced that He does not play dice.

Albert Einstein

However radical Newton's clockwork universe seemed to his contemporaries, by the early twentieth century it had become a sort of smugly accepted dogma. Luckily for us, this deterministic picture of the universe breaks down at the atomic level. The clearest demonstration that the laws of physics contain elements of randomness is in the behavior of radioactive atoms. Pick two identical atoms of a radioactive isotope, say the naturally occurring uranium 238, and watch them carefully. They will undergo fission at different times, even though there was no difference in their initial behavior.

We would be in big trouble if these atoms' behavior was as predictable as expected in the Newtonian world-view, because radioactivity is an important source of heat for our planet. In reality, each atom chooses a random moment at which to release its energy, resulting in a nice steady heating effect. The earth would be a much colder planet if only sunlight heated it and not radioactivity. Probably there would be no volcanoes, and the oceans would never have been liquid. The deep-sea geothermal vents in which life first evolved would never have existed. But there would be an even worse consequence if radioactivity was deterministic: after a few billion years of peace, all the uranium 238 atoms in our planet would presumably pick the same moment to decay. The huge amount of stored nuclear energy, instead of being spread out over eons, would all be released at one instant, blowing our whole planet to Kingdom Come.

The new version of physics, incorporating certain kinds of randomness, is called quantum physics (for reasons that will become clear later). It represented such a dramatic break with the previous, deterministic tradition that everything that came before is considered "classical," even the theory of relativity. The remainder of this book is a basic introduction to quantum physics.

3.1 Randomness Isn't Random

Einstein's distaste for randomness, and his association of determinism with divinity, goes back to the Enlightenment conception of the universe as a gigantic piece of clockwork that only had to be set in motion initially by the Builder. Many of the founders of quantum mechanics were interested in possible links between physics and Eastern and Western religious and philosophical thought, but every educated person has a different concept of religion and philosophy. Bertrand Russell remarked, "Sir Arthur Eddington deduces religion from the fact that atoms do not obey the laws of mathematics. Sir James Jeans deduces it from the fact that they do."

Russell's witticism, which implies incorrectly that mathematics cannot describe randomness, reminds us how important it is not to oversimplify this question of randomness. You should not simply surmise, "Well, it's all random, anything can happen." For one thing, certain things simply cannot happen, either in classical physics or quantum physics. The conservation laws of mass, energy, momentum, and angular momentum are still valid, so for instance processes that create energy out of nothing are not just unlikely according to quantum physics, they are impossible.

A useful analogy can be made with the role of randomness in evolution. Darwin was not the first biologist to suggest that species changed over long periods of time. His two new fundamental ideas were that (1) the changes arose through random genetic variation, and (2) changes that enhanced the organism's ability to survive and reproduce would be preserved, while maladaptive changes would be eliminated by natural selection. Doubters of evolution often consider only the first point, about the randomness of natural variation, but not the second point, about the systematic action of natural selection. They make statements such as, "the development of a complex organism like Homo sapiens via random chance would be like a whirlwind blowing through a junkyard and spontaneously assembling a jumbo jet out of the scrap metal.” The flaw in this type of reasoning is that it ignores the deterministic constraints on the results of random processes. For an atom to violate conservation of energy is no more likely than the conquest of the world by chimpanzees next year.

Discussion Question

Economists often behave like wannabe physicists, probably because it seems prestigious to make numerical calculations instead of talking about human relationships and organizations like other social scientists. Their striving to make economics work like Newtonian physics extends to a parallel use of mechanical metaphors, as in the concept of a market's supply and demand acting like a self-adjusting machine, and the idealization of people as economic automatons who consistently strive to maximize their own wealth. What evidence is there for randomness rather than mechanical determinism in economics?

3.2 Calculating Randomness

You should also realize that even if something is random, we can still understand it, and we can still calculate probabilities numerically. In other words, physicists are good bookmakers. A good bookmaker can calculate the odds that a horse will win a race much more accurately that an inexperienced one, but nevertheless cannot predict what will happen in any particular race.

Statistical independence

As an illustration of a general technique for calculating odds, suppose you are playing a 25-cent slot machine. Each of the three wheels has one chance in ten of coming up with a cherry. If all three wheels come up cherries, you win $100. Even though the results of any particular trial are random, you can make certain quantitative predictions. First, you can calculate that your odds of winning on any given trial are 1/10x1/10x1/10=1/1000=0.001. Here, I am representing the probabilities as numbers from 0 to 1, which is clearer than statements like “The odds are 999 to 1,” and makes the calculations easier. A probability of 0 represents something impossible, and a probability of 1 represents something that will definitely happen.

Also, you can say that any given trial is equally likely to result in a win, and it doesn't matter whether you have won or lost in prior games. Mathematically, we say that each trial is statistically independent, or that separate games are uncorrelated. Most gamblers are mistakenly convinced that, to the contrary, games of chance are correlated. If they have been playing a slot machine all day, they are convinced that it is “getting ready to pay,” and they do not want anyone else playing the machine and “using up” the jackpot that they “have coming.” In other words, they are claiming that a series of trials at the slot machine is negatively correlated, that losing now makes you more likely to win later. Craps players claim that you should go to a table where the person rolling the dice is “hot,” because she is likely to keep on rolling good numbers. Craps players, then, believe that rolls of the dice are positively correlated, that winning now makes you more likely to win later.

My method of calculating the probability of winning on the slot machine was an example of the following important rule for calculations based on independent probabilities:

> **The Law of Independent Probabilities**
> If the probability of one event happening is P_A, and the probability of a second statistically independent event happening is P_B, then the probability that they will both occur is the product of the probabilities, $P_A P_B$. If there are more than two events involved, you simply keep on multiplying.

Note that this only applies to independent probabilities. For instance, if you have a nickel and a dime in your pocket, and you randomly pull one out, there is a probability of 0.5 that it will be the nickel. If you then replace the coin and again pull one out randomly, there is again a probability of 0.5 of coming up with the nickel, because the probabilities are independent. Thus, there is a probability of 0.25 that you will get the nickel both times.

Suppose instead that you do not replace the first coin before pulling out the second one. Then you are bound to pull out the other coin the second time, and there is no way you could pull the nickel out twice. In this situation, the two trials are not independent, because the result of the first trial has an effect on the second trial. The law of independent probabilities does not apply, and the probability of getting the nickel twice is zero, not 0.25.

Experiments have shown that in the case of radioactive decay, the probability that any nucleus will decay during a given time interval is unaffected by what is happening to the other nuclei, and is also unrelated to how long it has gone without decaying. The first observation makes sense, because nuclei are isolated from each other at the centers of their respective atoms, and therefore have no physical way of influencing each other. The second fact is also reasonable, since all atoms are identical. Suppose we wanted to believe that certain atoms were “extra tough,” as demonstrated by their history of going an unusually long time without decaying. Those atoms would have to be different in some physical way, but nobody has ever succeeded in detecting differences among atoms. There is no way for an atom to be changed by the experiences it has in its lifetime.

Addition of probabilities

The law of independent probabilities tells us to use multiplication to calculate the probability that both A and B will happen, assuming the probabilities are independent. What about the probability of an “or” rather than an “and”? If two events A and B are mutually exclusive, then the probability of one or the other occurring is the sum P_A+P_B. For instance, a bowler might have a 30% chance of getting a strike (knocking down all ten pins) and a 20% chance of knocking down nine of them. The bowler's chance of knocking down either nine pins or ten pins is therefore 50%.

It does not make sense to add probabilities of things that are not mutually exclusive, i.e. that could both happen. Say I have a 90% chance of eating lunch on any given day, and a 90% chance of eating dinner. The probability that I will eat either lunch or dinner is not 180%.

Normalization

If I spin a globe and randomly pick a point on it, I have about a 70% chance of picking a point that's in an ocean and a 30% chance of picking a point on land. The probability of picking either water or land is 70%+30%=100%. Water and land are mutually exclusive, and there are no other possibilities, so the probabilities had to add up to 100%. It works the same if there are more than two possibilities — if you can classify all possible outcomes into a list of mutually exclusive results, then all the probabilities have to add up to 1, or 100%. This property of probabilities is known as normalization.

Averages

Another way of dealing with randomness is to take averages. The casino knows that in the long run, the number of times you win will approximately equal the number of times you play multiplied by the probability of winning. In the game mentioned above, where the probability of winning is 0.001, if you spend a week playing, and pay $2500 to play 10,000 times, you are likely to win about 10 times (10,000x0.001=10), and collect $1000. On the average, the casino will make a profit of $1500 from you. This is an example of the following rule.

> **Rule for Calculating Averages**
> If you conduct *N* identical, statistically independent trials, and the probability of success in each trial is *P*, then on the average, the total number of successful trials will be *NP*. If *N* is large enough, the relative error in this estimate will become small.

The statement that the rule for calculating averages gets more and more accurate for larger and larger *N* (known popularly as the "law of averages") often provides a correspondence principle that connects classical and quantum physics. For instance, the amount of power produced by a nuclear power plant is not random at any detectable level, because the number of atoms in the reactor is so large. In general, random behavior at the atomic level tends to average out when we consider large numbers of atoms, which is why physics seemed deterministic before physicists learned techniques for studying atoms individually.

We can achieve great precision with averages in quantum physics because we can use identical atoms to reproduce exactly the same situation many times. If we were betting on horses or dice, we would be much more limited in our precision. After a thousand races, the horse would be ready to retire. After a million rolls, the dice would be worn out.

Self-Check

Which of the following things *must* have independent, which *could* be independent, and which definitely are *not* independent?
(1) the probability of successfully making two free-throws in a row in basketball
(2) the probability that it will rain in London tomorrow and the probability that it will rain on the same day in a certain city in a distant galaxy
(3) your probability of dying today and of dying tomorrow

Discussion questions

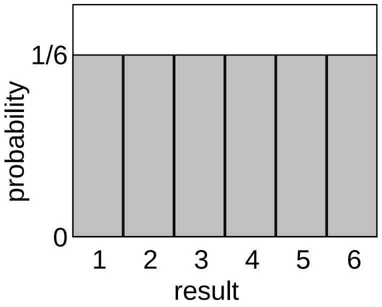
A. Newtonian physics is an essentially perfect approximation for describing the motion of a pair of dice. If Newtonian physics is deterministic, why do we consider the result of rolling dice to be random?

B. Why isn't it valid to define randomness by saying that randomness is when all the outcomes are equally likely?

C. The sequence of digits 121212121212121212 seems clearly nonrandom, and 41592653589793 seems random. The latter sequence, however, is the decimal form of pi, starting with the third digit. There is a story about the Indian mathematician Ramanujan, a self-taught prodigy, that a friend came to visit him in a cab, and remarked that the number of the cab, 1729, seemed relatively uninteresting. Ramanujan replied that on the contrary, it was very interesting because it was the smallest number that could be represented in two different ways as the sum of two cubes. The Argentine author Jorge Luis Borges wrote a short story called "The Library of Babel," in which he imagined a library containing every book that could possibly be written using the letters of the alphabet. It would include a book containing only the repeated letter "a"; all the ancient Greek tragedies known today, all the lost Greek tragedies, and millions of Greek tragedies that were never actually written; your own life story, and various incorrect versions of your own life story; and countless anthologies containing a short story called "The Library of Babel." Of course, if you picked a book from the shelves of the library, it would almost certainly look like a nonsensical sequence of letters and punctuation, but it's always possible that the seemingly meaningless book would be a science-fiction screenplay written in the language of a Neanderthal tribe, or a set of incomparably beautiful love poems written in a language that never existed. In view of these examples, what does it really mean to say that something is random?

3.3 Probability Distributions

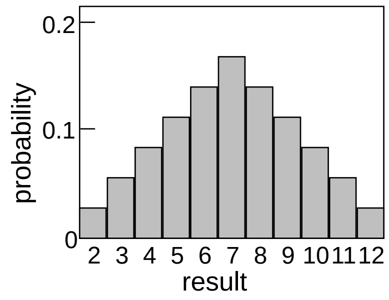
So far we've discussed random processes having only two possible outcomes: yes or no, win or lose, on or off. More generally, a random process could have a result that is a number. Some processes yield integers, as when you roll a die and get a result from one to six, but some are not restricted to whole numbers, for example the number of seconds that a uranium-238 atom will exist before undergoing radioactive decay.



(a) Probability distribution for the result of rolling a single die.

Consider a throw of a die. If the die is "honest," then we expect all six values to be equally likely. Since all six probabilities must add up to 1, then probability of any particular value coming up must be 1/6. We can summarize this in a graph, (a). Areas under the curve can be interpreted as total probabilities. For instance, the area under the curve from 1 to 3 is 1/6+1/6+1/6=1/2, so the probability of getting a result from 1 to 3 is 1/2. The function shown on the graph is called the probability distribution.

(1) Most people would think they were positively correlated, but it's possible that they're independent. (2) These must be independent, since there is no possible physical mechanism that could make one have any effect on the other. (3) These cannot be independent, since dying today guarantees that you won't die tomorrow.

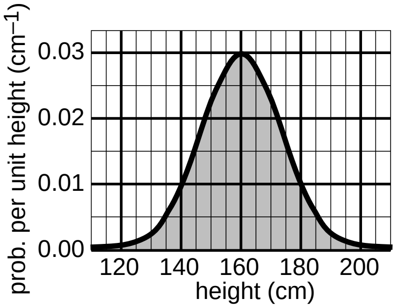


(b) Rolling two dice and adding them up.

Figure (b) shows the probabilities of various results obtained by rolling two dice and adding them together, as in the game of craps. The probabilities are not all the same. There is a small probability of getting a two, for example, because there is only one way to do it, by rolling a one and then another one. The probability of rolling a seven is high because there are six different ways to do it: 1+6, 2+5, etc.

If the number of possible outcomes is large but finite, for example the number of hairs on a dog, the graph would start to look like a smooth curve rather than a ziggurat.

What about probability distributions for random numbers that are not integers? We can no longer make a graph with probability on the *y* axis, because the probability of getting a given exact number is typically zero. For instance, there is zero probability that a radioactive atom will last for *exactly* 3 seconds, since there is are infinitely many possible results that are close to 3 but not exactly three: 2.99999999999999996876876587658465436, for example. It doesn't usually make sense, therefore, to talk about the probability of a single numerical result, but it does make sense to talk about the probability of a certain range of results. For instance, the probability that an atom will last more than 3 and less than 4 seconds is a perfectly reasonable thing to discuss. We can still summarize the probability information on a graph, and we can still interpret areas under the curve as probabilities.



(c) A probability distribution for height of human adults. (Not real data.)

But the y axis can no longer be a unitless probability scale. In radioactive decay, for example, we want the x axis to have units of time, and we want areas under the curve to be unitless probabilities. The area of a single square on the graph paper is then

(unitless area of a square)
= (width of square with time units) x (height of square) .

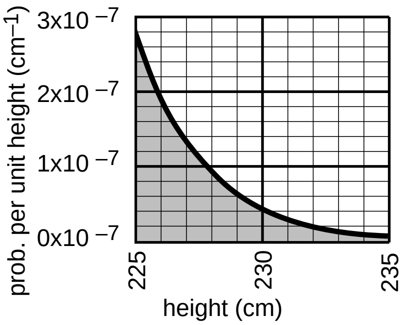
If the units are to cancel out, then the height of the square must evidently be a quantity with units of inverse time. In other words, the y axis of the graph is to be interpreted as probability per unit time, not probability.

Figure (c) shows another example, a probability distribution for people's height. This kind of bell-shaped curve is quite common.

Self-Check

Compare the number of people with heights in the range of 130-135 cm to the number in the range 135-140.

The area under the curve from 130 to 135 cm is about 3/4 of a rectangle. The area from 135 to 140 cm is about 1.5 rectangles. The number of people in the second range is about twice as much. We could have converted these to actual probabilities (1 rectangle = 5 cm x 0.005 cm-1 = 0.025), but that would have been pointless because we were just going to compare the two areas.



(d) A close-up of the right-hand tail of the distribution shown in the previous figure.

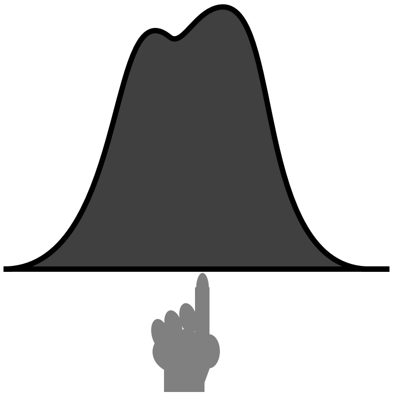
Example: Looking for tall basketball players

Question: A certain country with a large population wants to find very tall people to be on its Olympic basketball team and strike a blow against western imperialism. Out of a pool of 10^8 people who are the right age and gender, how many are they likely to find who are over 225 cm (7'4") in height? Figure (d) gives a close-up of the “tails” of the distribution shown previously.

Solution: The shaded area under the curve represents the probability that a given person is tall enough. Each rectangle represents a probability of $0.2\text{x}10^{-7}$ cm^{-1} x 1 cm = $2\text{x}10^{-9}$. There are about 35 rectangles covered by the shaded area, so the probability of having a height greater than 230 cm is $7\text{x}10^{-8}$, or just under one in ten million. Using the rule for calculating averages, the average, or expected number of people this tall is $(10^8)\text{x}(7\text{x}10^{-8})=7$.

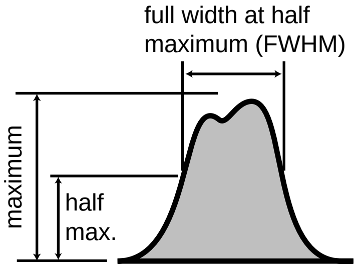
Average and width of a probability distribution

If the next Martian you meet asks you, “How tall is an adult human?,” you will probably reply with a statement about the average human height, such as “Oh, about 5 feet 6 inches.” If you wanted to explain a little more, you could say, “But that's only an average. Most people are somewhere between 5 feet and 6 feet tall.” Without bothering to draw the relevant bell curve for your new extraterrestrial acquaintance, you've summarized the relevant information by giving an average and a typical range of variation.



(e) The average of a probability distribution.

The average of a probability distribution can be defined geometrically as the horizontal position at which it could be balanced if it was constructed out of cardboard. A convenient numerical measure of the amount of variation about the average, or amount of uncertainty, is the full width at half maximum, or FWHM, shown in the figure.



(f) The full width at half maximum (FWHM) of a probability distribution.

A great deal more could be said about this topic, and indeed an introductory statistics course could spend months on ways of defining the center and width of a distribution. Rather than force-feeding you on mathematical detail or techniques for calculating these things, it is perhaps more relevant to point out simply that there are various ways of defining them, and to inoculate you against the misuse of certain definitions.

The average is not the only possible way to say what is a typical value for a quantity that can vary randomly; another possible definition is the median, defined as the value that is exceeded with 50% probability. When discussing incomes of people living in a certain town, the average could be very misleading, since it can be affected massively if a single resident of the town is Bill Gates. Nor is the FWHM the only possible way of stating the amount of random variation; another possible way of measuring it is the standard deviation (defined as the square root of the average squared deviation from the average value).

3.4 Exponential Decay and Half-Life

Most people know that radioactivity "lasts a certain amount of time," but that simple statement leaves out a lot. As an example, consider the following medical procedure used to diagnose thyroid function. A very small quantity of the isotope ^{131}I, produced in a nuclear reactor, is fed to or injected into the patient. The body's biochemical systems treat this artificial, radioactive isotope exactly the same as ^{127}I, which is the only naturally occurring type. (Nutritionally, iodine is a necessary trace element. Iodine taken into the body is partly excreted, but the rest becomes concentrated in the thyroid gland. Iodized salt has had iodine added to it to prevent the nutritional deficiency known as goiters, in which the iodine-starved thyroid becomes swollen.) As the ^{131}I undergoes beta decay, it emits electrons, neutrinos, and gamma rays. The gamma rays can be measured by a detector passed over the patient's body. As the radioactive iodine becomes concentrated in the thyroid, the amount of gamma radiation coming from the thyroid becomes greater, and that emitted by the rest of the body is reduced. The rate at which the iodine concentrates in the thyroid tells the doctor about the health of the thyroid.

If you ever undergo this procedure, someone will presumably explain a little about radioactivity to you, to allay your fears that you will turn into the Incredible Hulk, or that your next child will have an unusual number of limbs. Since iodine stays in your thyroid for a long time once it gets there, one thing you'll want to know is whether your thyroid is going to become radioactive forever. They may just tell you that the radioactivity "only lasts a certain amount of time," but we can now carry out a quantitative derivation of how the radioactivity really will die out.

Let $P_{\text{surv}}(t)$ be the probability that an iodine atom will survive without decaying for a period of at least t. It has been experimentally measured that half all ^{131}I atoms decay in 8 hours, so we have

$$P_{\text{surv}}(8\text{ hr}) \quad = 0.5 \quad .$$

Now using the law of independent probabilities, the probability of surviving for 16 hours equals the probability of surviving for the first 8 hours multiplied by the probability of surviving for the second 8 hours,

$$\begin{aligned} P_{\text{surv}}(16\text{ hr}) \quad &= 0.5\times 0.5 \\ &= 0.25 \quad . \end{aligned}$$

Similarly we have

$$\begin{aligned} P_{\text{surv}}(24\text{ hr}) \quad &= 0.5\times 0.5\times 0.5 \\ &= 0.125 \quad . \end{aligned}$$

Generalizing from this pattern, the probability of surviving for any time t that is a multiple of 8 hours is

$$P_{\text{surv}}(\text{t}) \quad = 0.5^{\text{t} / (8\text{ hr})}$$

We now know how to find the probability of survival at intervals of 8 hours, but what about the points in time in between? What would be the probability of surviving for 4 hours? Well, using the law of independent probabilities again, we have

$$P_{\text{surv}}(8\text{ hr}) = P_{\text{surv}}(4\text{ hr}) \text{ x } P_{\text{surv}}(4\text{ hr}) \quad ,$$

which can be rearranged to give

$$\begin{aligned} P_{\text{surv}}(4\text{ hr}) &= \sqrt{P_{\text{surv}}(8\text{ hr})} \\ &= \sqrt{0.5} \\ &= 0.707 \quad . \end{aligned}$$

This is exactly what we would have found simply by plugging in $P_{\text{surv}}(\text{t}) = 0.5^{\text{t} / (8\text{ hr})} = 0.5^{1/2}$ and ignoring the restriction to multiples of 8 hours. Since 8 hours is the amount of time required for half of the atoms to decay, it is known as the half-life, written $t_{1/2}$. The general rule is as follows:

Exponential Decay Formula

$$P_{\text{surv}}(\text{t}) = 0.5^{t / t_{1/2}}$$

Using the rule for calculating averages, we can also find the number of atoms, $N(t)$, remaining in a sample at time t:

$$N(t) = N(0) \times 0.5^{t / t_{1/2}}$$

Both of these equations have graphs that look like dying-out exponentials, as in the example below.

Example: Radioactive contamination at Chernobyl

Question: One of the most dangerous radioactive isotopes released by the Chernobyl disaster in 1986 was ^{90}Sr, whose half-life is 28 years. (a) How long will it be before the contamination is reduced to one tenth of its original level? (b) If a total of 10^{27} atoms was released, about how long would it be before not a single atom was left?

Solution: (a) We want to know the amount of time that a ^{90}Sr nucleus has a probability of 0.1 of surviving. Starting with the exponential decay formula,

$$P_{\text{surv}} = 0.5^{t / t_{1/2}} \quad ,$$

we want to solve for t. Taking natural logarithms of both sides,

$$\ln P = \frac{t}{t_{1/2}} \ln 0.5 \quad ,$$

so

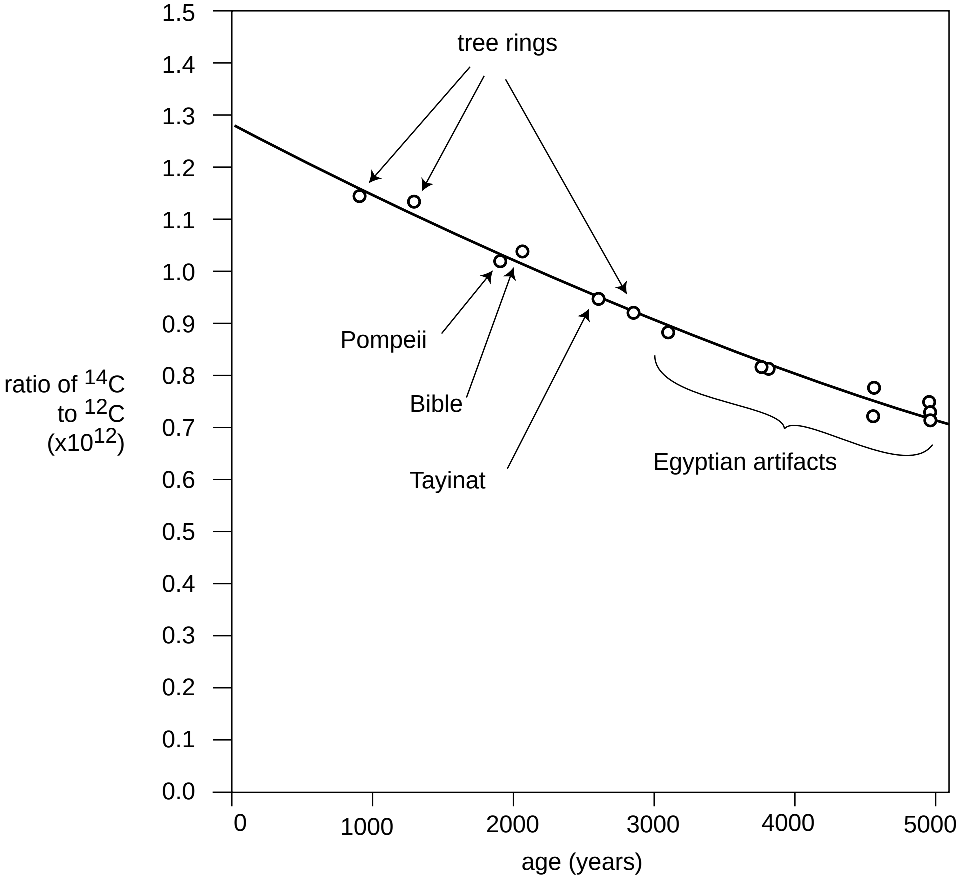
$$t = \frac{t_{1/2} \ln P}{\ln 0.5}$$

Plugging in P=0.1 and $t_{1/2}$=28 years, we get t=93 years.

(b) This is just like the first part, but P=10^{-27}. The result is about 2500 years.

Example: ^{14}C Dating

Almost all the carbon on Earth is ^{12}C, but not quite. The isotope ^{14}C, with a half-life of 5600 years, is produced by cosmic rays in the atmosphere. It decays naturally, but is replenished at such a rate that the fraction of ^{14}C in the atmosphere remains constant, at 1.3×10^{-12}. Living plants and animals take in both ^{12}C and ^{14}C from the atmosphere and incorporate both into their bodies. Once the living organism dies, it no longer takes in C atoms from the atmosphere, and the proportion of ^{14}C gradually falls off as it undergoes radioactive decay. This effect can be used to find the age of dead organisms, or human artifacts made from plants or animals. The following graph shows the exponential decay curve of ^{14}C in various objects. Similar methods, using longer-lived isotopes, provided the first firm proof that the earth was billions of years old, not a few thousand as some had claimed on religious grounds.



Calibration of the ^{14}C dating method using tree rings and artifacts whose ages were known from other methods. Redrawn from Emilio Segrè, **Nuclei and Particles**, 1965.

Rate of decay

If you want to find how many radioactive decays occur within a time interval lasting from time t to time $t+\Delta t$, the most straightforward approach is to calculate it like this:

$$\begin{aligned}
&(\text{number of decays between } t \text{ and } t+\Delta\text{t})\\
&\quad = N(t) - N(t+\Delta t)\\
&\quad = N(0)\left[P_{\text{surv}}(t) - P_{\text{surv}}(t+\Delta t)\right]\\
&\quad = N(0)\left[0.5^{t/t_{1/2}} - 0.5^{(t+\Delta t)/t_{1/2}}\right]\\
&\quad = N(0)\left[1 - 0.5^{\Delta t/t_{1/2}}\right]0.5^{t/t_{1/2}}
\end{aligned}$$

A problem arises when Δt is small compared to $t_{1/2}$. For instance, suppose you have a hunk of 10^{22} atoms of ^{235}U, with a half-life of 700 million years, which is $2.2\text{x}10^{16}$ s. You want to know how many decays will occur in Δt=1 s. Since we're specifying the current number of atoms, t=0. As you plug in to the formula above on your calculator, the quantity $0.5^{\Delta t/t_{1/2}}$ comes out on your calculator to equal one, so the final result is zero. That's incorrect, though. In reality, $0.5^{\Delta t/t_{1/2}}$ should equal 0.999999999999999968, but your calculator only gives eight digits of precision, so it rounded it off to one. In other words, the probability that a ^{235}U atom will survive for 1 s is very close to one, but not equal to one. The number of decays in one second is therefore $3.2\text{x}10^5$, not zero.

Well, my calculator only does eight digits of precision, just like yours, so how did I know the right answer? The way to do it is to use the following approximation:

$$a^b \approx 1 + b \ln a, \text{ if } b << 1$$

(The symbol << means "is much less than.") Using it, we can find the following approximation:

$$\begin{aligned}
&(\text{number of decays between } t \text{ and } t+\Delta t)\\
&\quad = N(0)\left[1 - 0.5^{\Delta t/t_{1/2}}\right]0.5^{t/t_{1/2}}\\
&\quad \approx N(0)\left[1 - \left(1 + \frac{\Delta t}{t_{1/2}}\ln 0.5\right)\right]0.5^{t/t_{1/2}}, \text{ if } \Delta t << t_{1/2}\\
&\quad = (\ln 2)N(0)\left(0.5^{t/t_{1/2}}\right)\frac{\Delta t}{t_{1/2}}
\end{aligned}$$

This also gives us a way to calculate the rate of decay, i.e. the number of decays per unit time. Dividing by Δt on both sides, we have

$$(\text{decays per unit time}) \approx \frac{(\ln 2)N(0)}{t_{1/2}}0.5^{t/t_{1/2}}, \text{ if } \Delta t << t_{1/2}$$

Example: The hot potato

Question: A nuclear physicist with a demented sense of humor tosses you a cigar box, yelling “hot potato.” The label on the box says “contains 10^{20} atoms of ^{17}F, half-life of 66 s, produced today in our reactor at 1 p.m.” It takes you two seconds to read the label, after which you toss it behind some lead bricks and run away. The time is 1:40 p.m. Will you die?

Solution: The time elapsed since the radioactive fluorine was produced in the reactor was 40 minutes, or 2400 s. The number of elapsed half-lives is therefore $t / t_{1/2} = 36$. The initial number of atoms was $N(0)=10^{20}$. The number of decays per second is now about $10^7\ \text{s}^{-1}$, so it produced about $2\text{x}10^7$ high-energy electrons while you held it in your hands. Although twenty million electrons sounds like a lot, it is not really enough to be dangerous.

By the way, none of the equations we’ve derived so far was the actual probability distribution for the time at which a particular radioactive atom will decay. That probability distribution would be found by substituting $N(0)=1$ into the equation for the rate of decay.

If the sheer number of equations is starting to seem formidable, let’s pause and think for a second. The simple equation for P_surv is something you can derive easily from the law of independent probabilities any time you need it. From that, you can quickly find the exact equation for the rate of decay. The derivation of the approximate equations for $\Delta t << t$ is a little hairier, but note that except for the factors of ln 2, everything in these equations can be found simply from considerations of logic and units. For instance, a longer half-life will obviously lead to a slower rate of decays, so it makes sense that we divide by it. As for the ln 2 factors, they are exactly the kind of thing that one looks up in a book when one needs to know them.

Discussion Questions

A. In the medical procedure involving ^{131}I, why is it the gamma rays that are detected, not the electrons or neutrinos that are also emitted?

B. For 1 s, Fred holds in his hands 1 kg of radioactive stuff with a half-life of 1000 years. Ginger holds 1 kg of a different substance, with a half-life of 1 min, for the same amount of time. Did they place themselves in equal danger, or not?

C. How would you interpret it if you calculated $N(t)$, and found it was less than one?

D. Does the half-life depend on how much of the substance you have? Does the expected time until the sample decays completely depend on how much of the substance you have?

3.5∫ Applications of Calculus

The area under the probability distribution is of course an integral. If we call the random number x and the probability distribution $D(x)$, then the probability that x lies in a certain range is given by

$$\text{(probability of } a \le x \le b\text{)} = \int_a^b D(x)\,dx \quad .$$

What about averages? If x had a finite number of equally probable values, we would simply add them up and divide by how many we had. If they weren't equally likely, we'd make the weighted average $x_1P_1+x_2P_2+\ldots$ But we need to generalize this to a variable x that can take on any of a continuum of values. The continuous version of a sum is an integral, so the average is

$$\text{(average value of } x\text{)} = \int x\,D(x)\,dx \quad ,$$

where the integral is over all possible values of x.

Example: Probability distribution for radioactive decay

Here is a rigorous justification for the statement in the previous section that the probability distribution for radioactive decay is found by substituting $N(0)=1$ into the equation for the rate of decay. We know that the probability distribution must be of the form

$$D(x) \quad = k\,0.5^{t/t_{1/2}} \quad ,$$

where k is a constant that we need to determine. The atom is guaranteed to decay eventually, so normalization gives us

$$\begin{aligned}&\text{(probability of } 0 \le t \le \infty\text{)}\\ &\quad = 1\\ &\quad = \int_0^\infty D(t)\,dt\end{aligned}$$

The integral is most easily evaluated by converting the function into an exponential with e as the base

$$\begin{aligned}D(x) &= k\exp\left[\ln\left(0.5^{t/t_{1/2}}\right)\right]\\ &= k\exp\left[\frac{t}{t_{1/2}}\ln\left(0.5\right)\right]\\ &= k\exp\left(-\frac{\ln 2}{t_{1/2}}t\right) \quad ,\end{aligned}$$

which give an integral of the familiar form $\int e^{cx}dx = \frac{1}{c}e^{cx}$. We thus have

$$1 \quad = -\frac{k\,t_{1/2}}{\ln 2}\exp\left(-\frac{\ln 2}{t_{1/2}}t\right)\Bigg]_0^\infty \quad ,$$

which gives the desired result:

$$k \quad = \frac{\ln 2}{t_{1/2}} \quad .$$

Example: Average lifetime

You might think that the half-life would also be the average lifetime of an atom, since half the atoms' lives are shorter and half longer. But the half whose lives are longer include some that survive for many half-lives, and these rare long-lived atoms skew the average. We can calculate the average lifetime as follows:

(average lifetime)

$$= \int_0^\infty t\, D(t)\, dt$$

Using the convenient base-*e* form again, we have

(average lifetime)

$$= \frac{\ln 2}{t_{1/2}} \int_0^\infty t \exp\left(-\frac{\ln 2}{t_{1/2}} t\right) dt \quad .$$

This integral is of a form that can either be attacked with integration by parts or by looking it up in a table. The result is $\int x e^{cx} dx = \frac{x}{c} e^{cx} - \frac{1}{c^2} e^{cx}$, and the first term can be ignored for our purposes because it equals zero at both limits of integration. We end up with

(average lifetime)

$$= \frac{\ln 2}{t_{1/2}} \left(\frac{t_{1/2}}{\ln 2}\right)^2$$

$$= \frac{t_{1/2}}{\ln 2}$$

$$= 1.443\, t_{1/2} \quad ,$$

which is, as expected, longer than one half-life.

Summary

Selected Vocabulary

probability .... the likelihood that something will happen, expressed as a number between zero and one

normalization .... the property of probabilities that the sum of the probabilities of all possible outcomes must equal one

independence .... the lack of any relationship between two random events

probability distribution ..... a curve that specifies the probabilities of various random values of a variable; areas under the curve correspond to probabilities

FWHM .... the full width at half-maximum of a probability distribution; a measure of the width of the distribution

half-life .... the amount of time that a radioactive atom has a probability of 1/2 of surviving without decaying

Notation

P	probability
$t_{1/2}$	half-life
D	a probability distribution (used only in optional section 3.5)

Summary

Quantum physics differs from classical physics in many ways, the most dramatic of which is that certain processes at the atomic level, such as radioactive decay, are random rather than deterministic. There is a method to the madness, however: quantum physics still rules out any process that violates conservation laws, and it also offers methods for calculating probabilities numerically.

In this chapter we focused on certain generic methods of working with probabilities, without concerning ourselves with any physical details. Without knowing any of the details of radioactive decay, for example, we were still able to give a fairly complete treatment of the relevant probabilities. The most important of these generic methods is the law of independent probabilities, which states that if two random events are not related in any way, then the probability that they will both occur equals the product of the two probabilities,

$$\text{probability of A and B} = P_A P_B\text{, if A and B are independent} \quad .$$

The most important application is to radioactive decay. The time that a radioactive atom has a 50% chance of surviving is called the half-life, $t_{1/2}$. The probability of surviving for two half-lives is (1/2)(1/2)=1/4, and so on. In general, the probability of surviving a time t is given by

$$P_{surv} = 0.5^{t/t_{1/2}} \quad .$$

Related quantities such as the rate of decay and probability distribution for the time of decay are given by the same type of exponential function, but multiplied by certain constant factors.

Homework Problems

1. If a radioactive substance has a half-life of one year, does this mean that it will be completely decayed after two years? Explain.

2. What is the probability of rolling a pair of dice and getting “snake eyes,” i.e. both dice come up with ones?

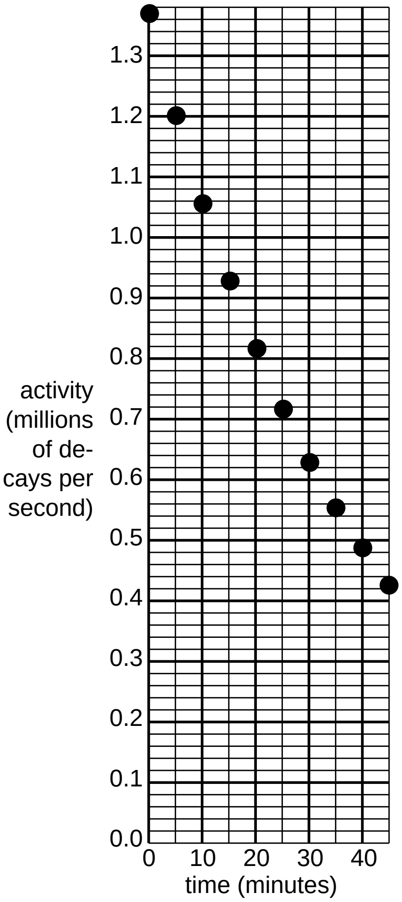
3. Use a calculator to check the approximation that $a^b \approx 1 + b \ln a$, if $b << 1$ using some arbitrary numbers. See how good the approximation is for values of b that are not quite as small compared to one.

4. Make up an example of a numerical problem involving a rate of decay where $\Delta t << t_{1/2}$, but $0.5^{t/t_{1/2}}$ can still be evaluated on a calculator without getting something that rounds off to one. Check that you get approximately the same result using both methods to calculate the number of decays between t and $t+\Delta t$.

5. (a) A nuclear physicist is studying a nuclear reaction caused in an accelerator experiment, with a beam of ions from the accelerator striking a thin metal foil and causing nuclear reactions when a nucleus from one of the beam ions happens to hit one of the nuclei in the target. After the experiment has been running for a few hours, a few billion radioactive atoms have been produced, embedded in the target. She does not know what nuclei are being produced, but she suspects they are an isotope of some heavy element such as Pb, Bi, Fr or U. Following one such experiment, she takes the target foil out of the accelerator, sticks it in front of a detector, measures the activity every 5 min, and makes a graph (figure). The isotopes she thinks may have been produced are:

isotope	half-life (minutes)
^{211}Pb	36.1
^{214}Pb	26.8
^{214}Bi	19.7
^{223}Fr	21.8
^{239}U	23.5

Which one is it?



Problem 5.

(b) Having decided that the original experimental conditions produced

S A solution is given in the back of the book.
✓ A computerized answer check is available.
★ A difficult problem.
∫ A problem that requires calculus.

one specific isotope, she now tries using beams of ions traveling at several different speeds, which may cause different reactions. The following table gives the activity of the target 10, 20 and 30 minutes after the end of the experiment, for three different ion speeds.

	activity (millions of decays/s) after...		
	10 min	20 min	30 min
first ion speed	1.933	0.832	0.382
second ion speed	1.200	0.545	0.248
third ion speed	6.544	1.296	0.248

Since such a large number of decays is being counted, assume that the data are only inaccurate due to rounding off when writing down the table. Which are consistent with the production of a single isotope, and which imply that more than one isotope was being created?

6. Devise a method for testing experimentally the hypothesis that a gambler's chance of winning at craps is independent of her previous record of wins and losses.

7. Refer to the probability distribution for people's heights in section 3.3.

(a) Show that the graph is properly normalized.

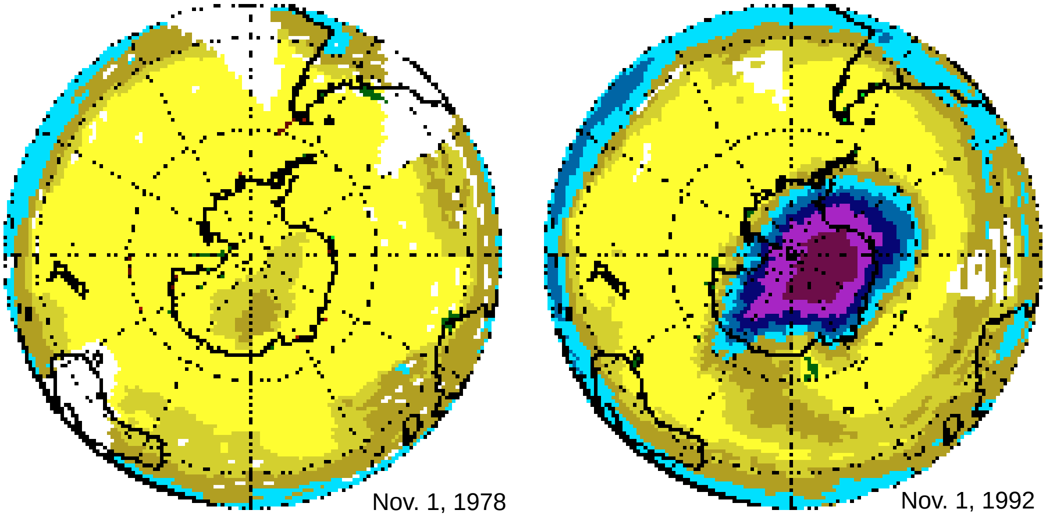
(b) Estimate the fraction of the population having heights between 140 and 150 cm.

8. ^{238}U decays be alpha emission, with a half-life of 4.5x10^{9} years. The subsequent chain of alpha and electron (beta) decays involves much shorter half-lives, and terminates in the stable nucleus ^{206}Pb. Almost all natural uranium is ^{238}U. All helium on earth is from the decay chain that leads from ^{238}U to ^{206}Pb.

(a) How many alphas are emitted per decay chain? [Hint: Use conservation of mass.]

(b) How many electrons are emitted per decay chain? [Hint: Use conservation of charge.]

(c ✓) Each alpha particles ends up claiming two electrons and becoming a helium atom. If the original ^{238}U atom is in solid rock (as opposed to the earth's molten regions), the He atoms are unable to diffuse out of the rock. Suppose a geologist finds a sample of hardened lava, melts it in a furnace, and finds that it contains 1230 mg of uranium and 2.3 mg of helium. How long has it been since the lava originally hardened?



In recent decades, a huge hole in the ozone layer has spread out from Antarctica.

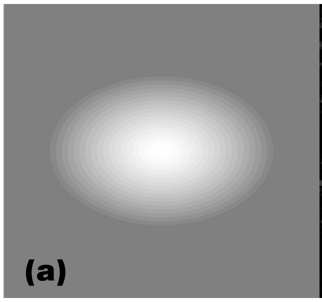
4 Light as a Particle

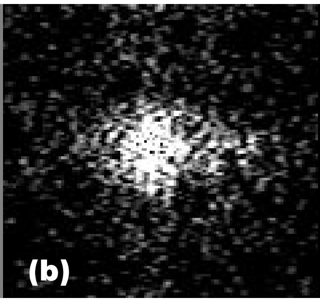
> The only thing that interferes with my learning is my education.
>
> Albert Einstein

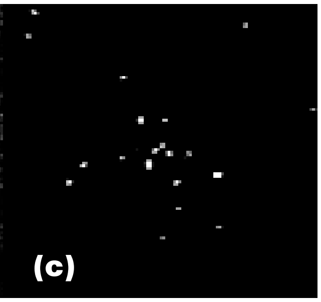
Radioactivity is random, but do the laws of physics exhibit randomness in other contexts besides radioactivity? Yes. Radioactive decay was just a good playpen to get us started with concepts of randomness, because all atoms of a given isotope are identical. By stocking the playpen with an unlimited supply of identical atom-toys, nature helped us to realize that their future behavior could be different regardless of their original identicality. We are now ready to leave the playpen, and see how randomness fits into the structure of physics at the most fundamental level.

The laws of physics describe light and matter, and the quantum revolution rewrote both descriptions. Radioactivity was a good example of matter's behaving in a way that was inconsistent with classical physics, but if we want to get under the hood and understand how nonclassical things happen, it will be easier to focus on light rather than matter. A radioactive atom such as uranium-235 is after all an extremely complex system, consisting of 92 protons, 143 neutrons, and 92 electrons. Light, however, can be a simple sine wave.

However successful the classical wave theory of light had been — allowing the creation of radio and radar, for example — it still failed to describe many important phenomena. An example that is currently of great interest is the way the ozone layer protects us from the dangerous short-wavelength ultraviolet part of the sun's spectrum. In the classical description, light is a wave. When a wave passes into and back out of a medium, its frequency is unchanged, and although its wavelength is altered while it is in the medium, it returns to its original value when the wave reemerges. Luckily for us, this is not at all what ultraviolet light does when it passes through the ozone layer, or the layer would offer no protection at all!



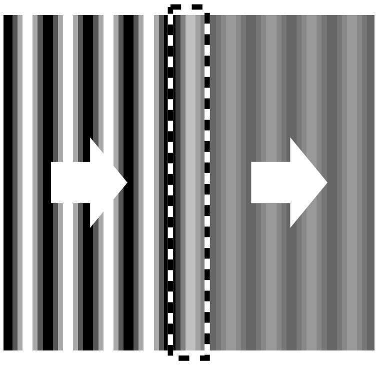




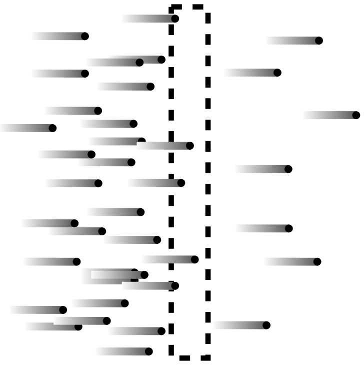
4.1 Evidence for Light as a Particle

For a long time, physicists tried to explain away the problems with the classical theory of light as arising from an imperfect understanding of atoms and the interaction of light with individual atoms and molecules. The ozone paradox, for example, could have been attributed to the incorrect assumption that one could think of the ozone layer as a smooth, continuous substance, when in reality it was made of individual ozone molecules. It wasn't until 1905 that Albert Einstein threw down the gauntlet, proposing that the problem had nothing to do with the details of light's interaction with atoms and everything to do with the fundamental nature of light itself.

In those days the data were sketchy, the ideas vague, and the experiments difficult to interpret; it took a genius like Einstein to cut through the thicket of confusion and find a simple solution. Today, however, we can get right to the heart of the matter with a piece of ordinary consumer electronics, the digital camera. Instead of film, a digital camera has a computer chip with its surface divided up into a grid of light-sensitive squares, called "pixels." Compared to a grain of the silver compound used to make regular photographic film, a digital camera pixel is activated by an amount of light energy orders of magnitude smaller. We can learn something new about light by using a digital camera to detect smaller and smaller amounts of light, as shown in figures (a) through (c) above. Figure (a) is fake, but (b) and (c) are real digital-camera images made by Prof. Lyman Page of Princeton University as a classroom demonstration. Figure (a) is what we would see if we used the digital camera to take a picture of a fairly dim source of light. In figures (b) and (c), the intensity of the light was drastically reduced by inserting semitransparent absorbers like the tinted plastic used in sunglasses. Going from (a) to (b) to (c), more and more light energy is being thrown away by the absorbers.


(d)

The results are drastically different from what we would expect based on the wave theory of light. If light was a wave and nothing but a wave, (d), then the absorbers would simply cut down the wave's amplitude across the whole wavefront. The digital camera's entire chip would be illuminated uniformly, and weakening the wave with an absorber would just mean that every pixel would take a long time to soak up enough energy to register a signal.


(e)

But figures (b) and (c) show that some pixels take strong hits while others pick up no energy at all. Instead of the wave picture, the image that is naturally evoked by the data is something more like a hail of bullets from a machine gun, (e). Each "bullet" of light apparently carries only a tiny

Einstein and Seurat: twins separated at birth.
Seine Grande Jatte by Georges Seurat (19th century)

amount of energy, which is why detecting them individually requires a sensitive digital camera rather than an eye or a piece of film.

Although Einstein was interpreting different observations, this is the conclusion he reached in his 1905 paper: that the pure wave theory of light is an oversimplification, and that the energy of a beam of light comes in finite chunks rather than being spread smoothly throughout a region of space.

We now think of these chunks as particles of light, and call them “photons,” although Einstein avoided the word “particle,” and the word “photon” was invented later. Regardless of words, the trouble was that waves and particles seemed like inconsistent categories. The reaction to Einstein’s paper could be kindly described as vigorously skeptical. Even twenty years later, Einstein wrote, “There are therefore now two theories of light, both indispensable, and — as one must admit today despite twenty years of tremendous effort on the part of theoretical physicists — without any logical connection.” In the remainder of this chapter we will learn how the seeming paradox was eventually resolved.

Discussion Questions

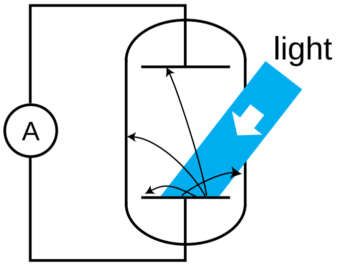
A. Suppose someone rebuts the digital camera data, claiming that the random pattern of dots occurs not because of anything fundamental about the nature of light but simply because the camera’s pixels are not all exactly the same. How could we test this interpretation?

B. Discuss how the correspondence principle applies to the observations and concepts discussed so far.

4.2 How Much Light Is One Photon?

The photoelectric effect

We have seen evidence that light energy comes in little chunks, so the next question to be asked is naturally how much energy is in one chunk. The most straightforward experimental avenue for addressing this question is a phenomenon known as the photoelectric effect. The photoelectric effect occurs when a photon strikes the surface of a solid object and knocks out an electron. It occurs continually all around you. It is happening right now at the surface of your skin and on the paper or computer screen from which you are reading these words. It does not ordinarily lead to any observable electrical effect, however, because on the average free electrons are wandering back in just as frequently as they are being ejected. (If an object did somehow lose a significant number of electrons, its growing net positive charge would begin attracting the electrons back more and more strongly.)



(a) Apparatus for observing the photoelectric effect. A beam of light strikes a capacitor plate inside a vacuum tube, and electrons are ejected (black arrows).

Figure (a) shows a practical method for detecting the photoelectric effect. Two very clean parallel metal plates (the electrodes of a capacitor) are sealed inside a vacuum tube, and only one plate is exposed to light. Because there is a good vacuum between the plates, any ejected electron that happens to be headed in the right direction will almost certainly reach the other capacitor plate without colliding with any air molecules.

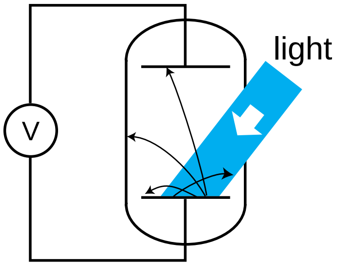
The illuminated (bottom) plate is left with a net positive charge, and the unilluminated (top) plate acquires a negative charge from the electrons deposited on it. There is thus an electric field between the plates, and it is because of this field that the electrons' paths are curved, as shown in the diagram. However, since vacuum is a good insulator, any electrons that reach the top plate are prevented from responding to the electrical attraction by jumping back across the gap. Instead they are forced to make their way around the circuit, passing through an ammeter. The ammeter allows a measurement of the strength of the photoelectric effect.

An unexpected dependence on frequency

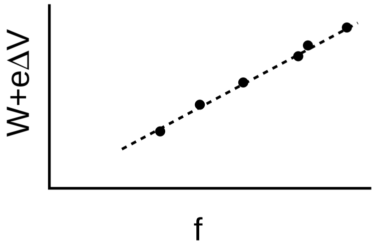
The photoelectric effect was discovered serendipitously by Heinrich Hertz in 1887, as he was experimenting with radio waves. He was not particularly interested in the phenomenon, but he did notice that the effect was produced strongly by ultraviolet light and more weakly by lower frequencies. Light whose frequency was lower than a certain critical value did not eject any electrons at all. (In fact this was all prior to Thomson's discovery of the electron, so Hertz would not have described the effect in terms of electrons — we are discussing everything with the benefit of hindsight.) This dependence on frequency didn't make any sense in terms of the classical wave theory of light. A light wave consists of electric and magnetic fields. The stronger the fields, i.e. the greater the wave's amplitude, the greater the forces that would be exerted on electrons that found themselves bathed in the light. It should have been amplitude (brightness) that was relevant, not frequency. The dependence on frequency not only proves that the wave model of light needs modifying, but with the proper interpretation it allows us to determine how much energy is in one photon, and it also leads to a connection between the wave and particle models that we need in order to reconcile them.

To make any progress, we need to consider the physical process by which a photon would eject an electron from the metal electrode. A metal contains electrons that are free to move around. Ordinarily, in the interior of the metal, such an electron feels attractive forces from atoms in every direction around it. The forces cancel out. But if the electron happens to find itself at the surface of the metal, the attraction from the interior side is not balanced out by any attraction from outside. Bringing the electron out through the surface therefore requires a certain amount of work, W, which depends on the type of metal used.

Suppose a photon strikes an electron, annihilating itself and giving up all its energy to the electron. (We now know that this is what always happens in the photoelectric effect, although it had not yet been established in 1905 whether or not the photon was completely annihilated.) The



(b) A different way of studying the photoelectric effect.



(c) The quantity $W+e\Delta V$ indicates the energy of one photon. It is found to be proportional to the frequency of the light.

Historical Note

What I'm presenting in this chapter is a simplified explanation of how the photon *could* have been discovered. The actual history is more complex.

Max Planck (1858-1947) began the photon saga with a theoretical investigation of the spectrum of light emitted by a hot, glowing object. He introduced quantization of the energy of light waves, in multiples of *hf*, purely as a mathematical trick that happened to produce the right results. Planck did not believe that his procedure could have any physical significance. In his 1905 paper Einstein took Planck's quantization as a description of reality, and applied it to various theoretical and experimental puzzles, including the photoelectric effect.

Millikan then subjected Einstein's ideas to a series of rigorous experimental tests. Although his results matched Einstein's predictions perfectly, Millikan was skeptical about photons, and his papers conspicuously omit any reference to them. Only in his autobiography did Millikan rewrite history and claim that he had given experimental proof for photons.

electron will (1) lose kinetic energy through collisions with other electrons as it plows through the metal on its way to the surface; (2) lose an amount of kinetic energy equal to W as it emerges through the surface; and (3) lose more energy on its way across the gap between the plates, due to the electric field between the plates. Even if the electron happens to be right at the surface of the metal when it absorbs the photon, and even if the electric field between the plates has not yet built up very much, W is the bare minimum amount of energy that it must receive from the photon if it is to contribute to a measurable current. The reason for using very clean electrodes is to minimize W and make it have a definite value characteristic of the metal surface, not a mixture of values due to the various types of dirt and crud that are present in tiny amounts on all surfaces in everyday life.

We can now interpret the frequency dependence of the photoelectric effect in a simple way: apparently the amount of energy possessed by a photon is related to its frequency. A low-frequency red or infrared photon has an energy less than W, so a beam of them will not produce any current. A high-frequency blue or violet photon, on the other hand, packs enough of a punch to allow an electron to make it to the other plate. At frequencies higher than the minimum, the photoelectric current continues to increase with the frequency of the light because of effects (1) and (3).

Numerical relationship between energy and frequency

Prompted by Einstein's photon paper, Robert Millikan (whom we encountered in book 4 of this series) figured out how to use the photoelectric effect to probe precisely the link between frequency and photon energy. Rather than going into the historical details of Millikan's actual experiments (a lengthy experimental program that occupied a large part of his professional career) we will describe a simple version, shown in figure (b), that is used sometimes in college laboratory courses. The idea is simply to illuminate one plate of the vacuum tube with light of a single wavelength and monitor the voltage difference between the two plates as they charge up. Since the resistance of a voltmeter is very high (much higher than the resistance of an ammeter), we can assume to a good approximation that electrons reaching the top plate are stuck there permanently, so the voltage will keep on increasing for as long as electrons are making it across the vacuum tube.

At a moment when the voltage difference has a reached a value ΔV, the minimum energy required by an electron to make it out of the bottom plate and across the gap to the other plate is $W+e\Delta V$. As ΔV increases, we eventually reach a point at which $W+e\Delta V$ equals the energy of one photon. No more electrons can cross the gap, and the reading on the voltmeter stops rising. The quantity $W+e\Delta V$ now tells us the energy of one photon. If we determine this energy for a variety of wavelengths, (c), we find the following simple relationship between the energy of a photon and the frequency of the light:

$$E = hf \quad ,$$

where h is a constant having a numerical value of $6.63\text{x}10^{-34}$ J·s. Note how the equation brings the wave and particle models of light under the same roof: the left side is the energy of one *particle* of light, while the right side is the frequency of the same light, interpreted as a *wave*. The constant h is known as Planck's constant (see historical note).

Self-Check

How would you extract h from the graph in figure (c)?

Since the energy of a photon is hf, a beam of light can only have energies of hf, $2hf$, $3hf$, etc. Its energy is quantized — there is no such thing as a fraction of a photon. Quantum physics gets its name from the fact that it quantizes quantities like energy, momentum, and angular momentum that had previously been thought to be smooth, continuous and infinitely divisible.

Example: number of photons emitted by a lightbulb per second
Question: Roughly how many photons are emitted by a 100-W lightbulb in 1 second?
Solution: People tend to remember wavelengths rather than frequencies for visible light. The bulb emits photons with a range of frequencies and wavelengths, but let's take 600 nm as a typical wavelength for purposes of estimation. The energy of a single photon is

$$\begin{aligned} E_{\text{photon}} &= hf \\ &= hc/\lambda \end{aligned}$$

A power of 100 W means 100 joules per second, so the number of photons is

$$\begin{aligned} (100\ \text{J})/E_{\text{photon}} & \\ &= (100\ \text{J}) / (hc/\lambda) \\ &\approx 3\text{x}10^{20} \end{aligned}$$

Example: Momentum of a photon
Question: According to the theory of relativity, the momentum of a beam of light is given by $p=E/c$ (see ch. 2, homework problem #6). Apply this to find the momentum of a single photon in terms of its frequency, and in terms of its wavelength.
Solution: Combining the equations $p=E/c$ and $E=hf$, we find

$$\begin{aligned} p &= E/c \\ &= \frac{h}{c}f \quad . \end{aligned}$$

To reexpress this in terms of wavelength, we use $c=f\lambda$:

$$\begin{aligned} p &= \frac{h}{c}\cdot\frac{c}{\lambda} \\ &= \frac{h}{\lambda} \end{aligned}$$

The second form turns out to be simpler.

Discussion Questions

A. Only a very tiny percentage of the electrons available near the surface of an object is ever ejected by the photoelectric effect. How well does this agree the wave model of light, and how well with the particle model?

B. What is the significance of the fact that Planck's constant is numerically very small? How would our everyday experience of light be different if it was not so small?

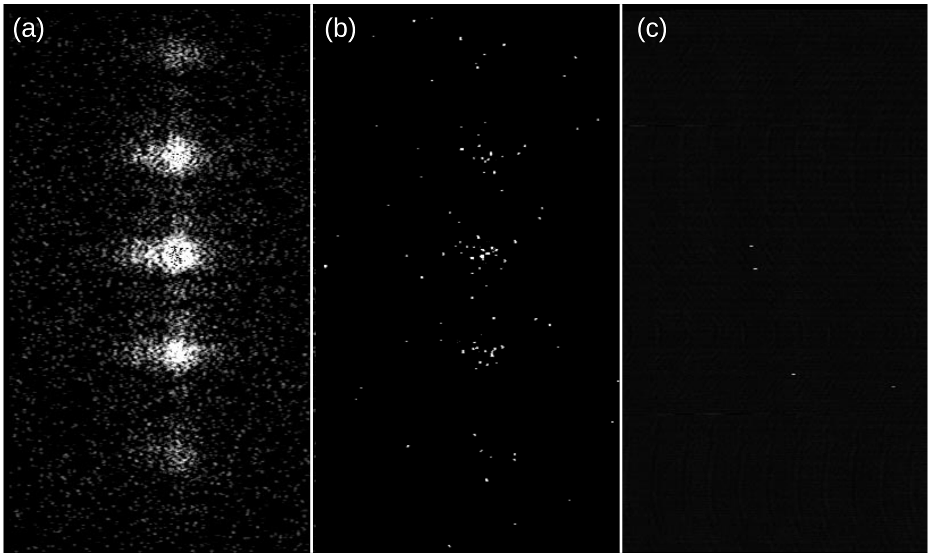
C. How would the experiments described above be affected if electrons were likely to get hit by more than one photon?

D. Draw some representative trajectories of electrons for $\Delta V=0$, ΔV less than the maximum value, and ΔV greater than the maximum value.

E. Explain based on the photon theory of light why ultraviolet light would be more likely than visible or infrared light to cause cancer by damaging DNA molecules.

F. Does $E=hf$ imply that a photon changes its energy when it passes from one transparent material into another substance with a different index of refraction?

The axes of the graph are frequency and photon energy, so its slope is Planck's constant.

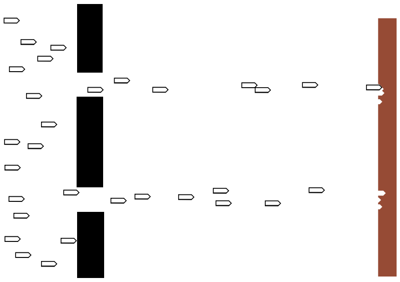


Wave interference patterns photographed by Prof. Lyman Page with a digital camera. Laser light with a single well-defined wavelength passed through a series of absorbers to cut down its intensity, then through a set of slits to produce interference, and finally into a digital camera chip. (A triple slit was actually used, but for conceptual simplicity we discuss the results in the main text as if it was a double slit.) In figure (b) the intensity has been reduced relative to (a), and even more so for figure (c).

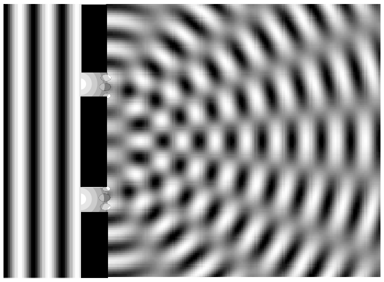
4.3 Wave-Particle Duality

How can light be both a particle and a wave? We are now ready to resolve this seeming contradiction. Often in science when something seems paradoxical, it's because we (1) don't define our terms carefully, or (2) don't test our ideas against any specific real-world situation. Let's define particles and waves as follows:

Waves exhibit superposition, and specifically interference phenomena.
Particles can only exist in whole numbers, not fractions



(d) Bullets pass through a double slit.



(e) A water wave passes through a double slit.

As a real-world check on our philosophizing, there is one particular experiment that works perfectly. We set up a double-slit interference experiment that we know will produce a diffraction pattern if light is an honest-to-goodness wave, but we detect the light with a detector that is capable of sensing individual photons, e.g. a digital camera. To make it possible to pick out individual dots from individual photons, we must use filters to cut down the intensity of the light to a very low level, just as in the photos by Prof. Page in section 4.1. The whole thing is sealed inside a light-tight box. The results are shown in figures (a), (b), and (c) above. (In fact, the similar figures in section 4.1 are simply cutouts from these figures.)

Neither the pure wave theory nor the pure particle theory can explain the results. If light was only a particle and not a wave, there would be no interference effect. The result of the experiment would be like firing a hail of bullets through a double slit, (d). Only two spots directly behind the slits would be hit.

If, on the other hand, light was only a wave and not a particle, we would get the same kind of diffraction pattern that would happen with a water wave, (e). There would be no discrete dots in the photo, only a diffraction pattern that shaded smoothly between light and dark.

Applying the definitions to this experiment, light must be both a particle and a wave. It is a wave because it exhibits interference effects. At the same time, the fact that the photographs contain discrete dots is a direct demonstration that light refuses to be split into units of less than a single photon. There can only be whole numbers of photons: four photons in figure (c), for example.

A wrong interpretation: photons interfering with each other

One possible interpretation of wave-particle duality that occurred to physicists early in the game was that perhaps the interference effects came from photons interacting with each other. By analogy, a water wave consists of moving water molecules, and interference of water waves results ultimately from all the mutual pushes and pulls of the molecules. This interpretation was conclusively disproved by G.I. Taylor, a student at Cambridge. The demonstration by Prof. Page that we've just been discussing is essentially a modernized version of Taylor's work. Taylor reasoned that if interference effects came from photons interacting with each other, a bare minimum of two photons would have to be present at the same time to produce interference. By making the light source extremely dim, we can be virtually certain that there are never two photons in the box at the same time. In figure (c), however, the intensity of the light has been cut down so much by the absorbers that if it was in the open, the average separation between photons would be on the order of a kilometer! At any given moment, the number of photons in the box is most likely to be zero. It is virtually certain that there were never two photons in the box at once.

A single photon can go through both slits.

The concept of a photon's path is undefined.

If a single photon can demonstrate double-slit interference, then which slit did it pass through? The unavoidable answer must be that it passes through both! This might not seem so strange if we think of the photon as a wave, but it is highly counterintuitive if we try to visualize it as a particle. The moral is that we should not think in terms of the path of a photon. Like the fully human and fully divine Jesus of Christian theology, a photon is supposed to be 100% wave and 100% particle. If a photon had a well defined path, then it would not demonstrate wave superposition and interference effects, contradicting its wave nature. (In the next chapter we will discuss the Heisenberg uncertainty principle, which gives a numerical way of approaching this issue.)

Another wrong interpretation: the pilot wave hypothesis

A second possible explanation of wave-particle duality was taken seriously in the early history of quantum mechanics. What if the photon *particle* is like a surfer riding on top of its accompanying *wave*? As the wave travels along, the particle is pushed, or "piloted" by it. Imagining the particle and the wave as two separate entities allows us to avoid the seemingly paradoxical idea that a photon is both at once. The wave happily does its wave tricks, like superposition and interference, and the particle acts like a respectable particle, resolutely refusing to be in two different places at once. If the wave, for instance, undergoes destructive interference, becoming nearly zero in a particular region of space, then the particle simply is not guided into that region.

The problem with the pilot wave interpretation is that the only way it can be experimentally tested or verified is if someone manages to detach the particle from the wave, and show that there really are two entities involved, not just one. Part of the scientific method is that hypotheses are supposed to be experimentally testable. Since nobody has ever managed to separate the wavelike part of a photon from the particle part, the interpretation is not useful or meaningful in a scientific sense.

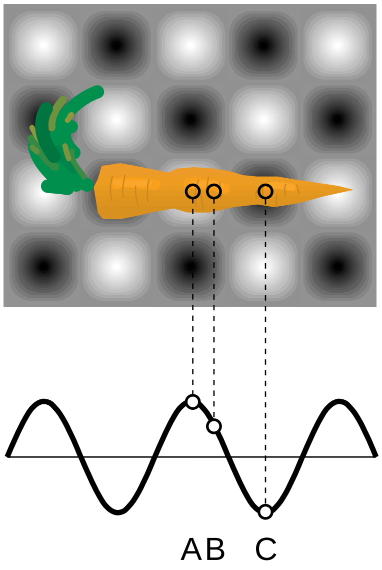
The probability interpretation

The correct interpretation of wave-particle duality is suggested by the random nature of the experiment we've been discussing: even though every photon wave/particle is prepared and released in the same way, the location at which it is eventually detected by the digital camera is different every time. The idea of the probability interpretation of wave-particle duality is that the location of the photon-particle is random, but the probability that it is in a certain location is higher where the photon-wave's amplitude is greater.

More specifically, the probability distribution of the particle must be proportional to the *square* of the wave's amplitude,

$$\text{(probability distribution)} \propto \text{(amplitude)}^2 \quad .$$

This follows from the correspondence principle and from the fact that a wave's energy density is proportional to the square of its amplitude. If we run the double-slit experiment for a long enough time, the pattern of dots fills in and becomes very smooth as would have been expected in classical physics. To preserve the correspondence between classical and quantum physics, the amount of energy deposited in a given region of the picture over the long run must be proportional to the square of the wave's amplitude. The amount of energy deposited in a certain area depends on the number of photons picked up, which is proportional to the probability of finding any given photon there.



Example: a microwave oven
Question: The figure shows two-dimensional (top) and one-dimensional (bottom) representations of the standing wave inside a microwave oven. Gray represents zero field, and white and black signify the strongest fields, with white being a field that is in the opposite direction compared to black. Compare the probabilities of detecting a microwave photon at points A, B, and C.
Solution: A and C are both extremes of the wave, so the probabilities of detecting a photon at A and C are equal. It doesn't matter that we have represented C as negative and A as positive, because it is the square of the amplitude that is relevant. The amplitude at B is about 1/2 as much as the others, so the probability of detecting a photon there is about 1/4 as much.

The probability interpretation was disturbing to physicists who had spent their previous careers working in the deterministic world of classical physics, and ironically the most strenuous objections against it were raised by Einstein, who had invented the photon concept in the first place. The probability interpretation has nevertheless passed every experimental test, and is now as well established as any part of physics.

An aspect of the probability interpretation that has made many people uneasy is that the process of detecting and recording the photon's position seems to have a magical ability to get rid of the wavelike side of the photon's personality and force it to decide for once and for all where it really wants to be. But detection or measurement is after all only a physical process like any other, governed by the same laws of physics. We will postpone a detailed discussion of this issue until the following chapter, since a measuring device like a digital camera is made of matter, but we have so far only discussed how quantum mechanics relates to light.

Example: What is the proportionality constant?

Question: What is the proportionality constant that would make an actual equation out of (probability distribution)$\propto$(amplitude)^2?

Solution: The probability that the photon is in a certain small region of volume v should equal the fraction of the wave's energy that is within that volume:

$$\begin{aligned} P &= \frac{\text{energy in volume } v}{\text{energy of photon}} \\ &= \frac{\text{energy in volume } v}{hf} \end{aligned}$$

We assume v is small enough so that the electric and magnetic fields are nearly constant throughout it. We then have

$$P = \frac{\left(\frac{1}{8\pi k}\left|\mathbf{E}\right|^2 + \frac{1}{2\mu_o}\left|\mathbf{B}\right|^2\right)v}{hf} \quad .$$

We can simplify this formidable looking expression by recognizing that in an electromagnetic wave, $|\mathbf{E}|$ and $|\mathbf{B}|$ are related by $|\mathbf{E}|=c|\mathbf{B}|$. With some algebra, it turns out that the electric and magnetic fields each contribute half the total energy (see book 4, ch. 6, homework problem #5), so we can simplify this to

$$\begin{aligned} P &= \frac{2\left(\frac{1}{8\pi k}\left|\mathbf{E}\right|^2\right)v}{hf} \\ &= \frac{v}{4\pi khf}\left|\mathbf{E}\right|^2 \quad . \end{aligned}$$

As advertised, the probability is proportional to the square of the wave's amplitude.

Discussion Questions

A. Referring back to the example of the carrot in the microwave oven, show that it would be nonsensical to have probability be proportional to the field itself, rather than the square of the field.

B. Einstein did not try to reconcile the wave and particle theories of light, and did not say much about their apparent inconsistency. Einstein basically visualized a beam of light as a stream of bullets coming from a machine gun. In the photoelectric effect, a photon "bullet" would only hit one atom, just as a real bullet would only hit one person. Suppose someone reading his 1905 paper wanted to interpret it by saying that Einstein's so-called particles of light were simply short wave-trains that only occupy a small region of space. Comparing the wavelength of visible light (a few hundred nm) to the size of an atom (on the order of 0.1 nm), explain why this poses a difficulty for reconciling the particle and wave theories.

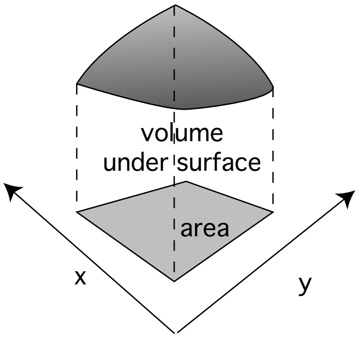
C. Can a white photon exist?

D. In double-slit diffraction of photons, would you get the same pattern of dots on the photo if you covered one slit? Why should it matter whether you give the photon two choices or only one?

4.4 Photons in Three Dimensions

Up until now I've been sneaky and avoided a full discussion of the three-dimensional aspects of the probability interpretation. The example of the carrot in the microwave oven, for example, reduced to a one-dimensional situation because we were considering three points along the same line and because we were only comparing ratios of probabilities. The purpose of bringing it up now is to head off any feeling that you've been cheated conceptually rather than to prepare you for mathematical problem solving in three dimensions, which would not be appropriate for the level of this course.

A typical example of a probability distribution in chapter 3 was the distribution of heights of human beings. The thing that varied randomly, height, *h*, had units of meters, and the probability distribution was a graph of a function *D*(*h*). The units of the probability distribution had to be m^{-1} (inverse meters) so that areas under the curve, interpreted as probabilities, would be unitless (area = width x height = m x m^{-1}).



Now suppose we have a two-dimensional problem, e.g. the probability distribution for the place on the surface of a digital camera chip where a photon will be detected. The point where it is detected would be described with two variables, *x* and *y*, each having units of meters. The probability distribution will be a function of both variables, *D*(*x*,*y*). A probability is now visualized as the volume under the surface described by the function *D*(*x*,*y*), as shown in the figure. The units of *D* must be m^{-2} so that probabilities will be unitless (area = width x depth x height = m x m x m^{-2}).

Generalizing finally to three dimensions, we find by analogy that the probability distribution will be a function of all three coordinates, *D*(*x*,*y*,*z*), and will have units of m^{-3}. It is unfortunately impossible to visualize the graph unless you are a mutant with a natural feel for life in four dimensions. If the probability distribution is nearly constant within a certain volume of space *v*, the probability that the photon is in that volume is simply *vD*. If you know enough calculus, it should be clear that this can be generalized to $P=\int D\,\text{d}x\,\text{d}y\,\text{d}z$ if *D* is not constant.

Summary

Selected Vocabulary

photon	a particle of light
photoelectric effect	the ejection, by a photon, of an electron from the surface of an object
wave-particle duality	the idea that light is both a wave and a particle

Summary

Around the turn of the twentieth century, experiments began to show problems with the classical wave theory of light. In any experiment sensitive enough to detect very small amounts of light energy, it becomes clear that light energy cannot be divided into chunks smaller than a certain amount. Measurements involving the photoelectric effect demonstrate that this smallest unit of light energy equals hf, where f is the frequency of the light and h is a number known as Planck's constant. We say that light energy is quantized in units of hf, and we interpret this quantization as evidence that light has particle properties as well as wave properties. Particles of light are called photons.

The only method of reconciling the wave and particle natures of light that has stood the test of experiment is the probability interpretation. It states that the probability that the particle is at a given location is proportional to the square of the amplitude of the wave at that location.

One important consequence of wave-particle duality is that we must abandon the concept of the path the particle takes through space. To hold on to this concept, we would have to contradict the well established wave nature of light, since a wave can spread out in every direction simultaneously.

Homework Problems

1. When light is reflected from a mirror, perhaps only 80% of the energy comes back. One could try to explain this in two different ways: (1) 80% of the photons are reflected, or (2) all the photons are reflected, but each loses 20% of its energy. Based on your everyday knowledge about mirrors, how can you tell which interpretation is correct? [Based on a problem from PSSC Physics.]

2. Suppose we want to build an electronic light sensor using an apparatus like the one described in the section on the photoelectric effect. How would its ability to detect different parts of the spectrum depend on the type of metal used in the capacitor plates?

3. The photoelectric effect can occur not just for metal cathodes but for any substance, including living tissue. Ionization of DNA molecules in can cause cancer or birth defects. If the energy required to ionize DNA is on the same order of magnitude as the energy required to produce the photoelectric effect in a metal, which of these types of electromagnetic waves might pose such a hazard? Explain.

60 Hz waves from power lines
100 MHz FM radio
microwaves from a microwave oven
visible light
ultraviolet light
x-rays

4✓. The beam of a 100-W overhead projector covers an area of 1 m x 1 m when it hits the screen 3 m away. Estimate the number of photons that are in flight at any given time. (Since this is only an estimate, we can ignore the fact that the beam is not parallel.)

5✓. In the photoelectric effect, electrons are observed with virtually no time delay (~10 ns), even when the light source is very weak. (A weak light source does however only produce a small number of ejected electrons.) The purpose of this problem is to show that the lack of a significant time delay contradicted the classical wave theory of light, so throughout this problem you should put yourself in the shoes of a classical physicist and pretend you don't know about photons at all. At that time, it was thought that the electron might have a radius on the order of 10^{-15} m. (Recent experiments have shown that if the electron has any finite size at all, it is far smaller.)

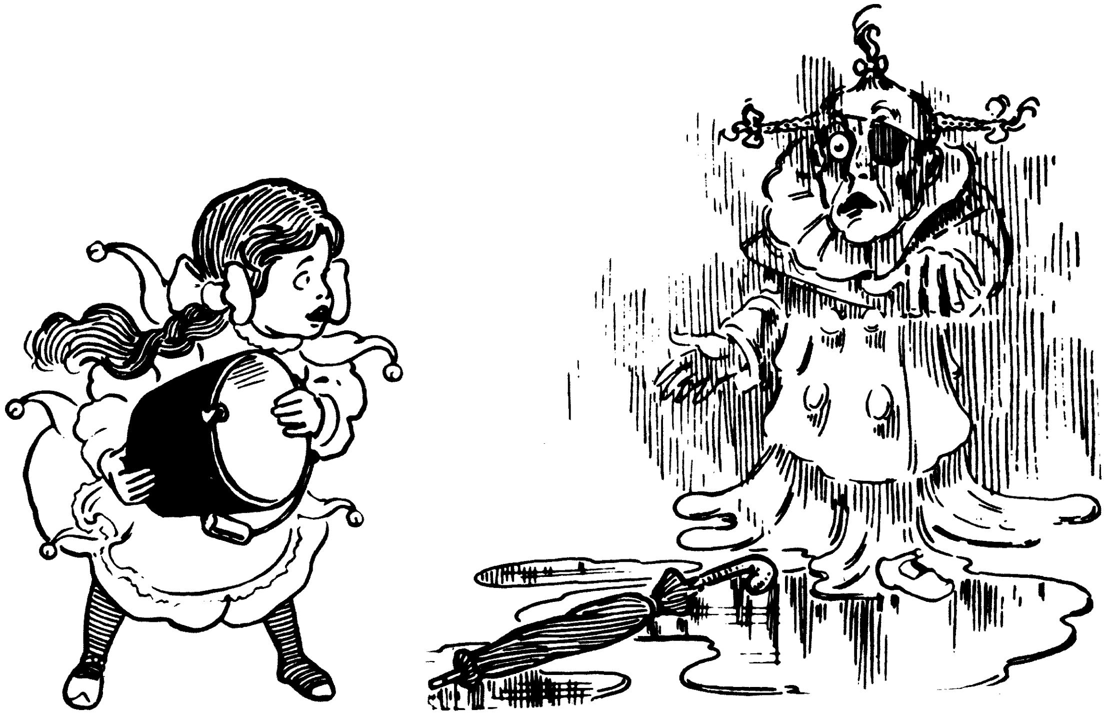
(a) Estimate the power that would be soaked up by a single electron in a beam of light with an intensity of 1 mW/m^2.

(b) The energy, *W*, required for the electron to escape through the surface of the cathode is on the order of 10^{-19} J. Find how long it would take the electron to absorb this amount of energy, and explain why your result constitutes strong evidence that there is something wrong with the classical theory.

S A solution is given in the back of the book.	★ A difficult problem.
✓ A computerized answer check is available.	∫ A problem that requires calculus.

6. A photon collides with an electron and rebounds from the collision at 180 degrees, i.e. going back along the path on which it came. The rebounding photon has a different energy, and therefore a different frequency and wavelength. Show that, based on conservation of energy and momentum, the difference between the photon's initial and final wavelengths must be $2h/mc$, where m is the mass of the electron. The experimental verification of this type of "pool-ball" behavior by Arthur Compton in 1923 was taken as definitive proof of the particle nature of light.

7★. Generalize the result of the previous problems to the case where the photon bounces off at an angle other than 180° with respect to its initial direction of motion.



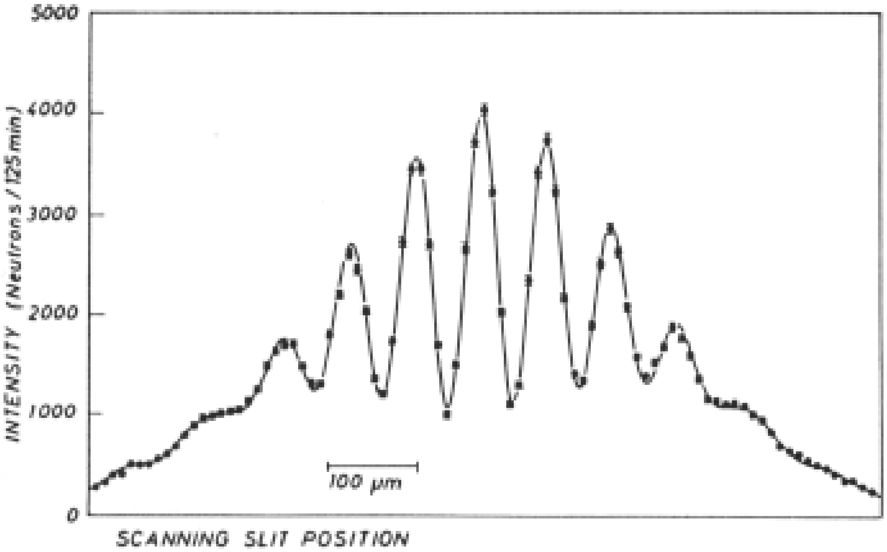
5 Matter as a Wave

> [In] a few minutes I shall be all melted... I have been wicked in my day, but I never thought a little girl like you would ever be able to melt me and end my wicked deeds. Look out — here I go!
>
> The Wicked Witch of the West

As the Wicked Witch learned the hard way, losing molecular cohesion can be unpleasant. That's why we should be very grateful that the concepts of quantum physics apply to matter as well as light. If matter obeyed the laws of classical physics, molecules wouldn't exist.

Consider, for example, the simplest atom, hydrogen. Why does one hydrogen atom form a chemical bond with another hydrogen atom? Roughly speaking, we'd expect a neighboring pair of hydrogen atoms, A and B, to exert no force on each other at all, attractive or repulsive: there are two repulsive interactions (proton A with proton B and electron A with electron B) and two attractive interactions (proton A with electron B and electron A with proton B). Thinking a little more precisely, we should even expect that once the two atoms got close enough, the interaction would be repulsive. For instance, if you squeezed them so close together that the two protons were almost on top of each other, there would be a tremendously strong repulsion between them due to the $1/r^2$ nature of the electrical force. The repulsion between the electrons would not be as strong, because each electron ranges over a large area, and is not likely to be found right on top of the other electron. Thus hydrogen molecules should not exist according to classical physics.

Quantum physics to the rescue! As we'll see shortly, the whole problem is solved by applying the same quantum concepts to electrons that we have already used for photons.



A double-slit interference pattern made with neutrons. (A. Zeilinger, R. Gähler, C.G. Shull, W. Treimer, and W. Mampe, *Reviews of Modern Physics*, Vol. 60, 1988.)

5.1 Electrons as Waves

We started our journey into quantum physics by studying the random behavior of *matter* in radioactive decay, and then asked how randomness could be linked to the basic laws of nature governing *light*. The probability interpretation of wave-particle duality was strange and hard to accept, but it provided such a link. It is now natural to ask whether the same explanation could be applied to matter. If the fundamental building block of light, the photon, is a particle as well as a wave, is it possible that the basic units of matter, such as electrons, are waves as well as particles?

A young French aristocrat studying physics, Louis de Broglie (pronounced "broylee"), made exactly this suggestion in his 1923 Ph.D. thesis. His idea had seemed so farfetched that there was serious doubt about whether to grant him the degree. Einstein was asked for his opinion, and with his strong support, de Broglie got his degree.

Only two years later, American physicists C.J. Davisson and L. Germer confirmed de Broglie's idea by accident. They had been studying the scattering of electrons from the surface of a sample of nickel, made of many small crystals. (One can often see such a crystalline pattern on a brass doorknob that has been polished by repeated handling.) An accidental explosion occurred, and when they put their apparatus back together they observed something entirely different: the scattered electrons were now creating an interference pattern! This dramatic proof of the wave nature of matter came about because the nickel sample had been melted by the explosion and then resolidified as a single crystal. The nickel atoms, now nicely arranged in the regular rows and columns of a crystalline lattice, were acting as the lines of a diffraction grating. The new crystal was analogous to the type of ordinary diffraction grating in which the lines are etched on the surface of a mirror (a reflection grating) rather than the kind in which the light passes through the transparent gaps between the lines (a transmission grating).

Although we will concentrate on the wave-particle duality of electrons because it is important in chemistry and the physics of atoms, all the other "particles" of matter you've learned about show wave properties as well. The figure above, for instance, shows a wave interference pattern of neutrons.

It might seem as though all our work was already done for us, and there would be nothing new to understand about electrons: they have the same kind of funny wave-particle duality as photons. That's almost true, but not quite. There are some important ways in which electrons differ significantly from photons:

(1) Electrons have mass, and photons don't.

(2) Photons always move at the speed of light, but electrons can move at any speed less than c.

(3) Photons don’t have electric charge, but electrons do, so electric forces can act on them. The most important example is the atom, in which the electrons are held by the electric force of the nucleus.

(4) Electrons cannot be absorbed or emitted as photons are. Destroying an electron or creating one out of nothing would violate conservation of charge.

(In chapter 6 we will learn of one more fundamental way in which electrons differ from photons, for a total of five.)

Because electrons are different from photons, it is not immediately obvious which of the photon equations from the previous chapter can be applied to electrons as well. A particle property, the energy of one photon, is related to its wave properties via $E=hf$ or, equivalently, $E=hc/\lambda$. The momentum of a photon was given by $p=hf/c$ or $p=h/\lambda$. Ultimately it was a matter of experiment to determine which of these equations, if any, would work for electrons, but we can make a quick and dirty guess simply by noting that some of the equations involve c, the speed of light, and some do not. Since c is irrelevant in the case of an electron, we might guess that the equations of general validity are those that do not have c in them:

$$E = hf$$
$$p = h/\lambda$$

This is essentially the reasoning that de Broglie went through, and experiments have confirmed these two equations for all the fundamental building blocks of light and matter, not just for photons and electrons.

The second equation, which I soft-pedaled in the previous chapter, takes on a greater important for electrons. This is first of all because the momentum of matter is more likely to be significant than the momentum of light under ordinary conditions, and also because force is the transfer of momentum, and electrons are affected by electrical forces.

Discussion Question

Frequency is oscillations per second, whereas wavelength is meters per oscillation. How could the equations $E = hf$ and $p = h/\lambda$ be made to look more alike by using quantities that were more closely analogous? How would this more symmetric treatment relate to incorporating relativity into quantum mechanics?

Example: the wavelength of an elephant

Question: What is the wavelength of a trotting elephant?

Solution: One may doubt whether the equation should be applied to an elephant, which is not just a single particle but a rather large collection of them. Throwing caution to the wind, however, we estimate the elephant's mass at 10^3 kg and its trotting speed at 10 m/s. Its wavelength is therefore roughly

$$\begin{aligned}\lambda &= h/p \\ &= \frac{h}{mv} \\ &= \frac{6.63\times10^{-34}\,\mathrm{J{\cdot}s}}{(10^3\ \mathrm{kg})(10\ \mathrm{m/s})} \\ &\sim 10^{-37}\frac{\left(\mathrm{kg{\cdot}m^2/s^2}\right)\cdot\mathrm{s}}{\mathrm{kg}\cdot\mathrm{m/s}} \\ &= 10^{-37}\ \mathrm{m}\end{aligned}$$

The wavelength found in this example is so fantastically small that we can be sure we will never observe any measurable wave phenomena with elephants or any other human-scale objects. The result is numerically small because Planck's constant is so small, and as in some examples encountered previously, this smallness is in accord with the correspondence principle.

Although a smaller mass in the equation $\lambda=h/mv$ does result in a longer wavelength, the wavelength is still quite short even for individual electrons under typical conditions, as shown in the following example.

Example: the typical wavelength of an electron

Question: Electrons in circuits and in atoms are typically moving through potential differences on the order of 1 V, so that a typical energy is (e)(1 V), which is on the order of 10^{-19} J. What is the wavelength of an electron with this amount of kinetic energy?

Solution: This energy is nonrelativistic, since it is much less than mc^2. Momentum and energy are therefore related by the nonrelativistic equation $KE=p^2/2m$. Solving for p and substituting in to the equation for the wavelength, we find

$$\begin{aligned}\lambda &= \frac{h}{\sqrt{2m{\cdot}KE}} \\ &= 1.6\mathrm{x}10^{-9}\ \mathrm{m}\end{aligned}$$

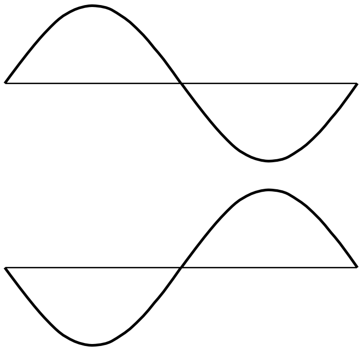
This is on the same order of magnitude as the size of an atom, which is no accident: as we will discuss in the next chapter in more detail, an electron in an atom can be interpreted as a standing wave. The smallness of the wavelength of a typical electron also helps to explain why the wave nature of electrons wasn't discovered until a hundred years after the wave nature of light. To scale the usual wave-optics devices such as diffraction gratings down to the size needed to work with electrons at ordinary energies, we need to make them so small that their parts are comparable in size to individual atoms. This is essentially what Davisson and Germer did with their nickel crystal.

Self-Check

These remarks about the inconvenient smallness of electron wavelengths apply only under the assumption that the electrons have typical energies. What kind of energy would an electron have to have in order to have a longer wavelength that might be more convenient to work with?

What kind of wave is it?

If a sound wave is a vibration of matter, and a photon is a vibration of electric and magnetic fields, what kind of a wave is an electron made of? The disconcerting answer is that there is no experimental “observable,” i.e. directly measurable quantity, to correspond to the electron wave itself. In other words, there are devices like microphones that detect the oscillations of air pressure in a sound wave, and devices such as radio receivers that measure the oscillation of the electric and magnetic fields in a light wave, but nobody has ever found any way to measure the electron wave directly.



These two electron waves are not distinguishable by any measuring device.

We can of course detect the energy (or momentum) possessed by an electron just as we could detect the energy of a photon using a digital camera. (In fact I’d imagine that an unmodified digital camera chip placed in a vacuum chamber would detect electrons just as handily as photons.) But this only allows us to determine where the wave carries high probability and where it carries low probability. Probability is proportional to the square of the wave’s amplitude, but measuring its square is not the same as measuring the wave itself. In particular, we get the same result by squaring either a positive number or its negative, so there is no way to determine the positive or negative sign of an electron wave.

Most physicists tend toward the school of philosophy known as operationalism, which says that a concept is only meaningful if we can define some set of operations for observing, measuring, or testing it. According to a strict operationalist, then, the electron wave itself is a meaningless concept. Nevertheless, it turns out to be one of those concepts like love or humor that is impossible to measure and yet very useful to have around. We therefore give it a symbol, Ψ (the capital Greek letter psi), and a special name, the electron *wavefunction* (because it is a function of the coordinates x, y, and z that specify where you are in space). It would be impossible, for example, to calculate the shape of the electron wave in a hydrogen atom without having some symbol for the wave. But when the calculation produces a result that can be compared directly to experiment, the final algebraic result will turn out to involve only Ψ^2, which is what is observable, not Ψ itself.

Since Ψ, unlike $\boldsymbol{E}$ and $\boldsymbol{B}$, is not directly measurable, we are free to make the probability equations have a simple form: instead of having the probability density equal to some funny constant multiplied by Ψ^2, we simply define Ψ so that the constant of proportionality is one:

$$\text{(probability density)} = \Psi^2 \quad .$$

Since the probability density has units of m^{-3}, the units of Ψ must be $\text{m}^{-3/2}$.

Wavelength is inversely proportional to momentum, so to produce a large wavelength we would need to use electrons with very *small* momenta and energies. (In practical terms, this isn’t very easy to do, since ripping an electron out of an object is a violent process, and it’s not so easy to calm the electrons down afterward.)

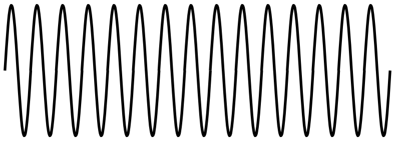
5.2*∫ Dispersive Waves

A colleague of mine who teaches chemistry loves to tell the story about an exceptionally bright student who, when told of the equation $p=h/\lambda$, protested, “But when I derived it, it had a factor of 2!” The issue that’s involved is a real one, albeit one that could be glossed over (and is, in most textbooks) without raising any alarms in the mind of the average student. The present optional section addresses this point; it is intended for the student who wishes to delve a little deeper.

Here’s how the now-legendary student was presumably reasoning. We start with the equation $v=f\lambda$, which is valid for any sine wave, whether it’s quantum or classical. Let’s assume we already know $E=hf$, and are trying to derive the relationship between wavelength and momentum:

$$\begin{aligned}\lambda &= v/f \\ &= \frac{vh}{E} \\ &= \frac{vh}{\frac{1}{2}mv^2} \\ &= \frac{2h}{mv} \\ &= \frac{2h}{p} \quad .\end{aligned}$$

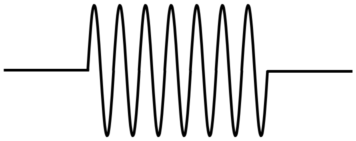
The reasoning seems valid, but the result does contradict the accepted one, which is after all solidly based on experiment.



(a) Part of an infinite sine wave.

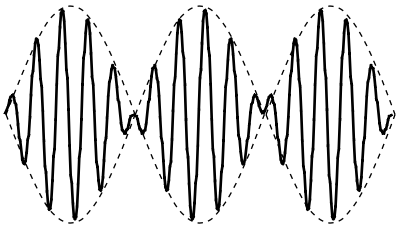
The mistaken assumption is that we can figure everything out in terms of pure sine waves. Mathematically, the only wave that has a perfectly well defined wavelength and frequency is a sine wave, and not just any sine wave but an infinitely long sine wave, (a). The unphysical thing about such a wave is that it has no leading or trailing edge, so it can never be said to enter or leave any particular region of space. Our derivation made use of the velocity, v, and if velocity is to be a meaningful concept, it must tell us how quickly stuff (mass, energy, momentum,...) is transported from one region of space to another. Since an infinitely long sine wave doesn’t remove any stuff from one region and take it to another, the “velocity of its stuff” is not a well defined concept.

Of course the individual wave peaks do travel through space, and one might think that it would make sense to associate their speed with the “speed of stuff,” but as we will see, the two velocities are in general unequal when a wave’s velocity depends on wavelength. Such a wave is called a *dispersive* wave, because a wave pulse consisting of a superposition of waves of different wavelengths will separate (disperse) into its separate wavelengths as the waves move through space at different speeds. Nearly all the waves we have encountered have been nondispersive. For instance, sound waves and light waves (in a vacuum) have speeds independent of wavelength. A water wave is one good example of a dispersive wave. Long-wavelength water waves travel faster, so a ship at sea that encounters a storm typically sees the long-wavelength parts of the wave first. When dealing with dispersive waves, we need symbols and words to distinguish the two speeds. The speed at which wave peaks move is called the phase velocity, v_p, and the speed at which "stuff" moves is called the group velocity, v_g.



(b) A finite-length sine wave.

An infinite sine wave can only tell us about the phase velocity, not the group velocity, which is really what we would be talking about when we refer to the speed of an electron. If an infinite sine wave is the simplest possible wave, what's the next best thing? We might think the runner up in simplicity would be a wave train consisting of a chopped-off segment of a sine wave, (b). However, this kind of wave has kinks in it at the end. A simple wave should be one that we can build by superposing a small number of infinite sine waves, but a kink can never be produced by superposing any number of infinitely long sine waves.



(c) A beat pattern created by superimposing two sine waves with slightly different wavelengths.

Actually the simplest wave that transports stuff from place to place is the pattern shown in figure (c). Called a beat pattern, it is formed by superposing two sine waves whose wavelengths are similar but not quite the same. If you have ever heard the pulsating howling sound of musicians in the process of tuning their instruments to each other, you have heard a beat pattern. The beat pattern gets stronger and weaker as the two sine waves go in and out of phase with each other. The beat pattern has more "stuff" (energy, for example) in the areas where constructive interference occurs, and less in the regions of cancellation. As the whole pattern moves through space, stuff is transported from some regions and into other ones.

If the frequency of the two sine waves differs by 10%, for instance, then ten periods will be occur between times when they are in phase. Another way of saying it is that the sinusoidal "envelope" (the dashed lines in figure (c)) has a frequency equal to the difference in frequency between the two waves. For instance, if the waves had frequencies of 100 Hz and 110 Hz, the frequency of the envelope would be 10 Hz.

To apply similar reasoning to the wavelength, we must define a quantity $z=1/\lambda$ that relates to wavelength in the same way that frequency relates to period. In terms of this new variable, the z of the envelope equals the difference between the z's of the two sine waves.

The group velocity is the speed at which the envelope moves through space. Let Δf and Δz be the differences between the frequencies and z's of the two sine waves, which means that they equal the frequency and z of the envelope. The group velocity is $v_g = f_{\text{envelope}} \cdot \lambda_{\text{envelope}} = \Delta f/\Delta \mathrm{z}$. If Δf and Δz are sufficiently small, we can approximate this expression as a derivative,

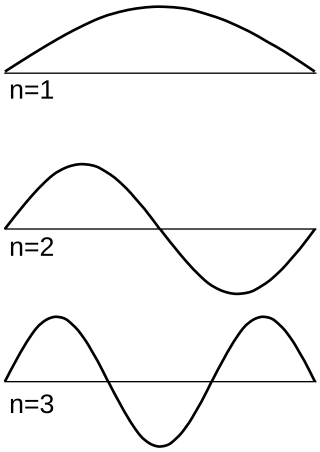
$$v_g = \frac{\mathrm{d}f}{\mathrm{d}z} .$$

This expression is usually taken as the definition of the group velocity for wave patterns that consist of a superposition of sine waves having a narrow range of frequencies and wavelengths. In quantum mechanics, with $f=E/h$ and $z=p/h$, we have $v_g=\mathrm{d}E/\mathrm{d}p$. In the case of a nonrelativistic electron the relationship between energy and momentum is $E=p^2/2m$, so the group velocity is $\mathrm{d}E/\mathrm{d}p=p/m=v$, exactly what it should be. It is only the phase velocity that is different from what we would have expected by a factor of two, but the phase velocity is not what is physically important.

5.3 Bound States

Electrons are at their most interesting when they're in atoms, that is, when they are bound within a small region of space. We can understand a great deal about atoms and molecules based on simple arguments about such bound states, without going into any of the realistic details of atom. The simplest model of a bound state is known as the particle in a box: like a ball on a pool table, the electron feels zero force while in the interior, but when it reaches an edge it encounters a wall that pushes back inward on it with a large force. In particle language, we would describe the electron as bouncing off of the wall, but this incorrectly assumes that the electron has a certain path through space. It is more correct to describe the electron as a wave that undergoes 100% reflection at the boundaries of the box.

Like a generation of physics students before me, I rolled my eyes when initially introduced to the unrealistic idea of putting a particle in a box. It seemed completely impractical, an artificial textbook invention. Today, however, it has become routine to study electrons in rectangular boxes in actual laboratory experiments. The "box" is actually just an empty cavity within a solid piece of silicon, amounting in volume to a few hundred atoms. The methods for creating these electron-in-a-box setups (known as "quantum dots") were a by-product of the development of technologies for fabricating computer chips.



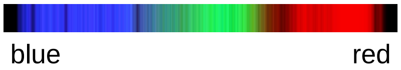
For simplicity let's imagine a one-dimensional electron in a box, i.e. we assume that the electron is only free to move along a line. The resulting standing wave patterns, of which the first three are shown in the figure, are just like some of the patterns we encountered with sound waves in musical instruments. The wave patterns must be zero at the ends of the box, because we are assuming the walls are impenetrable, and there should therefore be zero probability of finding the electron outside the box. Each wave pattern is labeled according to *n*, the number of peaks and valleys it has. In quantum physics, these wave patterns are referred to as "states" of the particle-in-the-box system.

The following seemingly innocuous observations about the particle in the box lead us directly to the solutions to some of the most vexing failures of classical physics:

The particle's energy is quantized (can only have certain values). Each wavelength corresponds to a certain momentum, and a given momentum implies a definite kinetic energy, $E=p^2/2m$. (This is the second type of energy quantization we have encountered. The type we studied previously had to do with restricting the number of particles to a whole number, while assuming some specific wavelength and energy for each particle. This type of quantization refers to the energies that a single particle can have. Both photons and matter particles demonstrate both types of quantization under the appropriate circumstances.)

The particle has a minimum kinetic energy. Long wavelengths correspond to low momenta and low energies. There can be no state with an energy lower than that of the n=1 state, called the ground state.

The smaller the space in which the particle is confined, the higher its kinetic energy must be. Again, this is because long wavelengths give lower energies.



The spectrum of the light from the star Sirius.
Photograph by the author.

Example: spectra of thin gases

A fact that was inexplicable by classical physics was that thin gases absorb and emit light only at certain wavelengths. This was observed both in earthbound laboratories and in the spectra of stars. The figure on the left shows the example of the spectrum of the star Sirius, in which there are “gap teeth” at certain wavelengths. Taking this spectrum as an example, we can give a straightforward explanation using quantum physics.

Energy is released in the dense interior of the star, but the outer layers of the star are thin, so the atoms are far apart and electrons are confined within individual atoms. Although their standing-wave patterns are not as simple as those of the particle in the box, their energies are quantized.

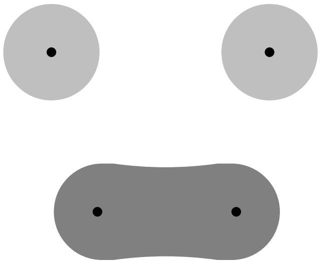
When a photon is on its way out through the outer layers, it can be absorbed by an electron in an atom, but only if the amount of energy it carries happens to be the right amount to kick the electron from one of the allowed energy levels to one of the higher levels. The photon energies that are missing from the spectrum are the ones that equal the difference in energy between two electron energy levels. (The most prominent of the absorption lines in Sirius’s spectrum are absorption lines of the hydrogen atom.)

Example: the stability of atoms

In many Star Trek episodes the Enterprise, in orbit around a planet, suddenly lost engine power and began spiraling down toward the planet’s surface. This was utter nonsense, of course, due to conservation of energy: the ship had no way of getting rid of energy, so it did not need the engines to replenish it.

Consider, however, the electron in an atom as it orbits the nucleus. The electron *does* have a way to release energy: it has an acceleration due to its continuously changing direction of motion, and according to classical physics, any accelerating charged particle emits electromagnetic waves. According to classical physics, atoms should collapse!

The solution lies in the observation that a bound state has a minimum energy. An electron in one of the higher-energy atomic states can and does emit photons and hop down step by step in energy. But once it is in the ground state, it cannot emit a photon because there is no lower-energy state for it to go to.



Two hydrogen atoms bond to form an H_2 molecule. In the molecule, the two electrons’ wave patterns overlap, and are about twice as wide.

Example: chemical bonds

I began this chapter with a classical argument that chemical bonds, as in an H_2 molecule, should not exist. Quantum physics explains why this type of bonding does in fact occur. When the atoms are next to each other, the electrons are shared between them. The “box” is about twice as wide, and a larger box allows a smaller energy. Energy is required in order to separate the atoms. (A qualitatively different type of bonding is discussed in section 6.6.)

Discussion Question

A. Neutrons attract each other via the strong nuclear force, so according to classical physics it should be possible to form nuclei out of clusters of two or more neutrons, with no protons at all. Experimental searches, however, have failed to turn up evidence of a stable two-neutron system (dineutron) or larger stable clusters. Explain based on quantum physics why a dineutron might spontaneously fly apart.

B. The following table shows the energy gap between the ground state and the first excited state for four nuclei in units of picojoules. (The nuclei have been chosen to be ones that have similar structures, e.g. they are all spherical nuclei.)

nucleus	**energy gap**
^{4}He	3.234 pJ
^{16}O	0.968
^{40}Ca	0.536
^{208}Pb	0.418

Explain the trend in the data.

5.4 The Uncertainty Principle and Measurement

The uncertainty principle

Eliminating randomness through measurement?

A common reaction to quantum physics, among both early-twentieth-century physicists and modern students, is that we should be able to get rid of randomness through accurate measurement. If I say, for example, that it is meaningless to discuss the path of a photon or an electron, one might suggest that we simply measure the particle's position and velocity many times in a row. This series of snapshots would amount to a description of its path.

A practical objection to this plan is that the process of measurement will have an effect on the thing we are trying to measure. This may not be of much concern, for example, when a traffic cop measure's your car's motion with a radar gun, because the energy and momentum of the radar pulses are insufficient to change the car's motion significantly. But on the subatomic scale it is a very real problem. Making a videotape through a microscope of an electron orbiting a nucleus is not just difficult, it is theoretically impossible. The video camera makes pictures of things using light that has bounced off them and come into the camera. If even a single photon of visible light was to bounce off of the electron we were trying to study, the electron's recoil would be enough to change its behavior completely.

The Heisenberg uncertainty principle

This insight, that measurement changes the thing being measured, is the kind of idea that clove-cigarette-smoking intellectuals outside of the physical sciences like to claim they knew all along. If only, they say, the physicists had made more of a habit of reading literary journals, they could have saved a lot of work. The anthropologist Margaret Mead has recently been accused of inadvertently encouraging her teenaged Samoan informants to exaggerate the freedom of youthful sexual experimentation in their society. If this is considered a damning critique of her work, it is because she could have done better: other anthropologists claim to have been able to eliminate the observer-as-participant problem and collect untainted data.

The German physicist Werner Heisenberg, however, showed that in quantum physics, *any* measuring technique runs into a brick wall when we try to improve its accuracy beyond a certain point. Heisenberg showed that the limitation is a matter of *what there is to be known*, even in principle, about the system itself, not of the ability or inability of a specific measuring device to ferret out information that is knowable but not previously hidden.

Suppose, for example, that we have constructed an electron in a box (quantum dot) setup in our laboratory, and we are able adjust the length L of the box as desired. All the standing wave patterns pretty much fill the box, so our knowledge of the electron's position is of limited accuracy. If we write Δx for the range of uncertainty in our knowledge of its position, then Δx is roughly the same as the length of the box:

$$\Delta x \approx L \qquad (1)$$

If we wish to know its position more accurately, we can certainly squeeze it into a smaller space by reducing L, but this has an unintended side-effect. A standing wave is really a superposition of two traveling waves going in opposite directions. The equation $p=h/\lambda$ really only gives the magnitude of the momentum vector, not its direction, so we should really interpret the wave as a 50/50 mixture of a right-going wave with momentum $p=h/\lambda$ and a left-going one with momentum $p=-h/\lambda$. The uncertainty in our knowledge of the electron's momentum is $\Delta p=2h/\lambda$, covering the range between these two values. Even if we make sure the electron is in the ground state, whose wavelength $\lambda=2L$ is the longest possible, we have an uncertainty in momentum of $\Delta p=h/L$. In general, we find

$$\Delta p \geq h/L \quad , \qquad (2)$$

with equality for the ground state and inequality for the higher-energy states. Thus if we reduce L to improve our knowledge of the electron's position, we do so at the cost of knowing less about its momentum. This trade-off is neatly summarized by multiplying equations (1) and (2) to give

$$\Delta p \, \Delta x \gtrsim h$$

Although we have derived this in the special case of a particle in a box, it is an example of a principle of more general validity:

> **The Heisenberg uncertainty principle:**
> It is not possible, even in principle, to know the momentum and the position of a particle simultaneously and with perfect accuracy. The uncertainties in these two quantities are always such that $\Delta p \, \Delta x \gtrsim h$.

(This approximation can be made into a strict inequality, $\Delta p \, \Delta x \geq h/4\pi$, but only with more careful definitions, which we will not bother with.)

Note that although I encouraged you to think of this derivation in terms of a specific real-world system, the quantum dot, no reference was ever made to any specific laboratory equipment or procedures. The argument is simply that we cannot *know* the particle's position very accurately unless it *has* a very well defined position, it cannot have a very well defined position unless its wave-pattern covers only a very small amount of space, and its wave-pattern cannot be thus compressed without giving it a short wavelength and a correspondingly uncertain momentum. The uncertainty principle is therefore a restriction on how much there is to know about a particle, not just on what we can know about it with a certain technique.

Example: an estimate for electrons in atoms

Question: A typical energy for an electron in an atom is on the order of 1 volt · e, which corresponds to a speed of about 1% of the speed of light. If a typical atom has a size on the order of 0.1 nm, how close are the electrons to the limit imposed by the uncertainty principle?

Solution: If we assume the electron moves in all directions with equal probability, the uncertainty in its momentum is roughly twice its typical momentum. This only an order-of-magnitude estimate, so we take Δp to be the same as a typical momentum:

$$\begin{aligned} \Delta p\,\Delta x &= p_{\text{typical}}\,\Delta x \\ &= (m_{\text{electron}})\,(0.01c)\,(0.1\text{x}10^{-9}\text{ m}) \\ &= 3\text{x}10^{-34}\text{ J}\cdot\text{s} \end{aligned}$$

This is on the same order of magnitude as Planck’s constant, so evidently the electron is “right up against the wall.” (The fact that it is somewhat less than h is of no concern since this was only an estimate, and we have not stated the uncertainty principle in its most exact form.)

Self-Check

If we were to apply the uncertainty principle to human-scale objects, what would be the significance of the small numerical value of Planck’s constant?

Measurement and Schrödinger’s cat

In the previous chapter I briefly mentioned an issue concerning measurement that we are now ready to address carefully. If you hang around a laboratory where quantum-physics experiments are being done and secretly record the physicists’ conversations, you’ll hear them say many things that assume the probability interpretation of quantum mechanics. Usually they will speak as though the randomness of quantum mechanics enters the picture when something is measured. In the digital camera experiments of the previous chapter, for example, they would casually describe the detection of a photon at one of the pixels as if the moment of detection was when the photon was forced to “make up its mind.” Although this mental cartoon usually works fairly well as a description of things they experience in the lab, it cannot ultimately be correct, because it attributes a special role to measurement, which is really just a physical process like all other physical processes.

If we are to find an interpretation that avoids giving any special role to measurement processes, then we must think of the entire laboratory, including the measuring devices and the physicists themselves, as one big quantum-mechanical system made out of protons, neutrons, electrons, and photons. In other words, we should take quantum physics seriously as a description not just of microscopic objects like atoms but of human-scale (“macroscopic”) things like the apparatus, the furniture, and the people.

The most celebrated example is called the Schrödinger's cat experiment. Luckily for the cat, there probably was no actual experiment — it was

Under the ordinary circumstances of life, the accuracy with which we can measure position and momentum of an object doesn’t result in a value of $\Delta p \Delta x$ that is anywhere near the tiny order of magnitude of Planck’s constant. We run up against the ordinary limitations on the accuracy of our measuring techniques long before the uncertainty principle becomes an issue.

simply a "thought experiment" that the physicist the German theorist Schrödinger discussed with his colleagues. Schrödinger wrote:

> One can even construct quite burlesque cases. A cat is shut up in a steel container, together with the following diabolical apparatus (which one must keep out of the direct clutches of the cat): In a Geiger tube [radiation detector] there is a tiny mass of radioactive substance, so little that in the course of an hour perhaps one atom of it disintegrates, but also with equal probability not even one; if it does happen, the counter [detector] responds and ... activates a hammer that shatters a little flask of prussic acid [filling the chamber with poison gas]. If one has left this entire system to itself for an hour, then one will say to himself that the cat is still living, if in that time no atom has disintegrated. The first atomic disintegration would have poisoned it.

Now comes the strange part. Quantum mechanics describes the particles the cat is made of as having wave properties, including the property of superposition. Schrödinger describes the wavefunction of the box's contents at the end of the hour:

> The wavefunction of the entire system would express this situation by having the living and the dead cat mixed ... in equal parts [50/50 proportions]. The uncertainty originally restricted to the atomic domain has been transformed into a macroscopic uncertainty...

At first Schrödinger's description seems like nonsense. When you opened the box, would you see two ghostlike cats, as in a doubly exposed photograph, one dead and one alive? Obviously not. You would have a single, fully material cat, which would either be dead or very, very upset. But Schrödinger has an equally strange and logical answer for that objection. In the same way that the quantum randomness of the radioactive atom spread to the cat and made its wavefunction a random mixture of life and death, the randomness spreads wider once you open the box, and your own wavefunction becomes a mixture of a person who has just killed a cat and a person who hasn't.

Discussion Questions

A. Compare Δp and Δx for the two loest energy levels of the one-dimensional particle in a box, and discuss how this relates to the uncertainty principle.

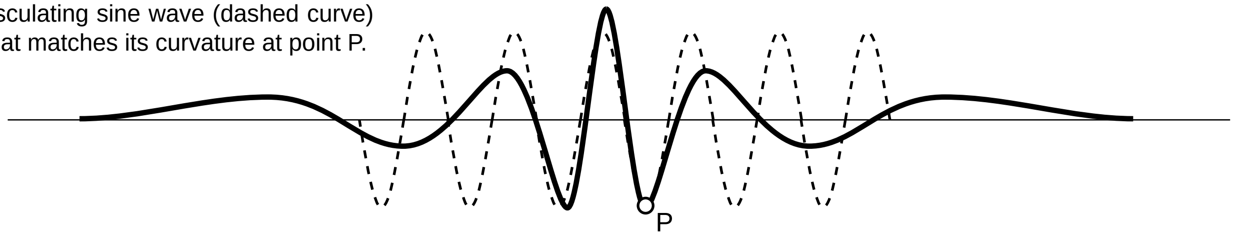
B. On a graph of Δp versus Δx, sketch the regions that are allowed and forbidden by the Heisenberg uncertainty principle. Interpret the graph: Where does an atom lie on it? An elephant? Can either p or x be measured with perfect accuracy if we don't care about the other?

5.5 Electrons in Electric Fields

So far the only electron wave patterns we've considered have been simple sine waves, but whenever an electron finds itself in an electric field, it must have a more complicated wave pattern. Let's consider the example of an electron being accelerated by the electron gun at the back of a TV tube. Newton's laws are not useful, because they implicitly assume that the path taken by the particle is a meaningful concept. Conservation of energy is still valid in quantum physics, however. In terms of energy, the electron is moving from a region of low voltage into a region of higher voltage. Since its charge is negative, it loses PE by moving to a higher voltage, so its KE increases. As its potential energy goes down, its kinetic energy goes up by an equal amount, keeping the total energy constant. Increasing kinetic energy implies a growing momentum, and therefore a shortening wavelength, (a).

The wavefunction as a whole does not have a single well-defined wavelength, but the wave changes so gradually that if you only look at a small part of it you can still pick out a wavelength and relate it to the momentum and energy. (The picture actually exaggerates by many orders of magnitude the rate at which the wavelength changes.)

But what if the electric field was stronger? The electric field in a TV is only ~10^5 N/C, but the electric field within an atom is more like 10^{12} N/C. In figure (b), the wavelength changes so rapidly that there is nothing that looks like a sine wave at all. We could get a rough idea of the wavelength in a given region by measuring the distance between two peaks, but that would only be a rough approximation. Suppose we want to know the wavelength at point P. The trick is to construct a sine wave, like the one shown with the dashed line, which matches the curvature of the actual wavefunction as closely as possible near P. The sine wave that matches as well as possible is called the "osculating" curve, from a Latin word meaning "to kiss." The wavelength of the osculating curve is the wavelength that will relate correctly to conservation of energy.

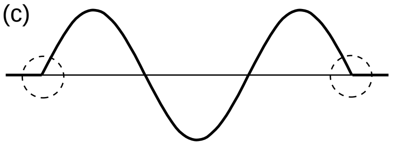


(b) A typical wavefunction of an electron in an atom (heavy curve) and the osculating sine wave (dashed curve) that matches its curvature at point P.

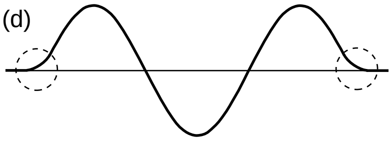
Tunneling

We implicitly assumed that the particle-in-a-box wavefunction would cut off abruptly at the sides of the box, (c), but that would be unphysical. A kink has infinite curvature, and curvature is related to energy, so it can't be infinite. A physically realistic wavefunction must always "tail off" gradually, (d). In classical physics, a particle can never enter a region in which its potential energy would be greater than the amount of energy it has available. But in quantum physics the wavefunction will always have a tail that reaches into the classically forbidden region. If it was not for this effect, called tunneling, the fusion reactions that power the sun would not occur due to the high potential energy nuclei need in order to get close together! Tunneling is discussed in more detail in the following section.

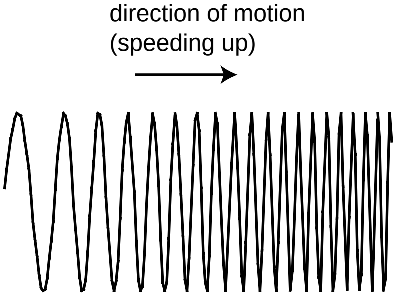
(c)



(d)



5.6*∫ The Schrödinger Equation



high PE	low PE
low KE	high KE
low momentum	high momentum
long wavelength	short wavelength
weak curvature	strong curvature

(a) An electron in a gentle electric field gradually shortens its wavelength as it gains energy.

In the previous section we were able to apply conservation of energy to an electron's wavefunction, but only by using the clumsy graphical technique of osculating sine waves as a measure of the wave's curvature. You have learned a more convenient measure of curvature in calculus: the second derivative. To relate the two approaches, we take the second derivative of a sine wave:

$$\begin{aligned}
&\frac{\mathrm{d}^2}{\mathrm{d}x^2}\sin\left(2\pi x/\lambda\right)\\
&\quad= \frac{\mathrm{d}}{\mathrm{d}x}\left[\frac{2\pi}{\lambda}\cos\left(2\pi x/\lambda\right)\right]\\
&\quad= -\left(\frac{2\pi}{\lambda}\right)^2\sin\left(2\pi x/\lambda\right)
\end{aligned}$$

Taking the second derivative gives us back the same function, but with a minus sign and a constant out in front that is related to the wavelength. We can thus relate the second derivative to the osculating wavelength:

$$\frac{\mathrm{d}^2\Psi}{\mathrm{d}x^2} = -\left(\frac{2\pi}{\lambda}\right)^2\Psi \qquad (1)$$

This could be solved for λ in terms of Ψ, but it will turn out below to be more convenient to leave it in this form.

Applying this to conservation of energy, we have

$$\begin{aligned}
E &= KE + PE\\
&= \frac{p^2}{2m} + PE\\
&= \frac{\left(h/\lambda\right)^2}{2m} + PE \qquad (2)
\end{aligned}$$

Note that both equation (1) and equation (2) have λ^2 in the denominator. We can simplify our algebra by multiplying both sides of equation (2) by Ψ to make it look more like equation (1):

$$\begin{aligned}
E\cdot\Psi &= \frac{\left(h/\lambda\right)^2}{2m}\Psi + PE\cdot\Psi\\
&= \frac{1}{2m}\left(\frac{h}{2\pi}\right)^2\left(\frac{2\pi}{\lambda}\right)^2\Psi + PE\cdot\Psi\\
&= -\frac{1}{2m}\left(\frac{h}{2\pi}\right)^2\frac{\mathrm{d}^2\Psi}{\mathrm{d}x^2} + PE\cdot\Psi
\end{aligned}$$

Some simplification is achieved by using the symbol $\hbar$ (h with a slash through it, read "h-bar") as an abbreviation for $h/2\pi$. We then have the important equation known as the Schrödinger equation:

$$E\cdot\Psi = -\frac{\hbar^2}{2m}\frac{\mathrm{d}^2\Psi}{\mathrm{d}x^2} + PE\cdot\Psi$$

(Actually this is a simplified version of the Schrödinger equation, applying only to standing waves in one dimension.) Physically it is a statement of conservation of energy. The total energy E must be constant, so the equation tells us that a change in potential energy must be accompanied by a change in the curvature of the wavefunction. This change in curvature relates to a change in wavelength, which corresponds to a change in momentum and kinetic energy.

Self-Check

Considering the assumptions that were made in deriving the Schrödinger equation, would it be correct to apply it to a photon? To an electron moving at relativistic speeds?

Usually we know right off the bat how PE depends on x, so the basic mathematical problem of quantum physics is to find a function $\Psi(x)$ that satisfies the Schrödinger equation for a given potential-energy function $PE(x)$. An equation, such as the Schrödinger equation, that specifies a relationship between a function and its derivatives is known as a differential equation.

The study of differential equations in general is beyond the mathematical level of this book, but we can gain some important insights by considering the easiest version of the Schrödinger equation, in which the potential energy is constant. We can then rearrange the Schrödinger equation as follows:

$$\frac{\mathrm{d}^2\Psi}{\mathrm{d}x^2} = \frac{2m\left(PE-E\right)}{\hbar^2}\Psi \quad ,$$

which boils down to

$$\frac{\mathrm{d}^2\Psi}{\mathrm{d}x^2} = a\Psi \quad ,$$

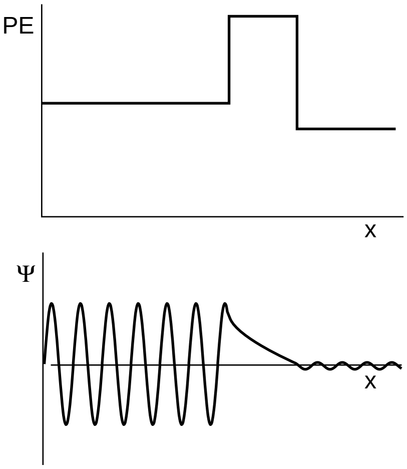
where, according to our assumptions, a is independent of x. We need to find a function whose second derivative is the same as the original function except for a multiplicative constant. The only functions with this property are sine waves and exponentials:

$$\frac{\mathrm{d}^2}{\mathrm{d}x^2}\left[q\sin\left(rx+s\right)\right] = -qr^2\sin\left(rx+s\right)$$

$$\frac{\mathrm{d}^2}{\mathrm{d}x^2}\left[qe^{rx+s}\right] = qr^2e^{rx+s}$$

The sine wave gives negative values of a, $a=-r^2$, and the exponential gives

No. The equation $KE=p^2/2m$ is nonrelativistic, so it can't be applied to an electron moving at relativistic speeds. Photons always move at relativistic speeds, so it can't be applied to them either.



positive ones, $a=r^2$. The former applies to the classically allowed region with $PE<E$, the latter to the classical forbidden region with $PE>E$.

This leads us to a quantitative calculation of the tunneling effect discussed briefly in the previous section. The wavefunction evidently tails off exponentially in the classically forbidden region. Suppose, as shown in the figure, a wave-particle traveling to the right encounters a barrier that it is classically forbidden to enter. Although the form of the Schrödinger equation we're using technically does not apply to traveling waves (because it makes no reference to time), it turns out that we can still use it to make a reasonable calculation of the probability that the particle will make it through the barrier. If we let the barrier's width be w, then the ratio of the wavefunction on the left side of the barrier to the wavefunction on the right is

$$\frac{qe^{rx+s}}{qe^{r(x+w)+s}} = e^{-rw} \quad .$$

Probabilities are proportional to the squares of wavefunctions, so the probability of making it through the barrier is

$$\begin{aligned} P &= e^{-2rw} \\ &= \exp\left(-\frac{2w}{\hbar}\sqrt{2m\left(PE-E\right)}\right) \end{aligned}$$

Self-Check

If we were to apply this equation to find the probability that a person can walk through a wall, what would the small value of Planck's constant imply?

Use of complex numbers

In a classically forbidden region, a particle's total energy, $PE+KE$, is less than its PE, so its KE must be negative. If we want to keep believing in the equation $KE=p^2/2m$, then apparently the momentum of the particle is the square root of a negative number. This is a symptom of the fact that the Schrödinger equation fails to describe all of nature unless the wavefunction and various other quantities are allowed to be complex numbers. In particular it is not possible to describe traveling waves correctly without using complex wavefunctions.

This may seem like nonsense, since real numbers are the only ones that are, well, real! Quantum mechanics can always be related to the real world, however, because its structure is such that the results of measurements always come out to be real numbers. For example, we may describe an electron as having non-real momentum in classically forbidden regions, but its average momentum will always come out to be real (the imaginary parts average out to zero), and it can never transfer a non-real quantity of momentum to another particle.

A complete investigation of these issues is beyond the scope of this book, and this is why we have normally limited ourselves to standing waves, which can be described with real-valued wavefunctions.

Dividing by Planck's constant, a small number, gives a large negative result inside the exponential, so the probability will be very small.

Summary

Selected Vocabulary

wavefunction .... the numerical measure of an electron wave, or in general of the wave corresponding to any quantum mechanical particle

Notation

$\hbar$ .... Planck's constant divided by 2π (used only in optional section 5.6)

Ψ .... the wavefunction of an electron

Summary

Light is both a particle and a wave, and matter is also both a particle and a wave. The equations that connect the particle and wave properties are the same in all cases:

$$E = hf$$

$$p = h/\lambda$$

Unlike the electric and magnetic fields that make up a photon-wave, the electron wavefunction is not directly measurable. Only the square of the wavefunction, which relates to probability, has direct physical significance.

A particle that is bound within a certain region of space is a standing wave in terms of quantum physics. The two equations above can then be applied to the standing wave to yield some important general observations about bound particles:

(1) The particle's energy is quantized (can only have certain values).

(2) The particle has a minimum energy.

(3) The smaller the space in which the particle is confined, the higher its kinetic energy must be.

These immediately resolve the difficulties that classical physics had encountered in explaining observations such as the discrete spectra of atoms, the fact that atoms don't collapse by radiating away their energy, and the formation of chemical bonds.

A standing wave confined to a small space must have a short wavelength, which corresponds to a large momentum in quantum physics. Since a standing wave consists of a superposition of two traveling waves moving in opposite directions, this large momentum should actually be interpreted as an equal mixture of two possible momenta: a large momentum to the left, or a large momentum to the right. Thus it is not possible for a quantum wave-particle to be confined to a small space without making its momentum very uncertain. In general, the Heisenberg uncertainty principle states that it is not possible to know the position and momentum of a particle simultaneously with perfect accuracy. The uncertainties in these two quantities must satisfy the approximate inequality

$$\Delta p \Delta x \gtrsim h$$

When an electron is subjected to electric forces, its wavelength cannot be constant. The "wavelength" to be used in the equation $p = h/\lambda$ should be thought of as the wavelength of the sine wave that most closely approximates the curvature of the wavefunction at a specific point.

Infinite curvature is not physically possible, so realistic wavefunctions cannot have kinks in them, and cannot just cut off abruptly at the edge of a region where the particle's energy would be insufficient to penetrate according to classical physics. Instead, the wavefunction "tails off" in the classically forbidden region, and as a consequence it is possible for particles to "tunnel" through regions where according to classical physics they should not be able to penetrate. If this quantum tunneling effect did not exist, there would be no fusion reactions to power our sun, because the energies of the nuclei would be insufficient to overcome the electrical repulsion between them.

Homework Problems

1✓. In a television, suppose the electrons are accelerated from rest through a potential difference of 10^4 V. What is their final wavelength?

2. Use the Heisenberg uncertainty principle to estimate the minimum velocity of a proton or neutron in a ^{208}Pb nucleus, which has a diameter of about 13 fm (1 fm=10^{-15} m). Assume that the speed is nonrelativistic, and then check at the end whether this assumption was warranted.

3. Find the energy of a particle in a one-dimensional box of length L, expressing your result in terms of L, the particle's mass m, the number of peaks and valleys n in the wavefunction, and fundamental constants.

4. A free electron that contributes to the current in an ohmic material typically has a speed of 10^5 m/s (much greater than the drift velocity).

(a✓) Estimate its de Broglie wavelength, in nm.

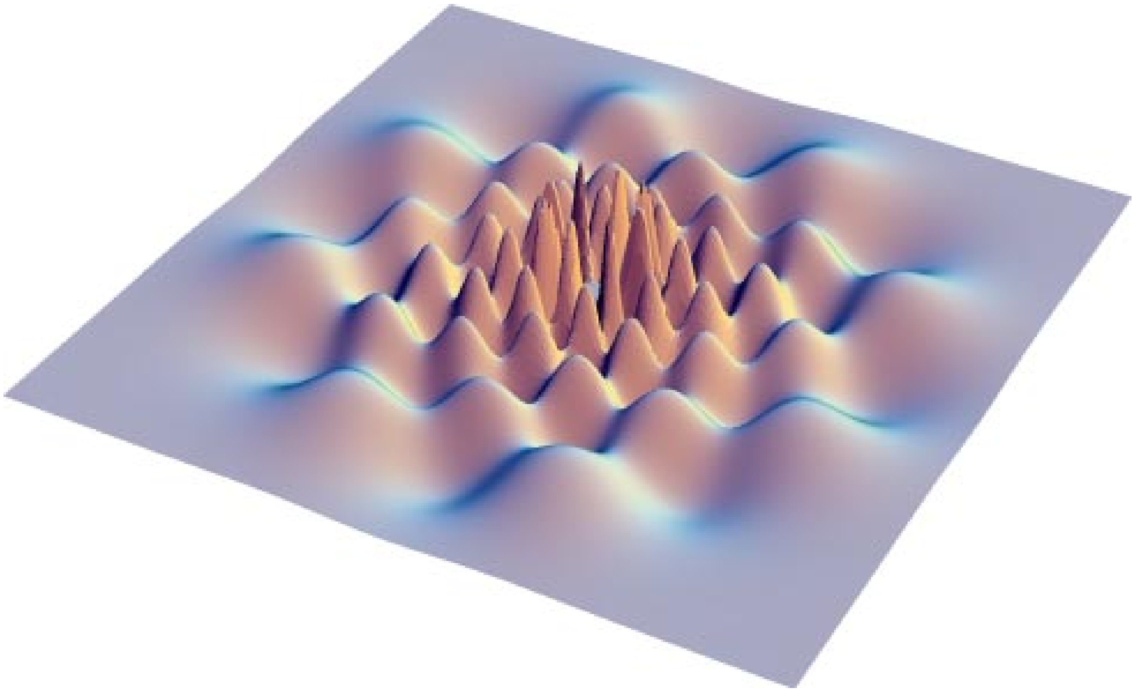
(b✓) If a computer memory chip contains 10^8 electric circuits in a 1 cm^2 area, estimate the linear size, in nm, of one such circuit.

(c) Based on your answers from parts a and b, does an electrical engineer designing such a chip need to worry about wave effects such as diffraction?

(d) Estimate the maximum number of electric circuits that can fit on a 1 cm^2 computer chip before quantum-mechanical effects become important.

5★∫. In section 5.6 we derived an expression for the probability that a particle would tunnel through a rectangular potential barrier. Generalize this to a barrier of any shape. [Hints: First try generalizing to two rectangular barriers in a row, and then use a series of rectangular barriers to approximate the actual curve of an arbitrary potential. Note that the width and height of the barrier in the original equation occur in such a way that all that matters is the area under the *PE*-versus-*x* curve. Show that this is still true for a series of rectangular barriers, and generalize using an integral.] If you had done this calculation in the 1930's you could have become a famous physicist.

S A solution is given in the back of the book.	★ A difficult problem.
✓ A computerized answer check is available.	∫ A problem that requires calculus.



6 The Atom

You can learn a lot by taking a car engine apart, but you will have learned a lot more if you can put it all back together again and make it run. Half the job of reductionism is to break nature down into its smallest parts and understand the rules those parts obey. The second half is to show how those parts go together, and that is our goal in this chapter. We have seen how certain features of all atoms can be explained on a generic basis in terms of the properties of bound states, but this kind of argument clearly cannot tell us any details of the behavior of an atom or explain why one atom acts differently from another.

The biggest embarrassment for reductionists is that the job of putting things back together job is usually much harder than the taking them apart. Seventy years after the fundamentals of atomic physics were solved, it is only beginning to be possible to calculate accurately the properties of atoms that have many electrons. Systems consisting of many atoms are even harder. Supercomputer manufacturers point to the folding of large protein molecules as a process whose calculation is just barely feasible with their fastest machines. The goal of this chapter is to give a gentle and visually oriented guide to some of the simpler results about atoms.

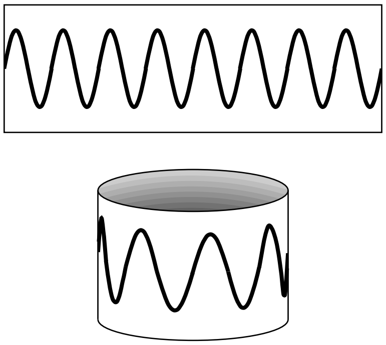
6.1 Classifying States

We'll focus our attention first on the simplest atom, hydrogen, with one proton and one electron. We know in advance a little of what we should expect for the structure of this atom. Since the electron is bound to the proton by electrical forces, it should display a set of discrete energy states, each corresponding to a certain standing wave pattern. We need to understand what states there are and what their properties are.

What properties should we use to classify the states? The most sensible approach is to used conserved quantities. Energy is one conserved quantity, and we already know to expect each state to have a specific energy. It turns out, however, that energy alone is not sufficient. Different standing wave patterns of the atom can have the same energy.

Momentum is also a conserved quantity, but it is not particularly appropriate for classifying the states of the electron in a hydrogen atom. The reason is that the force between the electron and the proton results in the continual exchange of momentum between them. (Why wasn't this a problem for energy as well? Kinetic energy and momentum are related by $KE=p^2/2m$, so the much more massive proton never has very much kinetic energy. We are making an approximation by assuming all the kinetic energy is in the electron, but it is quite a good approximation.)

Angular momentum does help with classification. There is no transfer of angular momentum between the proton and the electron, since the force between them is a center-to-center force, producing no torque.



Eight wavelengths fit around this circle (ℓ =8).

Like energy, angular momentum is quantized in quantum physics. As an example, consider a quantum wave-particle confined to a circle, like a wave in a circular moat surrounding a castle. A sine wave in such a "quantum moat" cannot have any old wavelength, because an integer number of wavelengths must fit around the circumference, C, of the moat. The larger this integer is, the shorter the wavelength, and a shorter wavelength relates to greater momentum and angular momentum. Since this integer is related to angular momentum, we use the symbol ℓ for it:

$$\lambda = C/\ell$$

Its angular momentum is

$$L = rp \quad .$$

Here, $r=C/2\pi$, and $p = h/\lambda = h\ell/C$, so

$$\begin{aligned} L &= \frac{C}{2\pi} \cdot \frac{h\ell}{C} \\ &= \frac{h}{2\pi}\ell \end{aligned}$$

In the example of the quantum moat, angular momentum is quantized in units of $h/2\pi$. This makes $h/2\pi$ a pretty important number, so we define the abbreviation $\hbar = h/2\pi$. This symbol is read "h-bar."

In fact, this is a completely general fact in quantum physics, not just a fact about the quantum moat:

> **Quantization of angular momentum**
> The angular momentum of a particle due to its motion through space is quantized in units of $\hbar$.

Self-Check

What is the angular momentum of the wavefunction shown at the beginning of the chapter?

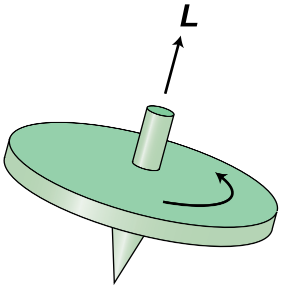
6.2 Angular Momentum in Three Dimensions

A more complete discussion of angular momentum in three dimensions is given in supplement 2-7, which can be downloaded from www.lightandmatter.com.

Up until now we've only worked with angular momentum in the context of rotation in a plane, for which we could simply use positive and negative signs to indicate clockwise and counterclockwise directions of rotation. A hydrogen atom, however, is unavoidably three-dimensional. Let's first consider the generalization of angular momentum to three dimensions in the classical case, and then consider how it carries over into quantum physics.

Three-dimensional angular momentum in classical physics

If we are to completely specify the angular momentum of a classical object like a top, (a), in three dimensions, it's not enough to say whether the rotation is clockwise or counterclockwise. We must also give the orientation of the plane of rotation or, equivalently, the direction of the top's axis. The convention is to specify the direction of the axis. There are two possible directions along the axis, and as a matter of convention we use the direction such that if we sight along it, the rotation appears clockwise.



(a) The angular momentum vector of a spinning top.

Angular momentum can, in fact, be defined as a vector pointing along this direction. This might seem like a strange definition, since nothing actually moves in that direction, but it wouldn't make sense to define the angular momentum vector as being in the direction of motion, because every part of the top has a different direction of motion. Ultimately it's not just a matter of picking a definition that is convenient and unambiguous: the definition we're using is the only one that makes the total angular momentum of a system a conserved quantity if we let "total" mean the vector sum.

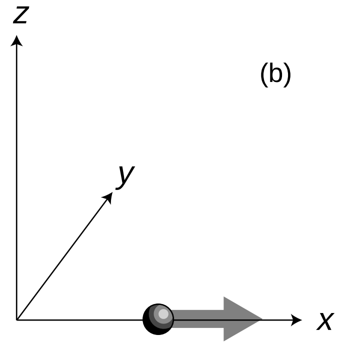
As with rotation in one dimension, we cannot define what we mean by angular momentum in a particular situation unless we pick a point as an axis. This is really a different use of the word "axis" than the one in the previous paragraphs. Here we simply mean a point from which we measure the distance *r*. In the hydrogen atom, the nearly immobile proton provides a natural choice of axis.

If you trace a circle going around the center, you run into a series of eight complete wavelengths. Its angular momentum is $8\hbar$.

Three-dimensional angular momentum in quantum physics

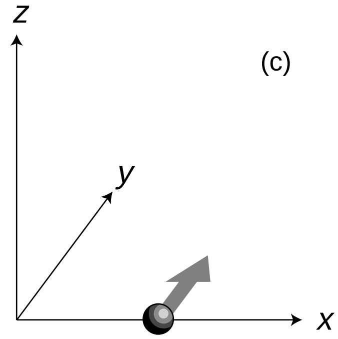
Once we start to think more carefully about the role of angular momentum in quantum physics, it may seem that there is a basic problem: the angular momentum of the electron in a hydrogen atom depends on both its distance from the proton and its momentum, so in order to know its angular momentum precisely it would seem we would need to know both its position and its momentum simultaneously with good accuracy. This, however, might seem to be forbidden by the Heisenberg uncertainty principle.

Actually the uncertainty principle does place limits on what can be known about a particle's angular momentum vector, but it does not prevent us from knowing its magnitude as an exact integer multiple of $\hbar$. The reason is that in three dimensions, there are really three separate uncertainty principles:



$$\Delta p_x \, \Delta x \geq h$$
$$\Delta p_y \, \Delta y \geq h$$
$$\Delta p_z \, \Delta z \geq h$$

Now consider a particle, (b), that is moving along the *x* axis at position *x* and with momentum p_x. We may not be able to know both *x* and p_x with unlimited accurately, but we can still know the particle's angular momentum about the origin exactly: it is zero, because the particle is moving directly away from the origin.



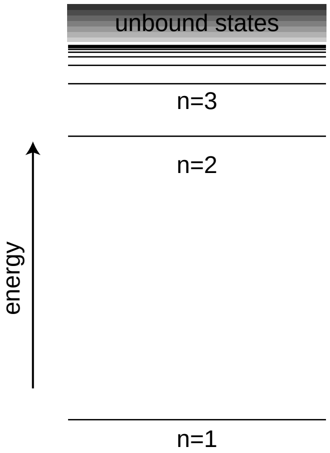
Suppose, on the other hand, a particle finds itself, (c), at a position *x* along the *x* axis, and it is moving parallel to the *y* axis with momentum p_y. It has angular momentum xp_y about the *z* axis, and again we can know its angular momentum with unlimited accuracy, because the uncertainty principle on relates *x* to p_x and *y* to p_y. It does not relate *x* to p_y.

As shown by these examples, the uncertainty principle does not restrict the accuracy of our knowledge of angular momenta as severely as might be imagined. However, it does prevent us from knowing all three components of an angular momentum vector simultaneously. The most general statement about this is the following theorem, which we present without proof:

> **The angular momentum vector in quantum physics**
> The most the can be known about an angular momentum vector is its magnitude and one of its three vector components. Both are quantized in units of $\hbar$.

6.3 The Hydrogen Atom

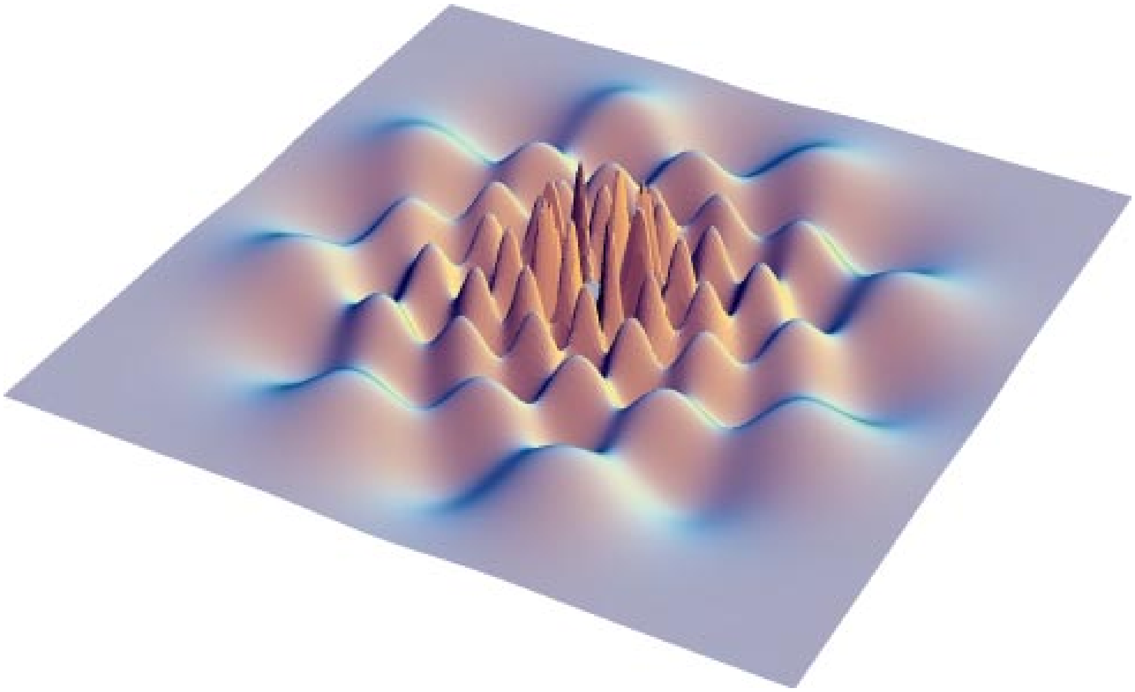
Deriving the wavefunctions of the states of the hydrogen atom from first principles would be mathematically too complex for this book, but it's not hard to understand the logic behind such a wavefunction in visual terms. Consider the wavefunction from the beginning of the chapter, which is reproduced below. Although the graph looks three-dimensional, it is really only a representation of the part of the wavefunction lying within a two-dimensional plane. The third (up-down) dimension of the plot represents the value of the wavefunction at a given point, not the third dimension of space. The plane chosen for the graph is the one perpendicular to the angular momentum vector.



The energy of a state in the hydrogen atom depends only on its n quantum number.

Each ring of peaks and valleys has eight wavelengths going around in a circle, so this state has $L=8\hbar$, i.e. we label it ℓ =8. The wavelength is shorter near the center, and this makes sense because when the electron is close to the nucleus it has a lower *PE*, a higher *KE*, and a higher momentum.

Between each ring of peaks in this wavefunction is a nodal circle, i.e. a circle on which the wavefunction is zero. The full three-dimensional wavefunction has nodal spheres: a series of nested spherical surfaces on which it is zero. The number of radii at which nodes occur, including $r=\infty$, is called n, and n turns out to be closely related to energy. The ground state has $n=1$ (a single node only at $r=\infty$), and higher-energy states have higher n values. There is a simple equation relating n to energy, which we will discuss in section 6.4.



The numbers n and ℓ, which identify the state, are called its quantum numbers. A state of a given n and ℓ can be oriented in a variety of directions in space. We might try to indicate the orientation using the three quantum numbers $\ell_x = L_x/\hbar$, $\ell_y = L_y/\hbar$, and $\ell_z = L_z/\hbar$. But we have already seen that it is impossible to know all three of these simultaneously. To give the most complete possible description of a state, we choose an arbitrary axis, say the z axis, and label the state according to n, ℓ, and ℓ_z.

Angular momentum requires motion, and motion implies kinetic energy. Thus it is not possible to have a given amount of angular momentum without having a certain amount of kinetic energy as well. Since energy relates to the n quantum number, this means that for a given n value there will be a maximum possible ℓ. It turns out that this maximum value of ℓ equals n–1.

In general, we can list the possible combinations of quantum numbers as follows:

n can equal 1, 2, 3, ...
ℓ can range from 0 to n–1, in steps of 1
ℓ_z can range from $-\ell$ to ℓ, in steps of 1

Applying these, rules, we have the following list of states:

n=1, ℓ=0, ℓ_z=0	one state
n=2, ℓ=0, ℓ_z=0	one state
n=2, ℓ=1, ℓ_z=–1, 0, or 1	three states
etc.	

Self-Check

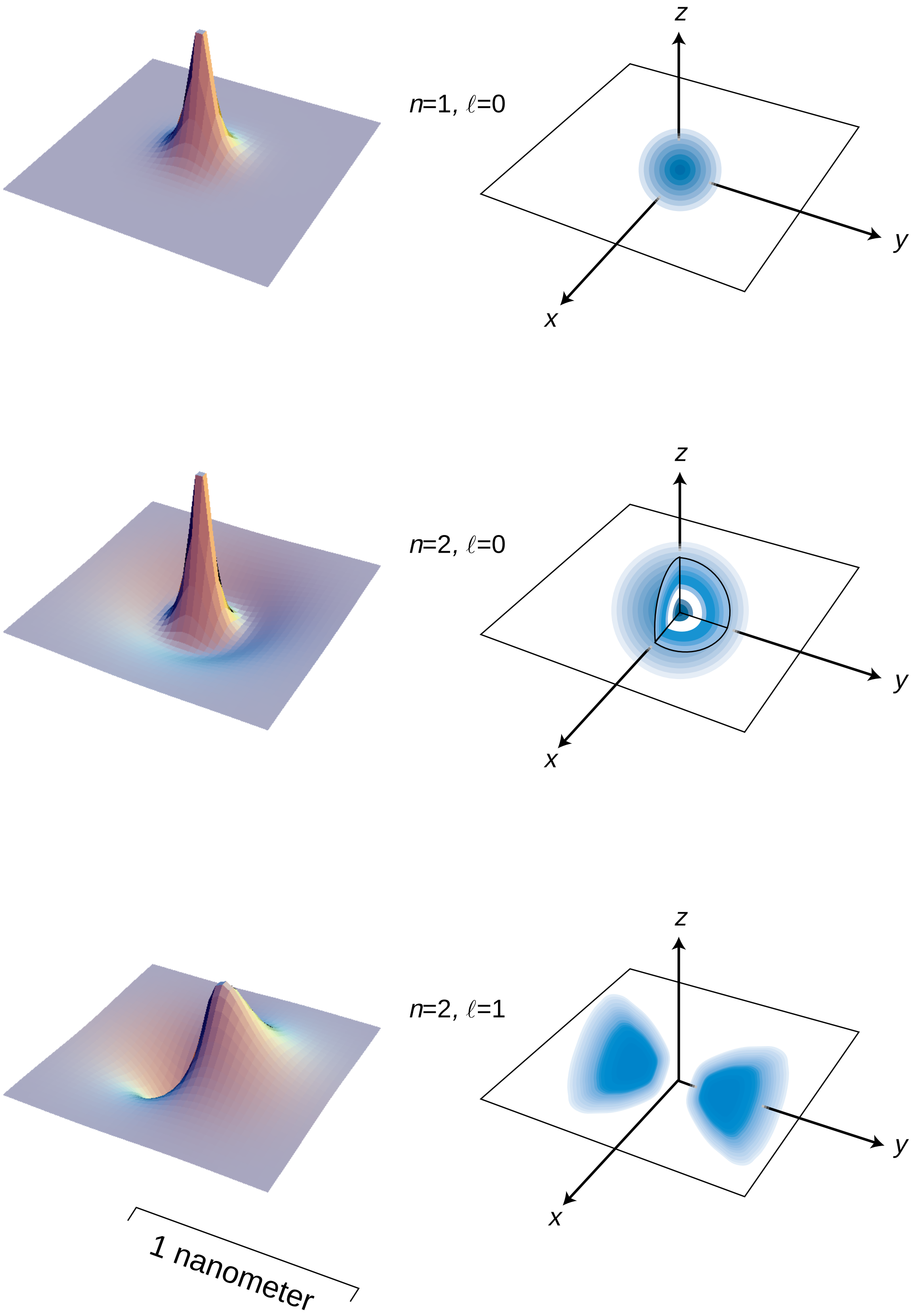
Continue the list for n=3.

The figures on the facing page show the lowest-energy states of the hydrogen atom. The left-hand column of graphs displays the wavefunctions in the x-y plane, and the right-hand column shows the probability density in a three-dimensional representation.

Discussion Questions

A. The quantum number n is defined as the number of radii at which the wavefunction is zero, including r=∞. Relate this to the features of the figures on the facing page.
B. Based on the definition of n, why can't there be any such thing as an n=0 state?
C. Relate the features of the wavefunction plots on the facing page to the corresponding features of the probability density pictures.
D. How can you tell from the wavefunction plots on the right which ones have which angular momenta?
E. Criticize the following incorrect statement: "The ℓ=8 wavefunction on the previous page has a shorter wavelength in the center because in the center the electron is in a higher energy level."
F. Discuss the implications of the fact that the probability cloud in of the n=2, ℓ=1 state is split into two parts.

n=3, ℓ=0, ℓ_z=0: one state; n=3, ℓ=1, ℓ_z=–1, 0, or 1: three states; n=3, ℓ=2, ℓ_z=–2, –1, 0, 1, or 2: five states


z
n=1, ℓ=0
y
x
z
n=2, ℓ=0
y
x
z
n=2, ℓ=1
y
x
1 nanometer

6.4* Energies of States in Hydrogen

The experimental technique for measuring the energy levels of an atom accurately is spectroscopy: the study of the spectrum of light emitted (or absorbed) by the atom. Only photons with certain energies can be emitted or absorbed by a hydrogen atom, for example, since the amount of energy gained or lost by the atom must equal the difference in energy between the atom's initial and final states. Spectroscopy had actually become a highly developed art several decades before Einstein even proposed the photon, and the Swiss spectroscopist Johann Balmer determined in 1885 that there was a simple equation that gave all the wavelengths emitted by hydrogen. In modern terms, we think of the photon wavelengths merely as indirect evidence about the underlying energy levels of the atom, and we rework Balmer's result into an equation for these atomic energy levels:

$$E_\mathrm{n} = -\frac{2.2\times10^{-18}\,\mathrm{J}}{n^2} \quad ,$$

where we have made use of the electron-volt (1 volt multiplied by the charge of the electron) as a convenient energy unit on the atomic scale. This energy includes both the kinetic energy of the electron and the electrical potential energy. The zero-level of the potential energy scale is chosen to be the energy of an electron and a proton that are infinitely far apart. With this choice, negative energies correspond to bound states and positive energies to unbound ones.

Where does the mysterious numerical factor of $2.2\mathrm{x}10^{-18}$ J come from? In 1913 the Danish theorist Niels Bohr realized that it was exactly numerically equal to a certain combination of fundamental physical constants:

$$E_\mathrm{n} = -\frac{mk^2e^4}{2\hbar^2}\cdot\frac{1}{n^2} \quad ,$$

where m is the mass of the electron, k is the Coulomb force constant for electric forces, and $\hbar$ is the abbreviation for $h/2\pi$ that we have already mentioned in passing.

Bohr was able to cook up a derivation of this equation based on the incomplete version of quantum physics that had been developed by that time, but his derivation is today mainly of historical interest. It assumes that the electron follows a circular path, whereas the whole concept of a path for a particle is considered meaningless in our more complete modern version of quantum physics. Although Bohr was able to produce the right equation for the energy levels, his model also gave various wrong results, such as predicting that the atom would be flat, and that the ground state would have ℓ =1 rather than the correct ℓ =0.

A full and correct treatment is impossible at the mathematical level of this book, but we can provide a straightforward explanation for the form of the equation using approximate arguments. A typical standing-wave pattern for the electron consists of a central oscillating area surrounded by a region in which the wavefunction tails off. As discussed in section 5.5, the oscillating type of pattern is typically encountered in the classically allowed region, while the tailing off occurs in the classically forbidden region where the electron has insufficient kinetic energy to penetrate according to classical physics. We use the symbol *r* for the radius of the spherical boundary between the classically allowed and classically forbidden regions.

When the electron is at the distance *r* from the proton, it has zero kinetic energy — in classical terms, this would be the distance at which the electron would have to stop, turn around, and head back toward the proton. Thus when the electron is at distance *r*, its energy is purely potential:

$$E = -\frac{ke^2}{r} \qquad (1)$$

Now comes the approximation. In reality, the electron's wavelength cannot be constant in the classically allowed region, but we pretend that it is. Since *n* is the number of nodes in the wavefunction, we can interpret it approximately as the number of wavelengths that fit across the diameter 2*r*. We are not even attempting a derivation that would produce all the correct numerical factors like 2 and π and so on, so we simply make the approximation

$$\lambda \sim \frac{r}{n} \quad . \qquad (2)$$

Finally we assume that the typical kinetic energy of the electron is on the same order of magnitude as the absolute value of its total energy. (This is true to within a factor of two for a typical classical system like a planet in a circular orbit around the sun.) We then have

$$\begin{aligned} &\text{absolute value of total energy} \\ &\quad = \frac{ke^2}{r} \\ &\quad \sim KE \\ &\quad = p^2/2m \\ &\quad = (h/\lambda)^2 / 2m \\ &\quad \sim hn^2 / 2mr^2 \qquad (3) \end{aligned}$$

We now solve the equation $ke^2/r{\sim}hn^2 / 2mr^2$ for *r* and throw away numerical factors we can't hope to have gotten right, yielding

$$r \sim \frac{h^2n^2}{mke^2} \qquad (4)$$

Plugging n=1 into this equation gives r=2 nm, which is indeed on the right order of magnitude. Finally we combine equations (4) and (1) to find

$$E \sim -\frac{mk^2e^4}{h^2n^2} \quad , \qquad (5)$$

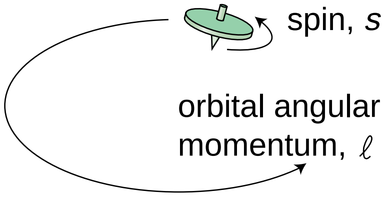
which is correct except for the numerical factors we never aimed to find.

Discussion Questions

A. States of hydrogen with *n* greater than about 10 are never observed in the sun. Why might this be?

B. Sketch graphs of *r* and *E* versus *n* for the hydrogen, and compare with analogous graphs for the one-dimensional particle in a box.

6.5 Electron Spin



The top has angular momentum both because of the motion of its center of mass through space and due to its internal rotation. Electron spin is roughly analogous to the intrinsic spin of the top.

It's disconcerting to the novice ping-pong player to encounter for the first time a more skilled player who can put spin on the ball. Even though you can't see that the ball is spinning, you can tell something is going on by the way it interacts with other objects in its environment. In the same way, we can tell from the way electrons interact with other things that they have an intrinsic spin of their own. Experiments show that even when an electron is not moving through space, it still has angular momentum amounting to $\hbar/2$.

This may seem paradoxical because the quantum moat, for instance, gave only angular momenta that were integer multiples of $\hbar$, not half-units, and I claimed that angular momentum was always quantized in units of $\hbar$, not just in the case of the quantum moat. That whole discussion, however, assumed that the angular momentum would come from the motion of a particle through space. The $\hbar/2$ angular momentum of the electron is simply a property of the particle, like its charge or its mass. It has nothing to do with whether the electron is moving or not, and it does not come from any internal motion within the electron. Nobody has ever succeeded in finding any internal structure inside the electron, and even if there was internal structure, it would be mathematically impossible for it to result in a half-unit of angular momentum.

We simply have to accept this $\hbar/2$ angular momentum, called the "spin" of the electron, as an experimentally proven fact. Protons and neutrons have the same $\hbar/2$ spin, while photons have an intrinsic spin of $\hbar$.

As was the case with ordinary angular momentum, we can describe spin angular momentum in terms of its magnitude, and its component along a given axis. The usual notation for these quantities, in units of $\hbar$, are s and s_z, so an electron has s=1/2 and s_z=+1/2 or –1/2.

Taking electron spin into account, we need a total of four quantum numbers to label a state of an electron in the hydrogen atom: n, ℓ , ℓ_z , and s_z. (We omit s because it always has the same value.) The symbols ℓ and ℓ_z include only the angular momentum the electron has because it is moving through space, not its spin angular momentum. The availability of two possible spin states of the electron leads to a doubling of the numbers of states:

n=1, ℓ =0, ℓ_z =0, s_z=+1/2 or –1/2	two states
n=2, ℓ =0, ℓ_z =0, s_z=+1/2 or –1/2	two states
n=2, ℓ =1, ℓ_z =–1, 0, or 1, s_z=+1/2 or –1/2	six states
...	

6.6 Atoms With More Than One Electron

What about other atoms besides hydrogen? It would seem that things would get much more complex with the addition of a second electron. A hydrogen atom only has one particle that moves around much, since the nucleus is so heavy and nearly immobile. Helium, with two, would be a mess. Instead of a wavefunction whose square tells us the probability of finding a single electron at any given location in space, a helium atom would need to have a wavefunction whose square would tell us the probability of finding two electrons at any given combination of points. Ouch! In addition, we would have the extra complication of the electrical interaction between the two electrons, rather than being able to imagine everything in terms of an electron moving in a static field of force created by the nucleus alone.

Despite all this, it turns out that we can get a surprisingly good description of many-electron atoms simply by assuming the electrons can occupy the same standing-wave patterns that exist in a hydrogen atom. The ground state of helium, for example, would have both electrons in states that are very similar to the n=1 states of hydrogen. The second-lowest-energy state of helium would have one electron in an n=1 state, and the other in an n=2 states. The relatively complex spectra of elements heavier than hydrogen can be understood as arising from the great number of possible combinations of states for the electrons.

A surprising thing happens, however, with lithium, the three-electron atom. We would expect the ground state of this atom to be one in which all three electrons settle down into n=1 states. What really happens is that two electrons go into n=1 states, but the third stays up in an n=2 state. This is a consequence of a new principle of physics:

The Pauli Exclusion Principle
Only one electron can ever occupy a given state.

There are two n=1 states, one with s_z=+1/2 and one with s_z=–1/2, but there is no third n=1 state for lithium's third electron to occupy, so it is forced to go into an n=2 state.

It can be proven mathematically that the Pauli exclusion principle applies to any type of particle that has half-integer spin. Thus two neutrons can never occupy the same state, and likewise for two protons. Photons, however, are immune to the exclusion principle because their spin is an integer.

Deriving the periodic table

We can now account for the structure of the periodic table, which seemed so mysterious even to its inventor Mendeleev. The first row consists of atoms with electrons only in the n=1 states:

I	II	III	IV	V	VI	VII	0
1 H							2 He
3 Li	4 Be	5 B	6 C	7 N	8 O	9 F	10 Ne
11 Na	...						

The beginning of the periodic table.

H	1 electron in an n=1 state
He	2 electrons in the two n=1 states

The next row is built by filling the n=2 energy levels:

Li	2 electrons in n=1 states, 1 electron in an n=2 state
Be	2 electrons in n=1 states, 2 electrons in n=2 states
...	
O	2 electrons in n=1 states, 6 electrons in n=2 states
F	2 electrons in n=1 states, 7 electrons in n=2 states
Ne	2 electrons in n=1 states, 8 electrons in n=2 states

In the third row we start in on the n=3 levels:

Na	2 electrons in n=1 states, 8 electrons in n=2 states, 1 electron in an n=3 state
...	

We can now see a logical link between the filling of the energy levels and the structure of the periodic table. Column 0, for example, consists of atoms with the right number of electrons to fill all the available states up to a certain value of n. Column I contains atoms like lithium that have just one electron more than that.

This shows that the columns relate to the filling of energy levels, but why does that have anything to do with chemistry? Why, for example, are the elements in columns I and VII dangerously reactive? Consider, for example, the element sodium (Na), which is so reactive that it may burst into flames when exposed to air. The electron in the n=3 state has an unusually high energy. If we let a sodium atom come in contact with an oxygen atom, energy can be released by transferring the n=3 electron from the sodium to one of the vacant lower-energy n=2 states in the oxygen. This energy is transformed into heat. Any atom in column I is highly reactive for the same reason: it can release energy by giving away the electron that has an unusually high energy.

Hydrogen is highly reactive. (Actually it has been claimed recently that what exploded in the Hindenburg disaster was aluminum powder on the outside of the blimp’s skin.)

Column VII is spectacularly reactive for the opposite reason: these atoms have a single vacancy in a low-energy state, so energy is released when these atoms steal an electron from another atom.

It might seem as though these arguments would only explain reactions of atoms that are in different rows of the periodic table, because only in these reactions can a transferred electron move from a higher-n state to a lower-n state. This is incorrect. An n=2 electron in fluorine (F), for example, would have a different energy than an n=2 electron in lithium (Li), due to the different number of protons and electrons with which it is interacting. Roughly speaking, the n=2 electron in fluorine is more tightly bound (lower in energy) because of the larger number of protons attracting it. The effect of the increased number of attracting protons is only partly counteracted by the increase in the number of repelling electrons, because the forces exerted on an electron by the other electrons are in many different directions and cancel out partially.

Summary

Selected Vocabulary

quantum number .... a numerical label used to classify a quantum state

spin .... the built-in angular momentum possessed by a particle even when at rest

Notation

n .... the number of radial nodes in the wavefunction, including the one at $r=\infty$

$\hbar$ .... $h/2\pi$

$\boldsymbol{L}$ .... the angular momentum vector of a particle, not including its spin

ℓ .... the magnitude of the $\boldsymbol{L}$ vector, divided by $\hbar$

ℓ_z .... the z component of the $\boldsymbol{L}$ vector, divided by $\hbar$; this is the standard notation in nuclear physics, but not in atomic physics

s .... the magnitude of the spin angular momentum vector, divided by $\hbar$

s_z .... the z component of the spin angular momentum vector, divided by $\hbar$; this is the standard notation in nuclear physics, but not in atomic physics

Notation Used in Other Books

m_ℓ .... a less obvious notation for ℓ_z , standard in atomic physics

m_s .... a less obvious notation for s_z, standard in atomic physics

Summary

Hydrogen, with one proton and one electron, is the simplest atom, and more complex atoms can often be analyzed to a reasonably good approximation by assuming their electrons occupy states that have the same structure as the hydrogen atom's. The electron in a hydrogen atom exchanges very little energy or angular momentum with the proton, so its energy and angular momentum are nearly constant, and can be used to classify its states. The energy of a hydrogen state depends only on its n quantum number.

In quantum physics, the angular momentum of a particle moving in a plane is quantized in units of $\hbar$. Atoms are three-dimensional, however, so the question naturally arises of how to deal with angular momentum in three dimensions. In three dimensions, angular momentum is a vector in the direction perpendicular to the plane of motion, such that the motion appears clockwise if viewed along the direction of the vector. Since angular momentum depends on both position and momentum, the Heisenberg uncertainty principle limits the accuracy with which one can know it. The most the can be known about an angular momentum vector is its magnitude and one of its three vector components, both of which are quantized in units of $\hbar$.

In addition to the angular momentum that an electron carries by virtue of its motion through space, it possesses an intrinsic angular momentum with a magnitude of $\hbar$ /2. Protons and neutrons also have spins of $\hbar$ /2, while the photon has a spin equal to $\hbar$.

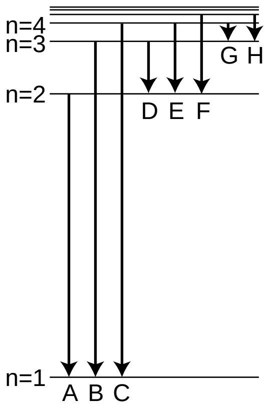
Particles with half-integer spin obey the Pauli exclusion principle: only one such particle can exist is a given state, i.e. with a given combination of quantum numbers.

We can enumerate the lowest-energy states of hydrogen as follows:

n=1, ℓ =0, ℓ_z =0, s_z=+1/2 or –1/2	two states
n=2, ℓ =0, ℓ_z =0, s_z=+1/2 or –1/2	two states
n=2, ℓ =1, ℓ_z =–1, 0, or 1, s_z=+1/2 or –1/2	six states
...	

The periodic table can be understood in terms of the filling of these states. The nonreactive noble gases are those atoms in which the electrons are exactly sufficient to fill all the states up to a given n value. The most reactive elements are those with one more electron than a noble gas element, which can release a great deal of energy by giving away their high-energy electron, and those with one electron fewer than a noble gas, which release energy by accepting an electron.

Homework Problems



Problem 2.

1. (a) A distance scale is shown below the wavefunctions and probability densities illustrated in section 6.3. Compare this with the order-of-magnitude estimate derived in section 6.4 for the radius *r* at which the wavefunction begins tailing off. Was the estimate in section 6.4 on the right order of magnitude? (b) Although we normally say the moon orbits the earth, actually they both orbit around their common center of mass, which is below the earth's surface but not at its center. The same is true of the hydrogen atom. Does the center of mass lie inside the proton or outside it?

2. The figure shows eight of the possible ways in which an electron in a hydrogen atom could drop from a higher energy state to a state of lower energy, releasing the difference in energy as a photon. Of these eight transitions, only D, E, and F produce photons with wavelengths in the visible spectrum. (a) Which of the visible transitions would be closest to the violet end of the spectrum, and which would be closest to the red end? Explain. (b) In what part of the electromagnetic spectrum would the photons from transitions A, B, and C lie? What about G and H? Explain. (c) Is there an upper limit to the wavelengths that could be emitted by a hydrogen atom going from one bound state to another bound state? Is there a lower limit? Explain.

3. Before the quantum theory, experimentalists noted that in many cases, they would find three lines in the spectrum of the same atom that satisfied the following mysterious rule: $1/\lambda_1=l/\lambda_2+1/\lambda_3$. Explain why this would occur. Do not use reasoning that only works for hydrogen — such combinations occur in the spectra of all elements. [Hint: Restate the equation in terms of the energies of photons.]

4. Find an equation for the wavelength of the photon emitted when the electron in a hydrogen atom makes a transition from energy level n_1 to level n_2. [You will need to have read optional section 6.4.]

5. (a) Verify that Planck's constant has the same units as angular momentum. (b) Estimate the angular momentum of a spinning basketball, in units of $\hbar$.

6. Assume that the kinetic energy of an electron the n=1 state of a hydrogen atom is on the same order of magnitude as the absolute value of its total energy, and estimate a typical speed at which it would be moving. (It cannot really have a single, definite speed, because its kinetic and potential energy trade off at different distances from the proton, but this is just a rough estimate of a typical speed.) Based on this speed, were we justified in assuming that the electron could be described nonrelativistically?

S A solution is given in the back of the book. ★ A difficult problem.
✓ A computerized answer check is available. ∫ A problem that requires calculus.

7. The wavefunction of the electron in the ground state of a hydrogen atom is

$$\Psi = 2\, a^{-3/2}\, e^{-r/a}$$

where r is the distance from the proton, and $a=\hbar^2/kme^2 = 5.3\text{x}10^{-11}$ m is a constant that sets the size of the wave.

(a) Calculate symbolically, without plugging in numbers, the probability that at any moment, the electron is inside the proton. Assume the proton is a sphere with a radius of b=0.5 fm. [Hint: Does it matter if you plug in r=0 or r=b in the equation for the wavefunction?]

(b) Calculate the probability numerically.

(c) Based on the equation for the wavefunction, is it valid to think of a hydrogen atom as having a finite size? Can a be interpreted as the size of the atom, beyond which there is nothing? Or is there any limit on how far the electron can be from the proton?

8 ★. Use physical reasoning to explain how the equation for the energy levels of hydrogen,

$$E_n = -\frac{mk^2e^4}{2\hbar^2}\cdot\frac{1}{n^2} \quad ,$$

should be generalized to the case of a heavier atom with atomic number Z that has had all its electrons stripped away except for one.

9. This question requires that you read optional section 6.4. A muon is a subatomic particle that acts exactly like an electron except that its mass is 207 times greater. Muons can be created by cosmic rays, and it can happen that one of an atom's electrons is displaced by a muon, forming a muonic atom. If this happens to a hydrogen atom, the resulting system consists simply of a proton plus a muon. (a) How would the size of a muonic hydrogen atom in its ground state compare with the size of the normal atom? (b) If you were searching for muonic atoms in the sun or in the earth's atmosphere by spectroscopy, in what part of the electromagnetic spectrum would you expect to find the absorption lines?

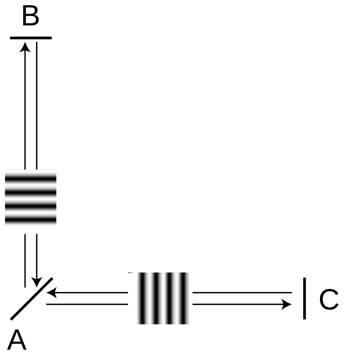
10. Consider a classical model of the hydrogen atom in which the electron orbits the proton in a circle at constant speed. In this model, the electron and proton can have no intrinsic spin. Using the result of problem 17 from book 4, ch. 6, show that in this model, the atom's magnetic dipole moment D_m is related to its angular momentum by $D_m=(-e/2m)L$, regardless of the details of the orbital motion. Assume that the magnetic field is the same as would be produced by a circular current loop, even though there is really only a single charged particle. [Although the model is quantum-mechanically incorrect, the result turns out to give the correct quantum mechanical value for the contribution to the atom's dipole moment coming from the electron's orbital motion. There are other contributions, however, arising from the intrinsic spins of the electron and proton.]

Exercises

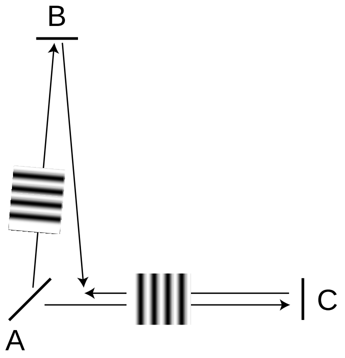
Ex. 1A: The Michelson-Morley Experiment

In this exercise you will analyze the Michelson-Morley experiment, and find what the results should have been according to Galilean relativity and Einstein's theory of relativity. A beam of light coming from the west (not shown) comes to the half-silvered mirror A. Half the light goes through to the east, is reflected by mirror C, and comes back to A. The other half is reflected north by A, is reflected by B, and also comes back to A. When the beams reunite at A, part of each ends up going south, and these parts interfere with one another. If the time taken for a round trip differs by, for example, half the period of the wave, there will be destructive interference.

The point of the experiment was to search for a difference in the experimental results between the daytime, when the laboratory was moving west relative to the sun, and the nighttime, when the laboratory was moving east relative to the sun. Galilean relativity and Einstein's theory of relativity make different predictions about the results. According to Galilean relativity, the speed of light cannot be the same in all reference frames, so it is assumed that there is one special reference frame, perhaps the sun's, in which light travels at the same speed in all directions; in other frames, Galilean relativity predicts that the speed of light will be different in different directions, e.g. slower if the observer is chasing a beam of light. There are four different ways to analyze the experiment:



laboratory's x,t frame of reference



sun's x',t' frame (lab moving to the right)

1. *Laboratory's frame of reference, Galilean relativity*. This is not a useful way to analyze the experiment, since one does not know how fast light will travel in various directions.

2. *Sun's frame of reference, Galilean relativity*. We assume that in this special frame of reference, the speed of light is the same in all directions: we call this speed c. In this frame, the laboratory moves with velocity v, and mirrors A, B, and C move while the light beam is in flight.

3. *Laboratory's frame of reference, Einstein's theory of relativity*. The analysis is extremely simple. Let the length of each arm be L. Then the time required to get from A to either mirror is L/c, so each beam's round-trip time is $2L/c$.

4. *Sun's frame of reference, Einstein's theory of relativity*. We analyze this case by starting with the laboratory's frame of reference and then transforming to the sun's frame.

Groups 1-4 work in the sun's frame of reference according to Galilean relativity.

Group 1 finds time AC. Group 2 finds time CA. Group 3 finds time AB. Group 4 finds time BA.

Groups 5 and 6 transform the lab-frame results into the sun's frame according to Einstein's theory.

Group 5 transforms the x and t when ray ACA gets back to A into the sun's frame of reference, and group 6 does the same for ray ABA.

Discussion:

Michelson and Morley found no change in the interference of the waves between day and night. Which version of relativity is consistent with their results?

What does each theory predict if v approaches c?

What if the arms are not exactly equal in length?

Does it matter if the "special" frame is some frame other than the sun's?

Ex. 2A: Sports in Slowlightland

In Slowlightland, the speed of light is 20 mi/hr = 32 km/hr = 9 m/s. Think of an example of how relativistic effects would work in sports. Things can get very complex very quickly, so try to think of a simple example that focuses on just one of the following effects:

- relativistic momentum
- relativistic kinetic energy
- relativistic addition of velocities
- time dilation and length contraction
- Doppler shifts of light
- equivalence of mass and energy
- time it takes for light to get to an athlete’s eye
- deflection of light rays by gravity

Ex. 6A: Quantum Versus Classical Randomness

1. Imagine the *classical* version of the particle in a one-dimensional box. Suppose you insert the particle in the box and give it a known, predetermined energy, but a random initial position and a random direction of motion. You then pick a random later moment in time to see where it is. Sketch the resulting probability distribution by shading on top of a line segment. Does the probability distribution depend on energy?

2. Do similar sketches for the first few energy levels of the *quantum mechanical* particle in a box, and compare with 1.

3. Do the same thing as in 1, but for a classical hydrogen atom in two dimensions, which acts just like a miniature solar system. Assume you're always starting out with the same fixed values of energy and angular momentum, but a position and direction of motion that are otherwise random. Do this for L=0, and compare with a real L=0 probability distribution for the hydrogen atom.

4. Repeat 3 for a nonzero value of L, say L=$\hbar$.

5. Summarize: Are the classical probability distributions accurate? What qualitative features are possessed by the classical diagrams but not by the quantum mechanical ones, or vice-versa?

Solutions to Selected Problems

Chapter 2

11. (a) The factor of 2 comes from the reversal of the direction of the light ray's momentum. If we pick a coordinate system in which the force on the surface is in the positive direction, then $\Delta p = (-p)-p = -2p$. The question doesn't refer to any particular coordinate system, and is only talking about the magnitude of the force, so let's just say $\Delta p=2p$. The force is $F=\Delta p/\Delta t=2p/\Delta t=2E/c\Delta t=2P/c$. (b) $mg=2P/c$, so $m=2P/gc=70$ nanograms.

12.

$$a = \frac{\mathit{force}}{(\mathit{mass\ of\ payload}) + (\mathit{mass\ of\ sail})}$$

$$= \frac{2(\mathit{flux})(\mathit{area})\ /\ c}{(\mathit{mass\ of\ payload}) + (\mathit{area})(\mathit{thickness})(\mathit{density})}$$

$$= \frac{2(1400\ W/\text{m}^2)(600\ \text{m}^2)\ /\ (3.0\times10^8\ \text{m/s})}{(40\ kg) + (600\ \text{m}^2)(5\times10^{-6}\ \text{m})(1.40\times10^3\ \text{kg/m}^3)}$$

$= 1.3\text{x}10^{-4}\ \text{m/s}^2$

Glossary

FWHM. The full width at half-maximum of a probability distribution; a measure of the width of the distribution.

Half-life. The amount of time that a radioactive atom has a probability of 1/2 of surviving without decaying.

Independence. The lack of any relationship between two random events.

Invariant. A quantity that does not change when transformed.

Lorentz transformation. The transformation between frames in relative motion.

Mass. What some books mean by "mass" is our *m*g.

Normalization. The property of probabilities that the sum of the probabilities of all possible outcomes must equal one.

Photon. A particle of light.

Photoelectric effect. The ejection, by a photon, of an electron from the surface of an object.

Probability. The likelihood that something will happen, expressed as a number between zero and one.

Probability distribution. A curve that specifies the probabilities of various random values of a variable; areas under the curve correspond to probabilities.

Quantum number. A numerical label used to classify a quantum state.

Rest mass. Referred to as mass in this book; written as m_0 in some books.

Spin. The built-in angular momentum possessed by a particle even when at rest.

Transformation. The mathematical relationship between the variables such as *x* and *t*, as observed in different frames of reference.

Wave-particle duality. The idea that light is both a wave and a particle.

Wavefunction. The numerical measure of an electron wave, or in general of the wave corresponding to any quantum mechanical particle.

Index

Photo Credits

All photographs are by Benjamin Crowell, except as noted below. In some cases I have used historical photographs for which I know the date the picture was taken but not the photographer; in these cases I have simply listed the dates, which indicate the copyrights have expired. I will be grateful for any information that helps me to credit the photographers properly.

Chapter 1
Einstein: ca. 1905

Chapter 2
Eclipse: 1919
Large Hadron Collider: Courtesy of CERN.

Chapter 3
Mount St. Helens: Public-domain image by Austin Post, USGS.
Pu'u O'o: Public-domain image by Lyn Topinka, USGS.

Chapter 4
Ozone maps: NASA/GSFC TOMS Team.
Photon interference photos: Lyman Page.

Chapter 5
Wicked witch: W.W. Denslow, 1900. Quote from The Wizard of Oz, by L. Frank Baum, 1900.

Chapter 6
Hindenburg: Arthur Cofod Jr., U.S. Air Force, 1937, courtesy of the National Air and Space Museum Archives.

Useful Data

Metric Prefixes

M-	mega-	10^{6}
k-	kilo-	10^{3}
m-	milli-	10^{-3}
μ- (Greek mu)	micro-	10^{-6}
n-	nano-	10^{-9}
p-	pico-	10^{-12}
f-	femto-	10^{-15}

(Centi-, 10^{-2}, is used only in the centimeter.)

Notation and Units

quantity	unit	symbol
distance	meter, m	x, Δx
time	second, s	t, Δt
mass	kilogram, kg	m
density	kg/m^{3}	ρ
force	newton, 1 N=1 kg·m/s^{2}	$\boldsymbol{F}$
velocity	m/s	$\boldsymbol{v}$
acceleration	m/s^{2}	$\boldsymbol{a}$
energy	joule, J	E
momentum	kg·m/s	$\boldsymbol{p}$
angular momentum	kg·m^{2}/s	L
period	s	T
wavelength	m	λ
frequency	s^{-1} or Hz	f
focal length	m	f
magnification	unitless	M
index of refraction	unitless	n

Fundamental Constants

gravitational constant	G=6.67x10^{-11} N·m^{2}/kg^{2}
Coulomb constant	k=8.99x10^{9} N·m^{2}/C^{2}
quantum of charge	e=1.60x10^{-19} C
speed of light	c=3.00x10^{8} m/s
Planck's constant	h=6.63x10^{-34} J·s

Conversions

Conversions between SI and other units:

1 inch	=	2.54 cm (exactly)
1 mile	=	1.61 km
1 pound	=	4.45 N
(1 kg)·g	=	2.2 lb
1 gallon	=	3.78x10^{3} cm^{3}
1 horsepower	=	746 W
1 kcal*	=	4.18x10^{3} J

*When speaking of food energy, the word "Calorie" is used to mean 1 kcal, i.e. 1000 calories. In writing, the capital C may be used to indicate 1 Calorie=1000 calories.

Conversions between U.S. units:

1 foot	=	12 inches
1 yard	=	3 feet
1 mile	=	5280 ft

Some Indices of Refraction

substance	index of refraction
vacuum	1 by definition
air	1.0003
water	1.3
glass	1.5 to 1.9
diamond	2.4

Note that all indices of refraction depend on wavelength. These values are about right for the middle of the visible spectrum (yellow).

Subatomic Particles

particle	mass (kg)	charge	radius (fm)
electron	9.109x10^{-31}	$-e$	<~0.01
proton	1.673x10^{-27}	$+e$	~1.1
neutron	1.675x10^{-27}	0	~1.1
neutrino	~10^{-39} kg?	0	?

The radii of protons and neutrons can only be given approximately, since they have fuzzy surfaces. For comparison, a typical atom is about a million fm in radius.